Structural Studies of a Novel Adhesin Involved in Host-specific *Lactobacillus reuteri* Biofilm Formation in the Vertebrate Gastrointestinal Tract

AND

Purification Method Development for Structural Studies of WaaP: an Essential Gram-Negative Bacteria Lipopolysaccharide Kinase

Saannya Sequeira, B.Sc., M.Chem.

Thesis submitted for the degree of Doctor of Philosophy

TO

The University of East Anglia

Norwich Medical School
Faculty of Medicine and Health Sciences

September 2016

ABSTRACT

Structural Studies of a Novel Adhesin Involved in Host-specific *Lactobacillus reuteri* Biofilm Formation in the Vertebrate Gastrointestinal Tract

In vertebrates, strains of the gut symbiont, *Lactobacillus reuteri,* form biofilms on the proximal gastric stratified epithelium in a host-dependent manner. The gene for a novel serine-rich repeat (SRR) adhesin, conserved amongst rat and porcine *L. reuteri*, is essential for successful biofilm formation, suggesting initial adhesion to be the most crucial step for colonisation, and a possible driver of host-specificity. High resolution crystal structures of the partial binding regions (BR) of two SRR adhesins, BR55 and BR78, from porcine *L. reuteri* ATCC 53608 and rodent *L. reuteri* 100-23 strains, respectively, are reported. Although 50% similar, the BRs were highly conserved in overall structure and in the orientation of putative binding residues. BR55 and BR78's right-handed parallel β-sheet solenoid-type fold (typical of pectate lyase like proteins), and their aromatic residue-rich putative binding cleft indicated that glycans were the most likely binding substrates. Potential binding residues were identified, allowing mutagenesis for functional studies. SAXS analysis of the complete BR55 protein revealed two regions of intrinsic disorder at the N- and C-terminal, beyond the ordered β-solenoid core. Structural information from this investigation provided valuable insights for ongoing biochemical analyses.

Purification Method Development for Structural Studies of WaaP: an Essential Gram-Negative Bacteria Lipopolysaccharide (LPS) Kinase

LPS is a vital component of Gram-negative bacteria outer membrane (OM), playing a crucial role in maintaining OM integrity against host-immune defenses and antibiotics. In turn, phosphorylation of the LPS core oligosaccharide is essential for structural viability of LPS in *E. coli, S. enterica* and *P. aeruginosa.* The *waaP* gene, encoding a Kdo sugar kinase, is a proven virulence factor, essential for phosphorylating LPS. Furthermore, *waaP* is conserved amongst many Gram-negative pathogens, and evokes interest as a potential antibiotic target. This thesis reports purification method development of recombinant WaaP for structural studies, and proposes strategies for accessing other WaaP homologues for crystallography. *P. aeruginosa* WaaP was successfully purified in adequate amounts for crystallography.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My sincere thanks go to my principle advisor and supervisor, Prof. Changjiang Dong for his support and guidance through my PhD; also to Dr. Nathalie Juge from the Institute of Food Research, for an enjoyable collaboration, and for her prompt and sincere support in times when I needed it. I would also like to extend my thanks to other members of the Juge group at IFR for their support: Dr. Donald MacKenzie and Dr. Devon Kavanaugh, and to Prof. Tom Wileman for his supervision and being a source of reassurance and focus.

Next, I must thank the Trustees of the Gen Foundation, London, for their generous grant which helped in the continuation of my research.

I must not forget to thank Dr. Katsuaki Inoue, from the Diamond Light Source in Didcot, for coaching me through the basics of SAXS and for being receptive to the hundreds of questions I had, often at short notice! I would also like to thank Dr. Tom Clarke and Dr. Marcus Edwards for their cordial advice and support which helped me in my understanding of structural biology.

I am lucky to have had an opportunity to do more than a PhD during my time at UEA; for this, I would like to thank the Committee Members of the Faculty of Science Enterprise and Engagement Club, especially Nicolle and Lukas, the staff at UEA Outreach, and the BIO academic staff with whom I enjoyed demonstrating, Richard Bowater, Helen James, Harriet Jones and Kay Yeoman.

Then, to all the special people in my life who were rays of sunshine on the ever so many cloudy days I encountered during this challenging journey. My parents, who always encouraged my dreams, taught me to stay strong and offered their unconditional love during stressful times. My one-of-a-kind siblings, Sarah, Sharic and Sabby, without whom life would be dull. And the friends who were my family away from home: Saira, Bintu, Sonia, Sherly and Sumera – I cannot thank you enough for your care and support, the kind I could only expect from family. I would also like to thank Ellie, Monika, Saima, Kolja, Aya, Beth, Amina, Meanhaa, Hifsa and Tesbiha for their friendship and all the fun times.

Finally, I want to thank my husband, Loqmane: for being an anchor in turbulent times and for being the best of comforts any person could ask for.

ABBREVIATIONS

A_{260}	Absorbance at 260 nm
A_{280}	Absorbance at 280 nm
Abs	Absorbance
ACN	Acetonitrile
ADP	Adenosine Diphosphate
ATP	Adenosine Triphosphate
AUC	Analytical Ultracentrifugation
bp	Base pair(s)
BR	Binding region
D_{max}	Maximum particle dimension
DTT	Dithiothreitol
ECM	Extracellular matrix
EOM	Ensemble Optimisation Method
EPS	Exopolysaccharide
EtBr	Ethidium Bromide
GIT	Gastrointestinal Tract
GST	Glutathione S-transferase
HRV	Human rhinovirus
IM	Inner Membrane
IMAC	Immobilised Metal Affinity Chromatography
IPTG	Isopropyl β-D-thiogalactopyranoside
KT10	Keratin 10
LB	Luria Bertani
LDAO	Lauryldimethylamine-N-Oxide
LPS	Lipopolysaccharide
MALDI-MS	Matrix Assisted Laser Desorption Ionisation Mass Spectrometry
MBP	Maltose Binding Protein
MCS	Multiple cloning site
MWCO	Molecular Weight Cut-off
Ni-NTA	Nickel Nitrilotriacetic acid
OD_{600}	Optical Density at 600 nm
OM	Outer Membrane
PBS	Phosphate Buffered Saline

PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PEG	Polyethylene Glycol
PGA	Polygalacturonic Acid
pI	Isoelectric Point
R_g	Radius of gyration
SAD	Single-wavelength Anomalous Dispersion
SAXS	Small Angle X-ray Scattering
SDS-PAGE	Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis
SEC	Size-Exclusion Chromatography
SeMet	L-Selenomethionine
SRR	Serine-rich repeat or Serine-rich region
SRRP	Serine Rich Repeat Protein
TAE	Tris acetate EDTA
TB	Terrific Broth
TCEP	Tris(2-carboxyethyl)phosphine
TEV	Tobacco etch virus
TGA	Trigalacturonic Acid
XRD	X-Ray Diffraction
β-ME	β-mercaptoethanol

Recombinant gene/protein constructs

BR55	Binding region for the Lr0906 SRRP
BR78	Binding region for the Lr70902 SRRP
CFWaaP	*Citrobacter freundii* WaaP gene or protein
ECWaaP	*Escherichia coli* WaaP gene or protein
PAWaaP	*Pseudomonas aeruginosa* WaaP gene or protein
PFWaaP	*Pseudomonas fluorescens* WaaP gene or protein
PSWaaP	*Pseudomonas syringae* WaaP gene or protein
SEnWaaP	*Salmonella enterica* WaaP gene or protein

CHAPTER 1

GENERAL METHODS

1 GENERAL METHODS

This thesis describes experimental findings from two research projects. The first aims to provide a molecular understanding of the role of a novel *Lactobacillus reuteri* Serine-rich-repeat protein (SRRP) adhesin in host-driven biofilm formation; the second project details the method development for producing recombinant WaaP protein at adequate levels for structural studies. Similar molecular biology and protein biochemistry methods were employed for both projects, so they will be described together in this section. Where special conditions were used for an experiment they will be presented with the results in relevant sections of the thesis.

1.1 Materials

1.1.1 Recombinant plasmids, genomic DNA and vectors

1.1.1.1 For structural studies of *L. reuteri* SRRP

Domain boundaries for BR55 and BR78 gene inserts, representing the putative binding regions of *L. reuteri* SRRPs Lr0906 [3] and Lr70902 [4], respectively, were identified by Donald MacKenzie from the Institute of Food Research – IFR (Norwich, UK) and cloned into the pOPINF vector [5] for expression with a 3C-protease cleavable N-terminal 6-His tag. The primary sequence of Lr0906 and Lr70902 are provided in APPENDIX 2. The pOPINF vector map is given in APPENDIX 3.

1.1.1.2 For structural studies of WaaP proteins

Genomic DNA for *E. coli* K-12 (EC), *Pseudomonas aeruginosa* PAO1 (PA) and *Citrobacter freundii* ATCC 8090 (CF) were obtained from the genomic DNA collection in Changjiang Dong's laboratory (University of East Anglia, Norwich, UK). *Pseudomonas fluroscens* SBW25 (PF) and *Pseudomonas syringae* DC3000 (PS) genomic DNA was kindly given by Jacob Malone (John Innes Centre, Norwich, UK).

The **pET28 b** vector was purchased from Novagen (USA). The **pLou3 vector** [6] was originally provided by Louise Major (University of St. Andrews, St. Andrews, UK); it was derived from the New England Biolabs (NEB) vector pMAL-c2X for the over-expression of maltose-binding protein (MBP) fusion protein. Novel features of the pLou3 plasmid included an N-terminal 6xHis tag before MBP and a TEV protease cleavage site between MBP and the multiple cloning site (MCS). Plasmid maps of pET28 b and pLou3 are provided in APPENDIX 3. The **pWQ911 plasmid**, consisting of the Gro*ESL* chaperonin from *E. coli* W3110, cloned into the pBAD33 vector, was kindly provided by Chris Whitfield (University of Guelph, Ontario, Canada) [7]. The **ECWaaP_pLou3** and **SEnWaaP_pLou3** plasmids were constructed by Zhongshan Wang (Sichuan University, Chengdu, China), in the Dong laboratory at UEA (Norwich, UK). PCR amplification of WaaP fragments were from *Escherichia coli* K-12 (EC) and *Salmonella enterica* (SEn) genomic DNA, respectively.

1.1.2 Cell Culture Media

1.1.2.1 Media Composition for growth of *E. coli* cells

1) BR55 (wild type and mutants) and BR78: For protein expression, high salt Luria-Bertani (LB) broth consisting of 10 g/L tryptone, 5 g/L yeast extract and 10 g/L NaCl (Melford, Ipswich, UK) was autoclaved before being supplemented with 1% w/v glucose and 100 µg/mL ampicillin (Formedium, Hunstanton, UK). 1.5% LB-agar (Melford, Ipswich, UK) supplemented with 100 µg/mL ampicillin was used for the growth of single colonies of new transformants.

2) WaaP proteins: Unless specified, for recombinant protein production and growth of new transformants, LB broth or 1.5% LB-agar media were used, respectively. For C-6His-PAWaaP_C43(DE3) and C-6His-ECWaaP_C43(DE3) growth, Modified Terrific Broth (TB Broth) (Melford, Ipswich, UK) was used, consisting of 12 g/L casein digest peptone, 24 g/L yeast extract, 9.4 g/L dipotassium phosphate and 2.2 g/L monopotassium phosphate, and supplemented with 0.4% glycerol (Sigma-Aldrich,

Dorset, UK), prior to autoclaving. All media were supplemented with the appropriate antibiotics (Formedium, Hunstanton, UK) prior to use (see **Table 1.2**).

1.1.2.2 Minimal Media Composition for Expression of L-Selenomethionine (Se-Met) Labelled BR55

A detailed description of stock solution compositions for the preparation of minimal media is given in APPENDIX 1. The minimal media consisted of M9 salts solution, 2% w/v glucose, 0.03% w/v $MgSO_4$, 0.001% w/v $Fe_2(SO_4)_3$, 0.001% w/v thiamine-HCl, 0.0011% SelenoMethionine nutrient mix, 0.01% w/v L-lysine, 0.01% w/v L-phenylalanine, 0.01% w/v L-threonine, 0.005% w/v L-leucine, 0.005% w/v L-isoleucine, 0.005% w/v L-valine and 1% w/v SeMe, and was supplemented with 100 µg/mL ampicillin before culture growth.

1.1.3 Sample Storage

24% w/v glycerol cell stocks of all *E. coli* expression strains were prepared from saturated liquid LB cultures and flash frozen in liquid nitrogen for long term storage at -80^{o}C. Protein solutions for crystallisation were concentrated in their respective buffers to 10 mg/mL (unless specified) after size-exclusion chromatography (SEC) and flash frozen in liquid nitrogen in aliquots of 50 µL for storage at -80^{o}C.

1.1.4 Protease Stocks

Human rhinovirus (HRV) 3C protease was previously purified by immobilised metal affinity chromatography (IMAC) and size-exclusion chromatography (SEC), as a Glutathione S-transferase (3C-GST) fusion protein. It was stored at 1 mg/mL as 50% glycerol stocks and 1 mL aliquots. Recombinant Tobacco etch virus (TEV) protease, purified routinely in the Dong laboratory by IMAC, followed by SEC, was stored in concentrations of 2.5 to 3 mg/mL, as 50% glycerol stocks in 500 µL aliquots. TEV protease was prepared in 20 mM Tris /pH 8.0, 0.3 M NaCl buffer; the buffer composition of 3C-GST protease is unknown.

1.2 Molecular Biology

1.2.1 Polymerase Chain Reaction (PCR) for WaaP gene insert amplification

30 µL PCR reactions consisted of 500 µM of each primer, 200 µM of each dNTP, <2ng DNA template and 0.3 µL Q5 Hot-Start High-Fidelity DNA Polymerase, with its corresponding buffer. PCR was carried out in the T100^{TM} Thermal Cycler with the following steps: initial template denaturation at 98^{o}C for 30 sec; 33 amplification cycles, each consisting of denaturation at 98 ^{o}C for 10 sec, 20 sec annealing at the temperatures in **Table 1.1** and extension at 72^{o}C for 45 sec; this was followed by two cycles where the annealing temperature was 5^{o}C lower than those in **Table 1.1**, and then a final extension at 72^{o}C for 5 min. The PCR products were purified by agarose gel electrophoresis (see 1.2.6) and were then excised from the gel and isolated using the GeneJET Gel Extraction Kit (K0692, Thermo Scientific, UK), as per the manufacturer's instructions.

Table 1.1: PCR primer pairs for the amplification of WaaP genes from genomic DNA of selected Gram-negative bacteria species.

Sequences that belong to a restriction site are shown in **bold.** The start codon is indicated in green and the stop codon in red.

Construct	Vector		Restriction site	Primer Sequence (5' to 3')	T_m (°C)	Annealing Temp. (°C)
C-6His-PAWaaP	pET28b	*Fwd*	*Nco*I	CGC**CCATGG**TGAGGCTGGTGCTGGAAGAG	72	72
		Rev	*Not*I	**AATAGCGGCCGCCAGGTCTCCGTAACGCTGCTT**	83	
PAWaaP	pLou3	*Fwd*	*Eco*RI	**CGCTGAATTCATGAGGCTGGTGCTGGAAGAG**	68	71
		Rev	*Hin*dIII	**AATTAAGCTTTCACAGGTCTCCGTAACGCTGCTT**	68	
C-6His-ECWaaP	pET28b	*Fwd*	*Nco*I	CGCAA**CCATGG**TTGAATTAAAAGAGCCGCT	63	64
		Rev	*Xho*I	TTAAG**CTCGAG**TAATCCTTTGCGTTGTGTTCG	61	
PFWaaP	pLou3	*Fwd*	*Bam*HI	TTCCAA**GGATCC**ATGAAATTGATTCTTGCCGAACCG	64	67
		Rev	*Hin*dIII	AACCTT**AAGCTT**TCAGATCGCGTCCCCATAAC	65	
PSWaaP	pLou3	*Fwd*	*Bam*HI	TTCCAA**GGATCC**ATGAAGTTGTTCCTTGCTGAACCG	66	69
		Rev	*Hin*dIII	AACCTT**AAGCTT**TCAGAGCGCATCCCCGTAG	68	
CFWaaP	pLou3	*Fwd*	*Bam*HI	AACCAA**GGATCC**ATGATTGAATTAAAAGAACCATTT	54	57
		Rev	*Hin*dIII	AACCTT**AAGCTT**TTATAGCGCTTTTCTAATTGTTCT	58	

Table 1.2: Selected features of expression vectors used for *L. reuteri* SRRP BR and WaaP recombinant protein expression

Expression Vector	Origin of replication	Inducer	Antibiotic resistance
pOPINF	pUC	IPTG	Ampicillin
pET28b	pBR322	IPTG	Kanamycin
pLou3	pBR322	IPTG	Ampicillin
pWQ911	pACYC	L-arabinose	Chloramphenicol

1.2.2 Ligation Cloning of WaaP gene inserts into expression vectors

15 µL double-restriction reactions for amplified constructs and the vectors were set up using the respective FastDigest restriction enzymes (Thermo Scientific, UK) given in **Table 1.1**. 0.5 µL of each enzyme and 1.5 µL of FastDigest universal buffer was mixed with 12.5 µL of DNA solution and incubated at 37^{o}C for 1 hour. The enzymes were heat killed at 65^{o}C for 20 min, and 10 µL ligation reactions, consisting of 1 µL T4 DNA Ligase, 1 µL 10x T4 DNA Ligase Buffer (Thermo Scientific, UK) and 8 µL of the heat-killed 1:3 vector: insert mixture was incubated at 15^{o}C O/N. The ligation products were then transformed into *E. coli* cells, as described in Section 1.2.3.

1.2.3 Bacterial Transformation

A 0.5 µL aliquot of each ligation product (section 1.2.2) was mixed with 60 µL RapidTrans^{TM} TAM1 Competent *E. coli* cell suspension (Active Motif, California, USA). For expression strains, 75-100 µL of the relevant *E. coli* cell suspension was transformed with 50-100 ng of plasmid, as detailed in **Table 1.3**. Transformation mixtures were incubated on ice for 30 min, heat shocked in water at for 90 sec at 42^{o}C, followed by 5 min incubation on ice. 700 µL LB broth was then added to the mixture and incubated at 37^{o}C for 1 hour. Cells were pelleted by centrifugation at 4000 rpm for 5 min. Excess supernatant was discarded, saving about 100 µL in which the cell pellet was resuspended and spread on to LB-agar plates with suitable antibiotics (see), and incubated at 37 ^{o}C O/N.

Table 1.3: *E. coli* expression strains and their transforming plasmids

Expression plasmid	*E. coli* strain
ECWaaP_pET28b and PAWaaP_pET28b	OverExpress^{TM} C43 (DE3) (Lucigen, Wisconsin, USA)
pOPINF_BR55 and pOPINF_BR78; ECWaaP_pLou3, SEnWaaP_pLou3 and PAWaaP_pLou3	BL21 (DE3) (Novagen, USA)
ECWaaP_pLou3, PAWaaP_pLou3 and CFWaaP_pLou3, all with pWQ911	SoluBL21 (DE3) (Genlantis, California, USA)

1.2.4 Plasmid DNA Extraction

Single colonies were picked from LB-agar plates and inoculated into 10 mL LB broth with suitable antibiotics and incubated O/N at 37^{o}C. Plasmids were extracted from *E. coli* cells using the GeneJET Plasmid Miniprep Kit (K0503, Thermo Scientific, UK) as per the manufacturer's instructions.

1.2.5 Recombinant DNA Analysis

Cloned WaaP expression vectors were tested for the WaaP construct by double restriction analysis. About 1 µg of each plasmid was treated with 0.5 µL of the required FastDigest restriction enzymes (see **Table 1.1**) in FastDigest Green universal restriction buffer (ThermoFisher Scientific), at 37^{o}C for 1 hour. The reactions were resolved on a DNA agarose gel (see 1.2.6). Integrity of the cloned constructs were further verified by Sanger Sequencing (Source BioScience, Cambridge, UK).

1.2.6 DNA Agarose Gel Electrophoresis

PCR (1.2.1) and restriction samples (1.2.4) were resolved on 1% (w/v) agarose gels, with 0.5 µg/mL ethidium bromide (Bio-Rad, Hemel Hempstead, UK), and run in Tris acetate EDTA (TAE) buffer (40 mM Tris-base/ pH8, 2 mM sodium acetate, 1 mM EDTA) for 50 min at 100 V. DNA was detected under UV light with the Bio-Rad Gel Doc™ EZ Imager. The 1kb GeneRuler DNA Ladder (Thermo Scientific, UK) was used as a standard.

1.2.7 Estimation of DNA Concentration

DNA concentrations were measured at A_{260} nm on a NanoDrop 2000c UV-Vis Spectrophotometer (Thermo Scientific, UK). The output was in ng/µL and the absorbance ratio A_{260}/A_{280} was used to assess the purity of the plasmid; a ratio of 1.8 and above was deemed acceptable.

1.3 Protein Biochemistry

1.3.1 Recombinant Protein Expression and Extraction

1.3.1.1 BR55 and BR78 expression

O/N starter culture was inoculated into LB and 1% glucose media at 2% v/v and grown to an OD_{600} of 0.8 to 1.0. Recombinant protein expression was induced in *E. coli* strains BL21(DE3) (for BR55 and BR78) with 0.5 mM isopropyl β-D-thiogalactopyranoside (IPTG) (Formedium) at 20°C for 16-18 hours. *E. coli* cells were harvested by centrifugation at 4,500 rpm and used promptly, or stored at -20°C until further use.

1.3.1.2 C-6His-tagged WaaP expression

a) C-6His-PAWaaP: O/N starter culture was inoculated into TB broth (prepared as described in Section 1.1.2.1) at 3% v/v and grown to an OD_{600} of 0.8. Protein expression was induced in *E. coli* CHisPAWaaP_pET28b_C43(DE3) with 1mM IPTG at 20°C for 16-18 hours.
b) C-6His-ECWaaP with Gro*ESL* coexpression: O/N starter culture was inoculated into TB broth at 2% v/v and grown to an OD_{600} of 0.8. Gro*ESL* expression was induced in *E. coli* CHisECWaaP_ pWQ911_pET28b_C43(DE3) with 0.02% L-arabinose at 30°C; ECWaaP expression was induced after 1 hour with 0.5 mM IPTG at 22°C for 24 hours.

All *E. coli* cells were harvested by centrifugation at 4,500 rpm and used promptly, or stored at -20°C until further use.

1.3.1.3 Expression of MBP-tagged WaaP

a) Without Gro*ESL* co-expression: O/N starter culture was inoculated into LB broth at 2% v/v and grown to an OD_{600} of 0.5-0.6. Recombinant protein expression was induced with 0.1 mM IPTG at 20°C for 14-18 hours.
b) With Gro*ESL* co-expression: MBP-WaaP fusion proteins that were co-expressed with the Gro*ESL* chaperonin (expressed by the pWQ911 plasmid – section

1.1.1.2) were ECWaaP_pLou3, PAWaaP_pLou3 and CFWaaP_pLou3, as given in **Table 1.3**). O/N starter culture was inoculated into LB broth at 2% v/v and grown to an OD_{600} of 0.5-0.6 at 37°C. Gro*ESL* expression via the pWQ911 plasmid was induced with 0.015% L-Arabinose (Sigma-Aldrich) at 20°C for 40 min. After this, 6His-MBP-WaaP expression was induced with 0.05 mM IPTG, at 20°C for 18-22 hours, with shaking at 200 rpm.

In all cases the cells were harvested by centrifugation at 4,500 rpm and used promptly, or stored at -20°C until further use.

1.3.2 Recombinant Protein Extraction

Harvested cell pellets were re-suspended at room temperature in 2 v/w of the respective 'Lysis Buffer' (see **Table 1.4**) until a smooth and homogenous consistency was obtained. The suspension was subjected to 2 cycles of lysis on a cell disruptor (Constant Systems Ltd., Daventry, UK), which was pre-chilled to 5°C; the first cycle was performed at 30 kpsi and the second at 20 kpsi. The cell lysate was collected on ice. Clarified supernatant for IMAC purification was obtained by centrifuging the lysate at 20,000 rpm and 4°C for 30 min.

1.3.3 Protein Purification

For all protein extraction and purification procedures 6 buffers were used; a lysis buffer, binding buffer, wash buffer, elution buffer, protease buffer and SEC buffer. The buffers vary in composition for the purification of *L. reuteri* SRRP and WaaP recombinant proteins, although the purification procedures remain largely uniform. Buffer compositions for all purification trials are summarised in **Table 1.4** below.

Table 1.4: Protein purification buffers

Buffer compositions for BR55, BR78 and WaaP proteins which were extracted and purified according to the methods described in 1.3.2 and 1.3.3. At the bottom of this table, each protein is cross-referenced with their purification results, discussed in the relevant section of this thesis.

Protein	Base Buffer	Lysis	Binding	Wash	Elution	Protease	SEC Buffer
BR55[1]	20 mM Tris-HCl/pH 7.9, 0.5 M NaCl, 10% glycerol	Base buffer, 10 mM IMD, 1 µM DNAse, 1 µM lysozyme	Base buffer, 10 mM IMD	Base buffer, 20 – 30 mM IMD	Base buffer, 250 mM IMD	20 mM Tris-HCl/pH 7.9, 0.3 M NaCl, 10% glycerol	20mM Tris-HCl, pH 7.9, 0.15 M NaCl and 2.5 mM DTT
BR78[2]							
BR55 mutants[3]							
C-6His-ECWaaP[4]	20 mM Tris-HCl/pH 7.5, 0.5 M NaCl, 10% glycerol	Base buffer, 10 mM IMD, 1 µM DNAse, 1 µM lysozyme, 5 mM β-ME	Base buffer, 10 mM IMD, 5 mM β-ME	Base buffer, 25 mM IMD	Base buffer, 250 mM IMD	N/A	20 mM Tris-base, pH 7.5, 0.3 M NaCl, 10% glycerol
C-6His-PAWaaP[5]							
MBP-ECWaaP[6]	20 mM Tris-HCl/pH 7.5, 0.5 M NaCl, 10% glycerol	Base buffer, 10 mM IMD, 1 µM DNAse, 1 µM lysozyme, 5 mM β-ME	Base buffer, 10 mM IMD, 5 mM β-ME	Base buffer, 25 mM IMD	Base buffer, 250 mM IMD	20 mM Tris-HCl/pH 8.0,, 0.3 M NaCl, 10% glycerol, 2.5 mM β-ME	20 mM Tris-base, pH 7.5, 0.3 M NaCl, 10% glycerol
MBP-SEnWaaP[6,7]							
MBP-PAWaaP[6]							
MBP-PAWaaP, + Gro*ESL*[8]	20 mM Tris-HCl/pH 7.5, 10% glycerol	Base buffer, 0.75 M NaCl, 10 mM IMD, 1 µM DNAse, 1 µM lysozyme, 5 mM β-ME	Base buffer, 0.75 M NaCl, 10 mM IMD, 5 mM β-ME	Base buffer, 0.5 M NaCl, 20 mM IMD	Base buffer, 0.5 M NaCl. 250 mM IMD		20 mM Tris-base, pH 7.5 and 0.15 M NaCl; with 5mM DTT or 3 mM TCEP
MBP-ECWaaP + Gro*ESL*[8]							
MBP-CFWaaP + Gro*ESL*[9]							

[1] Results in section 4.1.1
[2] Results in section 4.1.2
[3] Results in APPENDIX 4
[4] Results in section 6.3.1
[5] Results in section 6.3.2
[6] Results in section 6.3.3
[7] Results in 6.3.3.1; additives to buffers for all SEnWaaP purifications trials are given in **Table *6.2***
[8] Results in section 6.3.4
[9] Results in section 6.3.6

1.3.3.1 IMAC Purification

6His-tagged fusion proteins were purified by IMAC using nickel-nitrilotriacetic acid (Ni-NTA) agarose resin (Qiagen, Hilden, Germany) in a gravity flow column (Bio-Rad, UK). Approximately 1 to 1.5 mL of resin was used per litre of cell culture, according to recombinant protein yield. The Ni-NTA matrix was pre-equilibrated with the appropriate 'Binding Buffer' before chilled clarified lysate, prepared as described in section 1.3.2, was applied onto the column at a flow rate of 2-3 mL/min. Non-specific proteins were washed off with 10 times the matrix's volume of 'Wash Buffer'. Bound proteins were then eluted in approximately 1.5 mL fractions with high imidazole containing 'Elution Buffer'; up to 12-15 fractions were collected and most proteins eluted between fractions 3 and 8, the concentrations of which were estimated as described in section 1.3.4. Proteins not requiring His-tag removal (C-6His-ECWaaP and C-6His-PAWaaP) were directly subjected to SEC purification (section 1.3.3.3).

1.3.3.2 Fusion tag removal and pull down purification on IMAC

Fractions containing the most protein were pooled together for buffer exchange into a no-imidazole 'Protease Buffer' for subsequent His-tag or His/MBP-tag removal. This was done using a 26/10 HiPrep Desalting column (GE Healthcare, Little Chalfont, UK) on an ÄKTApurifier UPC 100 chromatography system (GE Healthcare, Uppsala, Sweden). Proteins were collected in 1.5 mL fractions, and those containing the highest concentration of protein were pooled together. The desalted protein solution was supplemented with 2.5 mM β-mercaptoethanol (β-ME). 3C-protease or TEV-protease (as appropriate) was added to the protein solution in approximately 1:50 protease to protein ratio and incubated at room temperature O/N with gentle shaking. The protease treated protein solutions were supplemented with 15 mM imidazole from a 1M imidazole stock (pH 8.0) and then loaded onto the re-washed IMAC matrix, pre-equlibrated with the relevant 'Protease Buffer' for a pull-down purification. Where excessive MBP carry-over in the pull down was an issue, the flow through was passed through an additional Amylose-agarose resin (which selectively binds to MBP) or another IMAC resin; this especially was the case in purification of PAWaaP expressed

as an MBP-fusion with Gro*ESL* co-expression in *E. coli* SoluBL21(DE3) (results in section 6.3.4). Flow through containing the cleaved target protein was collected and concentrated to 2-4 mL using a 10,000 molecular weight cut-off (MWCO) Vivaspin 20 spin concentrator (Sartorius, Göttingen, Germany).

1.3.3.3 Size-exclusion Chromatography (SEC)

IMAC purified proteins were further purified by SEC, in the appropriate 'SEC Buffer', for subsequent crystallography and biophysical or biochemical characterisation. SEC was performed on a HiLoad™ 16/600 Superdex™ 200 pg prep grade column, fixed to the ÄKTApurifier UPC 100 system (GE Healthcare, Sweden).

1.3.4 Estimation of Protein Concentration

Protein concentrations were measured at A_{280} nm on a NanoDrop 2000c UV-Vis Spectrophotometer. The output was in mg/mL and the A_{260}/A_{280} ratio was used to assess protein quality; ratios under 0.6 were deemed acceptable.

1.3.5 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

For SDS-PAGE of protein samples, Bolt® 4-12% Bis-Tris Plus precast mini gels (12 well or 15 well, 1.0 mm) were used with Bolt® 2-(N-morpholino)ethanesulfonic acid (MES) SDS running buffer (Life Technologies, Paisley, UK). The gels were run for 35 min at 165 V, according to the manufacturer's instructions. Samples were prepared with 5x loading buffer (50% glycerol, 300 mM Tris-HCl pH 6.8, 10% SDS, 0.05% bromophenol blue and 5% β-ME – as reducing agent), and incubated at 90^{o}C for 5 min before loading onto gels. Either the 2-212 kDa broad range protein marker (NEB) or the Pierce™ unstained protein marker (Thermo Scientific, UK) was used as protein standard. SDS-PAGE gels were stained with Coomassie blue stain solution (0.025% w/v Brilliant Blue G, 7.5% v/v glacial acetic acid and 50% v/v ethanol, all from Sigma-Aldrich), or the Quick Coomassie Stain (Generon, Maidenhead, UK).

1.3.6 Se-Met Labelling of BR55 for X-ray Diffraction (XRD) Studies

Minimal media stock solution compositions and the sequence of preparation is given in APPENDIX 1. An amount equivalent to 5% v/v of the original overnight culture was inoculated into a litre of minimal essential media (M9 salts and essential media stock), supplemented with Glucose-free SelenoMethionine Nutrient Mix and incubated at 37^{o}C with shaking at 200 rpm till an OD_{600} around 0.3 was obtained. Amino acid supplement was then added to the cultures, with further incubation for 30 min, after which the L-Se-Met suspension was added. The cultures were grown to an OD_{600} of 0.5-0.6, where after protein production was induced with 1 mM IPTG at 20^{o}C for 14 hours. The wet cell yield after harvesting was 6.6 g/L, and Se-Met labelled BR55 was purified as described in section 1.3.3.

1.4 Bioinformatics

1.4.1 Sequence Analysis

The ExPasy ProtParam tool (http://web.expasy.org/protparam/) was used to calculate the theoretical molecular weight, isoelectric point (pI), protein extinction coefficient and Abs 0.1% value from the amino acid sequence of proteins. Clustal Omega [8] was used for multiple sequence alignment (MSA) of protein molecules and generating the percentage identity matrix of WaaP homologues (APPENDIX 6). MView [9] was used to produce the WaaP MSA graphics (see APPENDIX 5). The LALIGN pairwise alignment program [10] (http://embnet.vital-it.ch/software/LALIGN_form.html) was used to analyse DNA sequencing results. EggNOG 4.5 [11] and Pfam [12] were used for WaaP orthologue searches. The online 'GenScript Rare Codon Analysis Tool' was used to estimate the Codon Adaptation Index (CAI) of WaaP sequences (http://www.genscript.com/cgi-bin/tools/rare_codon_analysis).

1.4.2 Structural Modelling

The PSIPRED Protein Sequence Analysis Workbench [13] was used for protein secondary structure prediction, disorder prediction and homology modelling. Protein structure searches and predictive structural modelling against the Protein Data Bank (PDB) was performed using the Phyre2 server [14]. Pairwise structural alignment was performed on the DaliLite server [15], [16] and structural alignment based on secondary structure matching was performed by Superpose [17] within the CCP4 suite of programs. Structural superfamilies were identified on the CATH database [18]. To generate surface electrostatic maps, pdb files were first converted to pqr format by PDB2PQR [19] and then computed by the Adaptive Poisson-Boltzmann Solver (APBS) [20]. TopDraw was used for the construction of protein topology maps [21]. Structural images were generated using the MacPyMOL Molecular Graphics System, Version 1.7.4.5. Schrödinger, LLC.

CHAPTER 2

INTRODUCTION:

Structural Studies of a Novel *Lactobacillus Reuteri* Serine-rich Repeat Surface Adhesin

2 INTRODUCTION: STRUCTURAL STUDIES OF A NOVEL *L. REUTERI* SURFACE ADHESIN

Prokaryotes form a vast component of the earth's biota and can thrive in a range of environments such as the atmosphere, soil, water, sub-surfaces, in other organisms [22], and some are found in habitats as extreme as hot springs, with temperatures above 80^{o}C [23]. Many vertebrate and invertebrate animals possess large prokaryotic populations with roles in health and disease. Prokaryotic bacteria are often associated with disease as many of them are pathogenic, however, the majority are known to form complex mutually beneficial relationships with animals and insects [24].

In vertebrates the densest bacterial population is found in the GIT, as much as 10^{10}-10^{12} cells/mL, the highest known for any microbial ecosystem [2]. The community within vertebrate GITs comprise largely of mutualistic bacteria that make vital contributions to the host in the form of nutrition [25], pathogen defense [26] immunity development [27-29] and physiology [30]. Despite their importance, little is known on how gut bacterial populations are acquired, and how are maintained over extended periods of time. 16S rRNA surveys demonstrate that the fecal microbiota of mammals are generally specific to their host species [31] and are remarkably stable [32,33] indicating that special mechanisms exist to recruit and maintain selected bacterial populations in the GIT. Compared to microbial symbioses in invertebrates [24], the processes underlying recognition and capture of bacterial lineages in vertebrate GITs are relatively less understood, but an increasing number of investigations in the last decade has provided a better understanding of these host-selection factors [34,35].

This thesis aims to further elucidate the molecular mechanisms of mutualistic bacteria in forming stable and heritable populations in the GIT of their vertebrate hosts. The following subsections will, therefore, summarise the known evolutionary mechanisms of *Lactobacillus reuteri*: a popular model for understanding vertebrate gut symbiosis.

2.1 Evolutionary Strategies of Vertebrate Gut Symbionts

L. reuteri belongs to the order Lactobacillales, comprising lactic acid bacteria (LAB) which are tolerant to low pH conditions [36]. Out of hundreds of gut commensals, *L. reuteri* is regarded as an excellent model to investigate the evolutionary mechanisms of vertebrate gut symbiosis, as it is naturally found in the digestive tract of several vertebrates, but, in varying population percentages and niches [2]. *L. reuteri* has been found in the digestive tracts of humans, mice, rats, pigs, horses, dogs, goat, and birds such as chicken and turkey [1, 37-43]. How did *L. reuteri* come to populate such a range of hosts? Are their associations promiscuous or autochthonous? Are the host associations mutualistic or commensal? These questions have been increasingly investigated in the last decade at the genomic, proteomic and molecular levels.

2.1.1 *L. reuteri* phylogenetic clusters sort by host-origin rather than geographical location

To investigate the genetic diversity in global *L. reuteri* populations, Oh et. al. [1] analysed the genomes of 165 *L. reuteri* strains from rats, mice, pigs, humans, chickens and turkeys from worldwide locations, using amplified fragment length polymorphism analysis (AFLP) [44,45] and multilocus sequence analysis (MLSA) [46]. Results from both techniques identified six clusters with almost complete overlap, and indicated that divergent *L. reuteri* subpopulations sort by host origin rather than geographical provenance: 3 rodent, 2 pig, 1 human cluster and 1 cluster largely of poultry strains, with a few human isolates were distinguished.

The average age of the common ancestor of the 165 *L. reuteri* strains studied was estimated at about 7.60–12.95 million years, implying a long-term association of *L. reuteri* lineages with their host species [1]. It is worth noting that the phylogenetic clusters were not constrained to a single host, but to a group of closely-related species such as poultry (chicken and turkey) and rodents (rats and mice). Subsequently, it

was hypothesized that over millions of years, *L. reuteri* isolates may have evolved within groups of closely related hosts that possess similar niches in the gut (such as anatomy, GIT glycan composition, diet and innate immune receptors), and live in environments permitting horizontal transfer of microbes. Another suggestion was that the time allowing for the divergence of truly species-specific lineages was not sufficient [1, 2].

2.1.2 *L. reuteri* evolution is host driven and selective

Apart from genomic evidence, host specialization has been reflected and proven for the phenotypic traits of *L. reuteri* isolates via *in vivo* competition experiments [1, 4, 47-49]. For example, a study by Frese et. al. [4] investigating factors of host-specific biofilm formation in the murine GIT, demonstrated that all mouse and rat *L. reuteri* isolates were successful in forming biofilms on the stratified squamous epithelium of the germ-free mice's forestomach (the natural habitat of rodent *L. reuteri* [50]). Contrarily, populations of *L. reuteri* isolates from human, pig and chicken were profuse in the forestomach lumen, but showed limited adhesion to epithelial cells. This study is expounded later in this chapter (section 2.2), and the competition results are displayed in **Figure *2.7***. This denotes that although the murine GIT is not a hostile environment for exogenous *L. reuteri* strains, it is not conducive to long-term, self-sustainable habitation either. A similar observation was made by Lin and Savage [47] for the behavior of rodent versus non-rodent *L. reuteri* strains on murine forestomach epithelium, and by Tannock et. al. [49] for the colonization of the fowl crop epithelium by poultry and non-poultry *L. reuteri* isolates.

Studies have also identified selection factors unique to certain hosts [2, 48]. In a study, Walter et. al. [2] investigated two phenotypic survival factors within 32 *L. reuteri* isolates from rodent, human, porcine and poultry hosts: the ability to express urease [51] and reuterin [52]. Urease is an enzyme encoded by the urease cluster, of which the *ureC* gene is a component. It has been shown to confer *L. reuteri* acid resistance

in the gut by hydrolyzing urea [53]. Reuterin is an antimicrobial compound shown to inhibit pathogenic bacteria [54], [55], and the ability to express reuterin was assessed by detecting the presence of the *pduC* gene. The results () showed that all 9 rodent strains produced urease, but only one of 23 strains from other hosts was positive for this trait. Contrarily, only one rodent strain could produce reuterin, whereas all human and poultry strains possessed this phenotype.

Expanding this study further, Frese et. al. [48] identified a pattern of genomic features strongly associated with host type among 57 *L. reuteri* strains. Like above, the urease phenotype was solely found in rodent isolates and the reuterin expressing *pdu-cbi-cob-hem* cluster was identified as a defining feature of human *L. reuteri* genomes [48].

Four other genomic elements were found to be prominent in rodent *L. reuteri* lineages compared to non-rodent isolates; an illustration showing the distribution of genomic islands amongst the 57 strains is given in **Figure 2.2**. Firstly, up to 11 additional surface proteins were found in rodent strains, that were rare in pig and poultry isolates and absent in human strains. They were identified with possible roles in adhesion and biofilm formation as they possessed a C-terminal LPXTG-motif (IPR019931), a typical cell wall anchor motif in Gram-positive bacteria [56, 57]. Domain features of known surface adhesins were also identified, such as for mucus-binding proteins [58], extracellular matrix binding adhesins, and a serine-rich repeat surface adhesin [48].

Secondly, a SecA2 cluster was detected in most rodents and pig strains, while it was rare in human and poultry isolates. The SecA2 system is widely found in streptococci and staphylococci, where it exports glycosylated serine-rich surface adhesins – this system and its associated adhesin will be explored comprehensively in section 2.4.

Thirdly, the xylose operon was highly conserved in rodent and porcine strains, but absent in human and poultry strains. Walter et. al. [59] identified xylose isomerase (xylA) as one of the *L. reuteri* genes specifically induced in the murine GIT upon colonization. XylA catalyses xylose (a plant-based sugar found in bran and straw) to lactic acid [60, 61]. Therefore, a role in mutualistic nutrient acquisition from xylose

fermentation is implied, both for *L. reuteri* and its rodent host.

Finally, a two-component regulatory system (TCS) indicating a role in quorum sensing was detected in more than half of the rodent strains, but with limited occurrence in non-rodent isolates. Quorum sensing has been implicated in maintaining bacterial biofilm integrity [62], where the susceptibility of pathogenic biofilms to antibiotics was increased with the use of quorum sensing inhibitors [63, 64].

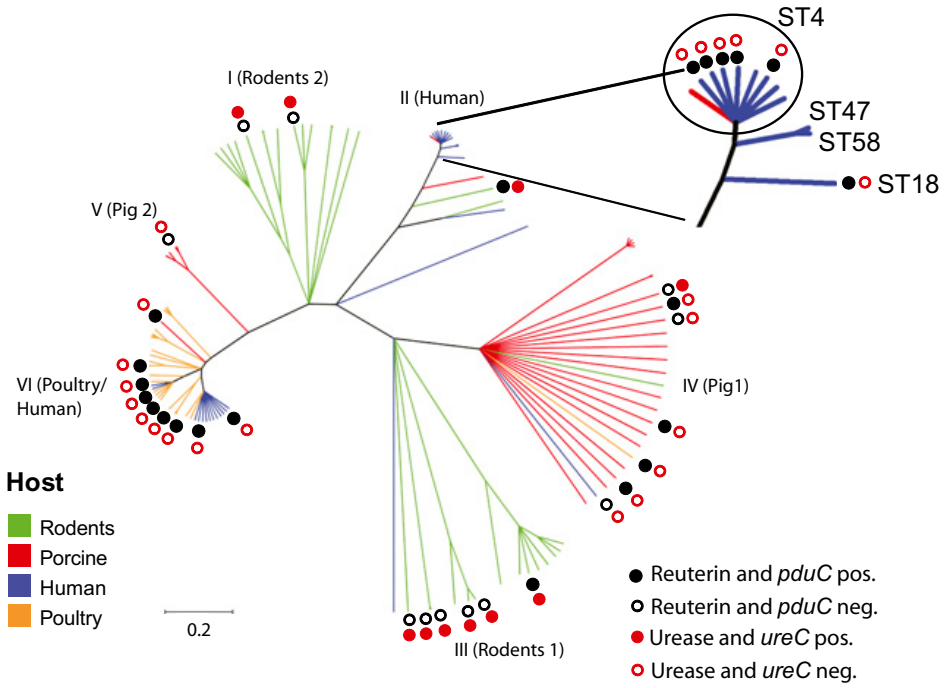


Figure 2.1: Genealogy of 116 *L. reuteri* strains inferred by Oh et. al. [1]**, from which 32 strains were analysed for *ureC* and *pduC* genes. Adapted from** [2].
The tree branches are coloured by host origin. Strains producing reuterin and having the *pduC* gene are marked by solid black circles and strains that do not are marked with hollow black circles. Strains producing urease and possessing the *ureC* gene are marked with solid red circles and those that do not are marked with hollow red circles

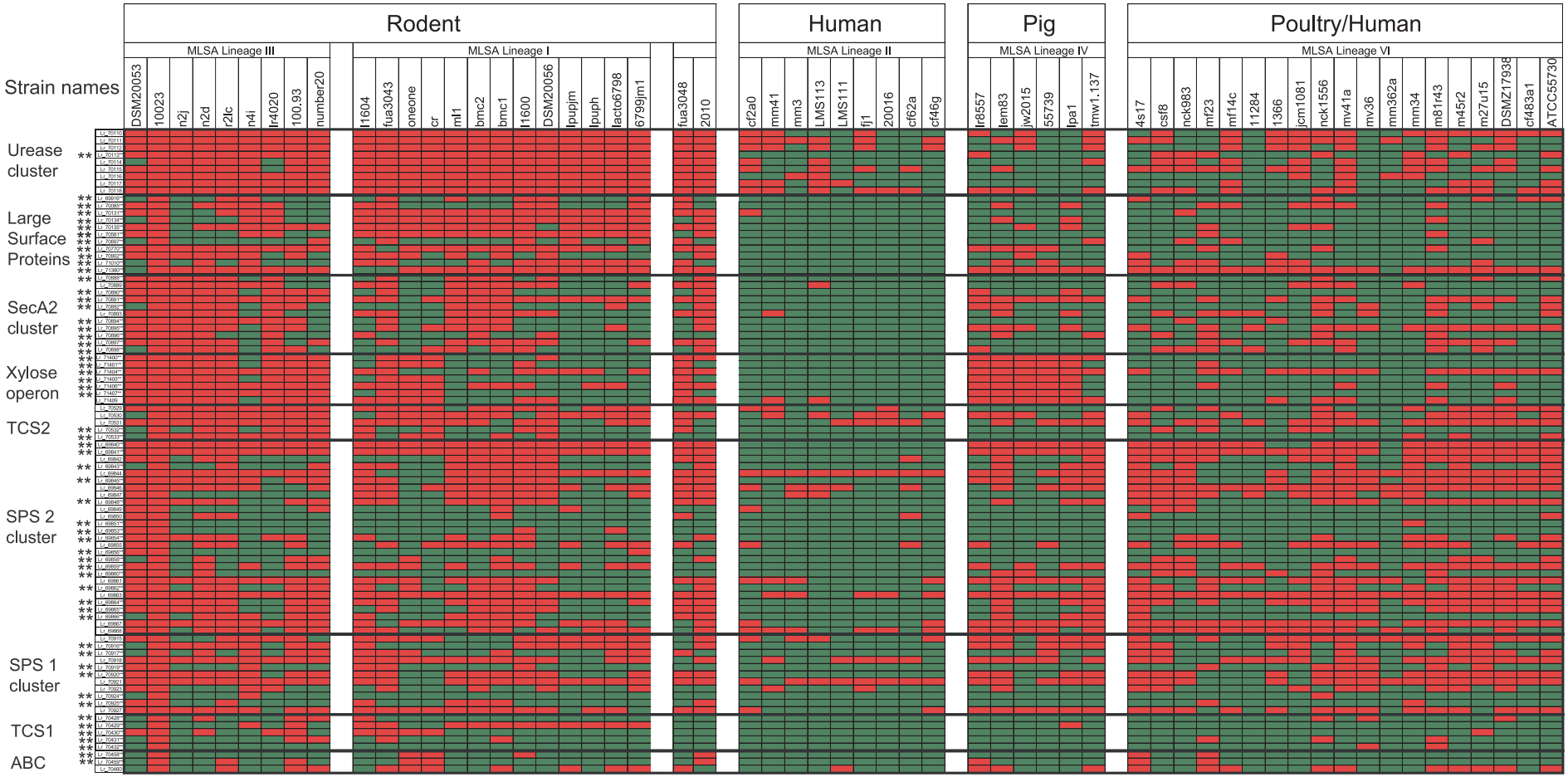


Figure 2.2: Distribution of putative rodent-specific genes amongst 57 L. reuteri strains. Adapted from [48]**.**

Strains are arranged by host and MLSA lineage (see section 2.1.1) as deciphered by Oh et. al. [1]. Red and green colours denote presence or absence, respectively. Genes identified as rodent-specific when compared to human lineage II strains are marked by double-asterisks (**).

2.1.3 How host-specific genes reflect niche features within the host GIT

Biofilms are populations of microbes that are concentrated at an interface and are surrounded by a self-produced extracellular matrix (ECM), composed of proteins and exopolysaccharides (EPS) [65]. Biofilm development is an important phenomenon for bacterial survival in hostile environments and has been postulated to be a hallmark of successful bacterial gut colonization [66].

In rodents, pigs, and poultry thick *L. reuteri* dominated biofilms form in proximal gastric regions of the gut, adhering directly to the non-secreting, stratified squamous epithelia on the forestomach [67], pars oesophagus [68] and crop [69], respectively. These non-glandular gastric regions form a type of transition site from the oesophagus to the true stomach, where it often functions as a temporary food storage site. Contrarily, the human stomach is devoid of stratified squamous epithelia, is entirely glandular and lacks a post-oesophagal food storage pocket . Lactobacillus biofilm formation has not yet been described in the human GIT, and the density and diversity of symbionts in the human gut is known to rise progressively from small intestine to colon [70].

Figure 2.3, **Figure 2.4** and **Figure 2.5** highlight relevant features of the anatomy of murine, human, porcine and poultry digestive systems that will be discussed further in this section in the context of their respective L. reuteri niches.

As described earlier in section 2.1.2, several rodent-specific surface proteins were identified by Frese et. al. [48], with proposed functions as surface adhesins for biofilm development. The SecA2 system is likely involved in the secretion of the newly detected putative serine-rich surface adhesin. Other factors such as the TCS and the urease cluster are implied in biofilm maintenance, via cell aggregation and pH regulation, respectively. Additionally, some of the rodent-associated genes were also measurable in selected pig and poultry isolates (see **Figure 2.2**) reflecting some similarities in the lifestyle of rodent, pig, and poultry *L. reuteri* which all form biofilm-like in the proximal gastric region.

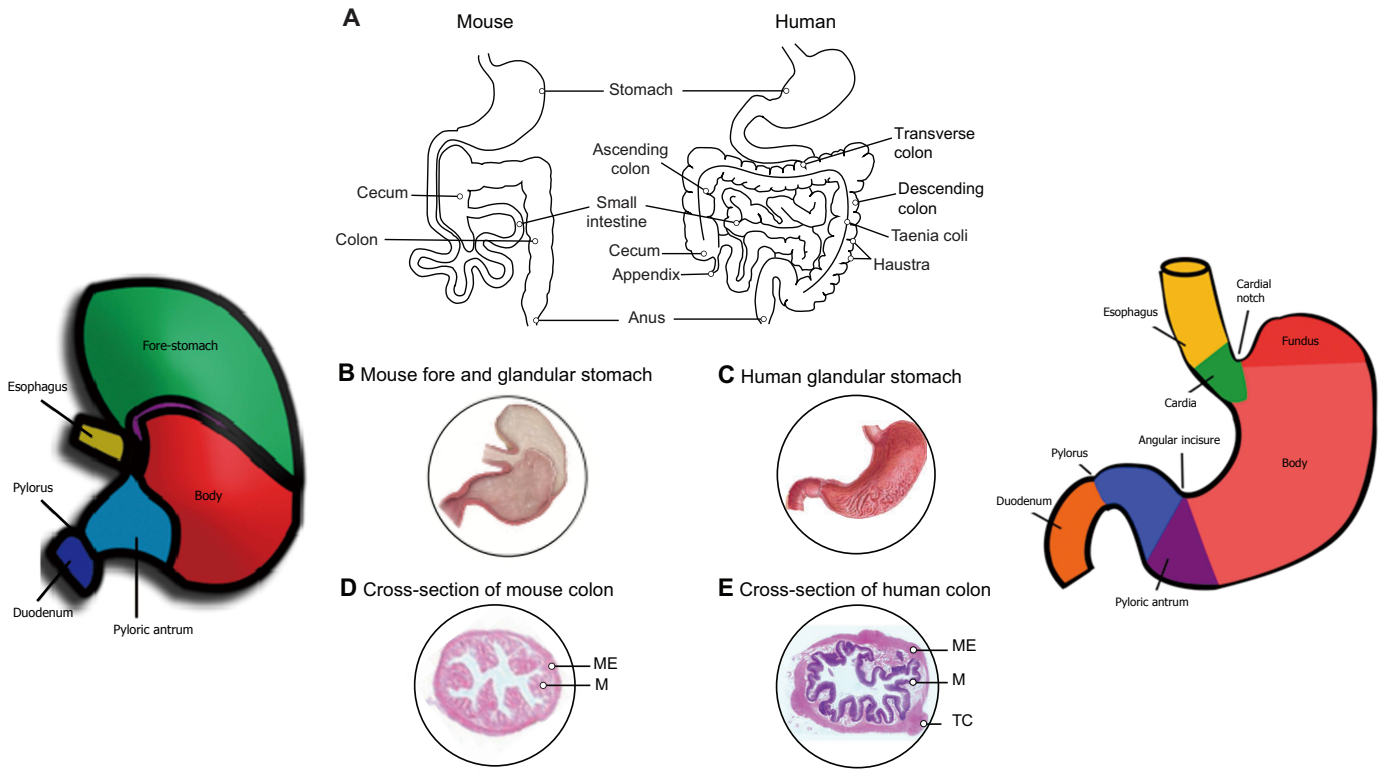


Figure 2.3: Comparative anatomy of the human and murine gastrointenstinal tracts. Adapted from [71] **and** [72]**.**
A. The human colon is more compartmentalised than the mouse colon, being divided into the ascending, transverse and descending colon, with the presence of taenia coli (TC) and haustra which are absent in the murine colon. **B.** The mouse glandular stomach secretes gastric acid and is divided into the body and pylorus. The non-glandular stomach (or forestomach) is lined with stratified squamous epithelial cells and functions as a temporary site of food storage and digestion; it is also the region of the stomach in which *L. reuteri* biofilms are found [67]. **C.** The human stomach is entirely glandular and is organised into four regions: 1) the cardia at the junction of the esophagus and the stomach; 2) the fundus and 3) body constitute the bulk of the stomach; 4) the pylorus leads into the duodenum (small intestine). **E.** The cross-section of a human colon, having a thicker muscular wall (ME) and mucosa (M) compared to the murine colon, **D.**

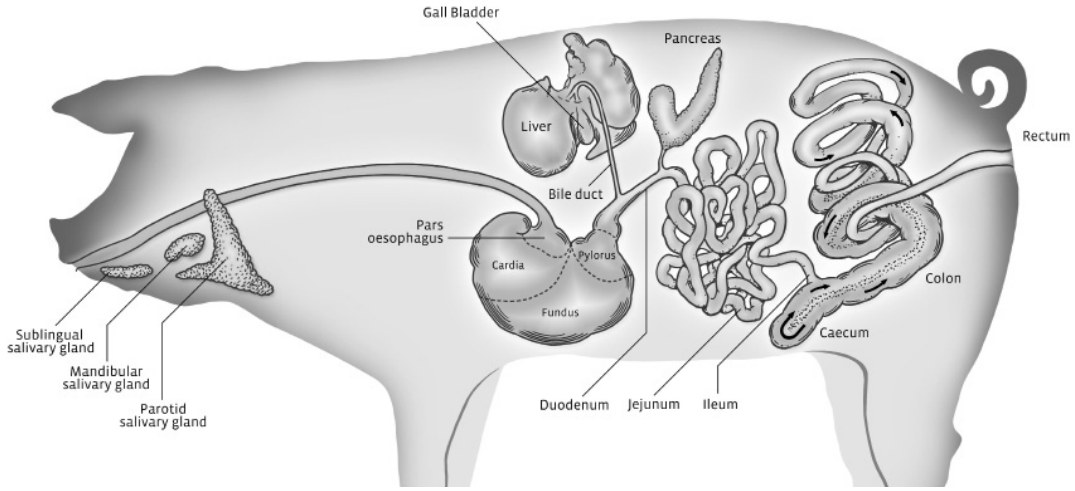


Figure 2.4: Anatomy of the pig gastrointestinal tract.
Taken from [73]. The pig's stomach is divided into four functionally and structurally different regions. 1) The pars esophagus is a non-glandular extension of the esophagus leading to the stomach, and is the region in which *L. reuteri* biofilms form [68]. 2) The cardia is much larger in pigs in contrast to other species and occupies a third of the glandular stomach, followed by the 3) fundus and 4) pylorus [73].

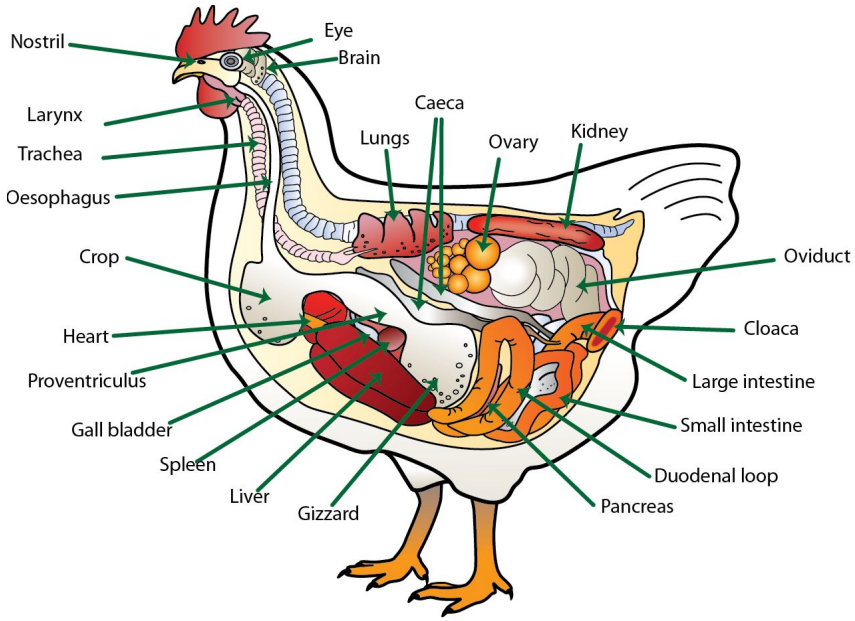


Figure 2.5: A hen's anatomy, including the digestive system.
Image borrowed from *PoultryHub* (http://www.poultryhub.org/physiology/body-systems/digestive-system/). 1) The crop is an out-pocket of the esophagus which functions as a temporary storage pouch, and it does not have any digestive function; *L. reuteri* biofilms in poultry are found in this part of the GIT [69]. 2) The proventriculus, also known as the 'true stomach' is the glandular stomach where gastric enzymes are secreted for chemical digestion of food. 3) The gizzard, or ventriculus is also referred to as 'the mechanical stomach' where physical digestion predominantly occurs [74].

The genome content of human *L. reuteri* isolates demonstrated a stark contrast to other *L. reuteri* lineages, especially with rodents. Despite high nutrient availability, microbial numbers are restricted in human gastric regions due to extremely low pH, toxic bile salts and the swift flow of digestive material [70]. Coupled with the fact that stratified squamous epithelia are absent in the human stomach, this would deter the genesis of biofilms like those in rodents, pigs and poultry. It is postulated that *L. reuteri* populations in the human GIT lead a semi-planktonic lifestyle in distal locations of the gut with few epithelial interactions [48] (some associated adhesins have been identified and characterised [75]). Frese et. al. [48] hypothesised that such a niche would demand quick multiplication rates, explaining the lack of large surface proteins that are likely to pose a large metabolic burden. Moreover, due to limited nutrient supplies in the colon, the ability of *L. reuteri* to use 1,2-propanediol as an energy source (via the *pdu-cbi-cob-hem* cluster) could be an important survival factor, and reuterin production may also serve to boost *L. reuteri* survival in the colon by inhibiting competitor colonisation within the nutrient deficient niche [2,55,76].

Thus the above studies [2,48] helped reveal the differing ecological strategies of *L. reuteri* in humans and animals, with human and rodent isolates posing nearly opposing selection factors, and pig and poultry strains being intermediate in possessing some human, but mostly rodent characteristics, or both. The distribution of the genomic features discussed above is summarised as follows:

1) the urease cluster was observed exclusively in rodent isolates
2) 11 extra surface adhesive proteins were identified, largely confined to rodent strains, and with limited occurrences in porcine and poultry isolates
3) The SecA2 cluster was detected in almost all rodent, but some porcine strains
4) the xylose operon was observed in almost all rodent and porcine isolates
5) The *pdu-cbi-cob-hem* was observed in all human lineages, almost all poultry strains and was rare in porcine strains.

2.2 Molecular Mechanisms of *L. reuteri* Biofilm Formation on the Proximal GIT of Vertebrates

In a study with rodent *L. reuteri* 100-23 as a model, Frese et. al. [4] performed a systematic assessment of factors underlying host-driven biofilm formation in the GIT of germ-free mice. The aims of this study were to identify the key genetic components driving biofilm formation. 11 genes in rodent *L. reuteri* 100-23 were identified or predicted to be playing a vital role in the formation and maintenance of biofilms, some of which were introduced in the previous sub-section. The impact of these genes on *in vivo* biofilms was assessed via functional assays, where germ-free mice were colonised by individual isogenic mutants and wild-type *L. reuteri* 100-23, and biofilm densities were quantified 2 days later by confocal microscopy [4]. The 11 genes were: **LysM2** and **LysM3** (roles in intra-species aggregation), **LrgA** (a putative biofilm regulator), **Lr70430** (histidine-kinase part of a TCS), **UreC** (alpha-subunit of urease enzyme), **LytS** (cell lysis regulator in biofilms), **CgI** (involved in reactive oxygen resistance), **Lr70532** (host-specific quorum sensing), the **SecA2** system, **Lr70902** (putative serine-rich surface adhesin) and **Lsp** (putative surface protein [77]).

Figure 2.6 presents the functional assay results, where six out of eleven mutants had over a 50% reduction in biofilm density compared to wild type 100-23. Variations in the degree of biofilm impairment for each mutant, therefore, helped to propose a hierarchy of molecular processes for step-by-step biofilm formation. *L. reuteri's* host-dependency in biofilm formation was also examined in a monoassociated mouse model, where the biofilm density of nine-wild type strains (isolated from mouse, rat, pig, chicken and human) was evaluated by confocal microscopy after an inoculation period of 2 days. The results are presented in **Figure 2.7**.

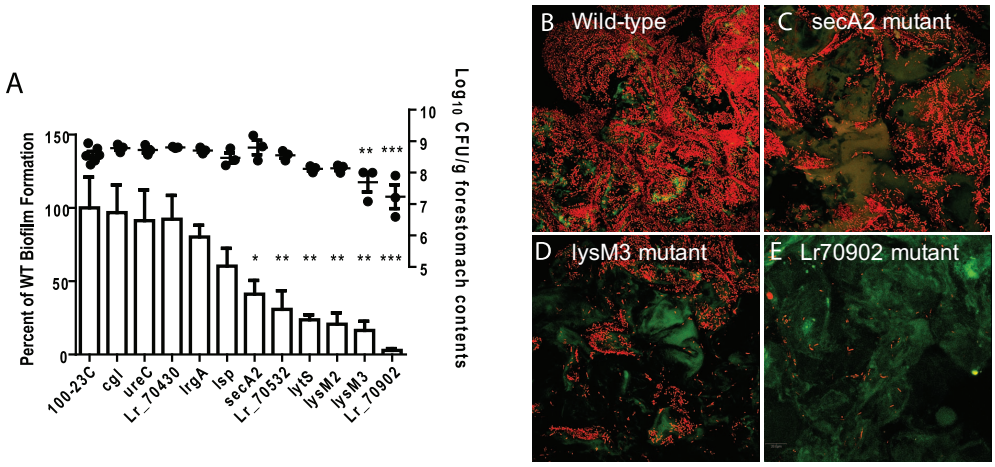


Figure 2.6: *In vivo* biofilm formation of mutant *L. reuteri* 100-23 strains
Taken from [4], **A.** Displays a graph quantifying the biofilm density of deletion mutants, relative to wild-type L. reuteri 100-23, measured by confocal microscopy. **B, C, D** and **E** are confocal micrographs of forestomach tissue from mice after two days of colonisation with the labelled *L. reuteri* genotypes.

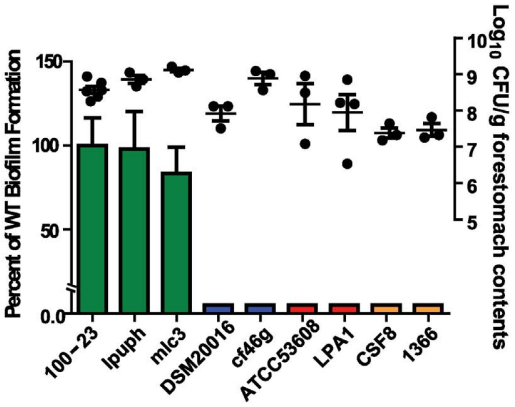


Figure 2.7: Biofilm formation of *L. reuteri* strains is highly host specific
Taken from [4], Biofilm density of various *L. reuteri* strains on forestomach epithelial tissue of monoassociated mice after two days of colonisation, quantified relative to 100-23. Bars are colour coded according to host origin: green – rodent, blue – human, red – pig, and orange – chicken.

2.2.1 Bacterial adhesion is the most critical step for biofilm formation

The loss of the Lr70902 gene, encoding a putative serine-rich repeat (SRR) surface adhesin, almost abrogated *L. reuteri* biofilm density in the mouse GIT, indicating that the initial process of host-surface recognition and attachment is the most vital step for

biofilm formation. When the genome of *L. reuteri* 100-23 was sequenced [48], Lr70902 was found to be unusually serine-rich (>35% serine content), having a similar domain architecture to known SRR adhesins of pathogenic streptococci and staphylococci (section 2.4.1). SRR proteins (SRRP) in pathogenic bacteria were identified as critical virulence factors for establishing infection in the host by initial adhesion to specific receptors, leading to the formation of stable biofilms within their host niches. An example is the Fap1 protein from *S. parasanguinis,* involved in dental plaque formation [78], and GspB form *S. gordonii* that binds to human platelets causing infective endocarditis [79]. The necessity of Lr70902 for biofilm formation was further reinforced by a clear loss in biofilm density for the secA2 deficient mutant. Genomic analyses [48] also revealed a specialised SecA2 secretion system in *L. reuteri* 100-23 for the possible production and surface export of Lr70902. The secA2 system is associated with SRRP biogenesis in pathogenic Gram-positive bacteria [80-83], and its conservation across known SRRP expressing bacteria is discussed in section 2.4.3. Thus, the observed impairment in biofilms for the SecA2 mutant was most likely caused by diminished Lr70902 transport to the cell surface.

2.2.2 Intra-species aggregation is important for biofilm integrity

After initial *L. reuteri* attachment, biofilm development is known to proceed with the formation of micro and macro cell colonies and their intra-cellular aggregation [84]. LysM proteins are implicated in inducing lactobacilli aggregation [85,86], which is proposed to be aided by the binding of LysM's N-terminal domain to *L. reuteri* cell wall peptidoglycan [87] and of its C-terminal YG-motif to carbohydrates [86]. Hence, the extent of biofilm atrophy noted for LysM2 and LysM3 deficient mutants suggest an influential role of auto aggregation in biofilm development.

2.2.3 Quorum sensing and cell lysis have roles in maintaining mature biofilms

The influence of two regulatory systems upon *L. reuteri* biofilms was made apparent in these experiments, although their mechanisms and genetic targets remain to be

understood. The first is the LytS system, which is orthologous to the one found in *S. aureus* [88] and shown to be a cell lysis regulator, contributing to the release of extracellular DNA that serves as a matrix in biofilms [89,90]. The second is a TCS associated with an ABC bacteriocin transporter (Lr70532) involved in quorum sensing; the proposed contribution of quorum sensing to biofilm preservation was covered in section 2.1.3.

2.2.4 Adhesion is host-specific and is the key step in biofilm formation

Although the isogenic mutants of *L. reuteri* 100-23 – shown in **Figure 2.6** – were impaired in forming biofilms, they achieved large populations in the forestomach after two days of colonization (>10^7cells/gram), implying that these genes are especially crucial to biofilm formation. Lr70902 (encoding a putative SRRP) had the most severe impact on biofilm density, while the other mutants formed biofilms to some extent. The authors, therefore, postulated that initial recognition and epithelial attachment is likely to be the most important step in biofilm development [4]. While most of the genes attributed to biofilm formation, that were studied by Frese et. al. [4], were not unique to rodent *L. reuteri* isolates, homologues of Lr70902 SRRP were only noted in rodent and pig isolates of *L. reuteri* (in which they always co-localised with the SecA2 gene cluster). This suggests that the SRR adhesins in rodent and porcine isolates may fulfill a vital role in defining the *L. reuteri* niche by recognising specific epithelium receptorsof their respective hosts. An SRRP like sequence was identified in the genome *L. reuteri* isolate ATCC 53608 [91], which was one of the pig isolates studied by Frese et. al. [4] in competition experiments with monoassociated mice (see **Figure *2.7***). ATCC 53608 clearly failed at forming biofilms on the murine forestomach epithelium, as shown in **Figure *2.7***. Supplementing this observation is the low sequence similarity of ATCC 53608 SRRP to Lr70902, which likely accounts for the observed host specificity. Furthermore, a recent comparative genomic analysis of six new porcine *L. reuteri* strains identified SRRPs as a potential molecular determinant of host-strain specificity in the pig pangenome [3].

2.3 Functionally characterised *L. reuteri* adhesins

Several *L. reuteri* cell-surface adhesins with distinct profiles have been identified and characterized from various strains [58,75,92-94]. The adhesins can be classified according to their binding targets in the GIT, their localisation on the bacterial surface, or to the way they are anchored to the bacterial surface [75]. Whilst there is notable structural diversity amongst known adhesins throughout lactobacilli, inter-strain variations in the number and types of adhesins found within a single strain also occur. For example, analysis of the *L. reuteri* 100-23 genome revealed 10 rodent-specific surface adhesins, each with different functional domains [48]; these are illustrated in **Figure 2.8**, and exclude non-rodent specific adhesins. Putative adhesins in lactobacill are usually identified via conserved Pfam [12] or InterPro [95] domains with known binding function. The presence of a distinct LPXTG cell wall anchor motif at the C-terminal, which is widely conserved in Gram-positive bacterial adhesins [57], is also a useful way of identifying potential surface adhesins, and this is often, but not always, found together with an N-terminal YSIRK-signal peptide motif [48], as seen in **Figure 2.8** below.

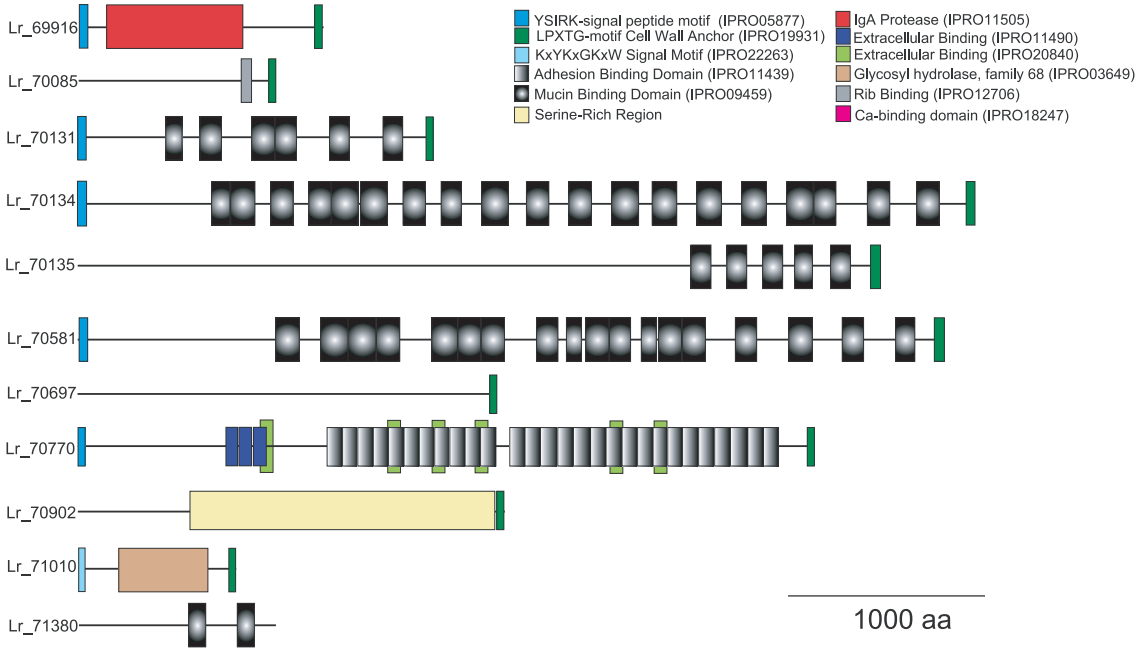


Figure 2.8: Rodent-specific surface proteins in *L. reuteri* 100-23 with InterPro functional domains drawn to scale; adapted from [48].

The first *L. reuteri* surface adhesin to be identified was an ABC transporter protein named CnBP from *L. reuteri* NCIB 11951, which binds to collagen type I [96,97]. This was followed by the identification of an adhesion promoting protein called MapA of *L. reuteri* 104R [98], which was found to bind to porcine GIT mucin [98] and to caco-2 human intestinal epithelial cells [99], in two separate studies. The receptor-like protein molecules on caco-2 cells that bind to MapA from *L. reuteri* DSM 20016 were identified as annexin A13 and paralemmin [100]. Due to high sequence similarity, CnBP and MapA are believed to be homologous adhesins.

A prominent type of *L. reuteri* mucus-binding surface adhesin was first identified by Roos and Jonsson from *L. reuteri* 1063 [101]. This mucus-binding protein was named MUB, and it facilitates lactobacilli adherence via multiple interactions with terminal sialylated glycans on the mucus lining the host's GIT [102]. MUBs are distinguished by the presence of one or more types of tandem Mub repeats (a Mub is a module, the type of which is determined by its amino acid sequence) [58,101], arranged like beads on a string to form protein fibrils of 100-200 nm [102]. Proteins with putative MUB domains have been detected in *L. reuteri* strains of all host origins, although varying greatly in composition and number between different strains [103]. These variations in MUB phenotypes also ensure host-specific *L. reuteri* adhesion to mucus in secretory regions of the host GIT, as demonstrated by MacKenzie et. al. [103]. Crystal structures for two Mub repeats constituting 14 repeat domains of a MUB from porcine *L. reuteri* ATCC 53608 [102,104] show an immunoglobulin (Ig) binding fold and binding activity, adding to the complexity of possible interactions between MUBs and the host GIT [104].

In the last 5 years, three additional types of surface adhesins from different *L. reuteri* strains have been identified and characterized to varying degrees. The first is the SRRP adhesin [48], which was introduced previously in this thesis, and will be explored in detail in the following section, 2.4. The second is a protein called elongation factor-Tu (EF-Tu), from *L. reuteri* JCM1081, which preferentially binds to sulfated carbohydrates on mucosal surfaces [105]. The third is a novel modular

protein found predominantly in human *L. reuteri* strains [106], with a similar tandem-repeat modular organisation like the MUB proteins, but much shorter in overall length and with a different Ig fold for the module. The first of its kind was a surface protein from human *L. reuteri* PTA 6475, with four repeat domains, which could bind to caco-2 cells and mucus, *in vitro*, and was hence named 'cell and mucus binding protein' (CmbA) [107]. A second homologous protein named Lar_0958, with six repeat domains was isolated from human *L. reuteri* JCM 112. The structure of its most common repeat unit was solved by X-ray crystallography, revealing an Ig-like beta-sandwich fold, that enables bacterial adhesion to GIT intestinal mucus [106].

2.4 Serine-rich Repeat Binding Proteins (SRRP)

Serine-rich repeat proteins (SRRPs) are a family of surface-exposed adhesive glycoproteins found in Gram-positive bacteria. Most SRRPs that have been biochemically or structurally characterised belong to pathogenic bacteria from the streptococcus and staphylococcus genera (see **Table 2.1**), whilst discoveries of SRRPs in commensal bacteria have been more limited; *Lactobacillus johnsonii* is the only commensal bacteria for which an SRRP type adhesin has been reported [108]. In pathogenic bacteria SRRPs have been reported to facilitate attachment to a range of host surfaces leading to biofilms that contribute to diseases like infective endocarditis [79,109], streptococcal meningitis [110], pneumococcal disease [111], neonatal disease [82,112] and dental plaques [78].

Recognising the fundamental and direct role of SRRP adhesins in host-specific biofilm formation, and identifying two SRRP proteins from divergent *L. reuteri* strains offered a novel model and an exciting opportunity to investigate the structural and biophysical features of these SRRPs. It was proposed that such a study could lead to an improved understanding towards how host-specific differences in epithelial adhesion are perpetuated at the molecular level. Domain boundaries of SRRPs from rodent *L.*

reuteri 100-23 and porcine *L. reuteri* ATCC 53608 were determined by comparative [48] [91]. From here onwards, *L. reuteri* 100-23 SRRP will be called Lr70902_SRRP, and its putative binding region will be named BR78; *L. reuteri* ATCC 53608 SRRP will be called Lr0906_SRRP (nomenclature derived from Wegmann et. al. [3]), and its putative binding domain will be called BR55.

2.4.1 The domain architecture of SRRP adhesins

Scanning electron microscopy suggests that SRRPs form fimbriae-like structures that extend outward from the bacteria cell surface to mediate adhesion (see **Figure 2.9**) [78,113,114]. SRRPs are a unique class of proteins, in that they all have a large domain consisting of hundreds of alternating serine residues, called serine-rich repeats (SRR). The SRR domains are extensively *O*-glycosylated, which occurs in the cytoplasm, and then transported to the cell surface thereafter [78,115].

From N- to C- termini, the defining domains of SRRPs include:

1) a signal sequence and export-targeting region at the *N*-terminus, denoted SS.

2) a short serine-repeat region ~50-170 residues in length, denoted SRR1

3) a unique non-repeat domain, usually 250-600 residues in length, that mediates adhesion, called the binding region - BR

4) a second lengthy SRR region about ~400-4000 residues, denoted SRR2

5) a highly conserved cell-wall anchoring motif and sorting signal (LPXTG) at the C-terminus, that forms part of the cell-wall anchor domain, denoted CWA [115]. An enzyme called Sortase catalyses the transpeptidation reaction between the LPXTG sorting signal and a polypeptide substrate within the cell wall or to another pilus [56].

Figure 2.10 provides a diagram of the domain architectures of known SRRPs drawn to scale, including those of the newly discovered SRRPs from *L. reuteri* being presented in this thesis: Lr0906_SRRP and Lr70902_SRRP, from ATCC 53608 and 100-23 strains, respectively. The SS-SRR1-BR-SRR2-CWA domain arrangement is consistently conserved amongst the diverse SRRPs. Despite their variable lengths

the individual binding domains are highly diverse, their structures and properties being dictated by their function, as will be discussed in section 2.4.3.1. A feature of SRRPs is that they have a strong net acidic charge, whereas the binding regions (BR) may have a charge that is either strongly acidic or basic (see **Table 2.2**) and this property may correlate to their specific adhesive function [108].

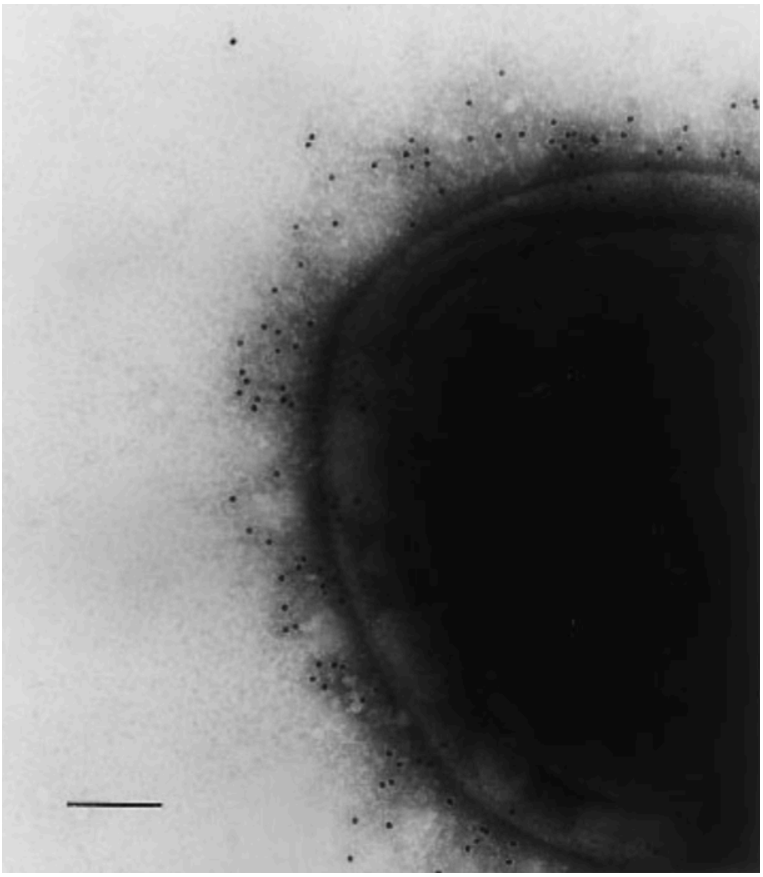


Figure 2.9: Electron micrograph showing localisation of the 200 kDa FapI SRRP on *S. parasanguinis* FW213 fimbriae [113]**.**
Fap1 was visualised on the fimbrae by immunogold electron microscopy, where FW213 was labelled with Fap1 antigen and 10nm gold-conjugated secondary antibody. The black dots represent the conjugated gold beads that are bound to the fimbriae, and the horizontal bar equals 0.1 µm.

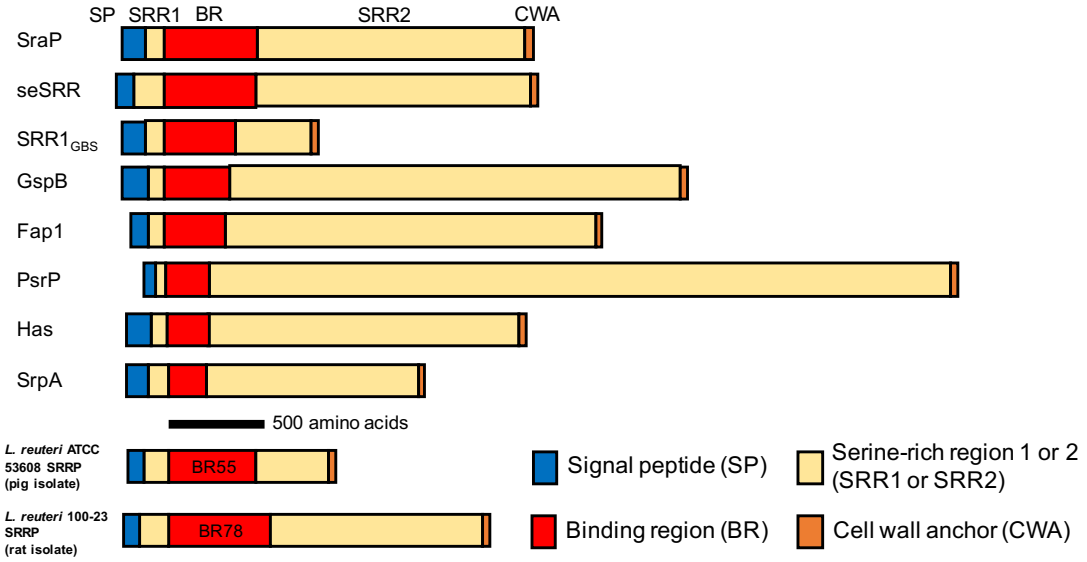


Figure 2.10: Schematic of SRRPs drawn to scale, adapted from [116]**.**
The domain architecture of selected SRRPs aligned by the N-terminus of their binding regions (BR) and listed in order of decreasing BR length. The last two SRRPs belong to *L. reuteri* ATCC 53608 and 100-23, with BRs BR55 and BR78, respectively. The SRRPs in this figure are *S. aureus* SraP [117]; *S. epidermidis* seSRR [116], *S. agalactiae* $SRR1_{GBS}$ [118], *S. gordonii* M99 GspB [116], *S. parasanguinis* Fap1 [113], [78], *S. pneumoniae* PsrP [111], *S. gordonii* Challis Hsa [119], and *S. sanguinis* SrpA [120].

Table 2.1: Serine repeat sequences in the family of SRRPs. Adapted from [108]**.**

Species and strain (protein name)	Serine-repeat sequence	
	SRR I	**SRR II**
Strep. parasanguinis (Fap1)	SESVSESVSI	SESVSESVS1
L. johnsonii NCC533 (LJ1711)	SQSESI	SESL(/I)SESVSESV
Strep. sanguinis Sk36 (SrpA)	SASISA	SASTSASTSASV
Strep. agalactiae 2603V/R (Srr-1)	SV(/A)SV(/A)SQ	SASASVSASQ
Strep. agalactiae J48 (Srr-2)	SVSM	SAST(/M)
Staph. aureus (SraP)	SSS	SESISSSESVSN
Staph. aureus NCTC 8325 (SAOUHSC_02990)	STST	STSLSD
Staph. epidermidis ATCC 12228 (SE2249)	SSEK	SDSNSASTSL
L. johnsonii NCC533 (LJ_0391)	SLS	SL(/M)SNSVSMSE
Strep. gordonii M99 (GspB)	SAS	SASTSASVSAST(/E)
Strep. gordonii DL1 (Hsa)	SA(/L)S	SAST(/E)SASTSASV
Strep. pneumonia TIGR4 (PsrP)	SASE(/T)	SASTSASA(/E)
Staph. haemolyticus JCSC1435 (SH0326)	STST	STSTSVSD
***L. reuteri* ATCC 53608 (Lr0906)**	**ST(/VGQ)S**	**SLSNSVSMSE**
***L. reuteri* 100-23 (Lr70902)**	**ST**	**SLSNSVSMSE**

Table 2.2: pI values of SRRPs.

This table has been adapted from [108]. Entries for *L. reuteri* Lr0906 and Lr70902 are emboldened and the pI of unglycosylated protein was calculated using the online ExPASy ProtParam tool [121].

Group	Species and strain (protein name)	pI	
		Full Protein	**Binding region (BR)**
Acidic	*S. parasanguinis* (Fap1)	3.26	4.85 (179-505)
	L. johnsonii NCC533 (LJ1711)	4.23	4.61 (181-1412)
	S. sanguinis Sk36 (SrpA)	4.96	5.41 (242-446)
	S. cristatus CC5A (SrpA)	5.60	6.21 (256-739)
	S. agalacticae 2603V/R (Srr-1)	4.94	4.63 (200-641)
	S. agalacticae J48 (Srr-2)	4.21	4.53 (212-531)
	S. aureus (SraP)	4.20	5.69 (231-751)
	S. aureus NCTC 8325 (SAOUHSC_02990)	4.04	5.61 (239-751)
	S. epidermidis ATCC 12228 (SE2249)	4.03	6.00 (263-790)
	***L. reuteri* ATCC 53608 (Lr0906)**	**4.38**	**4.76 (223-672)**
	***L. reuteri* 100-23 (Lr70902)**	**4.12**	**4.81 (202-687)**
Basic	*L. johnsonii* NCC533 (LJ_0391)	4.70	8.13 (232-361)
	S. gordonii M99 (GspB)	3.96	9.51 (237-603)
	S. gordonii DLi (Hsa)	4.15	8.74 (221-449)
	S. pneumonia TIGR4 (PsrP)	3.90	9.91 (122-379)
	S. haemolyticus JCSC1435 (SH0326)	3.58	9.25 (301-496)

2.4.2 The length of the SRR2 domain is vital for mediating bacterial adhesion

The SRR2 domain is often the largest in SRRP adhesins, largely dominating its overall size; the number and amino acid composition of serine repeats show a high variability from one SRRP to another, and is the main reason why SRRPs vary so much in length, as can be seen in **Figure 2.10**. The repetitive pattern of serine-repeats unique to each SRRP is shown in **Table 2.1**.

Due to the conserved C-terminal location of the CWA domain and the increased length of the SRR2 domain, it is postulated that the chief function of the SRR2 region is to extend the adhesive BR domain beyond the bacterial cell surface to enable attachment [111, 115]. A study that strongly supports this hypothesis was conducted with *S. pneumoniae* PsrP, in which a PsrP deficient mutant was complemented with a truncated version of the SRRP, consisting of only 33 serine repeats in SRR2 instead of the wild-type 539, resulting in non-adhesion of the transformed bacteria to lung

cells. However, the same truncated construct was able to restore bacterial adhesion upon transformation into an unencapsulated PsrP deficient mutant. This indicated that the length of the SRR2 domain was essential for exposing the adhesin's BR beyond the obstructive cell capsule to mediate adhesion to its substrate(s) [111]. Furthermore, it also suggests that the number of repeats in the SRR2 regions could have evolved to surpass the extracellular components, characteristic of each species and strain. For example, in **Figure 2.10** the *L. reuteri* SRRPs have comparable domain lengths, except for SRR2, which is about three times longer for Lr70902_SRRP compared to Lr0906_SRRP. This difference could have arisen due to a selection factor within the rodent forestomach requiring a longer SRR2 domain to project the BR beyond the bacterial cell surface for effective adhesion to the gut. This is another example of a divergent phenotypic trait in two *L. reuteri* isolates which has likely arisen due to host driven evolution, which was discussed above in section 2.1.3.

Another noteworthy feature of the SRR regions is the fundamental occurrence of alternating serine residues in the form of dipeptide repeats: S(X), where X is any amino acid (see **Table 2.1**). Ramboarina et. al. [78] suggest that the dipeptide repeats in SRR1 And, more importantly in SRR2 could allow the formation of extended coiled structures, such as that observed in known proteins with repetitive amino acid sequences, like collagen and silk. In their study on the Fap1 SRRP from *P. parasanguinis*, with the help of electron microscopy data, they have proposed that Fap1 strands form a super-helical extended structure with a principal repeat at 6.5 nm and serine side chains exposed for O-glycosylation. The first proposed model, described a super-coiled Fap1 monomer with 6.5 nm repeating units along the SRR regions, and protomer length of 250-300 nm. The second model represented multiple polymerised subunits in which Fap1 monomers undergo head to tail cross-linking. **Figure 2.11** below presents a reconstruction of the proposed models.

The above studies, therefore, provide evidences to support the significance of the amino acid composition and length of SRR regions in influencing the functionality of SRRP proteins in mediating bacterial adhesion to their host surfaces.

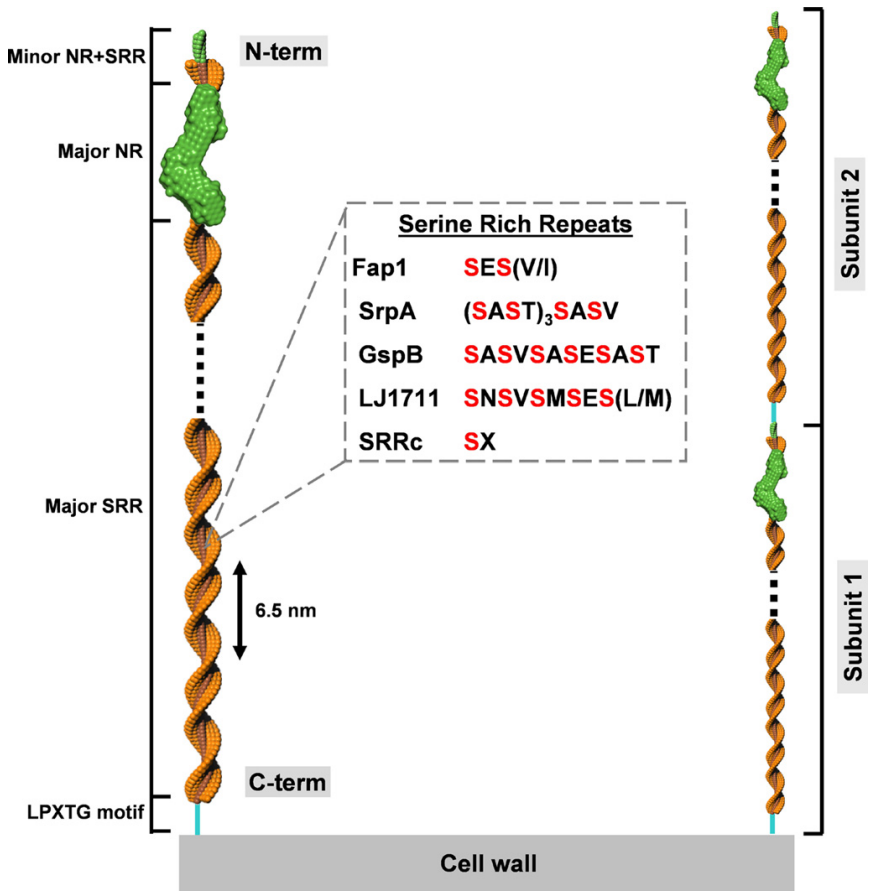


Figure 2.11: Models for the super-coiled strand of Fap1 fimbriae, taken from [78] The CWA is coloured cyan, the SRR domains are coloured dark orange (protein) and light orange (glycan), the BRs are green. The larger BR represents the SAXS envelope observed for Fap1 BR. The model on the left represents the first proposed super-coiled helix model of Fap1 protein; the model on the right depicts the polymerised Fap1 model, with only two subunits shown.

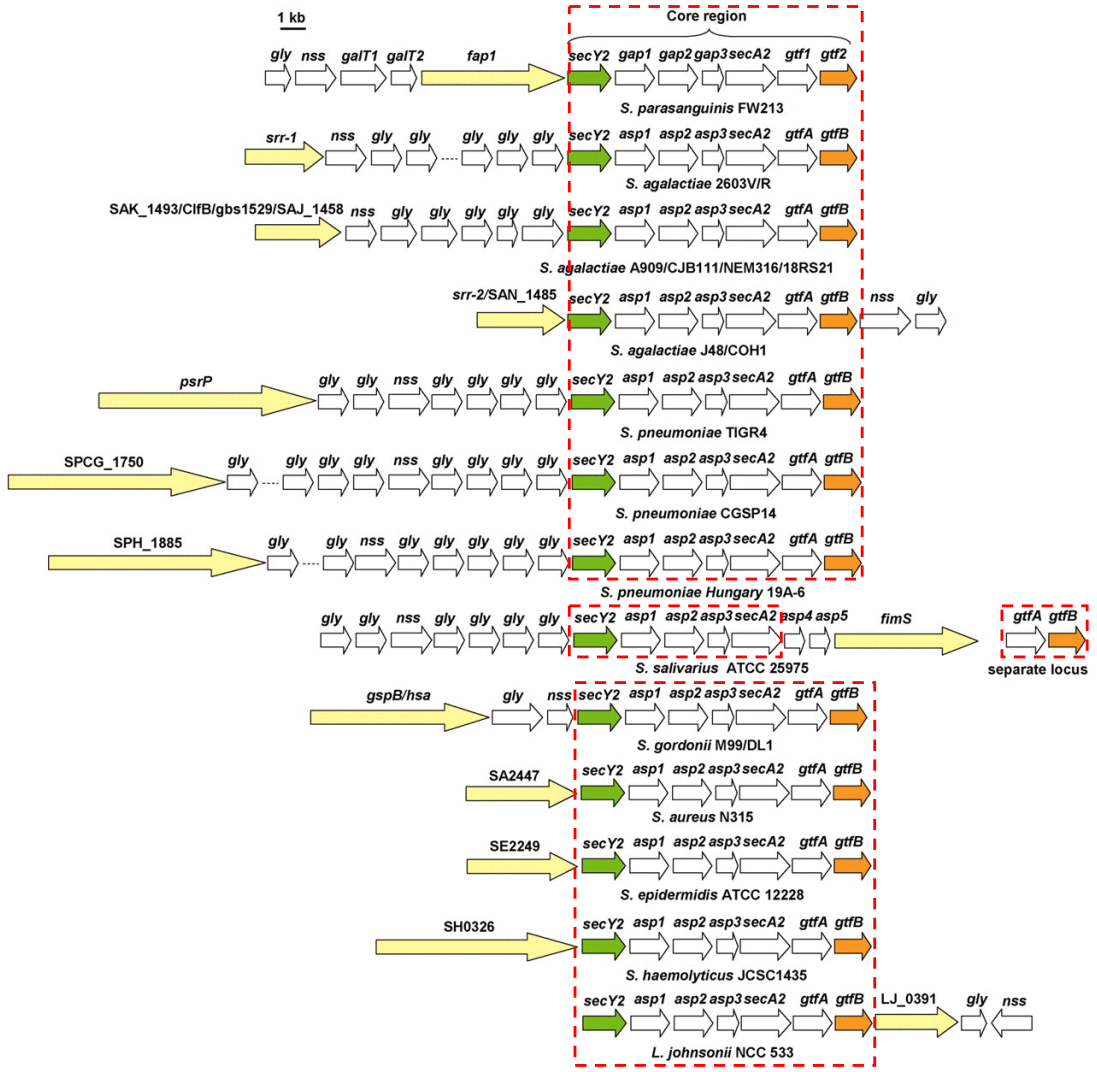
2.4.3 SRRP export is controlled by a conserved SecA2/Y2 secretion system

The secA2/Y2 accessory secretion system was introduced in section 2.1.2 as the secA2 system, which is found amongst selected species in class Bacilli. The first gene cluster associated with SRRP biogenesis and transport was identified in *S. gordonii* M99 for GspB, as a secA2/Y2 related secretion system [81,119,122-124]. Subsequently, similar glycosylation and secretion loci associated with SRRP-like proteins were discovered and characterised for oral streptococci like *S. parasanguis* [83], *S. sanguinis* [125] and *S. cristatus* [120], and in streptococcal and staphylococcal

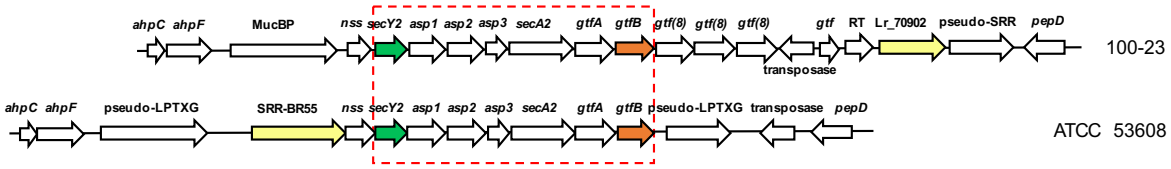
pathogens such as *S. pnuemoniae* [126], *S. aureus* [127] and *S. haemolyticus* [128]. The secA2/Y2 gene cluster consists of two regions:

1) A characteristic core of seven genes, conserved in all Gram-positive strains that express SRRPs. Five of these genes express proteins required for SRRP transport, namely secY2, asp1-3 (or gap1-3) and secA2. Two of the genes, gtfA and gtfB (or gtf1 and gtf2), are glycosyltransferases involved in the first step of SRRP glycosylation, likely conferring protein stability [115].
2) A region composed of glycosyltransferase genes that are highly diverse in content and organization for various SRRPs.

Figure 2.12A displays a scheme of secA2/Y2 clusters found in selected SRRP producing bacteria, which has been adapted from [108] and [115]. Zhou and Wu [108] have discussed that the core region of the secA2/Y2 gene cluster may have evolved to provide bacteria with a specific mechanism to produce surface SRR adhesins, while differences in glycosyl-transferases may allow them to modify SRRPs to suit their distinct ecological niches. Recognising the role of the conserved secA2/Y2 loci in SRRP biogenesis is allowing the identification of other SRRP expressing bacteria, by detecting homologous loci in their genomes. For example, **Figure 2.12B** shows the recently discovered secA2/Y2 gene cluster in *L. reuteri* ATCC 53608 [91] and 100-23 [48], which produce the SRR adhesins Lr0906 and Lr70902, respectively. Last year, Wegmann et. al. [3] reported similarly conserved, SRRP associated SecA2/Y2 clusters in the genomes of 5 other porcine *L. reuteri* isolates. In sections 2.1.2 and 2.1.3, it was highlighted that the SecA2 cluster was confined to only some rodent and porcine *L. reuteri* isolates, and it was not a core genomic feature found universally across all *L. reuteri.* Therefore, its limited distribution amongst lactobacilli species implies that this cluster was acquired by HGT by selected Lactobacillus lineages [3].



A. Genetic organisation of loci encoding SRRPs in streptococcus, staphylococcus and lactobacillus, adapted from [108] and [115]. SRRP genes are shown in yellow; the green arrow marks the first gene, and the orange arrow marks the last gene in the conserved core region (red dotted line).



B. A similarly conserved *sec*A2/Y2 secretion system in *L. reuteri* strains 100-23 (isolated from rat) [48] and ATCC 53608 (isolated from pig) [91].

Figure 2.12: Conservation of genes in the *sec*A2/Y2 loci of Gram positive bacteria species, involved in the biosynthesis and secretion of SRRP adhesins.

2.4.3.1 Role of the SRRP binding region in substrate recognition and adhesion

While the overall domain architecture is highly conserved amongst all known SRRPs (see **Figure 2.10**), their ligand binding regions (BR) are much more divergent in sequence and structure, and to the types of ligands to which they bind [115,129]. Atomic-resolution structures have been determined for the BRs of seven SRRPs to date and **Figure 2.13** compiles all the known BR structures to provide a visual comparison of their subdomains. **Table 2.3** is an extension of **Figure 2.13** and briefly describes the role of each BR in bacterial adhesion and pathogenicity.

The variation in secondary and tertiary structures of SRRP BRs, dictated by the nature of their binding ligands, can clearly be observed in **Figure 2.13**. The subdomain organisation in each BR has evolved to bestow unique structure-function properties that not only facilitate, but also regulate selective binding of SRRPs to host-receptors.

For example, the fimbriae-associated protein 1 (Fap1) is an *S. parasanguinis* adhesin, aiding the colonisation of oral cavity via dental plaque formation [130,131]. The BR of Fap1, spanning from residues 106-437, is composed of two subdomains, named Fap1-NRα and Fap1-NRβ that are protruded away from the bacterial surface by the SRR2 region described in section 2.4.2, which encourages adhesion to the salivary pellicle. The NMR solved structure of the Fap1-NRα subdomain from 129-206 comprises of three helices arranged in a bundle. The Fap1-NRβ subdomain structure, from 237-425, was solved by X-ray diffraction and reveals a CnaA type folding pattern [78]. It has been shown that Fap1's adhesive properties are controlled by a pH switch in which a reduction in pH from 8.0 to 5.0 results in a an 'open' conformation between Fap1-NRα and Fap1-NRβ, due to subtle rearrangements of the 3-helix bundle together with a change in electrostatic potential, which subsequently allows adhesion and biofilm formation of *S. parasanguinis* in acidic oral environments [78]. This pH dependent change is portrayed in **Figure 2.14**. Further experiments confirmed that an increase in biofilm density correlated to a decrease in pH [131]. As the theoretical pI of Fap1 BR is about 4.85, it has been proposed [131] that in addition to its activation

under low pH conditions, caused by fermentation of sugars in the oral cavity, the BR domain would likely have a net zero charge, in turn promoting inter-domain aggregation and further enhancing biofilm formation.

Another example of the unique structure-function characteristics of the BR subdomains in directing highly regulated ligand recognition and adhesion is SraP from *Staphylococcus aureus*. SraP is crucial for the pathogenesis of human endocarditis via the interaction of its BR with sialylated glycans on human platelets and/or epithelial cells [109,117]. The structure of its BR has been solved by X-ray crystallography, revealing that the BR consists of four discrete subdomains (see **Figure 2.15**): the first is a legume-lectin (L-lectin) type module, the second subdomain adopts by a ß -grasp fold (ß–GF) similar to Ig-binding proteins, and the last two cadherin-like (CDHL) subdomains occur in tandem, and are named as such because they resemble eukaryotic cadherins that coordinate to Ca^{2+} [117]. Ligand-binding and adhesion occur via the L-lectin module, where it has been shown to specifically bind to N-acetylneuraminic acid (Neu5Ac). SAXS and molecular dynamic simulations indicated that the following three subdomains act as a rigid stem extending the L-lectin module outwards to facilitate binding, and that the structural integrity and rigidity of the stem highly relies on Ca^{2+} complexation by the last two CDHL subdomains. In the absence of Ca^{2+} the SraP BR underwent significant conformational changes due to loss of rigidity in the CDHL regions, which then hindered ligand binding [117]; see **Figure 2.15C and D.**

Finally, PsrP from *Streptococcus pneumoniae* shows unique structure-mediated binding properties. PsrP is a vital virulence factor for pneumococcal invasive disease, by mediating bacterial adhesion onto keratin 10 (KRT10) molecules found on lung epithelium cells [111]. The PsrP BR (122-385) is different from BRs of other structurally characterized SRRPs in having a single structural module. This module adopts a novel variation of the Dev-IgG fold, typical for microbial surface proteins recognizing adhesive matrix molecules [132]. Residues 122-166 have been shown to promote intra-species aggregation, contributing to biofilm formation [133], whereas

region 187-385, for which a crystal structure has been obtained [132] consists of the KRT10 binding domain, having a Dev-IgG fold (**Figure 2.16**). Two mechanisms have been proposed for PsrP BR binding to KRT10 by Schulte et. al. [132]: 1) A paperclip-like region was observed in the KRT10 binding segment of PsrP BR, comprising of loops formed by residues 268-295 and 305-324. It was proposed that conformational rearrangements of these loops could facilitate KRT10 recognition and binding; 2) A basic-binding groove observed on the other side of the compressed protein barrel was postulated to facilitate KRT10 adhesion by accommodating an acidic helical rod domain of KRT10, via electrostatic interactions. The structural features supporting these mechanisms are highlighted in **Figure 2.16C** and **D**. Indeed, as has been presented previously in **Table 2.2**, full SRRPs are distinctive in having an unusually acidic pI, whereas their binding regions may have pIs that are highly acidic or basic. Together with self-aggregating properties conferred by residues 122-166, a positively charged BR could have evolved to support PsrP's adhesive function in binding to negatively charged regions of KRT10, causing *S. pneumoniae* infection and biofilm formation on the lung epithelium.

To conclude, supporting literature on structural characterisation of BRs have shown that relative to the entire SRRP, the BR domain is much more structurally diverse, and that it is this diversity that enables each SRRP to permit bacterial infection and biofilm formation in their respective niches.

Table 2.3: BR of SRRPs for which structures have been published (except for Hsa) , and their role in pathogenesis.

Species and SRRP	BR region	Role in disease	Binding ligands	PDB ID and reference
S. parasanguinis Fap1	106-437	Oral infectious disease; facilitates pH dependent dental plaque formation by adhesion to the salivary pellicle.	Exact binding ligand(s) unknown	2KUB, 2X12, [78], [131]
S. gordonii M99 GspB	233-617	Endocarditis; BR mediates binding to human platelets by interaction with sialyl-T antigen on the GPIbα receptor.	Sialyl-T anitgen	3QC6, [116], [134]
S. gordonii DL1 Hsa	219-454	Endocarditis; BR mediates attachment to human platelets by adhering to either sialyl-T antigen or sialyllactose.	Sialyl-T antigen and sialyllactose	PDB: N/A, [116], [134]
S. sanguinis SrpA	240-453	Endocarditis; mediates infection by pathogen attachment to host platelets via the BR.	Human Sialylated glycan(s)	5EQ2, [135]
S. agalactiae $SRR1_{GBS}$	303-641	Neonatal meningitis; the BR is postulated to employ a dock-lock-latch mechanism for adhesion. Keratin 4 binding mediates colonisation of the female genital tract, subsequently risking neonatal infection; fibrinogen binding mediates adhesion to human brain microvascular endothelial cells.	Human fibrinogen and keratin 4	4MBO, [136], [137]
S. agalactiae $SRR2_{GBS}$	303-641		Human fibrinogen	4MBR, [136]
S. aureus SraP	245-751	Endocarditis; the L-lectin subdomain promotes adhesion and invasion of host platelet cells via sialylated receptors.	Neu5Ac	4M0(0-3), [117]
S. pneumoniae PsrP	187-385	Pneumonia; specifically forms biofilms on lung epithelial cells via self-oligomerisation and adhesion to keratin 10, both facilitated by the BR.	Keratin 10	3ZGH/I, [132]

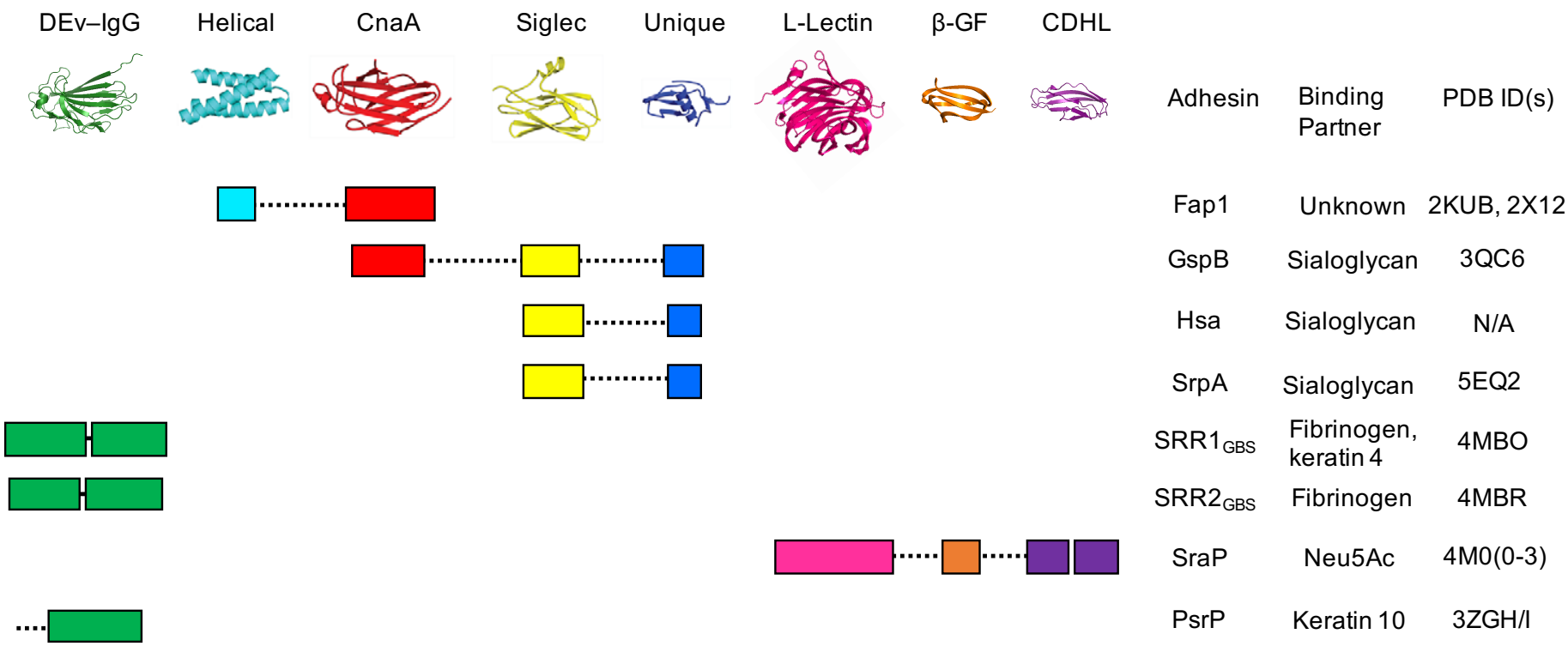


Figure 2.13: Modular organization of the binding regions within SRRP adhesins (updated from the scheme presented in [116]**).** Coloured rectangles represent the individual subdomains comprising the binding region of each SRRP, and are drawn to scale. For example, the binding region for *S. aureus* **SraP** [117] comprises of four subdomains, the first one having an L-lectin type fold, the second one with a ß-GF and the last two subdomains are each folded separately in a CDHL motif. Other SRRPs in this diagram are *S. gordonii* **GspB** [116], *S. parasanguinis* **Fap1** [78], *S. gordonii* **Hsa** [138], *S. sanguinis* **SrpA** [135], *S. agalactiae* **SRR1$_{GBS}$** and **SRR2$_{GBS}$** [136], and *S. pneumoniae* **PsrP** [132].

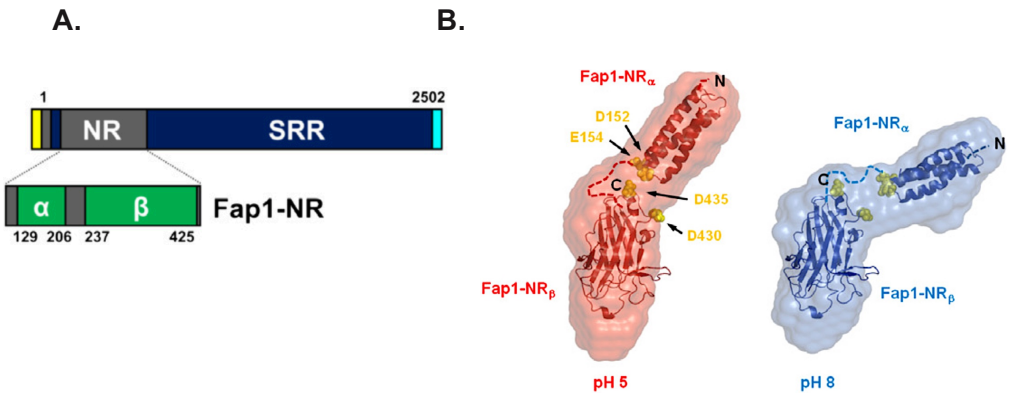


Figure 2.14: Subdomain boundaries and electrostatic activation of the Fap1 BR. From [131]. **A**. Fap1 SRRP scheme, with numbering for mature protein. The BR (grey) is expanded to show the NRα and NRβ subdomains. **B.** High-resolution structures of NRα and NRβ docked into low-resolution Fap1-BR SAXS models at pH 5.0 and 8.0, connected via a flexible linker (dashed lines). Yellow spheres denote residues D152, E154 (NRα), D430 and D435 (NRβ). At pH 5.0, the Fap1-BR conformation changes into an 'open' form, caused by a reduction in electrostatic repulsion, and possible formation of salt bridge(s) between D152/E154 and D430/D435.

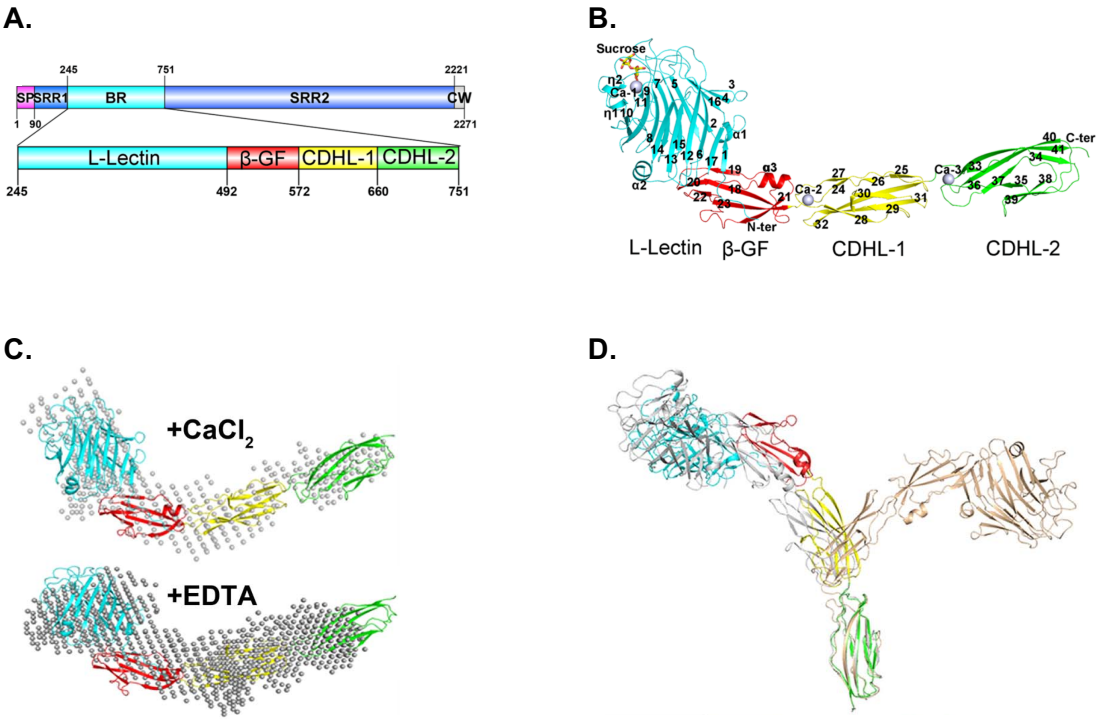


Figure 2.15: SraP BR subdomains and their Ca^{2+} driven conformational change Adapted from [117]. **A.** Scheme of SraP SRRP; the BR is expanded to show the four subdomains. **B.** Cartoon representation of overall SraP BR structure. **C.** SAXS data envelopes for SraP BR in the presence or absence of Ca^{2+}, showing an obvious conformational change. **D.** The final states of molecular dynamics simulations of SraP BR superimposed with its crystal structure. Ca^{2+} bound SraP BR (grey) remains extended, whereas removal of Ca^{2+} causes curling of the SraP BR (peach).

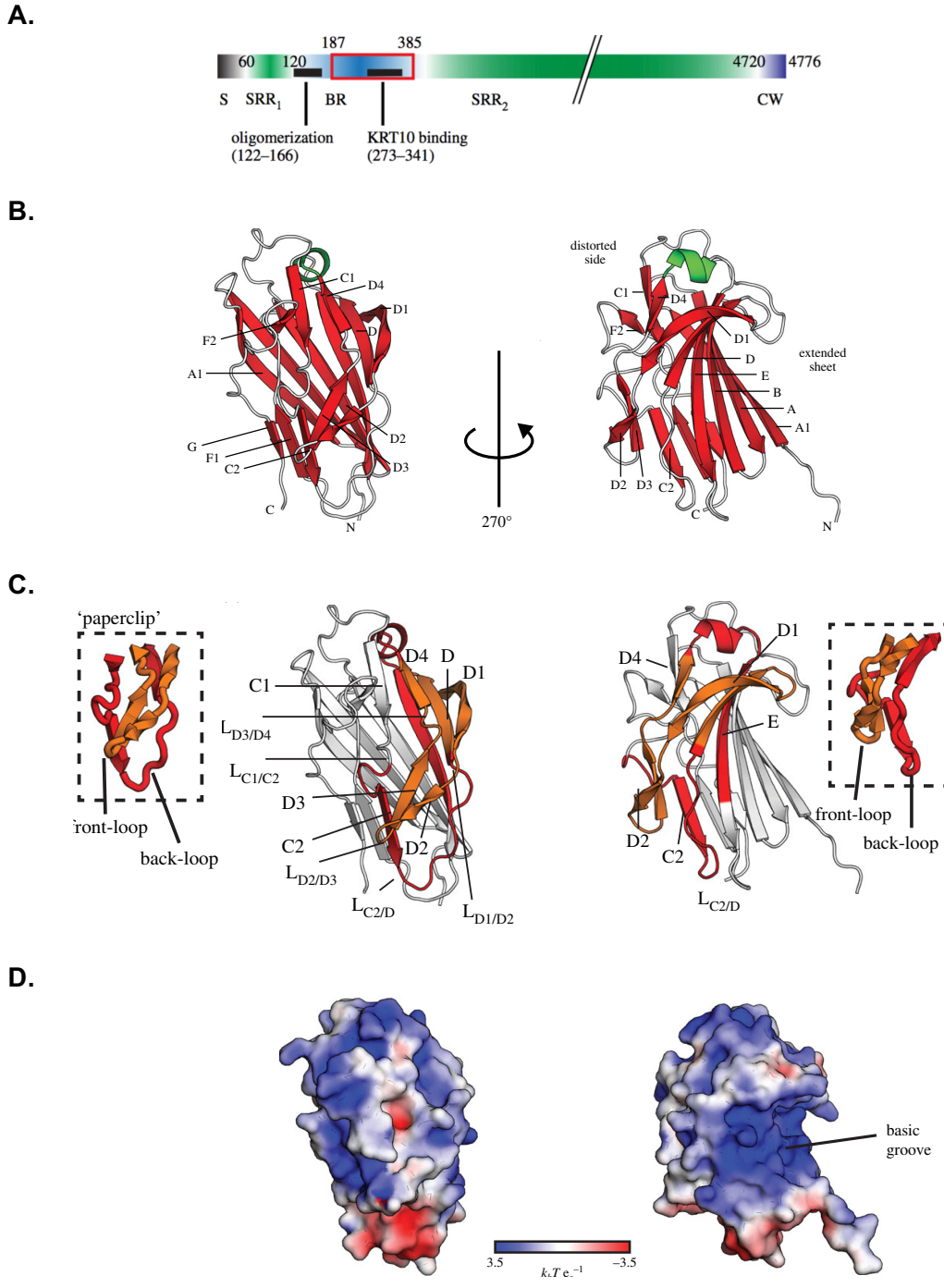


Figure 2.16: Representation of binding features in the crystal structure of the PsrP KRT10 binding subdomain.

Adapted from [132]. **A.** PsrP domain organisation, comprising the BR from 122-385 showing two distinct subregions for oligomerisation and KRT10 binding. **B.** Two orientations of the BR crystal structure shown as ribbon diagrams; the first side is defined by well-ordered loops and two sets of β-sheet belts, whereas the other side shows an extended and twisted antiparallel β-sheet. **C.** Diagram representing the putative paperclip region with front and back loops formed by residues 268-295 (red) and 305-324 (orange), respectively. **D**. Surface electrostatic potential map of PsrP BR showing a highly basic binding groove within the extended antiparallel β-sheet, that could accept the acidic rod domain of KRT10.

2.5 Research Aims

Biofilm development and maintenance is a complex process, with several molecular steps involved in its establishment, from initial host surface recognition mediated by specialised surface adhesins, to maintaining thriving macrocolonies of bacteria over extended time periods, which is proposed to be managed by mechanisms such as quorum sensing and cell lysis (see sections 2.1.3 and 2.2.3) [92]. Previous studies conducted with *L. reuteri* highlighted the importance of the SRRP adhesin in successful colonization of the non-glandular stomach epithelium in rodents, pigs and poultry, although other surface adhesins have been identified in *L. reuteri* strains isolated from different hosts.

The investigation aims to elucidate the structure of BR55 and BR78 binding domains by X-ray crystallography and small-angle X-ray scattering (SAXS), in attempts to answer the following questions:

- Can the 3D structure provide clues towards the potential ligands recognised by *L. reuteri* SRRPs?
- How divergent are the structural features between the two BRs?
- Are there any structural features in the full-length SRRP that may contribute to host-specific binding?
- What are the structural similarities between the BR of *L. reuteri* SRRPs and those characterised from pathogenic Gram-positive bacteria in **Table 2.3**.
- What is the likely binding mechanism of *L. reuteri* SRRPs?

CHAPTER 3

PRINCIPALS OF X-RAY CRYSTALLOGRAPHY

3 PRINCIPLES OF X-RAY CRYSTALLOGRAPHY

The process for obtaining a three-dimensional (3-D) protein structure can be divided into six steps: 1) expression and purification of the target protein molecules; 2) search of initial crystallization conditions; 3) crystal quality optimisation; 4) diffraction data collection; 5) data processing, model building and refinement; and 6) analysis of the refined model [139, 140]. This chapter will explore the process from steps 2 to 5, to provide a deeper understanding of the journey from crystal growth to final model.

3.1 Protein Crystallisation

A prerequisite for an X-ray diffraction experiment is the growth of high-quality single protein crystals. Protein crystallization is the controlled and gradual precipitation of homogenous protein from a supersaturated aqueous solution under conditions that do not denature or degrade the protein [139]. Typically, a crystal of 0.1 – 0.2 mm (in its shortest dimension) is needed for diffraction studies, however, modern methods of data collection are enabling analyses of increasingly smaller microcrystals, such as the VMXi [141] and VMXm [142] beamlines, under construction at the Diamond Light Source (DLS) synchrotron in Oxfordshire, UK.

Proteins crystallise due to weak surface interactions between protein molecules in solution, chiefly hydrogen bonding between hydrated protein surfaces, but also hydrodrophobic and electrostatic interactions. The interactions must occur specifically and consistently between all molecules, thus bringing them together in a repetitive and ordered fashion in a 3-D crystal lattice [140]. Thus, it is crucial for protein samples to be at a high concentration (between the 5-50 mg/mL range) and homogenous: i.e. of high purity (usually >90-95%) and as monodisperse as possible to facilitate ordered and regular inter-protein aggregation, leading to crystal nuclei formation.

The crystal growth process can be depicted in a two-dimensional phase diagram (**Figure *3.1***), where the undersaturated and supersaturated states of a protein are shown as a function of protein concentration against an adjustable parameter. A protein solution attains supersaturation when the protein concentration exceeds its solubility limit. The supersaturated region of the phase diagram can be further divided into three zones. The first is the “precipitation” zone in which a protein solution is highly supersaturated, leading to the formation of amorphous aggregates. The second region of intermediate supersaturation is called the “nucleation zone”, where protein nucleation and (often rapid) growth occurs – this is where tiny crystal clusters or microcrystals are usually found. The third region of lower supersaturation is the “metastable zone”, where the controlled growth of single crystals is favoured [140]. These regions are influenced by kinetic parameters and so the boundaries between them are not well defined, but they are relatively illustrated in **Figure 3.1.**

During a crystallization trial, the strategy most often employed is to induce nucleation at a lower level of supersaturation just within the “nucleation zone” (a higher level of supersaturation may lead to the formation of too many small nuclei, too fast, thus leading to the formation of precipitate or microcrystals). Following nuclei formation, the concentration of protein in the solution gradually decreases, driving the system into the metastable zone, where growth occurs slowly. However, controlling this process is not straightforward, as numerous factors affect protein crystal growth, as listed in **Table 3.1** below. Factors that can be controlled with relative ease, and are therefore, often altered in trials are protein concentration, precipitant type and concentration, pH, temperature, additives and the use of ligands [140].

One widely used crystallisation technique is the sitting-drop vapour diffusion method. In a typical vapour diffusion experiment, protein solution mixed with a crystallisation buffer is dispensed as a drop and placed in a sealed cell with a reservoir of the crystallisation buffer, as depicted in **Figure 3.2**. The concentration gradient created due to the dilution of crystallisation buffer by the protein sample in the drop, compared to its original concentration in the reservoir causes water vapour from the protein drop

to diffuse to the reservoir, until the precipitant concentration is the same in both solutions, thus increasing the concentration of protein and precipitant in the drop. Under ideal conditions, this gradual 'drying out' of the protein sample increases the frequency of specific contacts between protein molecules, permitting the formation of crystal nuclei [139]. Specific precipitants, pH, salts, protein concentrations and incubation temperatures are required for the successful crystallisation of each protein, which are often not possible to rationally predict, and so commercially available sparse-matrix screens and grid screens are often used to search for buffer conditions that would allow the formation of protein nuclei or single crystals, such as is the case of crystallisation trials reported in this thesis, in sections 4.2.1, 4.3.1 and 6.3.5.

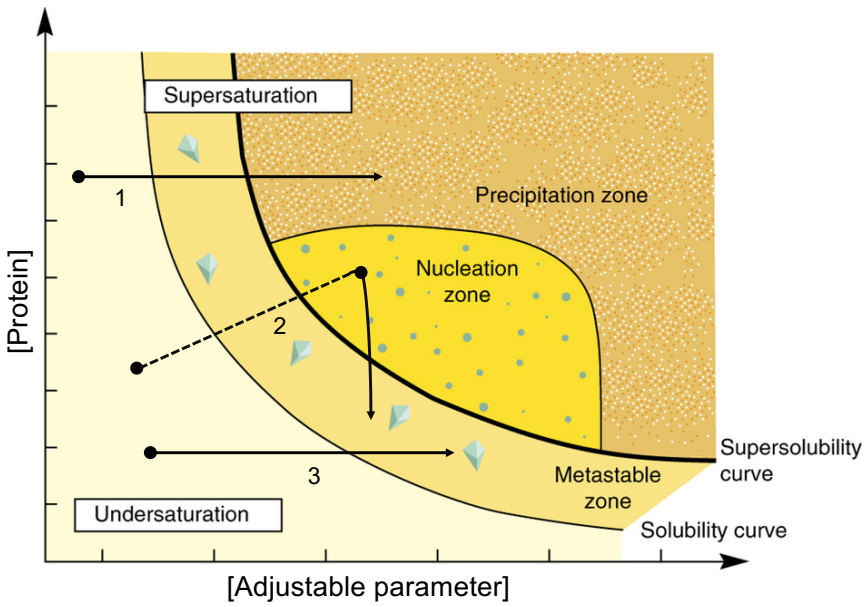


Figure 3.1: Schematic of a protein crystallization phase diagram.

Adapted from [143]. The solubility curve defines the concentration of protein in the solute that is in equilibrium with crystals. The supersolubility curve separates conditions where spontaneous nucleation or precipitation occurs from conditions where the crystallization solution remains clear if left undisturbed. The adjustable parameter can be precipitant, additive concentration, pH or temperature and so on. 1) Sometimes specific parameters may create conditions that drive the solution into the precipitation zone, where aggregates and amorphous precipitates are formed 2) A common strategy for growing large crystals is to allow nucleation to occur, and then to move to conditions in the metastable zone until crystal growth ceases 3) Under the ideal combination of parameters a supersaturated protein solution is within the metastable zone, where controlled crystal growth up to its maximum size occurs.

Table 3.1: Parameters affecting protein crystal growth.

The table has been adapted from [140]. Parameters in *italics* have been modified for crystallisation trials reported in this thesis, in Chapters 4 and 6.

Physical Factors	Chemical Factors	Biochemical factors
Temperature	*Precipitant type*	Sample purity
Time	*Precipitant concentration*	Sample homogeneity
Pressure	*pH*	*Intrinsic disorder*
Gravity	*Buffer type*	Sequence modifications
Magnetic fields	*Ionic strength*	Post-translational modifications
Electric fields	*Sample concentration*	Chemical modifications
Dielectric properties	Metal ions	Aggregation
Viscosity	Polyions	Proteolysis
Vibrations and sound	Detergents	Sample pI
Equilibration rate	Heavy metals	*Ligands, co-factors, inhibitors*
Methodology	Small molecule impurities	
Nucleants	Crosslinkers	
Crystallisation surface	Reagent source	
Sample handling	Reagent formulation	

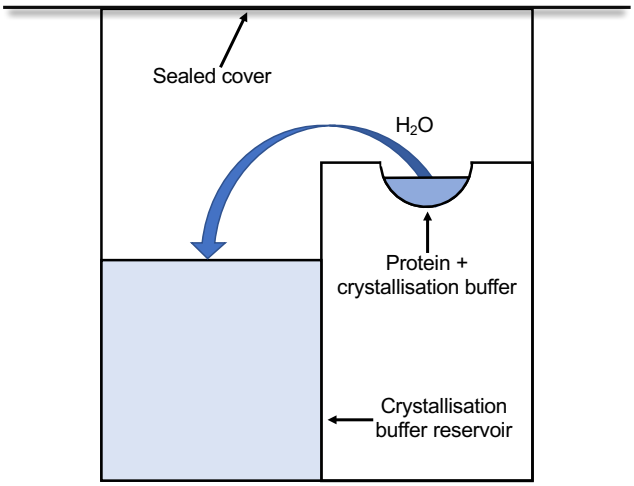


Figure 3.2: Scheme of sitting-drop vapour diffusion protein crystallisation

Sitting drop vapour diffusion uses a sealed environment in which solvent vapour diffuses from the protein/crystallisation buffer drop to the reservoir, until the concentrations of precipitant in the drop and reservoir solution attain equilibrium.

3.2 X-Ray Diffraction

3.2.1 Geometric principles of diffraction

The unit cell: this is the simplest repeating unit of atomic arrangement in a crystal, from which the entire lattice can be built up in three dimensions. Unit cell dimensions are given by six numbers: the lengths of three unique edges **a, b,** and **c**, and three unique angles *α*, *β*, and *γ*. A combination of these edges and angles gives rise to seven basic crystal systems: a cell in which $a \neq b \neq c$ and $\alpha \neq \beta \neq \gamma$ is called *triclinic*. If $a \neq b \neq c$ and $\alpha = \gamma = 90$, and $\beta > 90^{\circ}$,the cell is *monoclinic*. If $a = b$, $\alpha = \beta = 90^{\circ}$ and $\gamma = 120^{\circ}$, the cell is hexagonal. When all three cell angles are 90°: if $a = b = c$, the cell is *cubic*, if $a = b \neq c$, the cell is *tetragonal*, and if $a \neq b \neq c$, the cell is *orthorhombic.*

Space groups: during data collection, the internal symmetry of the unit cell is crucial. The symmetry of a unit cell together with its contents is defined by its space group, which describes the cell's internal symmetry elements. As proteins are chiral, symmetry elements in protein crystals include only translations, rotations and screw axes and so, exclude mirror and inversion symmetry operations. This results in 65 unique space groups for protein crystal systems, rather than the maximum predicted 230 space groups for non-chiral systems [139]. Space groups are represented in the form XY_z, where X denotes the lattice type (*P* for primitive, *I* for body-centered or *F* for face-centred), Y represents the point-group and $_z$ denotes the required translation through transflection or screw-axis symmetry operations. For instance, the space group $P2_12_12_1$ describes a primitive unit cell with three perpendicular twofold screw axes, where the unit cell can be rotated by 180^{o}C (2 fold-symmetry) around each axis and translated ½ of the repeat distance to complete a translation [139]. A full list of space groups can be found in the International Tables for X-ray Crystallography [144].

Miller indices: The most distinct sets of planes in a crystal lattice are those determined by the faces of the unit cells. Planes in a unit cell are defined by a set of numbers called Miller indices, represented as *hkl*. Index *h* represents the number of

planes per unit cell perpendicular to the *x* axis, or **a** edge of the cell. Likewise, the indices *k* and *l* represent the number of planes per unit cell perpendicular to the *y* and *z* axes, or the **b** and **c** edges of the unit cell, respectively. Thus, the Miller indices give the fractional spacing of lattice planes along the unit cell relative to the unit cell origin.

Bragg's law: W. L. Bragg showed that a set of parallel planes with index *hkl* and inter-planar distance d_{hkl} produces a diffracted beam when an incident X-ray beam of wavelength λ meets the planes at an angle θ and are reflected at the same angle, if θ meets the following condition (where *n* is an integer):

$$2d_{hkl} \sin\theta = n\lambda \qquad \text{Equation 3.1}$$

Protein crystals consist of many lattice planes, each with their own Miller indices and unique d_{hkl} values; atoms in a crystal can scatter incident X-rays in various directions and only X-rays reflected from successive planes, where $2d_{hkl} \sin\theta$ is equal to an integral number of wavelengths ($n\lambda$) of the incident X-ray beam, will emerge in phase from the crystal, interefering constructively to produce strong diffracted beams. For other angles of incidence where $2d_{hkl} \sin\theta$ does not equal $n\lambda$, the emergent waves are out of phase and interfere destructively, providing limited recordable diffractions at those angles. Hence, by knowing λ for the incident X-ray beam, and the angle of reflections given by the positions of Bragg spots in the diffraction pattern, one can compute the crystal's unit cell dimensions.

θ is inversely proportional to d_{hkl}, implying that larger unit cells (having many more atoms), with larger inter-planar distances give smaller angles of diffraction and so produce many measurable reflections. Conversely, smaller unit cells (fewer atoms) give larger angles of diffraction, producing fewer measurable reflections [139]. It is also important to briefly highlight the relationship between lattice plane origin and resolution. Planes closer to the origin (indices 000) have larger d_{hkl}s and scatter X-rays to smaller angles, thus providing lower resolution information on molecule, such as its shape and size. Conversely, parallel planes with larger Miller indices are closer to one another and so they scatter X-rays further from the angle of the incident beam, and bear higher resolution data on the finer details of the protein structure [139].

Resolution is measured in Å and can be defined as the minimum inter-planar distance in a lattice that can provide a measurable diffraction of X-rays. The smaller the d_{hkl} within a lattice, the higher the resolution of the data [145].

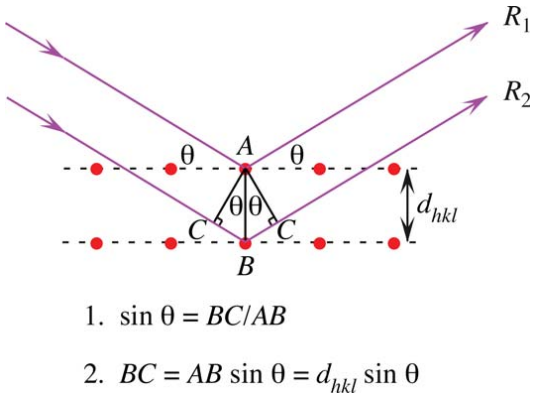


Figure 3.3: A visual representation explaining Bragg's Law
Taken from [139]. The red dots denote two parallel lattice planes with inter-planar spacing d_{hkl}. R_1 and R_2 represent two rays reflected at an angle θ from the planes. ABC is a right-angle triangle, so $\sin\theta = BC/AB$ or $\sin\theta = BC/d_{hkl}$. Therefore, $BC = d_{hkl} \sin\theta$. The additional distance $2BC$ travelled by R_2 thus equals $2d_{hkl} \sin\theta$. Therefore, if the additional distance travelled by R_2 is an integral multiple of λ, then rays R_1 and R_2 will interfere constructively, in accordance with Bragg's Law.

3.2.2 Wave descriptions of X-ray diffraction – Structure Factors

The sum of a diffracted ray for a certain reflection *hkl,* is defined by its structure-factor equation, F_{hkl}, which implies that each reflection, *hkl,* on the detector is the result of diffractive contributions from all atoms in the unit cell. Thus, F_{hkl}, is a complex periodic function, or a Fourier sum, where its wave is created by the superposition of many individual wave equations, each originating from diffraction by an single atom [139].

Since the diffractors of X-rays in a molecule are the electron-clouds of each individual atom, crystallography aims to obtain the function which would produce in an electron-density map of the molecule from the structure factors [139]. Another way of defining a structure-factor equation is a sum of the average electron density function of each volume element of the unit cell at position *(x,y,z)* These volume elements can be as small and numerous as desired and in this Fourier sum, each term contains the average numerical value of the desired electron density function $\rho(x,y,z)$:

$$F_{hkl} = \int_V \rho\,(x,y,z)\, e^{2\pi i(hx+ky+lz)}\, dV \qquad \text{Equation 3.2}$$

In equation 3.2, F_{hkl} is the Fourier transform of the electron density function $\rho(x,y,z)$. As Fourier transformation is reversible, electron density in turn can be expressed as a Fourier sum, where it is the transform of the structure factors as follows:

$$\rho(x,y,z) = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl} e^{-2\pi i(hx+ky+lz)} \qquad \text{Equation 3.3}$$

Where V is the volume of the unit cell, $\sum_h \sum_k \sum_l$, is the sum of the diffractions with index (*hkl*), and *i* is the type of atom. Therefore, this view of $\rho(x,y,z)$ implies that if three parameters can be measured for each reflection – amplitude, frequency and phase, they can be added together to obtain the function $\rho(x,y,z)$, from which an electron density map of the molecules in the unit cell can be obtained [139]. The measurable intensity of a reflection, I_{hkl}, gives the amplitude of its Fourier term, and the position *hkl* provides the frequency for that term, however, the phase α of each reflection cannot be measured on any kind of detector. Therefore, the phases need to be obtained separately, as will be elaborated in section **Error! Reference source not found.**, to complete the information needed to compute $\rho(x,y,z)$ [139].

3.2.3 X-ray data collection

Data collection aims to record the intensities of as many reflections as possible and rapidly, because protein crystals are prone to radiation damage by X-rays over prolonged exposure. Crystals are often analysed under a stream of nitrogen (boiling point -196^{o}C) to minimise this, and also because lowering the temperature could, theoretically, increase molecular order in the crystal, thereby improving diffraction. However, a drawback of freezing is the possibility of ice formation from solvent water which could damage the crystal. This is usually countered by using cryoprotectants, such as glycerol, whilst harvesting and freezing crystals for diffraction. If the crystal was not grown in cryoprotectant, harvesting usually involves soaking the crystal for 5-20 seconds in mother liquor to which cryoprotectant has been supplemented, picking the crystal with a loop of appropriate size and then flash freezing it [139].

Monochromatic X-ray radiation for crystal diffraction is generated by a synchrotron, which is a circular particle accelerator that works by accelerating electrons with a continuous array of magnets to produce radiation in the X-ray region. At the Diamond Synchotron, the photons produced from electron acceleration have an energy of 3 GeV. The X-rays are then channeled into individual beam lines and are utilised for data collection; a typical set up for a diffraction experiment is given in **Figure 3.4**.

Crystal quality is assessed by its diffracting power, which determines the number of available reflections that can be recorded. Crystals that produce reflections of 3 Å and below (often gauged from an initial diffraction screen producing 3-4 diffraction images) are considered suitable for structure determination [139]. Thereafter, a greater number of reflection images can be recorded (ranging between 1800-3600) from several parts of the good crystal, aided by the manually controlled goniometer (**Figure 3.4**) to enormously widen the scope of structural information that can be obtained.

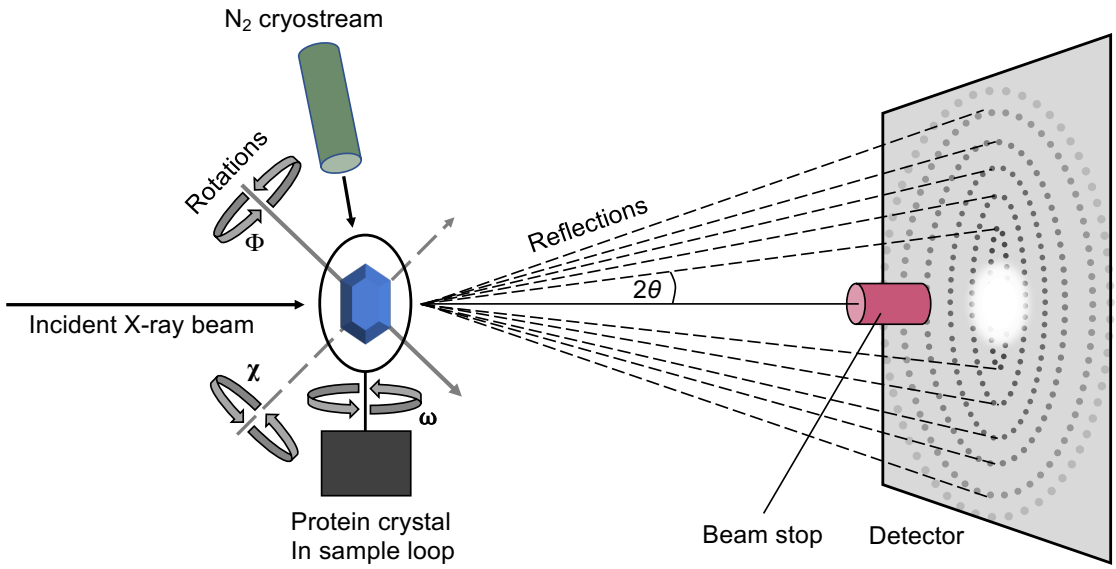


Figure 3.4: Set up of an X-ray diffraction experiment
Figure adapted from [146]. A protein crystal in a sample loop is mounted onto a goniometer for data collection, where it is held in a cryostream of nitrogen gas. The crystal is aligned and centred in the direction of the X-ray beam by controlling the angles of rotation on the goniometer. An incident monochromatic X-ray beam is scattered by the crystal and the positions and intensities of the resultant reflections are recorded by the detector. The positions of the Bragg spots give the deviation angle (2θ) of the reflection from the incident beam. Varying degrees of crystal rotation may be required to obtain a complete dataset, depending on the crystal's symmetry.

3.3 Diffraction data processing

3.3.1 Indexing, integration and scaling

After data collection, the next step is to process the data, wherein the *'hkl'* positions (indexing) and intensities (I_{hkl}) of all the reflections (integration) are computed, from which the crystal geometry can be determined. iMOSFLM is a program for the processing of diffraction images from a wide range of detectors [147] and it produces an MTZ file of reflection indices with intensities and standard deviations. However, indexing only provides information on the unit cell dimensions of the lattice, which may not reflect its true symmetry; for instance, a primitive hexagonal lattice may belong to point groups *C*222, 2, 1 3, 321, 312, 6 or 662. Although the true space group is a hypothesis until structure solution, it is useful to predict the point-group as accurately as possible at an earlier stage, to facilitate accuracy in subsequent stages of data reduction and interpretation, such as scaling [148]. Programs like POINTLESS are used after data indexing and integration to predict the space group of the crystal via estimating the most likely point-group symmetry or Laue group of the lattice [148].

Thereafter, scaling is performed to ensure internal consistency of all symmetry-reflected and duplicate reflections in the dataset, providing an estimate of the intensity of each averaged unique reflection together with an estimate of their error. This process accounts for the divergence in intensities of equivalent reflections, due to systematic errors associated with the incident beam, data collection process and the detector [148]. The output is a scaled and merged dataset, together with measures of the quality of scaled data, such as R factors and correlation coefficients, which are explained further under **Table 4.2**.

3.3.2 Phasing

As mentioned earlier in section 3.2.2, three key pieces of information are required to obtain the electron density map of a structure by X-ray crystallography: the frequencies, amplitudes and phases of reflections that contribute to the scattering

pattern. However, whilst recording diffraction intensities and positions, phase information is always lost and must be determined via phasing. Phasing can be of two main types, experimental phasing and molecular replacement (MR). Experimental phasing methods aim to resolve the phases of reflections using isomorphous replacement, single or multi-wavelength anomalous dispersion (SAD or MAD, respectively) or combinations of both.

Isomorphous replacement: this is the addition of one or more heavy atoms to a crystal lattice for phasing, and rests on the principle that a minor alteration in the diffraction pattern due to extra heavy atom(s) in the unit cell can provide initial estimates of phases. The most common method for preparing heavy-atom derivative crystals is to soak native protein crystals in solutions of heavy ions, such as Hg, Pt, Au, Br or I. The criteria for producing a good derivative crystal are that they must be isomorphous with native crystals, i.e. the extra heavy atom must not alter the unit cell from the native form, and that there must be measurable changes in a modest number of reflection intensities compared to the native, so that the phases can be resolved via a Patterson difference map revealing the heavy atoms' locations in the unit cell. The structure factor of the heavy atom crystal (F_{PH}) is given by the vector sum of the heavy atom itself (F_H) and the native crystal (F_P). Thus, the phase of the F_{PH} and F_P vectors can be derived geometrically from F_H via the relationship: $F_P = F_{PH} - F_H$ [139].

Anomalous dispersion: phases can also be obtained from heavy-atoms by utilising its ability to *absorb* X-rays of a certain wavelength at which Friedel's law does not apply, and the pair of reflections (*h, k, l*) and (*-h, -k, -l*) associated with the heavy-atom are not equal in intensity. This inequality in of symmetry-related reflections is called anomalous dispersion [139] and this non-equivalence of Friedel pairs is utilised to establish the phases of the reflections due to the heavy-atoms, and then use the derivative structure factors to solve the phase for native diffraction data.

All atoms can absorb and emit X-rays at wavelengths characteristic to the type of element they are, and absorption drops sharply just below each element's distinct emission wavelength K_β. The sudden change in absorption as a function of

wavelength (λ) is called an absorption edge, and an element displays anomalous scattering when the incident X-ray wavelength is near its absorption edge. Whilst absorption edges for the usual atoms in a protein molecule (such as carbon, oxygen and nitrogen) are not near the wavelength of X-rays routinely used in crystallography, the absorption edges of certain heavy atoms are within this range. Thus, the incident X-ray beam can be adjusted to maximise anomalous scattering by an auxiliary heavy atom by matching its wavelength to the heavy-atom's absorption edge [139].

Selenium is frequently used as an anomalous scatterer, as it has a K absorption edge maximum at 0.9793 Å and an absorption edge inflection point at 0.9794 Å [149]. Selenium is incorporated into a protein via selenomethione (SeMet) labelling during its *in vivo* expression (as detailed in 1.3.6 for the labelling of BR55), and then crystallising the protein after purification. This technique was used to solve the structure of BR55, and the phasing procedure is elaborated in section 4.2.4.

Molecular replacement utilises the phases from structure factors of a known protein as initial estimates of phases for a new protein [139]. This method requires a homologous protein structure with protein sequence identity of at least 35% [150]. Hence, it is a suitable phasing method for new proteins for which an orthologous PDB structure is available, or for new data on another form, conformation or portion of a protein that has an available structure. Structural homology databases such as DALI [15] and Phyre2 [14] are instrumental in identifying potential structural homologues that could be tested for MR. As will be presented in section 4.3.4 below, phases for the reflections from BR78 diffraction data were determined by MR, using the BR55 structure as the phasing model.

3.3.3 Model Building

After phasing, an electron density map can be derived from the complete structure factor of the reflections, given by:

$$\rho(x,y,z) = \frac{1}{V}\sum_h \sum_k \sum_l |F_{hkl}| e^{-2\pi i(hx+ky+lz-\alpha'_{hkl})} \qquad \text{Equation 3.4}$$

Where $|F_{hkl}|$denotes the amplitudes derived from the measured intensities in native diffraction data, and α'_{hkl} are the calculated phases [139]. An atomic model can be built into this electron density map, and nowadays this process is largely automated. For the structures in this thesis, two programmes called Buccaneer [151] and ARP/wARP [152] were used for model building. Both software use a chain-tracing approach to provide a partial trace of the protein backbone within the electron density map. As such, the quality and completeness of the trace depends on the quality of the initial phasing and phase improvements, and the resolution of the map [151].

Buccaneer uses a density likelihood target function to identify and score possible C^{α} positions and orientations in the map in a six-dimensional search, against a simulated density map for a solved structure [151]. At resolutions as low as 3.5 Å, the model building process involves finding initial C^{α} positions; building a C^{α} backbone form the initial C^{α} 'seeds', whilst adhering to Ramachandran bond angle constraints; joining the chain fragments into larger chains; and then curtailing any clashing C^{α} atoms within 2.0 Å of each other. Buccaneer does not refine or rebuild models, and performs limited sequence assignment and sidechain additions to the C^{α} chain trace [151].

ARP/wARP typically requires a data resolution of 2.8 Å or better. Driven by a sophisticated decision-making control module, ARP/wARP combines model building and refinement into a unified process that iteratively builds protein chains, secondary structures and solvent, providing the final macromolecular model. The number of iterative building/refinement cycles can be predetermined by the user. An advantage of this iterative process is that as the quality of the partially built model is gradually increased, the phases also improve and result in an enhanced density map where an increasingly accurate and complete updated model may be built with each cycle [153].

Automated model building, however, has its limitations and a manual examination of the model is necessary to look for any incorrectly traced features or amino-acid sidechains. Coot [154] is a graphic program for manual model building, allowing the moving of atoms, residues and fragments, amino acid modification, and electron density fitting to improve the model.

3.3.4 Structure refinement

The iterative process of refinement is the final step towards solving a crystal structure, and it is performed to maximise the agreement between the derivative model and the experimental diffraction data [155]. REFMAC5 was used for the refinement of BR55 and BR78 models after every building cycle, where model quality was judged by R_{factor} and R_{free} values. Model parameters that can be optimised include atomic coordinates, atomic displacement, scale factors and twin fractions (in the case of twinning) [155].

The R_{factor} measures the agreement between structure factors amplitudes calculated from a crystallographic model and those from the original diffraction data [156]:

$$R_{factor} = \frac{\Sigma\,||F_{obs}|-|F_{cal}||}{\Sigma\,|F_{obs}|} \qquad \text{Equation 3.5}$$

Where $|F_{obs}|$ is the structure factor of the experimental data and $|F_{cal}|$ is the structure factor of the model. The lower the R_{factor} is below 0.5, the better the agreement [139].

As the refinement process improves the atomic model of a structure to ensure it fits better to the experimental data, this introduces biases into the process. To counter this, the R_{free} is employed as a quality control measure. About 5-10% of the experimental reflections can be randomly selected and removed from the data set prior to refinement, and the remaining 90-95% of the observations are refined with the model. The R_{free} is calculated from this small subset of unrefined reflections in the same way as the R_{factor} to measure the degree of agreement between the built model and the unrefined experimental structure factors. For a model that is a good fit to the diffraction data, the R_{free} will be very close to, albeit higher than, the R_{factor} [139,156].

Other measures of a model's quality are the root mean square deviations (RMSD) of its bond lengths and angles from an accepted set of values derived from the geometric constraints of small organic molecules. Generally, for a refined model, RMSDs for bond lengths and angles must be no more than 0.02 Å and 4°, respectively [139]. All these parameters are reported for the structures of BR55 and BR78 in **Table 4.2** and **Table 4.3**, respectively.

CHAPTER 4

STRUCTURAL CHARACTERISATION OF BR55 AND BR78

4 STRUCTURAL CHARACTERISATION OF BR55 AND BR78

Identification of the BR55 and BR78 gene sequences, and PCR and cloning of the constructs were carried out by Donald MacKenzie at the IFR (Norwich, UK). BR55 and BR78 constructs was cloned into the pOPIN-F plasmid [5], which allows the expression of proteins with a 3C-protease cleavable N-terminal 6His tag. Initial small scale expression and purification trials (using IMAC) were performed at the IFR, with *E. coli* BL21(DE3) as expression host, in up to 200 mL of culture, where it was noted that BR55 gave yields in the order of 10 mg/L, with purity >90%, whereas BR78 afforded yields lower than 1 mg/L, with apparent co-eluting proteins of other sizes. Thereafter, the plasmids and expression strains for BR55 and BR78 were given to the Dong group to optimise the expression and purification of BR55 and BR78, in order to commence X-Ray crystallographic studies of BR55 and BR78.

4.1 Expression and Purification of BR55 and BR78 Proteins

4.1.1 BR55 Purification

BL21(DE3)-BR55 cultures for recombinant production of His-tagged BR55 were typically 4-10 L in quantity, and grown as 1L batches, as described in section 1.3.1.1. BR55 was subjected to a two-step purification process. In the first step His-tagged BR55 was largely captured by Ni-NTA IMAC, followed by the removal of the His tag by 3C-protease, and isolation of tag-free BR55 in a much more purified form due to two pulldowns on an IMAC column, as described in 1.3.3.1 and 1.3.3.2. The second step involved refining the purity and homogeneity of BR55, by getting rid of tiny amounts of contaminating protein and higher MW oligomers by SEC, as described in section 1.3.3.3.

Representative results of column flow through from each step of IMAC purification is displayed in **Figure 4.1C,** and this includes SDS-PAGE analysis of the clarified lysate before and after Ni-NTA IMAC pulldown, column washes, elution of His-tagged-BR55, the pulldown of BR55 after His-tag separation and column washes after BR55 isolation. The results from a typical SEC purification of BR55 is shown in **Figure 4.1B** and **D** below**.** The yield of crude protein was not noted after the IMAC stage, and any yields reported in this thesis is of final, purified protein after SEC.

For the first purification trial, the yield of monomeric BR55 after His tag removal and SEC was 0.35 mg/L of culture. A large amount of protein appeared to oligomerise during SEC, as can be seen in **Figure 4.1A** below. It was suspected that this could be caused by a cysteine residue at position 207 of the BR55 protein. Therefore, for the second purification, 2.5mM DTT was added to the SEC buffer. This increased the yield of monomeric protein to 0.69 mg per litre culture (see **Figure 4.1B**), and so DTT was included in SEC procedures for all subsequent purifications of BR55, BR55 mutants (unless specified) and BR78. For subsequent recombinant BR55 over-expression, supplementing LB media with 1% glucose (as per recommendations by Donald MacKenzie) increased the yield of BR55 up to 10 mg/L of culture.

SeMet-BR55 was purified similarly to native BR55, and the yield was 0.35 mg per litre of culture. A comparatively low yield for SeMet-BR55 is not strange, because SeMet-BR55 protein was expressed in minimal, inorganic media, in which bacterial multiplication rates and overall recombinant protein yields are typically lower. Results for SeMet-BR55 purification are presented in **Figure 4.2** below, where **Figure 4.2A** displays the SEC chromatogram of SeMet-BR55, and **Figure 4.2B** is an SDS-PAGE gel showing the protein content of samples during progressive stages of IMAC and SEC purification.

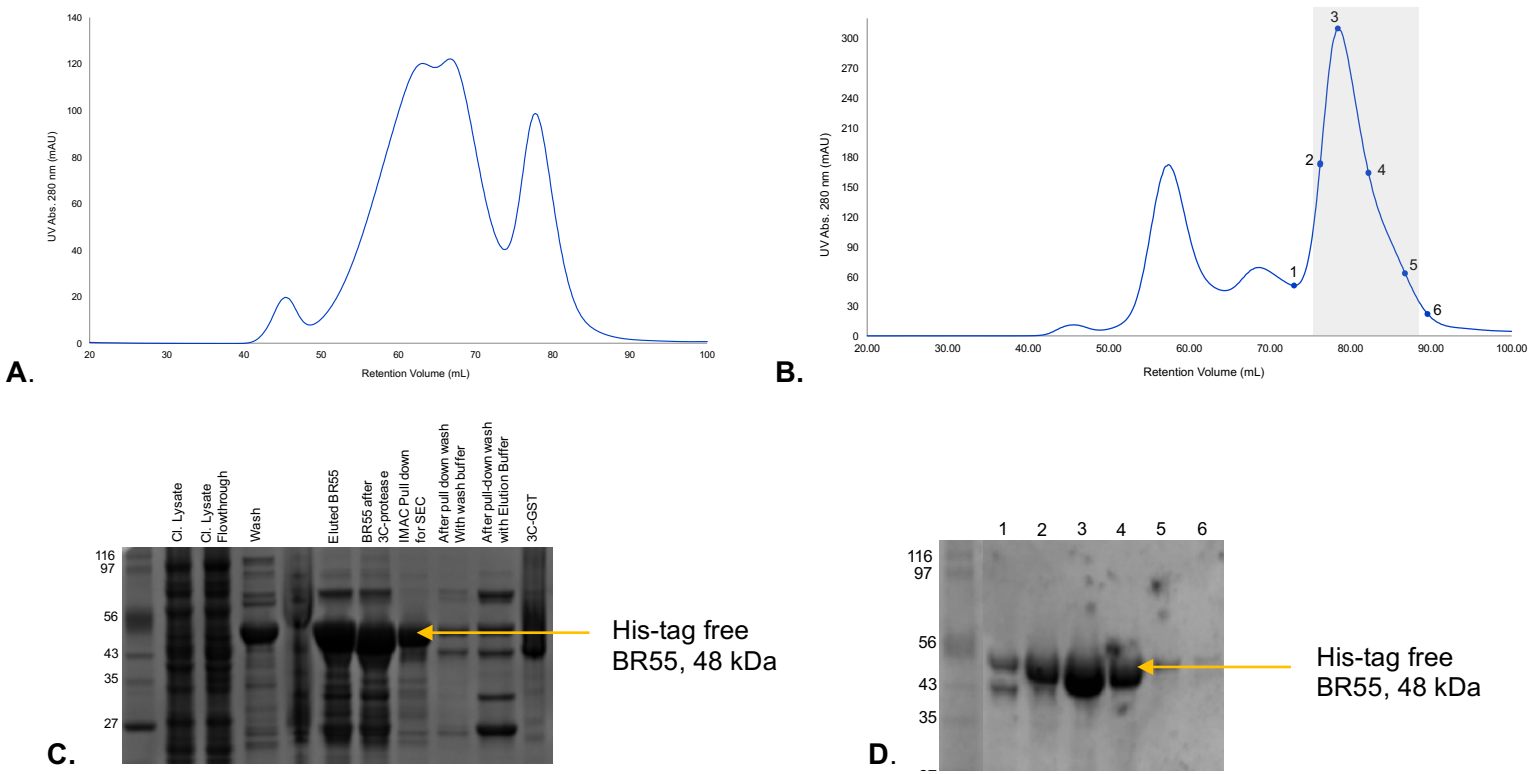


Figure 4.1: SEC chromatograms and SDS-PAGE gels for BR55 purification trials.
SEC Chromatograms of BR55 purification trials, **A.** without DTT in SEC buffer (yield: 0.35 mg/L of culture) and **B.** with 2.5 mM DTT in SEC buffer (yield: 0.69 mg/L of culture); the numbers correspond to the center of the fractions analysed by SDS-PAGE in **D**, and the grey area indicates fractions that were collected and concentrated for crystallography. **C.** SDS-PAGE gel of protein samples taken at selected stages of IMAC purification of BR55, and after 6His-tag removal by 3C-GST, followed by IMAC pull down (Cl. Lysate – Clarified Lysate). **D.** SDS-PAGE results of selected fractions of BR55 after SEC purification; the numbers labelling each well correspond to the fractions indicated in **B.**

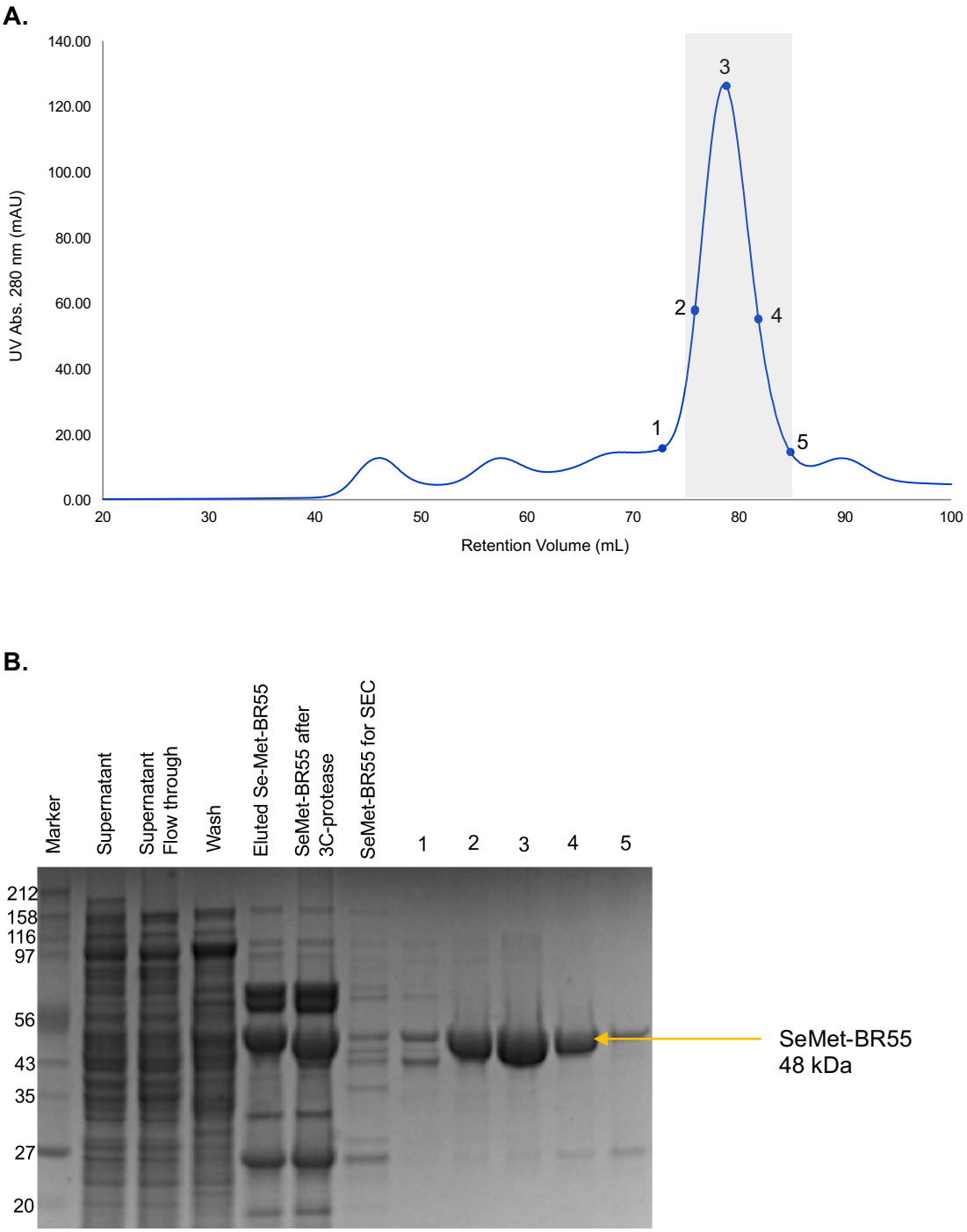


Figure 4.2: SEC Chromatogram and SDS-PAGE for SeMet-BR55 purification

A. SEC Chromatogram of SeMet-BR55 purification after 6His-tag removal and IMAC pulldown. The numbered points on the chromatogram correspond to the center of the fractions analysed on the SDS-PAGE gel in **B**. The shaded areas indicate the fractions that were collected for crystallography. **B.** SDS-PAGE results of selected protein samples during IMAC purification, followed by SEC.

4.1.2 BR78 Purification

Expression and purification procedures that were refined for BR55 production (i.e. by adding 1 % glucose to LB media) was applied to BR78. However, the yields for BR78 were significantly lower at about 0.8 mg/L of culture. Furthermore, while the SEC chromatogram for BR78 would show a single, sharp peak at the expected retention volume, SDS-PAGE analysis of the fractions revealed three bands at 70 kDa, 60 kDa and 55 kDa, respectively (see **Figure 4.4B**). MALDI-MS analysis of these bands (performed by Gerhard Saalbach, John Innes Centre, UK) revealed that both bands consist of peptide sequences from BR78 (**Table 4.1** and **Figure 4.3**). Nevertheless, despite the heterogeneity of the sample, the protein was concentrated to 9 mg/mL and subjected to a crystallisation trials, as discussed in sections 4.2.2 and 4.3.2.

Table 4.1: MALDI-MS Results of the 3 bands from SEC Purification of BR78

Sample No.	Peptides	BR Coverage	Full protein Coverage	Mascot Score*
1	8	22%**	5%	38
2	6	16%***	4%	46
3	5	13%****	3%	56

* Mascot scores >35 are significant matches ($p < 0.05$)
** Coverage highlighted in yellow in
*** Coverage shown in **bold** in
**** Coverage underlined in

```
                                STL SGLDNNAVAN SVTQNLETTT NTADLTNAAL ASAYQTNLIQ
241         LQNRTANAEN DTVQAEDIQA DATAANASEL KKALQDTSVH TIKLTDNITL TSAIELTNVS
301         RDVTIYGNGK YINATDGNGG IFIHNTKSYT VNLTIEKATL YNQSQYGFVH MNDEGTDNIT
361         YKNITAYGGT LVWSQTHVGT KTLSLEGTVN FYSVPSYTVG GQTYSTDAFK IGTHYPNGEN
421         KDTTPAIYVS NEINIADNAN IALENSATKI DIWMIADIGI HPHTTALTIG NNATLTMENG
481         NNSALNIKLD GDTSNSFTVG EGSTVKLSAK VDNVRILPYE DSNTANVSFA KGSDVTLHAG
541         TGSNLRMGAS ISNQIDFNGK ATFIKDSGAY ANTAYADQTR GNIEFDYYWN DQQKTGSTGV
601         ANFNPGSNVL FQAGPGASNV NTYKLDKNHS YVRTTVNINN PEKVTFDTDI IQNNKGVAYL
661         GNADNGVVVN VNNASVKVNN NNFSDVISSN TTTVTGSTVN VGSVSMASTN GSFTNASTSL
721         SGSVSFSYAT SMATSQLNAV DTNSTSTVVY QGSTVIS
```

Figure 4.3: Coverage of MALDI-MS peptides in BR78
The peptide sequence coverage details and font key are presented in , with the corresponding sample number.

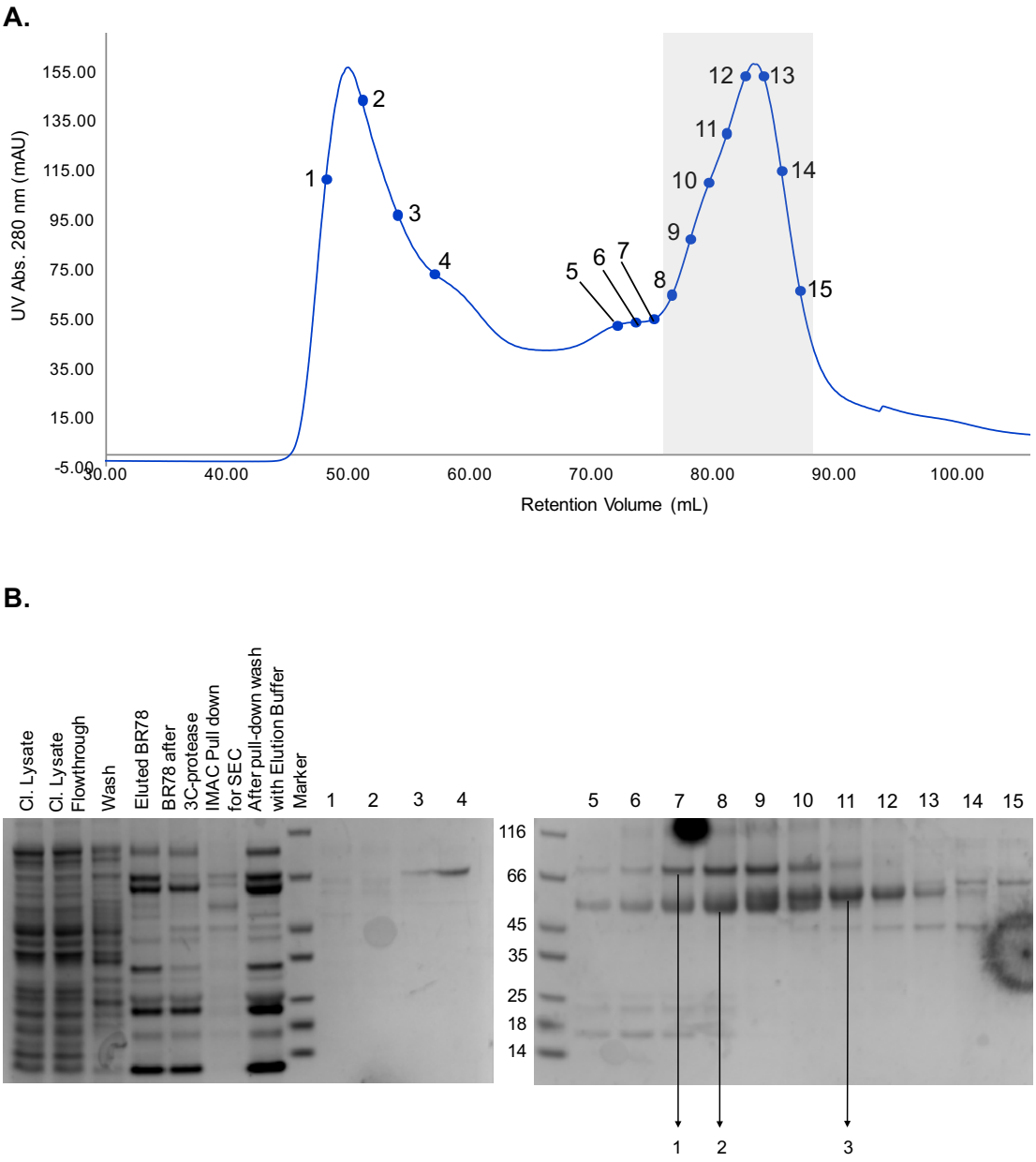


Figure 4.4: SEC Chromatogram and SDS-PAGE results for BR78 purification

A. SEC Chromatogram of BR78 purification after 6His-tag removal and IMAC pulldown. The numbered points on the chromatogram corresponds to the fractions analysed on the SDS-PAGE gels. The shaded areas indicate the fractions that were collected for crystallographic studies. **B.** SDS-PAGE results of selected protein samples during BR78 IMAC purification, followed by SEC. Arrows 1,2 and 3 indicate the gel bands that were sliced and prepared for MALDI-MS analysis, as samples 1,2 and 3, respectively; MALDI-MS results are in and , above.

4.2 X-Ray Crystallography: Method

4.2.1 *In-situ* limited proteolysis for crystallisation of BR55 and BR78

Protease screening for limited proteolysis of BR55 and BR78: To identify a suitable protease and protease dose for limited *in-situ* proteolysis of BR55 during crystallisation, an initial screen with 4 proteases at a range concentrations was performed, and the results were analysed by SDS-PAGE. The rationale behind adopting this approach is explained in section 4.3.1 below. Trypsin, α-chymotrypsin, thermolysin and V8-protease were each added in protease to protein ratios of 1:10, 1:100 or 1:1000 to 35 µg of BR55 (7 µg/µL), and in ratios of 1:100, 1:500 and 1:1000 to 28 µg samples of BR78 (9 µg/µL). They were then incubated at room temperature for 1.5 hours, after which the proteolysis reaction was quenched as described in [157] by adding HEPES buffer (10 mM HEPES, pH 7.5, 0.5 M NaCl) and 5x SDS-PAGE loading buffer. SDS-PAGE analysis was then performed as described in section 1.3.5. All proteases and the components of the buffers in which they were prepared were procured from Sigma-Aldrich. Protease stock solutions were prepared as described by Wernimont et. al. [157], and were diluted as required by HEPES buffer before use. Trypsin (2 mg/mL, in 1 mM HCl and 2 mM $CaCl_2$), α-chymotrypsin (2 mg/mL, in 1 mM HCl and 2 mM $CaCl_2$), thermolysin (1 mg/mL in water), and Endoproteinase-V8 protease (2 mg/mL in water).

4.2.2 BR55 and BR78 Protein Crystallisation

Proteins were crystallised by the sitting drop vapour diffusion method in 96 well crystallisation trays (Molecular Dimensions, Newmarket, UK). Protein solutions in their respective SEC buffers were either used fresh from purification or from storage at -80°C. Crystallisation trials were set up using the Gryphon crystallisation robot (Art Robbins Instruments, California, USA) with a 1:1 ratio of protein to precipitant solution, where the total drop size ranged from 1 µL to 2 µL.

For limited proteolysis crystallisation trials, BR55 (at 7 mg/mL) and Se-Met BR55 (at 15 mg/mL), were mixed with α-chymotrypsin at a 1:500 protease to protein ratio, and BR78 (9 mg/mL) was mixed with 1:700 thermolysin. These were incubated at room temperature for 1.5 hours prior to crystallisation screening with the PEG/Ion 1 and 2 screens (Hampton Research, California, USA). Crystals of BR55 (<3 weeks) were obtained in 0.1 M sodium malonate/pH 6.0, 12% PEG 3350. Se-Met BR55 crystals were grown in 0.03 M Citric acid, 0.07 M Bis-Tris propane/pH 7.6, 20% PEG 3350 (within 3 days). BR78 crystals grew in 0.2 M Sodium tartrate dibasic dehydrate/pH 7.3, 20% PEG 3350 (within 3 days).

Crystallisation trays were stored at 22^{o}C in a temperature controlled incubator and crystal growth was monitored with a light microscope. Crystals of adequate quality were harvested with a 0.2 to 0.3 mm loop, and then soaked in cryo-protectant (20% glycerol supplemented to the respective crystallisation mother liquors) and frozen in liquid nitrogen for X-ray diffraction at the Diamond Light Source (Harwell, UK).

4.2.3 Data collection and processing

Data sets were collected at beamlines I03 and I04 of the Diamond Light Source (Harwell, UK). 1800 or 3600 diffraction images for each protein crystal was collected at a wavelength of 0.9795 Å and under a cryostream of liquid nitrogen at -196^{o}C, with exposure set at 0.100 sec and 100% transmission. First an edge scan was performed on a crystal of SeMet-BR55, resulting in anomalous dispersion, giving an absorption peak at 0.9793 Å and inflection point at 0.9794 Å, characteristic of a Se atom (see section 3.3.2). The numbers and positions of Se atoms in the unit cell were estimated by the automated FastEP pipeline on the IsPyB interface, at Diamond. Thereafter, data was collected under cryogenic temperatures at a wavelength of 0.9792 Å, with exposure at 0.100 sec and 30% transmission. The diffraction images were indexed and integrated automatically by Xia2 [158], or manually using MOSFLM [147], with merging and scaling performed by SCALA [159] in the CCP4 suite [160]. Space groups were determined on POINTLESS [148]. A detailed account of BR55 and BR78

data collection and processing is provided in the results sections 4.3.3 and 4.3.4, respectively, together the relevant statistics.

4.2.4 Structure determination and refinement

Phasing of Se-Met BR55 diffraction data was first performed by the SHELXC/D/E programs. SHELXC was used to prepare the .sca file – obtained from scaling and merging of the Se-BR55 diffraction data – for the SHELXD (heavy atom location) and SHELXE (phasing and density modification (DM)) programs, for phasing by SAD (SAD was defined as the type of experiment in SHELXC). No native data sets were provided. The output fa.hkl and fa.ins files were used as input files for SHELXD, which was set to find 8 Se heavy atoms in the substructure, with resolution cutoff for the anomalous signal at 2.0 Å. The number of tries to find the correct solution was limited to 1000. The 'Site occupancy vs. Peak Number' output graph (**Figure 4.11A**) revealed 11 Se atomic positions in the substructure, which was also depicted in an output PDB file (**Figure 4.12A).** DM of the phased substructure was then performed by SHELXE, where the solvent content was defined as 0.5. The output 'Contrast vs. Cycles' graph for 200 DM cycles (**Figure 4.11B**) indicated that the original hand is the correct enantiomeric form for Se-BR55, and not the inverted hand.

Using the information obtained from SHELXC/D/E phasing, CRANK was used in its custom setting (AFRO for F_A calculation, CRUNCH2 for substructure determination, BP3 for phasing, SOLOMON for enantiomer detection and DM, and Buccaneer for model building [151]) to re-phase the merged diffraction data and build a structure of SeMet-BR55. CRANK is a program that automates the entire structure determination process, beginning from phasing to model building, drawing from various crystallographic programs for each step. CRANK was asked to search for 11 Se 'substructure atoms' during phasing and f'/ f'' values obtained from the Se edge scan were also input. 12 Se atomic positions were identified in the correct hand, and subsequently a partial model of Se-Met BR55 was obtained, which was used as the search model for molecular replacement of native BR55 using Phaser [160,161]. A

complete model of native BR55 was auto-built by ARP/wARP [152], and thereafter manually verified with Coot [154]. Each building cycle was subjected to structural refinements (rigid body and restrained) with REFMAC5 [162].

The structure of BR78 was determined by molecular replacement using Phaser, with the native BR55 structure as the search model. The complete model of BR78 was auto-built by ARP/wARP [152], and manually built with Coot [154], and each building cycle was subjected to structural refinements with REFMAC5 [162]. Both native BR55 and BR78 structures were validated with MolProbity [163]. A detailed account of BR55 and BR78 structural refinement is provided in the results sections 4.3.3 and 4.3.4, respectively, together with the refinement statistics.

4.3 X-Ray Crystallography: Results

4.3.1 Crystallisation of BR55

Initially, full length BR55 (7 mg/mL) was subjected to crystallisation trials using the PEG/Ion 1/2, JCSG+, Crystal Screen Cryo 1/2 and Structure Screen sparse matrix screens. BR55 crystals would take an average of 1-2 months to form in selected conditions, and were often of inadequate quality for X-ray diffraction analysis, even after repeated optimisation of the successful conditions. Full length BR55 crystals from these initial trials were typically fine, short needles, seen in **Figure 4.6A** below.

To improve the quality and size of BR55 crystals, an *in-situ* limited proteolysis approach to crystallisation was adopted. Dong et. al. [164] and Wernimont et. al. [157] have reported *in situ* proteolysis for crystallisation as an effective rescue strategy to resuscitate proteins that have previously resisted structure determination despite extensive trials. The respective success rates of this technique in both investigations were 14.5% and 12.6%, with further improvement ongoing. The proteases seem to work by removing disordered appendages of the intact protein, so that the intrinsic homogeneity of the trimmed protein fragment is increased, which can aid in the growth of good quality crystals [157]. The proteases used in BR55 were trypsin, α-chymotrypsin, thermolysin and Endoproteinase Glu-C V8 (V8-protease). The type of protease(s) and the best possible concentration(s) to be used for BR55 crystallography were determined from the results of a limited proteolysis trial, in which BR55 was treated with three concentrations of each protease at 1:10, 1:100 and 1:1000 protease to protein ratio. In theory, the condition(s) that gave the single largest, most stable domain are considered ideal, and those that show more than one proteolytic fragment and/or excessive degradation are considered least ideal [157], [164]. The method for this experiment is described in section 4.2.1 and the results are displayed as an SDS-PAGE gel image in **Figure 4.5** below.

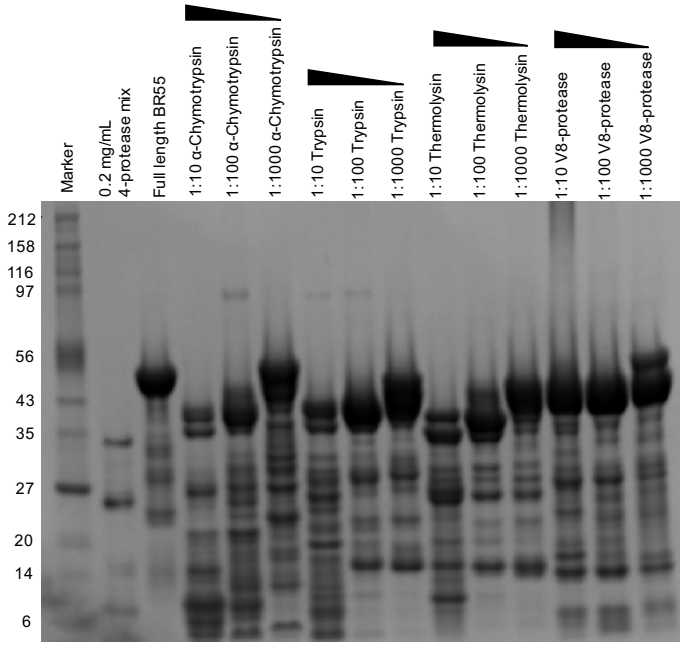


Figure 4.5: SDS-PAGE results for the initial screening of type and amount of protease appropriate for *in situ* limited proteolysis crystallisation of BR55.

A qualitative assessment of the degradation patterns observed in **Figure 4.5** helped to propose suitable protease to protein ratios for the initial BR55 crystallisation trials: α-chymotrypsin was chosen at 1:500, trypsin at 1:700, thermolysin at 1:1000, and V8 protease at 1:400. The limited proteolysis strategy was successful in allowing the growth of diffraction quality BR55 crystals, as displayed in **Figure 4.6B**. BR55 treated with 1:500 α-chymotrypsin afforded crystals in under 3 weeks from 0.1 M Sodium malonate, pH 7.0, 12% w/v PEG 3350, which were sent to the Diamond Light Source (Didcot, UK) for data collection. The method was repeated with SeMet-BR55 at 15 mg/mL and crystals were obtained within 3 days in 0.03M Citric acid, 0.07 M Bis-Tris propane/pH 7.6 and 20% PEG 3350 (see **Figure 4.6C**). Diffracted BR55 crystals returned from Diamond were analysed by SDS-PAGE (**Figure 4.7A**) and gel slices of the fragments were provided for orbitrap MS analysis (done by Fran Mulholland, IFR, Norwich) for amino acid identification. These data provided molecular weight (MW) and residue coverage information which was used in data processing (see section 4.3.3). The MW of the crystallised fragment was estimated as 36.3 kDa by SDS-PAGE analysis and calculated as 33.5 kDa from orbitrap MS identification of BR55 residues in the fragment, which was 70% of the full-length protein (**Figure 4.7B**).

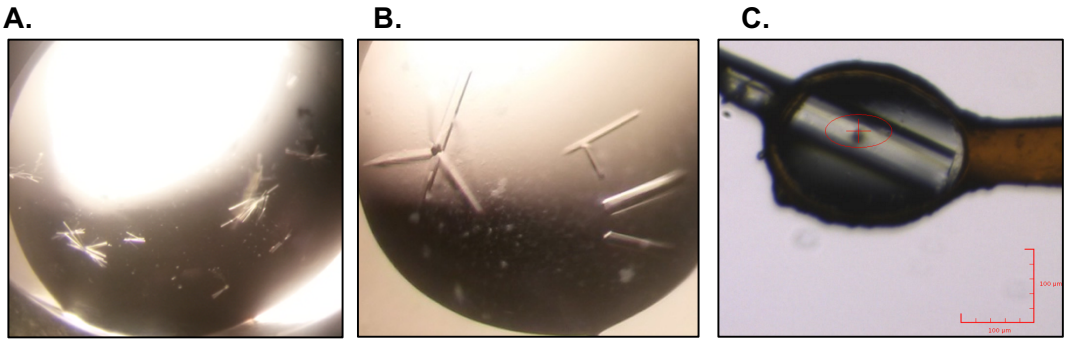
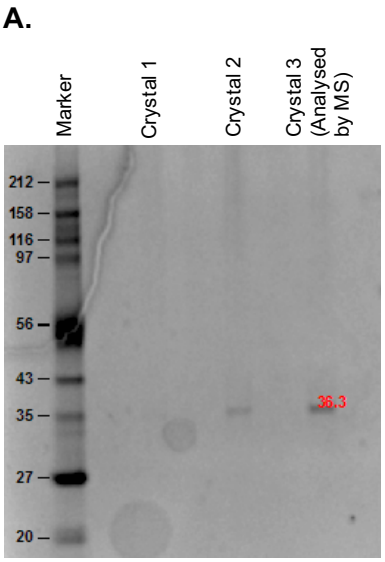


Figure 4.6: Protein crystal images for BR55 crystallisation
A. Fine, needles of BR55 (7 mg/mL) without protease treatment, grown over 6 weeks in optimised conditions of 0.2M ammonium tartrate dibasic, 20% w/v PEG 3350, pH 6.6. **B.** Diffraction quality crystals of BR55 (7 mg/mL) grown in <3 weeks from 0.1M sodium malonate, pH 7.0, 12% w/v PEG 3350. **C.** Diffraction quality crystal of SeMet-BR55 (15 mg/mL) in a sample loop, which was grown in <3 days in 0.03M Citric acid, 0.07 M Bis-Tris propane/pH 7.6 and 20% PEG 3350.



B.

MATILPATGNNDGETSVANRTVTEGNIVQ**LAENIEEVSNEEE**
LKAALRDASITTIKLKNNITLNNAITINNGNRNITIIGDGHY
IN**ALNSDGGIILNNR**GGSAK**IDLTIENATLYNTSKYGFVNMS**
SNGVDTVTYKDVTAYGGTLVWSKTGAGVK**TLNLVGNTTLNSV**
KSYEVDGQSCG**TEAFSHR**TPDGDKTTALYVSNAINIAENANV
VLNNSATDIDMWLLTAVPSTSGISTVTVGN**NASLTMENIGNT**
EYNIKLDGGRENHFIVNENAAVKMSAKVDNVRIIPQLENIFT
RGNIELAKGSNVHLEVITGSNFRVAGTVANRIDFNGTATLIK
QEGASGPTIYDKNSQANIEFNWLGIFENTVNFNPGADVTLIA
GKGASNIGSSSLVTATVNINDPLRVAFNTDKLETDNGVTYLG
NSRINVKVTNATIKVDDNDF

Primary sequence of BR55. The red fragments are those amino acids observed from Orbitrap MS results of proteolysed BR55 protein extracted from gel slice in A.

Full length BR55:	46.8 kDa
BR55 from MASCOT coverage:	33.5 kDa

Figure 4.7: Analysis of BR55 crystals from which diffraction data was collected
A. SDS-PAGE results of BR55 crystals from limited proteolysis, with an approximate molecular weight of 36.5 kDa. **B.** MASCOT coverage of BR55 protein extracted from SDS-PAGE gel slice and treated with trypsin before orbitrap MS analysis.

4.3.2 BR78 Crystallisation by *in situ* limited proteolysis treatment

The SDS-PAGE gels in show that BR78 was purified as more than one fragment. MS results in showed that all these fragments were composed of BR78 peptides with varying sequence coverage. Nevertheless, BR78 was subjected to a limited

proteolysis screen in a similar strategy to BR55. The results of this screen are shown in **Figure 4.8** below and for the initial *in situ* crystallisation trial three proteases were chosen: thermolysin, trypsin and α-chymotrypsin, all at 1:700 protease to protein ratio. A large single crystal of BR78 (9 mg/mL) treated with 1:700 thermolysin was obtained in 0.2 M Sodium tartrate dibasic dehydrate, pH 7.3, 20% PEG 3350; an image of this crystal is presented in **Figure 4.9** below.

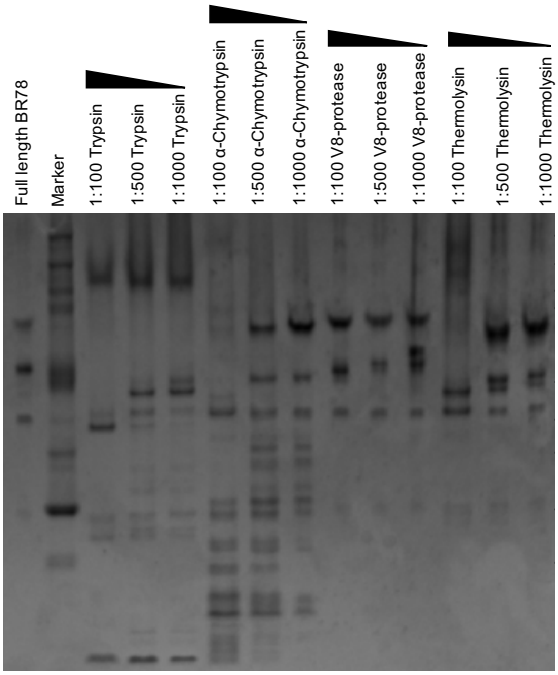


Figure 4.8: SDS-PAGE results for the initial screening of type and amount of protease for *in situ* limited proteolysis crystallisation of BR78.

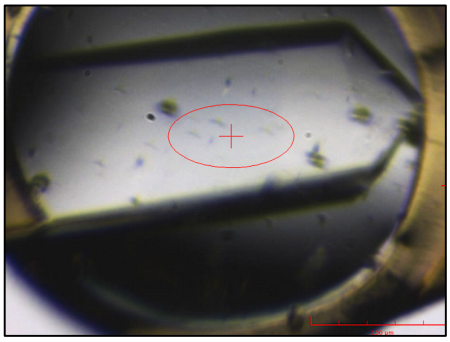


Figure 4.9: Single crystal of BR78 in a sample loop during data collection
The crystal was grown in 0.2 M Sodium tartrate dibasic dehydrate/pH 7.3 and 20% PEG 3350

4.3.3 BR55 Data collection and Structure Determination

3600 diffraction images of native BR55 crystals were obtained at temperatures around -196^{o}C, and an X-ray wavelength of 0.9763 Å and resolution 2.00 Å, which were integrated and scaled by XIA2, to a maximum resolution of 2.13 Å. The diffraction images were integrated in MOSFLM [147] and scaled using SCALA [159]. A space group of P $3_1$21 was determined by POINTLESS with average unit cell dimensions of a = b = 146.70, c = 110.42. A merged MTZ file was generated in which the resolution of the data was improved to 1.92 Å. The data processing statistics for BR55 are given in **Table 4.2**.

The Phyre^2 [14] protein structural homology search tool was used to identify a suitable search model for BR55 data phasing by molecular replacement. The highest scoring template (PDB entry 3ZSC) could model only 10% of the BR55 sequence with 87.7% confidence, giving an overall structural homology of 22%, which was inadequate for molecular replacement. Subsequent attempts at molecular replacement with models of the Fap1 (PDB ID 2X12) [131] and GpsB (PDB ID 3QC5) [116] SRRP binding domains also failed to provide a solution for a BR55 structural model. Therefore, SeMet-BR55 was produced and crystallised, as described in 1.3.1.1 and 4.2.2, respectively, to determine the phase for BR55 diffraction data by heavy atom replacement.

An edge scan was performed on a crystal of SeMet-BR55 (see **Figure 4.10A**), to identify the unique absorption edge of selenium. An absorption maximum at 0.9793 Å, and an absorption inflection point at 0.9794 Å was noted in the resulting X-ray fluorescence spectrum. Scattering coefficients f": 5.52 and f': -8.53 were computed from the edge scan on the ISPyB interface, indicating the presence of Se electron density in the crystal lattice. 3600 X-ray diffraction images were then collected at the SeMet wavelength of 0.9792 Å, which were integrated and scaled by XIA2 to a maximum resolution of 2.73 Å. While XIA2 suggested a space group of P$3_1$21 for the data, with a unit cell dimension of a = b = 148.40, c = 110.73, the in house FastEP auto processing software identified 5 Se atom positions with a space group of P$3_2$21

(see **Figure 4.10B**). This space group was considered more accurate as it was derived from experimental phasing of heavy atom diffraction data. Therefore, the resulting MTZ file from SeMet-BR55 diffraction, which was downloaded after XIA2 auto-processing on the Diamond website, had its space group changed from $P3_121$ to $P3_221$ using SORTMTZ, which is part of the CCP4 Suite [160]. The complete processing statistics for SeMet-BR55 are provided in .

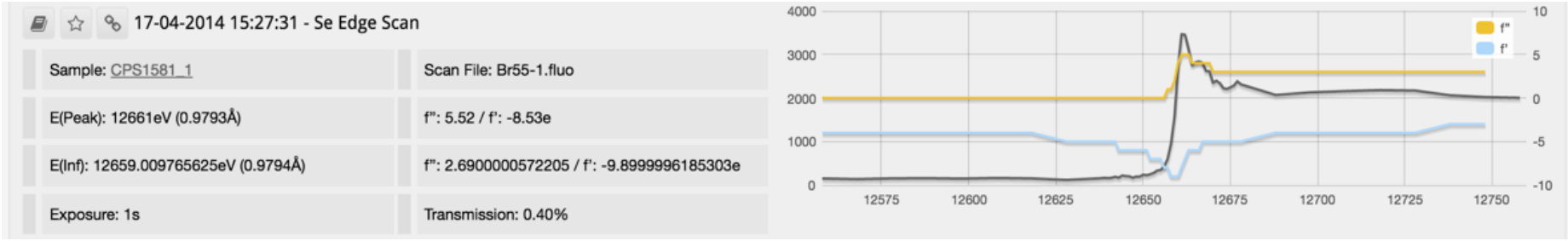
Se coordinates in the diffraction data and the correct enantiomer were more accurately determined with SHELXC/D/E, as described in section 4.2.4 above. SHELXD identified 11 Se atom coordinates in the substructure, as shown by the 'Site occupancy vs. Peak Number' output graph in **Figure 4.11A**. A sharp drop in occupancy after the expected number of number of sites that was initially input (8) indicated that the solution must be correct. A PDB file showing the positions of the 11 atoms in a partially built model (see **Figure 4.12**) was also output. Thereafter, DM by SHELXE helped to indicate the correct enantiomeric form of Se-BR55; a 'Contrast vs. Cycle' graph depicting the SHLEXE output is given in **Figure 4.11B**, where a higher pseudo-free CC of 0.73 was found for the original form (vs. 0.64 for the inverted form), indicating that the original form is the correct enantiomer.

Thereafter, CRANK was used, as described in section 4.2.4, to re-phase the diffraction data and build a partial structure of SeMet-BR55; a solution was obtained with R_{factor} 0.3841 and R_{free} 0.4769 and 12 Se atoms. The model was further built using Buccaneer, with 40 cycles of building and refinement, affording a model with R_{factor} 0.2334 and R_{free} 0.2887. An image of this model is presented in **Figure 4.12B** below, and was used as the search model for molecular replacement of the high resolution native BR55 data, with MOLREP. Before molecular replacement, the space group of native BR55 was changed to $P3_221$ using SORTMTZ, with cell dimensions of a = b = 140.7, c = 110.42. The Matthews coefficient revealed the presence of 3 monomers in the asymmetric unit, for a solvent content of 61%, using a protein molecular weight of 36.3 kDa (obtained from TOF-MS results of BR55 diffraction crystal, shown in **Figure 4.7**). A BR55 model was built with the resulting output coordinates using ARP/wARP

[152], with 10 building cycles, and 5 refinement cycles per building cycle; R_{factor} 0.1823 and R_{free} 0.220 for 6862 built atoms was attained. Subsequent cycles of manual building in COOT, and rigid body and restrained refinement of the model in REFMAC yielded a model with a final R_{factor} of 0.1614, R_{free} of 0.1935, $RMSD_{bond\ angle}$ 1.983 and $RMSD_{bond\ length}$ 0.0203. BR55 Refinement statistics are given in **Table 4.2**.

Validation of the structural model in MolProbity [163], showed no Ramachandran outliers, and 95.9% of Ramachandran positions were favoured. A clashscore of 7.52% was obtained, which is in the 89% percentile for 828 PDB entries of comparable resolution (100^{th} percentile being the best and 0^{th} percentile being the worst). Full validation statistics are provided in APPENDIX 5.

A.



B.

Figure of Merit: 0.62
Pseudo-free CC: 64.72

Map / Model Viewer | PDB/MTZ file | Log file

#	X	Y	Z	Occ
1	-21.623	43.655	1.488	1.000
2	-54.318	27.061	-2.734	0.762
3	-78.789	26.807	22.269	0.732
4	-81.733	13.135	37.016	0.728
5	-48.951	29.596	0.080	0.710

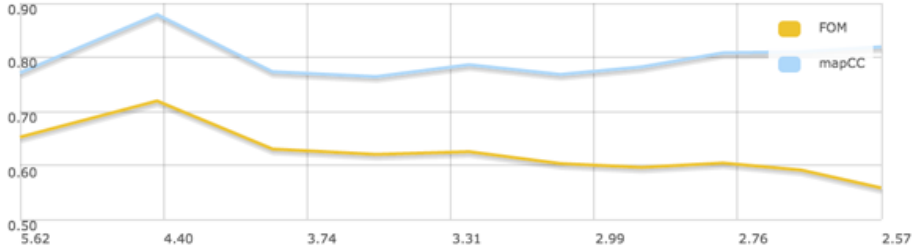
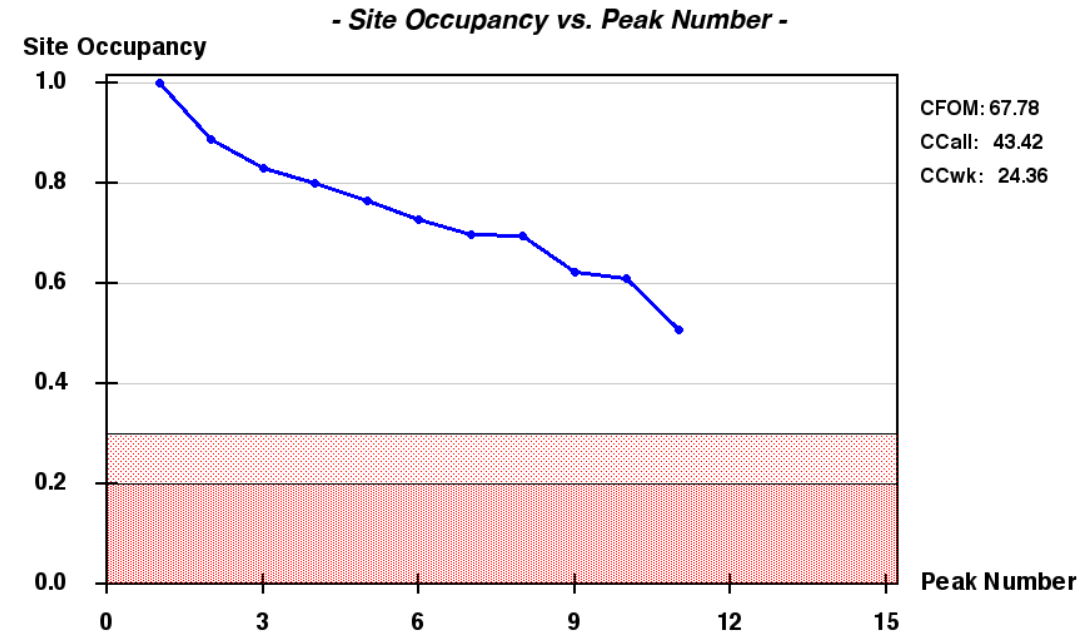


Figure 4.10: Se edge scan to identify selenium atoms in the BR55 crystal structure

A. Plot of the Se Edge scan for SeMet-BR55 (in grey), revealing the anomalous scattering coefficients, corresponding to the maxima (f'') and minima (f') of the diffracted wavelength, indicating the presence of Se atoms in the crystal lattice. **B.** Se atomic coordinates identified by the FastEP pipeline in IsPyb; an FOM of 0.62 indicates good quality data.

A.



B.

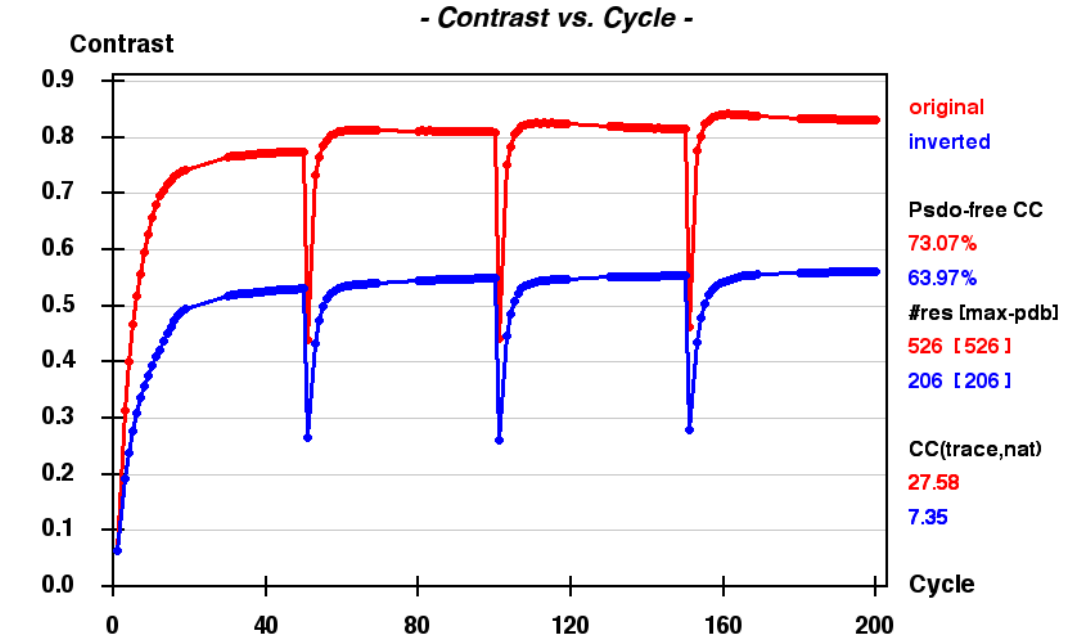
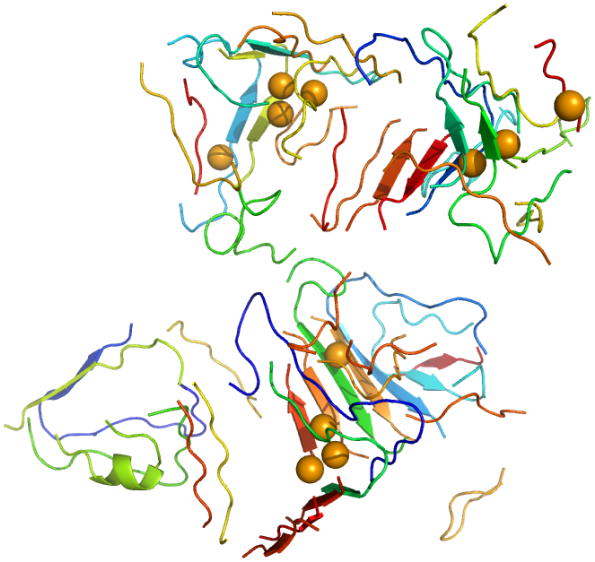


Figure 4.11: SeMet phasing results from SHELX

A. Graph of site occupancy vs. peak number, generated by SHELXD, showing the presence of 11 Se atoms in the asymmetric unit; in this case the FOM is higher than that shown in **Figure 4.10**B, at about 0.68. **B.** Plot of electron density contrast vs. density modification cycle generated by SHELXE, indicating that the original chirality of Se diffraction is the correct enantiomeric form for SeMet-BR55 diffraction.

A.



B.

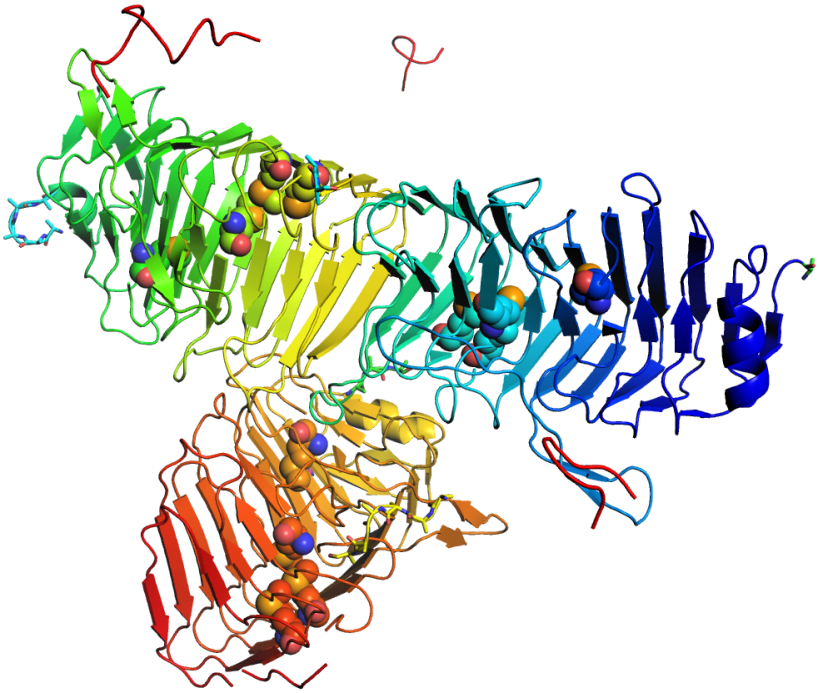


Figure 4.12: Initial SeMet-BR55 model versus Se atomic positions by SHELX

A. 11 Se atoms were identified by SHELXD, shown as ochre spheres in the output PDB file**. B.** An initial model of SeMet-BR55 built in Buccaneer after partial model building based on Se densities. The SeMet residues are shown as spheres and a total of 12 residues were built into the model. This was the search model for molecular replacement of native BR55 diffraction data with MOLREP.

Table 4.2: Data collection and refinement statistics for BR55 and SeMet-BR55

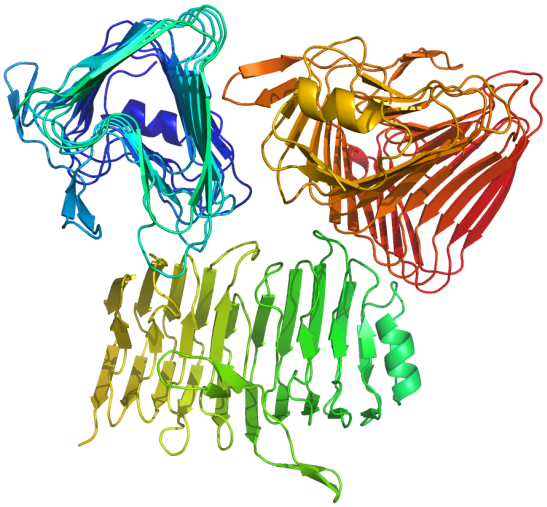
Statistics from 'Resolution' to '$CC_{1/2}$' are given in the format 'Overall (Outer shell)'. A brief definition of each statistical term is provided underneath the Table.

Protein	BR55	SeMet-BR55
Data Collection Statistics		
Beamline	i03	i04
Wavelength (Å)	0.9762	0.9792
Space group	P 3_2 2 1	P 3_2 2 1
Cell parameters: a, b, c (Å)	146.70, 146.70, 110.42	148.40, 148.40, 110.73
***α, β, γ* (°)**	90, 90, 120	90, 90, 120
Resolution (Å)	73.35 –1.92 (1.95 –1.92)	74.20 – 2.73 (2.80 – 2.73)
I/σI	9.9 (2.2)	29.7 (6.2)
Unique reflections	104897 (5168)	37723 (2781)
Completeness (%)	100.0 (99.6)	99.9 (99.9)
Multiplicity	18.0 (15.5)	39.0 (40.9)
R_{merge}	0.278 (3.087)	0.134 (1.134)
R_{meas}	0.286 (3.191)	0.136 (1.148)
R_{pim}	0.067 (0.797)	0.022 (0.178)
$CC_{1/2}$	0.997 (0.668)	1.000 (0.984)
Refinement Statistics		
Molecules per AU	3	3
Total atoms	7630	
Water molecules	733	
R_{factor}	0.1614	
R_{free}	0.1935	
Ramachandran analysis		
Most favoured	95.90	
Allowed	3.10	
Outliers	0.00	
R.M.S.D.		
Bonds (Å)	0.020	
Angles (°)	1.983	
Planes (Å)	0.009	
Mean atomic B-factor (Å^2)	32.00	

I/σI: signal to noise ratio; values >2 in the higher resolution shell is recommended [166]
Completeness: The number of reflections measured in a data set, expressed as a % of the total number of reflections at the specified resolution. Data completeness is crucial, especially in all resolution shells, although it is likely to be less in the outer (higher resolution) shells [148].
Multiplicity: Also called Redundancy, it is the average number of observations of each reflection; a higher multiplicity data set will innately be of better quality than one in which every reflection has only been measured once [165].
R_{merge}: A measure of precision for the multiple observations of each reflection, dependent on data multiplicity. [165].
R_{meas}: A measure of precision of the individual reflections, independent of its multiplicity [165].
R_{pim}: The precision-indicating merging factor. Describes the precision of the averaged reflection intensity measurements; it is the most informative R value for predicting the quality of a given data set. The lower the **R_{pim}**, the better the data performance [165].
$CC_{1/2}$: The correlation of one half of random observations (with approximately even distribution for each reflection) to the other half; it is a measure of internal consistency of the reflections. $CC_{1/2}$ tends to 1.0 at overall resolution and to 0.1 at higher resolutions [156,166].
R_{factor} is a measure of agreement between amplitudes of structure factors calculated from a crystallographic model and those from the original diffraction data [156].
R_{free} is a residual function calculated during structure refinement similarly to R_{factor} but applied to a small subset of reflections not used in refining the structural model [156].
B-factor: measure of atomic displacement from their mean position in a crystal structure. It is calculated as: $B = 8\pi^2\langle u^2\rangle$, where *u* is the displacement of a scattering center and $\langle u^2\rangle$ is the mean squared displacement. The B-factor for BR55 is within range (see APPENDIX 5).

Chain A (blue-cyan), B (green-yellow) and C (orange-red) were 304, 305 and 310 residues long, respectively. Chain A was missing 35 and 105 residues, Chain B was missing 33 and 106 residues, and Chain C was missing 33 and 101 residues at their N- and C-terminals, respectively. With 303 common residues amongst the 3 chains (264 to 566), the RMSDs between chain A and B is 0.25 Å, chain A and C is 0.29 Å, and chain B and C is 0.28 Å. As maximum sequence coverage is given by chain C, it is chosen as the standard model, named $BR55_{262\text{-}571}$, for subsequent structural analyses. The structure of $BR55_{262\text{-}571}$ is discussed in detail with BR78 in section 4.4.

A.



B.

```
229    ----------------------------------------------------MA TILPATGNND
241    GETSVANRTV TEGNIVQLAE NIEEVSNEEE LKAALRDASI TTIKLKNNIT LNNAITINNG
301    NRNITIIGDG HYINALNSDG GIILNNRGGS AKIDLTIENA TLYNTSKYGF VNMSSNGVDT
361    VTYKDVTAYG GTLVWSKTGA GVKTLNLVGN TTLNSVKSYE VDGQSCGTEA FSHRTPDGDK
421    TTALYVSNAI NIAENANVVL NNSATDIDMW LLTAVPSTSG ISTVTVGNNA SLTMENIGNT
481    EYNIKLDGGR ENHFIVNENA AVKMSAKVDN VRIIPQLENI FTRGNIELAK GSNVHLEVIT
541    GSNFRVAGTV ANRIDFNGTA TLIKQEGASG PTIYDKNSQA NIEFNWLGIF ENTVNFNPGA
601    DVTLIAGKGA SNIGSSSLVT ATVNINDPLR VAFNTDKLET DNGVTYLGNS RINVKVTNAT
661    IKVDDNDFSE LI----------------------------------------------------
```

Figure 4.13: Structural model of BR55 by X-ray Crystallography

A. Structure from a crystal obtained by the limited proteolysis of BR55 with 1:500 α-chymotrypsin. Three molecules were found in the asymmetric unit of the $P3_221$ cell, labelled as Chain A (blue-cyan), B (green-yellow) and C (orange-red). Chain A is $BR55_{264\text{-}567}$, Chain B is $BR55_{262\text{-}566}$, and Chain C is $BR55_{262\text{-}571}$. **B.** Protein sequence of BR55, where the residues in red indicate regions of the protein that may have been removed by proteolysis; the deleted regions are estimated from the sequence of Chain C for which the maximum number of residues are visualised by crystallography. This estimation is in good agreement with MALDI-MS data obtained from proteolysed BR55 crystals, presented in **Figure 4.7**.

4.3.4 BR78 Data collection and Structure Determination

1800 diffraction images of native BR78 were obtained at -196^{o}C and an X-ray wavelength of 0.9002 Å, and integrated by XIA2 to a maximum resolution of 2.14 Å. Indexing of the diffraction images in MOSFLM [147] suggested two possible space groups, C2 and H3, with very low penalty scores of 0 and 2, respectively. Cell refinement and integration of the data was performed for both space groups, however, POINTLESS determined the rhomohedral trigonal H 3 2 space group to be the best solution, with unit cell dimensions of a = b = 162.36, c = 146.78. This was followed by scaling of the data with SCALA [159]. In the resulting merged MTZ file, the resolution of the data was improved to 1.90 Å. Data processing statistics for BR78 are given in **Table 4.3**.

The Matthews coefficient indicated 1 monomer in the asymmetric unit cell of BR78, for 560 residues (of full-length BR78), and a solvent content of 58.4%. Phase identification of BR78 was done by molecular replacement using PHASER, with BR55 as search model, which had a primary sequence identity of 50%. Refinement of the PDB and MTZ output files from PHASER gave a solution with R_{factor} 0.4589 and R_{free} 0.4694, which was a sound result. A model of BR78 was then built by ARP/wARP [152], with 10 building cycles, and 5 refinement cycles per building cycle; a model with R_{factor} 0.2969 and R_{free} 0.3310 for 2575 built atoms was obtained. Subsequent cycles of manual building in COOT, and rigid body and restrained refinement in REFMAC yielded a final model with R_{factor} 0.2400, R_{free} 0.2842, $RMS_{bond\ angle}$ 1.905 and $RMS_{bond\ length}$ 0.029. The refinement statistics are given in **Table 4.3** below.

Validation of the model in MolProbity [163], showed 0.6 % Ramachandran outliers, and 94.1% favoured Ramachandran positions. The outliers corresponded to V378 (for which both, Ψ (Psi) and Φ (Phi) angles were out of the allowed range) and I411 (for which the Phi angle was out of range). However, a close re-examination of the model in COOT showed that the residue backbones were fitted appropriately in the density map. Additionally, it was noted that both these residues were found in disordered turn and loop regions of the BR78 protein, respectively. A clashscore of 7.88% was

obtained, which is in the 88th percentile amongst 793 PDB entries of similar resolution. Complete validation statistics are provided APPENDIX 5.

The detailed crystal structure of partially proteolysed BR78, as BR78$_{257-623}$, is given in **Figure 4.14B**. Despite 50% sequence similarity with BR55 it adopts an identical fold to BR55, with an RMSD of 0.435 Å. Like, BR55, it is missing residues at the N- and C-terminals, (59 and 134, respectively) due to limited proteolysis by 1:700 thermolysin. Additionally, there are two gaps in the model; the first gap is due to 9 missing residues, from 413-421 and the second gap is 30 residues from 568-597. These missing regions are due to diminished electron density in some areas of the crystal structure, which prevented the building of these residues.

Table 4.3: Data collection and refinement statistics for BR78
Statistics from 'Resolution' to '$CC_{1/2}$' are given in the format 'Overall (Outer shell)'.

Protein	BR78
Data Collection Statistics	
Beamline	i04
Wavelength (Å)	0.9002
Space group	H 3 2
Cell parameters: a, b, c (Å)	162.36, 162.36, 146.78
***α, β, γ* (°)**	90, 90, 120
Resolution (Å)	65.06 –1.90 (2.00 –1.90)
I/σI	14.4 (1.5)
Unique reflections	58298 (8418)
Completeness (%)	99.9 (99.5)
Multiplicity	9.10 (8.10)
R_{merge}	0.070 (1.396)
R_{meas}	0.074 (1.491)
R_{pim}	0.024 (0.517)
$CC_{1/2}$	0.999 (0.365)
Refinement Statistics	
Molecules per AU	1
Total atoms	2561
Water molecules	110
R_{factor}	0.2400
R_{free}	0.2842
Ramachandran analysis	
Most favoured	94.10
Allowed	5.30
Outliers	0.60
R.M.S.D.	
Bonds (Å)	0.029
Angles (°)	1.905
Planes (Å)	0.007
Mean atomic B-factor (Å^2)	44.40

4.4 Discussion: BR55 and BR78 adopt a right-handed parallel β-sheet solenoid fold

The BR55 and BR78 proteins have a solenoid type core comprising chiefly of β-strands folding together in a repetitive pattern to form a right-handed helix with 3 parallel β-sheets. This topology is very different to other SRRP binding region structures that were presented in **Figure 2.13**. A solenoid is described as comprising of repeating structural units arranged such that the polypeptide forms a continuous superhelix [167]. This arrangement is quite different to the structure of most globular proteins where the protein chain folds in more complex ways. 18 types of solenoid folds have been described and reviewed by Kobe and Kajava [167], based on the type of repeating structural unit and the arrangement of individual units in the solenoid. These include solenoids with α/α, α/α/α, α/β, α/α/β/β, β/β and β/β/β structural units. According to the classification convention of Kobe and Kajava [167] BR55 and BR78 are pectate lyase C (PelC) like, right-handed parallel β/β/β solenoid proteins; PelC from the plant pathogen *Erwinia chrysanthemi* (now known as *Dickeya didantii*) was the first parallel β-helix protein to be reported by Yoder et. al. in 1993 [168]. Since then the structures of several parallel β-helix proteins have been reported, with varying structural properties, such as those in **Table 4.8**, which is a compilation of PelC-like proteins from the same CATH superfamily as BR55 and BR78.

4.4.1 Nomenclature for structural descriptions of BR55 and BR78

The nomenclature and definitions in this subsection will be used in upcoming sections in discussions of BR55 and BR78.

BR55 and BR78 are composed of three parallel β-sheets, which will be referred to as PB 1, PB 2 and PB 3, according to the naming convention by Yoder. et. al. [169]. The disordered turns connecting consecutive β-strands in the individual structural unit are termed T1 (between PB1 and PB 2), T2 (between PB 2 and PB 3) and T3 (between PB3 and PB1); these structural features are shown in **Figure 4.14** and **Figure 4.15**. In this context, an individual structural unit is defined as the fundamental repeating unit of the β-solenoid core, which comprises of a complete PB1-T1-PB2-T2-PB3-T3 coil. Henceforth, 'structural unit' and 'coil' will be used interchangeably. The first β-strand in both structures, which precedes the α-helix, α2 (**Figure 4.14**) is counted as being part of PB1, as it is parallel to subsequent β -strands in the sheet, although it is not counted as a structural unit, as it is not part of the PB1-T1-PB2-T2-PB3-T3 sequence.

Individual secondary structure features, including β-strands have been numbered in order from N-terminal to C-terminal, as shown in **Figure 4.14**. The disordered regions have not been numbered.

In some places the individual structural units, or coils will be denoted as 'levels'. For both, BR55 and BR78 (**Figure 4.14A** and **B**, respectively**)** a flexible loop can be seen extending from PB2 of a coil in the middle of the solenoid body, and then folding back into the body, into PB3 of the same coil. This coil, from which the loop originates and culminates, will be considered as level '0'. Coils above '0' will be numbered positively (+1, +2 and so on) and those below '0' will be negatively numbered (-1, -2 and so on).

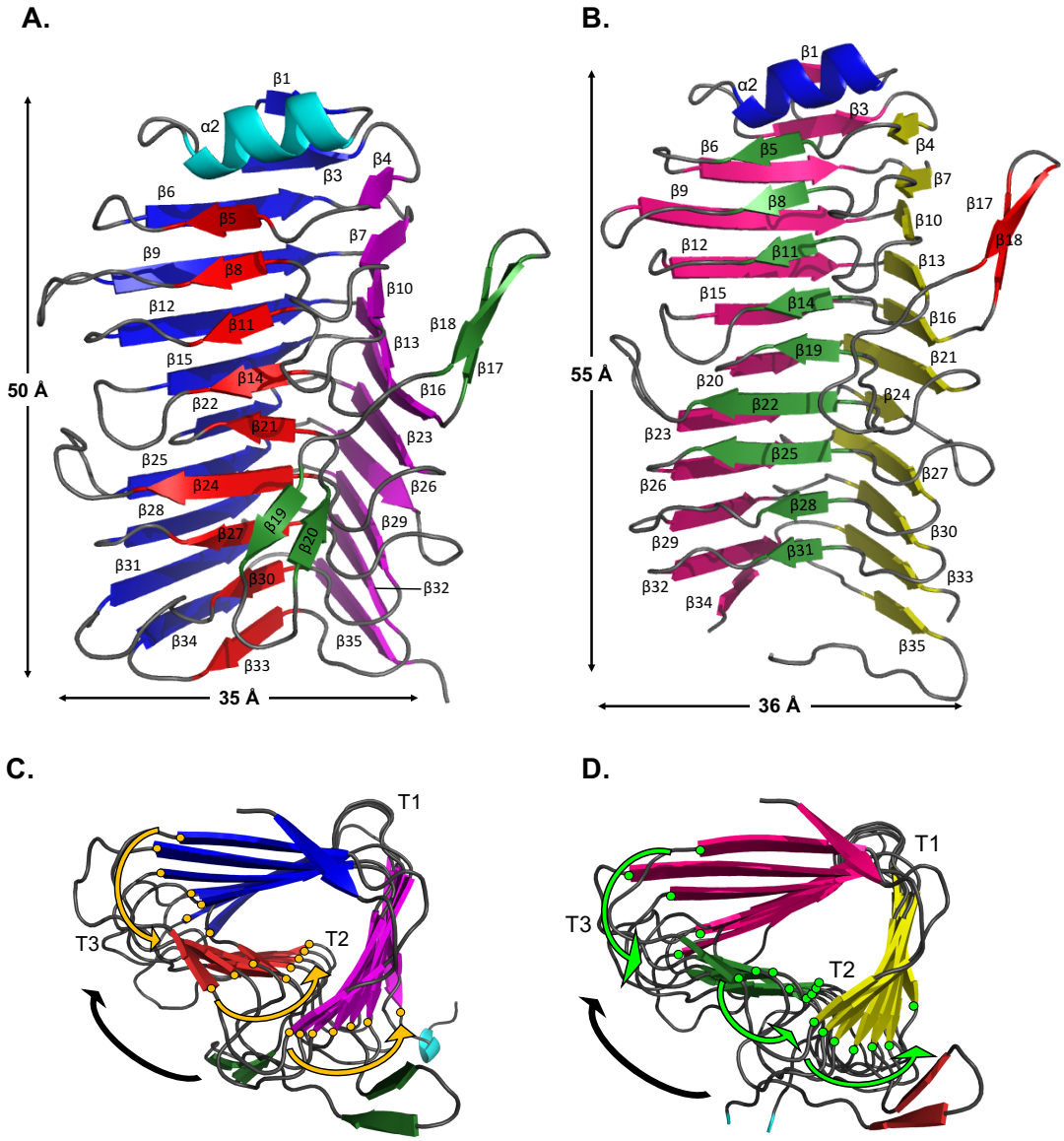


Figure 4.14: Crystal structures of BR55 and BR78

A. Longitudinal view of the structural model of $BR55_{262-571}$ where secondary structure features are colour coordinated with the topology diagram in **Figure 4.15**; β-strands in PB1, PB2 and PB3 are deep blue, magenta and red, respectively; α-helices are cyan **B.** Longitudinal view of the $BR78_{257-623}$ model. PB1, PB2 and PB3 β-strands are pink, yellow and green, respectively, and the α-helix is deep blue. There are two gaps in the model, the first of which is 9 residues, 413-421, from the lower half of the putative flexible binding loop between β-18 and β-19, and the second gap is 30 residues from 568-597, corresponding to a missing turn–β-sheet–turn region on PB3, between β-33 and β-34. Cross-section of $BR55_{262-571}$ **(C)** and $BR78_{257-623}$ **(D)** from N- to C-terminal. β1 and α2 in both are omitted for clarity. The black arrow shows the direction in which the polypeptide chains fold around the helical axis, yielding a right-handed superhelix. The helical twist down each β-sheet is indicated by yellow and light green dots and arrows in $BR55_{262-571}$ and $BR78_{262-571}$, respectively: β-strands in each parallel β-sheet twist leftwards with respect to each other down the helical axis.

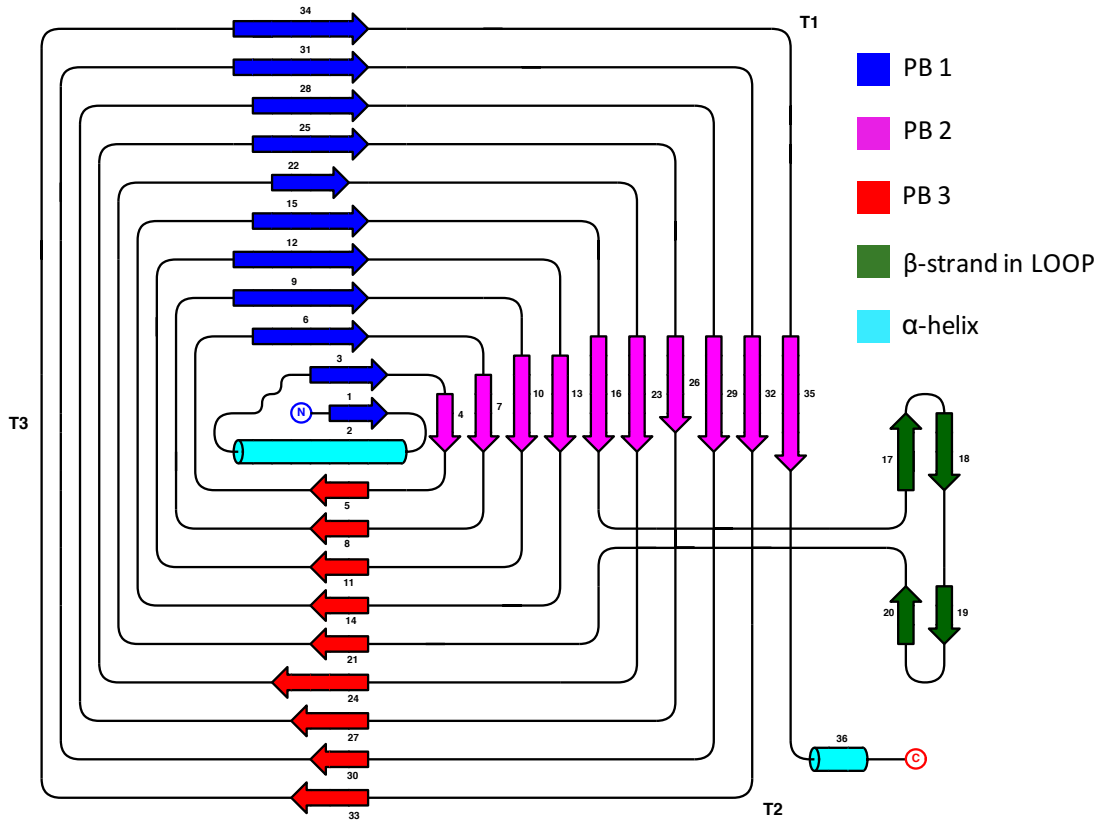


Figure 4.15: Topology diagram of BR55

A 2D representation of BR55's spatial folding in the form of a topology diagram, sketched on TopDraw within the CCP4 suite. The diagram shows BR55's architecture as a right-handed parallel β-helix solenoid, comprising of three of parallel β-sheets (PB1, PB 2 and PB 3), connected by disordered loops or turns, called T1, T2 and T3. The secondary structure features are drawn to scale and the relative alignment of individual β-strands in each parallel β-sheet are also represented to scale.

4.4.2 Comparing the topology and folding of BR55 and BR78

Despite a 43.2% sequence identity (between residues modelled by crystallography), the partial BR55 and BR78 structures in **Figure 4.14** share common structural features like: a right-handed parallel 3β-sheet superhelical solenoid core, an N-terminal α-helix that forms a 'cap' above the solenoid core, and a flexible loop that runs diagonally across PB2 and PB3 of the solenoid body. This loop extends from a β-strand in PB2 (β16 in BR55 and BR78) and folds back into a β-strand in PB3 (β21 in BR55 and β19 in BR78) of the same structural unit in at level 0 of the solenoid core. The differences noted between BR55 and BR78 structures are that BR78 has a longer lower solenoid body compared to BR55. As shown in **Figure 4.14**, up to 10 complete

solenoid turns or structural units were modelled for BR78, whereas 9 structural units were modelled for BR55. BR78 has a longer binding loop than BR55; the lower loop of BR78 has 5 more residues than that of BR55, and this may have a correspondence with the longer lower solenoid body of BR78. PB1, 2 and 3 in both proteins are of varying sizes, as summarised in **Table 4.4** below.

Table 4.4: Number of residues in parallel β-sheets of BR55 and BR78

Parallel β-sheets: BR55	PB1	PB2	PB3
No. of strands	10	10	9
No. of residues per strand	3,4,6,7,6,4,6,6,7,7	3,4,5,5,6,6,5,6,6,7	3,3,3,3,3,5,4,3,4
Mean residues per β-sheet	5.6	5.3	3.4

Parallel β-sheets: BR78	PB1	PB2	PB3
No. of strands	12	11	10 (9+ 1 missing)
No. of residues per strand	3,4,6,8,7,6,4,6,6,7, 6,4	3,4,5,5,6,6,5,7,6, 6,6	3,3,3,3,3,5,5,3,3,-
Mean residues per β-sheet	5.6	5.4	3.1

4.4.3 Structural homology of BR55

The Phyre^2 server uses advanced remote homology detection to build 3D models and predict ligand binding using the primary sequence of the target protein as input. Homology for regions of the sequence that can be modelled into a 3D structure are judged by % confidence of agreement to its model template [14]. The first ten hits for BR55 from Phyre^2 are given in **Table 4.5** below and demonstrate that structural homology for regions of BR55 are found with pectin/pectate lyases and pectinesterases, which are all enzymes that act on a substance called pectin, a structural polysaccharide found in plant cell walls [170]. The only exception was a putative A1 protease from *Bacteroides ovatus*, for which the substrate is human immunoglobulin A1. Nevertheless, all homologous structures, upon which BR55 was modelled had a right-handed parallel β-sheet solenoid fold.

A second homology search on the Dali server [15] provided results that were different from the Phyre^2 output, and consisted of proteins with a variety of functions and substrates, as presented in **Table 4.6**, including 4 polysaccharide lyases and

hydrolases, 2 extracellular pertactin adhesins, 2 phage tailspike proteins that degrade bacterial cell wall polysaccharides, 1 fructotransferase and 1 mannuroran epimerase. However, the common features of these proteins were their possession of a right-handed parallel β-solenoid fold and that their substrates are polysaccharides. Dali works by comparing the PDB coordinates of a query structure against existing ones in the PDB, and provides estimates of structural similarity primarily via a Z-score, for which Z-scores lower than 2 are spurious. The higher the Z-score the better the agreement between the query structure and the model [15].

Table 4.5: The top ten structural homology results from Phyre^2 [14]

The structural homology results are sorted by percentage confidence of amino acid coverage between BR55 and the model template.

Molecule	Confidence (%)*	Sequence ID (%)	PDB ID
Solanum lycopersicum Pectinesterase 1 [171]	35% residues at 93% confidence	13%	1XG2
Daucus carota Pectin methylesterase [172]	36% residues at 88% confidence	12%	1GQ8
Thermotoga maritima Pectate lyase 2 [173]	10% residues at 87% confidence	22%	3ZSC
Aspergillus niger pectin methylesterase [174]	27% residues at >85% confidence	15%	5C1C
Dickeya chrysanthemi Pectate lyase A [175]	25% residues at 79% confidence	21%	1PE9
Aspergillus niger Pectin lyase A [176]	17% residues at 78% confidence	23%	1IDK
Bacteroides ovatus immunoglobulin A1 protease [177]	25% residues at 74% confidence	15%	3NZ6
Bacillus subtilis Pectate lyase [178]	25% residues at 68% confidence	26%	1BN8
D. chrysanthemi pectate lyase E [169]	24% residues at 67% confidence	24%	1PCL
Xanthomonas campestris Pectate lyase [179]	24% residues at 45% confidence	15%	2QY1

* Confidence is the probability of the sequence and template being homologous and is not an indication of model quality. A match with >90% confidence generally implies that the protein core is modelled at high accuracy (2-4 Å rmsd from the protein's true structure) and that the protein is very likely to adopt the overall fold shown. Surface loops are likely to deviate from the native structure.

Table 4.6: The top ten structural homology results from the Dali Server

The results are sorted in order of decreasing Z-score. Additional measures of agreement between the input and template model is given via the RMSD. 'Lalign' indicates the number of residues that are aligned between the query and template, for which the modelling statistics are provided.

Molecule	Z-score	RMSD	Lalign	Sequence ID (%)	PDB ID
Bordetella bronchiseptica pertactin extracellular domain (Prn-E) [180]	18.5	2.8	233	13	2IOU
Pectobacterium carotovorum Poly-galacturonase [181]	18.0	3.1	242	13	1BHE
Shigella phage Sf6 tailspike protein [182]	17.9	3.1	243	8	4URR
Bordetella pertussis P.69 pertactin [183]	17.8	2.6	214	11	1DAB
E.coli phage HK620 tail spike protein [184]	17.3 – 17.8	3.1	245	11	2X6X, 4XKW, 4XMY, 4X19, 4XON, 4XNF, 4XGF, 4XN3, 2VJJ, 4XOP, 4XLH, 4XLC, 2VJI, 4XNO, 4YEJ, 4XKV, 4XOR, 4XM3, 4XLF, 4XOT, 4XGH, 2X6Y, 4XLE, 4XLA, 2X6W, 4YEL, 4XR6, 4XQI
T. maritima exopoly-galacturonase [185]	17.8	3.1	246	9	3JUR
Yersinia enterocolitica Family 28 glycoside hydrolase [186]	17.8	3.2	251	8	2UVE, 2UVF
Bacillus sp. inulin fructotransferase [187]	17.3	2.9	236	11	2INU, 2INV
Ruminiclostridium thermocellum poly-saccharide lyase [188]	17.2	2.6	206	12	4PHB
Azotobacter vinelandii mannuronan C5 epimerase [189]	17.1	3.0	238	12	2PYG, 2PYH

4.4.4 PelC-like structural features of BR55 and BR78

A search against BR55's PDB model on the CATH protein structure classification database [18] suggested that BR55, and thus, also BR78 belong to the 2.160.20.10 CATH superfamily of single-stranded, right-handed beta-helix, pectin lyase-like proteins. This superfamily consisted of 181 entries, including PelC. The CATH hierarchy is as follows:

C: Class (2) – Mainly Beta

A: Architecture (160) – Solenoid

T: Topology (20) – Pectate Lyase C-like

H: Homologous superfamily (10) – Single-stranded, right-handed β-helix, pectin lyase-like

The three key structural properties of a solenoid superhelix are its 'handedness', its solenoid twist and its curvature [167]. The 'handedness' is defined as the direction in which the polypeptide chain folds around the helical axis. Most structurally characterised solenoids to date are right-handed, as there is an energetic preference for right-handed connections between secondary structure units in a polypeptide chain [190]. BR55 and BR78 are right handed helices like all known PelC-like proteins (also see **Table 4.8**). When neighbouring repeats in a solenoid rotate relative to each other along the helical axis they produce an overall twist. A slight left-handed twist is noted through the cross-section of BR55 and BR78 structures, similar to other PelC-like proteins [167,191], and these are illustrated in **Figure 4.14C** and **D**. Helical curvature arises due to the rotation of adjacent structure units, relative to each other, in directions other than along the helical axis. However, because inter-coil distances in β-solenoids are determined by uniform hydrogen bonding between successive β-strands, β-solenoids usually do not have any curvature [167], which is the case for alll PelC like proteins [169], BR55 and BR78.

Of the three parallel β-sheets in BR55 and BR78, two (PB1 and PB3) form an antiparallel β-sandwich fold, to which the third β-sheet (PB2) is approximately perpendicular. This results in a V-shaped cross-section for the parallel β-helices of BR55 and BR78 (see **Figure 4.14C** and **D**), similar to other PelC-like proteins [169]. This β-sandwich is different to other known β-sandwich folds, in being composed entirely of parallel β-sheets, albeit them being anti-parallel relative to each other. The mean distance between the Cα of the two β-sheets in BR55 is 10.3 Å, which is within the range observed for anti-parallel β-sheet sandwiches [192].

Most structurally characterised PelC-like proteins to date have functions as pectate lyases, pectin methyl-esterases and glycan hydrolases, as summarised in **Table 4.8** in section 4.9. A common feature in them is the protrusion of flexible loops or domains from the core β-solenoid body – which vary in size and conformation – commonly extending over a cleft in the solenoid core which harbours the binding or catalytic site of the protein. Some examples are seen in the complexes of *E. chrysanthemi* PelC [193] with tetragalacturonate, XylC (a xylosidase) with xylobiose and xylose and another pectate lyase from *Thermatoga maritima* which is presented in **Figure 4.16** below. In BR55 and BR78, a paperclip-like loop originates from β16 in PB2 of coil 0 and folds back into β21 and β19, respectively, in PB3 of coil 0. This loop is oriented diagonally across the solenoid core and is divided into the upper loop and lower loop regions; the lower loop could not be modelled for BR78. The upper loop in BR55 is situated across levels +1 to +3 of PB2, and the lower loop extends across -1 to -4 in PB3, where a cleft in T2, between PB2 and PB3 is observed (see **Figure 4.14C**).

It is well known that the tertiary fold is one of the key aspects dictating the biological function of a protein or its domain [194]. Therefore, considering the significant structural homology of BR55 and BR78 with PelC like proteins, it can be proposed that the putative binding regions of *L. reuteri* SRRPs adhesins are likely to have similar binding ligands or mechanisms to PelC like proteins. Henceforth, further in depth structural comparisons of BR55 with PelC-like proteins are reported, which endeavour to provide further clues on the putative binding site and suggested binding ligands for BR55 (and BR78).

4.4.5 Identification of potential binding sites in BR55

Throughout the thesis the focus is mostly on the structural model of BR55, as the ordered solenoid core and flexible loop of $BR55_{262\text{-}571}$ has been built completely, without any gaps, unlike BR78. As explained in section 4.4.3, BR55 shows structural homology chiefly with polysaccharide lyases, hydrolases and esterases, all of which have a right handed parallel β-helix fold (see **Table 4.5** and **Table 4.6**). Some proteins

amongst the hits were co-crystallised with a ligand, and from these the structure that had the lowest RMSD (2.8 Å), with a Z-score of 16.3 was that of a pectate lyase (Pel) from *Thermotoga maritima* (TM) co-crystallised with trigalacturonic acid (TGA) at pH 4.2 (**Figure 4.16A**) [173]. Ten TMPel residues were observed to be involved in TGA binding (**Figure 4.16C**), including 6 with positively charged side chains: R110, K163, R197, R200, R202 and R264, one negatively charged residue D171 and 3 with polar side chains: Q196, N224 and Y229. The solvent accessible surface electrostatic potential (EP) map of TMPel, prepared with the ABPS programme [20], revealed a highly basic TGA binding pocket, which would likely enhance binding to an acidic substrate like TGA (see **Figure 4.16B**).

Superposition of this structure with BR55 was achieved with Superpose [17] to recognise the potential binding site of BR55, which is presented in **Figure 4.17.** The RMSD in this case was 2.63 Å. The binding region in BR55 is suggested to be underneath the flexible loop, which seems reasonable given that: **1)** The loop is the only region with apparent flexibility to accommodate a TGA-sized ligand, within the domain resolved by crystallography. As mentioned earlier, for all PelC-like proteins reported so far, flexible domains extending from the rigid solenoid core were usually sites for binding and catalysis; **2)** Like TMPel, the lower loop region was significantly basic versus the rest of the BR55 molecule (**Figure 4.21A)**; **3)** Compared to TMPel, the putative binding region in BR55 appeared to have a richness of aromatic residues not found in other parts of the molecule. This is a known feature of the binding sites of carbohydrate binding proteins [195-197]. Another exception is the upper loop region of BR55, in which four Tyr residues were found, implying that the upper loop may potentially be a second binding site in BR55, albeit it was not significantly basic. Aromatic residues in BR55 (and BR78) are discussed further in section 4.4.6.1.

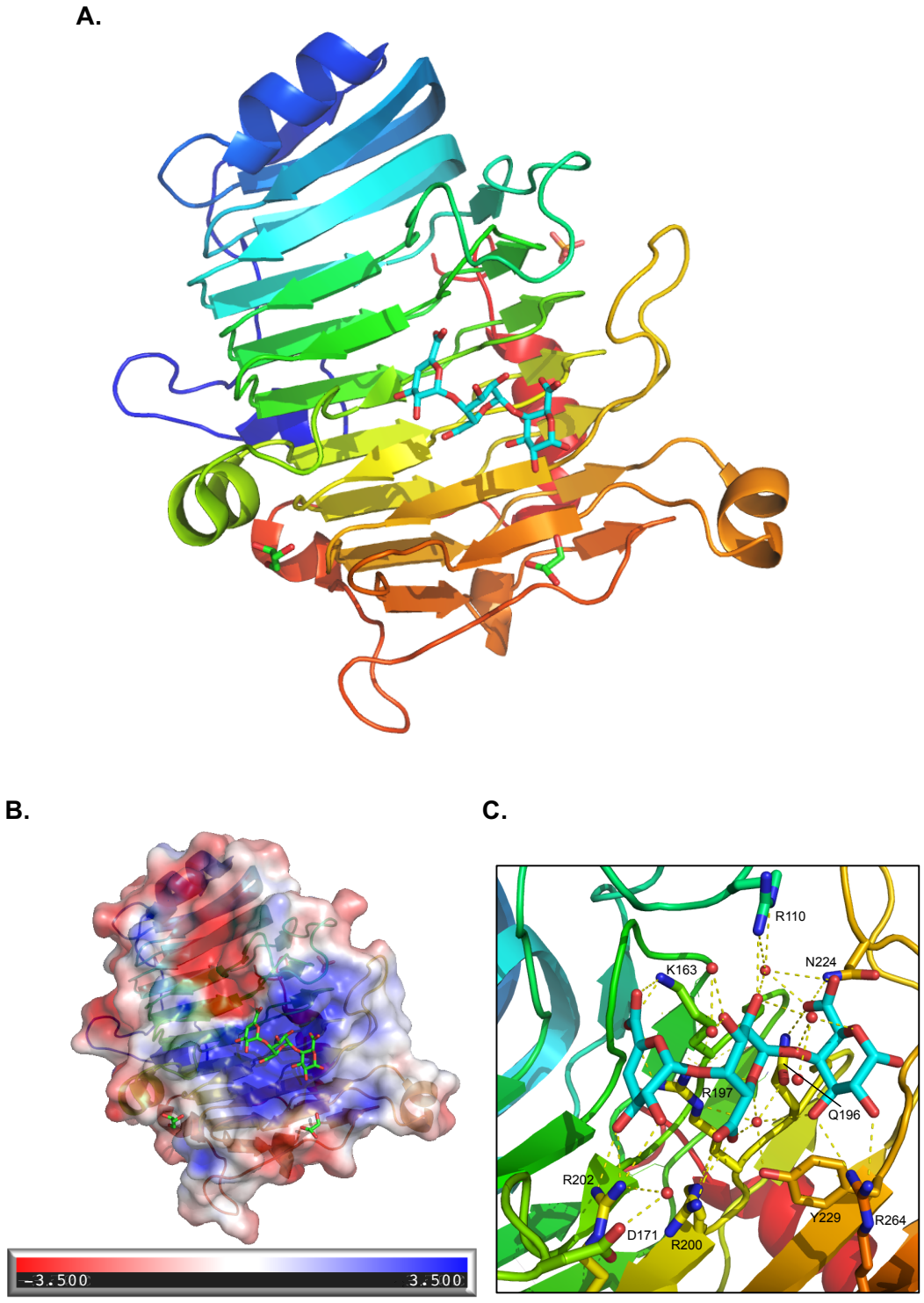


Figure 4.16: Crystal structure of *Thermotoga maritima* pectate lyase in complex with trigalacturonic acid (TGA) at pH 4.2

A. *T. maritima* pectate lyase is also a PelC like protein, from the same CATH superfamily as BR55. A crystal structure of the entire pectate lyase complexed with TGA (in cyan) is presented. **B.** Surface electrostatic potential map of the pectate lyase, showing a highly basic binding pocket **C.** A close-up view of the pectate lyase binding pocket, with residues involved in TGA binding shown as stick representations.

In accordance with the results of the superposition, BR55 residues were selected for mutagenesis from around the lower loop of BR55 for further functional studies. Four of them were aromatic residues, W375, Y425, W450 and Y482, and two were basic residues, K377 and R512A, as shown in **Figure 4.17** and **Figure 4.19**. The procedures for the generation, expression and purification of these mutants is beyond the scope of this thesis, however they are provided in APPENDIX 4.

Y482 and R512 in BR55 showed high structural conservation with respect to Y229 and R264 in TMPel (see **Figure 4.19B**). Below, amino acids surrounding the residues identified for mutagenesis (in magenta) are shown; boxed residues form part of a beta-strand and the unboxed residues are on the turns.

Residue number: 373.....423.....447.....482.....509.....
Amino acid: LVWSK.....ALYVS.....IDMWLLT.....YNIKLDG.....DNVRIIP.....

Except for K485 and D487, which were on beta-sheet 30, adjacent to Y482A, other residues in the vicinity were either non-polar, or if polar, the side chains were facing inside the superhelix, making them solvent inaccessible. K485 and D487 were the only charged residues that seemed reasonably placed within the binding region and also accessible to potential ligand binding. They were however, not chosen for mutagenesis in the first instance. Instead, a double-cysteine mutant was proposed, which was achieved by mutating P416 and Q516 to cysteine, theoretically placing them within a reasonable distance to form a disulphide bond that could potentially prevent the lower loop from opening for ligand binding, if this is indeed its mechanism of action. Furthermore, the lower loop (shown in green in **Figure 4.17B**) was removed in one mutant to assess whether the loop also has a role in ligand binding, or that binding is mainly dependent on the basic surface under the loop.

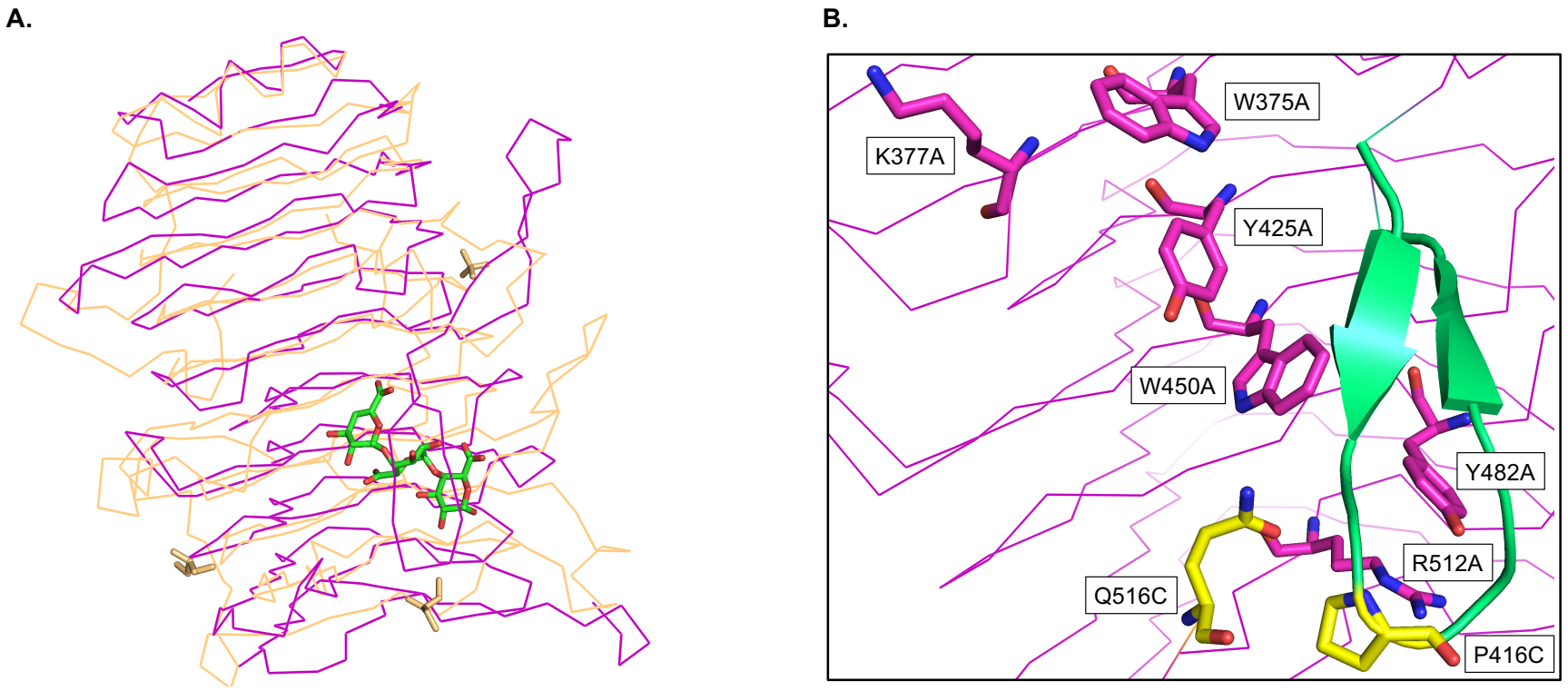


Figure 4.17 BR55 residues for mutagenesis.

A. Superposition of BR55 (purple) and TMPel (light orange) in complex with TGA (green), with an RMSD of 2.63; both are shown as cartoon representations. The proposed BR55 binding loop and pectate lyase binding region clearly overlap with one another, suggesting that the putative binding region and residues for BR55 may lie under the lower loop. **B.** Residues for BR55 site-directed mutagenesis. Residues in magenta shown as stick representations were mutated to alanine, as single point mutations; P416 and Q516, in yellow, were converted to generate a double cysteine mutant, with the potential to form a disulphide bond that would deter the opening of the binding loop; the lower half of the loop from F411 to T422, in green, was removed by mutagenesis to create a lower loop deficient mutant for functional studies

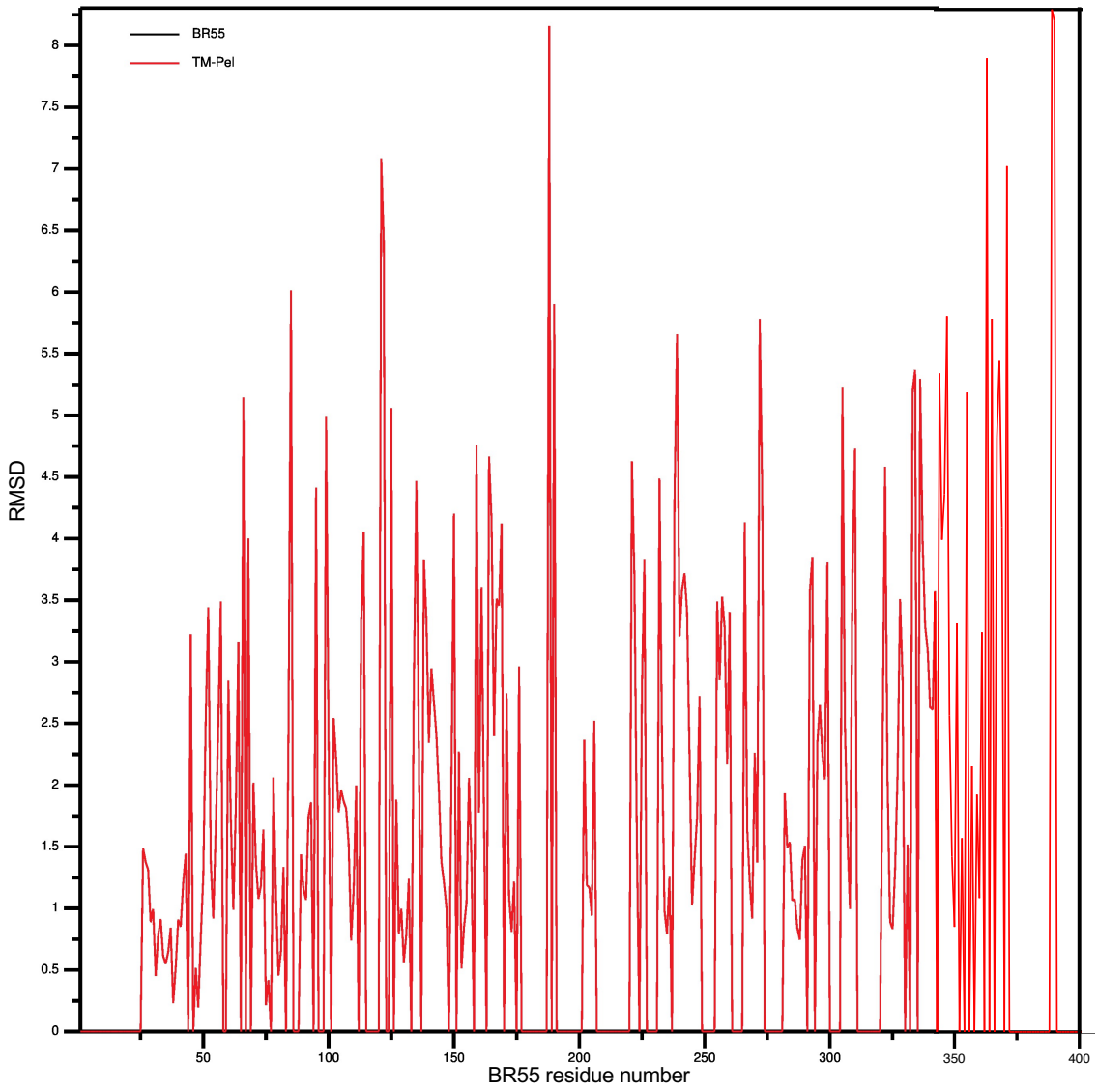


Figure *4.18*: RMSD per residue plot for TM-Pel *(in red)* against BR55 *(in black)*

Prepared using MultiSeq [198], which is included in the VMD (Visual Molecular Dynamics) suite of programs. The RMSD per residue values in this plot compare the amino acid residues of TM-Pel with BR55 and how well they align.

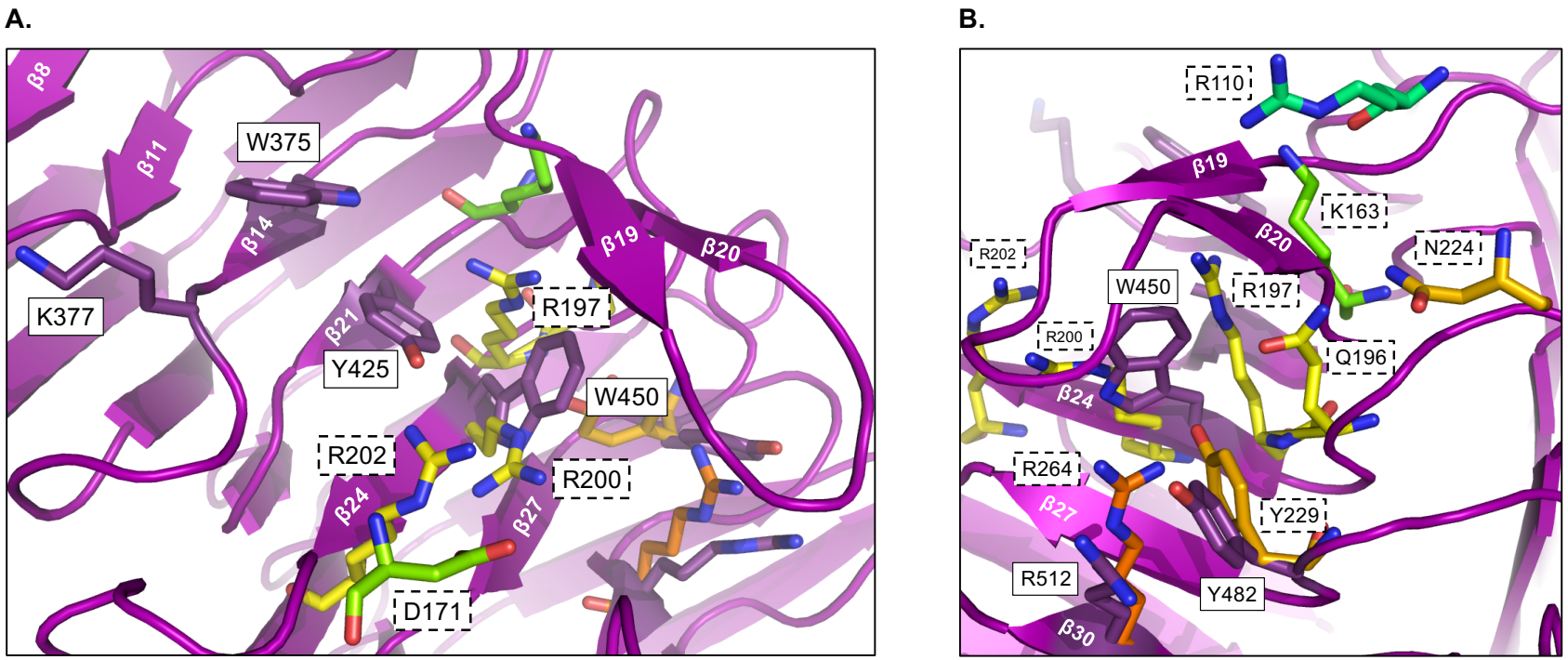


Figure 4.19: Identification of BR55 residues potentially involved in binding, superposed with TM pectate lyase binding residues

Only the TGA binding residues of pectate lyase are superimposed on the BR55 structure, to highlight its potential binding residues. Labels for pectate lyase residues are in dashed boxes, whilst those for BR55 are in solid-lined boxes. **A.** The lower loop binding region of BR55 from a left-hand view. W450A was the only polar (and aromatic) residue to be identified on β-24, corresponding to the TMPel β-strand on which R202, R200 and R197 were present; W375 and Y425A were other polar aromatic residues in the vicinity that were identified as potential binding residues, together with the charged K377. **B.** The lower loop BR55 binding region from a right-hand view; Y482 and R512A of BR55's putative binding region clearly show structural conservation with TMPel's TGA binding residues Y229 and R264, respectively, strongly implying their involvement in ligand binding.

4.4.6 Comparison of the putative binding region in BR55 and BR78

Superposition of BR78 upon the structure of BR55 allowed the putative binding regions in BR78 to be recognised, which showed a high degree of conservation with BR55. Of the residues with proposed roles in ligand binding in BR55, W375, Y425, W450 and R512 were conserved in BR78 as W373, Y428, W453 and R515, respectively, suggesting a role for them in ligand binding. The positive K377 was replaced by the polar Q375 residue in BR78, which would likely have a mild positive charge at native pH conditions in the acidic murine forestomach (ranging from 3.8 to 5.1, depending on food intake [199]). However, the corresponding residue for Y482 in BR78 was non-polar L485. Also noteworthy is the conservation of residues K488 and D490 in BR78, which correspond to K485 and D487 in BR55. **Figure 4.20** shows the primary sequence alignment of BR55 and BR78; secondary structure elements were denoted for regions underneath the lower and upper loops, and also of the loop region. Coil levels are also indicated.

Solvent accessible surface electrostatic potential (EP) maps were created for BR55 and BR78 from their PDB models using the ABP/PDBQ programme and these revealed stark differences in the charge distribution on the binding surface in BR55 and BR78. While the putative binding region in BR55 was highly basic, the corresponding region in BR78 was highly acidic. However, it must be recalled that the lower loop in BR78 was missing from the model and this may have had an effect on the surface charge in the putative binding region. To analyse the effect of the binding loop on charge distribution in the area, the lower loop was removed in the BR55 model from F411 to T422 (BR55-del). A significant reduction in the surface positive charge was noted in the putative binding region. **Figure 4.21** compares the surface EP maps for $BR55_{262\text{-}571,}$ BR55-del and $BR78_{257\text{-}623}$. This implies two things:

1) that the BR78 binding pocket could be less acidic than that observed in its EP map

2) that the proposed binding site of BR78 may be less basic than that of BR55, implying preference for a ligand with slightly different properties.

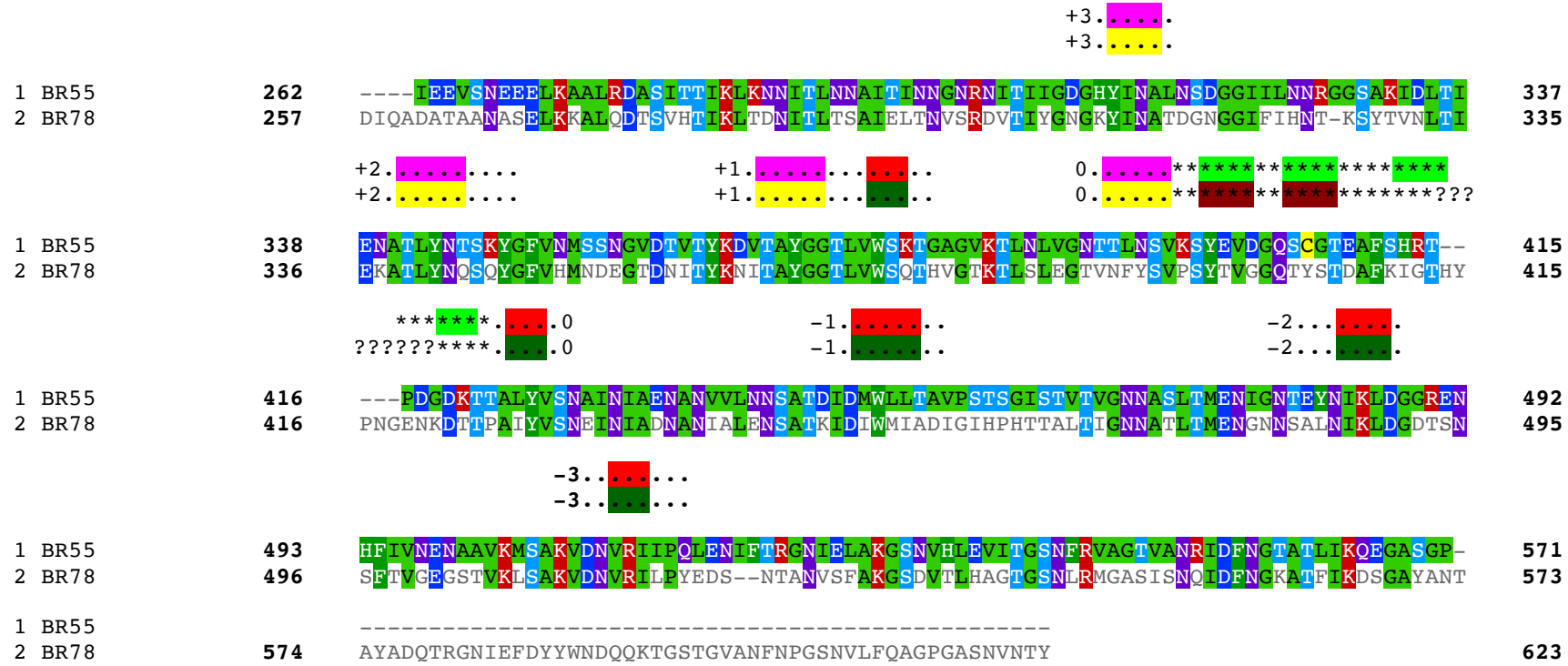


Figure 4.20: Sequence alignment of BR55$_{262\text{-}571}$ and BR78$_{257\text{-}623}$.

Residues in the upper (mostly PB2) and lower (mostly PB3) loop regions are highlighted. Dots (.) and asterisks (*) denote residues that are part of the solenoid core and the loop, respectively. Unmodelled BR78 residues are marked with '?'. BR55 PB2, PB3 and loop β-strands are highlighted in magenta, red and light green, respectively. Residues forming BR78 PB2, PB3 and loop β-strands are shown in yellow, dark green and brown, respectively. The numbers adjacent to each annotation indicate the coil 'levels', as discussed in 4.4.1.

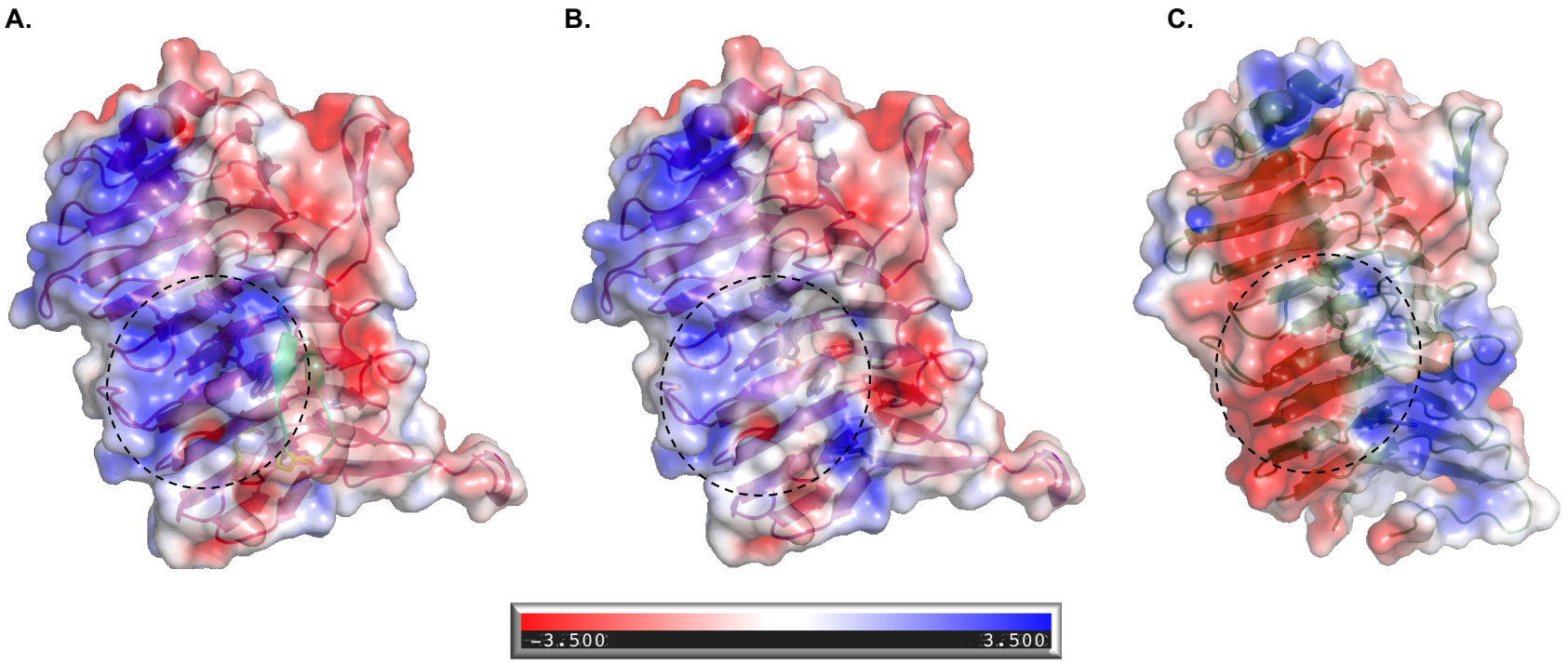


Figure 4.21: Solvent accessible surface electrostatic potential (EP) maps of BR55 and BR78

The putative, binding region underneath the lower loop is indicated with a dashed oval. **A.** $BR55_{262\text{-}571}$: the putative binding region under the lower loop shows a significantly basic charge distribution in the upper half, where W375, K377, Y425 and W450 are present. **B.** $BR55_{262\text{-}571}$–del, where the lower loop from F411 to T422 was removed prior to EP modelling. The basicity of the binding groove appears to be significantly reduced after this removal, thus implying that the lower loop plays a role in maintaining a basic binding pocket for pectate-like ligands that carry a net negative charge. **C.** BR78 $_{257\text{-}623}$, which was modelled without the lower putative binding loop (see **Figure 4.14B**), showing a notably acidic binding pocket, compared to $BR55_{262\text{-}571}$ -del.

4.4.6.1 Aromatic residues in the putative binding site

Solvent exposed aromatic residues are noted in particularly high occurrence, lined up within the loop region, compared to rest of the BR55 and BR78 molecules, many of which are engaged in stabilising aromatic interactions, as shown in **Figure 4.22**. Conservation of these residues between BR55 and BR78 in the loop region, and structural homology with PelC-like proteins indicate three things, that will be explored in further detail below: 1) That both, the lower and upper loop regions may be involved in ligand binding, 2) that the putative ligand is likely to be a plant-based carbohydrate with non-polar substituents 3) a longer β-solenoid core, a longer lower loop and a wider spread of aromatic residues within the putative binding groove of BR78 suggests that it may be able to bind larger glycans than BR55.

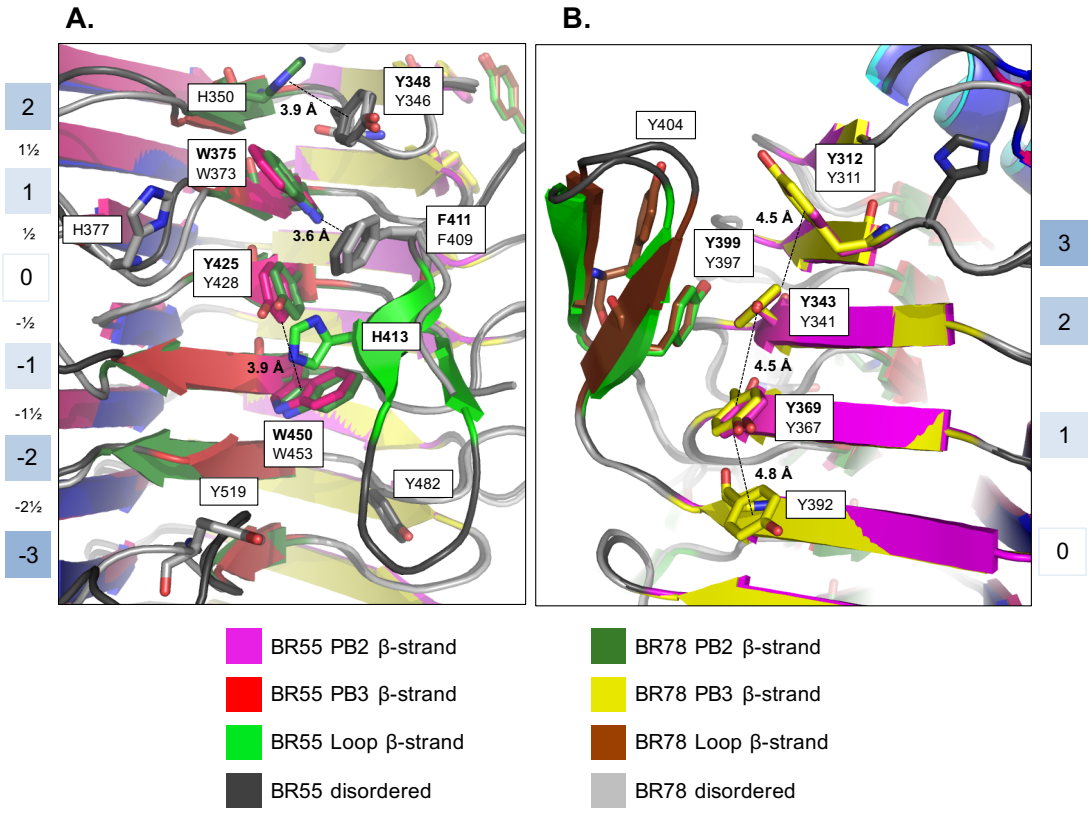


Figure 4.22: Solvent-exposed aromatic residues in the overlaid lower (A) and upper (B) loop regions of BR55 and BR78

Residues in bold are from BR55. It is seen that most residues in both loop regions are conserved, with some exceptions. Aromatic interactions are indicated with dashed lines. Residues on the lower loop intercalate with those from PB3 and these are labelled in halves.

Surface exposed aromatic residues that were conserved between BR55 and BR78 in the mid and lower loop region are Y348/Y346 (+2, on T2), W375/W373 (+1, on PB3), F411/F409 (+1½, mid-loop), Y425/Y428 (0, on PB3) and W450/W453 (-1, on PB3). In the upper loop region, residues Y312/Y311 (+3, on PB2), Y434/Y431 (+2, on PB2), Y369/Y367 (+1, on PB2) and Y399/Y397 (upper loop) are conserved between BR55/BR78, respectively.

Additional residues in BR78: H350 (+2, PB3), forming a face-to-face interaction with Y346 (T2). H377 (-1, on T3), in the lower loop binding region. H413 (-½) on the lower loop in BR55, interacting with Y425 (0, PB3); there is a histidine, H414, and a Y415 situated on the BR78 lower loop, which could not be modelled, however it could be in a different position to the one in BR55 as the lower BR78 loop is longer than BR55's, given that it has an additional 5 residues. Additionally, a corresponding aromatic residue to Y519 (-3, on T3) in BR78 is missing in BR55 (which is Q516 instead). It is proposed that Y519 may form some sort of interaction with the non-modelled H414 and Y415 residues in BR78, and may also have a role in ligand binding due to its proximity to the aromatic loop residues, further enlarging the putative binding site in BR78, compared to BR55.

4.4.7 Closing remarks for $BR55_{262-566}$ and $BR78_{256-623}$ X-ray crystallography

Following structural homology studies with the crystal structure of TMPeI complexed with TGA at pH 4.2 (section 4.4.5), TGA co-crystallisation and BR55 crystal soaking experiments were attempted from pH 4.0 to 8.0, however, none of the diffracted crystals showed evidence of TGA complexation. Sections 4.3 and 4.4 discussed findings from the partial crystal structure of $BR55_{262-566}$, and compared them to the structure of $BR78_{256-623}$. No structural information could be gained about the cleaved appendages in either BR55 or BR78. Section 4.5 will report the Small Angle X-ray Scattering (SAXS) results of full-length BR55, which provides information on the solution structure of BR55, from pH 4.0 to 7.0, and also provided information on the behavior of the appendages that were cleaved off during limited proteolysis treatment.

4.5 An Introduction to Small Angle X-Ray Scattering (SAXS)

SAXS is rapidly gaining prominence in structural biology for the low-resolution (50 Å to 10 Å) characterisation of biological molecules in solution [200]. It allows the analysis of the tertiary and quaternary structure of proteins averse to crystallization, like those that possess flexible domains or are intrinsically disordered. It can also be used for proteins and complexes over a broad range of sizes, from a few kDa to GDa, under a range of conditions, varying from extremely denaturing to near physiological [201]. Size is a limiting factor for structural methods, such as Nuclear Magnetic resonance spectroscopy (NMR), where proteins >20kDa are analysed with difficulty (data for a 40 kDa protein being the maximum achieved to date [202]). Cryo-electron microscopy (cryo-EM) is advantageous for medium to low resolution analysis of macromolecular complexes in the MDa range but challenging for structures <300 kDa [203, 204].

Data collection and sample characterisation is quicker in SAXS, compared to crystallography, NMR and cryo-EM and it can be jointly used with other structural and *in-silico* methods, to provide comprehensive structural information [201]. Like crystallography and NMR, SAXS requires mg amounts of pure and monodisperse protein at high concentration (1–10 mg/mL) [200]. Whilst stringent sample purity and high concentrations are not requirements for cryo-EM, sample preparation is relatively complicated and the technique is expensive compared to the others [203].

A scheme of a typical SAXS experiment set up is presented in **Figure 4.23**. Similar to X-ray crystallography, SAXS utilises the elastic scattering of incident X-rays to obtain a 2D scattering pattern containing information on the averaged electronic distribution of particles in the scattered X-ray beam. A monochromatic incident X-ray beam is passed through the sample and diffracted X-rays are measured by the detector. Often, a beam stop is employed to minimise detector damage (by absorbing the intensity of the incident beam behind the sample [205]) and to limit noise around the center of the detector so that the signal-to-noise ratio of the scattered reflections are improved. The intensity of diffracted X-rays is measured as a function of their

scattering vector or momentum transfer, *q*, which is the difference in momenta between the scattered wave vector and the incident wave vector [206]. It has units of inverse wavelength, nm^{-1} or $Å^{-1}$ and is derived from the wavelength of the incident beam (λ) and the half scattering angle (θ):

$$q = \frac{4\pi(\sin\theta)}{\lambda} \qquad \text{Equation 4.1}$$

Scattering Intensity is measured against *q*, instead of the scattering angle (2θ) so that intensity measurements become independent of the incident wavelength (λ). Sometimes, the momentum transfer, *q* is expressed in another form, *s* (also in units of nm^{-1} or $Å^{-1}$) where:

$$s = \frac{2\sin\theta}{\lambda},\ and\ so\ s = q/2\pi \qquad \text{Equation 4.2}$$

In this thesis, all scattering data is reported in *s.* A scattering curve, *I*(*s*) for the sample is obtained by subtracting a buffer blank of the solvent from samples in the same solvent. By normalising these intensities to the flux density (photons/s/mm^2) of the incident beam and the exposed sample volume it is possible to present these data as absolute intensities. Primary data reduction on the scattering patterns by radially averaging the recorded intensities as a function of momentum transfer, *q,* produces a 1D plot of absolute intensity as a function of momentum transfer (*I*/*q*). This is usually plotted as log (I)/*s*. This modification allows the subsequent analyses of data, which reveal information about the size and shape of the protein in solution [200, 201].

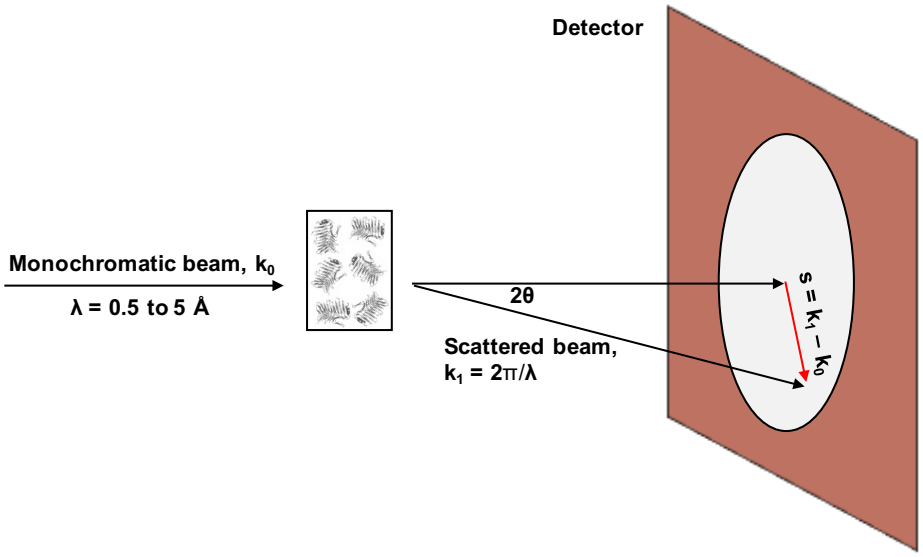


Figure 4.23: Schematic depiction of a typical SAXS experiment set up and radially averaged data.
Adapted from [200] and [201].

4.5.1 Information from scattering data in order of data processing

1) Inter particle attraction (aggregation) or repulsion effects

Inter particle attraction or repulsion effects in a sample can be determined by examining the shape of the scattering curve at low *s*. Aggregation effects often cause a slight increase in intensity at low *s* (upward smiling of the scattering curve), whereas inter particle repulsion causes a decrease in low *s* scattering intensity (downward smiling of the scattering curve); these effects are depicted in **Figure 4.24** below. Inter particle effects often manifest at higher concentrations, and for any given protein sample, a concentration dependent change in the shape of the low *s* region is often noted from high to low concentration. Merging low *s* scattering data from low concentrations (where the inter particle effects are much smaller), with data from higher *s* regions for high concentration data, is often used as a processing strategy to eliminate effects of aggregation without compromising on the enhanced signal-to-noise ratio afforded by a higher sample concentration.

2) The radius of gyration (R_g)

Low angle data can be plotted as ln(*Is*) versus s^2 to produce a Guinier plot [207], from which the radius of gyration (R_g) can be estimated . R_g is a measure of spatial particle size, which is different from molecular mass. A line of best fit is plotted such that the gradient, sR_g is <1.0-1.3, and can be extrapolated to estimate the scattering intensity at zero angle, *I*(0), which is given by the y-intercept [200]. A representation of this is given in **Figure 4.24B,** where it is shown that the altered shape of a scattering curve, due to aggregation or repulsion effects, can provide inaccurate estimates of *I*(0) and R_g, if the data is not treated carefully. For a monodisperse solution of globular macromolecules the Guinier equation is given as:

$$I(s) = I(0)\exp(-\frac{1}{3}R_g^2 s^2) \qquad \text{Equation 4.3}$$

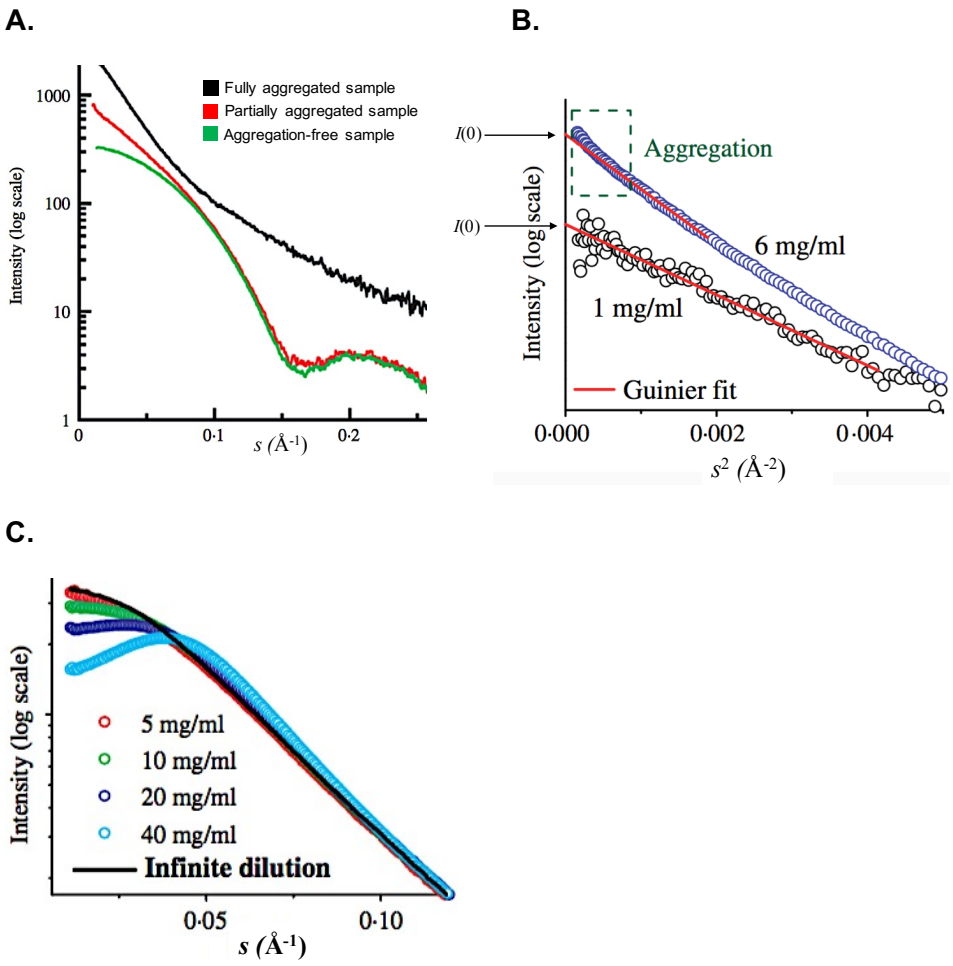


Figure 4.24: Interpretation of plot features in scattering data from SAXS.
Adapted from Putnam et. al. [208] **A.** Typical scattering of aggregates in SAXS. A **fully aggregated or unfolded sample** (in black) lacks defining features in its scattering curve and implies poor sample behaviour. In a **partially aggregated sample** (in red) the lowest resolution scattering is affected the most, showing an upward distortion of the curve closer to *I*(0), the scattering at zero intensity. A notable disagreement between *I*(0) and R_g values calculated from the Guinier plot and from the P(r) function often depends on the degree of aggregation in the sample. An **aggregation-free sample** is indicated by the green plot, where scattering at low *s* is relatively unaffected, and closer to the true *I*(0). **B**. A ln (*Is*) versus s^2 representation of the low *s* region, called a Guinier plot. A non-linear dependence of ln (*Is*) against s^2 indicates aggregation and may result in inaccurate estimates of *I*(0). **C.** Superposition of scaled scattering curves from a range of concentrations show concentration dependent inter particle repulsion for the sample at 40 mg/mL., indicated by the downward smiling curve with a reduction in intensity in the low *s.*

3) Protein folding – Kratky plots

An indication of protein folding/unfolding can be obtained by creating a Kratky plot of $s^2I(s)$ versus s. Globular, folded proteins generally give a prominent bell-shaped curve at low scattering angles, tending towards zero at higher s. On the other hand, unfolded proteins show a continuous increase in $s^2I(s)$ with rising scattering intensity, s. Flexible or partially disordered proteins can also be distinguished from Kratky plots, where they depict intermediate characteristics of folded and unfolded protein plots [209].

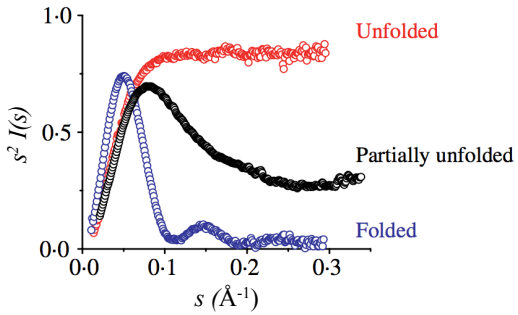


Figure 4.25: Kratky plot representations of the scattering intensities of three types of proteins

4) Pair-distance distribution function and maximum particle diameter

Indirect Fourier transformation of scattering intensities as a function of s, yields the pair-distance distribution function, P(r), a real space representation of the scattering data, providing information on the shape of the biomolecule, and an estimate of the maximum particle diameter, D_{max}. The P(r) is estimated by a program called GNOM [210,211], by producing a reciprocal-space intensity profile ($I(s)$) with the closest match to the experimentally observed intensity profile as determined by the value of χ^2 of fit between estimated P(r) and experimental intensities. $I(0)$ and R_g determined via the indirect Fourier transform method are usually more accurate than those derived from the Guinier approximation, as almost the whole scattering curve is used for estimating P(r), as opposed to only a portion of the low s region for the Guinier plot [200]. An indication of the overall protein shape can also be derived from P(r) profiles. For example, globular proteins show a bell-shaped profile, with a maximum at about $D_{max}/2$, whereas biomolecules with an elongated prolate shape produce profiles

where the maximum is skewed towards the low *s* region, decaying towards zero in the high *s* area. Multi-domain proteins often display curves with multiple maxima, that correspond to intra and inter subunit distances; therefore, as summarised in **Figure 4.26** below. P(r) curves provide valuable information on the actual shape of the protein, together with the maximum particle dimension, D_{max} [200].

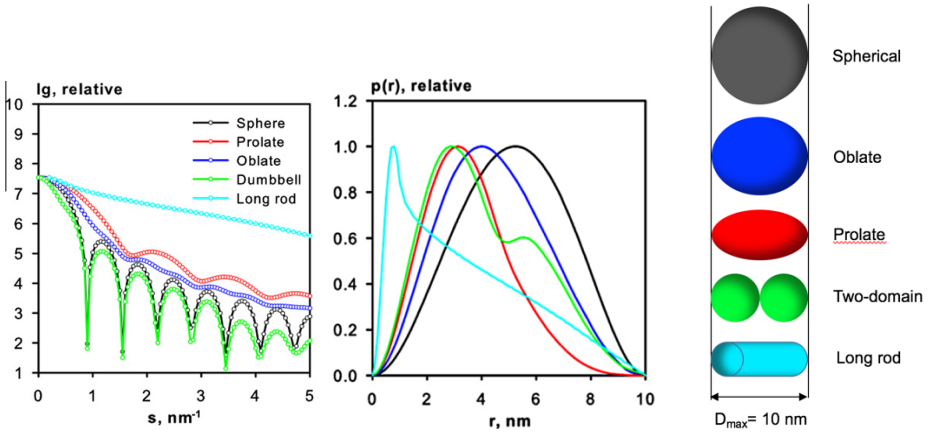


Figure 4.26: Representative scattering intensities and P(r) profiles for typical geometric shapes. Adapted from [200].

4.5.2 *Ab initio* model building.

Using the P(r) distribution it is possible to produce an ab initio bead model of the protein structure with a maximum diameter, defined by the experimentally determined D_{max}, by performing dummy atom model building on a programme called DAMMIF [212]. This procedure is usually repeated, producing 10-20 independent bead models, which are averaged and filtered to produce a model in which low-occupancy atoms are removed according to a defined experimental cut-off volume, and a bead model structure of the most probable distribution of atoms is produced. If a high-resolution structure is available, it could be fit into the SAXS bead model using. Furthermore, the integrity of the SAXS data can be validated by the producing a predicted SAXS curve from the crystal or NMR structure of the protein and fitting it to the experimental SAXS curve. A lower χ^2 value shows stronger agreement between the two curves, and implies a more accurate prediction for the low resolution bead model.

4.6 Method SAXS analysis of BR55

4.6.1 Buffers and Sample preparation

All buffer solutions were made with deionised water and the pH was adjusted with a 10 M stock solution of NaOH. Buffers were filtered with a 45 µm membrane before use. The pH 7.0 buffer consisted of 50 mM sodium phosphate, 0.2 M NaCl and 2.5 mM DTT. Buffers at pH 6.0, 5.0 and 4.0 consisted of 50 mM sodium acetate, 0.2M NaCl and 2.5 mM DTT.

BR55 protein (at 10 mg/mL) was dialysed O/N into each of the four buffers, and 5 mg/mL and 2 mg/mL stock solutions were prepared for SAXS analysis, by diluting the original sample with same batch of dialysis buffer. These concentrations were chosen following a preliminary screening of BR55 polydispersity at 10, 5, 2 and 1 mg/mL using Dynamic Light Scattering (DLS) (results not shown). BR55 samples appeared monodisperse at 2 mg/mL and lower, possible due to lower interparticle aggregation effects, however, samples at 5 mg/mL were also included in the SAXS analysis as they provided an improved signal-to-noise ratio of BR55 scattering.

4.6.2 Data Collection

The scattering curves of BR55 at 2.0 and 5.0 mg/mL and pH 4.0, 5.0, 6.0 or 7.0 were measured. SAXS data was collected on the HATSAXS beamline B21 at Diamond Light Source using a PILATUS 2M detector (DECTRIS, Switzerland). The X-ray beam was tuned to 12.4 keV (1.000 Å in wavelength) and the camera distance was set at 4014mm covering a q-range of 0.0038 to 0.42 $Å^{-1}$. The temperature of the samples was kept at 20^{o}C. All measurements were performed on an automated sample changer (BioSAXS robot), recording 18 x 10 sec frames. To avoid systematic errors, sample and buffer solutions were measured alternately. The voltage of the beamstop diode which was monitoring the intensity of the incident X-ray beam was used to normalize scattering intensities. Stability of SAXS profiles showed that there was no radiation damage to any of the samples.

4.6.3 Data Analysis

The ScÅtter (version 3beta) [213] and ATSAS (version 2.7.1) [214] software were used for SAXS data analysis. Two-dimensionally recorded scattering patterns were converted to a one-dimensional profile by sector integration. Buffer solution scattering was subtracted as background from the sample scattering. Scattering data in the small-angle region was analysed by Guinier's approximation for monodispersive systems, and this provided a value for the radius of gyration (R_g), such that $q \times R_g \leq 1.3$. The reciprocal vector, q, was calibrated by reflection rings from silver behenate. The P(r) and D_{max} was determined by indirect Fourier transformation of scattering profiles using GNOM [210] from the ATSAS package. Molecular structures of the polypeptides were predicted by applying the ab initio structure determination program DAMMIF [212]. Because ab initio simulations do not provide a unique solution for 3D structure, 10 molecular structures of proteins were created. Discrepancies were examined with the χ^2 values defined as:

$$\chi^2 = \Sigma((I_{exp}(q_j) - K_{model}I_{model}(q_j))/\sigma(q_j))^2/(N-1) \qquad \text{Equation 4.4}$$

where N is the number of experimental datum points, K_{model} is a scaling factor and $\sigma(q_j)$ is the statistical error of $I_{exp}(q_j)$ at the scattering vector q_j. Modelling calculations were repeated until χ^2 for each model became close to 1.0. The 10 predicted molecular structures having almost the same overall structures were averaged with DAMAVER, filtered for low occupancy atoms using DAMFILT [215] and then aligned with SUPCOMB [216]. The high resolution BR55 structure was fitted into the low-resolution SAXS envelopes, also using SUPCOMB.

4.7 Results

SAXS experiments of BR55 were performed at concentrations of 5.0 and 2.0 mg/mL and at pH 4.0, 5.0, 6.0, 7.0, to investigate pH dependent conformational changes in the protein. BR55 samples were prepared at 5 mg/mL to provide scattering data for BR55 with higher signal/noise levels, which could be merged with the low concentration BR55 scattering data if the quality of the latter was insufficient. However, at progressively lower pHs, 5.0 and 4.0, aggregation effects appeared to increase at 5 mg/mL.

4.7.1 The Low-resolution structure of full-length BR55 at pH 7.0

SAXS data analysis and ab initio modelling for BR55 at pH 7.0 is presented first as a standard reference for the low-resolution structure of full-length BR55. Buffer subtracted data of BR55 is displayed as a plot of log (*I*) versus *s* ($Å^{-1}$) in **Figure 4.27A** below, and shows evidence of minor aggregation at the lowest scattering angles (the low *s* region) for both 5.0 and 2.0 mg/mL samples, in red and blue, respectively. Converting a portion of this plot, from *s*=0 to *s*=0.2, into a Guinier plot of log (*I*) versus s^2 make the aggregation effects more apparent (**Figure 4.27B**), owing to the appearance of an upward smiling curve, surprisingly more pronounced for the 2.0 mg/mL sample (in blue) – perhaps the aggregation effect was more pronounced due to a lower S/N ratio. Hence, rather than merging low *q* data from the 2.0 mg/mL sample with that from the 5.0 mg/mL data, approximately 0.04 $Å^{-1}$ was trimmed from the low *s* region of both samples to minimise errors in the estimation of *I*(0) and R_g, which was performed manually on the ScÅtter 3 beta program [213]. An R_g of 27.82 Å and 26.98 Å was estimated for BR55 at 5.0 and 2.0 mg/mL, respectively. *I*(0)/C for 5.0 and 2.0 mg/mL samples was 7.0 x 10^{-3}, implying no concentration dependence on the Guinier estimation of R_g. All these results are presented in **Table 4.7** below. Kratky plots of the untrimmed scattering profiles reveal bell-shaped plots with a gradually rising tail at higher *s*, suggesting that BR55 is folded, with globular

characteristics but with some flexible regions in the protein (**Figure 4.27C**). A P(r) function was generated for both samples by the GNOM program [210], providing bell-shaped curves with a single maxima that were skewed to left, prior to $D_{max}/2$, and tailing off towards the right end. An identical profile was obtained for samples at both concentrations, suggesting a monomeric particle, with a molecular shape that is prolate, with a D_{max} of 102 Å (**Figure 4.27D**). The D_{max} calculated for BR55 is approximately double the particle length of BR55 from its crystal structure (50$\pm$2 Å), which was estimated in PyMol; whereas the crystal structure represents 70% of the entire protein sequence. This implies that the proteolysed appendages of BR55, which were excluded from the crystal structure must be relatively extended and flexible. R_g estimated by the P(r) plot was 28.55 Å and 28.06 Å for 5.0 and 2.0 mg/mL samples, respectively, which is in close agreement with R_g estimates from the Guinier approximation. Inspection of the results in **Table 4.7** also reveal good agreement of the *I*(0) values estimated from Guinier and P(r) analyses.

Due to a higher S/N ratio in scattering intensity, and a lack of concentration dependent variation in R_g and *I*(0) estimates between BR55 samples at 5.0 and 2.0 mg/mL, the P(r) for BR55 at 5.0 mg/mL was used to generate an Ab initio bead model in the online DAMMIF [212] server, producing ten unique structure predictions that were averaged in DAMAVER [215]. The average bead model was filtered using DAMFILT [215], omitting low occupancy atoms to yield a final, more compact model with a total volume of 84,398 $Å^3$. Further information on the averaging and filtering results can be found in **Table 4.7**. A surface envelope representation of the filtered BR55 ab initio model fitted with the high-resolution crystal structure of BR55 is given in **Figure 4.28,** and the features observed in the SAXS envelope is in good agreement with many of the structural deductions made from the manipulated SAXS scattering plots in **Figure 4.27.** Corresponding to the Kratky plot and P(r) profiles, the bead model appears prolate in shape, with partial flexibility implied in the extremities of the model. The position of the relatively ordered BR55 core in the bead model is deduced from

SUPCOMB fitting of the high resolution BR55 atomic structure into the SAXS envelope, with NSD of 3.06 $\pm$ 0.18, which is a reasonable fit (**Figure 4.28A.**).

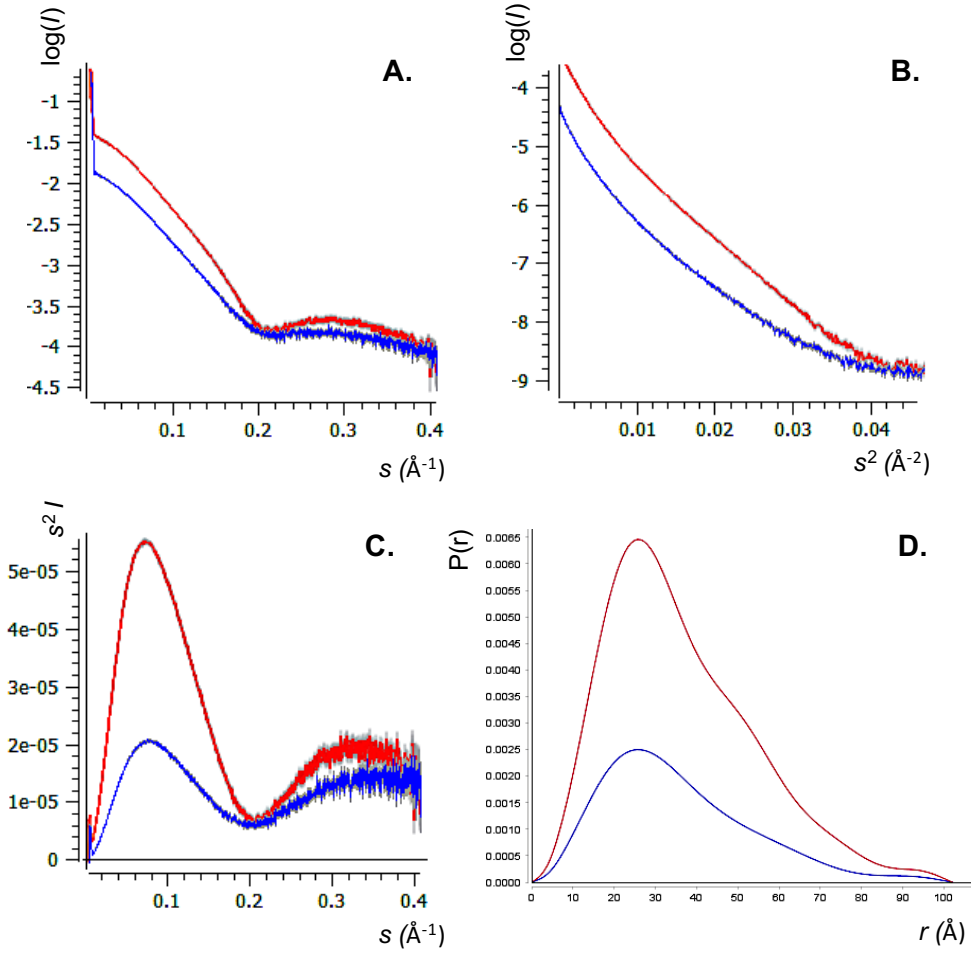


Figure 4.27: SAXS data analysis for BR55 at pH 7.0, at concentrations of 5.0 mg/mL (red) and 2.0 mg/mL (blue).

A. Scattering plot of log (*I*) against momentum transfer, *s,* displays inter particle aggregation effects at both concentrations in the low *s* region; approximately 0.04 Å of the initial lows region was trimmed for Guinier analysis. **B.** Guinier plot up to *s*=0.2 also reveals aggregation effects at lower *s*. **C.** Kratky plots of data for each BR55 concentration produced curves characteristic of folded, single domain proteins with partial flexibility. **D**. P(r) plot displaying a profile typical for a monomeric, single domain prolate shaped particle with D_{max} at 102 Å estimated for both concentrations.

Verification of the bead model against experimental scattering data was performed using CRYSOL [217], by generating a theoretical scattering curve for the filtered BR55 bead model, and comparing it to the experimental SAXS curve for BR55 at pH 7.0 and 5 mg/mL, as depicted in **Figure 4.28B**. A fit of χ^2 = 130.51 was obtained, which is not surprising, given that the P(r), and subsequently the Ab initio models were

derived from truncated scattering data, obtained by trimming off the low *s* Guinier region, and a major portion of the high *s* region beyond s = 0.26, as the sample quality was not ideal owing to aggregation. The best fitting is observed in the area of the experimental scattering curve that was included in subsequent data processing. Fitting of the theoretical scattering curve generated from high resolution BR55 structure to the experimental scattering curve afforded a considerably poorer fit of χ^2 = 880.74, which is expected as the high-resolution structure of $BR55_{262\text{-}566}$ represents only the ordered, globular domain of full-length BR55, accounting for 70% of the overall structure (data not shown). An R_g of 21.69 Å was estimated for $BR55_{262\text{-}571}$.

As seen in **Figure 4.28,** the upper and lower loops of BR55 appear outside the filtered cyan envelope, but within the averaged white envelope, implying that they occur within the low occupancy region of the SAXS envelope, thus confirming their flexibility, relative to the rest of the ordered core. Some T1 loops in the lower half of the $BR55_{262\text{-}571}$ model also appear within the low occupancy region, possibly meaning they are slightly more disordered than the rest of the molecule.

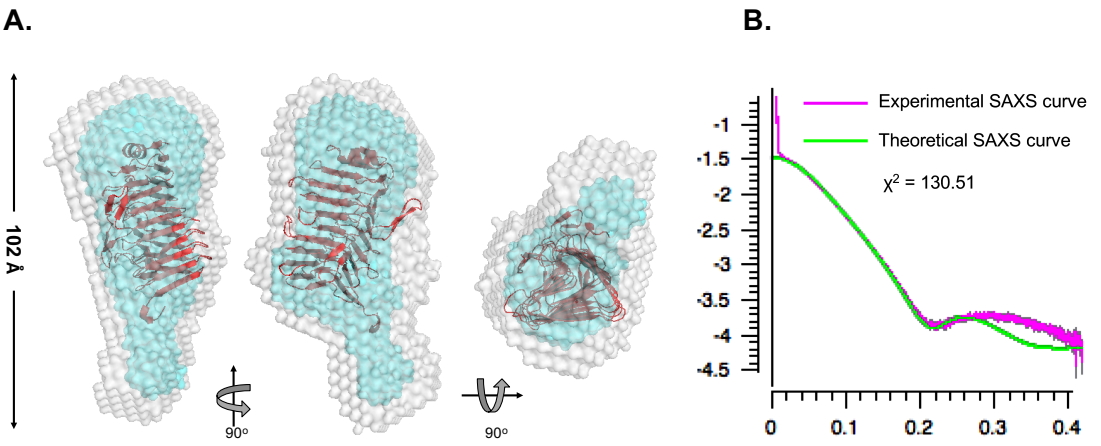


Figure 4.28: Low resolution SAXS model for BR55 at pH 7.0
A. Low resolution SAXS envelopes of BR55, collected at pH 7.0 and 5 mg/mL, fitted with the high-resolution model of $BR55_{262\text{-}571}$. The external envelope in white is the averaged bead model produced by DAMAVER, and the interior envelope in cyan is the filtered bead model after removal of low occupancy dummy atoms by DAMFILT, and the final surface model of BR55. **B.** Experimental SAXS curve of BR55 at pH 7.0 and 5 mg/mL overlaid with theoretical scattering curve calculated from the filtered model of BR55 using CRYSOL [217].

4.7.2 Low resolution SAXS model of full-length BR55 at pH 6.0, 5.0 and 4.0

Scattering data of BR55 in pH 6.0, 5.0 and 4.0 buffers, at 5.0 and 2.0 mg/mL were also collected to evaluate any pH driven conformation changes in the molecule. The SAXS envelopes for these are provided in **Figure 4.29** below, and have been fitted to the high resolution structure of $BR55_{262\text{-}571}$ using SUPCOMB; the statistics for which are provided in **Table 4.7**, and are comparable to the NSD values for $BR55_{262\text{-}571}$ fitting to the BR55 SAXS envelope obtained at pH 7.0. Scattering curves at 2.0 mg/mL for the protein at pH 5.0 and 4.0 was used for subsequent analysis and modelling, as the scattering curves at 5.0 mg/mL showed increased signs of aggregation, compared to BR55 at pH 7.0 and 6.0 at the same concentration. No major conformational changes can be observed in these SAXS envelopes compared to that of BR55 at pH 7.0, in **Figure 4.28**. The envelopes at all four pHs are approximately prolate, with a tail leading out to form the presumably disordered C-terminal of the protein, comprising of the missing 97 amino acids from 572 to 667. The putative binding loops and T1 turns in the lower half of the $BR55_{262\text{-}571}$ model are all found beyond the threshold of the filtered bead models, indicating disorder. The accompanying SAXS scattering curves, Kratky plots and P(r) curves for BR55 at pH 6.0, 5.0 and 4.0 are provided in APPENDIX 6.

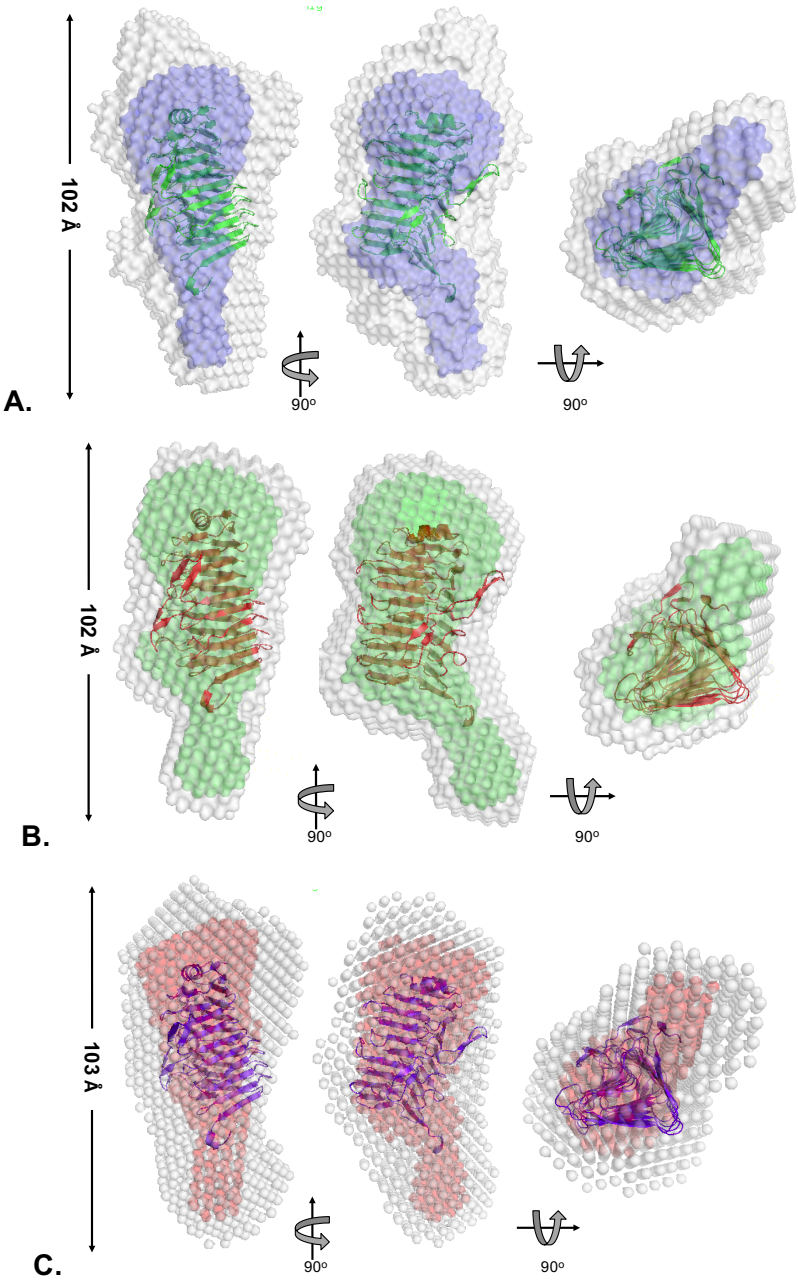


Figure 4.29: Low resolution SAXS models for BR55 at pH 6.0, 5.0 and 4.0

SAXS bead models generated from scattering curves of BR55 at **A.** pH 6.0 and 5.0 mg/mL (in blue), **B.** pH 5.0 and 2 mg/mL (in green), and **C.** pH 4.0 and 2.0 mg/mL (in red). The outer white envelopes are DAMAVER averaged bead models and the interior coloured models are DAMFILT filtered surface models.

Table 4.7: SAXS data processing and model building statistics for BR55 at pH 7.0, 6.0, 5.0 and 4.0, and 2.0 and 5.0 mg/mL

DAMMIF parameters are provided for scattering curves that were processed to produce dummy atom models; the χ^2 value is a measure of agreement between the dummy atom model and the experimental scattering curve, with lower χ^2 values denoting a better fit. The normalised spatial discrepancy (NSD) is a measure of the quantitative similarity between sets of 3D points, and was computed by SUPCOMB while averaging the 10 Ab inito models generated by DAMMIF; an NSD below 1 for superimposed objects imply that they tend towards similarity. NSDs above 1 for fitting of the high resolution $BR55_{262\text{-}571}$ model into the SAXS envelope are obtained because the partial crystal structure omits 30% of the protein molecule from its structure, implying significant differences in overall structural properties. The volume is the overall envelope volume computed from the final filtered SAXS model.

Structural parameters	**BR55 pH7.0**		**BR55 pH6.0**		**BR55 pH 5.0**		**BR55 pH4.0**	
	5 mg/mL	**2 mg/mL**	**5 mg/mL**	**2 mg/mL**	**5 mg/mL**	**2 mg/mL**	**5 mg/mL**	**2 mg/mL**
I(0) (cm^{-1}) from Guinier	0.033	0.013	0.034	0.012	0.037	0.013	0.043	0.012
R_g (Å) from Guinier	28.55	28.06	28.7	27.9	29.4	28.4	32.0	27.4
I(0) (cm^{-1}) from P(r)	0.035	0.014	0.037	0.014	0.041	0.015	0.038	0.010
R_g (Å) from P(r)	27.82	26.98	27.4	27.1	28.8	28.1	31.2	28.6
D_{max} (Å) from P(r)	102	102	102	101	102	102	112	103
DAMMIF *Ab initio* modelling								
χ^2	3.39		4.60			3.48		1.87
NSD	0.63±0.047		0.81±0.013			0.63±0.083		0.85±0.042
Discarded	None		None			None		None
Volume $Å^3$ (from DAMFILT)	84389		77302			75744		73740
$BR55_{262\text{-}571}$ Fitting (SUPCOMB)								
NSD	3.75±0.31		3.54±0.28			3.98±0.34		3.65±0.29

4.8 Discussion: Information from the low resolution SAXS model of BR55

4.8.1 Limitations of the SAXS Data

Signs of aggregation were consistently observed in all samples, but were particularly obvious for samples at lower pH and higher concentrations, such as BR55 at pH 4.0 or 5.0 at 5.0 mg/mL (see scattering plots in APPENDIX 6). BR55 has a tendency to oligomerise, as evidenced by SEC chromatograms of BR55 and mutants. However, DTT addition seems to significantly improve the problem, and this was included in all SAXS samples at 2.5 mM, minimizing the effects of oligomerization and X-ray damage. BR55 aggregation and precipitation at lower pH were reported often by collaborators at the IFR during their biochemical studies. Samples were centrifuged at high speed before dispensing them in the SAXS sample holders, to remove any solids or fine particles.

Oligomerisation of BR55 was assessed briefly with equilibrium sedimentation analytical ultracentrifugation (AUC), with 1, 2.5 and 5 mg/mL BR55 at pH 7.0, and 1 and 5 mg/mL BR55 at pH4 over a period of two days. The results indicated a monomeric protein with MW of approximately 46-47 kDa, which is very close to the actual MW of BR55, at 48 kDa. There was no obvious evidence of dimer or trimer formation, or any significant reduction in peak size during the experiment implying that the aggregation noted during SAXS was probably due to a tiny proportion of high MW oligomers (data not shown). Higher MW oligomers can be eliminated from an AUC sample due to sedimentation at ultra-high speeds employed during the analysis, however, SAXS is extremely sensitive to even the tiniest form of aggregation and this may affect the data quality.

This was indeed the case for most BR55 samples, where an upward curve was noted in the scattering curves at the low *s* regions (see **Figure 4.27** and APPENDIX 6) and

the data beyond q = 0.25 was eliminated during *Ab inito* model building due to poor sample quality, very likely due to unspecific aggregation. It is likely that the truncation of the latter part of the scattering curve may have omitted some information on the solution structure of BR55; the effects of this can be observed in **Figure 4.28B**, where the theoretical SAXS curve derived from the processed bead model fits poorly with the experimental SAXS curve. Another indication of non-ideal data are inconsistencies in the volume of the SAXS envelope, which is given in **Table 4.7**. Similarly, inconsistencies were also observed in the Porod volume estimated in ScÅtter, which are both likely to be the result of aggregation artifacts. The bead model volume and Porod volume can be used to estimate the MW of a protein, and a close match to the actual MW is a sign of a good data set without bias [218]. Such inconsistencies made it a challenge to accurately estimate the MW of BR55 measured by SAXS. In cases of a non-ideal sample type, measuring the scattering of an external protein standard, such as BSA, at various concentrations would have been a reliable comparator for estimating BR55 MW, and therefore the accuracy of the bead model. In future experiments this must be included.

In short, the data quality is not the best, but was treated with care to extract as much useful information as possible. Due to a generally good agreement between the R_g and $I(0)$ values estimated from Guinier and P(r) analysis, and the consistency of the SAXS envelopes, the data was deemed as a reasonable representation of BR55 in solution, despite its limitations. The information obtained from SAXS is as follows:

1) No major pH dependent conformational changes were apparent

2) Kratky plots and P(r) curves for all samples were suggestive of a folded, prolate protein with some degree of disorder and flexibility

3) A good visual fit of the high-resolution $BR55_{262-571}$ model into the SAXS envelopes (which produced an NSD >1 due to it being a partial match) indicated that the N-terminal and C-terminal of BR55 may be highly disordered.

4.8.2 Intrinsically disordered domains in BR55

The PSIPRED server was used to assess secondary structure features in BR55, including any disordered domains. A high degree of disorder was noted at the N- and C-terminals of BR55, from about 1 to 40 and 425 to 450 residues, respectively, which is represented in **Figure 4.30** below (any region above a confidence score of 0.5 is deemed to be disordered). This roughly corresponds to the region which was removed by proteolysis prior to crystallization. These results, together with the SAXS data is an indication that the appendages of BR55 beyond the solenoid core are likely to be intrinsically disordered.

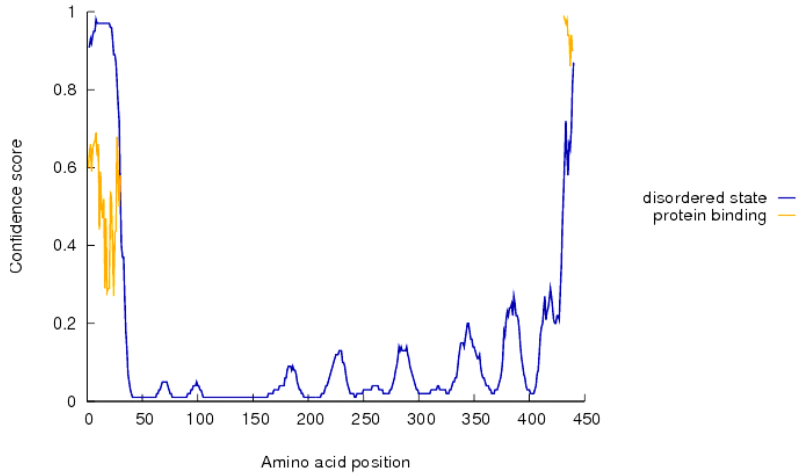


Figure 4.30: Intrinsic disorder profile for full length BR55
This was computed by the Disorder Prediction Server, DISOPRED, on the PSIPRED [13] online server. Amino acid stretches >0.5 confidence score are predicted to be intrinsically disordered. Possible protein binding sites within the disordered regions, also predicted by the server, are shown in orange. The regions of $BR55_{262\text{-}571}$, that were solved by crystallography correspond to residues 34 to 348. The first 33 and last 97 residues of BR55 are missing from the $BR55_{262\text{-}571}$ crystal structure, and as this profile shows, this corresponds to most of the highly disordered portion of BR55.

An observation of the SAXS envelopes fitted with the crystal structure of $BR55_{262\text{-}571}$ show that the envelopes beyond the core solenoid occupied region, especially the tail at the C-terminal of $BR55_{262\text{-}571}$, appears too small to accommodate 97 missing residues. Such an observation is not unlikely if a certain domain is disordered, as a highly flexible, unconstrained domain in solution has numerous configurations that

can be averaged out by SAXS [201]. An ensemble approach towards constructing the possible conformations of a flexible domain from the experimental SAXS profile can offer some insights into the intrinsically disordered domain. The Ensemble Optimisation Method (EOM) [219] from ATSAS online was used to model the disordered appendages of BR55 from the SAXS scattering profile for BR55 at pH 7.0 and 5.0 mg/mL, and from the primary sequence of BR55. EOM also allows the input of high resolution atomic coordinates for any domains that have been structurally characterised, and this would help in orienting the disordered domains more accurately. Hence, the structural model of $BR55_{262-571}$ was provided to the algorithm. The resulting output file provided a representation of one of the common configurations expected of full length BR55, and this is displayed in **Figure 4.31A** below. Using SUPCOMB [216] the ensemble was fitted into the SAXS envelope derived from the scattering curve of BR55 at pH 7.0 and 5.0 mg/mL, giving an NSD of 3.00±0.32 (**Figure 4.31B**).

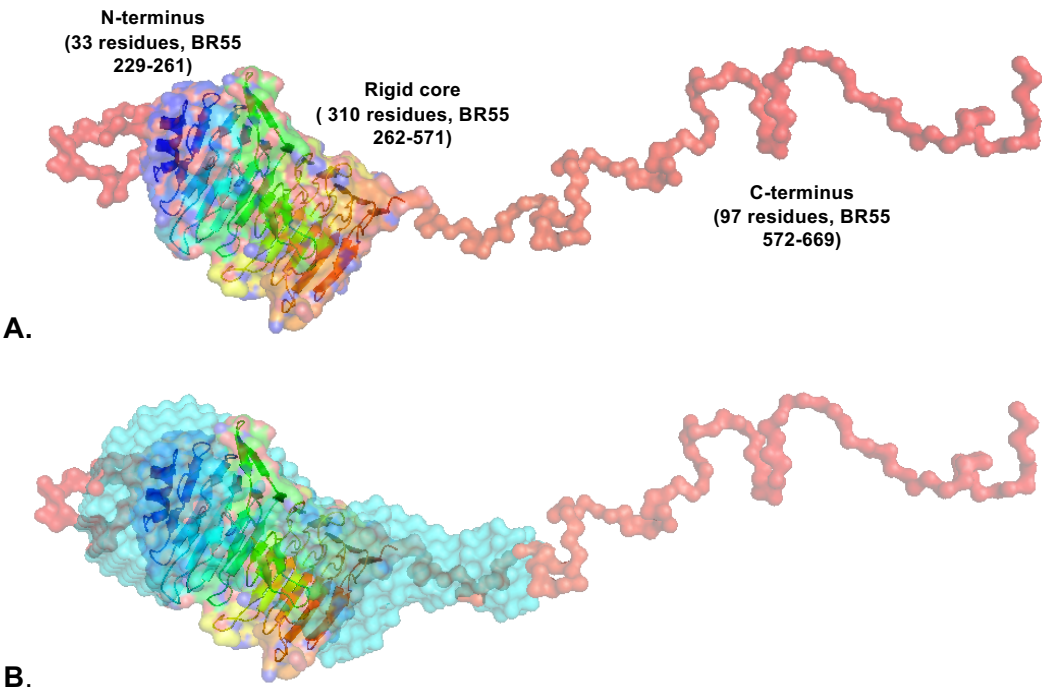


Figure 4.31: An ensemble of the model for full length BR55

A. An ensemble of the BR55 model, showing the disordered regions of BR55 as surface models. **B.** The BR55 ensemble fitted into the filtered SAXS bead model envelope of BR55 at pH 7.0 using SUPCOMB, where an NSD of 3.00±0.32 was obtained.

4.9 General Discussion: Proposed biological significance of BR55 and BR78 structure

As explained earlier in section 4.4.4, BR55 and BR78 were deduced to belong to the 2.160.20.10 CATH superfamily of single-stranded, right-handed β-helix, PelC-like proteins. **Table 4.8** below lists 33 proteins from this CATH superfamily sorted by their functions, and these include 9 pectate lyases, 2 pectin lyases, 4 pectin-methyl esterases, 5 polygalacturonases, 8 lyases or hydrolases with other plant glycan substrates, 2 mannuronan epimerases involved in alginate synthesis (an EPS aiding biofilm formation in pathogenic Pseudomonads [220]*)*, 2 adhesin associated proteins, and one each of an IgE-binding pollen allergen, and a Bacillus phage tail spike protein. Apart from PfbA, FMA-TpsA and Jun a 1 (see **Table 4.8**), the remaining 30 structurally characterised PelC-like proteins have carbohydrate substrates; of them, 20 have pectic substrates, 8 degrade other plant or algae based polysaccharides, and 2 have pathogenic biofilm EPS substrates. 4 out of 33 (PfbA, FMA-TpsA, AlgE4 and AlgG) have functions associated with adhesion and biofilm formation (**Table 4.8**).

Plant cell walls are composed of a complex network of polysaccharides, mainly cellulose, hemicelluloses and pectic sugars. Pectins are a heterogenous group of polysaccharides consisting of galacturonic acid as the fundamental monosaccharide, and can be divided into three types: 1) linear homogalacturonan pectin (also called pectate or polygalacturonate – pGal) and consisting only of galacturonic acid residues, 2) branched rhamnogalacturonan type-I and 3) branched rhamno-galacturonan type-II pectins (also called 'hairy' pectins) which consist of galacturonate and rhamnose residues linked together in different repeating patterns, and with varying side chain substitution [170, 221]. It is therefore, not surprising that most PelC-like proteins characterised to date originate from plant pathogens, such as the *Aspergillus* fungi and the enteric bacteria *Erwinia chrysanthemi,* for which pectic cell wall degradation plays a crucial role in establishing host infection.

Table 4.8: PelC-like Proteins from CATH superfamily 2.160.20.10 sorted by biological function.

Protein Name	Species	Substrate	Function	PDB ID (s)	Reference
Pectate Lyase					
PelC	*Erwinia chrysanthemi*	Polygalacturonic acid (PGA), also called polygalacturonate, pectate, pectic acid or GalA	Cleaves the a-1,4 linkage in the smooth, unbranched, methyl ester-free homogalacturonan region of plant cell wall pectin; often Ca^{2+} dependent	1AIR, 1PLU, 1O88, 1O8D, 1O8E, 1O8F, 1O8G, 1O8H, 1O8I, 1O8J, 1O8K, 1O8L, 1O8M, 2EWE, 2PEC	[169], [222], [193], [223]
BsPel	*Bacillus subtilis*			1BN8, 2NZM, 2O04, 2O0V, 2O0W, 2O17, 2O1D, 3KRG	[178], [224]
AcPel	*Acidovorax avanae sb. citrulli*			4HWV	[225]
Pel9A	*Erwinia chrysanthemi*			1RU4	[226]
PelI catalytic domain	*Erwinia chrysanthemi*			3B4N, 3B8Y, 3B90	[227]
TMPelA	*Thermotoga maritima*			3ZSC	[173]
Pel-15	*Bacillus sp.* KSM-P15			1EE6	[228]
CB-PL3	*Caldicellulosiruptor bescii*			3T9G	[229]
Bsp165PelA	*Bacillus sp.* N16-5		Alkaliphilic pectate lyase; Ca^{2+} independent	3VMV, 3VMW	[230]
Pectin Lyase					
PLB	*Aspergillus niger*	Highly esterified pectins	Pectinase; cleaves the a-1,4-glycosidic bond in methyl esterfied pectates, which are called pectins	1QCX	[231]
PLA	*Aspergillus niger*			1IDJ, 1IDK	[176]
Pectin methyl-esterase					
PemA	*Erwinia chrysanthemi*	O6 methyl-esterified $_D$-galacturonate residues	Aspartic esterases; removes methyl groups from methyl-esterified pectate in plant cell walls	1QJV, 2NSP, 2NST, 2NT6, 2NT9, 2NTB, 2NTP, 2NTQ	[232], [233]
YeCE8, PME	*Yersinia enterocolitica*			3UW0	[234]
DC-PME	*Daucus carota*			1GQ8	[172]
LE-PME	*Lycopersicon esculentum*			1XG2	[171]

Protein Name	Species	Substrate	Function	PDB ID (s)	Reference
Homo and Hetero Polygalacturonases					
Tm ExoPG	*Thermotoga maritima*	Exo of non-methylated PGA	Exopolygalacturonase that cleaves non-methylated saturated galacturonate at the non-reducing end	3JUR	[185]
PehA	*Erwinia carotovora*	Endo of the a-1,4-glycosidic linkage of PGA	Endo-polygalacturonase; hydrolyses the a-1,4 linkage in the homogalacturonan region of plant cell wall pectin	1BHE	[181]
EndoPG1	*Stereum purpureum*			1K5C, 1KCC, 1KCD	[235]
RGase A	*Aspergillus aculeatus*	Rhamnogalacturan, composed of alternating rhamnose and GalA residues (RG-I)	Rhamnogalacturanase, hydrolysing glycosidic bonds within RG-I	1RMG	[236]
XghA	*Aspergillus tubingensis*	Xylosylated PGA	Endo-xylogalacturonase; glycosidic bond hydrolysis of two β-xylose substituted PGA residues in xylosylated pectate	4C2L	[237]
Other glycan lyases and hydrolases					
Chondroitinase B	*Flavobacterium heparinum*	Glycosaminoglycan: Dermatan sulphate	Glycosaminoglycan lyase; cleaves the beta-1,4 linkage of dermatan sulfate in a random manner	1DBG, 1DBO	[238]
	Pedobacter heparinus			1OFL, 1OFM	[239]
Dex49A	*Talaromyces minioluteus*	Dextran	Dextranase; hydrolysis of the α-1,6-glycosidic linkage in dextran polymers	1OGM, 1OGO	[240]
IPU	*Aspergillus niger ATCC9642*	Pullulan	Isopullulanase; hydrolysis of the α-1,4-glucosidic linkages of pullulan to produce isopanose	1WMR, 1X0C, 2Z8G, 3WWG	[241]

Protein Name	Species	Substrate	Function	PDB ID (s)	Reference
XylC	*Thermoanaerobacterium saccharolyticum*	Xylanose	Xylan-1,4-beta xylosidase	3VST, 3VSU, 3VSV	[242]
Lam55A	*Phanerochaete chrysosporium*	beta-1,3-glucan oligosaccharides	Hydrolyzes -1,3-glucans in the exo-mode, producing gentiobiose and glucose from -1,3/1,6-glucans	3EQN, 3EQO	[243]
Iota-Carrageenase	*Alteromonas fortis*	Sulphated 1,3-a-1,4-b-galactans (carrageen) from red algae cell walls	Digestion of iota-carrageenan in red algae cell wall	1H80, 1KTW, 3LMW	[244], [245], [246]
IFTase	*Bacillus sp. snu-7*	b-2,1-linked inulin (a fructan)	Inulin fructotransfarase; catalyses depolymerisation of inulin by successively removing terminal difructose units	2INU, 2INV	[187]
Glycan epimerases					
AlgE4	*Azetobacter vinelandii*	Beta-1,4-$_D$-Mannuronanate	Mannuronan C5 epimerase, in the production of the EPS alginate for biofilm fomation	2PYG, 2PYH	[189]
AlgG	*Pseudomonas syringae*			4NK6, 4NK8	[247]
Miscellaneous functions, including adhesins, allergen and tail spike protein					
PfbA	*Streptococcus pneumoniae*	Human plasminogen and fibronectin in the host ECM	Plasmin and fibronectin binding protein	3ZPP	[248], [249]
FMA-TpsA	*Bordetella pertussis*	TpsB transporter	Filamentous hemagglutinin adhesion section domain of the two-partner secretion (TPS) pathway	1RWR	[250]
Jun a 1	*Juniperus ashei*	Proposed to be IgE	Cedar allergen from pollen	1PXZ	[251]
phi29	*Bacillus phage phi29*	*B. subtilis* cell wall glucosylated teichoic acid	Tail spike protein	3GQ7, 3GQ8, 3GQ9, 3GQA, 3GQH, 3GQK, 3SUC	[252]

This thesis reports the first instance of a PelC structure-like protein originating from a mammalian lactobacillus commensal. This does not imply that the putative binding ligand for BR55 and BR78 is pGal, but rather suggests that its binding ligand is very likely to be a carbohydrate moiety. Apart from structural homology, aromatic residues within the putative binding loops further consolidate the hypothesis that BR55 and BR78 may have a carbohydrate substrate [195, 253].

A detailed survey by Hudson et. al. [195] on binding site patterns in carbohydrate binding protein structures across the PDB revealed that non-covalently bound sugars make more plentiful and specific contacts with protein side chains than covalently attached carbohydrates, like glycoprotein substituents. In the binding sites of these non-covalently bound sugars, polar residues were noted to occur at frequencies expected at random, aliphatic hydrophobic residues were underrepresented and electron-dense aromatic side chains were highly favoured, especially Trp, showing unusually high representation, with respect to its normal frequency of occurrence in a typical protein sequence. The most prominent interaction between aromatic residues and sugars, was the CH-π interaction, where electropositive saccharide C-H bonds would engage with electron-rich π-orbitals on aromatic side chains. Aromatic residues were found to participate in CH-π interactions to varying extents, in the order Trp>Tyr>Phe>His, which interestingly correlates to the decreasing electronegativity of the four sidechains, with Trp being the most electron-dense, and His being least electronegative [195].

For BR55 and BR78, the only two Trp residues (W375, W450 and W373, W453, respectively) to occur in the entire binding region are structurally conserved, and found underneath the lower loop, corresponding to a 100% Trp representation in the putative binding cleft of BR55 and BR78. The second most abundant aromatic residue to be found in the putative binding regions is Tyr, with several of them being conserved, as expounded in section 4.4.6.1 above. Whilst structural homology analyses point towards the lower loop region being a putative binding site in BR55 and BR78, the profusion of aromatic residues, mostly Tyr, in the upper loop area may

warrant a further investigation of their involvement (if any) in adhesion.

Two aspects of the new discovered *L. reuteri* SRRP that make it different from the relatively well known MUB proteins [102, 104] (see section 2.3) are:

1) The indication that SRRP-type adhesins are specific to rodent and porcine isolates of *L. reuteri*, (discussed in section 2.1.3), whereas MUB proteins are expressed in varying numbers across all *L. reuteri* host strains.

2) *L. reuteri* biofilms form on the non-secretory epithelial surface of pars esophageal and forestomach regions of the pig and rodent digestive tract, respectively (section 2.1.3). These regions link the esophagus to the glandular stomach, and function as temporary food storage sites. As elaborated in section 2.2.4, the fact that the forestomach epithelium in mice is not covered by mucus could imply that SRRP may mediate *L. reuteri* attachment to another type of substrate within the proximal stomach, such as receptors on epithelial cells, or a certain substance in the ECM. Suggestions for possible binding ligands of *L. reuteri* SRRPs, based on the structural information obtained from partial crystal structures BR55 and BR78 are proposed.

4.9.1 Possible binding ligands of *L. reuteri* SRRPs

Interaction with the extracellular matrix (ECM) in maintaining biofilms: The ECM is a structurally complex, non-cellular component surrounding and supporting cells within all tissues and organs. Depending on species and tissue type, ECM can have their own unique composition. It also participates in instigating crucial biological processes, such as tissue morphogenesis, differentiation and homeostasis [254]. The two types of macromolecules in ECM are fibrous proteins, like collagen, fibronectin and laminin, and glycoproteins [254, 255]. Examples of ECM binding adhesins from Lactobacilli, such as CnBP, MapA and Lps, have been discussed in section 2.3. It is quite likely that *L. reuteri* SRRPs recognise and bind some macromolecular component(s) of the ECM. Based on deductions from the structural analysis of BR55 and BR78, it seems more likely to bind to a glycan belonging to a glycoprotein, however, its potential binding to a fibrous protein must not be ruled out unless tested.

Interaction with exopolysaccharide (EPS) in maintaining biofilms: EPS are produced by a wide range of bacteria, including lactobacilli, where they play a role in biofilm formation and cell aggregation (section 2.2.2). In *L. reuteri,* sucrose is an important nutrient for EPS production [256]. A sourdough strain of *L. reuteri,* TMW1.106, forms large amounts of a high MW glucan from sucrose, and expresses a glucosyltransferase (*gtfA)* and an inulosucrase (*inu*). Competition experiments in germ-free mice with isogenic mutants of *gtfA* and *inu* showed impaired ecological performance in the presence of the wild-type strain [84]; in the absence of competition the *gtfA* mutants exhibited delayed colonisation of the murine GIT, therefore, demonstrating a role of EPS in biofilm maintainence. *L. reuteri* 100-23 produces EPS predominantly composed of levan, which is a β-$_D$-2,6-linked fructan, however, it was not shown to be crucial for maintaining biofilm integrity, to the extent exhibited by the former *L. reuteri* strain [257]. Although this may contradict the statement presented earlier in section 2.2.4, that SRRPs may have a vital role in initial epithelial adherence, leading to biofilm formation, (whereas EPS production starts after initial adherence) it may still be prudent to investigate any synergistic interaction of SRRPs with glycans in the EPS towards maintaining cell-to-cell adhesion and biofilm formation.

Interaction with ingested food material: selected commensal bacteria in mammals have been known to participate in the digestion of ingested plant material that cannot be treated by host enzymes [258]. For example, a recent metagenomics study of the human gut microbiome revealed a large range of genes encoding hydrolases and lyses that could degrade plant polysaccharides [259]. Ruminal flora in the first stomach of ruminants mediate the initial digestion of plant polysaccharides via the involvement of plant cell wall degrading enzymes, including pectinases for pectin digestion [260]. It is probable that that the BR of SRRPs may have a dual function in the *L. reuteri*-host symbiotic relationship: that of aiding initial plant digestion in the proximal GIT, the products of which may be incorporated by *L. reuteri*, facilitating its adhesion. However, this hypothesis can only be verified by testing for BR55 or BR78 glycolytic activity in the presence of a suitable sugar substrate (yet to be identified).

BR55 and BR78 are likely to prefer hydrophobic esterified sugars, over highly charged sugars. This proposal stems from a comparison of the conserved binding clefts of *A. niger* pectin lyases, PelA [176] and PelB [231] with pectate lyases from *B.s subtilis* (BsPel) [178] and *E. chrysanthemi* (PelC) [223], respectively. Pectin lyases act upon heavily esterified pectins, whereas pectate lyases degrade negatively charged GalpA (**Table 4.8**). PelA and PelB's suggested catalytic sites have a dense network of aromatic residues compared to those of BsPel and PelC which have highly charged catalytic sites, in which Ca^{2+} binding was a requirement for effective pectate lyase activity. PelA and PelB did not show a need for Ca^{2+} in catalysis, and their binding sites were predominated by an extensive distribution of Trp/His and Trp/Tyr residues, respectively [176, 231]. While the binding clefts of BR55 and BR78 are different to those of the Pel proteins, the profusion of aromatic, rather than polar residues in the binding clefts is indicative of preferential binding to increasingly hydrophobic sugar substrates.

4.9.2 Possible binding modes of *L. reuteri* SRRPs

SRRP binding may be pH dependent: pH in the rodent forestomach region ranges between 3.8 to 5.1, depending on the occurrence or absence of feeding; generally the pH slightly rises after feeding [199]. The pH gradient in a healthy porcine stomach, from esophagus to pylorus has not yet been established, however, it is known that pH is relatively low at the esophageal terminus of the stomach [261]. To facilitate selective binding in the proximal stomach, pH controlled SRRP adhesion could be a useful mechanism. Examples of such a mechanism are reported in the literature, both for structurally characterised SRRPs and Pel-C type proteins. An example of the former is the pH dependent conformational change of the two-domain binding region of S. parasanguinus Fap1 SRRP, which favours binding at an acidic pH (see section 2.4.3.1) [78]. For PelC-like proteins, crystal structures of the pectin lyase, PelA, at pH 6.5 and 8.0, showed a conformational change, which was attributed to be pH driven. The structure at pH 6.5 was deemed the active conformation of the enzyme [176].

SAXS analysis of BR55 from pH 7.0 to 4.0, (see section 4.7) revealed no obvious pH driven conformation changes in overall structure. pH dependent structural changes in BR55 could be best tested once a suitable binding ligand is identified.

The possible role of disordered domains in aiding BR55 binding: SAXS analysis revealed the degree of disorder in BR55, with approximately 33 residues from the N-terminal and 97 residues from the C-terminal forming intrinsically disordered domains. It is possible for these parts of the protein to play a role in binding, however, this would be best proven by comparative binding studies between full length BR55 and a construct of BR55 comprising only the solenoid core, from about 262-571 residues.

4.9.3 Binding studies of BR55 and mutants

Identification of BR55 residues potentially involved in binding was carried out as described in section 4.4.5, to produce mutants for functional and biochemical studies at the IFR, by Nathalie Juge's group. These mutants were W375A, K377A, Y425A, W450A and Y482A; the double-cysteine mutant BR55-P416C/Q516C or BR55-CC and the mutant in which the lower loop had been removed, BR55-ΔF411-T422 or BR55-del. All mutants, apart from W375A could purified for subsequent studies, which is reported in APPENDIX 4. Purification of BR55-W375A had been attempted three times (data not shown) but no protein bands at the expected MW could be visually detected on the SDS-PAGE gel. Perhaps this point mutation may have affected expression or folding of this protein for reasons not yet understood. Binding studies with wild-type and mutant BR55, at low and high pH are ongoing at the IFR to identify the type or types of binding substrate(s); these are being screened using a range of mammalian, mucin and plant-based glycan arrays. Together, the structural findings presented in this thesis and the biochemical work performed at the IFR are being prepared for publication. A new angle to supplement these ongoing biochemical studies is proposed for future investigations. This involves assessing the role of the upper loop region and the role of the intrinsically disordered regions of BR55 in ligand recognition and binding, via mutagenesis and functional assays.

CHAPTER 5

INTRODUCTION:

PURIFICATION OF GRAM-NEGATIVE BACTERIA WaaP FOR STRUCTURAL STUDIES

5 INTRODUCTION: PURIFICATION OF WaaP - AN ESSENTIAL LPS CORE OLIGOSACCHARIDE SUGAR KINASE IN GRAM-NEGATIVE BACTERIA

5.1 Gram-negative and Gram-positive bacteria

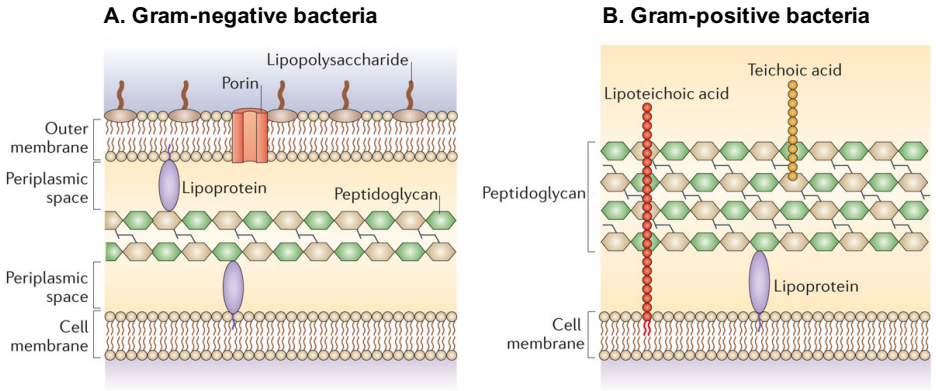


Figure 5.1: Gram-negative and Gram-positive bacteria cell envelope structure. Adapted from [262].

Most bacteria are classified as either Gram-positive or Gram-negative, depending on the structural organisation of their outer cell envelope. As presented in **Figure 5.1**, Gram-positive bacteria have a single lipid-bilayer membrane surrounded by a thick peptidoglycan cell wall [262]. Peptidoglycans are composed of linear polymers of *N*-acetylglucosamine and *N*-acetylmuramic acid (NAG-NAM) cross linked by peptides attached to the lactate residue of NAM [263]. Molecules of lipoteichoic acid and teichoic acid can also be found in the peptidoglycan layer, which are absent in the cell wall of Gram-negative bacteria. The cell envelope of Gram-negative bacteria differs from Gram-positive bacteria in consisting of a thinner peptidoglycan layer around the cell membrane, and an additional outer lipid membrane surrounding it. The peptidoglycan layer runs along the periplasmic space between the inner membrane

(IM) and outer membrane (OM) of Gram-negative bacteria [262]. Gram-negative bacteria are so called as they cannot retain the deep violet colour during Gram staining reactions with crystal violet. This happens because during the procedure, the OM lipid bilayer is dissolved by ethanol, increasing the permeability of the cell wall and enabling the thin peptidoglycan to be decolourised by ethanol. Subsequent counter-staining with safranin then causes the bacteria to turn pink. Since Gram-positive bacteria have a thicker peptidoglycan layer, ethanol is unable to permeate effectively through the cell wall to decolourise the crystal violet dye, and so the cell retains the violet colour. Although Gram-positive bacteria also counter-stain with safranin the deep violet stain supersedes it. Thus, the Gram staining method is a useful tool for initial identification of Gram-positive and Gram-negative bacteria [264].

5.2 Structure of Gram-negative bacteria cell envelope

Gram-negative cell envelopes have three layers: the outer membrane (OM), the periplasm, and the inner membrane (IM) (**Figure 5.1**).

5.2.1 The Outer Membrane

The OM is an asymmetric lipid bilayer, where phospholipids form the inner leaflet, (facing the periplasm), and the outer leaflet (facing the cell surface) comprises of glycolipids called lipopolysaccharides (LPS) [265, 266]. The OM also consists of a range of dedicated proteins that regulate vital cellular functions, like nutrient uptake, signaling, adhesion, waste export, and for pathogenic strains, can be important virulence factors contributing to nutrition, antibiotic resistance and evasion of host defences [266]. Two types of outer membrane proteins (OMP) are known; lipoproteins (LPP) and β-barrel proteins. LPPs are thought to play a key role in maintaining OM-peptidoglycan interactions and preserving OM integrity; certain studies have found that mutants lacking LPPs had a weak OM with a tendency to form vesicles, with severe periplasmic enzyme leakage [267,268]. Some LPPs, are complexed with

transmembrane β-barrel proteins, such as the LptDE complex [269] and the BamABCDE complex [270], as will be discussed shortly. LPPs are composed of a fatty acid moiety that helps in anchorage to the OM inner leaflet, and a third of the LPPs are also covalently linked to peptidoglycan in the periplasm [271].

β-barrel proteins are OM transmembrane proteins, with key functions in transport, assembly, signaling, and OM biogenesis. Structures of many Gram-negative β-barrel proteins have been elucidated in the last two decades, many of which are transporters, porins, channel proteins and adhesins; a comprehensive review describing them is available [272]. β-barrel proteins typically consist of an even number of β-strands (8 to 26) arranged in an antiparallel pattern [266], as seen in **Figure 5.2B**. Despite their profusion, only two OMPs have been found to be essential for bacterial survival [266]; the structures for both as multi-protein complexes have been published by the Dong group in the last 3 years.

The first vital OMP is the LptD protein: a 26-stranded β-barrel protein complexed with an LPP called LptE, which together insert LPS molecules into the OM outer leaflet [269]. They form part of the larger LptA-G LPS transport machinery, which is illustrated in **Figure 5.4** and further elaborated in section 5.3.2. The second essential OMP is part of the BAM five protein complex, responsible for assembling other β-barrel OMPs into the OM outer leaflet. In this complex, BamA is the OMP and BamBCDE are LPPs anchored to the OM inner leaflet [270]. The BAM complex is the final set of proteins in the sophisticated process of OMP biogenesis, as illustrated in **Figure 5.2** below.

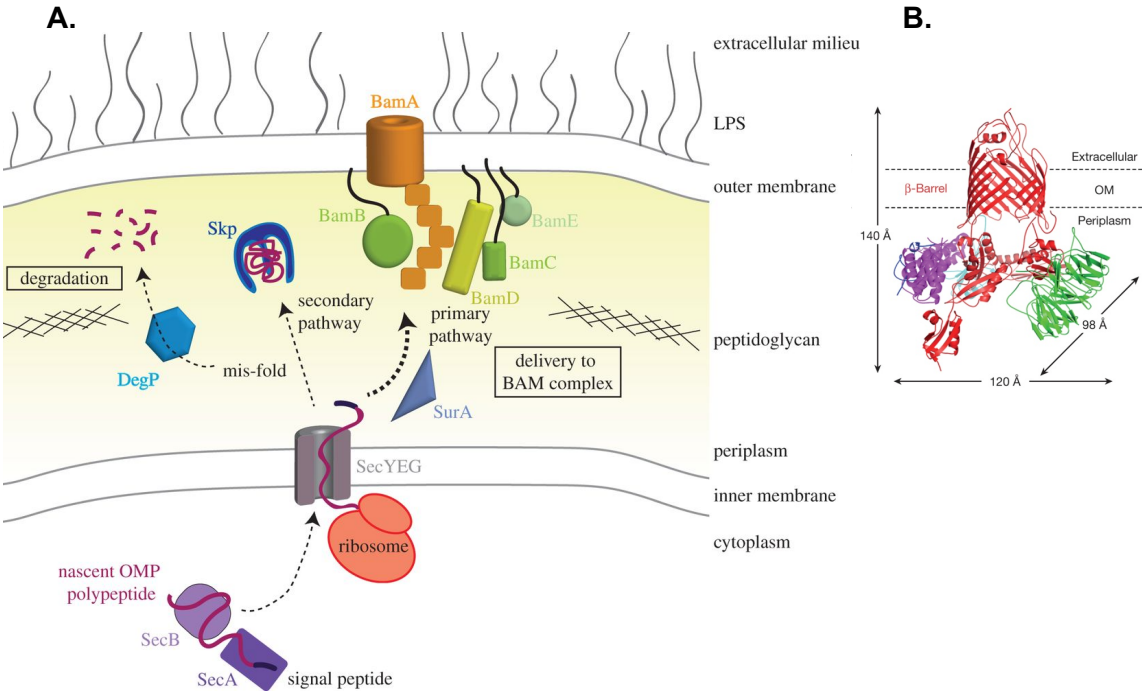


Figure 5.2: OMP biosynthesis and assembly.

A. From [266]. Inactive OMPs are first synthesised in the cytoplasm and moved by SecA/B chaperones (purple) or by the ribosome (red) to the Sec translocase (grey) for transport across the IM. In the periplasm, inactive OMPs are delivered to the BAM complex by SurA (blue-grey) or Skp (blue) chaperones. OMPs misfolding during transport are degraded by DegP (blue hexagon) to avert aggregation in the periplasm [266]. **B.** Adapted from [270]. High-resolution structural model of BamABDE complex. BamA is a β-barrel protein (red), with 16 antiparallel β-strands. LPPs BamB (green), BamD (purple) and BamE (cyan) are part of the BAM complex (BamC is not shown).

5.2.2 The Periplasm

The periplasm is as an aqueous layer between the bacterial outer membrane (OM) and inner membrane (IM) consisting of various periplasmic proteins [273]. The periplasm plays a vital role in sequestering degradative enzymes, like RNAses and alkaline phosphatases, to prevent them from harming the bacterial cell [274]. Periplasmic binding proteins for sugar and amino acid transport, and chaperones involved in OM biogenesis [275], as seen in **Figure 5.2** above, are mainly found here.

5.2.3 The Inner Membrane

Unlike eukaryotic cells, bacteria lack intracellular organelle, and so many membrane-associated functions of eukaryotic organelles are performed in the bacterial inner membrane (IM). Many membrane proteins associated with lipid biosynthesis, energy production, protein secretion, and transport are found in the IM [275]. The IM's cytoplasmic interface is a production site for membrane lipids lipoproteins and proteins, many of which are transported to the OM [265, 276] as discussed in section 5.3.2 below. The IM is a symmetric phospholipid bilayer, consisting of phosphatidyl-ethanolamine, phosphatidyl glycerol and minor amounts of cardiolipin and phosphatidyl serine. Phospholipids are suggested to maintain the IM permeability barrier and serve as a support for membrane proteins [275]. In short, IM proteins are synthesized on ribosomes where the signal recognition particle (SRP) binds to the new peptide at its N-terminal signal sequence, enabling its delivery to the Sec machinery, which then inserts it directly into the IM [277].

5.3 LPS Structure, Synthesis and Assembly

5.3.1 LPS Structure and Function

A major component of Gram-negative bacteria's OM outer leaflet [278], LPS is distinguished by three structural components: A hydrophobic membrane anchor portion called lipid A, a hydrophilic polysaccharide core and a hydrophilic O-antigenic oligosaccharide chain that is specific to bacterial serotype, and extends from the cell surface [265, 279]. A diagram of the general structure of bacterial LPS is provided in **Figure 5.3** below. Lipid A's basic structure consists of β-glucosamine-(1→6)-glucosamine-1-phosphate with fatty acid esters attached to both sugars. The length and number of acyl chains and groups vary between bacterial species [265], and shows a negative correlation with the degree of pathogenic virulence [279]. Lipid A is

a pathogen-associated molecular pattern (PAMP) moiety, responsible for evoking immune responses in mammalian host cells [279]; its associated pattern-recognition receptor in the host is the Toll-like receptor 4 (TLR4) [265, 279].

The oligosaccharide core is divided into two regions – the inner core (proximal to lipid A) and the outer core. The inner core structure can be relatively well conserved within an Order or even a Class of bacteria, which reflects its importance in maintaining OM integrity. The inner core typically contains between 1-4 molecules of Kdo (3-deoxy-α-D-*manno*-octulosonic acid), in addition to L-*glycero*-α-D-*manno*-heptose ($_{L,D}$-Hep). Kdo is the only component found in all known cores, and is solely associated with bacterial LPS. Inner core glycans are typically phosphorylated or modified with phosphate bearing groups. These modifications increase the overall negative charge of the cell membrane, which is vital for structural stability of the LPS, and the fitness of the bacterial OM against antibiotics and detergents [265, 280-282] – this shall be elaborated in section 5.5, as it is pivotal to the research discussed in this thesis [265]. The outer core consists mainly of hexoses, including glucose, galactose, and *N*-acetylglucosamine, and is structurally more diverse than the inner core [265]; this will be explored in section 5.4.

The O-antigen is the primary structural constituent of LPS that defines a specific bacterial strain and exhibits the most structural diversity. Over 60 monosaccharides and 30 non-carbohydrate constituents for O-antigens have been identified [265], and repeat units of different LPS can comprise varying number of monosaccharides, may be linear or branched and form homo- or heteropolymers with varying patterns of *O*-acetylations or glycosylations. O-antigen structures have been used to identify and assign serogroups to *E. coli*, *S. enterica*, and *P. aeruginosa* [265], [279]. Its location at the cell surface exposes it to the bacterium's habitat, where their main role is suggested to be protective, such as host immune response evasion [283, 284].

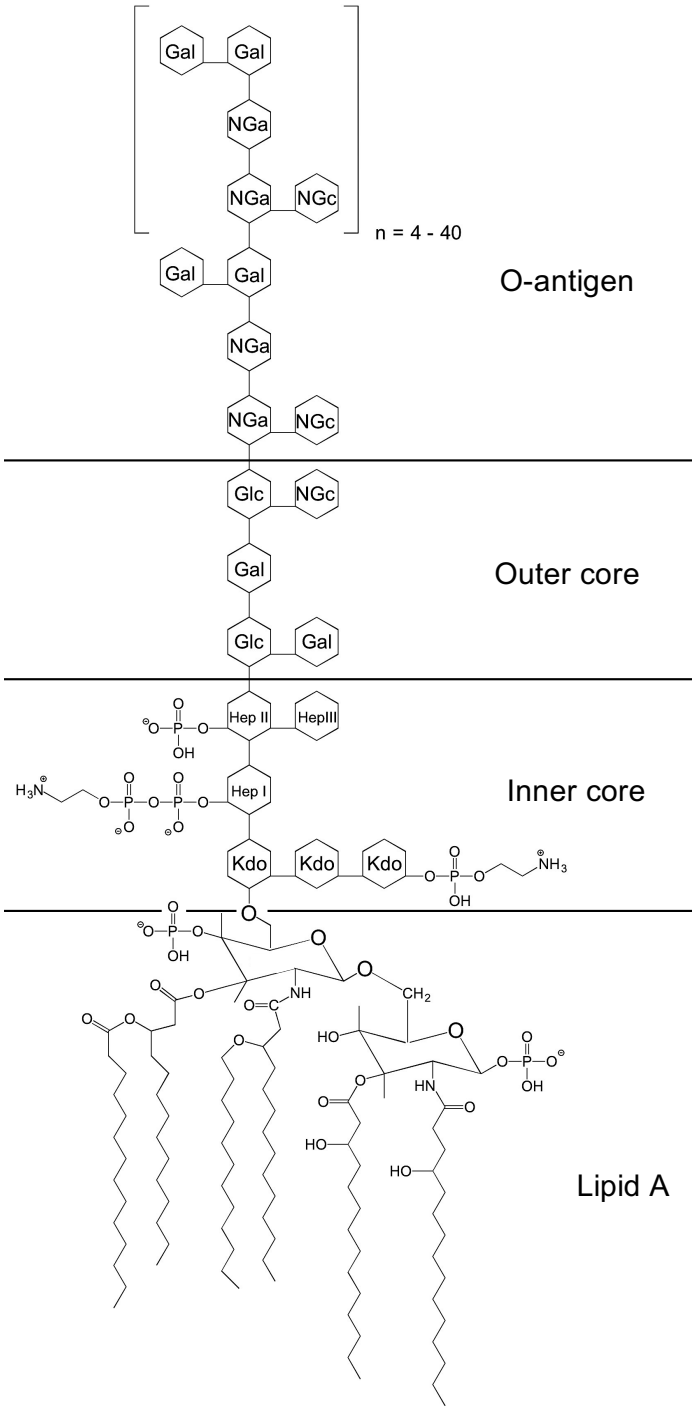


Figure 5.3: General structure for Gram-negative Lipopolysaccharide molecule
Kdo = 3-deoxy-α-$_D$-*manno*-octulosonic acid; **Hep** = (heptulose) $_L$-*glycero*-α-$_D$-*manno*-heptopyranose; **Glc** = β,$_D$-glucose; **Gal** = β,$_D$-galactose; **NGc** = Glucosamine; **NGa** = Galactosamine. Kdo III and the phosphate group on Hep I are modified by 2-aminoethyl phosphate (PEtN). Lipid A is consists of a β-glucosamine-(1→6)-glucosamine-1-phosphate base with fatty acid esters.

5.3.2 LPS Synthesis and Assembly

LPS components are synthesised at the IM prior to transport and assembly to the OM. LPS localisation to the OM from the IM can be divided into three key stages: 1) flipping of the lipid A-core segment from the cytoplasmic interface across the IM onto the periplasmic interface, 2) attachment of the O-antigen segment to the outer core in the periplasm, and 3) transport of the mature LPS across the periplasm to the OM.

The lipid A and core oligosaccharide sections of LPS are synthesised in the cytoplasmic side of the IM, and then flipped across the IM by a flippase named MsbA [285], which belongs to the ATP-binding cassette (ABC) family of proteins [286, 287]. The O-antigen is assembled separately at another site in the cytoplasm by a selection of glycosyltransferases and then translocated across the IM to the periplasm where it is ligated onto the outer core segment to produced mature LPS [288, 289]. Currently, there three known pathways for O-antigen biosynthesis in Gram-negative bacteria, which are distinguished by the membrane translocation mechanisms and the corresponding enzymes that facilitate them, although, the initial cytoplasmic O-antigen assembly and final ligation processes remain quite similar. These are the Wzy-dependent, ABC-transporter dependent and synthase-dependent pathways [265]. These pathways will not be discussed further as it is beyond the scope of this thesis, but a comprehensive review is provided by Raetz and Whitfield [265].

LPS transport to the OM occurs via protein bridges that span from the cytoplasm to the OM [290, 291]. These bridges are formed by a series of proteins, named the Lpt proteins (LptA to LptG). Lpt has been most extensively studied in *E. coli* and *S. enterica* and is widely conserved amongst LPS producers LptB, C, F and G are form an ABC transporter complex, and are transmembrane proteins spanning the IM [291, 292]. LptA is a purely periplasmic protein that relays LPS across the periplasm to the OM [293], and links the IM spanning LptBCFG complex to the OM spanning LptDE complex; LptA forms a bridge from LptC to the N-terminal domain of LptD. The OM proteins LptD and LptE form a stable complex and mediates LPS assembly at the cell

surface [269, 276]. **Figure 5.4** below summarises the LPS transport mechanism across the cell envelope and indicates the proteins involved in the process.

Synthesis of the oligosaccharide core, and the enzymes involved in the process, are discussed more comprehensively in the upcoming section, as it is relevant to the aims of the research described in this thesis.

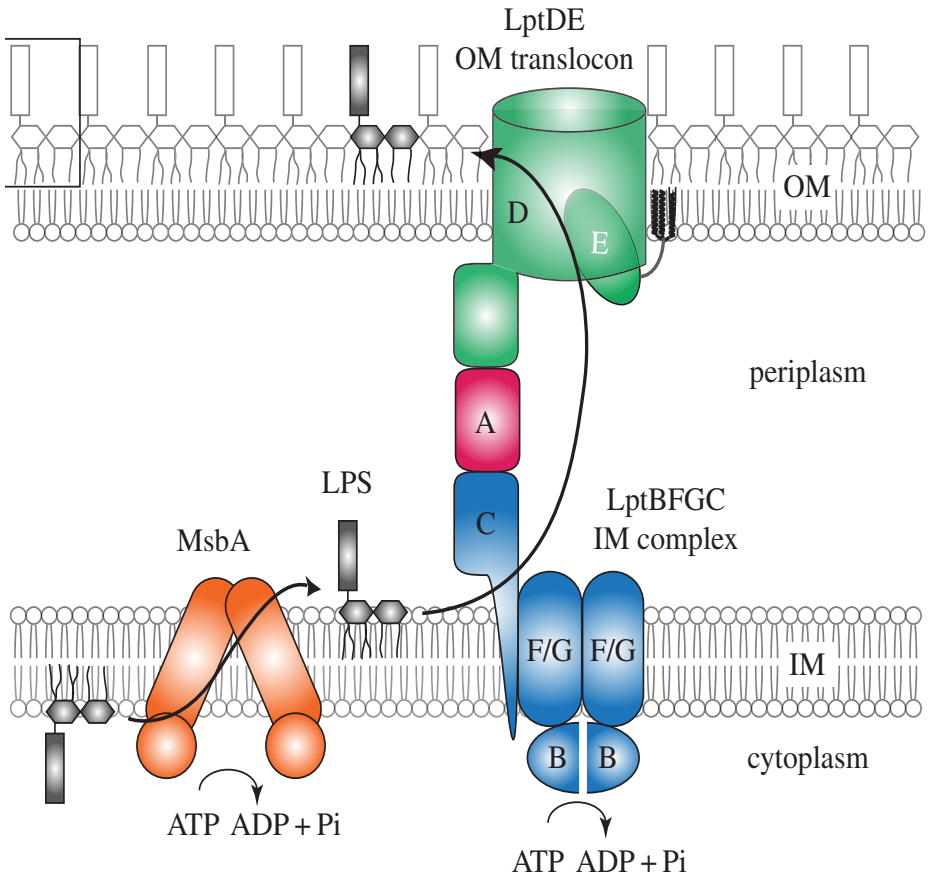


Figure 5.4: A scheme depicting LPS transport across the cell envelope

Image adapted from [294]. 1) After cytoplasmic synthesis, the lipid A-core molecule is flipped across the IM by MsbA, utilising ATP for the process. 2) At the IM periplasmic side, the O-antigen subunit is ligated onto the outer core to give mature LPS (process not shown). 3) LPS transport to the OM is mediated by the LptA-G proteins. First, the LptBFG ABC transporter extracts LPS from the IM in an ATP dependent manner, and then relays it across the periplasmic bridge, comprising of LptC, LptA and LptD's N-terminal domain. At the OM, the LptDE plug and barrel translocon inserts the LPS into the outer leaflet.

5.4 Proteins involved in core oligosaccharide (OS) synthesis

It was mentioned in 5.3.1 that core OS can be subdivided into two regions, the lipid A connected inner core and the O-antigen linked outer core. The inner core structure is typically conserved amongst Gram-negative bacteria and consists of 3-deoxy-$_{D}$-*manno*-octulosonic acid (Kdo) and $_{L}$-glycero-$_{D}$-manno-heptose (Hep) residues. The outer cores exhibit greater structural diversity and are made of varying amounts of $_{D}$-glucose (Glc) and $_{D}$-galactose (Gal) hexoses, linked together differingly from one strain to another [295]. For example, structures of core OS in *E. coli* serotypes R1, R2, R3, R4 and K-12 are different, albeit their backbones are all linear, six-unit OS to which other substituents are attached [296,297]. In Gram-negative bacteria, multiple chromosomal loci encode clusters of proteins for the conversion of core OS sugar precursors, and synthesis and assembly of core OS sugars. However, this thesis will focus on the *waa* or *rfa* locus which encode enzymes needed chiefly for OS assembly [298]. The names *waa* and *rfa* have been used interchangeably in different studies and genomes, but in this thesis *waa* will be used. Structural differences in core OS of various Gram-negative strains are reflected in the variety of genes controlling their biosynthesis. For example, a comparative study of the genomes of *E. coli* K-12, three *S. enterica* serovars and one *Klebsiella pneumoniae* serovar, revealed eight K-12 *waa* genes that had no close homologues in *S. enterica* or *K. pneumoniae* genomes [299].

5.4.1 The *waa* locus in *Escherichia coli* and *Salmonella enterica*

A browse through the literature suggested that the core OS of *E. coli* serotypes have been most extensively characterised, followed by *S. enterica* sv. typhimurium, which share some conserved features with the *E. coli* core OS. Both *E. coli* and *S. enterica* belong to the Order Enterobacteriales within Class Gammaproteobacteria and will be the standard models to understand LPS core OS biogenesis in this thesis. In *E. coli* and *S. enterica*, genes required for core OS biogenesis are found in three operons within the chromosomal *waa* locus: *gmhD, waaQ and waaA (or kdtA)*.

The *gmhD* operon has four genes: *gmhD-waaF-waaC-waaL* [298]. The *gmhD* gene, also named *rfaD* or *waaD,* encodes ADP-$_{L}$-glycero-$_{D}$-mannoheptose-6-epimerase, which epimerises ADP-heptose to Hep [300]. *waaF and waaC* encode the heptosyl transferases, WaaF and WaaC, that catalyse the assembly of Hep I and Hep II, respectively, onto Kdo [301]. The *waaL* gene expresses the O-antigen ligase WaaL, which ligates the lipid A-core complex onto O-antigen at the periplasmic interface in the IM to form mature LPS [302] (see **Figure 5.4**). **The *waaQ* operon** bears 7-11 genes (depending on enteric bacteria serotype) which encode enzymes for outer core OS assembly, and outer and inner core sugar modification [298], [301]. The names and functions of enteric bacteria waaQ proteins are compiled in **Table 5.1** below. **The *waaA* operon** bears a single gene expressing WaaA, a bifunctional Kdo transferase that sequentially adds two Kdo molecules to lipid A [301], [303]. The sequence of core OS assembly in *E. coli* and *S. enterica* will now be described briefly.

1) Inner core assembly: the inner core's four sugar backbone consists of 2 Kdo residues (Kdo I and Kdo II) and 2 Hep residues (Hep I and Hep II). Hep I is added to Kdo I, followed by Hep II onto Hep I by WaaC and WaaF, respectively. WaaP, a sugar kinase, then phosphorylates O4 of Hex I [281]. A Hep III substituent is then added to Hep II by WaaQ, followed by phosphorylation of Hep II O4 by WaaY [281]. This inner core structure and its related glycosyltransferases and kinases are conserved across *E. coli* and *S. enterica* serotypes [298], [301], as shown in **Figure 5.5** below.

2) Hex I addition by WaaG: assembly of the outer core backbone, comprising of three hexoses, begins with the α-1,3 addition of a Glc residue onto Hep II by WaaG. This initial Hex I and its glycosyltransferase are conserved across all *E. coli and S. enterica* serotypes. The degree of divergence increases further along the outer core.

3) Hex I substitution by WaaB: the conserved Glc I residue of *S. enterica* and *E. coli* K-12 and R2 are substituted by a galactosyl residue (Gal I) via an α-1,6 bond. *E. coli* R1, R4 and R3 core types lack Hex I substitution (but are Hex II substituted instead).

4) Hex II addition by WaaI/O: as summarized in **Table 5.1**, there are two types of Waa enzymes associated with Hex II addition to Hex I, which are differentiated

according to their glycosyltransferase activity: WaaI in *S. enterica* (and also *E. coli* R3 [304]) for the α-1,3 addition of Gal onto Glc I (Hex I), and WaaO in other *E. coli* serotypes for the α-1,3 linkage of a Glc residue onto Glc I [305].

5) Hex III addition by WaaJ/R/T: the enzymes are named distinctly due to differences in the types of glycosidic bonds they form. All of them facilitate an α-1,2 glycosidic bond, however, WaaJ does this between O1 of Glc II (Hex III) and O2 of Gal I (Hex II), WaaR - between Glc III O1 (Hex III) and Glc II O2 (Hex II) and WaaT - between Gal I O1 (Hex III) and Glc II O2 (Hex II). WaaJ is associated with *S. enterica* and *E. coli* R3, WaaR with *E. coli* K-12 and R2, and WaaT with *E. coli* R1 and R4.

6) Hex II substitution by WaaV/X: a divergence in Hex II substituents is observed in *E. coli* strains, where two types of β-glycosyltransferases, WaaV from R1 and WaaX from R4, add β1,3-linked Glc and β1,4-linked Gal onto Glc II (Hex II), respectively [306]. In R3 serotypes, GlcNAc substitution of its Gal Hex II has been proposed, but the enzyme remains unknown [304], [307]. *E. coli* R1, R4 and *S. enterica* strains are devoid of Hex II substitution (however their Hex I residues bear a Gal substituent).

7) Hex III substitution by WaaK/U/W: the sugar type of substituents on Hex III by far show the most divergence. WaaK forms an α-1,2 linkage between GlcNAc and Glc III in *S. enterica* and *E. coli* R2; WaaU adds a Hep IV residue onto Glc III in *E. coli* K-12 via an α-1,6 linkage; finally, WaaW forms an α-1,2 bond between a Gal residue and Gal I in *E. coli* R1 and R4. The addition of an α-1,2-linked Glc onto Glc II of the *E. coli* R3 core is proposed, but the enzyme involved remains to be identified [304, 307].

8) Kdo attachment to lipid A by WaaA and **9) Outer core ligation to O-antigen by WaaL:** These steps are conserved across all *E. coli* and *S. enterica* serotypes [298, 301]. However, the sequence of WaaL enzymes amongst various strains show divergence, corresponding to the nature of their outer core constituents [304].

The core OS assembly of selected *E. coli* strains, *S. typhimurium* and *P. aeruginosa* PAO1 O5 are summarised and depicted in **Figure 5.5**, offering a visual example of the core's structural divergence in Gammaproteobacteria.

Table 5.1: *E. coli* and *S. enterica* genes of the *waa* locus involved in core oligosaccharide assembly.

Adapted from [301]. Genes are listed in the order in which their products have been discovered to act.. Genes from the ***gmhD***, ***waaQ*** *and* ***waaA*** operons are indicated in **blue**, **red** and **green** bold font, respectively. Gene products are classified by function, indicated by the colour of the row: Dark grey = inner core backbone Hep heptosyltransferases; Light orange = Hep kinases; Green = substituent heptosyl-transferase; Blue = outer core backbone Hex glycosyltransferases; Yellow = Hex substituent glycosyltransferases; Light grey = enzymes in unique cases; Pink = Kdo transferase; White = enzymes beyond core OS assembly. **Abbreviations**: K-12 = *E. coli* K-12; R1,R2,R4 = *E. coli* R1, R2 or R4 serotype; *ST* = *Salmonella typhimurium.*

Gene	Function of gene product	Reference(s)
waaD	Epimerisation of ADP-heptose to the inner core Hep sugar (epimerase)	K-12 [300], [308]; *ST* [309]
waaC	Addition of Hep I to Kdo I of inner core	K-12 [310]; *ST* [311]
waaF	Addition of Hep II to Hep I to complete inner core backbone	K-12 [312]; *ST* [309]
waaP	A sugar kinase required for transfer of phosphate from ATP to O4 of Hep I	K-12 [313]; *ST* [314]
waaQ	Addition of a Hep III substituent to Hep II	R1 [281]; K-12 [315]
waaY	A sugar kinase similar to waaP that phosphorylates O4 of Hep II	R1 core [281]; *ST* [316]
waaG	Addition of outer core Hex I (Glc I) to Hep II	K-12 [313]; *ST* [317]
waaB	Addition of branch Gal to Glc I	K-12 [318]; *ST* [317]
waaI/O	Addition of Hex II to outer core (WaaO for Glc II in *E. coli* and WaaI for Gal I in *Salmonella* spp.)	K-12 [305]; *ST* [317]
waaJ/R/T	Addition of Hex III to outer core (WaaR for Glc III in *E. coli* K-12; WaaT for Gal I in *E. coli* R1 and R4; WaaJ for Glc II in *Salmonella* spp.)	K-12 [319]; R1 and R4[306] *ST* [317]
waaV/X	Addition of β-linked hexose substituent onto Hex II (Glc II) (WaaV adds Glc in R1 and WaaX adds Gal in R4 core types)	R1 and R4 [306]
waaK	Addition of GlcNAc to Hex III (Glc II) in *Salmonella* spp. and to Hex III in *E. coli* R2 serotypes	*ST* [316]; R2 [320]
waaU	Addition of terminal Hep IV to Hex III (Glc III) in *E. coli* K-12 serotypes.	K-12 [321]
waaW	Addition of Gal II to Hex III (Gal I) in *E. coli* R1 and R4 serotypes	R1 and R4 [306]
waaS	Rha transferase required to synthesise rough type LPS, without O-antigen	K-12 [322]
waaZ	Addition of Kdo III to synthesise rough type LPS, without O-antigen	K-12 [322]
waaA	Sequentially transfers Kdo I and Kdo II to lipid IV_A, the lipid A precursor	K-12 [323]
waaL	Ligation of O-antigen to core (ligase)	K-12 [302]; *ST* [316]

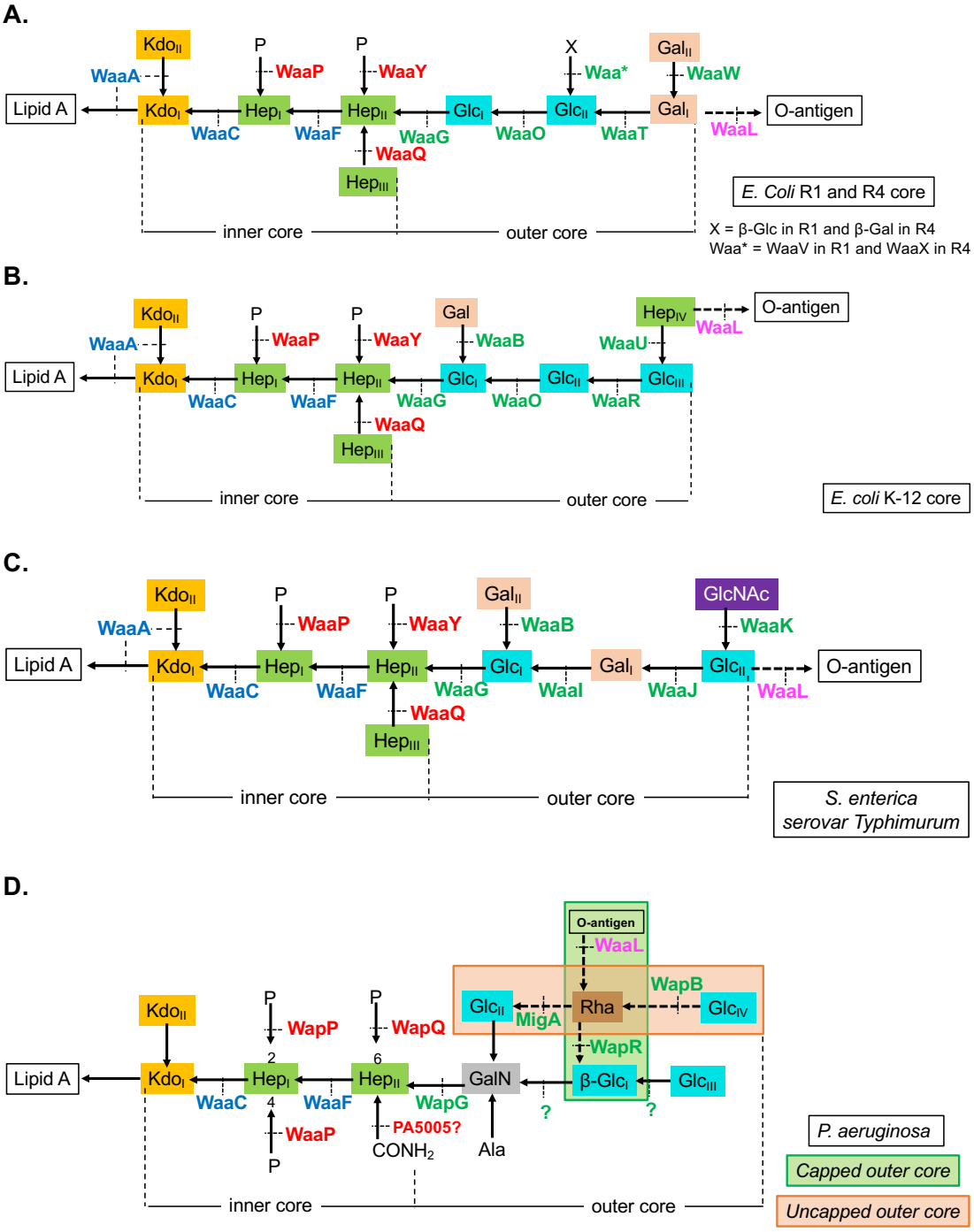


Figure 5.5: Structure and assembly of the LPS core oligosaccharide in selected Gram-negative bacteria strains

Gene products in blue font are involved in inner core assembly, those in green are for outer core assembly, enzymes that modify the structure of the inner core are in red font and the enzyme that ligates the lipid A-core onto the O-antigen is in magenta. The biosynthesis schemes are listed in decreasing order of *waa* operon conservation between the four strains. *E. coli* R1/R4 serotype (**A**), *E. coli* K-12 (**B**), *S. typhimurium* (**C**) and *P. aeruginosa* PAO1 O5 (**D**).

Kdo =3-deoxy-α-D-*manno*-octulosonic acid; Hep = L-*glycero*-α-D-*manno*-heptose; Glc =$_{D}$-glucose; Gal =$_{D}$-galactose; GlcNAc =*N*-acetylglucosamine; Rha =rhamnose; GalN =Galactosamine; $CONH_2$=carbamoyl; Ala= alanyl; P= phosphate group.

5.4.2 The core oligosaccharide in *Pseudomonas aeruginosa*

Key differences in core OS structure and biosynthetic enzymes are obvious between enteric bacteria and *P. aeruginosa*, as shown in the core OS diagrams in **Figure 5.5**. Such a divergence is not surprising, because although *E. coli, S. enterica* and *P. aeruginosa* are from Class Gammaproteobacteria, they belong to different Orders. In this section, *P. aeruginosa* core OS structure will be introduced as an example from Order Pseudomonadales. *P. aeruginosa* strains are classified into 20 International Antigenic Typing Scheme (IATS) serotypes, largely driven by variations in sugar constituents and O-antigen linkages [324]. Although, some structural features in the inner and outer cores are strongly conserved across *P. aeruginosa*, as discussed:

1) **The inner core** in *P. aeruginosa* consists of a four sugar backbone, with two Kdo residues (Kdo I and Kdo II) and two Hep residues (Hep I and Hep II), much like *E. coli* and *S. enterica*. However, the *P. aeruginosa* inner core is more phosphorylated and a Hep III branch on Hep II is absent. The three phosphorylation sites are O2 and O4 of Hep I and O6 of Hep II. Also, Hep II bears a carbamoyl group at O7 [324].
2) **The outer core** comprises a five sugar backbone of one $_{D}$-galactosamine (GalN), one $_{L}$-Rhamnose (Rha) and three Glc residues. N2 of GalN is alanyl substituted [324].

P.aeruginosa produces two basic core OS glycoforms: capped and uncapped. The capped glycoform is linked to O-antigen, whereas the uncapped core lacks it. Another difference in the two glycoforms lie in the position and linkage of the Rha residue; in capped cores, O-antigen ligates to Rha which is 1,3-linked to Glc I, and in uncapped cores Rha is 1,6-linked to Glc II, proceeded by the addition of β-Glc to Rha O2. Nine of the 20 IATS serotypes (O2, O5, O7, O8, O10, O16, O18, O19 and O20) express uncapped LPS, whereas the others produce capped LPS [324, 325]. Other variability in the core OS of *P. aeruginosa* serotypes is due to non-glycosyl substitution, such as phosphorylation and O-acetylation [324]. Although the function and sequential action of the gene products remain the same, the *waaF, C, G* and *P* genes in *P. aeruginosa* are found on the same operon, unlike those in enteric core clusters, where *waaF* and

C occur in divergent operons from *waaP* and *G*, as mentioned earlier in 5.4.1 [312, 326]. *wapP* and *wapQ*, genes expressing sugar kinases unique to *P. aeruginosa*, were discovered immediately downstream on the same operon and were found to phosphorylate Hep I O2 and Hep II O6, respectively [280, 326]. WaaG adds GalN to Hep II (named WapG in some studies) [280, 327]. In capped cores, WapR catalyses Rha linkage to Glc I [328, 329], followed by ligation of O-antigen to Rha by WaaL [330]. In uncapped cores, MigA catalyses Rha linkage to Glc II [328, 329] followed by attachment of β-Glc to Rha by WapB [331]. As seen in **Figure 5.5**, some enzymes in the *P. aeruginosa* core OS assembly pathway remain to be identified.

5.5 Inner core phosphorylation is vital for cell viability

The Gram-negative bacteria OM acts as a barrier to many antibiotics and host defense factors [271, 275] largely due to features of the LPS molecules that compose the OM outer leaflet [332]. Isogenic mutations in the *waa* locus of several glycosyl transferase genes, producing variably weakened OM phenotypes is reported in a comprehensive *E. coli* K-12 *waa* functional study [321]. This has been attributed to the production of O-antigen deficient LPS, termed rough-LPS (R-LPS), as O-antigen cannot be ligated onto an incomplete lipid A-core OS, due stringent WaaL substrate specificity [333]. However, mutants specifically deterring phosphorylation of the inner core were found to express a pleiotropic phenotype called deep rough with more detrimental OM defects. Deep rough mutants are discerned by hypersensitivity to antibiotics and detergents, phospholipid accumulation in the OM outer leaflet, periplasmic protein leakage, and a reduction in OM protein content. This inner core phosphoryl-dependent OM viability was observed in *E. coli, S. enterica* and *P. aeruginosa* [280-282,313,326,334], and also briefly reported for *Shigella flexneri* [335]. Phosphoryls are postulated to be vital for OM stability as their negative charge may aid in cross-linking between adjacent LPS molecules via divalent cations [336].

LPS from phosphate deficient deep rough mutants were also shown to be unable to support the proper folding of some OM proteins [282, 337].

5.5.1 WaaP is critical for *E. coli* and *S. enterica* viability

In an *E. coli* strain with R1 core type, WaaP was demonstrated to be essential for phosphoryl substitution to Hep I, and a prerequisite for the action of WaaQ and WaaY towards Hep III addition and Hep II phosphorylation, respectively [281] (section 5.4.1 and **Table 5.1**).

In each of the three isogenic depleted mutants, sugar composition analysis detected the presence of all outer core OS sugars implying that the absence of *waaP, Q* and *Y* had no notable influence on the glycosyl transferase activities of genes beyond *waaQ*. However, the mutant phenotypes varied in the inner core chemistry.

The core OS of *waaY* deficient mutant was devoid of Hep II O4 phosphorylation, thereby confirming its activity as a Hep II sugar kinase. The core OS of *waaQ* deficient mutant lacked a Hep III residue and the phosphoryl at Hep II, despite chromosomal expression of *waaY*; this demonstrated that the addition of Hep III onto Hep II is a prerequisite for WaaY activity, possibly driven by WaaY's fastidious substrate specificity for its sugar kinase activity upon Hep II. The deletion of *waaP* resulted in an inner core phenotype completely devoid of phosphorylation and the Hep III residue, despite the normal expression of WaaQ and WaaY proteins. This implied that phosphorylation at Hep I was essential for *waaQ* Hep III substitution upon Hex II.

Furthermore, amounts of the outer core OS residues, β-Glc and Glc II, were noticeably decreased in the *waaP* isogenic mutant, suggesting that lack of WaaP activity may also affect the functioning of enzymes further down the core OS synthetic pathway. Thus, this study was paramount in differentiating the functions of *E. coli* WaaP, WaaQ and WaaY, establishing their sequence in the inner core OS biosynthetic pathway, and in demonstrating their interdependence [281].

These isogenic mutants were also tested for their resistance towards the hydrophobic antimicrobial, novobiocin, and SDS. It was established that WaaP is essential for LPS integrity, leading to OM stability: *waaY* and *waaQ* depleted strains exhibited only a minor increase in sensitivity towards the hydrophobic agents, quite unlike the hypersensitivity expected of deep-rough phenotypes. However, the lack of core OS phosphoryl substituents in the *waaP* depleted strain resulted in a significant increase in sensitivity towards novobiocin and SDS [281].

The importance of *waaP* in OM viability was similarly demonstrated in *S. enterica* sv. Typhimurium, where an isogenic *waaP* depleted strain produced an inner core devoid of phosphoryl substituents, resulting in hypersensitivity towards the cationic antimicrobial, polymyxin, and loss of virulence *in vivo* in mice models. Although, the mutant could produce full length capped LPS molecules, a portion of them were found to be prematurely truncated in the outer core. The consequences for WaaP suppression are therefore, very similar in *S. enterica* and *E. coli,* which is expected as their WaaP, Q and Y enzymes, and the inner core OS structures are significantly conserved (see **Figure 5.5**).

5.5.2 WaaP and WapP are critical for *P. aeruginosa* viability

P. aeruginosa has an exceptionally low OM permeability compared to *E. coli* and *S. enterica*, and coupled with secondary resistance mechanisms, such as multidrug efflux pumps, it makes the OM an increasingly effective barrier against antibiotics and detergents [338], [339]. It is suggested that the role of inner core phosphoryls and associated divalent cations in cross-linking LPS molecules is especially vital for *P. aeruginosa* OM integrity due to the following reasons:

1) **Hypersensitivity to lysis in the presence of divalent metal ion chelators**, **like ethylenediamine-acetic acid (EDTA):** The susceptibility of certain Gram-negative bacteria towards lysis by lysozyme in the presence of EDTA was first demonstrated by Repaske [340], who postulated that EDTA sequestered metal ions involved in binding the cell wall together. Another study [341] found a correlation between EDTA

sensitivity and cell wall composition, where EDTA's bactericidal potency was greater towards bacteria with a higher cell wall phosphorous content; i.e. *P. aeruginosa* showed greater susceptibility towards EDTA + lysozyme compared to *E. coli* [341]. Furthermore, reduced susceptibility of *P. aeruginosa* towards EDTA + lysozyme was noted in the presence of Mg^{2+} ions, possibly due to Mg^{2+} binding with the EDTA permeabiliser, or replacing metal ions removed from the OM [342]. These findings therefore suggest that metal cations in the OM of *P. aeruginosa* play a crucial role in preserving its structural integrity.

2) **An unusually high phosphoryl content in the LPS**: *P. aeruginosa* has the most densely phosphorylated of known Gram-negative bacteria cores, with between 3 to 6 predicted [343]. The positions of three phosphoryl substituents have been confirmed, two being on Hep I (added by WaaP and WapP) and one on Hep II (added by WapQ). Also, a putative LPS kinase gene (PA5006) was found downstream of the *wapQ* gene on the same operon – however, this protein remains to be characterised [280].

3) ***P. aeruginosa* mutants lacking inner core Hep or phosphate have never been isolated**: while *waaP* (and *waaY)* depleted mutants expressing deep rough phenotypes in *E. coli* and *S. enterica* were previously isolated and reported [281], [282], the inability to generate *waaP* and *wapP* knockout mutants of *P. aeruginosa* clearly demonstrate the importance of inner core Hep I phosphorylation in cell survival, even under benign conditions [280, 326].

4**) *waaP* and *wapP* LPS kinases are essential for a healthy OM:** outcomes of the inactivation of *waaP* and *wapP* gene effects were tested in knockout *P. aeruiginosa* strains complemented with an L-arabinose inducible *waaP* and *wapP* expression system [280]. Uninduced strains stopped growing after several generations. An analysis of their LPS molecules revealed severe outer core OS truncations, with a tendency to localise in the IM, reflecting an inability of the OM transport enzymes in recognising the defective LPS molecules. These observations suggest that *waaP* and *wapP* are vital for producing full length, functional LPS, to an extent that has not been observed for *E. coli* and *S. enterica* LPS production. WaaP and WapP depleted cells

also exhibited significant morphological defects, such as IM invaginations [280]. However, WapQ, the third known kinase, was shown to be much less critical for *P. aeruginosa* survival, as a WapQ depleted mutant could be isolated, which had an unaffected minimum inhibitory concentration (MIC) for novobiocin, relative to the wild-type, and a minor decrease in MIC for SDS [326]. This information goes further in distinguishing the differing significance of Hep I phosphorylation versus Hep II phosphorylation in the *P. aeruginosa* inner core OS, affecting its survival.

The following subsection will describe all WaaP kinases characterised to date.

5.5.3 WaaP proteins: Function and Biochemistry

WaaP protein from *E. coli* and *P. aeruginosa* are the most biochemically and functionally characterised to date, and offer useful examples for understanding WaaP activity across enteric bacteria and Pseudomonads. The following subsections will summarise the findings for each type of WaaP.

5.5.3.1 *E. coli* WaaP kinase (ECWaaP)

Purification of functionally active ECWaaP has been described in [7], where it was reported to catalyse the incorporation of ^{33}P from [^{33}P]ATP into acceptor LPS that was extracted from a *waaP* deficient *E. coli* mutant. Mg^{2+} was required for kinase activity and its optimum pH was from 8.0 to 9.0. The Michealis-Menten constant (K_m) was estimated as 0.13 mM for ATP and 76 μM for LPS, providing evidence that EC WaaP is an LPS kinase. Residues 159-171 (N**HRD**CYCHFLLH) and 184-192 (SVIDLHRAQ) were noted to be highly conserved across all known WaaP homologues (see APPENDIX 7), and whilst showing only 10-15% sequence similarity with eukaryotic kinases, the majority of its conserved amino acids coincided with residues that were known to be important for eukaryotic kinase catalytic activity, particularly the HRD motif, which is known to be conserved throughout the eukaryotic protein kinase family [344, 345]. Based on previous information from known kinases, it was postulated that D162 in EC WaaP would act as a proton acceptor, from the 4-OH of Hep I, hence,

producing a nucleophilic oxyanion that would attack a phosphate from ATP. Mutagenesis of D162 to Ala was performed, rendering the WaaP kinase inactive *in vitro* and *in vivo,* thus proving that this residue is essential for its kinase function.

5.5.3.2 *P. aeruginosa* WaaP kinase (PAWaaP)

The purification and biochemical characterisation of PAWaaP is reported in [346]. The sequence similarity between PAWaaP and ECWaaP is about 55%, as presented in **Table 5.2** below. Similar to ECWaaP, Mg^{2+} was an essential ingredient in biochemical assays of PAWaaP, at a of pH 8.0. Its K_m was estimated as 0.22 mM for ATP and 14.4 µM for dephosphorylated LPS, in the same order as the respective K_m*s* for ECWaaP. Its V_{max} was 408.24 pmol/min and k_{cat} was 27.23 min^{-1} for dephosphorylated LPS. The (N**HRD**CYCHFLLH) and (SVIDLHRAQ) residues, from 160-172 and 185-193, respectively were conserved with that of ECWaaP (see APPENDIX 7). What was found to be significantly different about PAWaaP, with respect to ECWaaP, is that it also demonstrated protein tyrosine kinase activity, in addition to Hep I kinase activity. PAWaaP showed increasing conservation with multiple catalytic motifs of eukaryotic tyrosine kinases. In addition to D163, site-directed mutagenesis was performed for K69 and D181, and the resultant mutants demonstrated loss of sugar kinase and protein tyrosine kinase activity. The authors also demonstrated that PAWaaP is localized in the cytoplasm of *P. aeruginosa*, which is consistent with the knowledge that the core OS is synthesised in the cytoplasm, close to the IM (see section 5.3.2).

Table 5.2 below presents the physical properties of EC, PA and *S. enterica* (SEn) WaaP proteins for comparison.

Table 5.2: Physical properties of some WaaP homologues

Species	Protein Length	Size (kDa)	pI	% Identity
E. coli	265	31.05	9.7	100.0
S. enterica	265	31.00	9.4	83.0
P. aeruginosa	268	31.31	9.9	55.13

5.6 Research Aims

Establishing unequivocally that WaaP is essential for LPS completion, OM viability and multidrug resistance in *E. coli, S. enterica* and more so in *P. aeruginosa,* implies that WaaP would be a promising target for the discovery of much needed novel therapeutics to combat antibiotic resistance Gram-negative infections. Therefore, a structural study of the WaaP enzyme would be of immense interest to fulfill this gap. The purification procedures for ECWaaP and PAWaaP reported so far afforded enough protein only for biochemical analyses. Therefore, the aim of this research was to improve the yield of recombinant WaaP for X-ray crystallography studies. As will be explored in the results section below, WaaP from four different homologues were investigated, with another two in the pipeline. It was reasoned that because WaaP is essential for the survival of most Gammaproteobacteria, and that key putative catalytic regions are highly conserved (see APPENDIX 7 and 8), the structure of any one WaaP homologue would be able to offer valuable insights into the general sugar kinase activity for all the other WaaP homologues.

CHAPTER 6

WaaP PURIFICATION METHOD DEVELOPMENT

6 WaaP PURIFICATION METHOD DEVELOPMENT

6.1 Previous work

ECWaaP_pLou3 and SEnWaaP_pLou3 plasmids, for the recombinant production of MBP-fused WaaP, were constructed by Zhongshan Wang (Dong lab, UEA) (see 1.1.1.2). 6His-tagged ECWaaP [7] and PAWaaP [346] purification methods are reported in the literature, and the experimental conditions described in 6.3.1 were adapted from the methods therein. The pWQ911 plasmid, expressing the Gro*ESL* chaperonin, and used in the experiments reported in sections 6.3.4 and 6.3.6 was provided by C. Whitfield (University of Guelph, Ontario, Canada), and its construction is reported by Yethon and Whitfield [7]. All experiments subsequently reported were conducted as part of the thesis.

6.2 PCR and ligation of WaaP inserts into expression vectors

Images of WaaP PCR products resolved on an ethidium bromide stained agarose gel are shown in **Figure 6.1** below. The WaaP gene from *P. aeuroginosa* PAO1 was chosen as an additional WaaP homologue for experimental testing. Mimicking the cloning and expression procedures given in the literature, ECWaaP [7], and PAWaaP [346] were amplified and cloned into the Novagen pET28b (+) vector (Merck Millipore, Massachusetts, USA) for expression as C-terminal 6His-tagged proteins (method described in sections 1.2.1 and 1.2.2) Another PAWaaP construct with different restriction sites (see **Table 1.1**) was designed for ligation into the pLou3 [6] vector, to enable expression with a TEV protease cleavable N-terminal 6-His-MBP tag. After several trials to optimise ECWaaP, SEnWaaP and PAWaaP expression and purification, favourable results were only obtained for PAWaaP (see 6.3.4). Henceforth, it was decided to test other WaaP homologues for structural biology

studies. Therefore, CFWaaP, PFWaaP and PSWaaP gene inserts (see section 1.2.1 and **Table 1.1**) were amplified for cloning into the pLou3 expression vector. Of these, only the expression and purification of CFWaaP is reported in this thesis; PFWaaP and PSWaaP are yet to be tested.

All gene inserts were produced with two different restriction sites to promote unidirectional ligation into the cloning site of their respective vectors. After the transformation of competent *E. coli* cells with the recombinant plasmids (described in 1.2.3), up to four colonies were selected for plasmid extraction (see 1.2.4) and for double restriction analysis (see 1.2.4) to screen for ligation success.

Figure 6.2 below shows DNA-agarose gels of the analytical double-digestion. A clear single band at the expected number of base pairs for each insert denoted a successfully cloned recombinant plasmid; (EC, SEn and CF WaaP inserts were expected to be around 795 bp long and PA, PF and PS WaaP inserts around 804 bp long). Recombinant plasmids with visible gene inserts were subjected to further verification by Sanger sequencing at Source BioScience. The absence of the insert in restriction digests of some plasmids indicated that the vector sample may not have been restricted sufficiently prior to ligation, causing increased vector background.

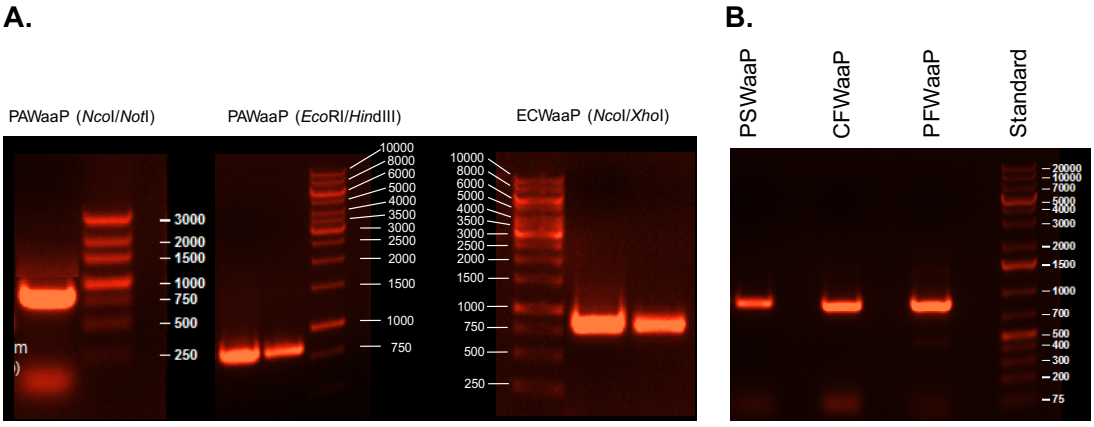


Figure 6.1: DNA agarose gel images of WaaP PCR products.
A. Agarose gel showing PCR amplified gene fragments of PAWaaP (NcoI/NotI), PAWaaP (EcoRI/HindIII) and ECWaaP (NcoI/XhoI), for ligation into pEt28b(+), pLou3 and pET28b(+), respectively. **B.** Agarose gel showing PCR amplified gene fragments of CFWaaP, PFWaaP and PSWaaP all with BamHI/HindIII restriction sites, for ligation into pLou3 plasmid.

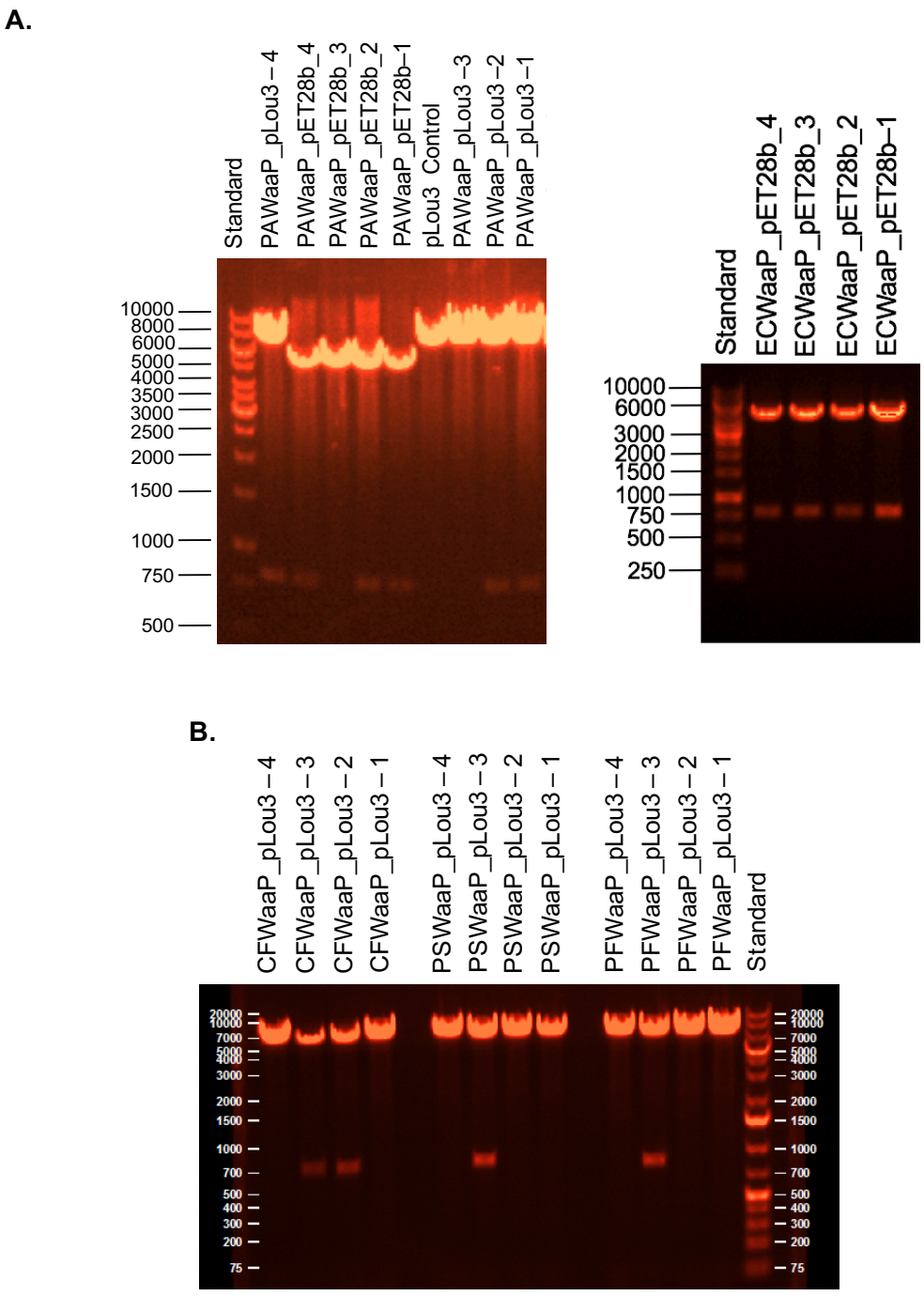


Figure 6.2: DNA agarose gel images showing double-restriction analysis of cloned WaaP plasmids.

Agarose gel analysis of: **A.** PAWaaP_pET28b, PAWaaP_pLou3 and ECWaaP_pET28b plasmids after double restriction analysis, the WaaP gene fragments are observed at the expected base pair length of 804 bp for PAWaaP and 795 bp for ECWaaP. **B.** CFWaaP, PSWaaP and PFWaaP in pLou3 after double restriction analysis, the WaaP gene fragments are observed at the expected base pair length of 795 bp for CFWaaP and 804 bp for PSWaaP and PFWaaP.

6.3 Purification and Characterisation of WaaP Proteins

Subsection synopsis: 6His-tagged ECWaaP [7] and PAWaaP [346] purifications are reported in the literature, and the experimental conditions described in 6.3.1 were adapted from the methods therein. Sufficient yields of WaaP could not be derived, so a new strategy was adopted to express EC, SEn and PA WaaP with a solubility enhancing MBP-fusion partner; although this strategy immensely improved protein yields, soluble aggregate formation was found to be a persistent issue, which could not be resolved even by altering buffer conditions. Buffers used for the purification of WaaP proteins are provided in **Table 1.4** above. Purification trials of 6His-tagged and MBP-fusion proteins for ECWaaP, SEnWaaP and PAWaaP were carried out in parallel and are presented in sections 6.3.1 to 6.3.3. The findings from these trials provided insights towards subsequent efforts to improve WaaP expression for structural studies, which are described from section 6.3.4 onwards. Successful results were obtained for the expression of PAWaaP as an MBP-fusion protein, together with Gro*ESL* chaperonin co-expression to assist folding, in a novel *E. coli* expression strain called SoluBL21(DE3). Average yields of MBP-free, monomeric PAWaaP with this method were 0.5 mg/L of culture, at >90% purity. Crystallisation trials with PAWaaP failed to provide diffraction quality crystals.

6.3.1 Purification of C-6His-ECWaaP: insufficient for crystallographic studies

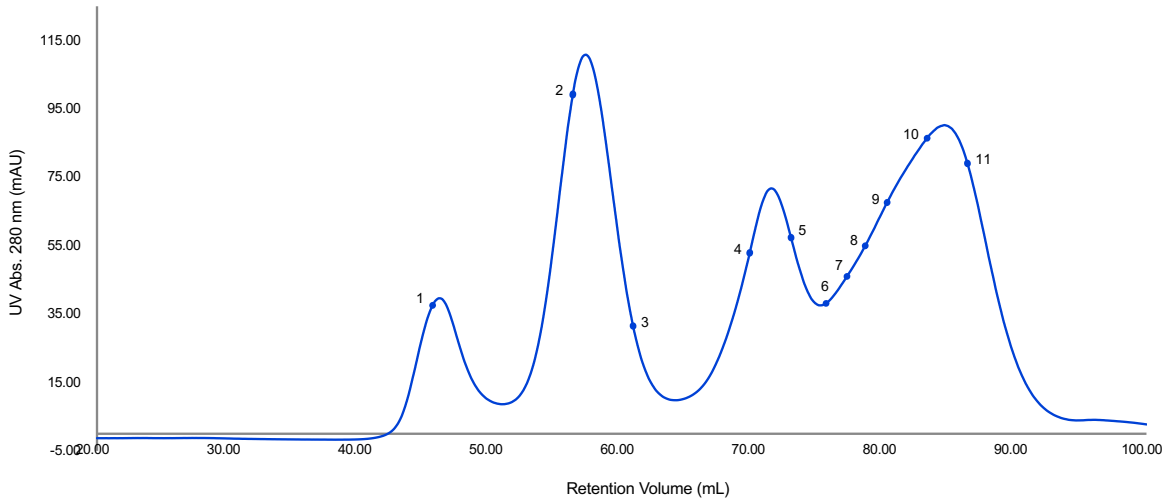
An ECWaaP expression and purification procedure is reported by Yethon and Whitfied [7], which was optimised after testing the usual well known variables for increasing protein yield, such as temperature, growth medium, time and duration of induction, inducer concentration(s), and expression from different promoters, none of which enhanced the yield of soluble ECWaaP. However, an improvement in yield was achieved by co-expressing ECWaaP with the solubility enhancing GroES and GroEL *E. coli* chaperones, which were cloned independently into the pBAD33 vector to

produce the pWQ911 plasmid [7]. Differences in origin of replication, antibiotic resistance marker and promoter between the pWQ911 and ECWaaP expression plasmids (see **Table 1.2**) allowed both plasmids to be maintained simultaneously in the *E. coli* cell, while controlling expression from each independently. In said method, ECWaaP was purified under native conditions by Nickel based IMAC, followed by subtractive anion exchange chromatography (AIC) (ECWaaP has a pI of 9.8). ECWaaP yield was not given, but purity was reported as >95%, with a 25.8-fold increase in specific activity of ECWaaP compared to the crude, cell-free extract.

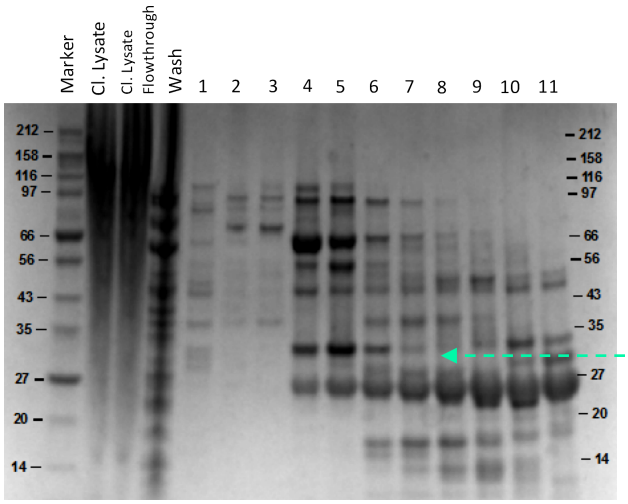
In this investigation the C-6His-ECWaaP_pET28b and pWQ911 plasmids were co-transformed into the C43(DE3) *E. coli* strain and protein was produced as described in section 1.3.1.2. C-6His-ECWaaP was purified by IMAC in the first instance, but instead of AIC, the eluted His-tagged protein was purified by SEC. In SEC biomolecules are separated by size, and it can also resolve various oligomeric forms of the same protein. It is often used as a final purification step to obtain highly pure and monodisperse proteins, suitable for crystallisation [347], and in the Dong group is the usual final purification step known to achieve proteins ideal for structural studies [6,269,348-350]. An SEC chromatogram and an SDS-PAGE gel dislaying the purification of C-6His-ECWaaP are given in **Figure 6.3**.

Bands attributed to C-6His-ECWaaP at about 31 kDa are indicated on the SDS-PAGE gel, and it can clearly be seen that its yield is extremely low and it has co-eluted with several other impurities. The band for ECWaaP appears in SEC fractions 4-6, which given the retention volume could be an oligomeric form of the protein, and in fractions 10-11 where it is most likely to be a monomeric form. The results from this trial clearly suggest that this method could not be replicated as in [7] and is not suitable to obtain ECWaaP for crystallography, so no further method development was performed. However, it must also be remembered that in this case, a larger scale of WaaP was required and even if ECWaaP could have been highly purified, its yield may not have been sufficient and economic in the long term for crystallographic studies.

A.



B.



The protein bands appearing at about 31 kDa were expected to be due to ECWaaP, and its appearance in two SEC peaks indicates different oligomeric forms of ECWaaP.

Figure 6.3: C-6His-ECWaaP IMAC and SEC purification results

A. SEC chromatogram of C-6His-ECWaaP after purification by IMAC. Numbered points corresponding to the center of fractions prepared for SDS-PAGE analysis are shown in **B**. **B.** SDS-PAGE gel showing results of C-6His-ECWaaP purification at selected stages of IMAC purification, followed by selected fractions from SEC.

6.3.2 Purification of C-6His-PAWaaP: improved yields of protein but inadequate for crystallography

C-6His-PAWaaP was purified by Nickel based IMAC followed by SEC purification of the eluted fractions, the results for which are in **Figure 6.4**. Purification of C-6His-PAWaaP was carried out more than once for optimisation. The yields were usually 0.4 to 0.8 mg/L of culture, which was deemed to be sufficient to work with for initial small-scale crystallisation trials. However, an impurity at about 25 kDa was consistently seen to be eluted with PAWaaP during all purifications, giving it an overall purity of 40-60%, and it could not be separated from PAWaaP even after SEC. Upon analysing the SDS-PAGE bands at 31 kDa and 25 kDa by MALDI-MS (see **Figure 6.4**), the results for which are given in **Table 6.1**, the former was identified as containing peptide sequences from PAWaaP, and the latter was identified as an endogenous *E. coli* impurity. It was therefore concluded that this, particular, expression and purification strategy for WaaP crystallisation was not viable for two reasons: 1) yields of C-6His-PAWaaP per expression (typically 12L cultures) and purification cycle were not sufficient for prolonged structural biology studies and, 2) The quality of C-6His-PAWaaP being compromised due to the consistent co-purification of an endogenous *E. coli* impurity with the target protein.

This impurity could not even be eliminated with a more stringent wash routine, involving higher concentrations of imidazole in the wash buffer and while the introduction of more stringent imidazole washes, and a step-wise elution gradient (such as from 100 mM to 250 mM imidazole, in 50 mM increments) could likely enhance PAWaaP purity, it would be offset by the greatly diminished yields of PAWaaP. Looking back at the literature method by Zhao and Lam [346], which describes the purification of C-his tagged PAWaaP, a single step procedure using cobalt based IMAC afforded PAWaaP at a yield of 0.5 mg per litre of culture, at >90% purity. Co^{2+} resins have been reported to have a lower affinity for polyhistidine tags and less non-specific protein binding compared to Ni^{2+} resins, enabling the elution of His-tagged proteins under milder conditions and in purer form [351]. It is not as

commonly used as Ni-NTA IMAC, unless required, and so retrospectively, the use of a Co^{2+} based resin in the original method should have indicated the behaviour of C-6His-PAWaaP during purification – as a protein that is expressed in relatively low yields and one that is likely to be eluted with non-specific proteins. The next logical step towards improving PAWaaP purification would have been to switch to Co^{2+} IMAC, however, a development in the MBP-PAWaaP purification procedure diverted the focus of the investigation, and is discussed in the following section (6.3.4).

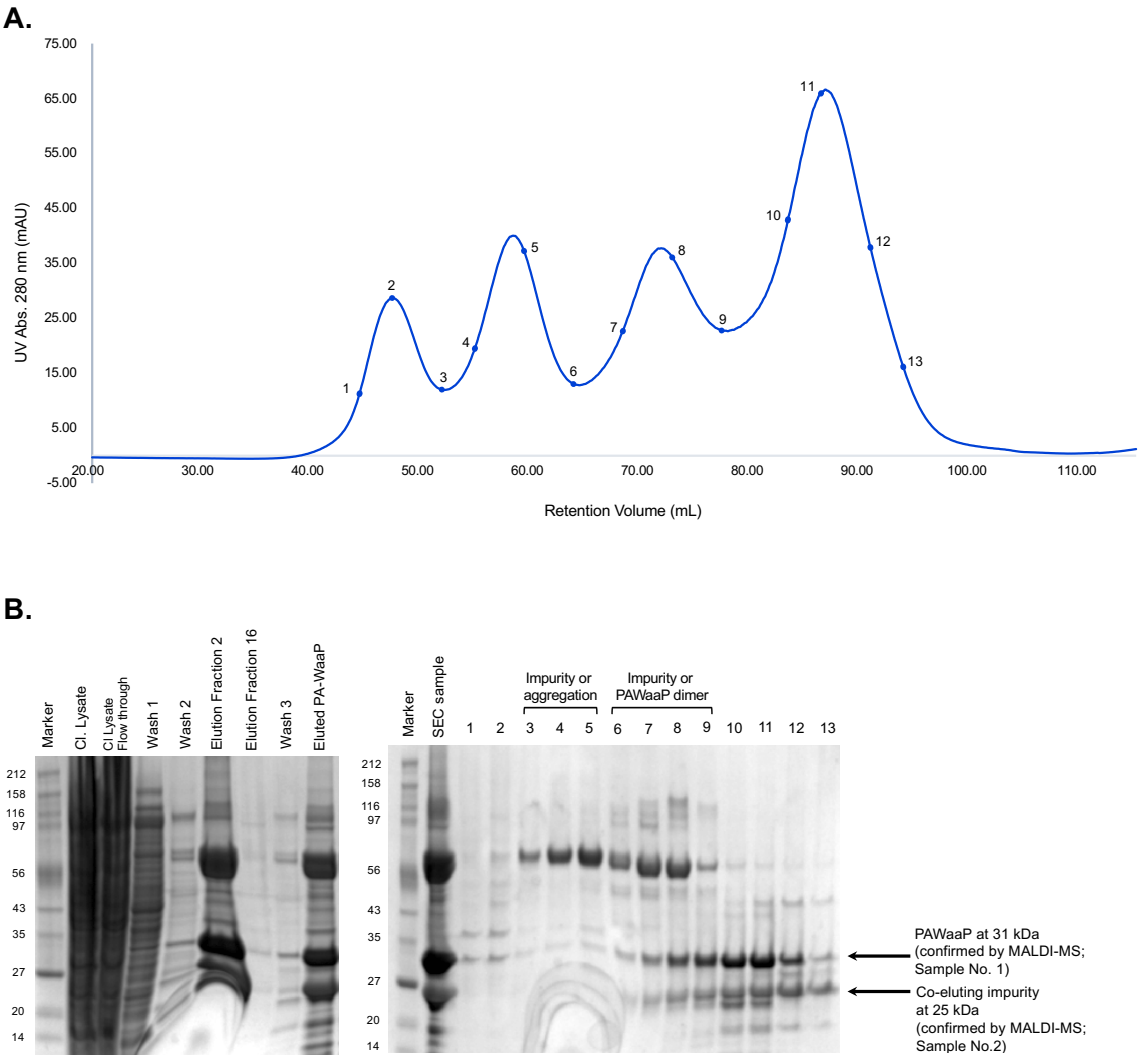


Figure 6.4: C-6His-PAWaaP IMAC and SEC purification results

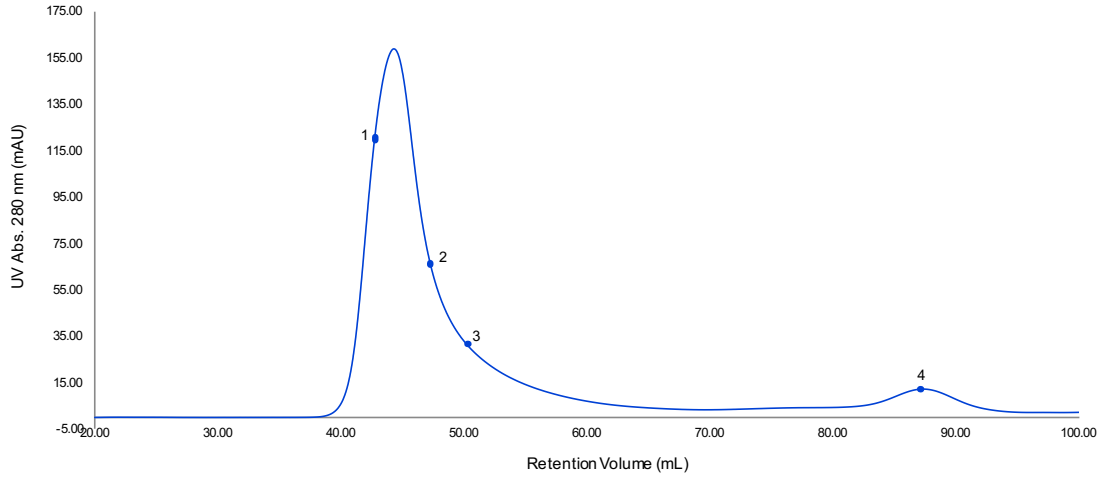
A. Typical SEC chromatogram of C-6His-PAWaaP after IMAC. The numbered points correspond to the center of fractions chosen for SDS-PAGE analysis, as indicated in **B** below. **B.** SDS-PAGE gel showing results of a typical C-6His-PAWaaP purification. Gels indicate the protein composition of samples from selected stages of IMAC, followed by selected SEC fractions. MALDI-MS results of the indicated bands are given in **Table 6.1**

6.3.3 Purification of WaaP as MBP fusion proteins

The drawback with WaaP expression and purification as His-tagged protein was low expression yields, which made isolation of the target protein with high purity quite challenging. In the reported methods for ECWaaP [7] and PAWaaP [346] purification, the authors remarked that most of the protein was found in the insoluble cell fraction. Therefore, to increase the expression yield of well folded, soluble WaaP, orthologues from *E. coli, S. enterica* and *P. aeruginosa* were cloned into the pLou3 plasmid, enabling the production of WaaP as 6-His-tagged MBP fusion proteins. The *E. coli* maltose-binding protein (MBP) is a popular solubility enhancing fusion partner, known to significantly improve the folding and expression of passenger proteins that otherwise express poorly, or aggregate in the host expression system [352,353]. It is also utilised as a purification tag via immobilised amylose affinity chromatography, however, the binding capacity of an amylose resin for MBP is lower than for other known affinity tags. Thus, the inclusion of a 6-His sequence N-terminal of MBP permits the purification of MBP-fused proteins by IMAC in higher yields [352,354]. MBP as an N-terminal fusion partner has been found to be more beneficial for enhancing solubility than as a C-terminal fusion [355].

Indeed, yields of the above mentioned WaaP homologues improved immensely when expressed with an MBP tag. While typical yields for C-6His-PAWaaP were 0.4 to 0.8 mg/L of culture, at 40-60% purity (as assessed qualitatively by SDS-PAGE in **Figure 6.4**) yields for MBP-fused WaaP were in the order of 10 mg/L of culture at >95% purity. However, after MBP tag removal and subsequent purification by SEC, WaaP proteins were obtained in a highly aggregated form, as they would elute in the void volume of the SEC column; this can be seen on the chromatograms in **Figure 6.5**, **Figure 6.6**, and **Figure 6.7,** respectively. The void volume was estimated as 45.5 mL, by SEC of Blue Dextran 2000 (GE Healthcare) using the same buffer conditions (see **Figure 6.8**). It was inferred that at some stage between expression and SEC purification, WaaP had a tendency to form soluble aggregates. To produce WaaP for crystallography this challenge had to be resolved.

A.



B.

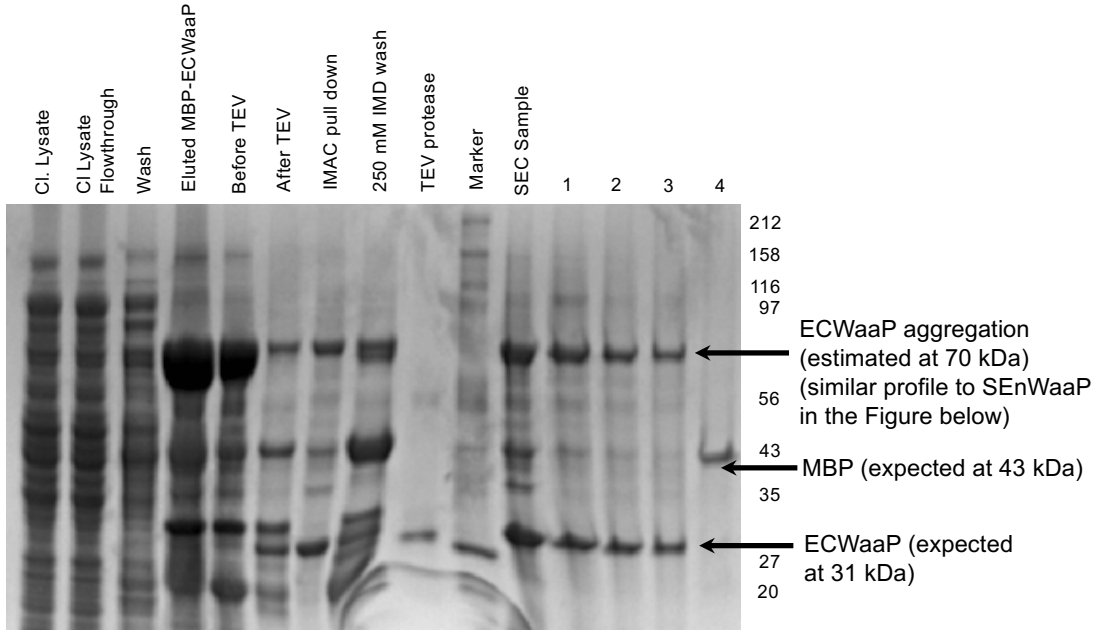
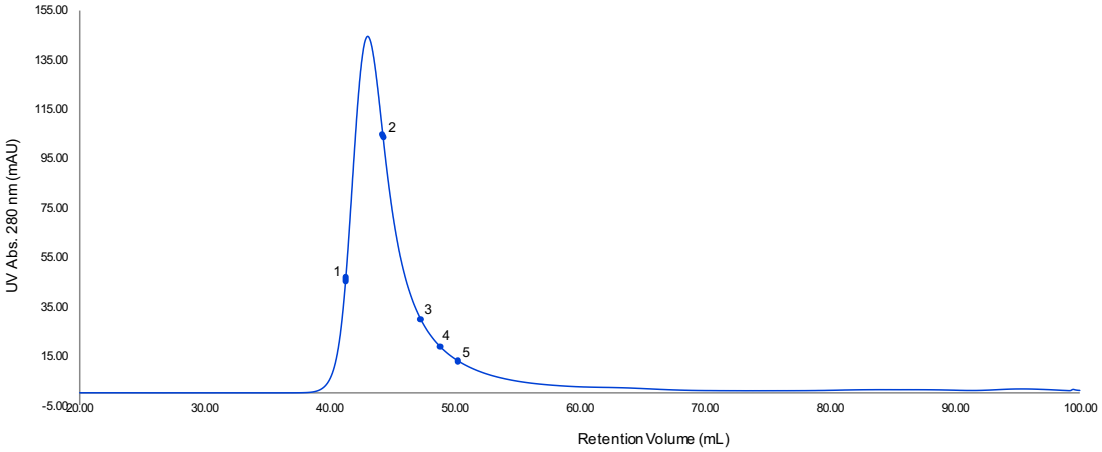


Figure 6.5: MBP-ECWaaP IMAC and SEC purifications

A. SEC purification of ECWaaP after proteolytic removal of MBP fusion protein and subsequent pull down by IMAC. The numbered points correspond to the center of the fractions prepared for SDS-PAGE analysis, as indicated in Figure **B** below. **B.** SDS-PAGE gel showing results of ECWaaP purification at selected stages of IMAC, followed by selected fractions from SEC.

A.



B.

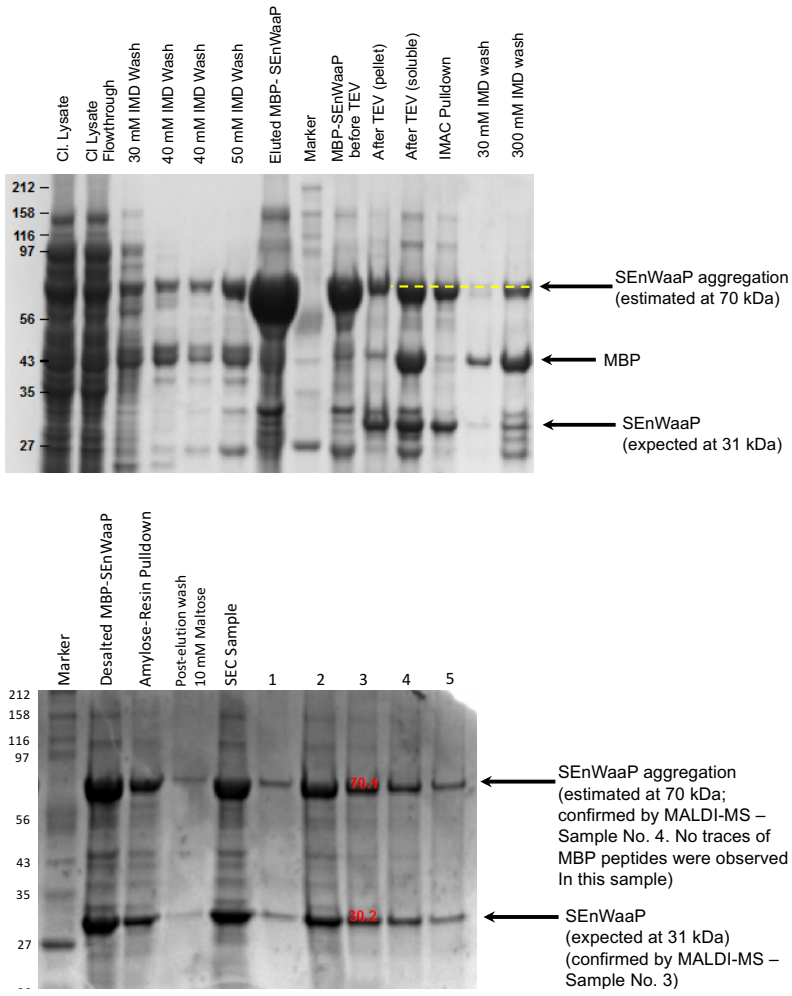


Figure 6.6: MBP-SEnWaaP IMAC and SEC purifications

A. SEC purification of SEnWaaP after proteolytic removal of MBP fusion protein and subsequent pull down by IMAC. The numbered points correspond to the center of fractions prepared for SDS-PAGE analysis, as indicated in **B** below. **B.** SDS-PAGE gel showing results of SEnWaaP purification at selected stages of IMAC, followed by selected fractions from SEC. MALDI-MS results of the labelled bands are provided in **Table 6.1**

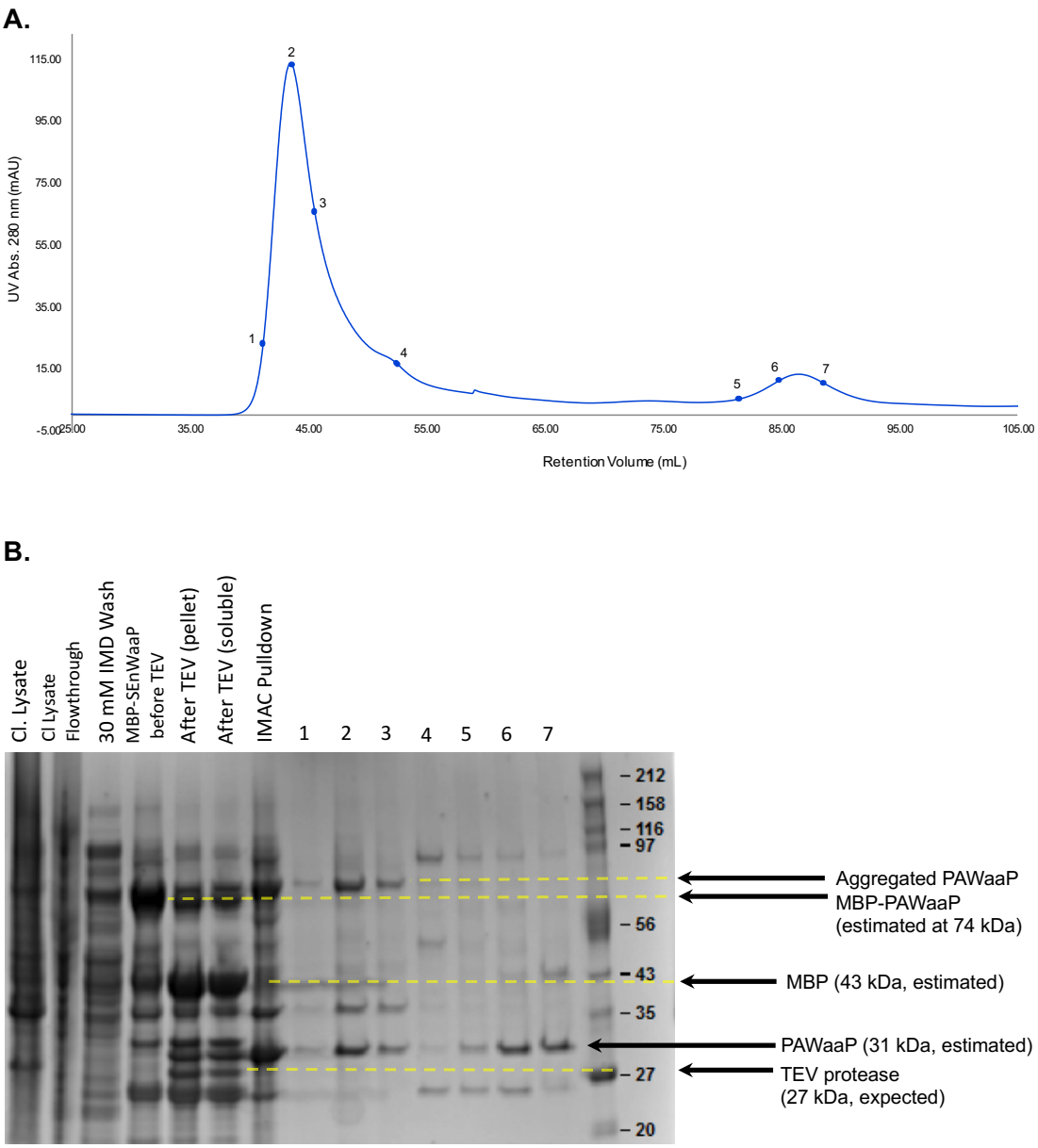


Figure 6.7: PAWaaP IMAC and SEC purifications

A. SEC purification of PAWaaP after proteolytic removal of MBP fusion protein and subsequent pull down by IMAC. The numbered points correspond to the center of fractions prepared for SDS-PAGE analysis, as indicated in Figure **B** below. **B.** SDS-PAGE gel showing results of PAWaaP purification at selected stages of IMAC, followed by selected fractions from SEC.

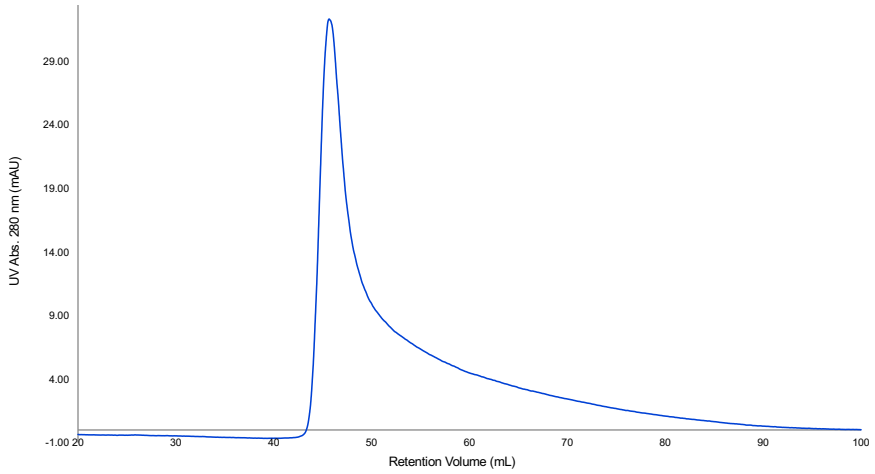


Figure 6.8: SEC Chromatogram of Blue Dextran 2000

With a MW of approximately 2000 kDa, Blue Dextran was used as a marker to determine the void volume of the HiLoad 16/600 Superdex 200 Column. The maximum of the peak corresponded to a retention volume of 45.5 mL, so this was considered the void volume of the column.

Table 6.1: Results of MALDI-MS analyses of selected protein bands from WaaP purification trials

Sample No.	Ref.	Protein Name	Uniprot No.	Peptides	Coverage	Mascot score
1	**Figure 6.4**	LPS core heptose (I) kinase RfaP; *Pseudomonas aeruginosa* PAO1	Q9HUF7	16	51%	138
2*	**Figure 6.4**	Gamma-butyrobetaine,2-oxoglutarate dioxygenase	F1ZCH4	4	12%	58
3	**Figure 6.6**	LPS core biosynthesis protein RfaP; *S. enterica*	E7Y039	23	65%	172
4**	**Figure 6.6**	LPS core heptose (I) kinase rfaP; *S. enterica*	H6NWS6	16	47%	118

* The band at 25 kDa was not attributed to PAWaaP, rather the peptides showed a limited match to existing proteins in the Uniprot database, indicating it could be an unknown impurity.

** No traces of *E. coli* MBP peptides could be observed in the band at 70 kDa, indicating the the band was not due to un-cleaved MBP-SEn, but rather a high MW SEnWaaP aggregate.

6.3.3.1 Attempts at suppressing WaaP aggregation during purification

Similar SEC purification profiles were observed for EC, SEn and PA WaaP derived from MBP-fused chimeras. After the purification of SEn-MBP (**Figure 6.6**), SDS-PAGE analysis of SEC fractions representing a single, sharp peak between 40-50 mL retention volume was carried out, revealing two bands at approximately 74 kDa and 31 kDa (shown as samples 1 to 5 in **Figure 6.6**). MALDI-MS analysis of these bands from SEnWaaP purification (see **Table 6.1**) revealed that the band at 31 kDa belonged to SEnWaaP, as expected, but so did the band at 74 kDa; no traces of *E. coli* MBP were found in the 74 kDa band, which is the expected MW of *E. coli* MBP. Given that SEnWaaP was purified as a single peak in SEC, eluting at a retention volume characteristic of high MW oligomers, it seemed likely that the two bands were by-products of the reducing conditions in the SDS-PAGE loading buffer (see section 1.3.5). As the purification profiles for EC and PA WaaP were similar, the same conclusions were applied to them.

MBP-SEnWaaP was subjected to preliminary purification trials with additives to find a method to lower the formation of soluble aggregates. Two types of additives were chosen, a detergent – LDAO, and an electrolytic mixture of 50 mM Arginine and Glutamic Acid (50 mM R/E), to investigate whether the aggregation of SEnWaaP was hydrophobic or hydrophilic in nature, or both.

1) Lauryldimethylamine-N-Oxide (LDAO): a zwitterionic detergent commonly used for solubilisation and crystallisation of membrane proteins [356-358]. Detergents can be of three types: ionic (cationic or anionic), zwitterionic or non-ionic. Ionic detergents, like SDS are harsh and denaturing as they effectively disrupt inter- and intra-protein interactions. Non-ionic detergents like Triton x-100 are milder and non-denaturing, as they can counter protein-lipid and lipid-lipid interactions without affecting protein-protein interactions – hence, their utility is mostly confined to solubilisation and protein extraction from membranes. Zwitterionic detergents are considered intermediate in harshness, since they are not as denaturing as ionic detergents, but are more effective at curbing inter-protein interactions compared to non-ionic detergents [359].

Thus, LDAO was first chosen for SEnWaaP purification, and added at 1.5 mM to all purification buffers, which is over 10 times the critical micellar concentration (CMC) of LDAO in aqueous NaCl (CMC in water: 1-2 mM; CMC in 0.1 M NaCl: 0.14 mM) [360].

2) 50 mM arginine/glutamic acid (50 mM R/E): An equimolar solute concentration of 50 mM R/E has been reported to have the following effects on the solution properties of proteins: increased solubility, up to a factor of 8 [361]; a decrease in protein aggregation and non-specific interactions, with the preservation of functional interactions [361]; increased thermal stability [362] and increased resistance to proteolytic degradation [361, 363]. Its use in lysis buffers to lower protein aggregation at the cell lysis stage has been investigated and reported [364]. Some explanations regarding R/E's efficacy as a stabilising and solubilising mixture include:

a) Polyelectrolyte Induced Compaction (PIC), where equimolar amounts of R/E affect disordered loops on the protein surface causing their compaction towards the relatively ordered protein core, consequently countering non-specific interactions between disordered regions on the protein surface [363];

b) The synergistic action of arginine and glutamic acid molecules, bridged via hydrogen-bonds, which then form clusters around the protein surface. This is postulated to cause a 'crowding effect' which suppresses protein aggregation. It was also proposed that the effectiveness of equimolar R/E mixtures in lowering protein-protein interactions would depend on protein surface composition – those with a charged and aromatic amino acid surface composition of 44-58% showing most enhanced suppression of aggregation [365].

Purification trials for SEnWaaP were carried out with under three buffer conditions, which are summarised in **Table 6.2** below. The lowered elution yields for MBP-fused SEnWaaP in these experiments, compared to those observed in the first purification (see **Figure 6.6**) can be explained due the the smaller *E. coli* cell quantities used for each purification trial (4 L versus 12 L cultures).

Table 6.2: SEnWaaP purification results under varying buffer conditions.
Compositions for the complete set of buffers are given in Table 1.4 in section 1.3.3.

1.5 mM LDAO in all buffers	1.5 mM LDAO and 50 mM R/E in all buffers	50 mM R/E in all buffers
MBP removal was not successful; no SEnWaaP obtained in IMAC pull down; Unknown if SEnWaaP aggregation was lowered		MBP tag was removed by TEV protease; SEC chromatogram shows that SEnWaaP is still aggregated
Figure 6.9	**Figure 6.10**	**Figure 6.11**

LDAO in purification buffers did not help to increase the yield of soluble, monomeric WaaP protein. Contrarily, the presence of LDAO seemed to have a negative effect on TEV-protease's activity as the absence of a band due to WaaP at 31 kDa can clearly be noted in **Figure 6.9** and ***Figure 6.10*** together with the absence of any protein in the IMAC pull down and SEC chromatograms. LDAO could therefore be ruled out as an appropriate buffer additive for SEnWaaP purification. Comparing this to the purification of SEnWaaP with only 50 mM R/E in the buffers, signs of protein aggregation in the IMAC pull down and during SEC purification remain. However, this trial experimentally confirms the negative effect of LDAO on TEV-protease activity, as the appearance of a new band at roughly 31 kDa in the 'After TEV' and 'IMAC pull down' lanes of the SDS-PAGE gel in **Figure 6.11** indicates that the MBP tag has been cleaved from SEnWaaP. Indeed, in two independent studies [366,367], the presence of LDAO in purification buffers was shown to partially or entirely inhibit TEV-protease's action on fusion tag cleavage. It was proposed that this could occur due to to steric obstruction of the cleavage site by the protein or detergent in the protein–detergent complex (PDC), or due to an inhibitory effect of LDAO upon protease activity [366].

A purification trial with DTT in the SEC stage showed a slight change in the profile of oligomeric species/soluble aggregates (see SEC chromatogram in **Figure 6.12A**), indicating that the cysteine residues on WaaP may have some degree of contribution to WaaP aggregation, however, the lack of an apparent monomeric SEnWaaP peak in the SEC chromatogram meant that purification under reducing conditions was not sufficient to curb aggregation entirely.

Figures A. SEC purification of SEnWaaP after proteolytic removal of MBP fusion protein and subsequent pull down by IMAC. The numbered points correspond to the center of fractions prepared for SDS-PAGE analysis. **B.** SDS-PAGE gel showing results of SEnWaaP purification.

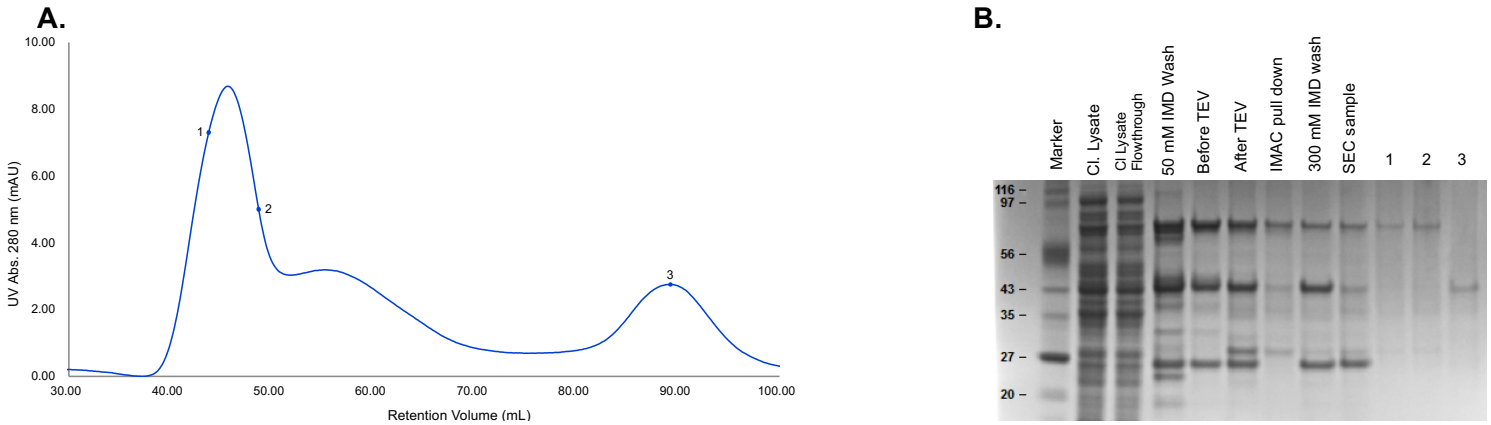


Figure 6.9: SEC Chromatogram and SDS-PAGE gel image from purification trial of MBP-SEnWaaP with 1.5 mM LDAO in buffers

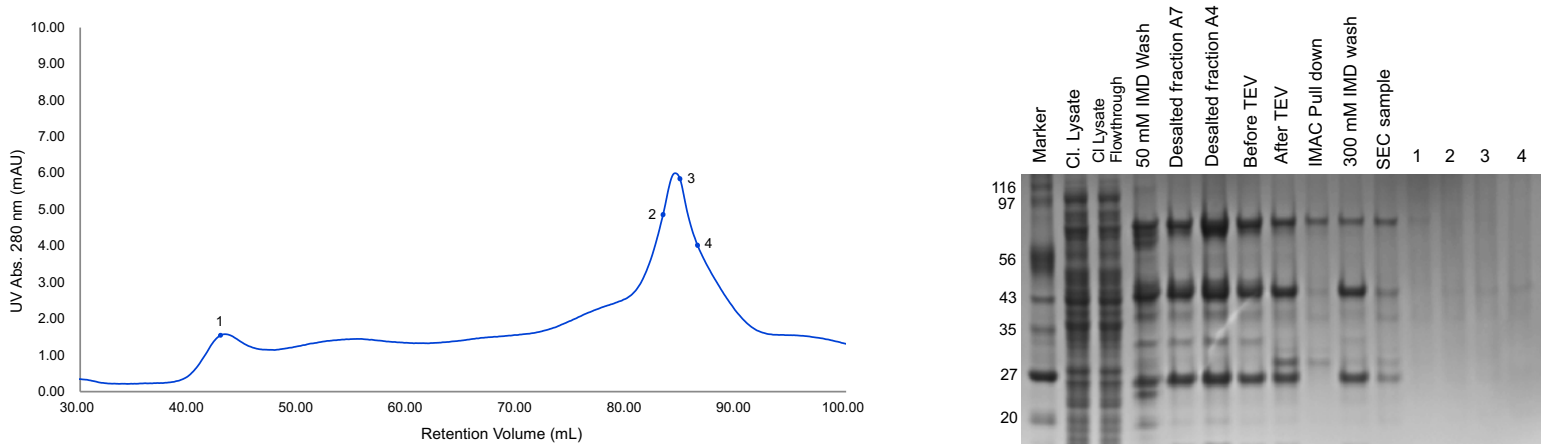


Figure 6.10: SEC Chromatogram and SDS-PAGE from purification trial of MBP-SEnWaaP with 1.5 mM LDAO and 50mM R/E in buffers

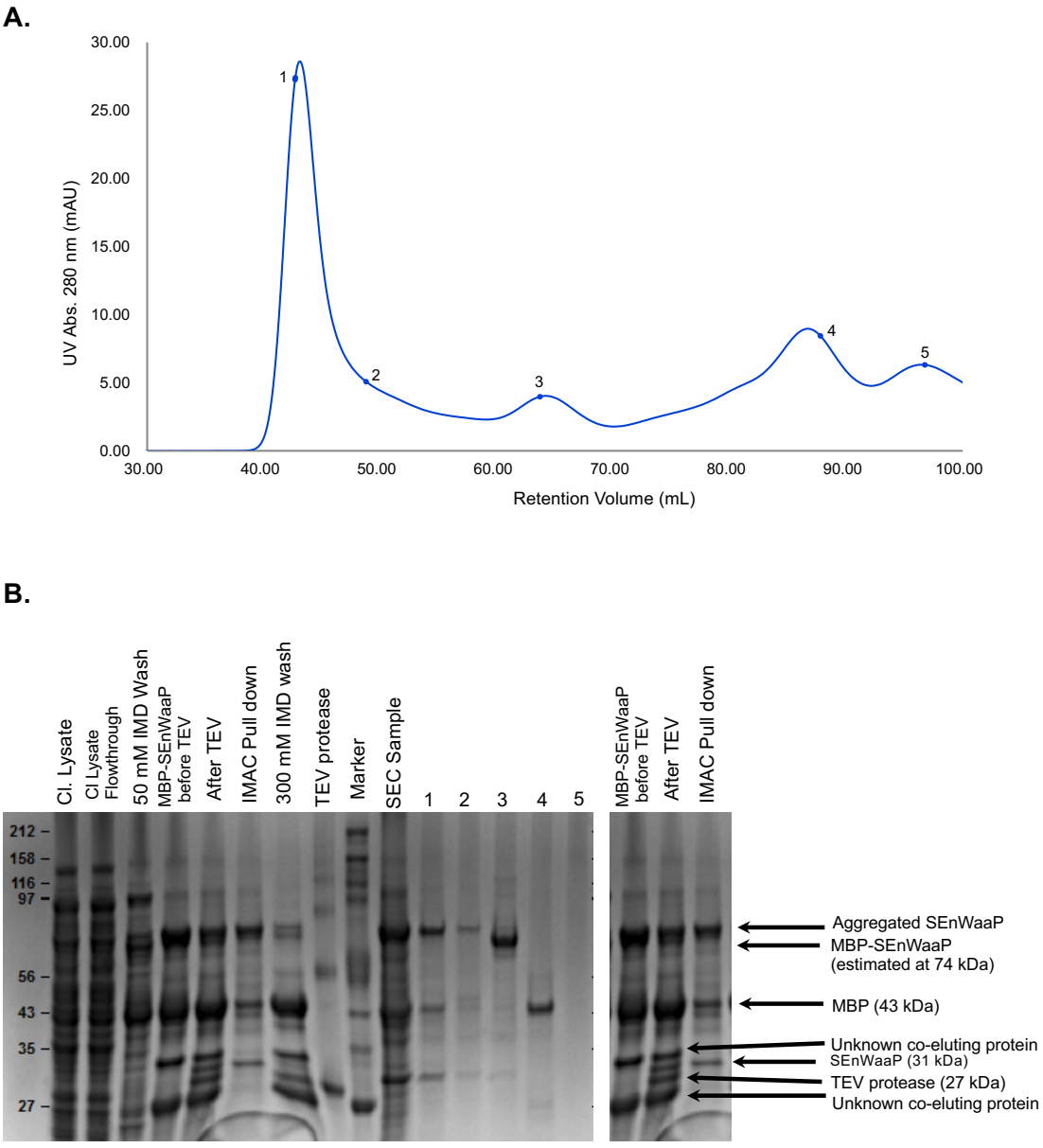
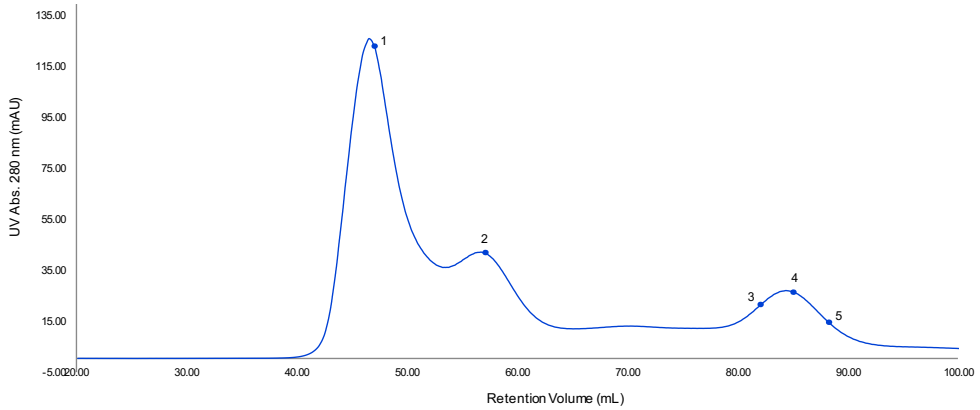


Figure 6.11: SEC Chromatogram and SDS-PAGE gel image from purification trial of MBP-SEnWaaP with 50mM R/E in selected buffers.

A. SEC purification of SEnWaaP after proteolytic removal of MBP fusion protein and subsequent pull down by IMAC. The numbered points correspond to the center of fractions prepared for SDS-PAGE analysis. **B.** SDS-PAGE gel showing results of SEnWaaP purification at selected stages of IMAC, followed by selected fractions from SEC.

A.



B.

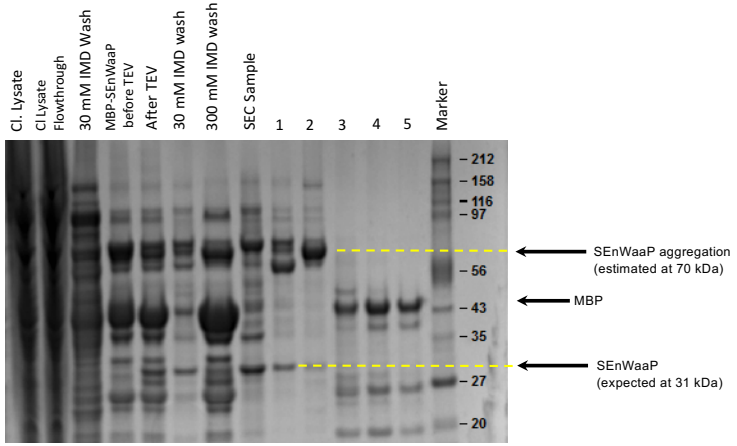
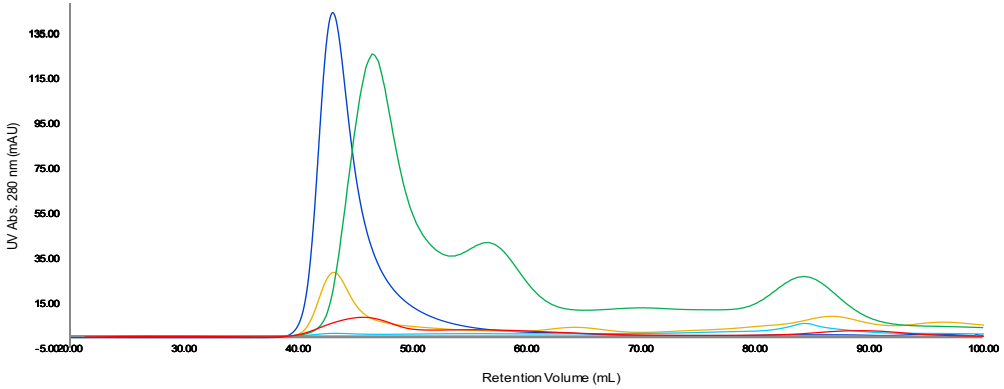


Figure 6.12: SEC Chromatogram and SDS-PAGE gel image from purification trial of MBP-SEnWaaP with DTT in SEC buffer.

A. SEC purification of SEnWaaP after proteolytic removal of MBP fusion protein and subsequent pull down by IMAC. The numbered points correspond to the center of fractions prepared for SDS-PAGE analysis in Figure **B,** showing results of SEnWaaP purification.



Blue: (from ***Figure 6.6***) without buffer additives
Red: (from **Figure 6.9**) with 1.5mM LDAO in buffers
Turquoise: (from ***Figure 6.10***) with 1.5mM LDAO and 50 mM R/E
Yellow: (from ***Figure 6.11***) with 50 mM R/E in buffers
Green: (from ***Figure 6.12***) with 2.5 mM DTT in SEC buffer only

Figure 6.13: Overlay of SEC Chromatograms for all SEnWaaP Purification Trials

It seemed very likely that MBP-WaaP was being produced as soluble aggregates within the *E. coli* cells during expression, to the extent that even after MBP tag removal by linker proteolysis, the WaaP proteins were obtained exclusively as high molecular weight (MW) oligomers, as evidenced by the SEC chromatogram profiles in **Figure 6.5**, **Figure 6.6** and **Figure 6.7**. This proposal would not be an unusual one, as a few studies have reported MBP-induced *in vivo* soluble aggregate formation, often resulting in the passenger protein precipitating, or losing activity upon removal of the MBP tag [368-372]. Different aspects of the challenges in protein purification that were discussed in these studies bore resemblance to the problem at hand, and so, provided useful insights towards developing an expression and purification protocol to obtain well folded, monomeric WaaP (discussed in 6.3.4).

For example, one such study describes the purification method development for E6 proteins from eight Human Papilloma Virus (HPV) species, as MBP fused chimeras. During purification trials these fusion proteins were observed to form large multimeric aggregates, the origins of which were proposed to be: a) *in vivo* aggregation during *E. coli* expression and, b) *in vitro* inter-molecular disulphide bridge formation during protein extraction and purification [371,372]. Mutating the non-conserved cysteine residues into serine eliminated intermolecular disulphide cross-linking, and slightly improved the yield of monomeric MBP-E6, however, the levels of soluble aggregates amongst the E6 proteins still varied between 64 to 94%, implying that the greater cause of aggregation would have been at the MBP-E6 expression stage.

Variations in soluble aggregate levels that were observed for the eight homologues did not correlate with percent surface hydrophobicity, which implied that the surface properties of the protein may not have driven large MW aggregate formation. Modification of expression conditions was suggested as the foremost way to troubleshoot this issue [371]. These could include screening for an appropriate *E. coli* expression strain, suitable induction temperature, duration and inducer concentration, and the use of *in vivo* chaperones to assist protein folding; these suggestions were adopted in the ensuing investigations.

6.3.4 Purification of PAWaaP and ECWaaP MBP fusion proteins from *E. coli* SoluBL21(DE3) with the Gro*ESL* chaperonin

Haacke et. al. [373] have investigated chaperone overexpression as a strategy to boost yields of ten tyrosine and tyrosine-kinase like proteins. They found that the kinases in their experiments behaved in three ways: first, those whose yields were unaffected by chaperone over-expression; second, those with increased yields upon chaperone over-expression, but mainly as soluble aggregates; third, kinases whose yields in the monomeric form were doubled, but were partially complexed with the chaperones. The results showed partial, case-specific improvements on quality and yield of protein kinases, but required further development.

Yethon and Whitfield [7] have previously described a protocol to co-express His-tagged ECWaaP with *E. coli* Gro*ES* and Gro*EL* chaperones, that were expressed from another plasmid – pWQ911 (see section 1.1.1.2). This protocol was attempted for C-6His-ECWaaP but without success (results in 6.3.1). However, it was proposed that if chaperone co-expression was combined with the alteration of another expression condition, improved yields of well folded WaaP could possibly be attained. An *E. coli* strain called SoluBL21™ (Genlantis, California, USA) was selected to test whether it is a suitable strain for improved MBP-WaaP expression with Gro*ESL* co-expression. According to the manufacturer's description, SoluBL21™ is a mutant strain of BL21(DE3) *E. coli* that was developed using a directed evolution approach, to enhance expression for toxic and insoluble proteins that were otherwise poorly expressed in the parent BL21(DE3) strain [374]. A few published studies [375-377] have used SoluBL21™ to improve protein expression quality and yields, one of which reports a crystal structure [377].

In the first instance, ECWaaP_pLou3 and PAWaaP_pLou3 were selected for co-transformation with pWQ911 into SoluBL21™. A rough indication of the relative behaviours of ECWaaP, SEnWaaP and PAWaaP proteins during purification could be gleaned from the trials conducted so far, where PAWaaP demonstrated better

behaviour during purification trials compared to EC and SEn WaaP. This can also be seen for the expression of EC and PA WaaP as C-6His proteins, where a much higher yield of monomeric PAWaaP could be obtained, whereas C-6His-ECWaaP could not be isolated at all (see sections 6.3.1 and 6.3.2). Likewise, for the expression of EC, SEn and PA WaaP as MBP-fusion proteins: almost all of the untagged WaaP protein was isolated as high MW soluble aggregates, but in the case of PAWaaP, a minute amount of protein could be isolated by SEC as a monomer, as indicated by the peak at roughly 87 mins in its SEC chromatogram and SDS-PAGE gel (**Figure 6.7**). In all the trials conducted, ECWaaP and SEnWaaP had demonstrated similar purification profiles, so for simplicity SEnWaaP was left out from this initial screen.

Indeed, the differences in ECWaaP and PAWaaP behaviour was corroborated by this batch of experiments. PAWaaP showed remarkable improvement in yield and quality; sufficient quantities of monomeric PAWaaP was successfully obtained for crystallisation trials, as indicated by the normal peak at approximately 87 minutes in the SEC chromatogram in **Figure 6.14A.** Unfortunately, this strategy did not afford the same outcome for ECWaaP expression and purification (see **Figure 6.15**) where expression based aggregation was still an unsolved problem. Furthermore, this set of experiments also indicated that there must be some key intrinsic differences in the properties of PAWaaP, compared to EC and SEn WaaP that drive these varied behaviours, which will be discussed shortly in Chapter 7.

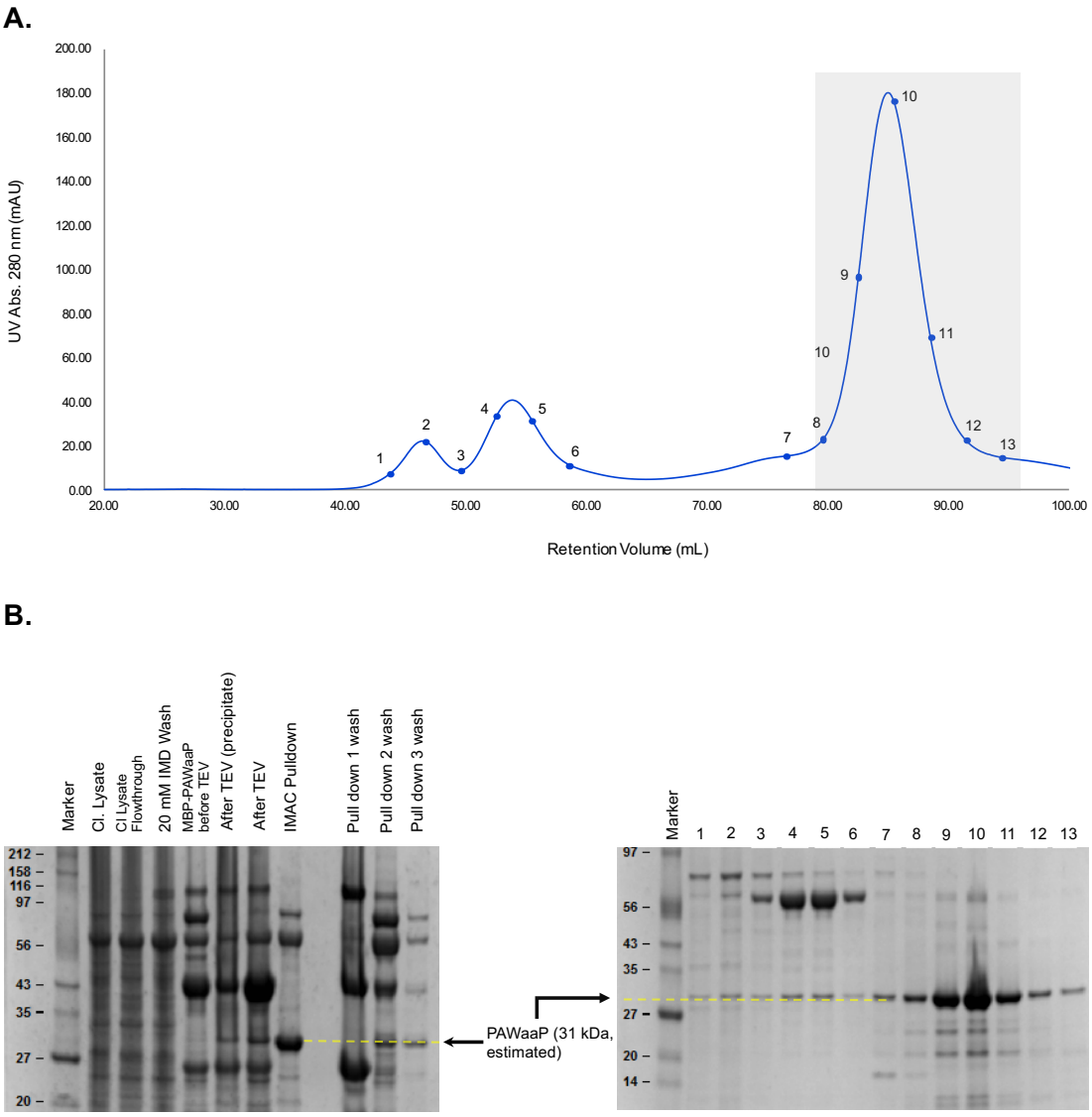
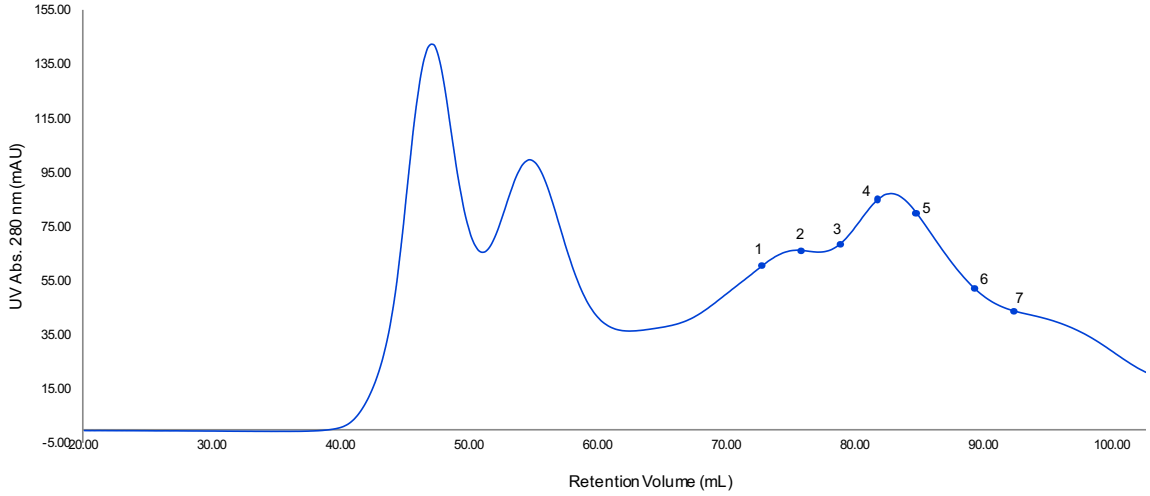


Figure 6.14: SEC Chromatogram and SDS-PAGE gel image of the purification of MBP-PAWaaP from *E. coli* SoluBL21™ with Gro*ESL* co-expression

A. A typical SEC chromatogram of PAWaaP following TEV-protease removal of the MBP tag after purification by IMAC. The numbered points correspond to the center of fractions prepared for SDS-PAGE analysis, as indicated in Figure **B** below. The grey shaded box indicate the fractions that were collected for crystallisation trials. **B.** SDS-PAGE gel showing results of SEnWaaP purification at selected stages of IMAC, followed by selected fractions from SEC.

A.



B.

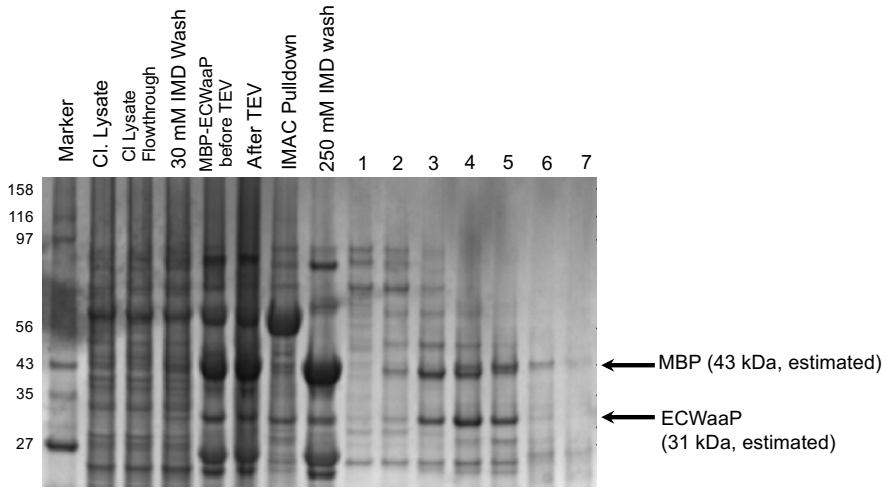


Figure 6.15: SEC Chromatogram and SDS-PAGE gel image of the purification of MBP-ECWaaP from *E. coli* SoluBL21™ with Gro*ESL* co-expression

A. SEC purification of ECWaaP after proteolytic removal of MBP fusion protein and subsequent pull down by IMAC. The numbered points correspond to the center fractions prepared for SDS-PAGE analysis, as indicated in Figure **B** below. **B.** SDS-PAGE gel showing results of ECWaaP purification at selected stages of IMAC, followed by selected fractions from SEC.

6.3.5 Crystallisation Trials for PAWaaP – No hits were obtained

Together with full length PAWaaP, the *in-situ* limited proteolysis approach (introduced in section 4.2.1) was utilised for PAWaaP crystallisation trials. To help identify suitable protease(s) for PAWaaP crystallisation a limited proteolysis screen was performed, the conditions and results of which are provided in **Figure 6.16** below. Thereafter,

PAWaaP was attempted to be crystallised in the presence and absence of its ligand – ATP, together with $MgCl_2$ as cofactor [7].

PAWaaP was concentrated to 10 mg/mL; when the extinction coefficient was applied this equated to 7 mg/mL. Crystallisation trials were dispensed by the Gryphon crystallisation robot (Art Robbins Instruments, California, USA) with a 1:1 ratio of protein to precipitant solution, giving a total drop size of 0.3 µL. The conditions and proteases for all the trials performed are given in in **Table 6.3**. No significant hits were obtained from any trial over a period of three months; most drops were clear or slightly hazy, indicating that the protein solution is undersaturated. Any drops that were unclear showed amorphous precipitate, lacking any microcrystalline growth that could have been salvaged to attempt seeding trials. These results indicated that three things could be attempted differently to obtain crystals for WaaP, in the following order:

1) Using a higher PAWaaP concentration, at around 20 mg/mL, for crystallization, as most drops were clear or slightly hazy
2) Investigating other WaaP species for purification and crystallisation, which is described below in Chapter 7.
3) Attempting purification and crystallisation of non-active mutant forms of PAWaaP.

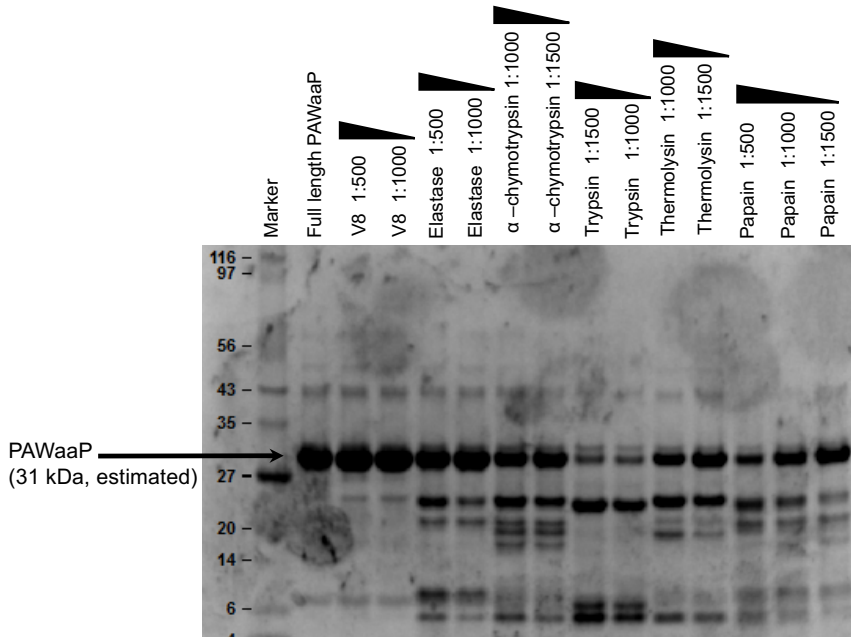


Figure 6.16: SDS-PAGE gel showing results for a limited proteolysis screen for the crystallisation of PAWaaP.

Table 6.3: Crystallisation Trials carried out for PAWaaP, co-expressed with *E. coli* GroESL chaperonin in *E. coli* strain SoluBL21(DE3)

Protein Buffer	Screens	Protease	Additives	Drop size
10 mM Tris, pH 7.5, 0.25 M NaCl, 2.5% glycerol, 5 mM DTT	PEGIon 1/2 ; JCSG+ ; CS Cryo 1/2	None	a. None b. 2.5 mM $MgCl_2$, 2 mM ATP	0.15 μL + 0.15 μL
10 mM Tris, pH 7.5, 0.6 M NaCl, 5 mM DTT	PEGIon 1/2 ; JCSG+ ; CS Cryo 1/2	a. 1:300 Thermolysin b. 1:500 Trypsin	None	0.15 μL + 0.15 μL
10 mM Tris, pH 7.5, 0.6 M NaCl, 7.5 mM DTT	PEGIon 1/2 ; JCSG+ ; CS Cryo 1/2	a. None b. 1:500 Trypsin c. 1:400 V8	a. None b. 2.5 mM $MgCl_2$, 2 mM ATP	0.15 μL + 0.15 μL
15 mM Tris, pH 7.5, 0.15 M NaCl, 2.5 mM TCEP	PEGIon 1/2 ; JCSG+; SG-1; MIDAS; MemGold 2		None	0.15 μL + 0.15 μL
15 mM Tris, pH 7.5, 0.15 M NaCl, 10% glycerol, 2.5 mM TCEP	SG-1; MIDAS; MemGold 2; Morpheus 1; Morpheus 2; PGA-LM	a. None b. 1:500 Trypsin	a. None b. 2.5 mM $MgCl_2$, 2 mM ATP	0.15 μL + 0.15 μL

6.3.6 Purification of CFWaaP MBP fusion proteins in *E. coli* SoluBL21(DE3) with the GroESL chaperonin

Other species of Gram-negative bacteria for which genomic DNA were readily available (see 1.1.1.2) were subjected to a genomic BLAST to search for the rfaP gene. From these, *C. freundii* ATCC 9080, *P. fluorescens* SBW25 and *P. syringae* DC3000 were found to carry an rfaP homologue, and these were amplified and cloned as shown in sections 0 and 1.2.2. CFWaaP was the first of the new WaaP proteins to be co-expressed with Gro*ESL* in SoluBL21™, and was purified on two occasions, as described in section 1.3.3. At both times, after O/N TEV-protease treatment for MBP tag removal, a higher amount of protein precipitation than usual was observed. Analysis of the pelleted precipitate by SDS-PAGE (**Figure 6.17B**, Lane 7) revealed a band at about 31 kDa, which is the expected MW for CFWaaP. The recovery of CFWaaP after IMAC pull down was almost nil, compared to EC, SEn or PA WaaP, as can be observed in its SEC chromatogram in **Figure 6.17 A** below. Thus, CFWaaP was ruled out as a candidate for further development. The investigation of PFWaaP and PSWaaP purification using this protocol will be pursued in the future.

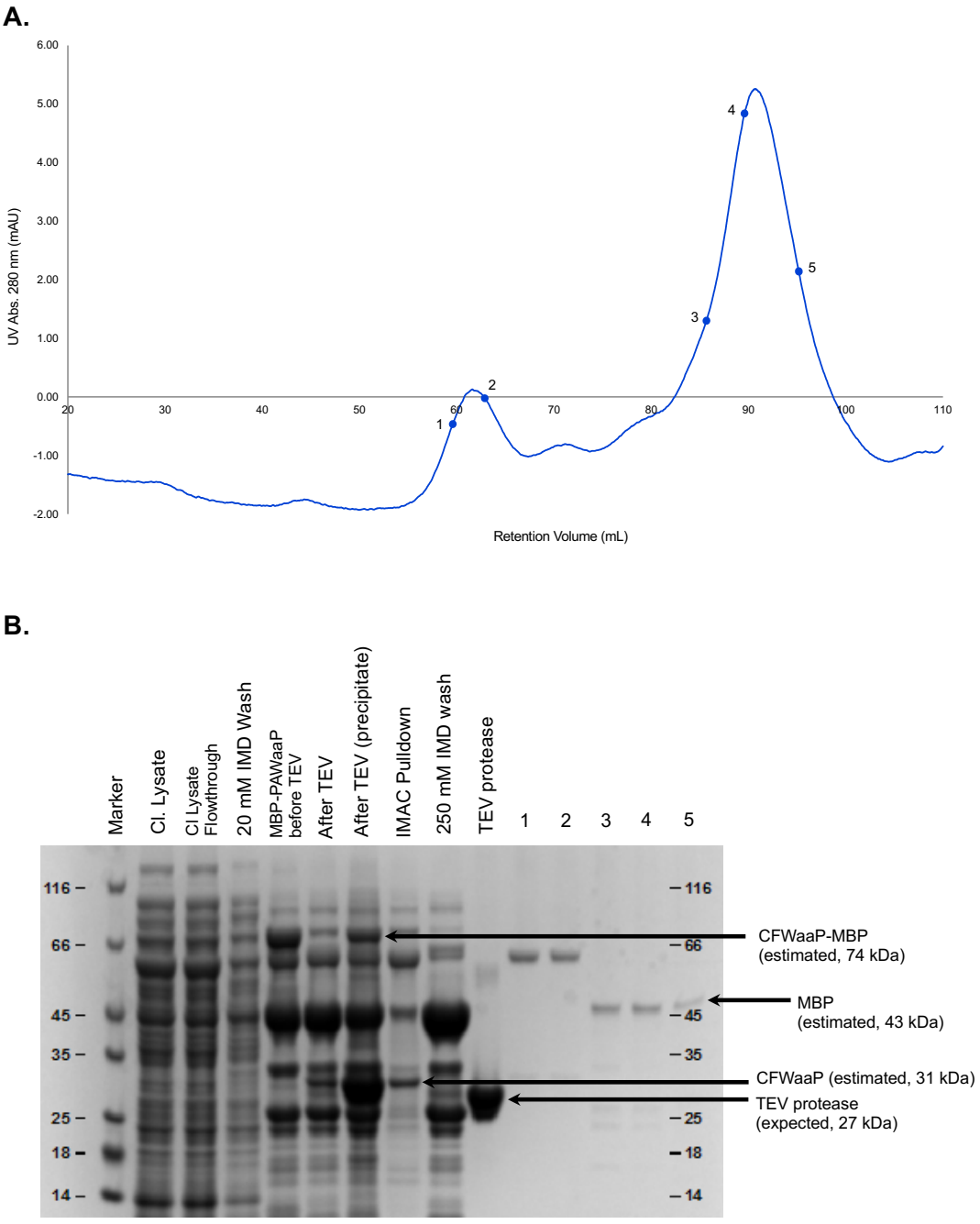


Figure 6.17: SEC Chromatogram and SDS-PAGE gel image of the purification of MBP-CFWaaP from *E. coli* SoluBL21™ with Gro*ESL* co-expression

A. SEC purification of CFWaaP after proteolytic removal of MBP fusion protein and subsequent pull down by IMAC. The numbered points correspond to the center of fractions prepared for SDS-PAGE analysis, as indicated in Figure **B** below. **B.** SDS-PAGE gel showing results of CFWaaP purification at selected stages of IMAC, followed by selected fractions from SEC.

6.4 Predictive Structural Homology of WaaP Proteins

Where an experimentally derived high-resolution structure is unavailable, predictive homology modelling is a helpful strategy to help understand the predicted folding patterns of the protein in question, by comparing it against known protein structures available in the Protein Data Bank (PDB). While not a substitute for experimental structural data, it is a useful tool towards understanding the structural homology and expected folding of the protein, providing clues to its functional mechanisms.

The templates upon which the six WaaP protein sequences were modelled show a highly conserved fold found across a range of eukaryotic protein kinases, where the N-terminal portion of the protein stacks above its C-terminal portion, held by flexible loops, which enable the kinase to 'open up' to accommodate its substrate(s). The predicted 3D models are shown below in **Figure 6.18** and **Figure 6.19**. The Kdo domain's putative catalytic sites are indicated as stick representations in the model images and show a high degree of structural conservation; individual amino acid residues conserved across a wider range of Kdo and eukaryotic kinases are highlighted in cyan. In addition to the D163 and K69 residues previously shown to be crucial for PAWaaP activity, an additional lysine residue, K51, has been identified as part of the proposed catalytic fold in PAWaaP, which was found to be strongly conserved in the predicted folds of the other five WaaP models, and also in the sequences of WaaP orthologues across Gammaproteobacteria (see APPENDIX 7), and in eukaryotic kinase structures available in the PDB.

Table 6.4: Structure homology statistics for predicted WaaP models built on Phyre[2]

Protein	Confidence (%)*	I.D. (%)	Ab initio residues	Templates (fold library ID)**
PAWaaP	89% residues at >90% conf.	12-16%	30	**c4wnkA**, **c3nyoB**, **c3qa8A**, **c4im2A**, c2acxB, **c3qa8H**
PSWaaP	99% residues at >90% conf.	12-15%	3	**c4wnkA**, **c3nyoB**, **c3qa8A**, **c4im2A**, c4g3dE, **c3qa8H**, c2fo0A
PFWaaP	87% residues at >90% conf.	13-16%	36	**c4wnkA**, **c3nyoB**, **c3qa8A**, **c4im2A**, c4yhjA, **c3qa8H**,
SEnWaaP	91% residues at >90% conf.	12-16%	23	**c4wnkA**, **c3nyoB**, **c3qa8A**, **c4im2A**, c2acxB, **c3qa8H**
ECWaaP	92% residues at >90% conf.	12-17%	20	**c4wnkA**, **c3nyoB**, **c3qa8A**, **c4im2A**, c3en9B, **c3qa8H**, c2pziA
CFWaaP	88% residues at >90% conf.	10-15%	32	**c4wnkA**, **c3nyoB**, **c3qa8A**, **c4im2A**, c3c4wB, **c3qa8H**

* Confidence is the probability of the sequence and template being homologous and is not an indication of model quality. However, a match with >90% confidence generally implies that the protein core is modelled at high accuracy (2-4 Å rmsd from the protein's true structure) and that the protein is very likely to adopt the overall fold shown. Surface loops are likely to deviate from the native structure.

** IDs shown in bold are common templates used for all six WaaP models; all of them belong to PDB structures of eukaryotic protein kinases.

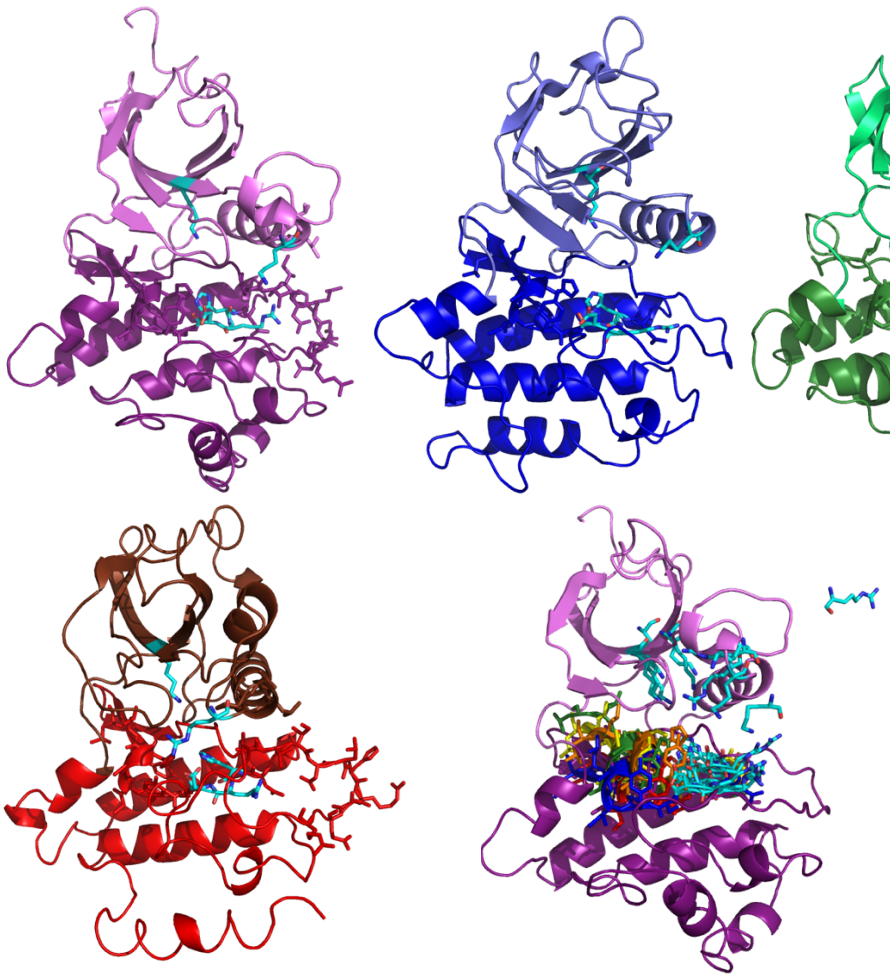


Figure 6.18: Predicted 3D models of WaaP orthologues, built on the Phyre[2] online server.

Top row, from L to R: PAWaaP (purple), PSWaaP (blue), PFWaaP (green), SEnWaaP (yellow) and ECWaaP (orange)
Botton row, from L to R: CFWaaP (red) and an overlay of stick representations of the NHRDCYICFLL catalytic motif from all six WaaP models, against the PAWaaP cartoon backbone.
The HRD triad, and putative catalytic/binding residues 50/51K and 68R/69K are enhanced in cyan and show a high level of structural conservation. **Figure 6.19** below shows a close-up of the catalytic folds of each of the predicted WaaP models.

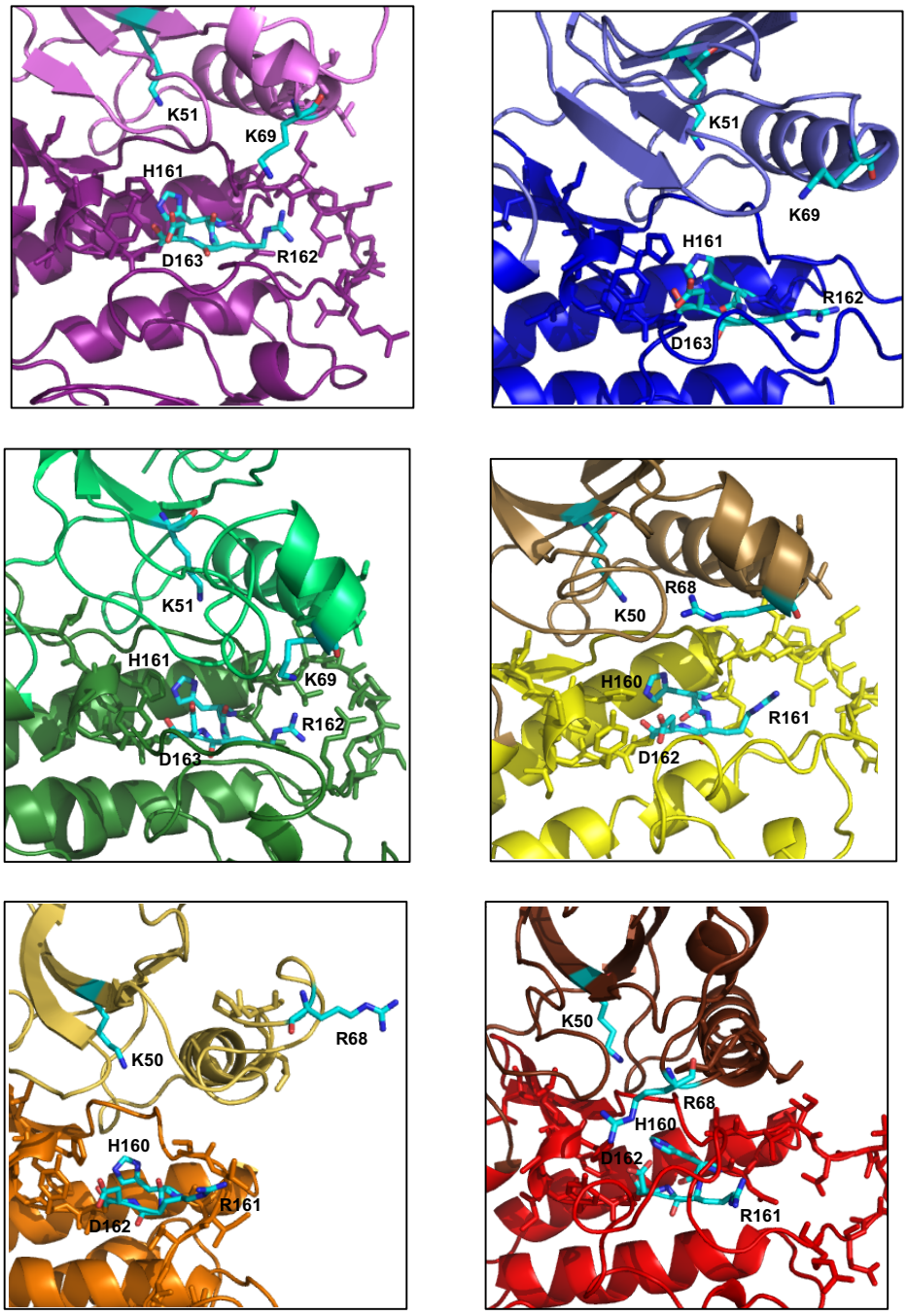


Figure 6.19: Close-up of the predicted 3D catalytic fold of six WaaP proteins

Proteins models are represented as follows: PAWaaP (purple), PSWaaP (blue), PFWaaP (green), SEnWaaP (yellow), ECWaaP (orange) and CFWaaP (red). Residues NHRDCYICFHLL of the conserved catalytic domain are shown as stick representations and can be seen to adopt a highly similar fold in all six predicted models. The HRD catalytic triad, and putative catalytic residues K50/51 and R68/K69 are enhanced in cyan. The 3D positions of the R68/K69 residues appear varied in the structural models, owing to their occurrence on flexible surface loops, however, the position of the K50/51 residue remains remarkably conserved relative to the HRD triad, suggesting a key role for this residue in WaaP's kinase activity, which was previously unidentified in reported homology studies of WaaP with eukaryotic kinases.

CHAPTER 7

DISCUSSION AND FUTURE WORK FOR WaaP STRUCTURAL STUDIES

7 DISCUSSION AND FUTURE WORK FOR WaaP STRUCTURAL STUDIES

7.1 The influence of WaaP gene sequences on its in vivo expression in *E. coli*

7.1.1 WaaP codon bias

The online Rare Codon Analysis Tool by GenScript (http://www.genscript.com/cgi-bin/tools/rare_codon_analysis) was used to assess codon bias in the WaaP gene sequences for expression in an *E. coli* host. The results are shown in **Table 7.1** below.

Table 7.1: Rare codon analysis of six WaaP gene sequences for recombinant expression in an *E. coli* host system.
WaaP entries in the table are sorted in decreasing order of codon compatibility for recombinant expression in an *E. coli* system. Entries in bold have been trialled for expression and purification, and are reported in Chapter 6. Non-bold entries are for WaaP gene homologues that have not yet been purified, but for which expression vectors have been constructed (see section 6.2).

WaaP species	CAI	Low frequency codons (%) (<30% usage)	GC Content (%)	Negative CIS elements	Negative repeat elements
P. aeruginosa	**0.73**	**6%**	**67.07 %**	**0**	**0**
P. syringae	0.72	6%	62.15 %	0	0
P. fluorescens	0.73	7%	62.73 %	0	0
S. enterica	**0.68**	**12%**	**46.89 %**	**0**	**0**
E. coli	**0.67**	**13%**	**43.58 %**	**0**	**0**
C. freundii	**0.66**	**12%**	**46.38 %**	**1**	**0**

The Codon Adaptation Index (CAI), is a measure of synonymous codon usage bias in a certain species. A score from 0 to 1 for a particular gene is calculated from the frequency of use of all codons in that species; the lower the CAI for a gene, the poorer the expression in the host organism. A CAI >0.8 implies that the gene could be

expressed very well in the desired organism. It is, therefore, a useful measure for assessing the likely success of heterologous gene expression [378]. Observations regarding the quality of EC, SEn and PA WaaP expressed and purified from *E. coli* in the experiments done so far (discussed in sections 6.3.3 and 6.3.4) are reflected in the CAI trends in **Table 7.1**; the outcomes are also reflected in the % of rare codons constituting each WaaP gene sequence – for example 13% of the ECWaaP gene consists of rare codons, compared to a 6% rare codon occurrence in the PAWaaP gene. Therefore, this is being suggested as one of the reasons why the expression quality of PAWaaP, either as His-tagged or MBP-fusion protein, was observed to be better than for ECWaaP.

The GC content of the target gene must be within an acceptable range of the expression host's genomic GC content [379]. For expression in *E. coli,* the ideal GC content of a gene should be 30 to 70%. Any GC-rich regions in the gene beyond this range could affect translational and transcriptional efficiency of the gene product. The average GC content of all the WaaP genes appear within this range, however, an exception with PAWaaP must be noted, which has two GC-rich regions above 70% between 200 – 300 bp and 375 – 430 bp. This suggests that, in this case, GC content may not be a very major influence on heterologous WaaP expression quality.

The number of negative cis elements seems to have a significant influence on the production of soluble WaaP: all WaaP proteins purified so far, except for CFWaaP, did not precipitate heavily after MBP removal with TEV-protease, irrespective of the oligomerisation state of the final protein. That a significant amount of CFWaaP had precipitated post MBP removal implies that CFWaaP expressed in *E. coli* is the least stable compared to other WaaP proteins. Considering all the above discussed factors, and the observed experimental outcomes, a clear correlation can be established. The relative impact each factor could have on heterologous WaaP expression in *E. coli* is proposed to be in the order: Negative cis elements > CAI > GC content.

PFWaaP and PSWaaP expression and purification is yet to be tested. Based on the results for EC, SEn, PA and CF WaaP purification, and their close correlation with the

codon usage trends demonstrated in **Table 7.1**, it is strongly recommended to pursue expression and purification trials for PF and PSWaaP homologues, as they are predicted to behave in a similar manner to PAWaaP. If successful, it would be a major step towards obtaining two other WaaP homologues for the much awaited structural studies of Gram negative bacteria WaaP activity and function.

7.1.2 Cysteine content

The inclusion of DTT in the SEC buffer for SEnWaaP (see section 6.3.3.1 and **Figure 6.12**) indicated that DTT may have had some effect, *in vitro*, in disrupting the high MW oligomerisation state of SEnWaaP protein, which means that undesirable intermolecular disulphide bridging could be another contributing factor to the patterns of high MW oligomerisation consistently observed across all WaaP purifications. All six WaaP species that were discussed in this thesis have two conserved cysteines, as displayed in **Figure 7.1**, which are a part of a highly conserved putative catalytic site. The conserved Asp162/163 residue, shown to be involved in ATP hydrolysis [346] is indicated in green. A third, partially conserved cysteine is also observed in EC, SEn and CF WaaP at position 126, which is absent from all three *Pseudomonas* WaaP proteins.

Cysteine residues are known to play critical roles in proteins and can be classified as structural cysteines (which aid in protein folding by forming disulphide linkages), metal coordinating, regulatory and catalytic redox and non-redox cysteines [380,381]. Given the known function of WaaP proteins as sugar kinases, the two conserved cysteines are likely to be non-redox catalytic, similar to those in protein tyrosine phosphatases and cys peptidases [381]. The role of the third cysteine in EC, SEn and CF WaaP cannot yet be deciphered. Moreover, given that the WaaP protein is naturally localised in the bacterial cytoplasm [346], the cysteines are less likely to be oxidised in their native form, implying that any undesired disulphide bonding *in vitro* could impair proper protein folding. While the clear differences in rare codon content between rfaP genes of the *Pseudomonas* species and that of EC, SEn and CF could have a major

influence on their over-expression in *an E. coli* host, the degree of influence of the cysteine content on *in vivo* and *in vitro* aggregation during expression and purification, still remains unclear and needs to be investigated further.

Revisiting the study by Zanier et. al [371] for the production of monomeric, soluble E6 protein from eight species of HPV (which was discussed earlier in section 6.3.3.1), it was proposed that soluble MBP-E6 aggregate formation was influenced by *in vivo* expression factors, and by intermolecular disulphide bridge cross-linking happening *in vitro* during extraction and purification. Mutation of the non-conserved cysteine residues into serine disrupted intermolecular disulphide cross-linking, and resulted in varied improvements in the yield of monomeric MBP-E6 during purification. The occurrence of high MW aggregates amongst the eight mutant E6 proteins varied between 64% for the most promising HPV E6 protein (E6 16) to 94% for one displaying the most tendency to aggregate (E6 45). A quick way to assess the effect of the extra Cys126 in *in vitro* (or even *in vivo*) soluble aggregate formation would be to generate C126S mutants and to express and purify these with Gro*ESL* co-expression in the *E. coli* SoluBL21™ strain. The results for PA, EC and CF WaaP purification in sections 6.3.4 and 6.3.5, would serve as comparative references for the quality of protein obtained.

```
P.fluorescens   121  ISLEDVTLNWVAEPPTPALRHALTAELARMVGDMHRGGVNHRDCYLCHFLLDTSSPIDAN  180
P.aeuroginosa   121  VDLEVFSQDWRERPPPPRLKRALVEAVARMVGDMHRAGVNHRDCYICHFLLHTDKPVSAD  180
P.syringae      121  TSLEDVSLNWRTEPPEPRLKRAFIAEVARLVGMMHRAGVNHRDCYICHFLLHTDKPVTAD  180
C.freundii      120  ISLEDFCSRWKVEPPDYVLKRTIITRVAEMVGKMHRGGVNHRDCYICHFLLQQPMPEDVA  179
E.Coli          120  ISLEDYCADWAVNPPDIRVKRMLIARVATMVRKMHTAGINHRDCYICHFLLHLPFTGRED  179
S.enterica      120  ISLEDYCADWAVNPPDAQVKWMIIKRVATMVRKMHAGGINHRDCYICHFLLHLPFTGREE  179
                       .**     *  .**   ::  :   :* :*  ** .*:***:*****.
```

Figure 7.1: Partial WaaP amino acid sequences of Gram negative bacteria species showing all their cysteine residues

7.2 Considerations for Future Studies

1) Confirming PAWaaP activity and folding

In the literature EC and PA WaaP, while purified in low yield, were generally reported to be monomeric and active, and their biochemical characterisation could be performed without further treatment or refolding of the protein. Whilst, SEC and SDS-PAGE results in this investigation demonstrate that PAWaaP was purified as a monomer at the expected MW of 31 kDa, the limitation lies in conclusively knowing if WaaP was purified in its active form. This could be verified by enzymatic assays: radiolabelled and chemiluminescence based ELISA assays for ECWaaP [7] and PAWaaP [346], [382], respectively, are reported. Another limitation is the lack of understanding on changes in polydispersity, oligomerisation state and stability of the protein over time, post purification. Yethon and Whitefield [7] report that ECWaaP retained its activity in 50% glycerol for up to 30 days at -20^{o}C; whereas PAWaaP was reported to retain about 70% of its activity after storage at -20^{o}C for a week. This indicates that the enzyme may lose its activity faster at room temperature, at which the crystallisation plates were incubated. In the future, it is recommended for plates to be incubated at 4^{o}C, instead of 22^{o}C, which is routine in our laboratory. Given that several crystallisation trials failed, it would be useful to test these factors to conclusively rule out if they are indeed limiting the crystallisation of WaaP.

2) Expression and purification of PSWaaP and PFWaaP proteins

The expression of these homologues should be carried out as described for PAWaaP, which revealed the best results for obtaining monomeric WaaP (see section 6.3.4).

3) Crystallisation trials of selected mutant forms of WaaP

As discussed in 7.1.2 above, mutation of the non-catalytic cysteine residues could be attempted to screen for a WaaP variant with better solution properties. Although not ideal, single mutations of any one conserved catalytic residue on PaWaaP could also be attempted to produce an inactive form, which may have a higher stability over an extended period of time, that would be helpful for crystallography.

4) Purification trials with codon optimised ECWaaP and SEnWaaP genes

Codon optimisation may facilitate better expression in an *E. coli* host; if this strategy works it would make additional WaaP orthologues available for crystallography.

5) Identification of other suitable WaaP homologues

A search with the *P. aeruginosa* PAO1 WaaP protein sequence on EggNOG 4.5 database gave 65 matches for existing WaaP orthologues across Gamma-proteobacteria in the Uniprot database. From these, 27 ortholgues showing 100% conservation of the NHRDCYICHFLL catalytic site were shortlisted and subjected to a rare codon analysis. The outcomes are presented in **Table 7.2** and listed in decreasing order of orthologues with the most favourable codon content for heterologous *E. coli* expression. This was done by sorting the WaaP entries by increasing number of negative CIS elements, and within the same group of negative CIS elements orthologues were sorted in increasing % of low frequency codons. The general trend was that WaaP from Order Pseudomoadales showed a lower rare codon distribution than those from Order Enterobacteriales; hence, most recommendations for future investigations of WaaP expression and purification are from Order Pseudomoadales. In some species, inter-strain variations in codon content are obvious. A percent identity matrix of the protein sequence of these orthologues are presented in APPENDIX 8, where the lowest inter-strain sequence conservation of 96.5% is observed within *P. syringae*, 96% within *P. putida*, and 72% within *P. fluorescens* strains.

6) Expression and purification of PA WapP and its homologues

In the introductory chapter for WaaP, in section 5.5.2 it was discussed that a second type of sugar kinase in *P. aeruginosa,* homologues for which were not found across enteric bacteria, was also found to be essential for *P. aeruginosa* viability. An initial expression and purification trial to test the suitability of this protein (and possibly even those from other Pseudomonads) for structural studies would be of great interest.

Table 7.2: Codon analysis of WaaP genes from selected Gammaproteobacteria

Grey rows include homologues that have been discussed in this thesis. The bold entries are WaaP homologues that have been trialled for purification.

WaaP species	CAI	Low frequency codons (%) (<30% usage)	GC Content (%)	Negative CIS elements	Negative repeat elements
P. aeruginosa PAO1	**0.73**	**6%**	**67.07**	**0**	**0**
P. syringae DC3000	0.72	6%	62.15	0	0
P. fluorescens SBW25	0.73	7%	62.73	0	0
E. coli W3110	**0.67**	**13%**	**43.58**	**0**	**0**
S. enterica	**0.68**	**12%**	**46.89**	**0**	**0**
C. freundii ATCC 8090	**0.66**	**12%**	**46.38**	**1**	**0**
P. putida S16	0.77	3%	66.60	0	0
P. putida F1	0.76	3%	64.44	0	0
P.syringae B728a	0.76	3%	62.14	0	0
P. mendocina (ymp)	0.76	4%	67.21	0	0
P. putida W619	0.74	4%	64.82	0	0
P. putida KT2440	0.75	4%	65.22	0	0
P. savastanoi 1448A	0.75	4%	62.02	0	0
P. putida GB-1	0.74	5%	64.84	0	0
P. mendocina NK-01	0.75	6%	65.05	0	0
P.fluorescens F113	0.73	6%	62.91	0	0
P. fluorescens WH6	0.74	7%	63.62	0	0
Azotobacter vinelandii ATCC BAA-1303	0.73	8%	71.88	0	0
P. fulva 12-X	0.68	8%	63.32	0	0
P. fluorescens pf-5	0.72	8%	65.38	0	0
Shigella Flexneri	0.71	9%	48.21	0	0
S. bongori ATCC43975	0.67	13%	44.38	0	0
P. alcalifaciens DSM30120	0.66	13%	39.16	0	0
Enterobacteriaceae bacterium 9_2_54FAA	0.64	15%	40.55	0	0
P. entomophila L48	0.75	4%	66.21	1	0
P.syringae ES4326	0.71	6%	62.91	1	0
C. rodentium ICC168	0.69	14%	50.09	1	0
P. rettgeri DSM1131	0.66	14%	39.62	1	0
P. rustigianni DSM4541	0.67	14%	39.16	1	0
C. youngae ATCC 29220	0.65	15%	47.22	1	0
E. albertii TW07627	0.63	17%	42.95	1	0
C. koseri ATCC BAA-895	0.68	12%	48.99	2	0
P. stuartii ATCC25827	0.62	15%	40.09	2	0

REFERENCES

[1] P. L. Oh, A. K. Benson, D. a Peterson, P. B. Patil, E. N. Moriyama, S. Roos, and J. Walter, “Diversification of the gut symbiont Lactobacillus reuteri as a result of host-driven evolution.,” *ISME J.*, vol. 4, no. 3, pp. 377–387, 2010.

[2] J. Walter, R. A. Britton, and S. Roos, “Host-microbial symbiosis in the vertebrate gastrointestinal tract and the Lactobacillus reuteri paradigm,” *Proc Natl Acad Sci U S A*, vol. 108 Suppl, pp. 4645–4652, 2011.

[3] U. Wegmann, D. A. MacKenzie, J. Zheng, A. Goesmann, S. Roos, D. Swarbreck, J. Walter, L. C. Crossman, and N. Juge, “The pan-genome of Lactobacillus reuteri strains originating from the pig gastrointestinal tract.,” *BMC Genomics*, vol. 16, no. 1, p. 1023, 2015.

[4] S. A. Frese, D. A. MacKenzie, D. A. Peterson, R. Schmaltz, T. Fangman, Y. Zhou, C. Zhang, A. K. Benson, L. A. Cody, F. Mulholland, N. Juge, and J. Walter, “Molecular Characterization of Host-Specific Biofilm Formation in a Vertebrate Gut Symbiont,” *PLoS Genet.*, vol. 9, no. 12, 2013.

[5] N. S. Berrow, D. Alderton, S. Sainsbury, J. Nettleship, R. Assenberg, N. Rahman, D. I. Stuart, and R. J. Owens, “A versatile ligation-independent cloning method suitable for high-throughput expression screening applications,” *Nucleic Acids Res.*, vol. 35, no. 6, 2007.

[6] X. Qi, S. Lan, W. Wang, L. M. Schelde, H. Dong, G. D. Wallat, H. Ly, Y. Liang, and C. Dong, “Cap binding and immune evasion revealed by Lassa nucleoprotein structure.,” *Nature*, vol. 468, no. 7325, pp. 779–83, 2010.

[7] J. A. Yethon and C. Whitfield, “Purification and Characterization of WaaP from Escherichia coli, a Lipopolysaccharide Kinase Essential for Outer Membrane Stability,” *J. Biol. Chem.*, vol. 276, no. 8, pp. 5498–5504, 2001.

[8] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins, “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.,” *Mol. Syst. Biol.*, vol. 7, no. 1, p. 539, 2011.

[9] N. P. Brown, C. Leroy, and C. Sander, “MView: a web-compatible database search or multiple alignment viewer.,” *Bioinformatics*, vol. 14, no. 4, pp. 380–1, 1998.

[10] X. Huang and W. Miller, “A time-efficient, linear-space local similarity algorithm.,” *Adv. Appl. Math.*, vol. 12, no. 3, pp. 337–357, 1991.

[11] J. Huerta-Cepas, D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, M. Kuhn, L. J. Jensen, C. von Mering, and P. Bork, “eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences,” *Nucleic Acids Res*, vol. 44, no. D1, pp. D286-93, 2016.

[12] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman, “The Pfam protein families database: towards a more sustainable future.,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D279–D285, 2015.

[13] D. W. A. Buchan, S. M. Ward, A. E. Lobley, T. C. O. Nugent, K. Bryson, and D. T. Jones, “Protein annotation and modelling servers at University College London,” *Nucleic Acids Res.*, vol. 38, no. SUPPL. 2, 2010.

[14] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, and M. J. E. Sternberg, “The Phyre2 web portal for protein modeling, prediction and analysis,” *Nat. Protoc.*, vol. 10, no. 6, pp. 845–858, 2015.

[15] L. Holm and P. Rosenstr??m, “Dali server: Conservation mapping in 3D,” *Nucleic*

Acids Res., vol. 38, no. SUPPL. 2, 2010.

[16] H. Hasegawa and L. Holm, “Advances and pitfalls of protein structural alignment,” *Current Opinion in Structural Biology*, vol. 19, no. 3. pp. 341–348, 2009.

[17] E. Krissinel and K. Henrick, “Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 60, no. 12 I, pp. 2256–2268, 2004.

[18] I. Sillitoe, T. E. Lewis, A. Cuff, S. Das, P. Ashford, N. L. Dawson, N. Furnham, R. A. Laskowski, D. Lee, J. G. Lees, S. Lehtinen, R. A. Studer, J. Thornton, and C. A. Orengo, “CATH: Comprehensive structural and functional annotations for genome sequences,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D376–D381, 2015.

[19] T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe, and N. A. Baker, “PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations,” *Nucleic Acids Res.*, vol. 35, no. SUPPL.2, 2007.

[20] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, “Electrostatics of nanosystems: application to microtubules and the ribosome.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 18, pp. 10037–41, 2001.

[21] C. S. Bond, “TopDraw: A sketchpad for protein structure topology cartoons,” *Bioinformatics*, vol. 19, no. 2, pp. 311–312, 2003.

[22] W. B. Whitman, D. C. Coleman, and W. J. Wiebe, “Prokaryotes: the unseen majority,” *Proc Natl Acad Sci U S A*, vol. 95, no. 12, pp. 6578–6583, 1998.

[23] K. O. Stetter, “Hyperthermophiles in the history of life,” *Ciba Found. Symp.*, vol. 202, no. September, pp. 1-10-18, 1996.

[24] C. Dale and N. A. Moran, “Molecular Interactions between Bacterial Symbionts and Their Hosts,” *Cell*, vol. 126, no. 3. pp. 453–465, 2006.

[25] H. J. Flint, K. P. Scott, P. Louis, and S. H. Duncan, “The role of the gut microbiota in nutrition and health,” *Nat. Rev. Gastroenterol. Hepatol.*, vol. 9, no. 10, pp. 577–589, 2012.

[26] B. Stecher and W. D. Hardt, “Mechanisms controlling pathogen colonization of the gut,” *Current Opinion in Microbiology*, vol. 14, no. 1. pp. 82–91, 2011.

[27] N. Kamada, S.-U. Seo, G. Y. Chen, and G. Núñez, “Role of the gut microbiota in immunity and inflammatory disease.,” *Nat. Rev. Immunol.*, vol. 13, no. 5, pp. 321–35, 2013.

[28] L. V. Hooper, D. R. Littman, and a. J. Macpherson, “Interactions Between the Microbiota and the Immune System,” *Science (80-.).*, vol. 336, no. 6086, pp. 1268–1273, 2012.

[29] H. Chung, S. J. Pamp, J. A. Hill, N. K. Surana, S. M. Edelman, E. B. Troy, N. C. Reading, E. J. Villablanca, S. Wang, J. R. Mora, Y. Umesaki, D. Mathis, C. Benoist, D. A. Relman, and D. L. Kasper, “Gut immune maturation depends on colonization with a host-specific microbiota,” *Cell*, vol. 149, no. 7, pp. 1578–1593, 2012.

[30] F. Sommer and F. Bäckhed, “The gut microbiota--masters of host development and physiology.,” *Nat. Rev. Microbiol.*, vol. 11, no. 4, pp. 227–38, 2013.

[31] R. E. Ley, M. Hamady, C. Lozupone, P. J. Turnbaugh, R. R. Ramey, J. S. Bircher, M. L. Schlegel, T. A. Tucker, M. D. Schrenzel, R. Knight, and J. I. Gordon, “Evolution of mammals and their gut microbes.,” *Science*, vol. 320, no. 5883, pp. 1647–1651, 2008.

[32] M. Rajili??-Stojanovi??, H. G. H. J. Heilig, S. Tims, E. G. Zoetendal, and W. M. De Vos, “Long-term monitoring of the human intestinal microbiota composition,” *Environ. Microbiol.*, vol. 15, no. 4, pp. 1146–1159, 2013.

[33] P. D. Schloss, A. M. Schubert, J. P. Zackular, K. D. Iverson, V. B. Young, and J. F. Petrosino, “Stabilization of the murine gut microbiome following weaning,” *Gut Microbes*, vol. 3, no. 4, 2012.

[34] K. A. Kline, S. F??lker, S. Dahlberg, S. Normark, and B. Henriques-Normark,

"Bacterial Adhesins in Host-Microbe Interactions," *Cell Host and Microbe*, vol. 5, no. 6. pp. 580–592, 2009.

[35] S. Lebeer, J. Vanderleyden, and S. C. J. De Keersmaecker, "Host interactions of probiotic bacterial surface molecules: comparison with commensals and pathogens.," *Nat. Rev. Microbiol.*, vol. 8, no. 3, pp. 171–184, 2010.

[36] O. Kandler, K.-O. Stetter, and R. Köhl, "Lactobacillus reuteri sp. nov., a New Species of Heterofermentative Lactobacilli," *Zentralblatt für Bakteriol. I. Abt. Orig. C Allg. Angew. und ökologische Mikrobiol.*, vol. 1, no. 3, pp. 264–269, 1980.

[37] G. Reuter, "The Lactobacillus and Bifidobacterium microflora of the human intestine: composition and succession.," *Curr. Issues Intest. Microbiol.*, vol. 2, no. 2, pp. 43–53, 2001.

[38] N. H. Salzman, H. de Jong, Y. Paterson, H. J. M. Harmsen, G. W. Welling, and N. A. Bos, "Analysis of 16S libraries of mouse gastrointestinal microflora reveals a large new group of mouse intestinal bacteria," *Microbiology*, vol. 148, no. 11. pp. 3651–3660, 2002.

[39] S. P. J. Brooks, M. McAllister, M. Sandoz, and M. L. Kalmokoff, "Culture-independent phylogenetic analysis of the faecal flora of the rat," *Can. J. Microbiol.*, vol. 49, no. 10, pp. 589–601, 2003.

[40] T. D. Leser, J. Z. Amenuvor, T. K. Jensen, R. H. Lindecrona, M. Boye, and K. Mo??ller, "Culture-independent analysis of gut bacteria: The pig gastrointestinal tract microbiota revisited," *Appl. Environ. Microbiol.*, vol. 68, no. 2, pp. 673–690, 2002.

[41] N. Yuki, T. Shimazaki, A. Kushiro, K. Watanabe, K. Uchida, T. Yuyama, and M. Morotomi, "Colonization of the stratified squamous epithelium of the nonsecreting area of horse stomach by Lactobacilli," *Appl. Environ. Microbiol.*, vol. 66, no. 11, pp. 5030–5034, 2000.

[42] S.-Y. Kim and Y. Adachi, "Biological and genetic classification of canine intestinal lactic acid bacteria and bifidobacteria.," *Microbiol. Immunol.*, vol. 51, no. 10, pp. 919–928, 2007.

[43] H. T. Abbas Hilmi, A. Surakka, J. Apajalahti, and P. E. J. Saris, "Identification of the most abundant Lactobacillus species in the crop of 1- and 5-week-old broiler chickens," *Appl. Environ. Microbiol.*, vol. 73, no. 24, pp. 7867–7873, 2007.

[44] P. Vos, R. Hogers, M. Bleeker, M. Reijans, T. Van De Lee, M. Hornes, A. Friters, J. Pot, J. Paleman, M. Kuiper, and M. Zabeau, "AFLP: A new technique for DNA fingerprinting," *Nucleic Acids Res.*, vol. 23, no. 21, pp. 4407–4414, 1995.

[45] J. J. Lin, J. Kuo, and J. Ma, "A PCR-based DNA fingerprinting technique: AFLP for molecular typing of bacteria.," *Nucleic Acids Res.*, vol. 24, no. 18, pp. 3649–50, 1996.

[46] S. P. Glaeser and P. K??mpfer, "Multilocus sequence analysis (MLSA) in prokaryotic taxonomy," *Systematic and Applied Microbiology*, vol. 38, no. 4. pp. 237–245, 2015.

[47] J. H.-C. Lin and D. C. Savage, "Host specificity of the colonization of murine gastric epithelium by lactobacilli," *FEMS Microbiol. Lett.*, vol. 24, no. 1, pp. 67–71, 1984.

[48] S. A. Frese, A. K. Benson, G. W. Tannock, D. M. Loach, J. Kim, M. Zhang, P. L. Oh, N. C. K. Heng, P. B. Patil, N. Juge, D. A. MacKenzie, B. M. Pearson, A. Lapidus, E. Dalin, H. Tice, E. Goltsman, M. Land, L. Hauser, N. Ivanova, N. C. Kyrpides, and J. Walter, "The evolution of host specialization in the vertebrate gut symbiont Lactobacillus reuteri," *PLoS Genet.*, vol. 7, no. 2, 2011.

[49] G. W. Tannock, O. Szylit, Y. Duval, and P. Raibaud, "Colonization of tissue surfaces in the gastrointestinal tract of gnotobiotic animals by lactobacillus strains.," *Can. J. Microbiol.*, vol. 28, no. 10, pp. 1196–8, 1982.

[50] D. C. Savage and R. V. H. Blumershine, "Surface-surface associations in microbial communities populating epithelial habitats in the murine gastrointestinal ecosystem: scanning electron microscopy," *Infect. Immun.*, vol. 10, no. 1, pp. 240–250, 1974.

[51] S. Kakimoto, K. Okazaki, T. Sakane, K. Imai, Y. Sumino, S. I. Akiyama, and Y. Nakao, "Isolation and Taxonomic Characterization of Acid Urease-Producing

Bacteria," *Agric. Biol. Chem.*, vol. 53, pp. 1111–1118, 1989.

[52] T. L. Talarico, I. A. Casas, T. C. Chung, and W. J. Dobrogosz, "Production and isolation of reuterin, a growth inhibitor produced by Lactobacillus reuteri," *Antimicrob. Agents Chemother.*, vol. 32, no. 12, pp. 1854–1858, 1988.

[53] C. M. Wilson, D. M. Loach, B. Lawley, T. Bell, I. M. Sims, P. W. O'Toole, A. Zomer, and G. W. Tannock, "Lactobacillus reuteri 100-23 Modulates Urea Hydrolysis in the Murine Stomach," *Appl. Environ. Microbiol.*, vol. 80, no. 19, pp. 6104–13, 2014.

[54] S. E. Jones and J. Versalovic, "Probiotic Lactobacillus reuteri biofilms produce antimicrobial and anti-inflammatory factors.," *BMC Microbiol.*, vol. 9, p. 35, 2009.

[55] J. K. Spinler, M. Taweechotipatr, C. L. Rognerud, C. N. Ou, S. Tumwasorn, and J. Versalovic, "Human-derived probiotic Lactobacillus reuteri demonstrate antimicrobial activities targeting diverse enteric bacterial pathogens," *Anaerobe*, vol. 14, no. 3, pp. 166–171, 2008.

[56] J. R. Davies, G. Svens??ter, and M. C. Herzberg, "Identification of novel LPXTG-linked surface proteins from Streptococcus gordonii," *Microbiology*, vol. 155, no. 6, pp. 1977–1988, 2009.

[57] A. P. A. Hendrickx, R. J. L. Willems, M. J. M. Bonten, and W. van Schaik, "LPxTG surface proteins of enterococci," *Trends in Microbiology*, vol. 17, no. 9. pp. 423–430, 2009.

[58] N. Juge, "Microbial adhesins to gastrointestinal mucus," *Trends in Microbiology*, vol. 20, no. 1. pp. 30–39, 2012.

[59] J. Walter, N. C. K. Heng, W. P. Hammes, D. M. Loach, G. W. Tannock, and C. Hertel, "Identification of Lactobacillus reuteri genes specifically induced in the mouse gastrointestinal tract," *Appl. Environ. Microbiol.*, vol. 69, no. 4, pp. 2044–2051, 2003.

[60] H. L. Carrell, J. P. Glusker, V. Burger, F. Manfre, D. Tritsch, and J. F. Biellmann, "X-ray analysis of D-xylose isomerase at 1.9 A: native enzyme in complex with substrate and with a mechanism-designed inactivator.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 86, no. 12, pp. 4440–4444, 1989.

[61] Y. X. Zhang, P. V Vadlani, A. Kumar, P. R. Hardwidge, R. Govind, T. Tanaka, and A. Kondo, "Enhanced D-lactic acid production from renewable resources using engineered Lactobacillus plantarum," *Appl. Microbiol. Biotechnol.*, vol. 100, no. 1, pp. 279–288, 2016.

[62] A. L. Spoering and M. S. Gilmore, "Quorum sensing and DNA release in bacterial biofilms," *Current Opinion in Microbiology*, vol. 9, no. 2. pp. 133–137, 2006.

[63] G. Brackman, P. Cos, L. Maes, H. J. Nelis, and T. Coenye, "Quorum sensing inhibitors increase the susceptibility of bacterial biofilms to antibiotics in vitro and in vivo," *Antimicrob. Agents Chemother.*, vol. 55, no. 6, pp. 2655–2661, 2011.

[64] G. Brackman, K. Breyne, R. De Rycke, A. Vermote, F. Van Nieuwerburgh, E. Meyer, S. Van Calenbergh, and T. Coenye, "The Quorum Sensing Inhibitor Hamamelitannin Increases Antibiotic Susceptibility of Staphylococcus aureus Biofilms by Affecting Peptidoglycan Biosynthesis and eDNA Release.," *Sci. Rep.*, vol. 6, p. 20321, 2016.

[65] L. Hall-Stoodley, J. W. Costerton, P. Stoodley, M. State, and B. Engineering, "Bacterial biofilms: from the natural environment to infectious diseases.," *Nat. Rev. Microbiol.*, vol. 2, no. 2, pp. 95–108, 2004.

[66] J. L. Sonnenburg, L. T. Angenent, and J. I. Gordon, "Getting a grip on things: how do communities of bacterial symbionts become established in our intestine?," *Nat. Immunol.*, vol. 5, no. 6, pp. 569–573, 2004.

[67] N. Suegara, M. Morotomi, T. Watanabe, Y. Kawal, and M. Mutai, "Behavior of microflora in the rat stomach: adhesion of lactobacilli to the keratinized epithelial cells of the rat stomach in vitro," *Infect. Immun.*, vol. 12, no. 1, pp. 173–179, 1975.

[68] R. Fuller, P. A. Barrow, and B. E. Brooker, "Bacteria associated with the gastric epithelium of neonatal pigs.," *Appl. Environ. Microbiol.*, vol. 35, no. 3, pp. 582–591, 1978.

[69] R. Fuller and B. E. Brooker, “Lactobacilli which attach to the crop epithelium of the fowl,” *American Journal of Clinical Nutrition*, vol. 27, no. 11. pp. 1305–1312, 1974.

[70] J. Walter, “Ecological role of lactobacilli in the gastrointestinal tract: Implications for fundamental and biomedical research,” *Applied and Environmental Microbiology*, vol. 74, no. 16. pp. 4985–4996, 2008.

[71] T. L. A. Nguyen, S. Vieira-Silva, A. Liston, and J. Raes, “How informative is the mouse for human gut microbiota research?,” *Dis. Model. Mech.*, vol. 8, no. 1, pp. 1–16, 2015.

[72] T. J. Caron, K. E. Scott, J. G. Fox, and S. J. Hagen, “Tight junction disruption: Helicobacter pylori and dysregulation of the gastric mucosal barrier,” *World Journal of Gastroenterology*, vol. 21, no. 40. pp. 11411–11427, 2015.

[73] H. N. Lærke and M. S. Hedemann, “The digestive System of the Pig,” in *Nutritional Physiology of Pigs*, Online Pub., K. E. B. Knudsen, N. J. Kjeldsen, H. D. Poulsen, and B. B. Jensen, Eds. Foulum: Videncenter for Svineproduktion, 2012.

[74] B. Svihus, “Function of the digestive system,” *J. Appl. Poult. Res.*, vol. 23, no. 2, pp. 306–314, 2014.

[75] M. P. Vélez, S. C. J. De Keersmaecker, and J. Vanderleyden, “Adherence factors of Lactobacillus in the human gastrointestinal tract,” *FEMS Microbiology Letters*, vol. 276, no. 2. pp. 140–148, 2007.

[76] V. Cleusix, C. Lacroix, S. Vollenweider, M. Duboux, and G. Le Blay, “Inhibitory activity spectrum of reuterin produced by Lactobacillus reuteri against intestinal bacteria.,” *BMC Microbiol.*, vol. 7, p. 101, 2007.

[77] J. Walter, P. Chagnaud, G. W. Tannock, D. M. Loach, F. Dal Bello, H. F. Jenkinson, W. P. Hammes, and C. Hertel, “A high-molecular-mass surface protein (Lsp) and methionine sulfoxide reductase B (MsrB) contribute to the ecological performance of Lactobacillus reuteri in the murine gut,” *Appl. Environ. Microbiol.*, vol. 71, no. 2, pp. 979–986, 2005.

[78] S. Ramboarina, J. A. Garnett, M. Zhou, Y. Li, Z. Peng, J. D. Taylor, W. C. Lee, A. Bodey, J. W. Murray, Y. Alguel, J. Bergeron, B. Bardiaux, E. Sawyer, R. Isaacson, C. Tagliaferri, E. Cota, M. Nilges, P. Simpson, T. Ruiz, H. Wu, and S. Matthews, “Structural insights into serine-rich fimbriae from gram-positive bacteria,” *J. Biol. Chem.*, vol. 285, no. 42, pp. 32446–32457, 2010.

[79] Y. Q. Xiong, B. A. Bensing, A. S. Bayer, H. F. Chambers, and P. M. Sullam, “Role of the serine-rich surface glycoprotein GspB of Streptococcus gordonii in the pathogenesis of infective endocarditis,” *Microb. Pathog.*, vol. 45, no. 4, pp. 297–301, 2008.

[80] M. E. Feltcher and M. Braunstein, “Emerging themes in SecA2-mediated protein export.,” *Nat. Rev. Microbiol.*, vol. 10, no. 11, pp. 779–89, 2012.

[81] B. A. Bensing and P. M. Sullam, “An accessory sec locus of Streptococcus gordonii is required for export of the surface protein GspB and for normal levels of binding to human platelets,” *Mol. Microbiol.*, vol. 44, no. 4, pp. 1081–1094, 2002.

[82] M. Y. Mistou, S. Dramsi, S. Brega, C. Poyart, and P. Trieu-Cuot, “Molecular dissection of the secA2 locus of group B streptococcus reveals that glycosylation of the Srr1 LPXTG protein is required for full virulence,” *J. Bacteriol.*, vol. 191, no. 13, pp. 4195–4206, 2009.

[83] H. Wu, S. Bu, P. Newell, Q. Chen, and P. Fives-Taylor, “Two gene determinants are differentially involved in the biogenesis of Fap1 precursors in Streptococcus parasanguis,” in *Journal of Bacteriology*, 2007, vol. 189, no. 4, pp. 1390–1398.

[84] J. Walter, C. Schwab, D. M. Loach, M. G. G??nzle, and G. W. Tannock, “Glucosyltransferase A (GtfA) and inulosucrase (Inu) of Lactobacillus reuteri TMW1.106 contribute to cell aggregation, in vitro biofilm formation, and colonization of the mouse gastrointestinal tract,” *Microbiology*, vol. 154, no. 1, pp. 72–80, 2008.

[85] Y. J. Goh and T. R. Klaenhammer, “Functional roles of aggregation-promoting-like

factor in stress tolerance and adherence of lactobacillus acidophilus NCFM," *Appl. Environ. Microbiol.*, vol. 76, no. 15, pp. 5005–5012, 2010.

[86] M. S. Turner, L. M. Hafner, T. Walsh, and P. M. Giffard, "Identification and characterization of the novel LysM domain-containing surface protein Sep from Lactobacillus fermentum BR11 and its use as a peptide fusion partner in Lactobacillus and Lactococcus," *Appl. Environ. Microbiol.*, vol. 70, no. 6, pp. 3673–3680, 2004.

[87] G. Buist, A. Steen, J. Kok, and O. P. Kuipers, "LysM, a widely distributed protein motif for binding to (peptido)glycans," *Molecular Microbiology*, vol. 68, no. 4. pp. 838–847, 2008.

[88] B. K. Sharma-Kuinkel, E. E. Mann, J. S. Ahn, L. J. Kuechenmeister, P. M. Dunman, and K. W. Bayles, "The Staphylococcus aureus LytSR two-component regulatory system affects biofilm formation," *J. Bacteriol.*, vol. 191, no. 15, pp. 4767–4775, 2009.

[89] K. W. Bayles, "The biological role of death and lysis in biofilm development.," *Nat. Rev. Microbiol.*, vol. 5, no. 9, pp. 721–726, 2007.

[90] E. E. Mann, K. C. Rice, B. R. Boles, J. L. Endres, D. Ranjit, L. Chandramohan, L. H. Tsang, M. S. Smeltzer, A. R. Horswill, and K. W. Bayles, "Modulation of eDNA release and degradation affects Staphylococcus aureus biofilm maturation," *PLoS One*, vol. 4, no. 6, 2009.

[91] D. Heavens, L. E. Tailford, L. Crossman, F. Jeffers, D. A. MacKenzie, M. Caccamo, and N. Juge, "Genome sequence of the vertebrate gut symbiont Lactobacillus reuteri ATCC 53608," *J. Bacteriol.*, vol. 193, no. 15, pp. 4015–4016, 2011.

[92] S. Lebeer, J. Vanderleyden, and S. C. J. De Keersmaecker, "Genes and molecules of lactobacilli supporting probiotic action.," *Microbiol. Mol. Biol. Rev.*, vol. 72, no. 4, p. 728–764, Table of Contents, 2008.

[93] M. L. van Tassell and M. J. Miller, "Lactobacillus adhesion to mucus," *Nutrients*, vol. 3, no. 5. pp. 613–636, 2011.

[94] S. Etzold and N. Juge, "Structural insights into bacterial recognition of intestinal mucins," *Current Opinion in Structural Biology*, vol. 28, no. 1. pp. 23–31, 2014.

[95] A. Mitchell, H. Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, A. Sangrador-Vegas, M. Scheremetjew, C. Rato, S. Y. Yong, A. Bateman, M. Punta, T. K. Attwood, C. J. A. Sigrist, N. Redaschi, C. Rivoire, I. Xenarios, D. Kahn, D. Guyot, P. Bork, I. Letunic, J. Gough, M. Oates, D. Haft, H. Huang, D. A. Natale, C. H. Wu, C. Orengo, I. Sillitoe, H. Mi, P. D. Thomas, and R. D. Finn, "The InterPro protein families database: The classification resource after 15 years," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D213–D221, 2015.

[96] P. Aleljung, W. Shen, B. Rozalska, U. Hellman, Å. Ljungh, and T. Wadström, "Purification of collagen-binding proteins of Lactobacillus reuteri NCIB 11951," *Curr. Microbiol.*, vol. 28, no. 4, pp. 231–236, 1994.

[97] S. Roos, P. Aleljung, N. Robert, B. Lee, T. Wadström, M. Lindberg, and H. Jonsson, "A collagen binding protein from Lactobacillus reuteri is part of an ABC transporter system?," *FEMS Microbiol. Lett.*, vol. 144, no. 1, pp. 33–38, 1996.

[98] M. Rojas, F. Ascencio, and P. L. Conway, "Purification and characterization of a surface protein from Lactobacillus fermentum 104R that binds to porcine small intestinal mucus and gastric mucin," *Appl. Environ. Microbiol.*, vol. 68, no. 5, pp. 2330–2336, 2002.

[99] Y. Miyoshi, S. Okada, T. Uchimura, and E. Satoh, "A mucus adhesion promoting protein, MapA, mediates the adhesion of Lactobacillus reuteri to Caco-2 human intestinal epithelial cells.," *Biosci. Biotechnol. Biochem.*, vol. 70, no. 7, pp. 1622–1628, 2006.

[100] Y. Matsuo, Y. Miyoshi, S. Okada, and E. Satoh, "Receptor-like Molecules on Human Intestinal Epithelial Cells Interact with an Adhesion Factor from Lactobacillus reuteri.," *Biosci. microbiota, food Heal.*, vol. 31, no. 4, pp. 93–102, 2012.

[101] S. Roos and H. Jonsson, “A high-molecular-mass cell-surface protein from Lactobacillus reuteri 1063 adheres to mucus components,” *Microbiology*, vol. 148, no. 2, pp. 433–442, 2002.

[102] S. Etzold, O. I. Kober, D. A. Mackenzie, L. E. Tailford, A. P. Gunning, J. Walshaw, A. M. Hemmings, and N. Juge, “Structural basis for adaptation of lactobacilli to gastrointestinal mucus,” *Environ. Microbiol.*, vol. 16, no. 3, pp. 888–903, 2014.

[103] D. A. MacKenzie, F. Jeffers, M. L. Parker, A. Vibert-Vallet, R. J. Bongaerts, S. Roos, J. Walter, and N. Juge, “Strain-specific diversity of mucus-binding proteins in the adhesion and aggregation properties of Lactobacillus reuteri,” *Microbiology*, vol. 156, no. 11, pp. 3368–3378, 2010.

[104] D. A. MacKenzie, L. E. Tailford, A. M. Hemmings, and N. Juge, “Crystal structure of a mucus-binding protein repeat reveals an unexpected functional immunoglobulin binding activity,” *J. Biol. Chem.*, vol. 284, no. 47, pp. 32444–32453, 2009.

[105] K. Nishiyama, A. Ochiai, D. Tsubokawa, K. Ishihara, Y. Yamamoto, and T. Mukai, “Identification and characterization of sulfated carbohydrate-binding protein from Lactobacillus reuteri,” *PLoS One*, vol. 8, no. 12, 2013.

[106] S. Etzold, D. A. Mackenzie, F. Jeffers, J. Walshaw, S. Roos, A. M. Hemmings, and N. Juge, “Structural and molecular insights into novel surface-exposed mucus adhesins from lactobacillus reuteri human strains,” *Mol. Microbiol.*, vol. 92, no. 3, pp. 543–556, 2014.

[107] H. Jensen, S. Roos, H. Jonsson, I. Rud, S. Grimmer, J. P. van Pijkeren, R. A. Britton, and L. Axelsson, “Role of Lactobacillus reuteri cell and mucus-binding protein A (CmbA) in adhesion to intestinal epithelial cells and mucus in vitro,” *Microbiology*, vol. 160, pp. 671–681, 2014.

[108] M. Zhou and H. Wu, “Glycosylation and biogenesis of a family of serine-rich bacterial adhesins,” *Microbiology*, vol. 155, no. 2. pp. 317–327, 2009.

[109] I. R. Siboo, H. F. Chambers, and P. M. Sullam, “Role of SraP, a Serine-Rich Surface Protein of Staphylococcus aureus, in Binding to Human Platelets,” *Infect. Immun.*, vol. 73, no. 4, pp. 2273–2280, 2005.

[110] N. M. van Sorge, D. Quach, M. a Gurney, P. M. Sullam, V. Nizet, and K. S. Doran, “The group B streptococcal serine-rich repeat 1 glycoprotein mediates penetration of the blood-brain barrier.,” *J. Infect. Dis.*, vol. 199, no. 10, pp. 1479–87, 2009.

[111] P. Shivshankar, C. Sanchez, L. F. Rose, and C. J. Orihuela, “The Streptococcus pneumoniae adhesin PsrP binds to Keratin 10 on lung cells,” *Mol. Microbiol.*, vol. 73, no. 4, pp. 663–679, 2009.

[112] K. N. Seifert, E. E. Adderson, A. A. Whiting, J. F. Bohnsack, P. J. Crowley, and L. J. Brady, “A unique serine-rich repeat protein (Srr-2) and novel surface antigen (??) associated with a virulent lineage of serotype III Streptococcus agalactiae,” *Microbiology*, vol. 152, no. 4, pp. 1029–1040, 2006.

[113] H. Wu, K. P. Mintz, M. Ladha, and P. M. Fives-Taylor, “Isolation and characterization of Fap1, a fimbriae-associated adhesin of Streptococcus parasanguis FW213,” *Mol. Microbiol.*, vol. 28, no. 3, pp. 487–500, 1998.

[114] H. Wu and P. M. Fives-Taylor, “Identification of dipeptide repeats and a cell wall sorting signal in the fimbriae-associated adhesin, Fap1, of *Streptococcus parasanguis*.,” *Mol. Microbiol.*, vol. 34, no. 5, pp. 1070–81, 1999.

[115] A. Lizcano, C. J. Sanchez, and C. J. Orihuela, “A role for glycosylated serine-rich repeat proteins in Gram-positive bacterial pathogenesis,” *Molecular Oral Microbiology*, vol. 27, no. 4. pp. 257–269, 2012.

[116] T. M. Pyburn, B. A. Bensing, Y. Q. Xiong, B. J. Melancon, T. M. Tomasiak, N. J. Ward, V. Yankovskaya, K. M. Oliver, G. Cecchini, G. A. Sulikowski, M. J. Tyska, P. M. Sullam, and T. M. Iverson, “A structural model for binding of the serine-rich repeat adhesin gspb to host carbohydrate receptors,” *PLoS Pathog.*, vol. 7, no. 7, 2011.

[117] Y. H. Yang, Y. L. Jiang, J. Zhang, L. Wang, X. H. Bai, S. J. Zhang, Y. M. Ren, N. Li,

Y. H. Zhang, Z. Zhang, Q. Gong, Y. Mei, T. Xue, J. R. Zhang, Y. Chen, and C. Z. Zhou, “Structural Insights into SraP-Mediated Staphylococcus aureus Adhesion to Host Cells,” *PLoS Pathog.*, vol. 10, no. 6, 2014.

[118] U. Samen, B. J. Eikmanns, D. J. Reinscheid, and F. Borges, “The surface protein Srr-1 of Streptococcus agalactiae binds human keratin 4 and promotes adherence to epithelial HEp-2 cells,” *Infect. Immun.*, vol. 75, no. 11, pp. 5405–5414, 2007.

[119] D. Takamatsu, B. A. Bensing, H. Cheng, G. A. Jarvis, I. R. Siboo, J. A. L??pez, J. M. Griffiss, and P. M. Sullam, “Binding of the Streptococcus gordonii surface glycoproteins GspB and Hsa to specific carbohydrate structures on platelet membrane glycoprotein Ib??,” *Mol. Microbiol.*, vol. 58, no. 2, pp. 380–392, 2005.

[120] P. S. Handley, F. F. Correia, K. Russell, B. Rosan, and J. M. DiRienzo, “Association of a novel high molecular weight, serine-rich protein (SrpA) with fibril-mediated adhesion of the oral biofilm bacterium Streptococcus cristatus,” *Oral Microbiol. Immunol.*, vol. 20, no. 3, pp. 131–140, 2005.

[121] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. . Wilkins, R. . Appel, and A. Bairoch, “Protein Identification and Analysis Tools on the ExPASy Server,” in *The Proteomics Protocols Handbook*, 2005, pp. 571–607.

[122] B. A. Bensing, B. W. Gibson, and P. M. Sullam, “The Streptococcus gordonii Platelet Binding Protein GspB Undergoes Glycosylation Independently of Export,” *J. Bacteriol.*, vol. 186, no. 3, pp. 638–645, 2004.

[123] D. Takamatsu, B. A. Bensing, and P. M. Sullam, “Four proteins encoded in the gspB-secY2A2 operon of Streptococcus gordonii mediate the intracellular glycosylation of the platelet-binding protein GspB,” *J. Bacteriol.*, vol. 186, no. 21, pp. 7100–7111, 2004.

[124] D. Takamatsu, B. A. Bensing, and P. M. Sullam, “Two additional components of the accessory Sec system mediating export of the Streptococcus gordonii platelet-binding protein GspB,” *J. Bacteriol.*, vol. 187, no. 11, pp. 3878–3883, 2005.

[125] P. Xu, J. M. Alves, T. Kitten, A. Brown, Z. Chen, L. S. Ozaki, P. Manque, X. Ge, M. G. Serrano, D. Puiu, S. Hendricks, Y. Wang, M. D. Chaplin, D. Akan, S. Paik, D. L. Peterson, F. L. Macrina, and G. A. Buck, “Genome of the opportunistic pathogen Streptococcus sanguinis,” *J. Bacteriol.*, vol. 189, no. 8, pp. 3166–3175, 2007.

[126] H. Tettelin, K. E. Nelson, I. T. Paulsen, J. A. Eisen, T. D. Read, S. Peterson, J. Heidelberg, R. T. DeBoy, D. H. Haft, R. J. Dodson, A. S. Durkin, M. Gwinn, J. F. Kolonay, W. C. Nelson, J. D. Peterson, L. A. Umayam, O. White, S. L. Salzberg, M. R. Lewis, D. Radune, E. Holtzapple, H. Khouri, A. M. Wolf, T. R. Utterback, C. L. Hansen, L. A. McDonald, T. V. Feldblyum, S. Angiuoli, T. Dickinson, E. K. Hickey, I. E. Holt, B. J. Loftus, F. Yang, H. O. Smith, J. C. Venter, B. A. Dougherty, D. A. Morrison, S. K. Hollingshead, and C. M. Fraser, “Complete Genome Sequence of a Virulent Isolate of Streptococcus pneumoniae,” *Science (80-.).*, vol. 293, no. 5529, pp. 498–506, 2001.

[127] T. Baba, F. Takeuchi, M. Kuroda, H. Yuzawa, K. I. Aoki, A. Oguchi, Y. Nagai, N. Iwama, K. Asano, T. Naimi, H. Kuroda, L. Cui, K. Yamamoto, and K. Hiramatsu, “Genome and virulence determinants of high virulence community-acquired MRSA,” *Lancet*, vol. 359, no. 9320, pp. 1819–1827, 2002.

[128] F. Takeuchi, S. Watanabe, T. Baba, H. Yuzawa, T. Ito, Y. Morimoto, M. Kuroda, L. Cui, M. Takahashi, A. Ankai, S. I. Baba, S. Fukui, J. C. Lee, and K. Hiramatsu, “Whole-genome sequencing of Staphylococcus haemolyticus uncovers the extreme plasticity of its genome and the evolution of human-colonizing staphylococcal species,” *J. Bacteriol.*, vol. 187, no. 21, pp. 7292–7308, 2005.

[129] B. A. Bensing, Z. Khedri, L. Deng, H. Yu, A. Prakobphol, S. J. Fisher, X. Chen, T. M. Iverson, A. Varki, and P. M. Sullam, “Novel aspects of sialoglycan recognition by the Siglec-like domains of streptococcal SRR glycoproteins,” *Glycobiology*, pp. 1–13, 2016.

[130] E. H. Froeliger and P. Fives-Taylor, “Streptococcus parasanguis fimbria-associated adhesin fap1 is required for biofilm formation,” *Infect. Immun.*, vol. 69, no. 4, pp.

2512–2519, 2001.

[131] J. A. Garnett, P. J. Simpson, J. Taylor, S. V. Benjamin, C. Tagliaferri, E. Cota, Y. Y. M. Chen, H. Wu, and S. Matthews, “Structural insight into the role of Streptococcus parasanguinis Fap1 within oral biofilm formation,” *Biochem. Biophys. Res. Commun.*, vol. 417, no. 1, pp. 421–426, 2012.

[132] T. Schulte, J. Löfling, C. Mikaelsson, A. Kikhney, K. Hentrich, A. Diamante, C. Ebel, S. Normark, D. Svergun, B. Henriques-Normark, and A. Achour, “The basic keratin 10-binding domain of the virulence-associated pneumococcal serine-rich protein PsrP adopts a novel MSCRAMM fold.,” *Open Biol.*, vol. 4, p. 130090, 2014.

[133] C. J. Sanchez, P. Shivshankar, K. Stol, S. Trakhtenbroit, P. M. Sullam, K. Sauer, P. W. M. Hermans, and C. J. Orihuela, “The pneumococcal serine-rich repeat protein is an intraspecies bacterial adhesin that promotes bacterial aggregation in Vivo and in biofilms,” *PLoS Pathog.*, vol. 6, no. 8, pp. 33–34, 2010.

[134] B. A. Bensing, J. A. L??pez, and P. M. Sullam, “The Streptococcus gordonii surface proteins GspB and Hsa mediate binding to sialylated carbohydrate epitopes on the platelet membrane glycoprotein Ib??,” *Infect. Immun.*, vol. 72, no. 11, pp. 6528–6537, 2004.

[135] B. A. Bensing, L. V. Loukachevitch, K. M. McCulloch, H. Yu, K. R. Vann, Z. Wawrzak, S. Anderson, X. Chen, P. M. Sullam, and T. M. Iverson, “Structural Basis for Sialoglycan Binding by the Streptococcus sanguinis SrpA Adhesin,” *J. Biol. Chem.*, no. 1, p. jbc.M115.701425, 2016.

[136] H. S. Seo, G. Minasov, R. Seepersaud, K. S. Doran, I. Dubrovska, L. Shuvalova, W. F. Anderson, T. M. Iverson, and P. M. Sullam, “Characterization of fibrinogen binding by glycoproteins Srr1 and Srr2 of Streptococcus agalactiae,” *J. Biol. Chem.*, vol. 288, no. 50, pp. 35982–35996, 2013.

[137] R. Sundaresan, U. Samen, and K. Ponnuraj, “Expression, purification, crystallization and preliminary X-ray diffraction studies of the human keratin 4-binding domain of serine-rich repeat protein 1 from Streptococcus agalactiae,” *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, vol. 67, no. 12, pp. 1582–1585, 2011.

[138] L. Deng, B. A. Bensing, S. Thamadilok, H. Yu, K. Lau, X. Chen, S. Ruhl, P. M. Sullam, and A. Varki, “Oral Streptococci Utilize a Siglec-Like Domain of Serine-Rich Repeat Adhesins to Preferentially Target Platelet Sialoglycans in Human Blood,” *PLoS Pathog.*, vol. 10, no. 12, 2014.

[139] G. Rhodes, *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*, vol. 35. 2006.

[140] I. R. Krauss, A. Merlino, A. Vergara, and F. Sica, “An overview of biological macromolecule crystallization,” *International Journal of Molecular Sciences*, vol. 14, no. 6. pp. 11643–11691, 2013.

[141] “Welcome to VMXi.” [Online]. Available: http://www.diamond.ac.uk/Beamlines/Mx/VMXi.html. [Accessed: 03-Mar-2017].

[142] “Welcome to VMXm.” [Online]. Available: http://www.diamond.ac.uk/Beamlines/Mx/VMXm.html. [Accessed: 03-Mar-2017].

[143] S. Khurshid, E. Saridakis, L. Govada, and N. E. Chayen, “Porous nucleating agents for protein crystallization.,” *Nat. Protoc.*, vol. 9, no. 7, pp. 1621–33, 2014.

[144] J. D. H. DONNAY and G. DONNAY, “International Tables for X-ray Crystallography: Symmetry Groups,” *International Union of Crystallography*, vol. 1. p. 558, 1969.

[145] A. Wlodawer, W. Minor, Z. Dauter, and M. Jaskolski, “Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures,” *FEBS Journal*, vol. 275, no. 1. pp. 1–21, 2008.

[146] C. R. Beckwith, “Characterisation of Cytochrome c MtoD from the Iron- Oxidising Autotroph Sideroxydans lithotrophicus ES-1,” University of East Anglia, 2015.

[147] T. G. G. Battye, L. Kontogiannis, O. Johnson, H. R. Powell, and A. G. W. Leslie, “iMOSFLM: A new graphical interface for diffraction-image processing with

MOSFLM," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 67, no. 4, pp. 271–281, 2011.

[148] P. R. Evans, "An introduction to data reduction: Space-group determination, scaling and intensity statistics," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 67, no. 4, pp. 282–292, 2011.

[149] C. W. Carter and R. M. sweet, "Macromolecular Crystallography," in *Methods in Enzymology, Volume 374, Part D.*, 2003, p. 326.

[150] C. Abergel, "Molecular replacement: Tricks and treats," in *Acta Crystallographica Section D: Biological Crystallography*, 2013, vol. 69, no. 11, pp. 2167–2173.

[151] K. Cowtan, "The Buccaneer software for automated model building. 1. Tracing protein chains," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 62, no. 9, pp. 1002–1011, 2006.

[152] G. Langer, S. X. Cohen, V. S. Lamzin, and A. Perrakis, "Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7.," *Nat. Protoc.*, vol. 3, no. 7, pp. 1171–9, 2008.

[153] G. Langer, S. X. Cohen, V. S. Lamzin, and A. Perrakis, "Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7.," *Nat. Protoc.*, vol. 3, no. 7, pp. 1171–9, 2008.

[154] P. Emsley and K. Cowtan, "Coot: Model-building tools for molecular graphics," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 60, no. 12 I, pp. 2126–2132, 2004.

[155] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, and A. A. Vagin, "REFMAC5 for the refinement of macromolecular crystal structures," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 67, no. 4, pp. 355–367, 2011.

[156] P. A. Karplus and K. Diederichs, "Linking Crystallographic Model and Data Quality," *Science (80-.).*, vol. 336, no. 6084, pp. 1030–1033, 2012.

[157] A. Wernimont and A. Edwards, "In Situ proteolysis to generate crystals for structure determination: An update," *PLoS One*, vol. 4, no. 4, 2009.

[158] G. Winter, "Xia2: An expert system for macromolecular crystallography data reduction," *J. Appl. Crystallogr.*, vol. 43, no. 1, pp. 186–190, 2010.

[159] P. Evans, "Scaling and assessment of data quality," in *Acta Crystallographica Section D: Biological Crystallography*, 2006, vol. 62, no. 1, pp. 72–82.

[160] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, and K. S. Wilson, "Overview of the CCP4 suite and current developments," *Acta Crystallographica Section D: Biological Crystallography*, vol. 67, no. 4. pp. 235–242, 2011.

[161] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read, "Phaser crystallographic software," *J. Appl. Crystallogr.*, vol. 40, no. 4, pp. 658–674, 2007.

[162] G. N. Murshudov, A. A. Vagin, and E. J. Dodson, "Refinement of macromolecular structures by the maximum-likelihood method," *Acta Crystallographica Section D: Biological Crystallography*, vol. 53, no. 3. pp. 240–255, 1997.

[163] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, "MolProbity: All-atom structure validation for macromolecular crystallography," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 66, no. 1, pp. 12–21, 2010.

[164] A. Dong, X. Xu, A. M. Edwards, C. Chang, M. Chruszcz, M. Cuff, M. Cymborowski, R. Di Leo, O. Egorova, E. Evdokimova, E. Filippova, J. Gu, J. Guthrie, A. Ignatchenko, A. Joachimiak, N. Klostermann, Y. Kim, Y. Korniyenko, W. Minor, Q. Que, A. Savchenko, T. Skarina, K. Tan, A. Yakunin, A. Yee, V. Yim, R. Zhang, H. Zheng, M. Akutsu, C. Arrowsmith, G. V Avvakumov, A. Bochkarev, L.-G. Dahlgren, S. Dhe-

Paganon, S. Dimov, L. Dombrovski, P. Finerty, S. Flodin, A. Flores, S. Gräslund, M. Hammerström, M. D. Herman, B.-S. Hong, R. Hui, I. Johansson, Y. Liu, M. Nilsson, L. Nedyalkova, P. Nordlund, T. Nyman, J. Min, H. Ouyang, H. Park, C. Qi, W. Rabeh, L. Shen, Y. Shen, D. Sukumard, W. Tempel, Y. Tong, L. Tresagues, M. Vedadi, J. R. Walker, J. Weigelt, M. Welin, H. Wu, T. Xiao, H. Zeng, and H. Zhu, "In situ proteolysis for protein crystallization and structure determination.," *Nat. Methods*, vol. 4, no. 12, pp. 1019–1021, 2007.

[165] M. S. Weiss, "Global indicators of X-ray data quality," *J. Appl. Crystallogr.*, vol. 34, no. 2, pp. 130–135, 2001.

[166] K. Diederichs and P. A. Karplus, "Better models by discarding data?," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 69, no. 7, pp. 1215–1222, 2013.

[167] B. Kobe and A. V. Kajava, "When protein folding is simplified to protein coiling: The continuum of solenoid protein structures," *Trends in Biochemical Sciences*, vol. 25, no. 10. pp. 509–515, 2000.

[168] M. D. Yoder, N. T. Keen, and F. Jurnak, "New domain motif: the structure of pectate lyase C, a secreted plant virulence factor," *Sci. (New York, NY)*, vol. 260, no. 5113, pp. 1503–1507, 1993.

[169] M. D. Yoder, S. E. Lietzke, and F. Jurnak, "Unusual structural features in the parallel ??-helix in pectate lyases," *Structure*, vol. 1, no. 4, pp. 241–251, 1993.

[170] D. Mohnen, "Pectin structure and biosynthesis," *Current Opinion in Plant Biology*, vol. 11, no. 3. pp. 266–277, 2008.

[171] A. Di Matteo, A. Giovane, A. Raiola, L. Camardella, D. Bonivento, G. De Lorenzo, F. Cervone, D. Bellincampi, and D. Tsernoglou, "Structural basis for the interaction between pectin methylesterase and a specific inhibitor protein," *Plant Cell*, vol. 17, no. 3, pp. 849–858, 2005.

[172] K. Johansson, M. El-Ahmad, R. Friemann, H. Jörnvall, O. Markovič, and H. Eklund, "Crystal structure of plant pectin methylesterase," *FEBS Lett.*, vol. 514, no. 2–3, pp. 243–249, 2002.

[173] M. A. McDonough, C. Ryttersgaard, M. E. Bjornvad, L. Lo Leggio, M. Schulein, S. O. Schroder Glad, and S. Larsen, "Crystallization and preliminary X-ray characterization of a thermostable pectate lyase from Thermotoga maritima," *Acta Crystallogr D Biol Crystallogr*, vol. 58, no. Pt 4, pp. 709–711, 2002.

[174] L. M. Kent, T. S. Loo, L. D. Melton, D. Mercadante, M. A. K. Williams, and G. B. Jameson, "Structure and properties of a non-processive, salt-requiring, and acidophilic pectin methylesterase from aspergillus Niger provide insights into the key determinants of processivity control," *J. Biol. Chem.*, vol. 291, no. 3, pp. 1289–1306, 2016.

[175] S. J. Dehdashti, C. N. Doan, K. L. Chao, and M. D. Yoder, "Effect of mutations in the T1.5 loop of pectate lyase A from Erwinia chrysanthemi EC16," *Acta Crystallogr. - Sect. D Biol. Crystallogr.*, vol. 59, no. 7, pp. 1339–1342, 2003.

[176] O. Mayans, M. Scott, I. Connerton, T. Gravesen, J. Benen, J. Visser, R. Pickersgill, and J. Jenkins, "Two crystal structures of pectin lyase A from Aspergillus reveal a pH driven conformational change and striking divergence in the substrate-binding clefts of pectin and pectate lyases.," *Structure*, vol. 5, no. 5, pp. 677–689, 1997.

[177] "Crystal structure of a putative immunoglobulin A1 protease from Bacteroides ovatus at 1.30 A resolution.," *PDB*, 2010. [Online]. Available: http://www.rcsb.org/pdb/explore/explore.do?structureId=3n6z. [Accessed: 08-Apr-2017].

[178] R. Pickersgill, J. Jenkins, G. Harris, W. Nasser, and J. Robert-Baudouy, "The structure of Bacillus subtilis pectate lyase in complex with calcium.," *Nat. Struct. Biol.*, vol. 1, no. 10, pp. 717–723, 1994.

[179] Z. Xiao, H. Bergeron, S. Grosse, M. Beauchemin, M.-L. Garron, D. Shaya, T. Sulea, M. Cygler, and P. C. K. Lau, "Improvement of the thermostability and activity of a pectate lyase by single amino acid substitutions, using a strategy based on melting-

temperature-guided sequence alignment.," *Appl. Environ. Microbiol.*, vol. 74, no. 4, pp. 1183–9, 2008.

[180] J. L. Miller, J. Le Coq, A. Hodes, R. Barbalat, J. F. Miller, and P. Ghosh, "Selective ligand recognition by a diversity-generating retroelement variable protein," *PLoS Biol.*, vol. 6, no. 6, pp. 1195–1207, 2008.

[181] R. Pickersgill, D. Smith, K. Worboys, and J. Jenkins, "Crystal structure of polygalacturonase from Erwinia caratovora ssp. carotovora," *J. Biol. Chem.*, vol. 273, no. 38, pp. 24660–24664, 1998.

[182] A. Freiberg, R. Morona, L. Van den Bosch, C. Jung, J. Behlke, N. Carlin, R. Seckler, and U. Baxa, "The tailspike protein of Shigella phage Sf6: A structural homolog of Salmonella phage P22 tailspike protein without sequence similarity in the β-helix domain," *J. Biol. Chem.*, vol. 278, no. 3, pp. 1542–1548, 2003.

[183] P. Emsley, I. G. Charles, N. F. Fairweather, and N. W. Isaacs, "Structure of Bordetella pertussis virulence factor P.69 pertactin," *Nature*, vol. 381, no. 6577. pp. 90–92, 1996.

[184] S. Steinbacher, U. Baxa, S. Miller, A. Weintraub, R. Seckler, and R. Huber, "Crystal structure of phage P22 tailspike protein complexed with Salmonella sp. O-antigen receptors," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 93, no. 20, pp. 10584–10588, 1996.

[185] T. Pijning, G. van Pouderoyen, L. Kluskens, J. van der Oost, and B. W. Dijkstra, "The crystal structure of a hyperthermoactive exopolygalacturonase from Thermotoga maritima reveals a unique tetramer," *FEBS Lett.*, vol. 583, no. 22, pp. 3665–3670, 2009.

[186] D. W. Abbott and A. B. Boraston, "The Structural Basis for Exopolygalacturonase Activity in a Family 28 Glycoside Hydrolase," *J. Mol. Biol.*, vol. 368, no. 5, pp. 1215–1222, 2007.

[187] W. S. Jung, C. K. Hong, S. Lee, C. S. Kim, S. J. Kim, S. Il Kim, and S. Rhee, "Structural and functional insights into intramolecular fructosyl transfer by inulin fructotransferase," *J. Biol. Chem.*, vol. 282, no. 11, pp. 8414–8423, 2007.

[188] D. W. Close, S. D'angelo, and A. R. M. Bradbury, "A new family of ??-helix proteins with similarities to the polysaccharide lyases," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 70, no. 10, pp. 2583–2592, 2014.

[189] H. J. Rozeboom, T. M. Bjerkan, K. H. Kalk, H. Ertesv??g, S. Holtan, F. L. Aachmann, S. Valla, and B. W. Dijkstra, "Structural and mutational characterization of the catalytic A-module of the mannuronan C-5-epimerase AlgE4 from Azotobacter vinelandii," *J. Biol. Chem.*, vol. 283, no. 35, pp. 23819–23828, 2008.

[190] a V Efimov, "Favoured structural motifs in globular proteins.," *Structure*, vol. 2, no. 11, pp. 999–1002, 1994.

[191] J. Jenkins and R. Pickersgill, "The architecture of parallel ??-helices and related folds," *Progress in Biophysics and Molecular Biology*, vol. 77, no. 2. pp. 111–175, 2001.

[192] F. E. Cohen, M. J. E. Sternberg, and W. R. Taylor, "Analysis of the tertiary structure of protein ??-sheet sandwiches," *J. Mol. Biol.*, vol. 148, no. 3, pp. 253–272, 1981.

[193] R. D. Scavetta, S. R. Herron, a T. Hotchkiss, N. Kita, N. T. Keen, J. a Benen, H. C. Kester, J. Visser, and F. Jurnak, "Structure of a plant cell wall fragment complexed to pectate lyase C.," *Plant Cell*, vol. 11, no. 6, pp. 1081–1092, 1999.

[194] J. M. Thornton, a E. Todd, D. Milburn, N. Borkakoti, and C. a Orengo, "From structure to function: approaches and limitations.," *Nat. Struct. Biol.*, vol. 7 Suppl, no. november, pp. 991–994, 2000.

[195] K. L. Hudson, G. J. Bartlett, R. C. Diehl, J. Agirre, T. Gallagher, L. L. Kiessling, and D. N. Woolfson, "Carbohydrate-Aromatic Interactions in Proteins," *J. Am. Chem. Soc.*, vol. 137, no. 48, pp. 15152–15160, 2015.

[196] W. Chen, S. Enck, J. L. Price, D. L. Powers, E. T. Powers, C. H. Wong, H. J. Dyson, and J. W. Kelly, "Structural and energetic basis of carbohydrate-aromatic packing

interactions in proteins," *J. Am. Chem. Soc.*, vol. 135, no. 26, pp. 9877–9884, 2013.

[197] C.-H. Hsu, S. Park, D. E. Mortenson, B. L. Foley, X. Wang, R. J. Woods, D. A. Case, E. T. Powers, C.-H. Wong, H. J. Dyson, and J. W. Kelly, "The Dependence of Carbohydrate–Aromatic Interaction Strengths on the Structure of the Carbohydrate," *J. Am. Chem. Soc.*, p. jacs.6b02879, 2016.

[198] E. Roberts, J. Eargle, D. Wright, and Z. Luthey-Schulten, "MultiSeq: unifying sequence and structure data for evolutionary analysis.," *BMC Bioinformatics*, vol. 7, p. 382, 2006.

[199] F. W. Ward and M. E. Coates, "Gastrointestinal pH measurement in rats: influence of the microbial flora, diet and fasting.," *Lab. Anim.*, vol. 21, no. 3, pp. 216–222, 1987.

[200] H. D. T. Mertens and D. I. Svergun, "Structural characterization of proteins and complexes using small-angle X-ray solution scattering," *J. Struct. Biol.*, vol. 172, no. 1, pp. 128–141, 2010.

[201] A. G. Kikhney and D. I. Svergun, "A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins," *FEBS Lett.*, vol. 589, no. 19, pp. 2570–2577, 2015.

[202] B. Liang and L. K. Tamm, "NMR as a tool to investigate the structure, dynamics and function of membrane proteins," *Nat. Struct. Mol. Biol.*, vol. 23, no. 6, pp. 468–474, 2016.

[203] J. L. S. Milne, M. J. Borgnia, A. Bartesaghi, E. E. H. Tran, L. A. Earl, D. M. Schauder, J. Lengyel, J. Pierson, A. Patwardhan, and S. Subramaniam, "Cryo-electron microscopy - A primer for the non-microscopist," *FEBS Journal*, vol. 280, no. 1. pp. 28–45, 2013.

[204] X. chen Bai, G. McMullan, and S. H. W. Scheres, "How cryo-EM is revolutionizing structural biology," *Trends in Biochemical Sciences*, vol. 40, no. 1. pp. 49–57, 2015.

[205] C. E. Blanchet, C. Hermes, D. I. Svergun, and S. Fiedler, "A small and robust active beamstop for scattering experiments on high-brilliance undulator beamlines," in *Journal of Synchrotron Radiation*, 2015, vol. 22, no. 2, pp. 461–464.

[206] M. Czjzek, H. P. Fierobe, and V. Receveur-Br??chot, "Small-angle x-ray scattering and crystallography: A winning combination for exploring the multimodular organization of cellulolytic macromolecular complexes," *Methods Enzymol.*, vol. 510, pp. 183–210, 2012.

[207] A. Guinier, "La diffraction des rayons X aux tres petits angles: applications a l'etude de phenomenes ultramicroscopiques," *Ann. Phys. (Paris).*, vol. 12, pp. 161–237, 1939.

[208] C. D. Putnam, M. Hammel, G. L. Hura, and J. a Tainer, "X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution.," *Q. Rev. Biophys.*, vol. 40, no. 3, pp. 191–285, 2007.

[209] S. Doniach, "Changes in biomolecular conformation seen by small angle X-ray scattering," *Chemical Reviews*, vol. 101, no. 6. pp. 1763–1778, 2001.

[210] A. V. Semenyuk and D. I. Svergun, "GNOM. A program package for small-angle scattering data processing," *J. Appl. Crystallogr.*, vol. 24, no. pt 5, pp. 537–540, 1991.

[211] D. I. Svergun, "Determination of the regularization parameter in indirect-transform methods using perceptual criteria," *J. Appl. Crystallogr.*, vol. 25, no. pt 4, pp. 495–503, 1992.

[212] D. Franke and D. I. Svergun, "DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering," *J. Appl. Crystallogr.*, vol. 42, no. 2, pp. 342–346, 2009.

[213] R. Rambo, "BIOISIS." [Online]. Available: http://www.bioisis.net/tutorial/9.

[214] EMBL, "BIOSAXS." [Online]. Available: http://www.embl-hamburg.de/biosaxs/software.html.

[215] V. V. Volkov and D. I. Svergun, “Uniqueness of ab initio shape determination in small-angle scattering,” in *Journal of Applied Crystallography*, 2003, vol. 36, no. 3 I, pp. 860–864.

[216] M. B. Kozin and D. I. Svergun, “Automated matching of high- and low-resolution structural models,” *J. Appl. Crystallogr.*, vol. 34, no. 1, pp. 33–41, 2001.

[217] D. Svergun, C. Barberato, and M. H. Koch, “CRYSOL - A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates,” *J. Appl. Crystallogr.*, vol. 28, no. 6, pp. 768–773, 1995.

[218] D. A. Jacques, J. M. Guss, D. I. Svergun, and J. Trewhella, “Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution,” in *Acta Crystallographica Section D: Biological Crystallography*, 2012, vol. 68, no. 6, pp. 620–626.

[219] G. Tria, H. D. T. Mertens, M. Kachala, and D. I. Svergun, “Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering,” *IUCrJ*, vol. 2, pp. 207–217, 2015.

[220] J. Yu, A. Penaloza-Vázquez, A. M. Chakrabarty, and C. L. Bender, “Involvement of the exopolysaccharide alginate in the virulence and epiphytic fitness of Pseudomonas syringae pv. syringae,” *Mol. Microbiol.*, vol. 33, no. 4, pp. 712–720, 1999.

[221] H. V. Scheller, J. K. Jensen, S. O. Sørensen, J. Harholt, and N. Geshi, “Biosynthesis of pectin,” *Physiologia Plantarum*, vol. 129, no. 2. pp. 283–295, 2007.

[222] M. D. Yoder and F. Jurnak, “Protein motifs. 3. The parallel beta helix and other coiled folds.,” *FASEB J.*, vol. 9, pp. 335–342, 1995.

[223] S. R. Herron, R. D. Scavetta, M. Garrett, M. Legner, and F. Jurnak, “Characterization and implications of Ca2+ binding to pectate lyase C,” *J. Biol. Chem.*, vol. 278, no. 14, pp. 12271–12277, 2003.

[224] A. Seyedarabi, T. T. To, S. Ali, S. Hussain, M. Fries, R. Madsen, M. H. Clausen, S. Teixteira, K. Brocklehurst, and R. W. Pickersgill, “Structural insights into substrate specificity and the anti beta-elimination mechanism of pectate lyase.,” *Biochemistry*, vol. 49, no. 3, pp. 539–46, 2010.

[225] Q. Tang, Y. P. Liu, Z. G. Ren, X. X. Yan, and L. Q. Zhang, “1.37 ?? Crystal structure of pathogenic factor pectate lyase from Acidovorax citrulli,” *Proteins Struct. Funct. Bioinforma.*, vol. 81, no. 8, pp. 1485–1490, 2013.

[226] J. Jenkins, V. E. Shevchik, N. Hugouvieux-Cotte-Pattat, and R. W. Pickersgill, “The Crystal Structure of Pectate Lyase Pel9A from Erwinia chrysanthemi,” *J. Biol. Chem.*, vol. 279, no. 10, pp. 9139–9145, 2004.

[227] C. Creze, S. Castang, E. Derivery, R. Haser, N. Hugouvieux-Cotte-Pattat, V. E. Shevchik, and P. Gouet, “The crystal structure of pectate lyase peli from soft rot pathogen Erwinia chrysanthemi in complex with its substrate.,” *J. Biol. Chem.*, vol. 283, no. 26, pp. 18260–18268, 2008.

[228] M. Akita, A. Suzuki, T. Kobayashi, S. Ito, and T. Yamane, “The first structure of pectate lyase belonging to polysaccharide lyase family 3,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 57, no. 12, pp. 1786–1792, 2001.

[229] M. Alahuhta, P. Chandrayan, I. Kataeva, M. W. W. Adams, M. E. Himmel, and V. V. Lunin, “A 1.5 Å resolution X-ray structure of the catalytic module of Caldicellulosiruptor bescii family 3 pectate lyase,” *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, vol. 67, no. 12, pp. 1498–1500, 2011.

[230] Y. Zheng, C. H. Huang, W. Liu, T. P. Ko, Y. Xue, C. Zhou, R. T. Guo, and Y. Ma, “Crystal structure and substrate-binding mode of a novel pectate lyase from alkaliphilic Bacillus sp. N16-5,” *Biochem. Biophys. Res. Commun.*, vol. 420, no. 2, pp. 269–274, 2012.

[231] J. Vitali, B. Schick, H. C. M. Kester, J. Visser, and F. Jurnak, “The three-dimensional structure of Aspergillus niger pectin lyase B at 1.7-angstrom resolution,” *Plant Physiol.*, vol. 116, no. 1, pp. 69–80, 1998.

[232] M. Fries, J. Ihrig, K. Brocklehurst, V. E. Shevchik, and R. W. Pickersgill, “Molecular basis of the activity of the phytopathogen pectin methylesterase.,” *EMBO J.*, vol. 26, no. 17, pp. 3879–87, 2007.

[233] J. Jenkins, O. Mayans, D. Smith, K. Worboys, and R. W. Pickersgill, “Three-dimensional structure of Erwinia chrysanthemi pectin methylesterase reveals a novel esterase active site.,” *J. Mol. Biol.*, vol. 305, pp. 951–60, 2001.

[234] A. B. Boraston and D. W. Abbott, “Structure of a pectin methylesterase from Yersinia enterocolitica,” *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, vol. 68, no. 2, pp. 129–133, 2012.

[235] T. Shimizu, T. Nakatsu, K. Miyairi, T. Okuno, and H. Kato, “Active-site architecture of endopolygalacturonase I from Stereum purpureum revealed by crystal structures in native and ligand-bound forms at atomic resolution,” *Biochemistry*, vol. 41, no. 21, pp. 6651–6659, 2002.

[236] T. N. Petersen, S. Kauppinen, and S. Larsen, “The crystal structure of rhamnogalacturonase A from Aspergillus aculeatus: a right-handed parallel beta helix.,” *Structure*, vol. 5, no. 4, pp. 533–544, 1997.

[237] H. J. Rozeboom, G. Beldman, H. A. Schols, and B. W. Dijkstra, “Crystal structure of endo-xylogalacturonan hydrolase from Aspergillus tubingensis,” *FEBS J.*, vol. 280, no. 23, pp. 6061–6069, 2013.

[238] W. Huang, a Matte, Y. Li, Y. S. Kim, R. J. Linhardt, H. Su, and M. Cygler, “Crystal structure of chondroitinase B from Flavobacterium heparinum and its complex with a disaccharide product at 1.7 A resolution.,” *J. Mol. Biol.*, vol. 294, no. 5, pp. 1257–69, 1999.

[239] G. Michel, K. Pojasek, Y. Li, T. Sulea, R. J. Linhardt, R. Raman, V. Prabhakar, R. Sasisekharan, and M. Cygler, “The structure of chondroitin B lyase complexed with glycosaminoglycan oligosaccharides unravels a calcium-dependent catalytic machinery,” *J. Biol. Chem.*, vol. 279, no. 31, pp. 32882–32896, 2004.

[240] A. M. Larsson, R. Andersson, J. Ståhlberg, L. Kenne, and T. A. Jones, “Dextranase from Penicillium minioluteum: Reaction course, crystal structure, and product complex,” *Structure*, vol. 11, no. 9, pp. 1111–1121, 2003.

[241] M. Mizuno, A. Koide, A. Yamamura, H. Akeboshi, H. Yoshida, S. Kamitori, Y. Sakano, A. Nishikawa, and T. Tonozuka, “Crystal Structure of Aspergillus niger Isopullulanase, a Member of Glycoside Hydrolase Family 49,” *J. Mol. Biol.*, vol. 376, no. 1, pp. 210–220, 2008.

[242] C.-H. Huang, Y. Sun, T.-P. Ko, C.-C. Chen, Y. Zheng, H.-C. Chan, X. Pang, J. Wiegel, W. Shao, and R.-T. Guo, “The substrate/product binding modes of a novel GH120 β-xylosidase (XylC) from Thermoanaerobacterium saccharolyticum JW/SL-YS485,” *Biochem. J.*, vol. 407, pp. 401–407, 2012.

[243] T. Ishida, S. Fushinobu, R. Kawai, M. Kitaoka, K. Igarashi, and M. Samejima, “Crystal structure of glycoside hydrolase family 55 {beta}-1,3-glucanase from the basidiomycete Phanerochaete chrysosporium.,” *J. Biol. Chem.*, vol. 284, no. 15, pp. 10100–10109, 2009.

[244] G. Michel, L. Chantalat, E. Fanchon, B. Henrissat, B. Kloareg, and O. Dideberg, “The ??-carrageenase of Alteromonas fortis: A ??-helix fold-containing enzyme for the degradation of a highly polyanionic polysaccharide,” *J. Biol. Chem.*, vol. 276, no. 43, pp. 40202–40209, 2001.

[245] G. Michel, W. Helbert, R. Kahn, O. Dideberg, and B. Kloareg, “The structural bases of the processive degradation of ??-carrageenan, a main cell wall polysaccharide of red algae,” *J. Mol. Biol.*, vol. 334, no. 3, pp. 421–433, 2003.

[246] E. Rebuffet, T. Barbeyron, A. Jeudy, M. Jam, M. Czjzek, and G. Michel, “Identification of catalytic residues and mechanistic analysis of family GH82 ???-carrageenases,” *Biochemistry*, vol. 49, no. 35, pp. 7590–7599, 2010.

[247] F. Wolfram, E. N. Kitova, H. Robinson, M. T. C. Walvoort, J. D. C. Cod??e, J. S. Klassen, and P. L. Howell, “Catalytic mechanism and mode of action of the

periplasmic alginate epimerase AlgG," *J. Biol. Chem.*, vol. 289, no. 9, pp. 6006–6019, 2014.

[248] M. D. Suits and A. B. Boraston, "Structure of the Streptococcus pneumoniae Surface Protein and Adhesin PfbA," *PLoS One*, vol. 8, no. 7, 2013.

[249] D. S. J. Beulin, M. Yamaguchi, S. Kawabata, and K. Ponnuraj, "Crystal structure of PfbA, a surface adhesin of Streptococcus pneumoniae, provides hints into its interaction with fibronectin," *Int. J. Biol. Macromol.*, vol. 64, pp. 168–173, 2014.

[250] B. Clantin, H. Hodak, E. Willery, C. Locht, F. Jacob-Dubuisson, and V. Villeret, "The crystal structure of filamentous hemagglutinin secretion domain and its implications for the two-partner secretion pathway," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 16, pp. 6194–6199, 2004.

[251] E. W. Czerwinski, T. Midoro-Horiuti, M. A. White, E. G. Brooks, and R. M. Goldblum, "Crystal structure of jun a 1, the major cedar pollen allergen from Juniperus ashei, reveals a parallel ??-helical core," *J. Biol. Chem.*, vol. 280, no. 5, pp. 3740–3746, 2005.

[252] Y. Xiang, P. G. Leiman, L. Li, S. Grimes, D. L. Anderson, and M. G. Rossmann, "Crystallographic Insights into the Autocatalytic Assembly Mechanism of a Bacteriophage Tail Spike," *Mol. Cell*, vol. 34, no. 3, pp. 375–386, 2009.

[253] J. L. Asensio, A. Ardá, F. J. Cañada, and J. Jiménez-Barbero, "Carbohydrate-aromatic interactions," *Acc. Chem. Res.*, vol. 46, no. 4, pp. 946–954, 2013.

[254] C. Frantz, K. M. Stewart, and V. M. Weaver, "The extracellular matrix at a glance.," *J. Cell Sci.*, vol. 123, pp. 4195–4200, 2010.

[255] R. Sengupta, E. Altermann, R. C. Anderson, W. C. McNabb, P. J. Moughan, and N. C. Roy, "The role of cell surface architecture of lactobacilli in host-microbe interactions in the gastrointestinal tract," *Mediators of Inflammation*, vol. 2013. 2013.

[256] C. Schwab, J. Walter, G. W. Tannock, R. F. Vogel, and M. G. Gänzle, "Sucrose utilization and impact of sucrose on glycosyltransferase expression in Lactobacillus reuteri," *Syst. Appl. Microbiol.*, vol. 30, no. 6, pp. 433–443, 2007.

[257] I. M. Sims, S. A. Frese, J. Walter, D. Loach, M. Wilson, K. Appleyard, J. Eason, M. Livingston, M. Baird, G. Cook, and G. W. Tannock, "Structure and functions of exopolysaccharide produced by gut commensal Lactobacillus reuteri 100-23," *Isme J*, vol. 5, no. 7, pp. 1115–1124, 2011.

[258] N. Hugouvieux-Cotte-Pattat, G. Condemine, and V. E. Shevchik, "Bacterial pectate lyases, structural and functional diversity," *Environmental Microbiology Reports*, vol. 6, no. 5. pp. 427–440, 2014.

[259] A. El Kaoutari, F. Armougom, J. I. Gordon, D. Raoult, and B. Henrissat, "The abundance and variety of carbohydrate-active enzymes in the human gut microbiota.," *Nat. Rev. Microbiol.*, vol. 11, no. 7, pp. 497–504, 2013.

[260] P. Yuan, K. Meng, Y. Wang, H. Luo, H. Huang, P. Shi, Y. Bai, P. Yang, and B. Yao, "Abundance and genetic diversity of microbial polygalacturonase and pectate lyase in the sheep rumen ecosystem," *PLoS One*, vol. 7, no. 7, 2012.

[261] Ê. V. F. Jubb, P. C. Kennedy, and N. Palmer, *Pathology of Domestic Animals: Fourth Edition*, vol. 1. 2013.

[262] L. Brown, J. M. Wolf, R. Prados-Rosales, and A. Casadevall, "Through the wall: extracellular vesicles in Gram-positive bacteria, mycobacteria and fungi.," *Nat. Rev. Microbiol.*, vol. 13, no. 10, pp. 620–30, 2015.

[263] M. Beeby, J. C. Gumbart, B. Roux, and G. J. Jensen, "Architecture and assembly of the Gram-positive cell wall," *Mol. Microbiol.*, vol. 88, no. 4, pp. 664–672, 2013.

[264] T. J. Beveridge, "Use of the gram stain in microbiology.," *Biotech. Histochem.*, vol. 76, no. 3, pp. 111–118, 2001.

[265] C. R. Raetz and C. Whitfield, "Lipopolysaccharide endotoxins," *Annu Rev Biochem*, vol. 71, pp. 635–700, 2002.

[266] S. E. Rollauer, M. A. Sooreshjani, N. Noinaj, and S. K. Buchanan, “Outer membrane protein biogenesis in Gram-negative bacteria.,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 370, no. 1679, p. 20150023-, 2015.

[267] Y. Hirota, H. Suzuki, Y. Nishimura, and S. Yasuda, “On the process of cellular division in Escherichia coli: a mutant of E. coli lacking a murein-lipoprotein.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 4, pp. 1417–20, 1977.

[268] H. Suzuki, Y. Nishimura, S. Yasuda, A. Nishimura, M. Yamada, and Y. Hirota, “Murein-lipoprotein of Escherichia coli: A protein involved in the stabilization of bacterial cell envelope,” *Mol. Gen. Genet. MGG*, vol. 167, no. 1, pp. 1–9, 1978.

[269] H. Dong, Q. Xiang, Y. Gu, Z. Wang, N. G. Paterson, P. J. Stansfeld, C. He, Y. Zhang, W. Wang, and C. Dong, “Structural basis for outer membrane lipopolysaccharide insertion.,” *Nature*, vol. 511, pp. 52–6, 2014.

[270] Y. Gu, H. Li, H. Dong, Y. Zeng, Z. Zhang, N. G. Paterson, P. J. Stansfeld, Z. Wang, Y. Zhang, W. Wang, and C. Dong, “Structural basis of outer membrane protein insertion by the BAM complex,” *Nature*, vol. advance on, no. 2, pp. 113–122, 2016.

[271] A. H. Delcour, “Outer membrane permeability and antibiotic resistance.,” *Biochim. Biophys. Acta*, vol. 1794, no. 5, pp. 808–816, 2009.

[272] J. W. Fairman, N. Noinaj, and S. K. Buchanan, “The structural biology of beta-barrel membrane proteins: a summary of recent reports,” *Curr Opin Struct Biol*, vol. 21, no. 4, pp. 523–531, 2011.

[273] C. W. Mullineaux, A. Nenninger, N. Ray, and C. Robinson, “Diffusion of green fluorescent protein in three cell environments in Escherichia coli,” *J. Bacteriol.*, vol. 188, no. 10, pp. 3442–3448, 2006.

[274] R. De Duve, Christian and Wattiaux, “FUNCTIONS OF LYSOSOMESl,” *Annu. Rev. Physiol.*, vol. 28, no. September, pp. 435–492, 1966.

[275] T. J. Silhavy, D. Kahne, and S. Walker, “The bacterial cell envelope.,” *Cold Spring Harbor perspectives in biology*, vol. 2, no. 5. 2010.

[276] M. P. Bos, V. Robert, and J. Tommassen, “Biogenesis of the gram-negative bacterial outer membrane.,” *Annu. Rev. Microbiol.*, vol. 61, pp. 191–214, 2007.

[277] R. E. Dalbey, P. Wang, and A. Kuhn, “Assembly of bacterial inner membrane proteins.,” *Annu. Rev. Biochem.*, vol. 80, pp. 161–187, 2011.

[278] P. F. Mühlradt and J. R. Golecki, “Asymmetrical distribution and artifactual reorientation of lipopolysaccharide in the outer membrane bilayer of Salmonella typhimurium.,” *Eur. J. Biochem.*, vol. 51, no. 2, pp. 343–352, 1975.

[279] M. Matsuura, “Structural modifications of bacterial lipopolysaccharide that facilitate gram-negative bacteria evasion of host innate immunity,” *Frontiers in Immunology*, vol. 4, no. MAY. 2013.

[280] A. M. Delucia, D. A. Six, R. E. Caughlan, P. Gee, I. Hunt, J. S. Lam, and C. R. Dean, “Lipopolysaccharide (LPS) inner-core phosphates are required for complete lps synthesis and transport to the outer membrane in pseudomonas aeruginosa PAO1,” *MBio*, vol. 2, no. 4, 2011.

[281] J. A. Yethon, D. E. Heinrichs, M. A. Monteiro, M. B. Perry, and C. Whitfield, “Involvement of waaY, waaQ, and waaP in the modification of Escherichia coli lipopolysaccharide and their role in the formation of a stable outer membrane,” *J. Biol. Chem.*, vol. 273, no. 41, pp. 26310–26316, 1998.

[282] J. A. Yethon, J. S. Gunn, R. K. Ernst, S. I. Miller, L. Laroche, D. Malo, and C. Whitfield, “Salmonella enterica serovar Typhimurium waaP mutants show increased susceptibility to polymyxin and loss of virulence in vivo,” *Infect. Immun.*, vol. 68, no. 8, pp. 4485–4491, 2000.

[283] J. Nesper, C. M. Lauriano, K. E. Klose, D. Kapfhammer, A. Krai??, and J. Reidl, “Characterization of Vibrio cholerae O1 El Tor galU and galE mutants: Influence on lipopolysaccharide structure, colonization, and biofilm formation,” *Infect. Immun.*, vol. 69, no. 1, pp. 435–445, 2001.

[284] J. Nesper, S. Schild, C. M. Lauriano, A. Kraiss, K. E. Klose, and J. Reidl, “Role of Vibrio cholerae O139 surface polysaccharides in intestinal colonization,” *Infect. Immun.*, vol. 70, no. 11, pp. 5990–5996, 2002.

[285] W. T. Doerrler, H. S. Gibbons, R. Christian, and H. Raetz, “MsbA-dependent translocation of lipids across the inner membrane of Escherichia coli,” *J. Biol. Chem.*, vol. 279, no. 43, pp. 45102–45109, 2004.

[286] Z. Zhou, K. A. White, A. Polissi, C. Georgopoulos, and C. R. H. Raetz, “Function of Escherichia coli MsbA, an essential ABC family transporter, in lipid A and phospholipid biosynthesis,” *J. Biol. Chem.*, vol. 273, no. 20, pp. 12466–12475, 1998.

[287] A. Polissi and C. Georgopoulos, “Mutational analysis and properties of the msbA gene of Escherichia coli, coding for an essential ABC family transporter,” *Mol. Microbiol.*, vol. 20, no. 6, pp. 1221–1233, 1996.

[288] M. J. Osborn, J. E. Gander, and E. Parisi, “Mechanism of assembly of the outer membrane of Salmonella typhimurium. Site of synthesis of lipopolysaccharide.,” *J. Biol. Chem.*, vol. 247, no. 12, pp. 3973–3986, 1972.

[289] M. J. Osborn, J. E. Gander, E. Parisi, and J. Carson, “Mechinism of assembly of the outer membrane of Salmonella typhimurium. Isolation and characterization of cytoplasmic and outer membrane.,” *J. Biol. Chem.*, vol. 247, no. 12, pp. 3962–3972, 1972.

[290] S. S. Chng, L. S. Gronenberg, and D. Kahne, “Proteins required for lipopolysaccharide assembly in escherichia coli form a transenvelope complex,” *Biochemistry*, vol. 49, no. 22, pp. 4565–4567, 2010.

[291] S. Okuda, E. Freinkman, and D. Kahne, “Cytoplasmic ATP hydrolysis powers transport of lipopolysaccharide across the periplasm in E. coli.,” *Science*, vol. 338, no. 6111, pp. 1214–7, 2012.

[292] R. Villa, A. M. Martorana, S. Okuda, L. J. Gourlay, M. Nardini, P. Sperandeo, G. Dehò, M. Bolognesi, D. Kahne, and A. Polissia, “The Escherichia coli lpt transenvelope protein complex for lipopolysaccharide export is assembled via conserved structurally homologous domains,” *J. Bacteriol.*, vol. 195, no. 5, pp. 1100–1108, 2013.

[293] M. D. L. Suits, P. Sperandeo, G. Deh??, A. Polissi, and Z. Jia, “Novel Structure of the Conserved Gram-Negative Lipopolysaccharide Transport Protein A and Mutagenesis Analysis,” *J. Mol. Biol.*, vol. 380, no. 3, pp. 476–488, 2008.

[294] B. W. Simpson, J. M. May, D. J. Sherman, D. Kahne, and N. Ruiz, “Lipopolysaccharide transport to the cell surface: biosynthesis and extraction from the inner membrane.,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 370, p. 20150029, 2015.

[295] X. Wang and P. J. Quinn, “Lipopolysaccharide: Biosynthetic pathway and structure modification,” *Progress in Lipid Research*, vol. 49, no. 2. pp. 97–107, 2010.

[296] S. M??ller-Loennies, B. Lindner, and H. Brade, “Structural analysis of deacylated lipopolysaccharide of Escherichia coli strains 2513 (R4 core-type) and F653 (R3 core-type),” *Eur. J. Biochem.*, vol. 269, no. 23, pp. 5982–5991, 2002.

[297] S. Müller-Loennies, B. Lindner, and H. Brade, “Structural Analysis of Oligosaccharides from Lipopolysaccharide (LPS) of Escherichia coli K12 Strain W3100 Reveals a Link between Inner and Outer Core LPS Biosynthesis,” *J. Biol. Chem.*, vol. 278, no. 36, pp. 34090–34101, 2003.

[298] C. Roncero and M. J. Casadaban, “Genetic analysis of the genes involved in synthesis of the lipopolysaccharide core in Escherichia coli K-12: Three operons in the rfa locus,” *J. Bacteriol.*, vol. 174, no. 10, pp. 3250–3260, 1992.

[299] M. McClelland, L. Florea, K. Sanderson, S. W. Clifton, J. Parkhill, C. Churcher, G. Dougan, R. K. Wilson, and W. Miller, “Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi.,” *Nucleic Acids Res.*, vol. 28, no. 24, pp. 4974–4986, 2000.

[300] J. C. Pegues, L. S. Chen, a W. Gordon, L. Ding, and W. G. Coleman, “Cloning, expression, and characterization of the Escherichia coli K-12 rfaD gene.,” *J. Bacteriol.*, vol. 172, no. 8, pp. 4652–60, 1990.

[301] C. A. Schnaitman and J. D. Klena, “Genetics of lipopolysaccharide biosynthesis in enteric bacteria.,” *Microbiol. Rev.*, vol. 57, no. 3, pp. 655–82, 1993.

[302] W. Han, B. Wu, L. Li, G. Zhao, R. Woodward, N. Pettit, L. Cai, V. Thon, and P. G. Wang, “Defining function of lipopolysaccharide O-antigen ligase waaI using chemoenzymatically synthesized substrates,” *J. Biol. Chem.*, vol. 287, no. 8, pp. 5357–5365, 2012.

[303] T. Clementz, “The gene coding for 3-deoxy-manno-octulosonic acid transferase and the rfaQ gene are transcribed from divergently arranged promoters in Escherichia coli,” *J. Bacteriol.*, vol. 174, no. 23, pp. 7750–7756, 1992.

[304] K. Amor, D. E. Heinrichs, E. Frirdich, K. Ziebell, R. P. Johnson, and C. Whitfield, “Distribution of core oligosaccharide types in lipopolysaccharides from Escherichia coli,” *Infect. Immun.*, vol. 68, no. 3, pp. 1116–1124, 2000.

[305] K. Shibayama, S. Ohsuka, T. Tanaka, Y. Arakawa, and M. Ohta, “Conserved structural regions involved in the catalytic mechanism of Escherichia coli K-12 WaaO (RfaI),” *J. Bacteriol.*, vol. 180, no. 20, pp. 5313–5318, 1998.

[306] D. E. Heinrichs, J. A. Yethon, P. A. Amor, and C. Whitfield, “The assembly system for the outer core portion of R1- and R4-type lipopolysaccharides of Escherichia coli: The R1 core-specific ??- glucosyltransferase provides a novel attachment site for O-polysaccharides,” *J. Biol. Chem.*, vol. 273, no. 45, pp. 29497–29505, 1998.

[307] C. Whitfield, N. Kaniuk, and E. Frirdich, “Molecular insights into the assembly and diversity of the outer core oligosaccharide in lipopolysaccharides from Escherichia coli and Salmonella.,” *J. Endotoxin Res.*, vol. 9, no. 4, pp. 244–249, 2003.

[308] B. Kneidinger, C. Marolda, M. Graninger, A. Zamyatina, F. McArthur, P. Kosma, M. A. Valvano, and P. Messner, “Biosynthesis pathway of ADP-L-glycero-beta-D-manno-heptose in Escherichia coli.,” *J. Bacteriol.*, vol. 184, no. 2, pp. 363–9, 2002.

[309] D. M. Sirisena, P. R. MacLachlan, S. L. Liu, A. Hessel, and K. E. Sanderson, “Molecular analysis of the rfaD gene, for heptose synthesis, and the rfaF gene, for heptose transfer, in lipopolysaccharide synthesis in Salmonella typhimurium,” *J. Bacteriol.*, vol. 176, no. 8, pp. 2379–2385, 1994.

[310] L. Chen and W. G. Coleman, “Cloning and characterization of the Escherichia coli K-12 rfa-2 (rfaC) gene, a gene required for lipopolysaccharide inner core synthesis,” *J. Bacteriol.*, vol. 175, no. 9, pp. 2534–2540, 1993.

[311] D. M. Sirisena, K. A. Brozek, P. R. MacLachlan, K. E. Sanderson, and C. R. H. Raetz, “The rfaC gene of Salmonella typhimurium: Cloning, sequencing, and enzymatic function in heptose transfer to lipopolysaccharide,” *J. Biol. Chem.*, vol. 267, no. 26, pp. 18874–18884, 1992.

[312] T. R. De Kievit and J. S. Lam, “Isolation and characterization of two genes, waaC (rfaC) and waaF (rfaF), involved in Pseudomonas aeruginosa serotype 05 inner-core biosynthesis,” *J. Bacteriol.*, vol. 179, no. 11, pp. 3451–3457, 1997.

[313] C. T. Parker, A. W. Kloser, C. A. Schnaitman, M. A. Stein, S. Gottesman, and B. W. Gibson, “Role of the rfaG and rfaP genes in determining the lipopolysaccharide core structure and cell surface properties of Escherichia coli K-12,” *J. Bacteriol.*, vol. 174, no. 8, pp. 2525–2538, 1992.

[314] I. M. Helander, I. Kilpeläinen, and M. Vaara, “Phosphate groups in lipopolysaccharides of Salmonella typhimurium rfaP mutants,” *FEBS Lett.*, vol. 409, no. 3, pp. 457–460, 1997.

[315] C. T. Parker, E. Pradel, and C. A. Schnaitman, “Identification and sequences of the lipopolysaccharide core biosynthetic genes rfaQ, rfaP, and rfaG of Escherichia coli K-12,” *J. Bacteriol.*, vol. 174, no. 3, pp. 930–934, 1992.

[316] P. R. MacLachlan, S. K. Kadam, and K. E. Sanderson, “Cloning, characterization,

and DNA sequence of the rfaLK region for lipopolysaccharide synthesis in Salmonella typhimurium LT2," *J. Bacteriol.*, vol. 173, no. 22, pp. 7151–7163, 1991.

[317] S. K. Kadam, A. Rehemtulla, and K. E. Sanderson, "Cloning of rfaG, B, I, and J genes for glycosyltransferase enzymes for synthesis of the lipopolysaccharide core of Salmonella typhimurium," *J. Bacteriol.*, vol. 161, no. 1, pp. 277–284, 1985.

[318] E. Pradel, C. T. Parker, and C. A. Schnaitman, "Structures of the rfaB, rfaI, rfaJ, and rfaS genes of Escherichia coli K- 12 and their roles in assembly of the lipopolysaccharide core," *J. Bacteriol.*, vol. 174, no. 14, pp. 4736–4745, 1992.

[319] K. Shibayama, S. Ohsuka, K. Sato, K. Yokoyama, T. Horii, and M. Ohta, "Four critical aspartic acid residues potentially involved in the catalytic mechanism of Escherichia coli K-12 WaaR," *FEMS Microbiol. Lett.*, vol. 174, no. 1, pp. 105–109, 1999.

[320] D. E. Heinrichs, M. A. Monteiro, M. B. Perry, and C. Whitfield, "The assembly system for the lipopolysaccharide R2 core-type of Escherichia coli is a hybrid of those found in Escherichia coli K-12 and Salmonella enterica. Structure and function of the R2 WaaK and WaaL homologs.," *J. Biol. Chem.*, vol. 10, no. 273, pp. 8849–59, 1998.

[321] Z. Wang, J. Wang, G. Ren, Y. Li, and X. Wang, "Influence of core oligosaccharide of lipopolysaccharide to outer membrane behavior of Escherichia coli.," *Mar. Drugs*, vol. 13, no. 6, pp. 3325–3339, 2015.

[322] G. Klein, B. Lindner, H. Brade, and S. Raina, "Molecular basis of lipopolysaccharide heterogeneity in Escherichia coli: Envelope stress-responsive regulators control the incorporation of glycoforms with a third 3-deoxy-α-D-manno-oct-2-ulosonic acid and rhamnose," *J. Biol. Chem.*, vol. 286, no. 50, pp. 42787–42807, 2011.

[323] C. J. Belunis and C. R. H. Raetz, "Biosynthesis of endotoxins: Purification and catalytic properties of 3-deoxy-D-manno-octulosonic acid transferase from Escherichia coli," *J. Biol. Chem.*, vol. 267, no. 14, pp. 9988–9997, 1992.

[324] D. Kocincova and J. S. Lam, "Structural diversity of the core oligosaccharide domain of Pseudomonas aeruginosa lipopolysaccharide," *Biochem.*, vol. 76, no. 7, pp. 755–760, 2011.

[325] O. V. Bystrova, Y. A. Knirel, B. Lindner, N. A. Kocharova, A. N. Kondakova, U. Zähringer, and G. B. Pier, "Structures of the core oligosaccharide and O-units in the R- and SR-type lipopolysaccharides of reference strains of Pseudomonas aeruginosa O-serogroups," *FEMS Immunol. Med. Microbiol.*, vol. 46, no. 1, pp. 85–99, 2006.

[326] A. G. Walsh, M. J. Matewish, L. L. Burrows, M. A. Monteiro, M. B. Perry, and J. S. Lam, "Lipopolysaccharide core phosphates are required for viability and intrinsic drug resistance in Pseudomonas aeruginosa," *Mol. Microbiol.*, vol. 35, no. 4, pp. 718–727, 2000.

[327] D. Perumal, K. R. Sakharkar, T. H. Tang, V. T. K. Chow, C. S. Lim, A. Samal, N. Sugiura, and M. K. Sakharkar, "Cloning and targeted disruption of two lipopolysaccharide biosynthesis genes, kdsA and waaG, of pseudomonas aeruginosa PAO1 by site-directed mutagenesis," *J. Mol. Microbiol. Biotechnol.*, vol. 19, no. 4, pp. 169–179, 2011.

[328] K. K. H. Poon, E. L. Westman, E. Vinogradov, S. Jin, and J. S. Lam, "Functional characterization of MigA and WapR: Putative rhamnosyltransferases involved in outer core oligosaccharide biosynthesis of Pseudomonas aeruginosa," *J. Bacteriol.*, vol. 190, no. 6, pp. 1857–1865, 2008.

[329] D. Kocíncová, S. L. Ostler, E. M. Anderson, and J. S. Lam, "Rhamnosyltransferase genes migA and wapR are regulated in a differential manner to modulate the quantities of core oligosaccharide glycoforms produced by Pseudomonas aeruginosa," *J. Bacteriol.*, vol. 194, no. 16, pp. 4295–4300, 2012.

[330] P. D. Abeyrathne, C. Daniels, K. K. H. Poon, M. J. Matewish, and J. S. Lam, "Functional characterization of WaaL, a ligase associated with linking O-antigen polysaccharide to the core of Pseudomonas aeruginosa lipopolysaccharide," *J. Bacteriol.*, vol. 187, no. 9, pp. 3002–3012, 2005.

[331] D. Kocíncová, Y. Hao, E. Vinogradov, and J. S. Lam, "Evidence that WapB Is a 1,2-

glucosyltransferase of Pseudomonas aeruginosa involved in lipopolysaccharide outer core biosynthesis," *J. Bacteriol.*, vol. 193, no. 11, pp. 2708–2716, 2011.

[332] H. Nikaido, "Molecular basis of bacterial outer membrane permeability revisited.," *Microbiol. Mol. Biol. Rev.*, vol. 67, no. 4, pp. 593–656, 2003.

[333] J. D. Klena, E. Pradel, and C. A. Schnaitman, "Comparison of lipopolysaccharide biosynthesis genes rfaK, rfaL, rfaY, and rfaZ of Escherichia coli K-12 and Salmonella typhimurium," *J. Bacteriol.*, vol. 174, no. 14, pp. 4746–4752, 1992.

[334] I. M. Helander, M. Vaara, S. Sukupolvi, M. Rhen, S. Saarela, U. Zahringer, and P. H. Makela, "rfaP mutants of Salmonella typhimurium," *Eur. J. Biochem.*, vol. 185, no. 3, pp. 541–546, 1989.

[335] M. Rathman, N. Jouirhi, A. Allaoui, P. Sansonetti, C. Parsot, and G. T. Van Nhieu, "The development of a FACS-based strategy for the isolation of Shigella flexneri mutants that are deficient in intercellular spread," *Mol. Microbiol.*, vol. 35, no. 5, pp. 974–990, 2000.

[336] M. Vaara, "Agents that increase the permeability of the outer membrane.," *Microbiol. Rev.*, vol. 56, no. 3, pp. 395–411, 1992.

[337] H. De Cock, K. Brandenburg, A. Wiesen, O. Holst, and U. Seydel, "Non-lamellar structure and negative charges of lipopolysaccharides required for efficient folding of outer membrane protein PhoE of Escherichia coli," *J. Biol. Chem.*, vol. 274, no. 8, pp. 5114–5119, 1999.

[338] R. E. W. Hancock, "Resistance Mechanisms in Pseudomonas aeruginosa and Other Nonfermentative Gram- Negative Bacteria," *Clin. Infect. Dis.*, vol. 27, pp. s93–s99, 1998.

[339] J. R. Aeschlimann, "The role of multidrug efflux pumps in the antibiotic resistance of Pseudomonas aeruginosa and other gram-negative bacteria. Insights from the Society of Infectious Diseases Pharmacists.," *Pharmacotherapy*, vol. 23, no. 7, pp. 916–924, 2003.

[340] R. REPASKE, "Lysis of gram-negative bacteria by lysozyme.," *Biochim. Biophys. Acta*, vol. 22, no. 1, pp. 189–91, Oct. 1956.

[341] G. W. GRAY and S. G. WILKINSON, "The Effect of Ethylenediaminetetra-acetic Acid on the Cell Walls of Some Gram-Negative Bacteria," *Microbiology*, vol. 39, no. 3, pp. 385–399, 1965.

[342] H. M. Ayres, J. R. Furr, and a D. Russell, "Effect of divalent cations on permeabilizer-induced lysozyme lysis of Pseudomonas aeruginosa.," *Lett. Appl. Microbiol.*, vol. 27, no. 6, pp. 372–4, 1998.

[343] B. Choudhury, R. W. Carlson, and J. B. Goldberg, "The structure of the lipopolysaccharide from a galU mutant of Pseudomonas aeruginosa serogroup-O11," *Carbohydr. Res.*, vol. 340, no. 18, pp. 2761–2772, 2005.

[344] L. Zhang, J.-C. Wang, L. Hou, P.-R. Cao, L. Wu, Q.-S. Zhang, H.-Y. Yang, Y. Zang, J.-P. Ding, and J. Li, "Functional Role of Histidine in the Conserved His-x-Asp Motif in the Catalytic Core of Protein Kinases.," *Sci. Rep.*, vol. 5, p. 10115, 2015.

[345] A. P. Kornev, N. M. Haste, S. S. Taylor, and L. F. Ten Eyck, "Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 47, pp. 17783–17788, 2006.

[346] X. Zhao and J. S. Lam, "WaaP of Pseudomonas aeruginosa is a novel eukaryotic type protein-tyrosine kinase as well as a sugar kinase essential for the biosynthesis of core lipopolysaccharide," *J. Biol. Chem.*, vol. 277, no. 7, pp. 4722–4730, 2002.

[347] K. C. Duong-Ly and S. B. Gabelli, "Gel filtration chromatography (size exclusion chromatography) of proteins," *Methods Enzymol.*, vol. 541, pp. 105–114, 2014.

[348] D. C. Gu, Yinghong, Li H, Dong H, Zeng Y, Zhang Z, Paterson NG, Stansfeld PJ, Wang Z, Zhang Y, Wang W, "Structural basis of outer membrane protein insertion by the BAM complex.," *Nature*, vol. 531, no. 7592, pp. 64–69, 2016.

[349] G. D. Wallat, Q. Huang, W. Wang, H. Dong, H. Ly, Y. Liang, and C. Dong, "High-

resolution structure of the N-terminal endonuclease domain of the Lassa virus L polymerase in complex with magnesium ions," *PLoS One*, vol. 9, no. 2, 2014.

[350] X. Jiang, Q. Huang, W. Wang, H. Dong, H. Ly, Y. Liang, and C. Dong, "Structures of arenaviral nucleoproteins with triphosphate dsRNA reveal a unique mechanism of immune suppression," *J. Biol. Chem.*, vol. 288, no. 23, pp. 16949–16959, 2013.

[351] J. Bornhost, Joshua; Falke, "Purification of Proteins Using Polyhistidine Affinity Tags," *Methods Enzymol.*, vol. 326, pp. 245–254, 2000.

[352] P. Sun, J. E. Tropea, and D. S. Waugh, "Enhancing the solubility of recombinant proteins in escherichia coli by using hexahistidine-tagged maltose-binding protein as a fusion partner," *Methods Mol. Biol.*, vol. 705, pp. 259–274, 2011.

[353] R. B. Kapust and D. S. Waugh, "Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused.," *Protein Sci.*, vol. 8, no. 8, pp. 1668–1674, 1999.

[354] B. P. Austin, S. Nallamsetty, and D. S. Waugh, "Hexahistidine-tagged maltose-binding protein as a fusion partner for the production of soluble recombinant proteins in Escherichia coli," *Methods Mol. Biol.*, vol. 498, pp. 157–172, 2009.

[355] S. Raran-Kurussi, K. Keefe, and D. S. Waugh, "Positional effects of fusion partners on the yield and solubility of MBP fusion proteins," *Protein Expr. Purif.*, vol. 110, pp. 159–164, 2015.

[356] P. A. Timmins, J. Hauk, T. Wacker, and W. Welte, "The influence of heptane-1,2,3-triol on the size and shape of LDAO micelles Implications for the crystallisation of membrane proteins," *FEBS Lett.*, vol. 280, no. 1, pp. 115–120, 1991.

[357] M. Zhou, J. H. Morais-Cabral, S. Mann, and R. MacKinnon, "Potassium channel receptor site for the inactivation gate and quaternary amine inhibitors.," *Nature*, vol. 411, no. 6838, pp. 657–661, 2001.

[358] D. D. Shultis, M. D. Purdy, C. N. Banchs, and M. C. Wiener, "Outer Membrane Active Transport: Structure of the BtuB:TonB Complex," *Science (80-.).*, vol. 312, no. 5778, pp. 1396–1399, 2006.

[359] A. M. Seddon, P. Curnow, and P. J. Booth, "Membrane proteins, lipids and detergents: Not just a soap opera," *Biochimica et Biophysica Acta - Biomembranes*, vol. 1666, no. 1–2. pp. 105–117, 2004.

[360] K. W. Herrmann, "NON-IONIC—CATIONIC MICELLAR PROPERTIES OF DIMETHYLDODECYLAMINE OXIDE," *J. Phys. Chem.*, vol. 66, no. 2, pp. 295–300, 1962.

[361] A. P. Golovanov, G. M. Hautbergue, S. A. Wilson, and L. Lian, "A simple method for improving protein solubility and long-term stability.," *J. Am. Chem. Soc.*, vol. 126, no. 29, pp. 8933–9, 2004.

[362] M. Vedadi, F. H. Niesen, A. Allali-Hassani, O. Y. Fedorov, P. J. Finerty, G. A. Wasney, R. Yeung, C. Arrowsmith, L. J. Ball, H. Berglund, R. Hui, M. Sundstrom, J. Weigelt, A. M. Edwards, B. D. Marsden, P. Nordlund, M. Sundstrom, J. Weigelt, and A. M. Edwards, "Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 43, pp. 15835–40, 2006.

[363] P. M. Blobel J, Brath U, Diehl C, Ballester L, Sornosa A, Akke M, "Protein loop compaction and the origin of the effect of arginine and glutamic acid mixtures on solubility, stability and transient oligomerization of proteins," *Eur. Biophys. J.*, vol. 40, no. 12, pp. 1327–1338, 2011.

[364] D. J. Leibly, T. N. Nguyen, L. T. Kao, S. N. Hewitt, L. K. Barrett, and W. C. van Voorhis, "Stabilizing Additives Added during Cell Lysis Aid in the Solubilization of Recombinant Proteins," *PLoS One*, vol. 7, no. 12, 2012.

[365] D. Shukla and B. L. Trout, "Understanding the synergistic effect of arginine and glutamic acid mixtures on protein solubility," *J. Phys. Chem. B*, vol. 115, no. 41, pp. 11831–11839, 2011.

[366] A. K. Mohanty, C. R. Simmons, and M. C. Wiener, “Inhibition of tobacco etch virus protease activity by detergents,” *Protein Expr. Purif.*, vol. 27, no. 1, pp. 109–114, 2003.

[367] A. K. Lundb??ck, S. van den Berg, H. Hebert, H. Berglund, and S. Eshaghi, “Exploring the activity of tobacco etch virus protease in detergent solutions,” *Anal. Biochem.*, vol. 382, no. 1, pp. 69–71, 2008.

[368] D. Sachdev and J. M. Chirgwin, “Properties of soluble fusions between mammalian aspartic proteinases and bacterial maltose-binding protein.,” *J. Protein Chem.*, vol. 18, no. 1, pp. 127–136, 1999.

[369] H. K. Lorenzo, D. Farber, V. Germain, O. Acuto, and P. M. Alzari, “The MBP fusion protein restores the activity of the first phosphatase domain of CD45,” *FEBS Lett.*, vol. 411, no. 2–3, pp. 231–235, 1997.

[370] S. E. Swalley, B. M. Baker, L. J. Calder, S. C. Harrison, J. J. Skehel, and D. C. Wiley, “Full-Length Influenza Hemagglutinin HA2 Refolds into the Trimeric Low-pH-Induced Conformation,” *Biochemistry*, vol. 43, no. 19, pp. 5902–5911, 2004.

[371] K. Zanier, Y. Nomin??, S. Charbonnier, C. Ruhlmann, P. Schultz, J. Schweizer, and G. Trav??, “Formation of well-defined soluble aggregates upon fusion to MBP is a generic property of E6 proteins from various human papillomavirus species,” *Protein Expr. Purif.*, vol. 51, no. 1, pp. 59–70, 2007.

[372] Y. Nominé, T. Ristriani, C. Laurent, J. F. Lefèvre, Weiss E, and G. Travé, “A strategy for optimizing the monodispersity of fusion proteins: application to purification of recombinant HPV E6 oncoprotein.,” *Protein Eng.*, vol. 14, no. 4, pp. 297–305, 2001.

[373] A. Haacke, G. Fendrich, P. Ramage, and M. Geiser, “Chaperone over-expression in Escherichia coli: Apparent increased yields of soluble recombinant protein kinases are due mainly to soluble aggregates,” *Protein Expr. Purif.*, vol. 64, no. 2, pp. 185–193, 2009.

[374] Genlantis, “SoluBL21 Competent E. coli White Paper: SoluBL21 Competent Cells for Toxic Clones,” 2016.

[375] S. Hata, F. Kitamura, and H. Sorimachi, “Efficient expression and purification of recombinant human μ-calpain using an Escherichia coli expression system.,” *Genes Cells*, vol. 18, no. 9, pp. 753–63, 2013.

[376] H. W. Kim, S. Y. Lee, H. Park, and S. J. Jeon, “Expression, refolding, and characterization of a small laccase from Thermus thermophilus HJ6,” *Protein Expr. Purif.*, vol. 114, pp. 37–43, 2015.

[377] J. S. Park, W. C. Lee, J. H. Song, S. Il Kim, J. C. Lee, C. Cheong, and H. Y. Kim, “Purification, crystallization and preliminary X-ray crystallographic analysis of diaminopimelate epimerase from Acinetobacter baumannii,” *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, vol. 69, no. 1, pp. 42–44, 2013.

[378] P. M. Sharp and W. H. Li, “The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications,” *Nucleic Acids Res.*, vol. 15, no. 3, pp. 1281–1295, 1987.

[379] M. D. Ermolaeva, “Synonymous codon usage in bacteria,” *Curr. Issues Mol. Biol.*, vol. 3, no. 4, pp. 91–97, 2001.

[380] D. E. Fomenko, S. M. Marino, and V. N. Gladyshev, “Functional diversity of cysteine residues in proteins and unique features of catalytic redox-active cysteines in thiol oxidoreductases.,” *Mol. Cells*, vol. 26, no. 3, pp. 228–35, 2008.

[381] S. M. Marino and V. N. Gladyshev, “Analysis and functional prediction of reactive cysteine residues,” *Journal of Biological Chemistry*, vol. 287, no. 7. pp. 4419–4425, 2012.

[382] X. Zhao, C. Q. Wenzel, and J. S. Lam, “Nonradiolabeling assay for WaaP, an essential sugar kinase involved in biosynthesis of core lipopolysaccharide of Pseudomonas aeruginosa,” *Antimicrob. Agents Chemother.*, vol. 46, no. 6, pp. 2035–2037, 2002.

[383] H. Liu and J. H. Naismith, “An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol.,” *BMC Biotechnol.*, vol. 8, p. 91, 2008.

[384] L. Urzhumtseva, P. V. Afonine, P. D. Adams, and A. Urzhumtsev, “Crystallographic model quality at a glance,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 65, no. 3, pp. 297–300, 2009.

APPENDICES

APPENDIX 1

Minimal Media Stock Solution Preparations for L-Selenomethionine Labelling of BR55

Stock solutions were combined in the following sequence:

1. **M9 salts solution:** 12.8 g $Na_2HPO_4.7H_2O$, 3.0 g KH_2PO_4 (Fisher Scientific, Loughborough, UK) and 1.0 g NH_4Cl (Sigma-Aldrich) in 900 mL of deionised water, prepared in 2L Erlenmeyer flasks prior to autoclaving.
2. **10x Essential Media Stock**: filter sterilised before addition to M9 salts solution and composed of 20% w/v glucose (Sigma-Aldrich), 0.3% w/v magnesium sulphate, 0.01% w/v iron (III) sulphate (Fisher Scientific) and 0.01% w/v thiamine-HCl (Duchefa Biochemie, Haarlem, Netherlands).
3. **20x Glucose-free SelenoMethionine Nutrient Mix** (MD12-502-GF, Molecular Dimensions, Suffolk, UK): prepared in deionised water at 0.022% w/v and filter-sterlised before addition
4. **1x Phosphate Buffered Saline (PBS) solution:** 12.8 g/L $Na_2HPO_4.7H_2O$ and 3.0 g/L KH_2PO_4, pH 7.2; 0.5 g/L NaCl and 1.0 g/L NH_4Cl.
5. **5% v/v starter culture suspension:** Overnight (O/N) starter culture grown in LB broth was washed to remove traces of organic media prior to inoculation into minimal media at an amount equivalent to 5% v/v. Starter culture was pelleted in 50 mL aliquots at 2,800 rpm for 10 minutes, re-suspended gently in 20 mL PBS and pelleted again before discarding the supernatant. This process was repeated twice more before the cells were re-suspended in 5 mL of PBS ready to be added to each litre of minimal essential media.
6. **100x Amino Acid Supplement:** prepared fresh before use and filter sterilised. Consisted of L-lysine, L-phenylalanine and L-threonine at 1% w/v and L-leucine, L-isoleucine and L-valine at 0.5% w/v (all from Duchefa Biochemie).
7. **10x Se-Met suspension:** prepared fresh at 10% w/v before use, in filter sterilised deionised water by mixing L-Se-Met (Molecular Dimensions, Newmarket, UK).

APPENDIX 2

A. Lr0906 protein sequence from *L. reuteri* ATCC 53806 (pig isolate)

```
1       MVKHKQAQET KSKVKLYKTK RGWFSALTRF FGLFSFMSKK EVHPIDLNDL DALKDADTSD
61      NYKKGIATVA TLFGMGAIGA VSPTEVHAAT TTVGYSSQVI GSGSTSVSTS TSQSMASTSG
121     STASQSTSTS SVSQSMQSNS LSVNSVSGST TSTTIFEGTT TSSMVTENSS NSISGSQAIE
181     NSDAASIDHS GSHSNSTTVD SITGSTSLDN EKTSIEESAS TSHLEENGMA TILPATGNND
241     GETSVANRTV TEGNIVQLAE NIEEVSNEEE LKAALRDASI TTIKLKNNIT LNNAITINNG
301     NRNITIIGDG HYINALNSDG GIILNNRGGS AKIDLTIENA TLYNTSKYGF VNMSSNGVDT
361     VTYKDVTAYG GTLVWSKTGA GVKTLNLVGN TTLNSVKSYE VDGQSCGTEA FSHRTPDGDK
421     TTALYVSNAI NIAENANVVL NNSATDIDMW LLTAVPSTSG ISTVTVGNNA SLTMENIGNT
481     EYNIKLDGGR ENHFIVNENA AVKMSAKVDN VRIIPQLENI FTRGNIELAK GSNVHLEVIT
541     GSNFRVAGTV ANRIDFNGTA TLIKQEGASG PTIYDKNSQA NIEFNWLGIF ENTVNFNPGA
601     DVTLIAGKGA SNIGSSSLVT ATVNINDPLR VAFNTDKLET DNGVTYLGNS RINVKVTNAT
661     IKVDDNDFSE LISSNETNIN GSSVGVGSVS LNSNSTSYAN ESESLSASLS FSVATSMAAS
721     QLAAVNNGSQ VIYGGSTSIY NSQSMSSSLS LSTSASLSES LSNSVSMSES MSNSVSISES
781     LSNSVSVSAS ESISASVSAS ESASTSLSAS ESTSASVSAS ESTSASVSAS ESTSTSVSTS
841     ESTSASVSAS ESTSASVSAS ESLSNSVSMS ESLSNSVSMS ESLSNSVSMS ESLSNSVSMS
901     ESLSNSVSMS ESLSNSVSMS ESLSNSVSMS ESLSNSVSMS ESLSNSVSMS ESLSNSVSMS
961     ESLSNSQSMS ISISELISLS SSESASTSLS LSNSTSISIS SSGNASESLS DSLSSISSSS
1021    SLSGSTSAVS DLGSTFSDNS INYSTSSSLS IMNNGNGLKT TNISESGNQT IMNLNETNSV
1081    NSQKPSQQNE KLPQTGNEKH STAKLLSLIS LEAILLLLIS KKKKG*
```

The secretion signal is indicated in orange

The SRR1 and SRR2 domains are indicated in turquoise

The BR55 binding domain is highlighted in green

The SRR2 domain consists of various serine-rich repeat patterns, the most abundant of which (SLSNSVSMSE x 11) is highlighted alternatingly in pink

PCR-amplified region boundaries indicated in **underlined-bold**

B. Lr70902 protein sequence from *L. reuteri* 100-23 (rat isolate)

```
1       MSKYKNEQEN GSKVKLYKTK HGWFSALTRF FKLFSFRSKK EVKPTDFNDL DSLKDKHDST
61      MPDAYKKGAA TMATLLGAGV IGTTSPTEAH AATTIVDQNS HVVGSGSTST STSTSQMTGS
121     TSTSTSTTSG STSTSTVSST SSSTSTTTST TGSTSTSQST GSTSASTSTS ITSGSGSTST
181     SNSNSTSFAS ASTSISNSTL SGLDNNAVAN SVTQNLETTT NTADLTNAAL ASAYQTNLIQ
241     LQNRTANAEN DTVQAEDIQA DATAANASEL KKALQDTSVH TIKLTDNITL TSAIELTNVS
301     RDVTIYGNGK YINATDGNGG IFIHNTKSYT VNLTIEKATL YNQSQYGFVH MNDEGTDNIT
361     YKNITAYGGT LVWSQTHVGT KTLSLEGTVN FYSVPSYTVG GQTYSTDAFK IGTHYPNGEN
421     KDTTPAIYVS NEINIADNAN IALENSATKI DIWMIADIGI HPHTTALTIG NNATLTMENG
481     NNSALNIKLD GDTSNSFTVG EGSTVKLSAK VDNVRILPYE DSNTANVSFA KGSDVTLHAG
541     TGSNLRMGAS ISNQIDFNGK ATFIKDSGAY ANTAYADQTR GNIEFDYYWN DQQKTGSTGV
601     ANFNPGSNVL FQAGPGASNV NTYKLDKNHS YVRTTVNINN PEKVTFDTDI IQNNKGVAYL
661     GNADNGVVVN VNNASVKVNN NNFSDVISSN TTTVTGSTVN VGSVSMASTN GSFTNASTSL
721     SGSVSFSYAT SMATSQLNAV DTNSTSTVVY QGSTVISQST STSVEQSQSV STEESQSAST
781     SESIVQSQSR STSTSESLSG SISESIANSQ STSTSIKESL SNSVSMSESM SGSVSMSESM
841     SNSVSMSESL SGSVSMSESM SNSVLMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
901     SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESM SNSVSMSESL
961     SNSVSMSESM SNSVSMSESL SNSVSMSESL SNSVSMSESM SNSVSMSESL SNSVSMSESL
1021    SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1081    SNSVSMSESL SNSVSMSESL SNSVTMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1141    SNSVSMSKSL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1201    SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1261    SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1321    SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1381    SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1441    SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1501    SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1561    SNSVCMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1621    SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1681    SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1741    SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSVSMSESL
1801    SNSVSMSESL SNSVSMSESL SNSVSMSESL SNSFVKSTST NDATSYSLES FPLSGLISES
1861    TSVSQSFSTS QSLSISESGS ASISISESAS HSTSLSESTS TSLSESSSTS ASLSSSESAS
1921    ISASESSSIS ASTSASISAS LSESVSVLQS ISTSQSLSTS ESDSASISAS AGASISTSLS
1981    TSASVSTSIS GSNSTSDSPV NPQPDNPSTS ISDSESMSIS TSESISVSAS GSTSISDSGS
2041    TSLSVSDSTS MSGSSSNGSG SISGSQSPVT PFPDNSGSTS MSTNGSGNGS ISLSLSITNS
3001    TNTASLENTS ASVTDKHMTV DLTAQNSVEN GNANKPSQQN KRLPQTGNAS QTTAELLGLM
3061    SLEAVLLLLI AKKKKNSEDK
```

The secretion signal is indicated in orange

The SRR1 domain is indicated in turquoise

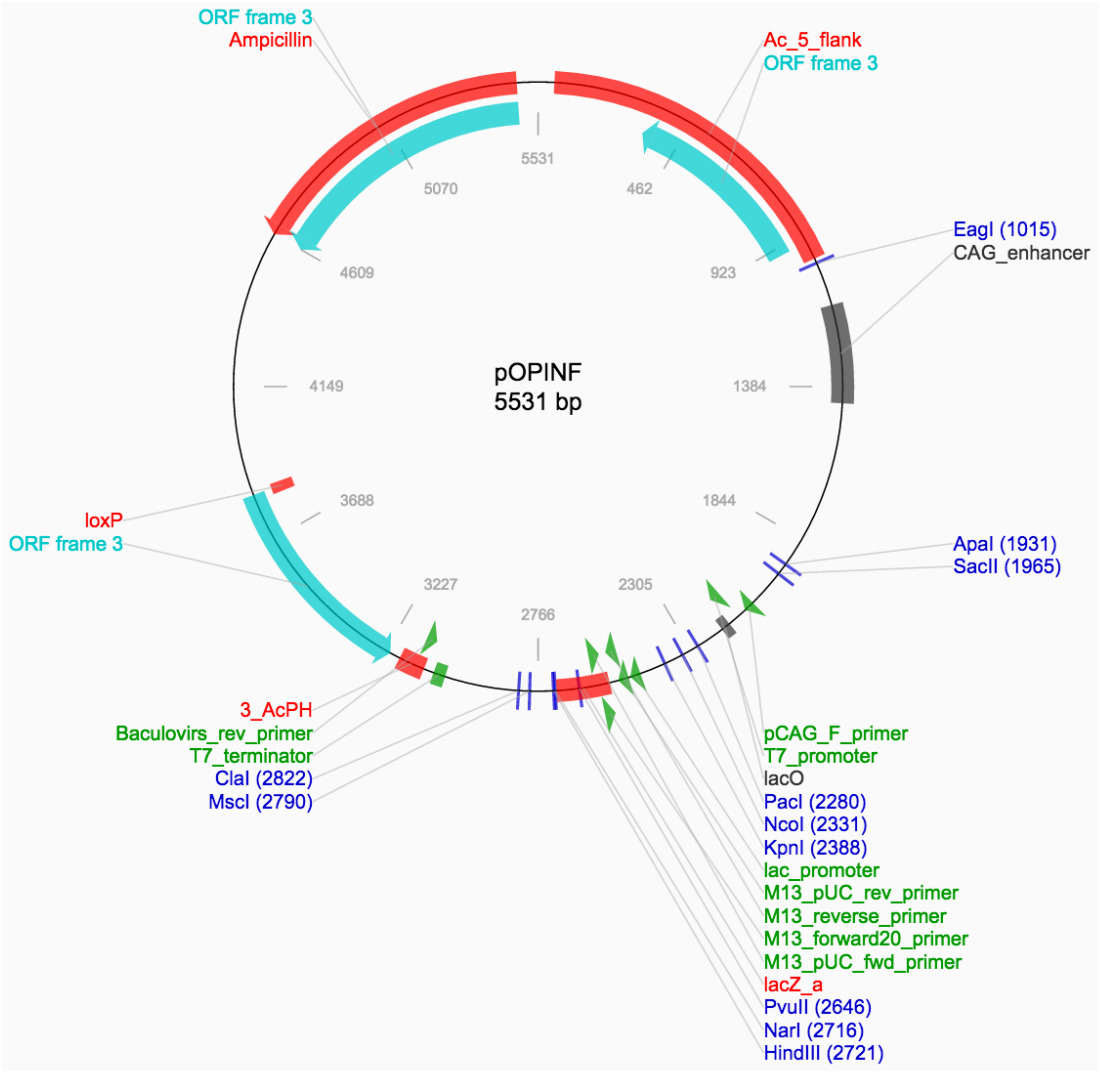
The BR55 binding domain is highlighted in green

The serine rich repeats in SRR2 are indicated in different colours according to sequence type, the most abundant of which (SLSNSVSMSE x 91) is highlighted alternatingly in mint green.

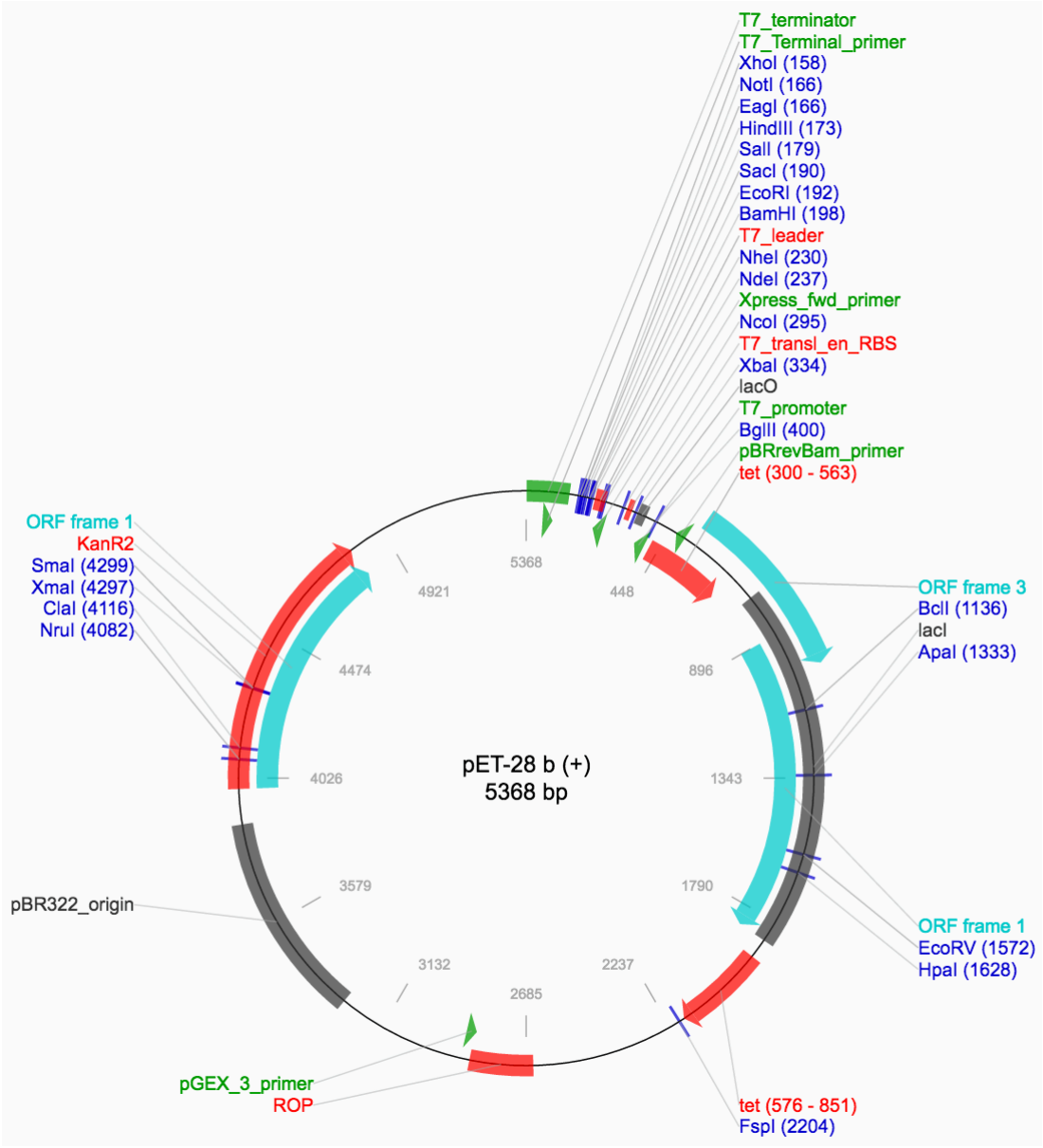
PCR-amplified region boundaries indicated in **underlined-bold**

APPENDIX 3

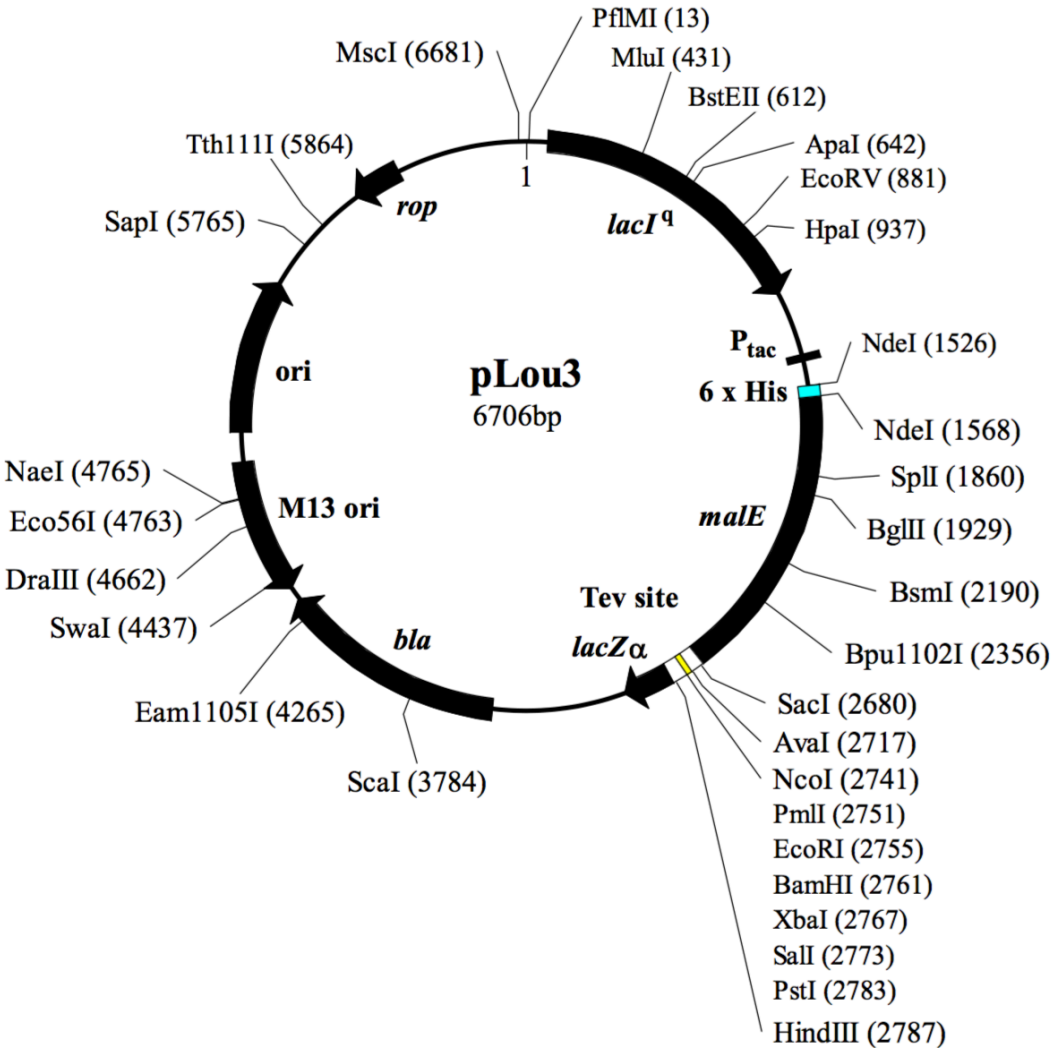
A. pOPINF Plasmid Map



B. pET28 b Plasmid Map



C. pLou3 Plasmid Map



pLOU3 Tev cleavage site and cloning sites:

AvaI		**TEV protease site**							**NcoI**			**PmlI**		**EcoRI**		**BamHI**	
CTC	GGG	GAA	AAC	CTG	TAT	TTT	CAG	GGC	GCC	ATG	GAT	CAC	GTG	GAA	TTC	GGA	TCC
L	G	E	N	L	Y	F	Q	G	A	M	D	H	V	E	F	G	S

XbaI		**SalI**		**PstI**		**HindIII**		
TCT	AGA	GTC	GAC	CTG	CAG	GCA	AGC	TTG	GCA
S	R	V	D	L	Q	A	S	L	A

Sequence for the 6xHis tag:

NdeI		**His-tag**										**NdeI**	
CAT	ATG	AAA	TAT	TAC	CAT	CAC	CAT	CAC	CAT	CAC	GAT	TAC	GAT	CAT	ATG
	M	K	Y	Y	H	H	H	H	H	H	S	Y	S	H	M

APPENDIX 4

Preparation of BR55 mutants for biochemical assays

Mutants of BR55 were prepared for biochemical analysis of the functional properties of BR55, and the identification of key binding residues in the putative binding region of BR55, which were determined as outlined in section 4.4.5. In total, eight mutants were to be produced, six of which were single alanine point mutants: W375A, K377A, Y425A, W450A, Y482A and R512A. One was a double-cysteine mutant, P416C/Q516C (or BR55-CC), to achieve a BR55 phenotype with a closed putative binding loop under oxidising conditions. The final mutant was one in which the lower putative binding loop was removed from residues 411 to 422, and thus this mutant is named ΔF411-T422, or BR55-del.

I. PCR for site-directed Mutagenesis of BR55 expression plasmids

A method by Liu and Naismith [383] was modified for primer design, for the creation of BR55 mutants. Primer pairs are given in and were supplied by Eurogentec (Liège, Belgium). 20 µL PCR reactions consisted of 500 µM of each primer, 200 µM of each dNTP, 2-10 ng DNA template and 0.2 µL Q5 Hot-Start High-Fidelity DNA Polymerase, with its corresponding buffer (NEB, Hitchin, UK). PCR was performed in the T100^{TM} Thermal Cycler (Bio-Rad, Hemel Hempstead, UK) with these steps: initial template denaturation at 98^{o}C for 30 sec; 14 amplification cycles, each with denaturation at 98 ^{o}C for 10 sec, annealing at 'Annealing Temp. 1' (see **Table A.1**) for 20 sec and extension at 72^{o}C for 7 min; the cycles were concluded with an annealing step at 'Annealing Temp. 2' (see) for 15 sec and a final extension at 72^{o}C for 10 min. The PCR products were then treated with 5 units of Fermentas *Dp*nI (Thermo Scientific, Loughborough, UK) at 37^{o}C for an hour, where after they were used for transformation of *E. coli* cells, as detailed in section 1.2.3.

Table A.1: Forward and reverse primer pairs for the creation of BR55 mutants.

The overlapping region (pp) is shown in *italics* and the non-overlapping region (no) has been underlined; the mutation sites are indicated in **bold**. The melting temperatures of each region, with their corresponding annealing temperatures (+3°C of the lowest T_m primer in the primer pair) are also listed and were calculated using the *NEB T_m Calculator* for Q5 High-fidelity DNA Polymerase (http://tmcalculator.neb.com/#!/). Annealing Temp. 1 corresponds to $\mathbf{T_{m\ no}}$ and Annealing Temp. 2 corresponds to $\mathbf{T_{m\ pp}}$.

Mutant		Primer Sequence (5' to 3')	$T_{m\ no}$ (°C)	$T_{m\ pp}$ (°C)	Annealing Temp. 1 (°C)	Annealing Temp. 2 (°C)
W375A	*Fwd*	*CATTAGTG**GCG**AGTAAAAC*TGGTGCTGGTGTTAAAACGTTGAAC	69	56	72	59
	Rev	*GTTTTACT**CGC**CACTAATG*TCCCACCGTAAGCCGTTAC	70	56		
K377A	*Fwd*	*GTGTGGAGT**GCA**ACTGGTG*CTGGTGTTAAAACGTTGAACTTAGTG	70	58	72	61
	Rev	*CAGT**TGC**ACTCCACAC*TAATGTCCCACCGTAAGCC	72	58		
Y425A	*Fwd*	*CACTGCCTTA**GCT**GTTTC*TAATGCCATTAATATTGCAGAAAATGCTAATGTAG	66	58	69	61
	Rev	*GAAAC**AGC**TAAGGCAGTG*GTCTTATCACCATCTGGTGTTCTATGAC	73	58		
W450A	*Fwd*	*CATG**GCG**CTGCTTAC*TGCTGTTCCGTCAACATCAG	70	58	69	61
	Rev	*GTAAGCAG**CGC**CATG*TCAATATCAGTTGCACTGTTATTTAATACTACATTAGC	66	58		
Y482A	*Fwd*	*GTAACACAGAG**GCC**AAC*ATTAAATTAGATGGTGGTAGGGAAAACCAC	67	58	70	61
	Rev	*GTT**GGC**CTCTGTGTTAC*CAATATTTTCCATCGTTAAACTAGCGTTATTAC	69	58		
R512A	*Fwd*	*GATAATGTA**GCA**ATCATACCAC*AACTTGAAAATATCTTTACACGAGGAAATATTG	67	56	70	59
	Rev	*GTGGTATGAT**TGC**TACATTATC*AACCTTAGCTGACATTTTAACAGCC	67	56		
Δ 411-422	*Fwd*	*GAGGCC__GCCTTATATG*TTTCTAATGCCATTAATATTGCAGAAAATGCTAATG	68	57	71	60
	Rev	*CATATAAGGC__GGCCTC*TGTACCACACGACTGTCCATC	73	57		
P416C	*Fwd*	*GTCATAGAACA**TGC**GATGG*TGATAAGACCACTGCCTTATATGTTTCTAATG	68	58	71	61
	Rev	*CCATC**GCA**TGTTCTATGAC*TAAAGGCCTCTGTACCACACG	71	58		
Q516C	*Fwd*	*CATACCA**TGC**CTTGAAAATATC*TTTACACGAGGAAATATTGAGCTTGC	67	57	70	60
	Rev	*GATATTTTCAAG**GCA**TGGTATG*ATTCGTACATTATCAACCTTAGCTGAC	67	57		

II. BR55 mutant expression and purification

Recombinant BR55 mutants were expressed and purified as described for wild type BR55 and BR78 in section 1.3.1 to 1.3.3. All mutants, apart from W375A could be successfully expressed and purified. The purification buffers are provided in **Table 1.4**. Yields of recombinant protein from each mutant's purification is given in **Table A.2** below. SEC chromatogram and SDS-PAGE gel images portraying the results of purification by IMAC and SEC are also provided below. Fractions containing the purest protein, as determined qualitatively by SDS-PAGE, were combined and concentrated to approximately 10 mg/mL for further biophysical or biochemical analyses (which have been carried out in Nathalie Juge's group at the IFR, Norwich, UK). The double-cysteine mutant (BR55-CC) was purified by SEC under reducing and non-reducing conditions to produce, in theory, three types of samples for binding studies:

1) Reducing conditions: reduced monomeric BR55- P416C/Q516C (reBR55-CC)

2a) Non-reducing conditions: monomeric BR55-P416C/Q516C (oxBR55-CC)

2b) Non-reducing conditions: dimeric or trimeric BR55-P416C/Q516C (ox2BR55-CC)

BR55-del was observed with a co-eluting protein in its SEC chromatogram, as seen in **Figure A.7**. Selected SEC fractions were pooled together, concentrated and subjected to a second round of SEC purification to rescue purer BR55-del. The purest BR55-del fractions were collected and concentrated to 10 mg/mL for future use.

Table A.2: Yields of recombinant protein from the purification of BR55 mutants

Mutant	Culture Scale	Yield (mg/L of culture)	Results
W375A	4 L to 12 L	Couldn't be isolated	None
K377A	4 L	4.0 mg/L	**Figure A.1**
Y425A	4 L	0.9 mg/L	**Figure A.2**
W450A	4 L	1.7 mg/L	**Figure A.3**
Y482A	4 L	3.0 mg/L	**Figure A.4**
R512A	4 L	2.9 mg/L	**Figure A.5**
P416C/Q516C (BR55-CC)	2 L	4.2 mg/L (with DTT)	**Figure A.6**
	2 L	1.3 mg/L (w/o DTT, higher MW)	
		1.8 mg/L (w/o DTT, lower MW)	
∆411-422 (BR55-del)	4 L	2.5 mg/L	**Figure A.7**

The numbered points on the SEC chromatograms correspond to the fractions analysed on the SDS-PAGE gels. The shaded areas indicate the fractions that were collected for subsequent biophysical and biochemical analyses.

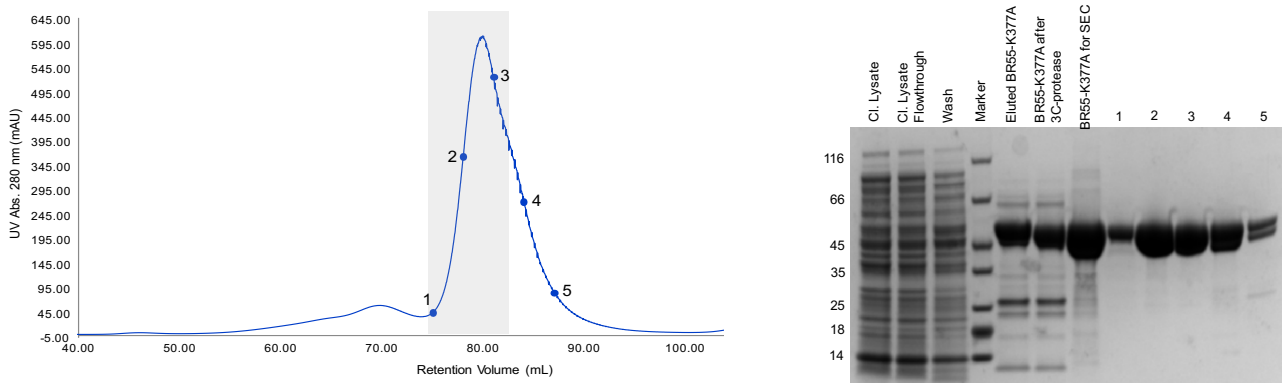


Figure A.1: SEC Chromatogram and SDS-PAGE gel image for BR55-K377A purification

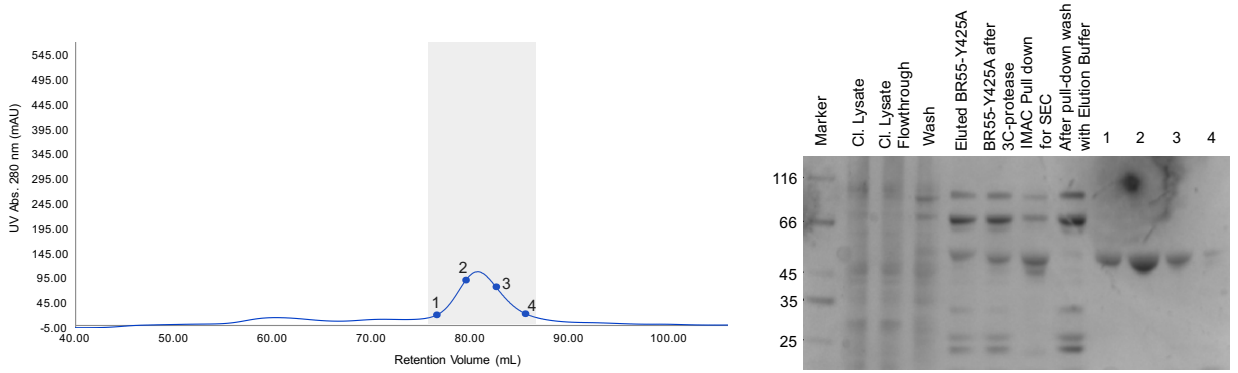


Figure A.2: SEC Chromatogram and SDS-PAGE gel image for BR55-Y425A purification

The numbered points on the SEC chromatograms correspond to the fractions analysed on the SDS-PAGE gels. The shaded areas indicate the fractions that were collected for subsequent biophysical and biochemical analyses.

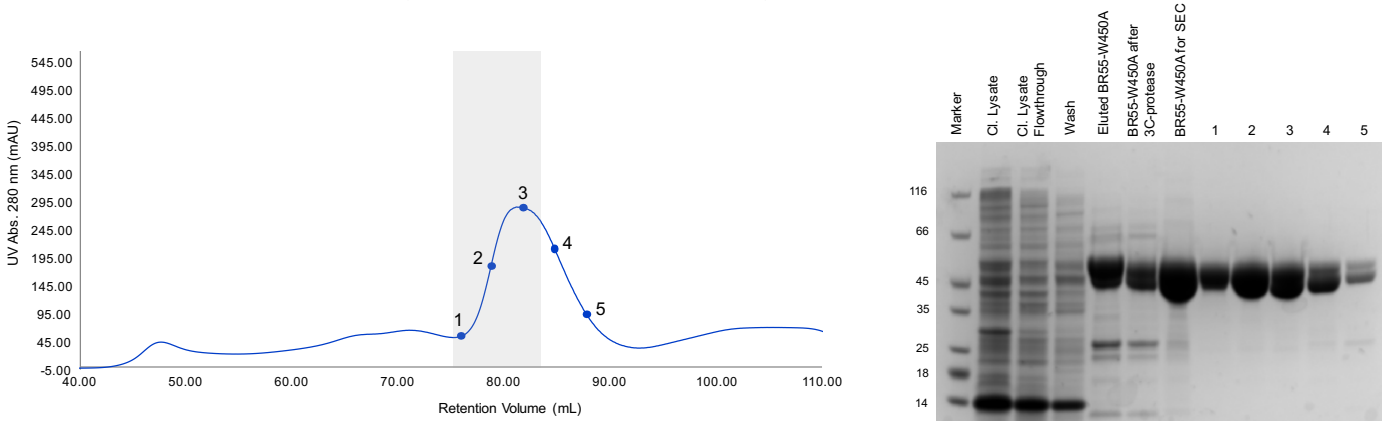


Figure A.3: SEC Chromatogram and SDS-PAGE gel image for BR55-W450A purification

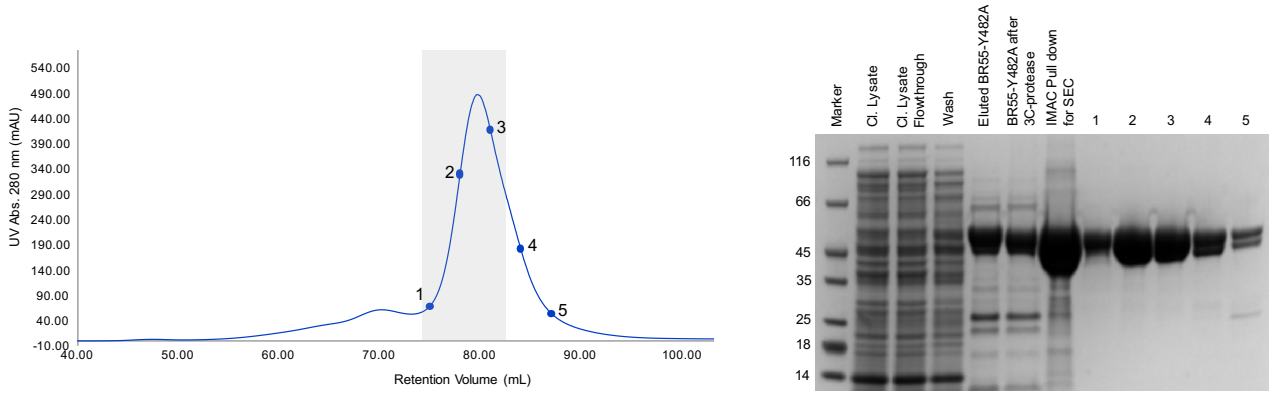
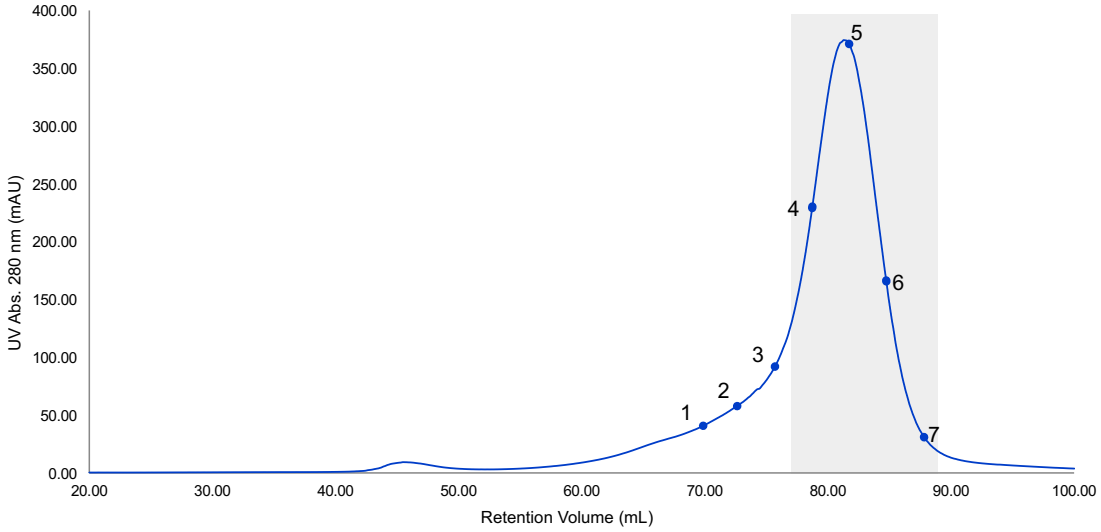


Figure A.4: SEC Chromatogram and SDS-PAGE gel image for BR55-Y482A purification

A.



B.

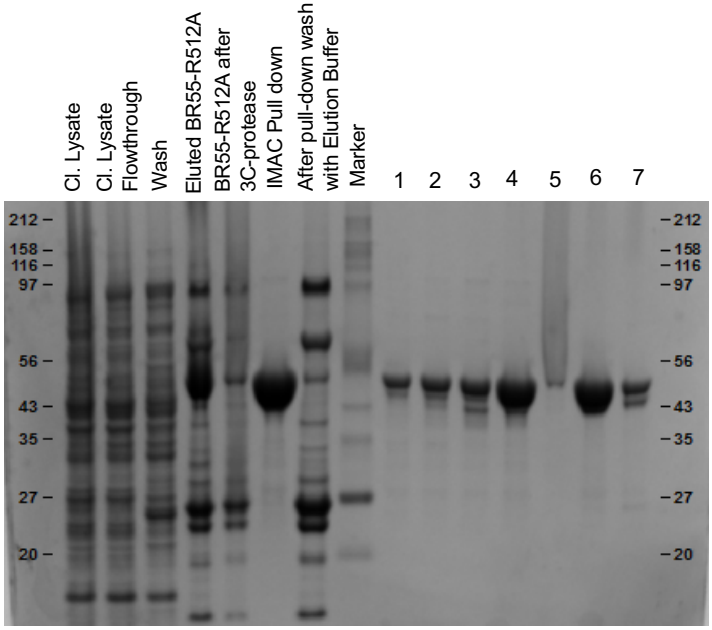
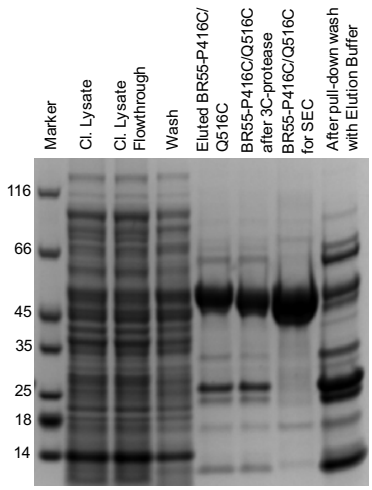
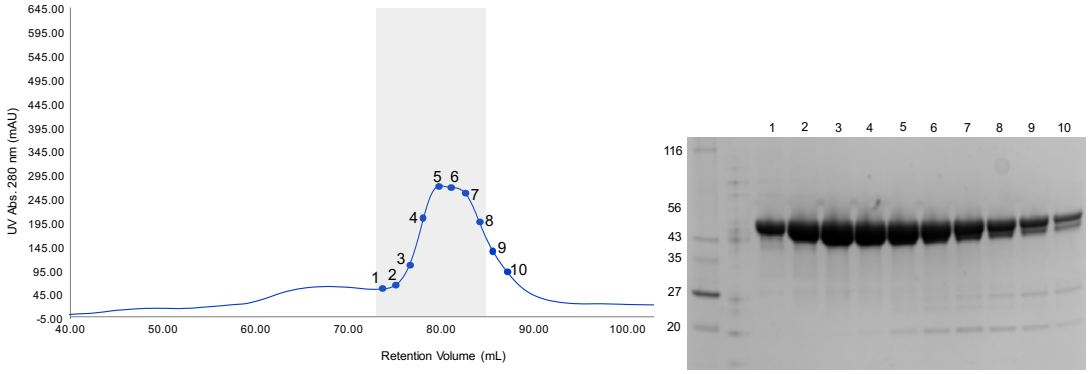


Figure A.5: SEC Chromatogram and SDS-PAGE gel image for BR55-R512A purification

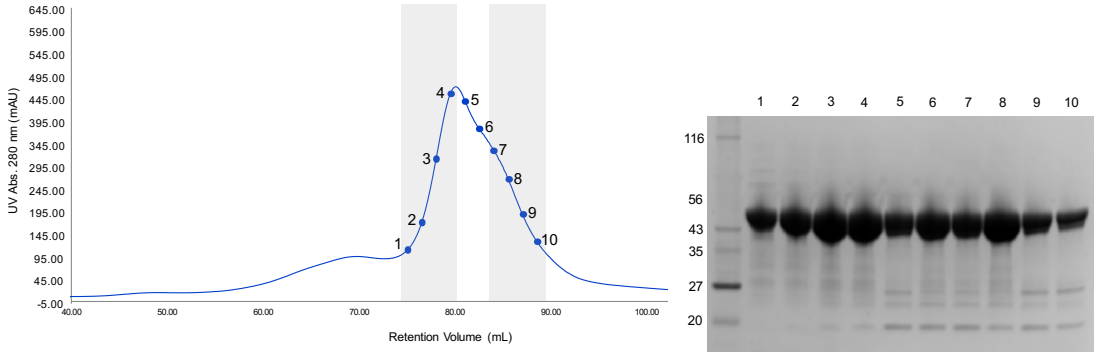
The numbered points on the SEC chromatogram in **A.** correspond to the fractions analysed on the SDS-PAGE gels in **B**. The shaded areas indicate the fractions that were collected for subsequent biophysical and biochemical analyses.



A. SDS-PAGE gel of IMAC Purification

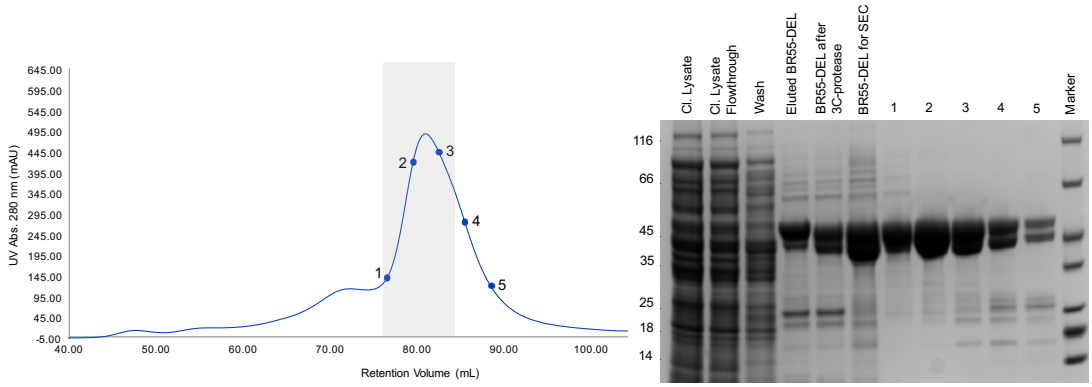


B. SEC Chromatogram and SDS-PAGE gel Results of BR55-P416C/Q516C purification with 2.5 mM DTT in SEC buffer.

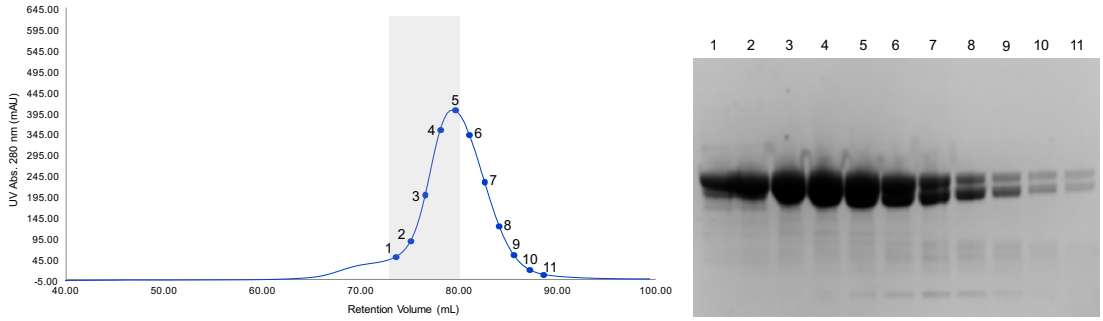


C. SEC Chromatogram and SDS-PAGE gel Results of BR55-P416C/Q516C purification without DTT in SEC buffer.

Figure A.6: SEC Chromatogram and SDS-PAGE gel image for BR55-P416C/Q516C purification (double cysteine mutant).



A. The numbered points on the SEC chromatograms correspond to the fractions analysed on the SDS-PAGE gels. The shaded areas indicate the fractions that were collected for another round of SEC (shown in **B**. below) in an attempt to separate the co-eluting impurity.



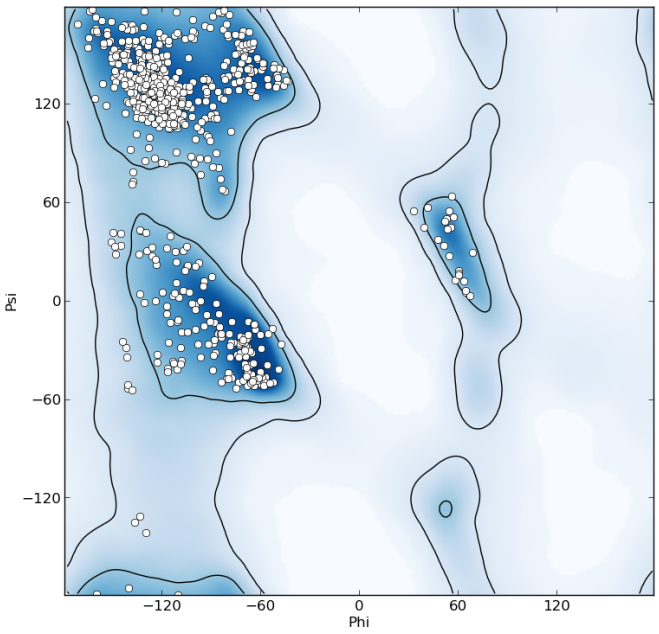
B. The numbered points on the SEC chromatograms correspond to the fractions analysed on the SDS-PAGE gels. The shaded areas indicate the fractions that were collected for subsequent biophysical and biochemical analyses.

Figure A.7: SEC Chromatogram and SDS-PAGE gel image for BR55-Δ411-422 purification (deletion mutant)

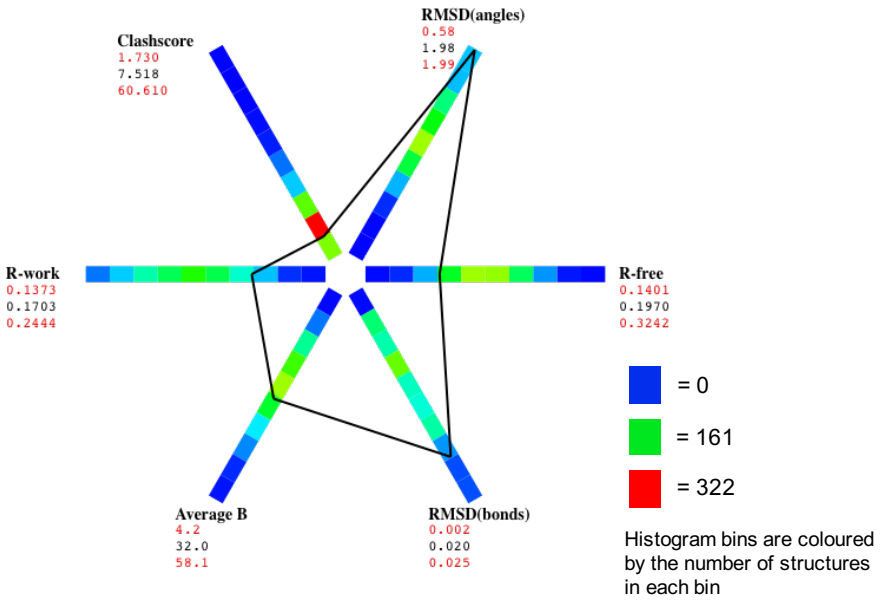
APPENDIX 5

1. BR55 Crystal Structure Validation Statistics

A.



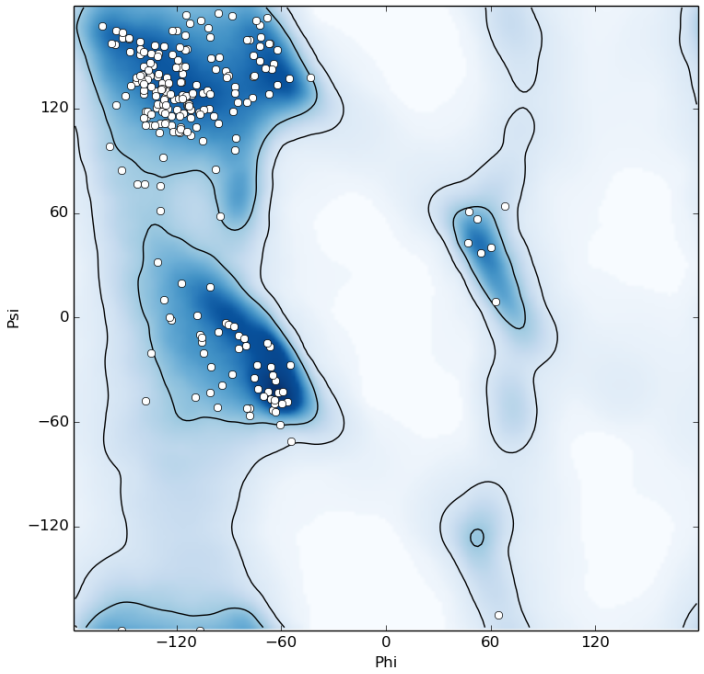
B.



A. Ramachandran plot for all non-Pro/Gly residues in the BR55 structure model. The backbone dihedral angles of all non-Pro/Gly residues are observed in the energetically allowed regions of the plot. **B.** Histograms of the distribution of selected refinement statistics across 828 PDB entries of comparable resolution, computed by POLYGON [384] in the PHENIX suite. Statistics for the BR55 model are in black, and the range of statistical data across 828 entries are shown in red. The connecting polygon show where the BR55 statistical values fall in the distribution.

2. BR78 Crystal Structure Validation Statistics

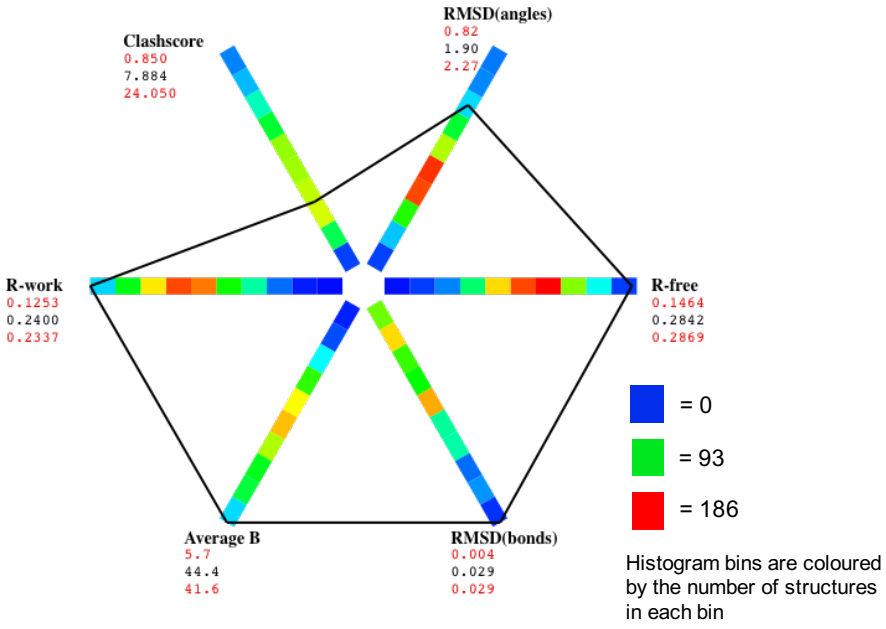
A.



Ramachandran outliers

Residue	Score	Φ	Ψ
V378	0.01	-138.0	-112.9
I411	0.09	-156.5	99.5

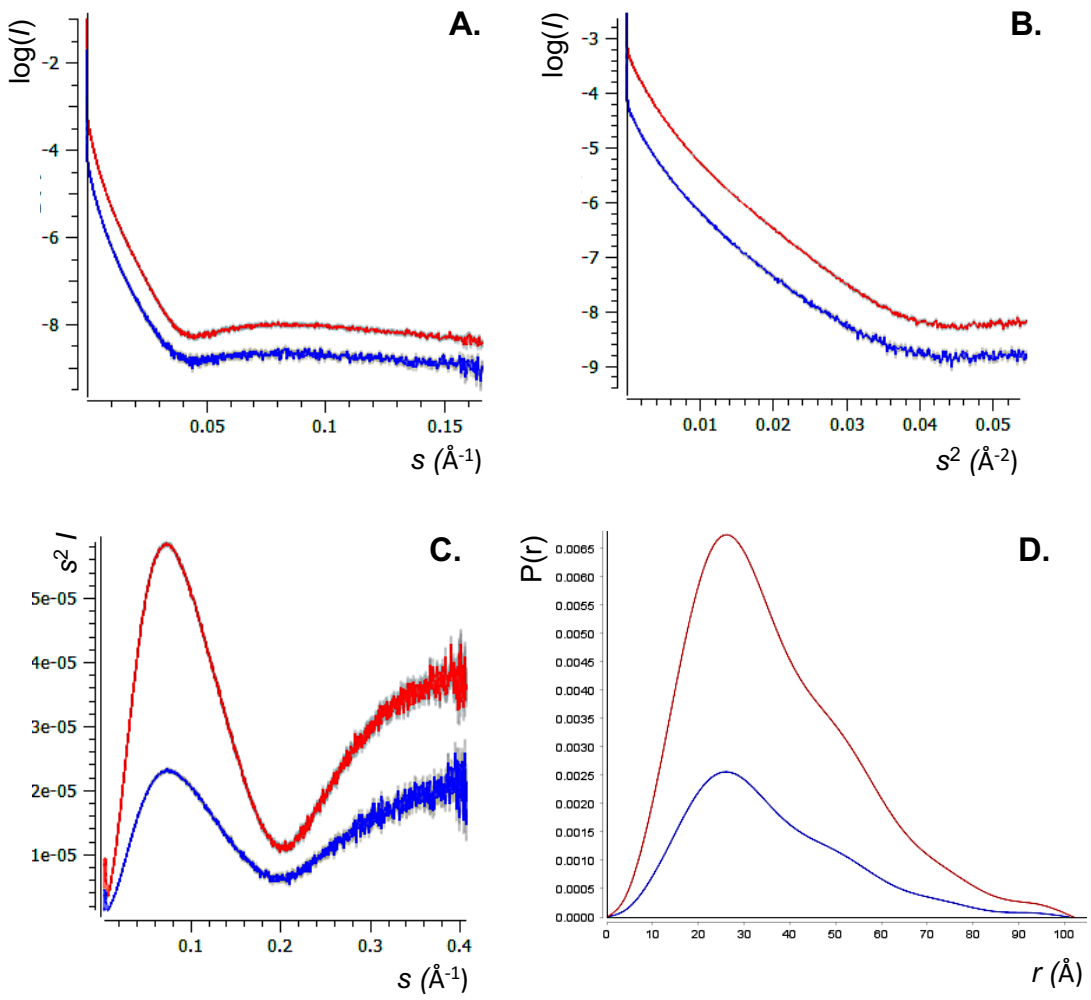
B.



A. Ramachandran plot for all non-Pro/Gly residues in the BR78 model. The backbone dihedral angles of all residues except V378 and I411 are observed in the energetically allowed regions of the plot. The adjacent table lists the dihedral angles of these outlier residues. **B.** Histograms of the distribution of selected refinement statistics across 793 PDB entries of comparable resolution to BR78, computed by POLYGON Statistics for the BR78 model are in black, and the range of statistical data across 793 entries are shown in red. The connecting polygon shows where the BR78 statistical values fall in the distribution.

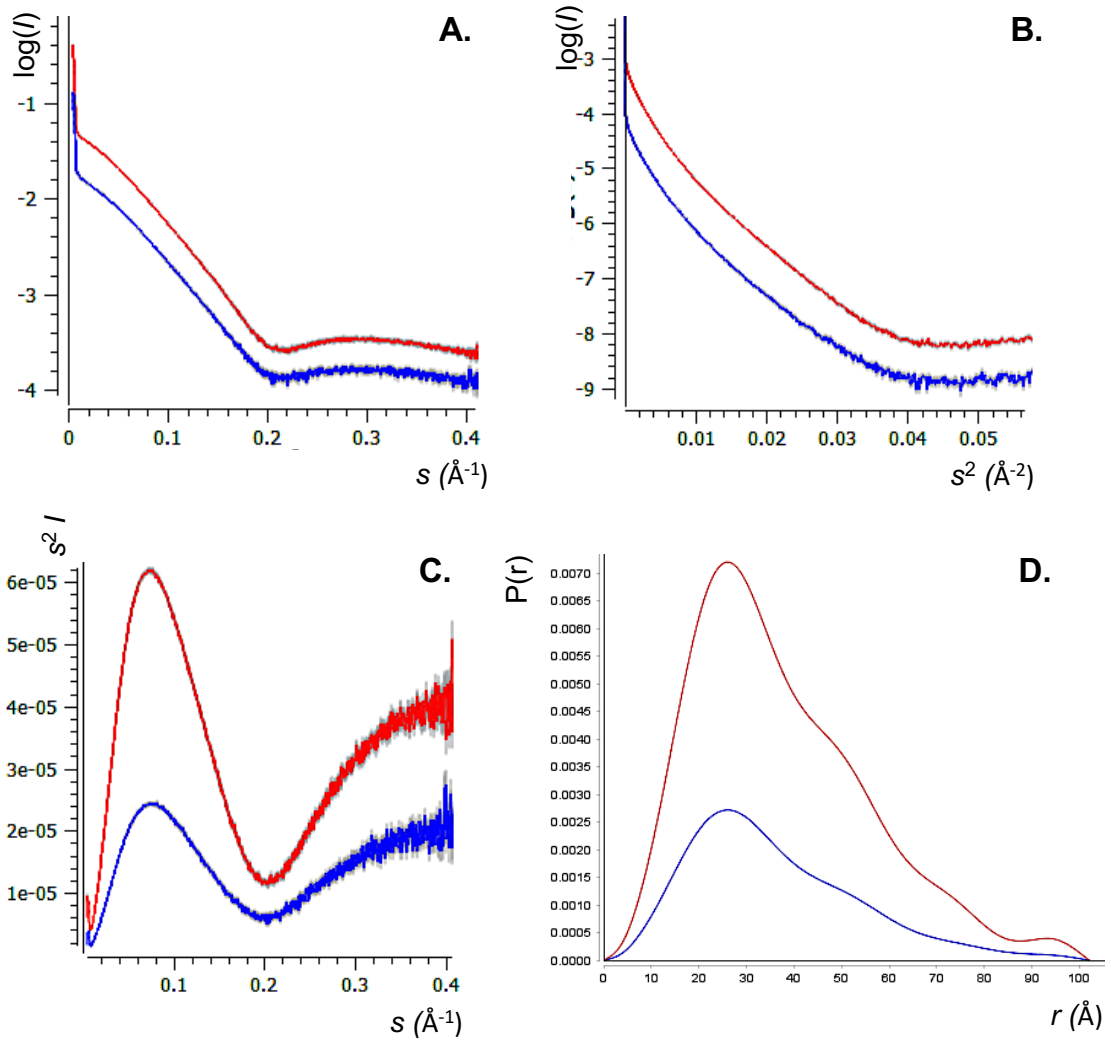
APPENDIX 6

1. SAXS data analysis for BR55 at pH 6.0, at concentrations of 5.0 mg/mL (red) and 2.0 mg/mL (blue).



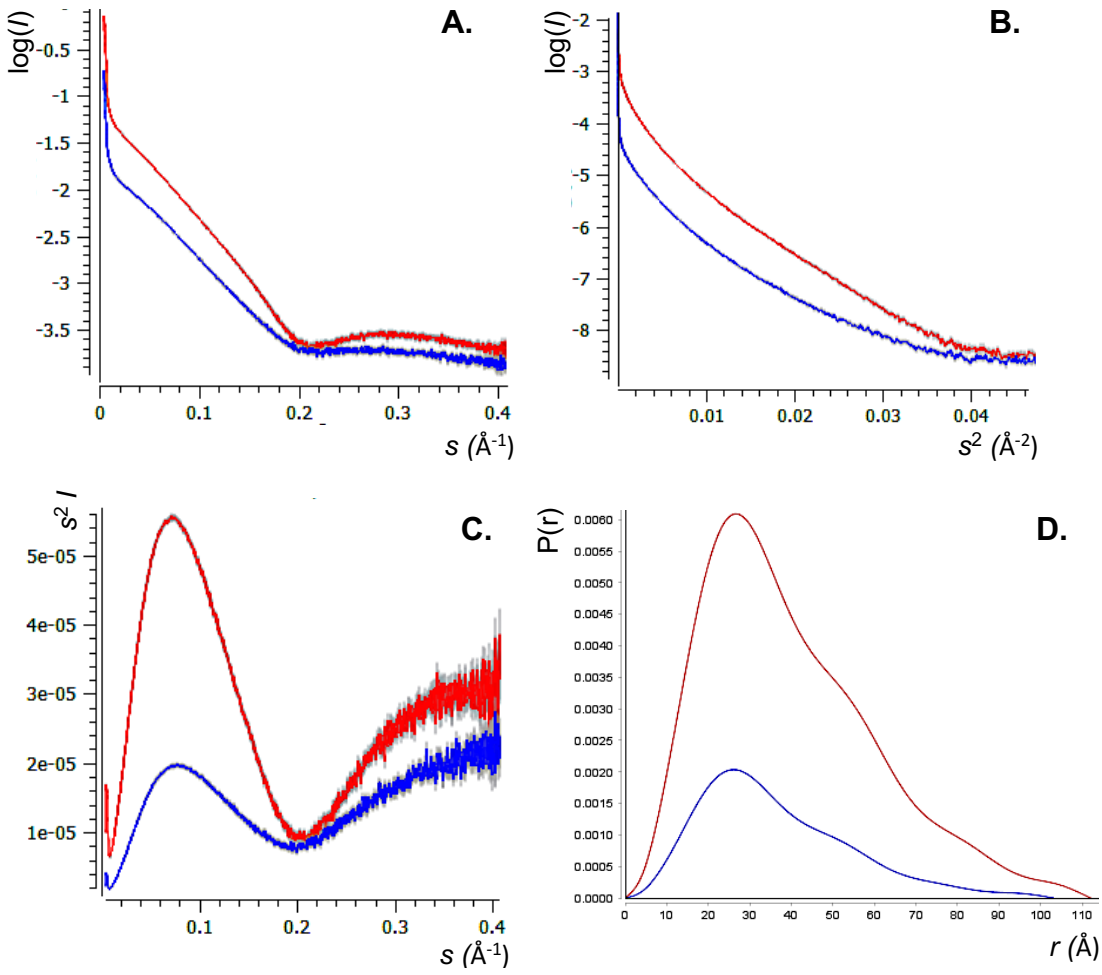
A. Scattering plot of log (*I*) against momentum transfer, *s,* displays inter particle aggregation effects at both concentrations in the low *s* region; approximately 0.05 Å of the initial low *s* region was trimmed for Guinier analysis. **B.** Guinier plot up to *s*=0.2 also reveals aggregation effects at lower *s*. **C.** Kratky plots of data for each BR55 concentration produced curves characteristic of folded, single domain proteins with partial flexibility. **D**. P(r) plot displaying a profile typical for a monomeric, single domain prolate shaped particle with D_{max} at 102 Å estimated for 5.0 mg/mL and 101 Å for 2.0 mg/mL concentrations.

2. SAXS data analysis for BR55 at pH 5.0, at concentrations of 5.0 mg/mL (red) and 2.0 mg/mL (blue).



A. Scattering plot of log (*I*) against momentum transfer, *s,* displays inter particle aggregation effects at both concentrations in the low *s* region; approximately 0.05 Å of the initial low *s* region was trimmed for Guinier analysis. **B.** Guinier plot up to *s*=0.2 also reveals aggregation effects at lower *s*. **C.** Kratky plots of data for each BR55 concentration produced curves characteristic of folded, single domain proteins with partial flexibility. **D**. P(r) plot displaying a profile typical for a monomeric, single domain prolate shaped particle with D_{max} at 102 Å estimated for both concentrations.

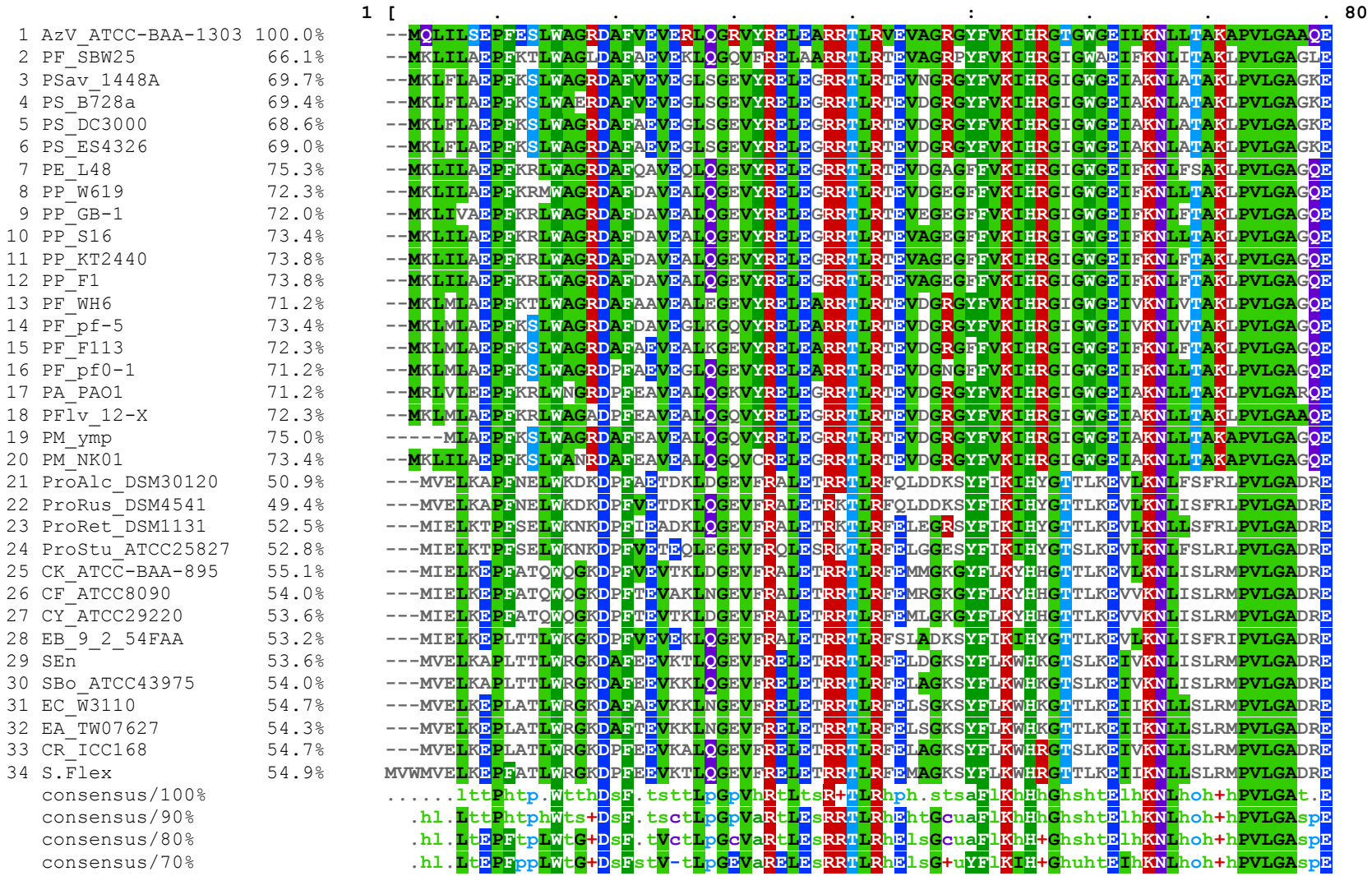
3. SAXS data analysis for BR55 at pH 4.0, at concentrations of 5.0 mg/mL (red) and 2.0 mg/mL (blue).

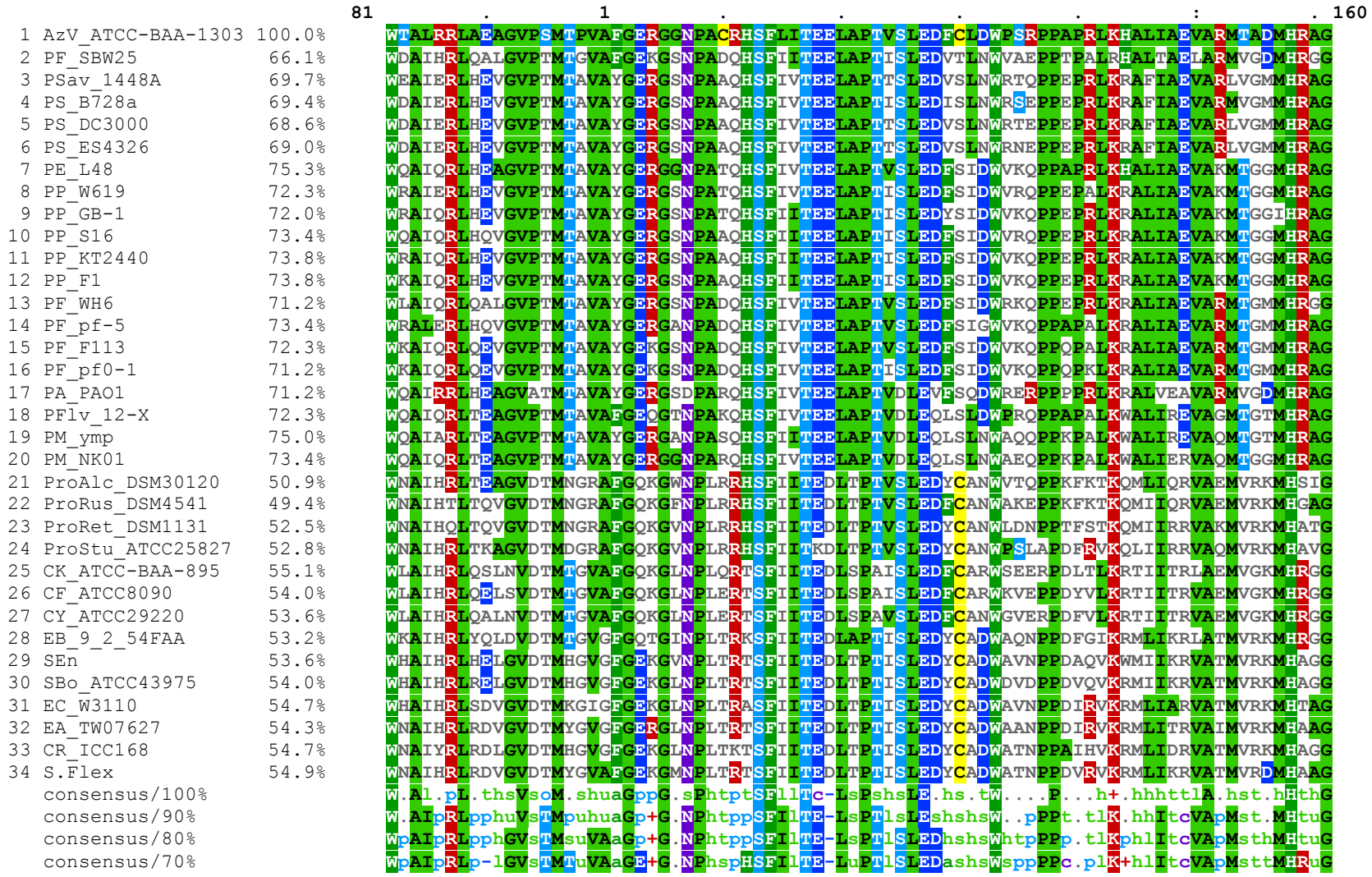


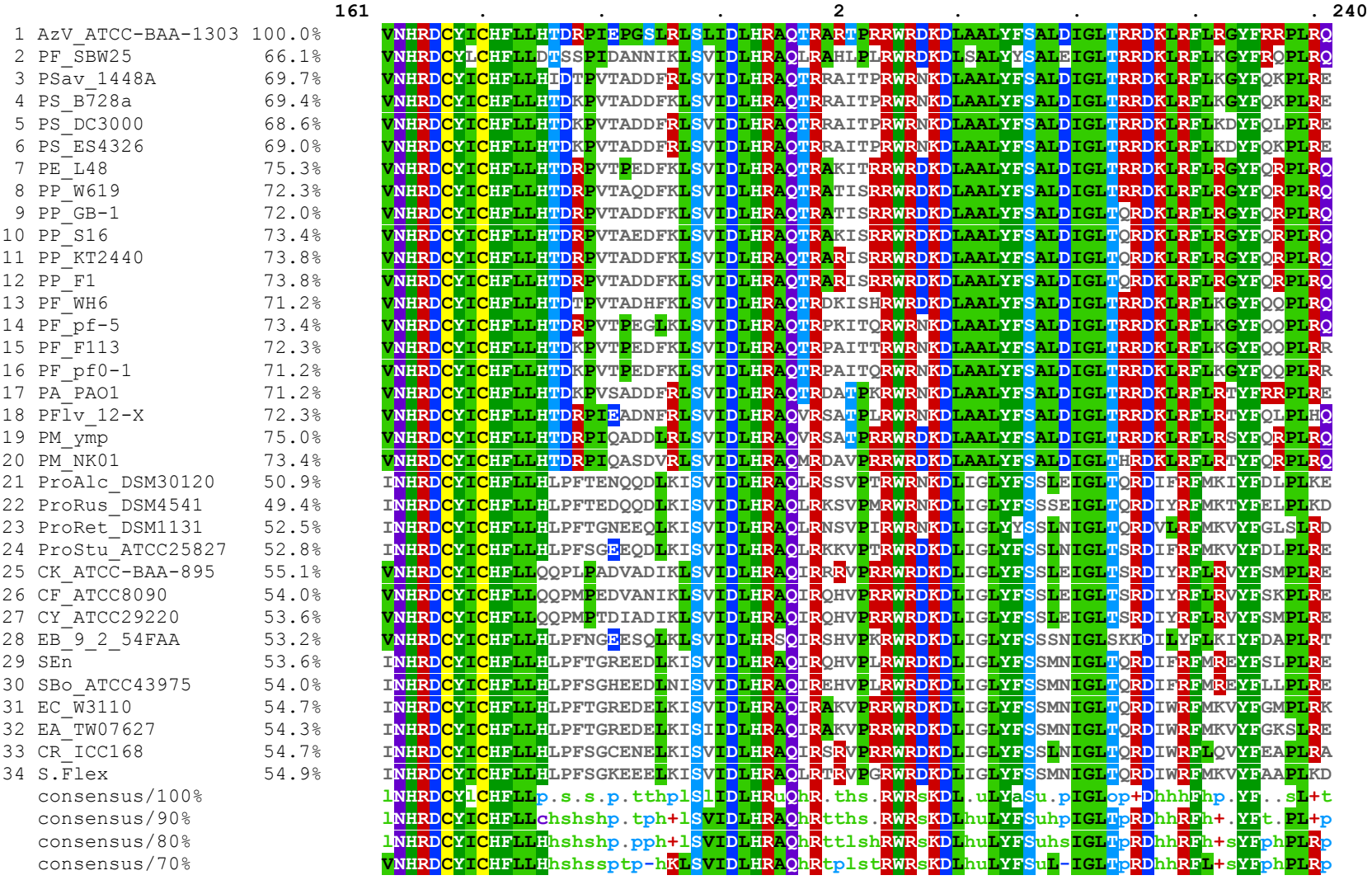
A. Scattering plot of log (*I*) against momentum transfer, *s,* displays inter particle aggregation effects at both concentrations in the low *s* region; approximately 0.05 Å of the initial low *s* region was trimmed for Guinier analysis. **B.** Guinier plot up to *s*=0.2 also reveals aggregation effects at lower *s*. **C.** Kratky plots of data for each BR55 concentration produced curves characteristic of folded, single domain proteins with partial flexibility. **D**. P(r) plot displaying a profile typical for a monomeric, single domain prolate shaped particle with D_{max} at 103 Å estimated for BR55 at 2.0 mg/mL, and 112 Å at 5.0 mg/mL.

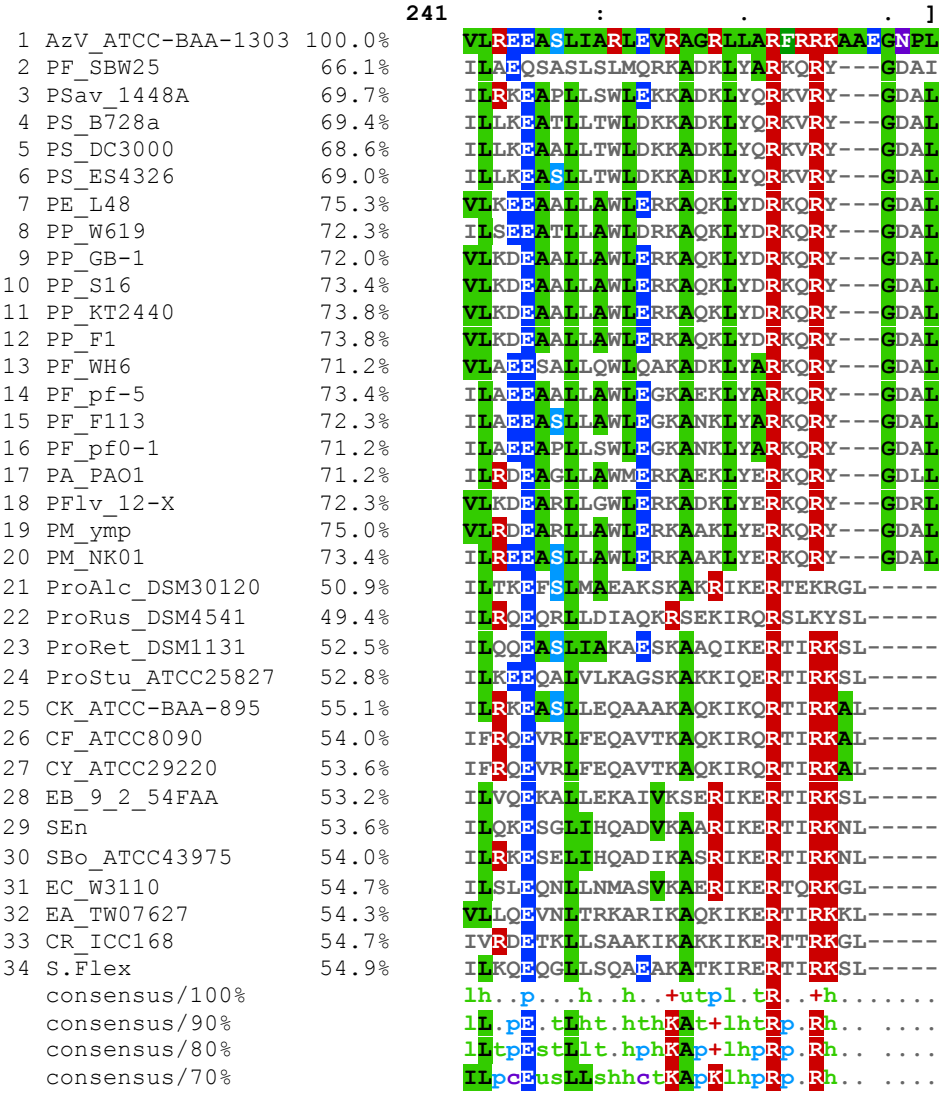
APPENDIX 7

Sequence Alignment for WaaP Proteins across selected Gammaproteobacteria




81 . 1 . . . . : . 160
1 AzV_ATCC-BAA-1303 100.0%
2 PF_SBW25 66.1%
3 PSav_1448A 69.7%
4 PS_B728a 69.4%
5 PS_DC3000 68.6%
6 PS_ES4326 69.0%
7 PE_L48 75.3%
8 PP_W619 72.3%
9 PP_GB-1 72.0%
10 PP_S16 73.4%
11 PP_KT2440 73.8%
12 PP_F1 73.8%
13 PF_WH6 71.2%
14 PF_pf-5 73.4%
15 PF_F113 72.3%
16 PF_pf0-1 71.2%
17 PA_PAO1 71.2%
18 PFlv_12-X 72.3%
19 PM_ymp 75.0%
20 PM_NK01 73.4%
21 ProAlc_DSM30120 50.9%
22 ProRus_DSM4541 49.4%
23 ProRet_DSM1131 52.5%
24 ProStu_ATCC25827 52.8%
25 CK_ATCC-BAA-895 55.1%
26 CF_ATCC8090 54.0%
27 CY_ATCC29220 53.6%
28 EB_9_2_54FAA 53.2%
29 SEn 53.6%
30 SBo_ATCC43975 54.0%
31 EC_W3110 54.7%
32 EA_TW07627 54.3%
33 CR_ICC168 54.7%
34 S.Flex 54.9%
consensus/100%
consensus/90%
consensus/80%
consensus/70%


161 . . . 2 240
1 AzV_ATCC-BAA-1303 100.0%
2 PF_SBW25 66.1%
3 PSav_1448A 69.7%
4 PS_B728a 69.4%
5 PS_DC3000 68.6%
6 PS_ES4326 69.0%
7 PE_L48 75.3%
8 PP_W619 72.3%
9 PP_GB-1 72.0%
10 PP_S16 73.4%
11 PP_KT2440 73.8%
12 PP_F1 73.8%
13 PF_WH6 71.2%
14 PF_pf-5 73.4%
15 PF_F113 72.3%
16 PF_pf0-1 71.2%
17 PA_PAO1 71.2%
18 PFlv_12-X 72.3%
19 PM_ymp 75.0%
20 PM_NK01 73.4%
21 ProAlc_DSM30120 50.9%
22 ProRus_DSM4541 49.4%
23 ProRet_DSM1131 52.5%
24 ProStu_ATCC25827 52.8%
25 CK_ATCC-BAA-895 55.1%
26 CF_ATCC8090 54.0%
27 CY_ATCC29220 53.6%
28 EB_9_2_54FAA 53.2%
29 SEn 53.6%
30 SBo_ATCC43975 54.0%
31 EC_W3110 54.7%
32 EA_TW07627 54.3%
33 CR_ICC168 54.7%
34 S.Flex 54.9%
consensus/100%
consensus/90%
consensus/80%
consensus/70%



Abbreviations of consensus characters

a => alcoholic	(S,T)
l => aliphatic	(I,L,V)
a => aromatic	(F,H,W,Y)
c => charged	(D,E,H,K,R)
h => hydrophobic	(A,C,F,G,H,I,L,M,T,V,W,Y)
- => negative	(D,E)
+ => positive	(H,K,R)
P => polar	(C,D,E,H,K,N,Q,R,S,T)
s => small	(A,C,D,G,N,P,S,T,V)
u => tiny	(A,G,S)
t => turnlike	(A,C,D,E,G,H,K,N,Q,R,S,T)

AzV = *Azotobacter vinelandii;* **PF** = *Pseudomonas fluorescens*
PSav = *Pseudomonas savastanoi;* **PS** = *Pseudomonas syringae*
PE = *Pseudomonas entomophila;* **PP** = *Pseudomonas putida*
PA = *Pseudomonas aeruginosa;* **PFlv** = *Pseudomonas fulva*
PM = *Pseudomonas mendocina;* **ProAlc** = *Providencia alcalifaciens*
ProRus = *Providencia rustigianii;* **ProRet** = *Providencia rettgeri*
ProStu = *Providencia stuartii;* **CK** = *Citrobacter koseri*
CF = *Citrobacter freundii;* **CY** = *Citrobacter youngae*
SEn = *Salmonella enterica;* **SBo** = *Salmonella bongori*
EC = *Escherichia coli;* **EA** = *Escherichia albertii*
CR = *Citrobacter rodentium*; **S.Flex** = *Shigella flexneri*

APPENDIX 8

Percent Identity Matrix (PIM) for selected WaaP proteins

1: AzV_ATCC-BAA1303	**100.00**	66.79	70.52	70.15	69.40	69.78	76.12	73.13	72.76	74.25	74.63	74.63	72.01	74.25	73.13	72.01	72.01
2: PF_SBW25	66.79	**100.00**	70.90	72.01	72.01	71.64	71.27	72.39	71.27	72.39	72.39	72.39	76.49	73.13	74.25	74.25	67.16
3: PSav_1448A	70.52	70.90	**100.00**	94.40	95.52	95.52	80.60	82.09	81.34	81.72	82.46	82.46	82.84	82.84	83.96	84.33	79.10
4: PS_B728a	70.15	72.01	94.40	**100.00**	96.27	96.64	80.97	83.96	81.72	82.09	82.84	82.84	83.21	83.21	84.70	84.70	79.10
5: PS_DC3000	69.40	72.01	95.52	96.27	**100.00**	98.88	80.60	82.46	80.97	81.34	82.09	82.09	83.21	82.84	84.33	83.96	79.48
6: PS_ES4326	69.78	71.64	95.52	96.64	98.88	**100.00**	80.22	82.46	80.60	80.97	81.72	81.72	82.84	82.46	84.70	83.96	79.48
7: PE_L48	76.12	71.27	80.60	80.97	80.60	80.22	**100.00**	91.42	92.16	92.91	92.91	92.91	85.45	86.57	87.31	86.19	79.48
8: PP_W619	73.13	72.39	82.09	83.96	82.46	82.46	91.42	**100.00**	93.66	94.03	94.03	93.66	85.07	85.82	86.57	86.19	79.85
9: PP_GB-1	72.76	71.27	81.34	81.72	80.97	80.60	92.16	93.66	**100.00**	95.90	97.76	97.39	85.07	83.58	85.07	84.33	78.73
10: PP_S16	74.25	72.39	81.72	82.09	81.34	80.97	92.91	94.03	95.90	**100.00**	97.76	97.76	85.82	85.07	85.45	85.45	79.85
11: PP_KT2440	74.63	72.39	82.46	82.84	82.09	81.72	92.91	94.03	97.76	97.76	**100.00**	99.63	86.19	84.70	86.19	85.45	79.85
12: PP_F1	74.63	72.39	82.46	82.84	82.09	81.72	92.91	93.66	97.39	97.76	99.63	**100.00**	86.19	84.33	86.57	85.82	79.85
13: PF_WH6	72.01	76.49	82.84	83.21	83.21	82.84	85.45	85.07	85.07	85.82	86.19	86.19	**100.00**	88.06	88.43	86.94	77.99
14: PF_pf-5	74.25	73.13	82.84	83.21	82.84	82.46	86.57	85.82	83.58	85.07	84.70	84.33	88.06	**100.00**	91.04	89.55	77.24
15: PF_F113	73.13	74.25	83.96	84.70	84.33	84.70	87.31	86.57	85.07	85.45	86.19	86.57	88.43	91.04	**100.00**	96.27	78.73
16: PF_pf0-1	72.01	74.25	84.33	84.70	83.96	83.96	86.19	86.19	84.33	85.45	85.45	85.82	86.94	89.55	96.27	**100.00**	78.36
17: PA_PAO1	72.01	67.16	79.10	79.10	79.48	79.48	79.48	79.85	78.73	79.85	79.85	79.85	77.99	77.24	78.73	78.36	**100.00**
18: PFlv_12-X	73.13	69.78	77.24	76.87	77.61	77.24	81.34	80.22	79.85	80.97	80.60	80.60	79.10	78.73	79.10	78.73	82.09
19: PM_ymp	75.85	70.19	79.25	78.49	78.87	78.87	82.64	82.26	82.26	83.02	83.02	83.02	79.62	81.13	80.00	78.49	82.26
20: PM_NK01	74.25	70.15	78.36	77.99	77.99	78.36	83.21	82.84	80.97	82.09	81.72	81.72	78.73	79.85	80.22	78.36	82.46
21: ProAlc_DSM30120	50.94	51.33	52.09	52.47	52.85	52.47	52.09	50.95	50.95	50.57	51.33	51.33	50.57	51.33	54.37	53.23	51.71
22: ProRus_DSM4541	49.43	51.33	51.71	52.85	52.09	51.71	50.19	49.81	49.05	50.19	50.19	50.19	49.81	51.33	51.71	52.47	50.95
23: ProRet_DSM1131	52.45	52.09	51.71	52.09	51.71	51.71	51.33	50.57	51.33	52.47	51.33	51.33	49.81	52.47	53.23	52.85	51.71
24: ProStu_ATCC25827	52.83	51.71	52.09	53.23	52.47	51.71	53.61	50.95	51.71	52.09	52.09	52.09	52.47	51.33	52.47	51.71	51.33
25: CK_ATCC-BAA895	55.09	57.41	55.89	56.65	55.51	55.89	54.75	55.51	54.37	55.51	55.89	55.89	55.51	52.85	55.13	55.51	54.37
26: CF_ATCC8090	53.96	57.41	54.75	55.51	54.37	54.75	54.37	55.13	53.99	54.75	55.13	55.13	55.51	52.85	54.37	55.13	53.61
27: CY_ATCC29220	53.58	57.03	54.37	54.75	54.37	54.37	54.37	54.37	53.23	54.37	54.37	54.37	55.89	52.85	54.37	54.37	53.61
28: EB_9_2_54FAA	53.21	54.75	50.95	51.71	50.95	50.57	51.71	51.71	52.09	52.85	52.47	52.85	52.85	51.71	51.71	52.85	51.33
29: SEn	53.58	53.99	52.47	53.99	53.61	53.23	53.99	53.99	53.61	53.61	53.61	53.61	53.23	51.71	52.85	53.23	52.85
30: SBo_ATCC43975	53.96	54.37	52.85	53.23	53.23	52.85	52.85	53.23	53.23	53.61	53.61	53.61	52.85	50.95	52.47	52.85	52.47
31: EC_W3110	54.72	54.37	53.61	54.37	53.99	53.99	54.75	55.51	55.89	56.27	55.89	55.89	55.13	53.99	54.75	55.13	52.85
32: EA_TW07627	54.34	52.47	53.61	54.75	53.61	53.99	54.75	54.37	55.89	56.27	55.89	55.89	54.37	52.47	52.47	52.85	52.09
33: CR_ICC168	54.72	55.13	53.61	53.23	52.85	52.47	54.37	55.13	55.51	56.27	56.27	56.27	55.13	53.23	53.99	55.51	54.75
34: S.Flex	55.26	55.30	54.92	54.55	54.17	53.79	55.30	54.92	56.06	57.20	56.82	56.82	54.55	54.17	55.30	57.20	55.30

1: AzV_ATCC-BAA1303	73.13	75.85	74.25	50.94	49.43	52.45	52.83	55.09	53.96	53.58	53.21	53.58	53.96	54.72	54.34	54.72	55.26	
2: PF_SBW25	69.78	70.19	70.15	51.33	51.33	52.09	51.71	57.41	57.41	57.03	54.75	53.99	54.37	54.37	52.47	55.13	55.30	
3: PSav_1448A	77.24	79.25	78.36	52.09	51.71	51.71	52.09	55.89	54.75	54.37	50.95	52.47	52.85	53.61	53.61	53.61	54.92	
4: PS_B728a	76.87	78.49	77.99	52.47	52.85	52.09	53.23	56.65	55.51	54.75	51.71	53.99	53.23	54.37	54.75	53.23	54.55	
5: PS_DC3000	77.61	78.87	77.99	52.85	52.09	51.71	52.47	55.51	54.37	54.37	50.95	53.61	53.23	53.99	53.61	52.85	54.17	
6: PS_ES4326	77.24	78.87	78.36	52.47	51.71	51.71	51.71	55.89	54.75	54.37	50.57	53.23	52.85	53.99	53.99	52.47	53.79	
7: PE_L48	81.34	82.64	83.21	52.09	50.19	51.33	53.61	54.75	54.37	54.37	51.71	53.99	52.85	54.75	54.75	54.37	55.30	
8: PP_W619	80.22	82.26	82.84	50.95	49.81	50.57	50.95	55.51	55.13	54.37	51.71	53.99	53.23	55.51	54.37	55.13	54.92	
9: PP_GB-1	79.85	82.26	80.97	50.95	49.05	51.33	51.71	54.37	53.99	53.23	52.09	53.61	53.23	55.89	55.89	55.51	56.06	
10: PP_S16	80.97	83.02	82.09	50.57	50.19	52.47	52.09	55.51	54.75	54.37	52.85	53.61	53.61	56.27	56.27	56.27	57.20	
11: PP_KT2440	80.60	83.02	81.72	51.33	50.19	51.33	52.09	55.89	55.13	54.37	52.47	53.61	53.61	55.89	55.89	56.27	56.82	
12: PP_F1	80.60	83.02	81.72	51.33	50.19	51.33	52.09	55.89	55.13	54.37	52.85	53.61	53.61	55.89	55.89	56.27	56.82	
13: PF_WH6	79.10	79.62	78.73	50.57	49.81	49.81	52.47	55.51	55.51	55.89	52.85	53.23	52.85	55.13	54.37	55.13	54.55	
14: PF_pf-5	78.73	81.13	79.85	51.33	51.33	52.47	51.33	52.85	52.85	52.85	51.71	51.71	50.95	53.99	52.47	53.23	54.17	
15: PF_F113	79.10	80.00	80.22	54.37	51.71	53.23	52.47	55.13	54.37	54.37	51.71	52.85	52.47	54.75	52.47	53.99	55.30	
16: PF_pf0-1	82.09	82.26	82.46	51.71	50.95	51.71	51.33	54.37	53.61	53.61	51.33	52.85	52.47	52.85	52.09	54.75	55.30	
17: PA_PAO1	78.73	78.49	78.36	53.23	52.47	52.85	51.71	55.51	55.13	54.37	52.85	53.23	52.85	55.13	52.85	55.51	57.20	
18: PFlv_12-X	**100.00**	89.43	85.45	52.09	50.57	52.09	52.09	52.09	52.85	52.09	50.95	52.47	51.71	51.71	51.33	54.75	54.55	
19: PM_ymp	89.43	**100.00**	92.08	52.87	51.34	52.87	52.49	54.02	54.02	54.41	52.11	54.41	53.26	54.02	54.02	56.32	55.17	
20: PM_NK01	85.45	92.08	**100.00**	53.23	51.71	53.61	53.23	54.75	53.61	53.99	51.33	53.99	52.85	53.23	52.47	55.13	54.55	
21: ProAlc_DSM30120	52.09	52.87	53.23	**100.00**	84.15	79.62	77.74	66.04	64.91	65.66	69.81	71.32	70.57	70.57	68.68	69.43	70.19	
22: ProRus_DSM4541	50.57	51.34	51.71	84.15	**100.00**	78.49	74.72	66.04	66.42	67.17	69.43	67.92	67.17	68.30	66.79	66.04	70.94	
23: ProRet_DSM1131	52.09	52.87	53.61	79.62	78.49	**100.00**	80.38	65.28	64.53	65.28	71.32	72.45	70.57	71.32	72.45	69.43	73.58	
24: ProStu_ATCC25827	52.09	52.49	53.23	77.74	74.72	80.38	**100.00**	66.79	65.66	67.17	69.81	72.08	71.70	70.57	70.94	70.19	73.21	
25: CK_ATCC-BAA895	52.09	54.02	54.75	66.04	66.04	65.28	66.79	**100.00**	89.43	90.57	70.19	70.19	70.94	69.06	69.43	70.19	70.19	
26: CF_ATCC8090	52.85	54.02	53.61	64.91	66.42	64.53	65.66	89.43	**100.00**	94.72	69.06	71.32	71.70	69.06	71.32	70.19	70.19	
27: CY_ATCC29220	52.09	54.41	53.99	65.66	67.17	65.28	67.17	90.57	94.72	**100.00**	68.68	70.57	70.94	68.30	69.81	69.43	69.43	
28: EnB_9_2_54FAA	50.95	52.11	51.33	69.81	69.43	71.32	69.81	70.19	69.06	68.68	**100.00**	73.96	73.58	73.58	72.08	73.21	72.83	
29: SEn	52.47	54.41	53.99	71.32	67.92	72.45	72.08	70.19	71.32	70.57	73.96	**100.00**	93.58	83.02	81.51	81.51	81.51	
30: SBo_ATCC43975	51.71	53.26	52.85	70.57	67.17	70.57	71.70	70.94	71.70	70.94	73.58	93.58	**100.00**	81.89	81.51	83.02	81.51	
31: EC_W3110	51.71	54.02	53.23	70.57	68.30	71.32	70.57	69.06	69.06	68.30	73.58	83.02	81.89	**100.00**	89.43	83.02	83.77	
32: EA_TW07627	51.33	54.02	52.47	68.68	66.79	72.45	70.94	69.43	71.32	69.81	72.08	81.51	81.51	89.43	**100.00**	83.02	84.53	
33: CR_ICC168	54.75	56.32	55.13	69.43	66.04	69.43	70.19	70.19	70.19	69.43	73.21	81.51	83.02	83.02	83.02	**100.00**	84.91	
34: S.Flex_2a_301	54.55	55.17	54.55	70.19	70.94	73.58	73.21	70.19	70.19	69.43	72.83	81.51	81.51	83.77	84.53	84.91	**100.00**	

AzV = *Azotobacter vinelandii;* **PF** = *Pseudomonas fluorescens;* **PSav** = *Pseudomonas savastanoi;* **PS** = *Pseudomonas syringae;* **PE** = *Pseudomonas entomophila;* **PP** = *Pseudomonas putida;* **PA** = *Pseudomonas aeruginosa;* **PFlv** = *Pseudomonas fulva;* **PM** = *Pseudomonas mendocina;* **ProAlc** = *Providencia alcalifaciens;* **ProRus** = *Providencia rustigianii;* **ProRet** = *Providencia rettgeri;* **ProStu** = *Providencia stuartii;* **CK** = *Citrobacter koseri;* **CF** = *Citrobacter freundii;* **CY** = *Citrobacter youngae;* **EnB** = *Enterobacteriaceae bacterium;* **SEn** = *Salmonella enterica;* **SBo** = *Salmonella bongori;* **EC** = *Escherichia coli;* **EA** = *Escherichia albertii;* **CR** = *Citrobacter rodentium*; **S.Flex** = *Shigella flexneri*