Structural and biophysical characterisation of the repetitive regions from biofilm-mediating cell-wall anchored proteins

Lotte van Beek, MSc

Doctor of Philosophy

University of York

Biology

May 2019

Abstract

Staphylococci and streptococci form microbial accumulations, defined as biofilms, on in-dwelling medical devices or damaged tissue. Biofilms are implicated in serious infections, such as infective endocarditis, with a mortality rate of 30%. Cell wall-anchored (CWA) proteins, containing a repetitive B region that putatively projects a functional A region from the bacterial surface, mediate biofilms, independent of other mechanisms. This highlights the need to better understand the mechanism of biofilm formation/accumulation by CWA proteins.

Staphylococcus aureus surface protein C (SasC) mediates biofilm accumulation, but the structure and function of the B region, containing domains of unknown function (DUF1542), remains undetermined. SGO0707 from *Streptococcus gordonii* mediates surface adhesion putatively via its A domains; the function of the B region remains unknown. Here, the B regions of SasC and SGO0707 are biophysically characterised and their ability to form elongated stalks is assessed.

Work presented in this thesis enables the redefinition of the domain boundaries for the repeats comprising the B region of SasC and renaming these as DUF1542 rigid extracellular surface structural (DRESS) domains. Tandem DRESS domains have tightly connected inter-domain interfaces that are essential for tandem domain stability and which mediate long-range stability. Importantly, DRESS domains form an extended, rigid rod and have remarkable mechanical stability, compared to other helical proteins.

The B region of SGO0707 comprises SGO0707 high identity repeat tandem (SHIRT) domains with an extended tandem domain architecture, determined prior to this work. Here, SHIRT domains are shown to lack flexible loops and notably, the linker between SHIRT domains has limited flexibility, implying that tandem SHIRT domains have an extended, rod-like conformation in solution.

Importantly, both domain architectures form a rigid rod, suggesting their ability to project the functional A region from the cell surface. Thus, staphylococci and streptococci have both evolved structurally distinct stalks-like CWA proteins to mediate biofilms.

Samenvatting

De Gram-positieve bacteriën staphylococci en streptococci kunnen een opeenhoping van micro-organismen vormen op niet-lichaamseigen materialen, zoals implantaten. Deze opeenhopingen heten biofilms en veroorzaken ernstige infecties, zoals een infectie op de hartkleppen, wat in 30% van de gevallen leidt tot overlijden. Eiwitten op het celoppervlak bevatten een repetitieve middenregio die het functionele deel wegprojecteert van het celoppervlak. Zo kunnen ze biofilms veroorzaken en vergroten. Dit geeft de noodzaak aan om het mechanisme waarop deze biofilm-vormende eiwitten werken, te onderzoeken.

SasC is een biofilm-vormend eiwit van *Staphylococcus aureus*, een bacterie die wordt aangetroffen bij een op drie mensen. Het heeft een repeterend deel met domeinen (DUF1542) waarvan de structuur en functie onbekend zijn. SGO0707, een eiwit van *Streptococcus gordonii*, vormt biofilms met het functionele deel, maar de functie van repetitieve deel (SHIRT domeinen) is ook onbekend. Dit proefschrift onderzoekt of de repetitieve regio's van beide eiwitten een staafachtige architectuur hebben, die het functionele deel van deze eiwitten wegprojecteert van het celoppervlak.

Dit werk bepaalt het correcte begin en einde van de repeterende eenheid in SasC en ontdekt hun structuur, wat het mogelijk maakt om DUF1542 te hernoemen naar DRESS. Dubbele DRESS domeinen hebben een stevig, gesloten oppervlak dat essentieel is voor hun stabiliteit. Ondanks hun karakteristiek zwakke secundaire structuur, hebben ze een uitgestrekte, rigide architectuur en zijn ze mechanisch verbazingwekkend sterk, in verhouding tot vergelijkbare eiwitten.

In de al beschreven structuur van dubbele SHIRT domeinen ontbreekt een oppervlak, wat de hypothese van een uitgestrekte architectuur in twijfel trekt. Dit werk toont aan dat SHIRT domeinen ondanks dit kleine oppervlak toch een gelimiteerde flexibiliteit hebben, wat suggereert dat ze ook uitgestrekt en rigide kunnen zijn in een repetitieve regio.

Beide repetitieve regio's met een verschillende secundaire structuur kunnen een rigide, staafachtige architectuur aannemen, die kan wijzen op hun functie om het functionele deel van deze eiwitten weg te projecteren van het celoppervlak. Extracellulaire eiwitten van staphylococci en streptococci zijn dus geëvolueerd met verschillende repetitieve structuren om biofilms te kunnen vormen.

Table of Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

Abbreviation	Meaning
(^{1}H, ^{15}N)-HSQC	(^{1}H, ^{15}N)-heteronuclear single quantum coherence
1D	One-dimensional
^{1}H-NMR	Proton NMR
αHb	α subunit of haemoglobin
A488	Alexa Fluor 488
Aap	Accumulation-associated protein
AdsA	Adenosine synthase A
AFM	Atomic force microscopy
ANK	Ankyrin
APS	Ammonium persulfate
ASA	Solvent accessible surface area
ATF	Amino-terminal fragment
ATP	Adenosine triphosphate
B_0	External magnetic field in Tesla
Bap	Biofilm-associated protein
Blast(p)	(Protein) basic local alignment search tool
BSA	Bovine serum albumin
BSSA	Buried surface area
c	Concentration
CbpA	Choline binding protein A
CC	Correlation coefficients
CCD	Charge-coupled device
CD	Circular dichroism
ClfA	Clumping factor A
ClfB	Clumping factor B
Cna	Collagen adhesin
CONTINLL	General constrained regularisation method with local linearisation
cv	Column volume
CWA	Cell wall-anchored
C_α RMSD	Root mean square deviation of C_α atoms
D0118	DRESS domains 1-18

D0118_2A488	DRESS domains 1-18 with two A488 fluorophores
D0310_scc	DRESS domains 3-10 with C-terminal Strep tag and C-terminal cysteine-cysteine
D0710	DRESS domains 7-10
D1417	DRESS domains 14-17
D1417M	Monomeric D1417
D1417O	Oligomeric D1417
D1617	DRESS domains 16-17
D17	DRESS domain 17
ΔL	Contour length
D_{max}	The largest dimension of a particle, from P(r)
DMSO	Dimethyl sulfoxide
DNA	deoxyribonucleic acid
DNAse	Deoxyribonuclease
dNTP	Deoxynucleotide 5'-triphosphate
DpnI	Restriction enzyme that digests methylated DNA
DRESS	DUF1542 rigid extracellular surface structural
DRI	Differential refractive index
DTT	1,4-dithiothreitol
DUF	Domain of unknown function
E. coli	*Escherichia coli*
E. faecalis	*Enterococcus faecalis*
Ebh	ECM-binding protein homologue
ECM	Extracellular matrix
EDTA	Ethylenediaminetetraacedic acid
EG	Ethylene glycol
EM	Electron microscopy
EMBL	European Molecular Biology Laboratory
emCCD	Electron multiplying CCD
EOM	Ensemble optimisation modelling
Epf	Extracellular protein factor
ε	Molar extinction coefficient
ESI MS	Electrospray ionisation mass spectrometry
et al.	*et alia*
ExPASy	Expert protein analysis system

F	Mechanical strength
FID	Free induction decay
FIVAR	Found in various architectures
FmtB	Factor which affects the methicillin resistance level and autolysis in the presence of Triton X-100 protein B
FnBPA	Fibronectin binding protein A
FnBPB	Fibronectin binding protein B
FRET	Fluorescence resonance energy transfer
GA	Protein G-related albumin binding module
GF	Growth factor, part of u-PA
Gnc*N*Ac-Mur*N*Ac	*N*-acetylglucosamine-*N*-acetylmurumic acid
GST	Glutathione S-transferase
HEPES	N-(2-hydroxyethyl)piperazine-N'-(2-ethanesulfonic acid)
His_{10}-tag	Decahistidine tag
His_6-tag	Hexahistidine tag
hnNOE	Heteronuclear Nuclear Overhauser Effect
HPSF	High purity salt free
HRV 3C protease	Human rhinovirus cysteine protease
I(q)	Scattering function
IgG	Immunoglobulin G
Im9	Immunity protein 9
IMAC	Immobilised metal affinity chromatography
IPTG	Isopropyl–β– thiogalactopyranoside
IsdA	Iron-regulated surface determinant A
IsdB	iron-regulated surface determinant B
IsdC	Iron-regulated surface determinant C
IsdH	Iron-regulated surface determinant H
I-TASSER	Iterative threading assembly refinement
JCSG	Joint Centre for Structural Genomics
JTT PAM	Jones, Taylor and Thornton accepted point mutation scoring matrix
LB	Lysogeny broth
LLG	Log likelihood gain
LTA	Lipoteichoic acids
MAD	Multiple -wavelength anomalous dispersion

MAFFT	Multiple alignment using fast Fourier transform
MALDI-ISD	Matrix-assisted laser desorption/ionisation with in-source delay
MBP	Maltose binding protein
MDD	Multi-dimensional decomposition
mdeg	Millidegrees
MES	2-(N-morpholino)ethanesulfonic acid
MG2	Salivary mucin
MLCT	Microlever AFM probe with soft silicon nitride tips
MOPS	3-(N-morpholino)propanesulfonic acid
MPD	2-methyl-2,4,pentanediol
MQ	MilliQ water
MR	Molecular replacement
MRE	Molar residual ellipticity
Mrp	Multi repeat polypeptide
MRSA	Methicillin-resistant *S. aureus*
MS	Mass spectrometry
MSA	Multiple sequence alignment
MSCRAMM	Microbial surface component recognising adhesive matrix molecules
MUSCLE	Multiple sequence comparison by Log-expectation
MW	Molecular weight
MWCO	Molecular weight cut-off
Nano-DSF	Nano-differential scanning fluorimetry
NCBI	National Centre for Biotechnology Information
NCL	Native chemical ligation
NCTC	National collection of type cultures
NEAT	Near iron-transport
NEB	New England Biolabs
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
NSD	Normalised spatial discrepancy
NUS	Non-uniform sampling
OD_{600}	Optical density measured at 600 nm
ω	Larmor frequency

ORF	Open reading frame
p	Persistence length
P(r)	Distance distribution function
PACT	pH, anion and cation testing screen
PCR	Polymerase chain reaction
PDB	Protein data bank
PEG	Polyethylene glycol
pET	Plasmid with T7 RNA polymerase
PFam	Protein family database
PG	Peptidoglycan
pI	Isoelectric point
PIA	Polysaccharide intercellular adhesin
Pls	Plasma-sensitive surface protein
PNAG	Poly-*N*-acetyl glucosamine
PPI	Protein-protein interaction
ppm	Parts per million
PSF	Point spread function
PSM	Phenol-soluble modulin
QELS	Quasi-elastic light scattering
r.f.	Radiofrequency
$R_{c,g}$	Radius of gyration of the cross-section of a rod
R_g	Radius of gyration
R_h	Radius of hydration
Rib	Resistance to proteases, immunity, group B
RICH	Rich in charged residues
rpm	Rotations per minute
RSA	Relative surface area
S. aureus	*Staphylococcus aureus*
S. carnosus	*Staphylococcus carnosus*
S. epidermidis	*Staphylococcus epidermidis*
S. gordonii	*Streptococcus gordonii*
S. pyogenes	*Streptococcus pyogenes*
S03	SHIRT domain 3
S0304	SHIRT domains 3 and 4

S0304_P704A,P706A	SHIRT domains 3 and 4 with mutations P704A, P706A
SAD	Single-wavelength anomalous dispersion
SAG	Salivary agglutinin glycoprotein
SasA	*S. aureus* surface protein A
SasB	*S. aureus* surface protein B
SasC	*S. aureus* surface protein C
SasD	*S. aureus* surface protein D
SasE	*S. aureus* surface protein E
SasF	*S. aureus* surface protein F
SasG	*S. aureus* surface protein G
SasH	*S. aureus* surface protein H
SasI	*S. aureus* surface protein I
SasJ	*S. aureus* surface protein J
SasK	*S. aureus* surface protein K
SasL	*S. aureus* surface protein L
SasX	*S. aureus* surface protein X
SAXS	Small-angle X-ray scattering
SdrC	Serine-aspartate repeat protein C
SdrD	Serine-aspartate repeat protein D
SdrE	Serine-aspartate repeat protein E
SDS	Sodium dodecyl sulfate
SDS PAGE	Sodium dodecyl sulfate poly-acrylamide gel electrophoresis
Sec	Secretion system
SEC	Size exclusion chromatography
SEC-MALLS	Size exclusion chromatography with multi-angle laser light scattering
SEC-SAXS	Size exclusion chromatography with small-angle X-ray scattering
SFDA	Single fluorophore detection algorithm
SGO0707	CWA protein from *S. gordonii* originating from gene with accession number *sgo_0707*
SH	Src homology
SHIRT	SGO0707 high identity repeat tandem
SHRImP	Single-molecule high resolution Imaging with photobleaching
SOC	Super optimal broth with catabolite repression

SpA	*S. aureus* protein A
SPI	Signal peptidase
SraP	Serine-rich adhesion for platelets
SRP	Signal recognition particle
Srr	Serine-rich repeat
SSM	Secondary structure matching
SspA	Streptococcal surface protein A
SspB	Streptococcal surface protein B
τ_C	The isotropic rotational correlation time; the time it takes for a N-H bond to rotate through one radian
TCEP	Tris(2-carboxyethyl) phosphine hydrochloride
TEMED	N,N,N',N'-tetramethylethylenediamine
TFZ	Translation function Z-score
TIGR	The Institute for Genomic Research
TIRF	Total internal reflection fluorescence
T_m	Melting temperature
tNCS	Translational non-crystallographic symmetry
Tris	Tris(hydroxymethyl)aminomethane
TPR	Tetratricopeptide repeat
TSB	Tryptic soy broth
Uniprot	Universal protein resource
u-PA	Urokinase-type plasminogen activator
UV	Ultraviolet
VCP	Vaccinia virus complement control protein
WLC	Worm-like chain
WTA	Wall teichoic acids

Author's declaration

I declare that this thesis is a presentation of original work and I am the sole author, except when clearly stated in the text and below. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as in Chapter 8: Bibliography.

I acknowledge Dr Alex Bateman for advice regarding the *in silico* redefinition of the domain boundaries of DRESS domains; Dr Huw Jenkins for assistance in running molecular replacement software; Dr Emanuele Paci for advice regarding obtaining the tilt and twist angles from crystal structures of tandemly arrayed domains (see section 3.3.9.4); William Rochira for event detection and analysis of mechanical unfolding data as presented in sections 4.3.8.2 and 6.1.7.3; Dr Fiona Whelan for the crystal structure of tandem SHIRT domains S0304, as shown in Figure 5.1 and Dr Michael Plevin for NMR data of SHIRT mutants in section 6.2.5.

Acknowledgements

First, I would like to thank my supervisors, Prof Jennifer Potts and Dr Christoph Baumann, for their continued support, guidance and feedback and for giving me the chance to work in the stimulating environment at York. I am also grateful for the feedback I received from Prof Lynne Regan, Dr Daniela Barillà and Dr Steven Johnson, in their respective roles as external examiner, internal examiner and members of my Thesis Advisory Panel, and for the collaborations with Dr Alex Bateman, Dr Michael Plevin and Prof Doug Barrick.

I am particularly thankful to members of the Potts, Plevin, Thomas and Barillà labs for their friendship and help. Special thanks go to Fiona Whelan, Andy Brentnall, Judith Hawkhead, Mike Hodgkinson, Laura Clark, Julie Tucker, Jean Wilkinson, Azhar Kabli, Herman Fung, James Gilburt, Reyme Herman, Sam Griffiths and to Clément Dégut for 3D-printing an amazing model of my structure. I would also like to thank the Bioscience Technology Facility, in particular Dr Andrew Leech and Dr Adam Dowle; along with Dr Alex Heyam and Dr Pedro Aguire for NMR-assistance, Dr Alexander Dulebo (Bruker) for AFM-support and Dr Huw Jenkins, Dr Johan Turkenburg and Sam Hart for X-ray support.

This work would not have been possible without the financial support I received from the British Heart foundation. I would also like to acknowledge Diamond Light Source and its staff for outstanding support regarding X-ray and SAXS data collections. Furthermore, I am grateful to the Biochemical and the Biophysical Society for funding to attend the exciting scientific meetings FEBS 2018 and the Astbury Conversation 2018.

My time in York has been unforgettable, thanks to the many friendships I have made. Thank you to Caroline Pearson, my Gradshare partner-in-crime; our pub quiz leader Jack Munns; Nathaniel Holman, Emma Stewart and Alex Haworth for mental thesis support; James Robson, Aritha Dornau, John Armstrong, Rachael Hallam, Stuart Graham and others for sunny lunch times regardless the weather and Sophie Rugg, Liam Chapman, Michelle Rudden, Banushan Balansethupathy and others for all the good times.

I would like to thank Pieternel, Abel, Carol, Johan, Henk and Lia for their visits, support, patience and love. Thank you, Casper and Favourite, for lots of horse hugs! Finally, an enormous heartfelt thank you to my partner Vincent, for always being there for me.

Chapter 1. Introduction

1.1 *Staphylococcus aureus*

Staphylococcus aureus is an opportunistic, Gram-positive bacterium. A reduction in hospital-associated infections by *S. aureus* has been achieved in recent years in developed countries[1], among other factors due to good hygiene practice[2]. Community-associated infections caused by *S. aureus* are also currently in decline, following a peak in infections around 2007[3]. Despite this reduction in prevalence in infections, *S. aureus* remains the major causative pathogen of prosthetic joint infections (27%)[4] and infective endocarditis (25-30%)[5,6], an infection of the heart valves[7] with an associated mortality rate of 30%[6]. This highlights the relevance to understand more about the mechanism of infection, that involves microbial adhesion and accumulation into biofilms[8].

Mostly, *S. aureus* is present in the nose[9] and the persistent nasal carriage rate of *S. aureus* was around 30% in humans in 2018[10]. Carriage of *S. aureus* is strongly associated with an increased risk of infection[11] and this poses a major burden on modern healthcare[12,13]. *S. aureus* can cause a wide range of infections, including but not limited to infective endocarditis[7], infections on in-dwelling medical devices[4,7] or aggressive periodontitis[14].

1.2 *Streptococcus gordonii*

Streptococcus gordonii, a member of the Viridans Group Streptococci[15], is a Gram-positive bacterium normally present among ~700 other bacterial species in the oral cavity[16]. Here, it initiates the formation of dental plaque[17], a multispecies community in which various bacterial species interact closely with each other[18,19]. *S. gordonii* employs streptococcal surface proteins to adhere to the salivary layer on the teeth and damaged epithelial tissue throughout the body[20,21]. The latter binding ability renders *S. gordonii* capable of causing infective endocarditis[22,23]. Incidence of infective endocarditis by *S. gordonii* was 13% in 1993[22,24]. More recently, the Viridans Group Streptococci have been overtaken by *S. aureus* as the leading cause of infective endocarditis[6]. Nevertheless, the overall mortality rate of infective endocarditis is around 30%[6], highlighting the need to better understand the mechanism of infections for both bacterial species.

1.3 Cell wall-anchored (CWA) proteins

1.3.1 Domain organisation

CWA proteins are abundant on the surface of Gram-positive bacteria[8]. At least two classes exist, where one is characterised by the peptidoglycan (PG) anchoring motif LPXTG[25] or NPQTN[26], where X encodes any amino acid[27]. Another class appears anchorless and associates with the cell surface by unknown mechanisms[28]. The proteins under study in this thesis are from the LPXTG class and briefly, they comprise a similar domain organisation, with an N-terminal signal sequence for secretion, a functional A region, a repetitive B region and the LPXTG sorting motif (Figure 1.1)[29,30]. It should be noted here that the A and B designations do not infer sequence similarity; rather, they refer to different architectural parts of CWA proteins.

In the literature, the classification of CWA proteins is based on structural and functional considerations[29], where the largest class comprises the microbial surface component recognising adhesive matrix molecules (MSCRAMM) family[30]. Here, CWA proteins are organised in different classes, based on the architecture of the B region (Figure 1.1). An example of a protein architecture is provided per class, where the "rod-like" property is represented by *S. aureus* surface protein G (SasG), the "beads-on-a-string" property by *S. aureus* protein A (SpA), the "aspartate-serine" property by clumping factor A (ClfA), the "disordered" property by fibronectin binding protein A (FnBPA), the "non-repetitive" class by iron-regulated surface determinant B (IsdB) and the "mixed" property by serine-aspartate repeat protein E (SdrE). Proteins *S. aureus* surface protein C (SasC) and SGO0707 are the subject of study in this thesis and cannot be assigned to a class yet.

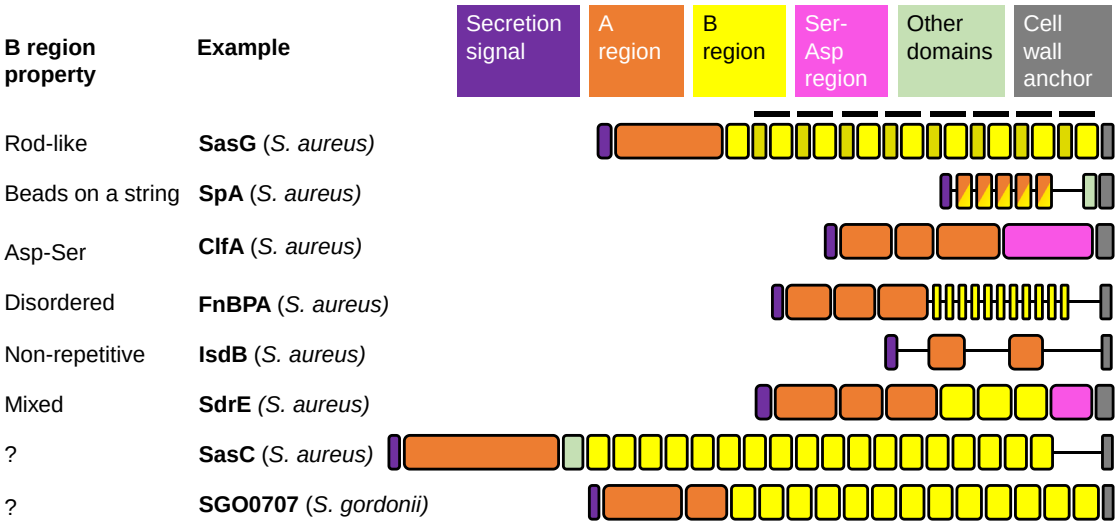


Figure 1.1: Schematic of the domain organisation of several LPXTG-containing CWA proteins. Top: the colour of the boxes matches specific regions in CWA proteins. If a B repeat consists of multiple domains, this is indicated by lines above the B region. Regions are drawn to relative scale according to the number of residues.

1.3.1.1 A region and functions of CWA proteins

The A region is defined as the most N-terminal region in the architecture of CWA proteins, following the secretion signal[29,30]. While the sequence of the signal peptides is mostly conserved in CWA proteins (see sections 1.3.3 and 1.3.4), the sequences of A regions vary[31]. This is reflected in the diversity that is observed in structure and function of the A region. Here, functions of the A regions are underlined and some structural examples of parts of the A regions are provided.

The second domain of the A region of IsdB binds haem (Figure 1.2D)[32,33]. Bacteria require iron for virulence, yet free iron in the host is very limited[34]. Therefore, iron uptake is an important function of CWA proteins. *S. aureus* lyses erythrocytes, releasing haemoglobin[34]. The family of near iron-transport (NEAT) proteins has strong binding affinity for haemoglobin; for example, the binding affinity of IsdB for haemoglobin is 0.42 μM[35], and comprises a well-described pathway to obtain iron[34,36].

The N2-N3 domains of SdrE (Figure 1.2A) bind to complement factor H as an immune evasion strategy[37]. Usually, neutrophils of the host immune system eliminate invading bacterial cells[38]. In this strategy, bacterial CWA proteins bind to host molecules, to disguise themselves as host cells and lower bacterial killing by the host immune system[39]. Other examples of a CWA proteins with immune evasion properties are SasX[40], with an unknown mechanism, and the five three-helix bundles of SpA binding antibodies[41].

SpA presents an example of a CWA protein with an interesting domain architecture. SpA was the first CWA protein of *S. aureus* to be reported[42] and it contains five repetitive sequences that bind antibodies[43]. The A region of SpA (Figure 1.2E) lacks a typical β-sheet rich IgG-like fold that is observed in many other CWA proteins (Figure 1.2A-D). Instead, the domains form three-helix bundles, connected by flexible linkers[44,45]. Here, the repetitive domains of SpA are interpreted as a hybrid merging the functionality of the A region with the repetitive architecture of the B region.

Many CWA proteins are involved in various aspects of adhesion, with a potential for virulence[29,30]. This includes adherence to proteins in the host extracellular matrix (ECM), such as fibronectin and fibrinogen, collagen and elastin. The N2-N3 domains of FnBPA (Figure 1.2B) bind fibrinogen[46] and disordered repeats in the B region of FnBPA bind fibronectin[47] (see section 1.3.1.2). SpA binds, among other things, the heavy and light chains of host antibodies[48] and the von Willebrand factor[49], present at sites of endothelial damage. Some CWA proteins mediate adherence to specific receptors on host cells, such as serine-aspartate repeat protein D (SdrD) to desmoglein 1[50]. Other proteins mediate adhesion to abiotic surfaces via hydrophobic interactions, such as serine-aspartate repeat protein C (SdrC)[51]. Cell-cell interactions can be mediated by the B regions of SasG and SdrC (see section 1.3.1.2). Finally, collagen adhesin (Cna) has immunoglobulin G (IgG)-like domains N1 and N2 domains (Figure 1.2C) that bind collagen[52], the most abundant protein in the human body[53,54]. Adhesion of CWA proteins is a major contributor to the formation of bacterial aggregates causing infections, defined as biofilms (see section 1.4)[8,29,30].

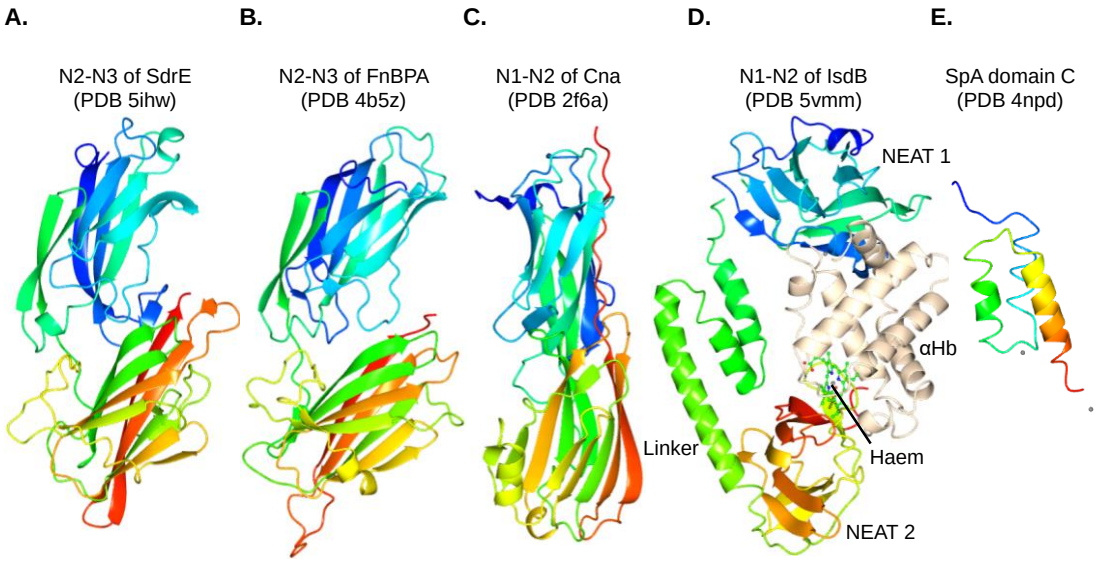


Figure 1.2: Crystal structures of A regions of CWA proteins. Models blended from N (blue) to C (red). **A.** N2-N3 of SdrE (Protein Data Bank (PDB) entry 5ihw)[55]. **B.** N2-N3 of FnBPA (PDB 4b5z)[46]. **C.** N1-N2 of Cna (PDB 2f6a)[52]. **D.** N1-N2 of IsdB with in beige, the α subunit of haemoglobin (αHb; PDB 5vmm)[32]. **E.** SpA domain C (PDB 4npd)[56]. Image was created using CCP4mg.

1.3.1.2 B region

Many CWA proteins comprise repeats in the B region that are proposed to serve a structural role in spatially separating the (putatively) functional A region from the cell surface[29,30,57,58]. Different architectures of B regions are present in the literature, on which the classification system of CWA proteins is based that is used in this thesis. In this section, an example is described from each class.

Rod-like: SasG

The B region of SasG contains G5 domains of 78 residues, separated by E domains of 50 residues[59]. Together, E and G5 domains form a B repeat, with a pairwise sequence conservation of 90-100%[59]. The repeat number varies from three to ten[60]. G5 domains fold independently into flat, highly elongated β-sheets and in the presence of folded G5 repeats, the E domains fold cooperatively from an intrinsically disordered state into a mechanically stable domain[57]. In isolation, E domains are disordered[59]. The interface contributes to the stability of the E domains and the folding dependence of the E domains generates long-range cooperativity and stability across the B region of SasG[57,59]. The B region of SasG has been proposed to mediate biofilm accumulation through dimerisation in the presence of zinc ions[61].

Beads on a string: SpA

The "beads-on-a-string" model is representative for repetitive regions, whose individual domains are monomeric and represented by beads, attached to a more flexible chain of amino acids[62]. The hybrid A/B region of SpA contains five triple-helical bundles that are connected by flexible, conserved linkers (Figure 1.2E)[56]. The stability of the bundles is thermodynamically decoupled and increases from the most N- to the most C-terminal domain[56,45]. SAXS analysis revealed that the flexibility of the linkers allows description of the triple-helical bundles as being connected by flexible linkers in a beads-on-a-string mode[45]. Following the triple-helical bundles, SpA contains a variable number of octapeptide repeats with the sequence KPGKEDGNK (Universal Protein Resource (Uniprot) entry P02976)[29] with approximately 83% sequence identity (Protein family (PFam) database accession number P02976). This is followed by a lysin motif, thought to be involved in binding to extracellular polysaccharides, such as the main chain of PG (see Figure 1.6A). Currently, no function is proposed for the octapeptide repeat region of SpA.

Asp-Ser: ClfA

The B region of ClfA contains an aspartate-serine repeat region comprising 156 DS repeats. The first 33 repeats contain some sequence divergence at the serine or aspartate position, such as DP, GS, AS and DN. The following 108 repeats have a conserved sequence of DS. The final 15 repeats of the repetitive region again show some sequence divergence, this time in the aspartate position, containing the sequences AS, SS, ES, NS and VS (Uniprot entry Q2G015). The B region of ClfA is predicted to lack secondary structure[29], it contains many negatively charged residues and comprises ~$^{1}/_{3}$ of the primary sequence, lowering the theoretical pI of ClfA to 3.22[63].

Disordered: FnBPA

The B repeats of FnBPA are intrinsically disordered and six of the eleven repeats fold upon binding to fibronectin[47]. Repeats are connected by coil structures that might allow FnBPA to flexibly bind multiple fibronectin molecules[47]. With ~40 residues, the repeats are short and have an average pairwise sequence identity of ~24%[47].

Non-repetitive: IsdB

IsdB from the NEAT family does not contain a repetitive region C-terminal to the functional A domains. Instead, the regions flanking iron-scavenging domains comprise 15-70 hydrophilic and charged residues, with low sequence conservation[36]. This might be to reduce cell surface hydrophobicity[29].

Mixed: SdrE

SdrE comprises three repetitive domains in the B region, followed by ~130-170 Asp-Ser repeats, prior to the characteristic cell wall attachment motif[64]. The repetitive domains have an increasing sequence conservation from 42% (B1) to 95% (B3)[64]. All repetitive domains rely on Ca^{2+} binding for folding[65]. Based on the structure of B1 from SdrD, comprising five repetitive domains, the B region has a proposed function as spacer and spring[66].

1.3.2 Secretion

Before CWA proteins can be displayed on the cell surface, they have to be secreted through the cell membrane into the extracellular environment. The secretion system of Gram-positive bacteria has not been studied in detail[67], but Gram-positive and Gram-negative bacteria are thought to share a similar general secretion system (Sec)[68], to secrete proteins in an unfolded state. Important elements of the Sec secretion system include at least a signal recognition particle (SRP)[69], signal peptidase (SPI)[69], an adenosine triphosphate (ATP) motor driving secretion (SecA), a translocation channel (SecYEG) and a complex that releases the transported protein from the Sec system (SecDF/YajC; see Figure 1.4)[67,68,70].

Proteins destined for the extracellular environment contain signal peptides, located at the N-terminus of the nascent polypeptide. Typically, the N-terminal part of the signal sequence is positively charged (Figure 1.3), which might favour electrostatically driven migration towards the cell membrane that contains negatively charged phospholipids[71] or recognition by SRPs[68]. The central hydrophobic region usually forms an α-helix of an appropriate length to reach the extracellular side of the cell membrane. A partly conserved glycine or proline residue is usually located four to six residues upstream of the Ala-X-Ala motif at or near the extracellular membrane surface[71], that can be hydrolysed by an SPI (Figure 1.3)[69].

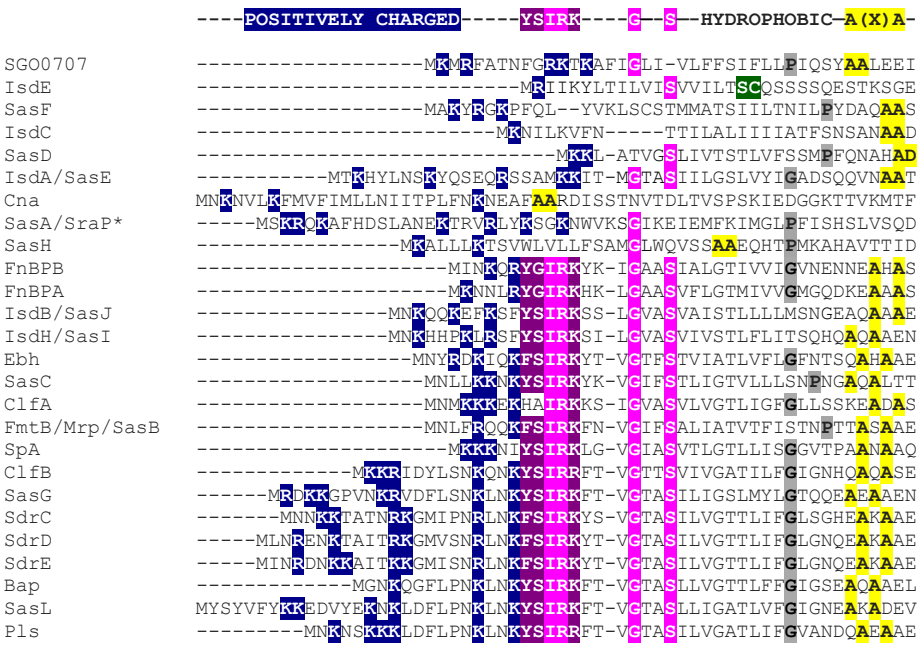


Figure 1.3: Multiple sequence alignment (MSA) of the YSIRK/GXXS region of precursor CWA proteins from *S. aureus* (Table 1.1) and SGO0707 (*S. gordonii*). For the full names of CWA proteins, please refer to Table 1.1. Residues 1-180 were aligned by Multiple Sequence Comparison by Log-Expectation (MUSCLE)[72,73] and the first 60 residues from the alignment containing the signal peptide are shown. Blue: positively charged residues. Pink: mostly conserved YSIRK/GXXS motif. Purple: strongly conserved YSIRK/GXXS motif. Grey: partly conserved glycine/proline residue. Yellow: putative recognition sequence for SPI cleavage, AXA. Green: putative recognition sequence for lipoprotein signal peptidase[74]. *No signal peptide predicted by SignalP-5.0[74].

In addition, ~$^{2}/_{3}$ of Gram-positive precursor CWA proteins from *S. aureus* comprise a mostly conserved signal peptide sequence between the charged and hydrophobic region, composed of two motifs close in sequence, YSIRK and GXXS, where X represents any amino acid (Figure 1.3)[70]. CWA proteins bearing the YSIRK/GXXS signal are incorporated into the growing cell wall (see section 1.3.3), adjacent to the newly forming septum of dividing cells[75,76] in a ring-like distribution (Figure 1.4B). Precursor CWA proteins lacking the YSIRK/GXXS signal are attached to punctate locations near the cell poles rather than in a ring-like distribution[76] (Figure 1.4B).

Underlined residues in the YSIRK/GXXS sequence are fully conserved in proteins bearing this motif[76] and the conserved residues are essential for secretion into the growing cell wall, as shown by mutational studies[70]. Maturation of precursor CWA proteins with the YSIRK/GXXS motif involves proteolytic cleavage of the peptide bond N-terminal to the GXXS motif, as shown by mass spectrometry (MS) experiments on secreted SpA with variants of this signal sequence[39]. However, it is unknown if this motif replaces the function of the Ala-

X-Ala motif, which motif is hydrolysed first and what the processing enzyme(s) and/or mechanism(s) are[70,77].

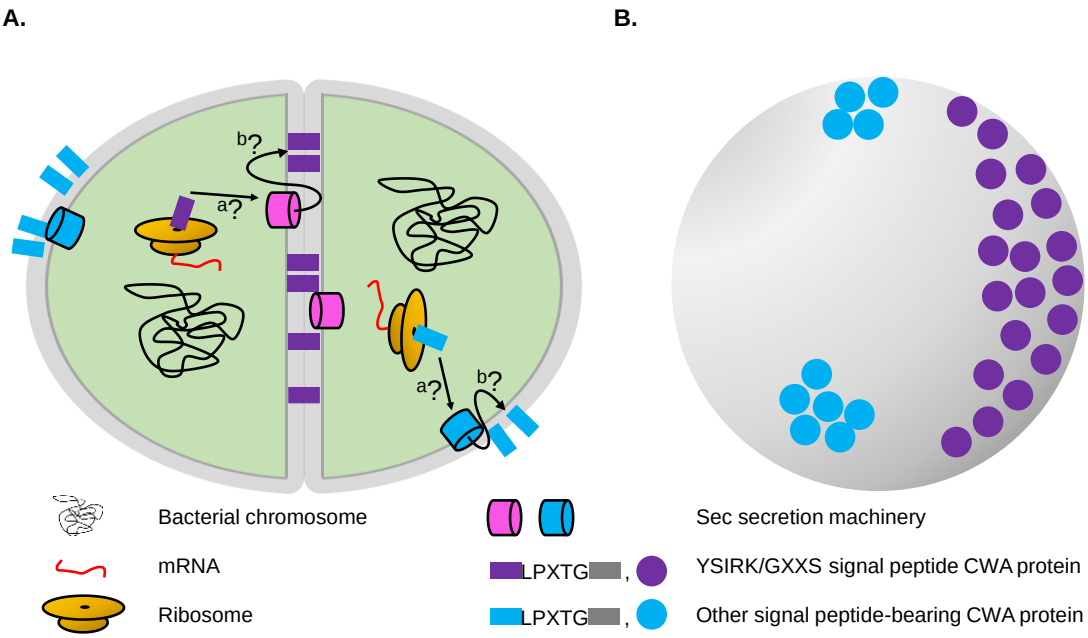
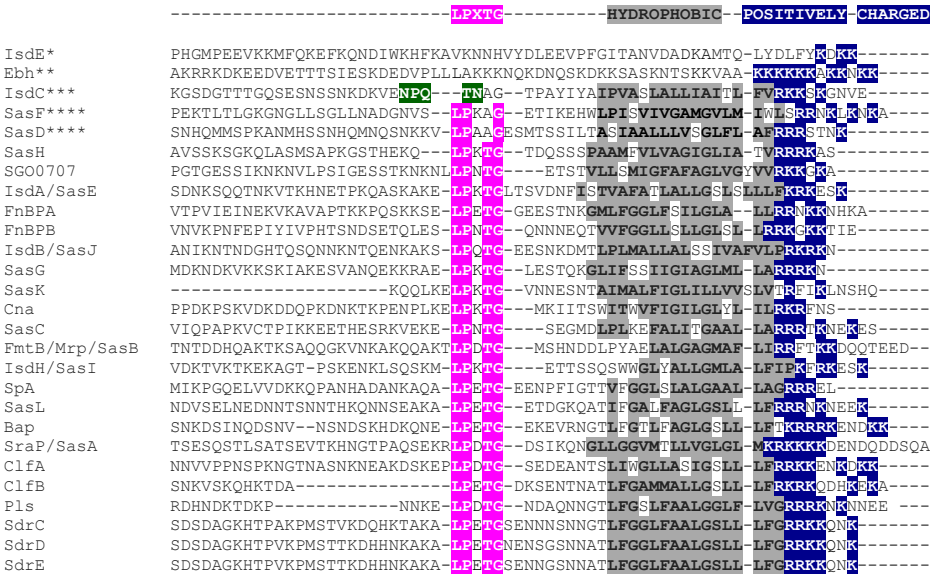


Figure 1.4: YSIRK/GXXS signal peptide directs secretion of CWA proteins towards the cross-wall [76]**. A.** Schematic of a possible translation, targeting and secretion pathway of precursor CWA proteins with and without the N-terminal signal peptide YSIRK/GXXS in dividing cells of *S. aureus*. **B.** Schematic of the surface coverage of CWA proteins

Many parts of this secretion process remain poorly understood. Several CWA proteins contain a highly extended, repetitive region[57,58]. To prevent potential adverse effects of such a CWA protein being folded in the cytoplasm, co-translational secretion might take place[68], where the SRP binds to the signal peptide of the nascent polypeptide and targets the ribosome to a Sec secretion system (Figure 1.4A question mark a)[78]. The presence of non-optimal codons in the signal peptide sequence may slow down translation and facilitate SRP recognition[79]. Although only one type of Sec secretion system is known, precursor CWA proteins are targeted to Sec systems in different locations in the cell; this might involve selective SRPs and/or SRP receptors[78]. Then, the secretion process itself remains elusive (Figure 1.4A question mark b): it is suggested that the N-terminal positively charged part of the signal peptide remains intracellular, while the rest of the polypeptide is translocated through SecYEG, forming a loop, followed by cleavage of a signal motif by an extracellular SPI[71]. This might be followed by LPXTG-mediated incorporation into the growing cell wall, where Sortase captures newly secreted CWA proteins directly from the Sec channel[80].

1.3.3 Covalent linkage to the cell wall

The cell wall serves a structural role to withstand the internal turgor pressure of the cell[81]. The precursor CWA proteins are secreted by the Sec secretion machinery in unfolded state[68]. Most CWA proteins contain an LPXTG motif (Figure 1.5), where underlined residues are mostly conserved and X represents any amino acid. The hydrophobic region following the conserved motif retards secretion[25], facilitating binding of the CWA protein by Sortase A[82]. The schematic process of the incorporation of CWA proteins into the PG structure of *S. aureus* is shown in Figure 1.6.



*membrane protein[83].
**lacks CWA anchor motif, yet is tightly associated with the cell wall[84].
***anchored by Sortase B[26].
****alternative LPXTG motif[80]. SasF is non-covalently associated with the cell wall[85].

Figure 1.5: MSA of the LPXTG region of CWA proteins from *S. aureus* (Table 1.1) and SGO0707 (*S. gordonii*). For the full names of CWA proteins, please refer to Table 1.1. The most C-terminal 120-180 residues were aligned by MUSCLE[72,73] and the most C-terminal 60-70 residues are shown. Blue: positively charged residues. Pink: conserved residues in LPXTG motif. Grey: hydrophobic residues. Green: NPQTN motif anchored by Sortase B[26].

Sortase A is an extracellular transpeptidase that is anchored in the cell membrane[82]. The catalytic Cys184, in conjunction with other catalytic residues[86], hydrolyses the peptide bond between threonine and glycine and forms a thio-ester acyl-enzyme intermediate[87] (Figure 1.6B, step 1). Subsequently, Sortase A catalyses the formation of an amide bond with the amino group of a free pentaglycine crossbridge[88], present in a unit of a lipid II PG precursor molecule (Figure 1.6B, step 2) [87,89]. Subsequently, this linked unit is incorporated in the growing PG cell wall (Figure 1.6B, step 3)[88]. The surface of *S. aureus* is approximated to be decorated by one CWA protein per surface *N*-acetylglucosamine-*N*-acetylmurumic

acid (Gnc*N*Ac-Mur*N*Ac) disaccharide, based on a cell wall thickness of ~20-40 nm thick[90–92] and ~20 layers of PG[93] (Figure 1.6A, C)[93].

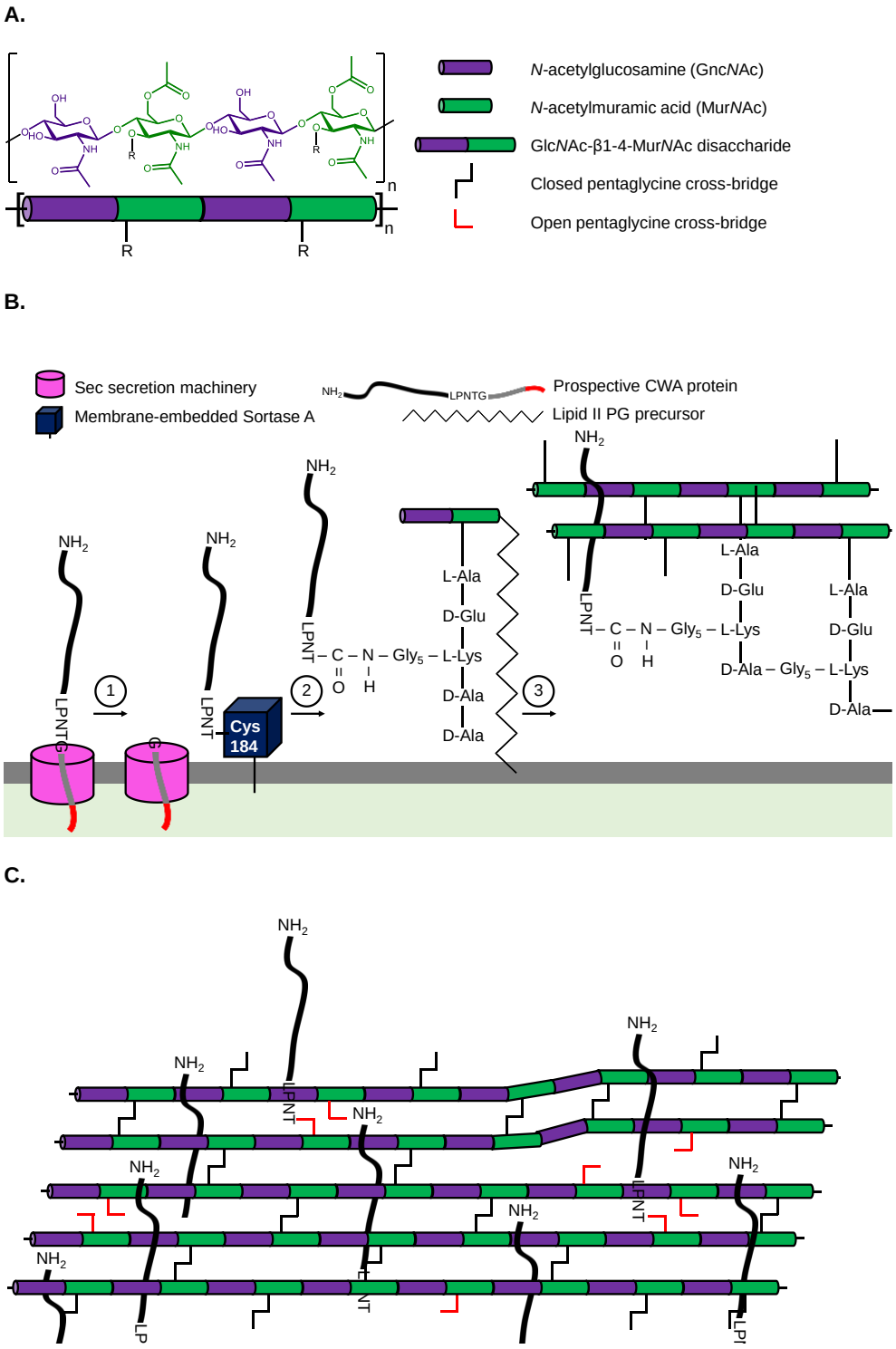


Figure 1.6: Schematic of the PG structure of *S. aureus* and the incorporation of CWA proteins, here SasC is shown. **A.** Molecular structure of a β1-4 linked disaccharide of Glc*N*Ac-Mur*N*Ac, where R is the peptide crossbridge (see B). **B.** Schematic of the secretion and incorporation of CWA proteins in the PG layer. **C.** Schematic of the resulting structure of PG containing CWA proteins.

The PG layer is constantly being remodelled[94]. Bacteria produce autolysins, which can create breaks in the PG layer and allow the insertion of new material, facilitating cell growth and surface remodelling[95]. Bacteria adhere to the "make-before-break" principle[96], maintaining their cell wall integrity by forming new strands before breaking old ones. New PG layers are thought to push old layers outwards, thereby stretching these layers, making the old layers more susceptible to autolysins[97]. The resulting Mur*N*Ac fragments are subsequently partly recycled, a process that is essential for survival in the stationary growth phase[98].

The expression level and display of specific proteins of *S. aureus* varies, depending on planktonic or biofilm state (see section 1.4)[99]; growth rate[100]; environmental conditions, for example, iron-regulated surface determinant A (IsdA) is over-produced in iron-limiting conditions[34], and whether it is a commensal or pathogenic strain[101]. Together with the remodelling process, this allows for the incorporation of different CWA proteins, that will be attached to the newly synthesised PG chains closest to the cell membrane. The domain architecture of many CWA proteins (see section 1.3.1) comprises a putatively extended repetitive B region (SasG[57], the ECM- binding protein homologue Ebh[58], SdrD[66], the clumping factor B (ClfB)[102], Cna[103], biofilm-associated protein Bap[104]). With an approximated pore size of ~2.2 nm for Gram-positive bacteria that would allow globular proteins of ~23 kDa to freely diffuse through the PG layer[91] and the notation that in stretched PG, the pores are larger[105]; an extended B region might penetrate the PG mesh. With a cell wall thickness of ~20-40 nm[90–92], projection beyond the cell surface will depend on the length of the B region, the relative depth of the CWA protein and the orientation of the B region with respect to the cell membrane.

1.3.4 CWA proteins of *S. aureus*

Currently, at least 27 CWA proteins are known that can be present on the cell wall of *S. aureus* (Table 1.1).

Table 1.1: CWA proteins on the cell wall of *S. aureus.* FIVAR: motif found in various architectures. GA: protein G-related albumin binding.

Protein	Full name	Function	B region	Ref.
SasA/ SraP	*S. aureus* surface protein A/ serine-rich adhesion for platelets	Adherence to salivary agglutinin gp340 in human saliva and human platelets.	SD repeats	76,106,107,108
FmtB/ SasB/ Mrp	Factor which affects the methicillin resistance level and autolysis in the presence of Triton X-100 protein B/ *S. aureus* surface protein B/ multiple repeat polypeptide	Mediates indirect resistance against oxacillin.	18 DUF1542 repeats	107,109,110
SasC	*S. aureus* surface protein C	Mediates biofilm aggregation, mechanism unknown.	18 DUF1542 repeats	111
SasD	*S. aureus* surface protein D	Structure and function unknown.	Unknown	76,112,107
SasE/StbA	*S. aureus* surface protein E/ staphylococcal transferrin-binding protein A	Implicated in transferrin binding for iron acquisition.	No repeats	107,113
SasF	*S. aureus* surface protein F	Structure and function unknown.	Unknown	107
SasG	*S. aureus* surface protein G	Adheres to nasal epithelial cells	3-10 (E-)G5 repeats	107,114
SasH/ AdsA	*S. aureus* surface protein H/ Adenosine synthase A	Converts adenosine monophosphate to adenosine, thereby implicated in immune evasion.	No repeats	107,115
SasI/IsdH	*S. aureus* surface protein I Iron-regulated surface determinant H	Binds haptoglobin-haemoglobin and haem	3 NEAT domains, no repeats	36,107
SasJ/IsdB	*S. aureus* surface protein J; Iron-regulated surface determinant B	Binds haem	2 NEAT domains, no repeats	36,107,113
SasK	*S. aureus* surface protein K	Structure and function unknown.	Unknown	76,107,112
SasL	*S. aureus* surface protein L	Structure and function unknown.	Unknown	112
SasX	*S. aureus* surface protein X	Nasal colonisation, promotes immune evasion	Unknown	40
Bap	Biofilm-associated protein	Prevents internalisation into host cells, binds host receptor GP96	Variable repeat number, SD-repeats	116,117

ClfA	Clumping factor A	Bind fibrinogen	SD-repeats	[63]
ClfB	Clumping factor B	Binds fibrinogen. Activity can be masked by ClfA.	SD-repeats	[118]
SdrC	Asp-Ser repeat protein C	N2 domain mediates homodimeric cell-cell interactions	2 B repeats, SD-repeats	[65,119]
SdrD	Asp-Ser repeat protein D	Adheres to host cell receptor desmoglein 1	5 B repeats, SD-repeats	[50,64]
SdrE	Asp-Ser repeat protein E	Binds human platelets	3 B repeats, SD-repeats	[64,120]
Cna	Collagen adhesin	A region binds collagen	1-4 B repeats	[52,121]
IsdA	Iron-regulated surface determinant A	Binds fibrinogen and fibronectin, binds haem.	1 NEAT domain, no repeats	[76,122]
IsdC	Iron-regulated surface determinant C	Binds haem.	1 NEAT domain, no repeats	[36,76]
Ebh	ECM-binding protein homologue	Binds fibronectin.	52 repeats of FIVAR-GA modules	[84]
FnBPA	Fibronectin-binding protein A	Binds fibronectin, fibrinogen and elastin	11 fibronectin-binding repeats	[123,124]
FnBPB	Fibronectin-binding protein B	Binds fibronectin, fibrinogen and elastin.	10 fibronectin-binding repeats	[124,125]
SpA	*S. aureus* protein A	Binds antibodies, promotes immune evasion.	5 repeats of three-helix bundles	[41]
Pls	Plasma-sensitive surface protein	Binds fibrinogen and fibronectin, following processing by plasmin.	5 B repeats, SD-repeats	[126,127]

1.4 Biofilms

1.4.1 Introduction to biofilms

Biofilms are microbial accumulations adhering to a biological or non-biological surface, for example damaged tissue or an in-dwelling medical device, respectively[8]. Many microorganisms thrive in a multispecies biofilm[128]; for example, *S. gordonii* is part of a multispecies biofilm in dental plaque[18]. Biofilms of *S. aureus* biofilms tend to be monospecies[129,130]. However, cases have been described where *S. aureus* was successfully co-cultured into biofilm-state with *Pseudomonas aeruginosa*[131], among other species[132].

1.4.2 Biofilms are partly responsible for antibiotic resistance

Planktonic bacteria have acquired antibiotic resistance through various mechanisms, such as adaptational mutations[133], acquisition of genetic material containing an antibiotic resistance gene[134,135], or alteration of gene expression[136]. When bacteria without acquired antibiotic resistance mechanisms transition into a biofilm state, they are still more resistant to antibiotics than their planktonic counterparts[137]. For example, the susceptibility of *S. aureus* against the last-resort antibiotic vancomycin is up to sixteen times lower, when *S. aureus* cells are resident in a biofilm environment[138]. The antibiotic resistance mechanism of cells in a biofilm is not yet well-understood. One hypothesis is that cells transition into a lower metabolic state[8] due to nutrient limitation[139] (due to lower diffusion of nutrients in the biofilm matrix[140,141]) and amplified by, for example, a thicker, finely meshed biofilm[137,142]. Finally, in nutrient-limiting conditions, a general stress response is activated[143], which might promote the lower metabolic state.

1.4.3 *S. aureus* biofilms

Biofilm formation by *S. aureus* can be regarded as a 'phased' process (Figure 1.7). During the initial adherence phase, planktonic cells adhere to a host tissue or in-dwelling medical device[144], which are typically covered in fibrinogen[145] and fibronectin[146]. In the adherence phase, attachment is mediated by hydrophobic interactions, such as via the intrinsically disordered region of SdrC[147], electrostatic interactions, such as via negatively charged (wall) teichoic acids[148], or specific protein-protein interactions (PPI) between CWA proteins[30] and proteinaceous surface components[149,150].

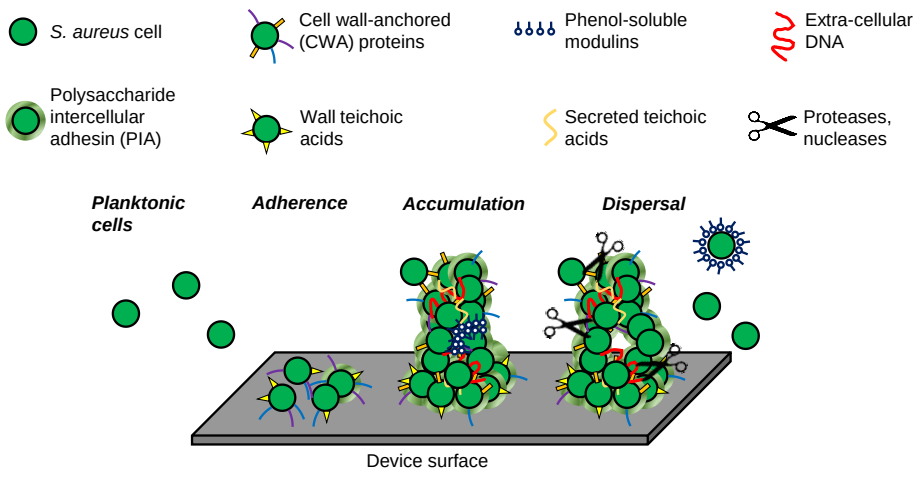


Figure 1.7: Biofilm formation and maturation by *S. aureus*.

Through an unknown mechanism, cells switch from the adherence phase to the accumulation phase, where the attached cells accumulate into a three-dimensional, mature biofilm, in which cell-cell interactions are key[150]. This state is achieved by growth and division of metabolically active, attached cells[8] and usually involves the production of a polysaccharide intercellular adhesin (PIA) matrix[151]. For a long time, the secretion of PIA was the only mechanism known to create a matrix, which allowed cells to accumulate into a mature biofilm[152] and the PIA matrix of *S. aureus* consists of poly-*N*-acetyl glucosamine (PNAG; see section 1.3.3)[153]. Importantly, CWA proteins can also mediate biofilm accumulation, independent of PIA[29,30,154]. For example, inter-bacterial interactions are mediated by the B region of SasG through dimerisation in the presence of zinc ions[57] or by dimerisation of the B region of SdrC[147]. Many other CWA proteins (SasC[111], SpA[155], FnBPA[156], FnBPB[124], ClfB[157], SasX[40], Bap[116]) are suggested to be involved in biofilm formation and/or accumulation, but information about some mechanisms remains unclear[30]. Finally, the expression and display of various CWA proteins can generate anomalous effects in terms of adherence: the display of Bap masked the adherence capacity of FnBPA, FnBPB, ClfA and ClfB[104] and the overexpression and display of SasG comprising ≥5 B repeats masked the binding abilities of SpA, ClfB, FnBPA and FnBPB, whereas no effect was observed for <5 B repeats[158]. Other contributing factors to the accumulation phase of *S. aureus* biofilm formation are secreted teichoic acids, wall teichoic acids (WTA), lipoteichoic acids (LTA)[159] and extracellular deoxyribonucleic acid (DNA) from lysed bacteria[160].

Together, the matrix components protect the bacteria from environmental stresses, such as dehydration, retard the diffusion of antimicrobial agents[8] and provide the growing maturing biofilm with nutrients by trapping extracellular enzymes[161], such as aminopeptidases that degrade proteins[162]. The accumulation phase is quite challenging for bacteria, due to environmental stresses such as dehydration, nutrient limitation and slow nutrient diffusion, shear stresses[163] and the host immune system[164]. The secretion of phenol-soluble modulins (PSM) creates nutrient channels through the biofilm via the disruption of non-covalent molecular interactions[150]. PSMs have also been reported as pro-inflammatory initiators[150,165] and virulence factors[166].

The dispersal phase enables bacteria to disseminate from the mature biofilm and spread the infection via the blood stream[163]. *S. aureus* actively secretes several proteases[167–169]

and employs secreted[170] and extracellular[171] nucleases to degrade specific components of the biofilm matrix[144]; CWA proteins[167–169], extracellular DNA[170,171] or PIA[172]. Furthermore, PSM surfactant activity contributes significantly to cell dispersal[173]. It seems likely that *S. aureus* has maintained several independent methods of biofilm formation, accumulation and dispersal to increase its probability of survival in very different environmentally challenging conditions[174].

1.4.4 *Streptococcus gordonii* biofilms

Biofilm formation by *S. gordonii* occurs on the tooth surface, where it initiates the formation of dental plaque (Figure 1.8). The adherence phase starts with the colonisation of *S. gordonii* on the pellicle, a layer containing saliva proteins such as salivary agglutinin glycoprotein (SAG), the salivary mucin MG2 and carbohydrates[21]. *S. gordonii* employs CWA proteins, such as streptococcal surface proteins A and B (SspA, SspB) that bind SAG[175] and the homologous serine-rich surface glycoproteins GspB and Hsa that bind SAG and MG2[176], forming a monolayer within two hours[20].

Following monolayer formation, the accumulation phase involves many other bacterial species[16], highlighting the need to form interactions with different bacterial species. The molecular basis for these cell-cell contacts is varied and can comprise the synthesis of an exopolysaccharide matrix[177], pili or fibrillary proteins promoting adhesion between cells[20,178]. For example, members of the serine-rich repeat (Srr) protein family bind carbohydrate motifs displayed on the surface of other bacteria in the oral cavity[21]. Furthermore, SspB also mediates coaggregation with *Actinomyces oris* by binding a receptor carbohydrate on its surface[179]. Finally, gene regulation responses between different bacterial species within dental plaque, such as *S. gordonii* and *Fusobacterium nucleatum*[180] can regulate biofilm formation processes[181].

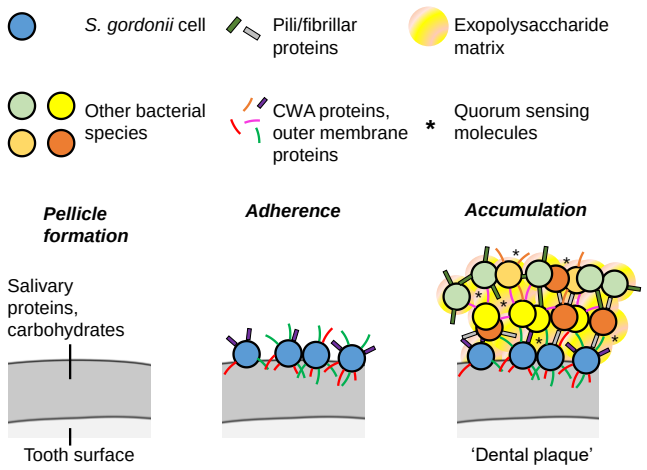


Figure 1.8: Dental plaque formation by *S. gordonii*.

1.5 Protein architectures

1.5.1 Protein domains

1.5.1.1 Definitions

The definition of a protein domain is varied and includes a structurally independent unit[182,183], a structurally compact semi-independent unit[182,184], a repeated sequence[185] or a unit of protein function[182,186]. Domains can be grouped into families depending on their sequence identity[187] and into superfamilies suggestive of a common evolutionary origin, depending on similar three-dimensional structures[187] or clans, based on similar sequences, consensus sequences or structures[188]. In 2018, nearly 18,000 protein families and 628 clans were known to PFam[189].

1.5.1.2 Structural components

The structure of protein domains is usually assembled from secondary structure motifs, such as α helices, β sheets and coils[190], although protein domains can also be devoid of structure[191]. Generally, most folded proteins comprise a hydrophobic core with apolar residues, surrounded by a hydrophilic surface with polar residues as a result of the hydrophobic effect[182]. Residues contributing to the stable formation of the domain fold tend to be more conserved; usually, these are located in the hydrophobic core[192]. Non-globular protein domains can break with the norm as described above. For example, the rod-like repeats in the B region of SasG have a smaller hydrophobic core and yet they assemble into stable folds, due to strong van der Waals interactions across the inter-domain interfaces[59].

1.5.1.3 Folding of protein domains

Protein folding is the transition from disorder to an organisation of secondary and tertiary structure, which usually results in the burial of hydrophobic residues into a hydrophobic core[182]. Generally, protein folding is driven by a lower total free energy for the native protein fold in the forming of a "folding funnel", in comparison to the denatured state[193]. The total free energy is the sum of enthalpic and entropic contributions[194]. Folding of a protein generates hydrogen bonds, salt bridges, electrostatic interactions, van der Waals interactions etc. that contribute to the enthalpy, at the cost of a loss in entropy[195]. The formation of the hydrophobic core has an entropic contribution to the total free energy. The net difference between the folded and unfolded state can be very small and thus, small disruptions can have large effects on the folding state[193,195].

1.5.2 Multi-domain proteins

1.5.2.1 Introduction to multi-domain proteins

Over 65% of 1.1 million proteins in the SCOP Superfamily database in 2007 comprise multiple domains[196]. New multi-domain proteins can be created via the assembly of existing domains[197]. This includes the duplication of domains, the recombination or domain shuffling of existing domains, the acquisition of mutations in existing domains or a combination of the above[198]. The inclusion of different domains into a single protein increases its functionality[198]. Although multi-domain proteins are more common in eukaryotes (~66-80%) than in prokaryotes (~40-66%)[196], all CWA proteins known to date in Gram-positive bacteria contain multiple domains/regions to achieve correct secretion, covalent cell wall anchoring and efficient functionality.

Generally, the sequential order of domains tends to be conserved in the emergence of new multi-domain proteins[198]. This might suggest that, during evolution, not just individual domains but also domain pairs can undergo duplication and recombination events[199]. This process maintains the interface between domains, which can be important for the function or stability of these domains[200,201].

1.5.2.2 Folding of multi-domain proteins

The folding of multi-domain proteins remains relatively poorly characterised, due to the presence of multiple domains and their interactions across the inter-domain interface[196]. In multi-domain proteins with small interfaces, domains tend to fold independently, have

longer linkers and behave according to the beads-on-a-string analogy[196,202]. An example where independent folding is obtained even in the absence of long linkers, is the folding of tandem IgG domains in titin. Tandem domains are in an extended conformation with some bending/twisting between individual repeats[203], featuring minimal inter-domain interactions[204].

In multi-domain proteins with large interfaces, the folding of the domains tends to be linked with the formation of a stabilising interface[196]. To satisfy the formation of these large, intricate interfaces, the linkers between domains tends to be shorter and stiffer[202]. As the stability of the domains depends on their connections, such proteins can have a rod-like architecture[57]. An example, where folding of multiple domains is dependent on the formation of an inter-domain interface, is spectrin. A single domain comprises a triple-helical bundle, that is connected to an adjacent repeat by a contiguous helix[205]. Although the interface between domains is relatively small[196], the stability of tandem domains is higher than of individual domains[205]. Moreover, the interface generates cooperativity in thermal and chemical denaturation[206]. Cooperativity in protein folding is defined as a coupling of interactions, resulting in all-or-none behaviour of a system, compared to the sum of its individual components[207].

1.5.3 Tandem repeats

Tandem repeat domains are the simplest form of a multi-domain organisation, where a single domain or sequence is repeated tandemly[208]. They vary in size from two amino acids, such as Asp-Ser repeats in the B region of ClfA, ClfB, SdrC, SdrD and SdrE[64] to repeats over one hundred residues in size, such as spectrin repeats (106 residues)[205,209]. Tandem repeats are abundant in bacteria and variations in repeat number can contribute to functional diversity and virulence[210]. An example of tandem variability is the repetitive region of SasG, where a minimum number of six B repeats was required for biofilm formation of *S. aureus* caused by SasG[158]. An example of a highly repeated tandem repeat in a single protein is Domain of Unknown Function 1542 (DUF1542), which is repeated 39 times in FmtB from *S. epidermidis* (Interpro accession number A0A0H2VJ55) and 18-19 times in the extracellular protein factor (Epf) from *S. pyogenes*[210].

Misfolding of proteins may lead to protein aggregation and the formation of protein amyloids, which are associated with disease[193]. The potential for misfolding in multi-

domain proteins appears to be mitigated by limiting the sequence identity of adjacent tandem domains to <40%[211,212]. For example, DUF1542 domains, with an average pairwise sequence identity of ~28%, are below this boundary and thus have a theoretical lower tendency to aggregate during unfolding/refolding events. However, a high pairwise identity in tandem repeats might be beneficial for the creation of B regions of varying length, thought to contribute to virulence[210].

1.5.4 Ising model, a model of tandem repeat folding

The Ising model is used for the interpretation of the folding behaviour of linear arrays of some tandemly arrayed, interacting domains[213]. Generally, describing the folding of multi-domain proteins is difficult, due to different thermodynamic properties of individual domains. For some tandemly repeated domains that are generally 20-40 residues in size with a sequence identity of around 25%[214], the folding can be quantitatively described, due to a simplification of the model to include only the intrinsic stability of a domain and the interfacial stability[214].

As an example, Mello and Barrick (2004)[215] applied this to the folding of Notch ankyrin (ANK) domains, which are tandem repeats with a large inter-domain interface. Zweifel and Barrick (2001)[216,217] showed that ANK repeats unfold cooperatively with a single transition and benefit from long-range interactions. Individual repeats were unstable, but stability of the multi-domain protein was provided by the highly favourable interactions at the inter-domain interfaces[215]. The instability of single repeats is in contrast with the definition of domains as a structurally independent unit, as stated in section 1.5.1.1. Rather, their stable structure is only obtained by a long-range stabilising effect of ANK domains assembled in a tandem repeat structure.

Another example of the application of the Ising model to protein folding involves tetratrico peptide repeat (TPR) proteins, assemblies of 3-16+ repeats of 34 amino acids with the scope to form a non-globular, elongated structure[213,218]. The thermal stability increases with the number of repeats, sharpening the unfolding transition[219]. This is quite remarkable, considering that such multi-domain proteins are merely stabilised by local stabilising contacts[220]. Kajander *et alia*[213] showed that the unfolding of TPR domains could be described using the 1D Ising model[213]. This implies that near the mid-point of the unfolding transition, the multi-domain protein comprises significantly populated partially

folded configurations, rather than an all-or-none unfolding transition observed for fewer domains and described for globular proteins[213].

1.6 SasC

1.6.1 Introduction to SasC

SasC is a CWA protein from *S. aureus* with characteristic sequence motifs for Sec secretion and cell wall attachment, along with an A region and a B region that comprises eighteen repetitive domains (Figure 1.9). SasC was identified *in silico* by Mazmanian *et al.* (2001)[221] from a basic local alignment search tool (blast) search on DNA from SpA, encoding the LPXTG motif, followed by 10< apolar residues[221]. Roche *et al*. (2003)[107] proposed a modular domain organisation for SasC, of which an updated version is shown in Figure 1.9. Briefly, SasC comprises an N-terminal YSIRK/GXXS signal peptide that likely directs SasC to the cross wall[76]; an A region with 53% homology to FmtB/Mrp that are implicated, indirectly, with reduced resistance against β-lactams[109]; a B region comprising eighteen DUF1542 domains and a C-terminal LPNTG signal peptide that likely results in covalent attachment to the PG layer[88].

SasC was detected in 90% of all *S. aureus* strains and 94% of invasive strains[107]. This is in approximate agreement with the prevalence of SasC reported by Schroeder *et al*. (2009)[111] of 97% in clinical strains. Schroeder *et al*. observed that SasC mediated strong cell aggregation, when it was overexpressed in *S. aureus* or heterologously expressed in *S. carnosus*[111]. The aggregation ability is proposed to be located in the A region and is somewhat enhanced in the presence of the B region[111]. Since SasC strongly mediates biofilm accumulation[111] and biofilms are implicated with increased antibiotic resistance, it is important to further understand the structure and function of SasC.

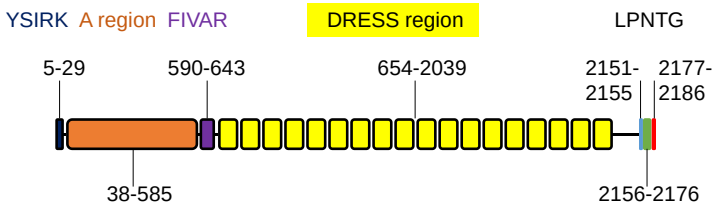


Figure 1.9: Schematic of the domain organisation of SasC. Top: signal peptides and domain annotations. Middle: domain organisation with domains to relative size. Numbers: residue numbers as in accession number C7BUR8.

1.6.2 Homologues to the A region of SasC

The A region of SasC has been shown to contain the adhesive region largely responsible for biofilm accumulation[111]. It has 40% sequence identity and 53% sequence similarity to the A region of FmtB and the full sequence of SasC shares 34% sequence identity with FmtB[111]. Homology of the A region of SasC with other CWA proteins could suggest a similar role.

FmtB, Multiple Repeat Polypeptide (Mrp)[110] and *S. aureus* surface protein B (SasB)[222] are highly homologous *S. aureus* CWA proteins of and shall hereafter be referred to as FmtB. FmtB was identified in 1997 by Komatsuzawa and co-workers[223] as an important factor for methicillin resistance of *S. aureus.* When an insertion was placed within the open reading frame (ORF) of FmtB, cells showed a reduced resistance to the methicillin oxacillin when treated with 0.02% Triton X-100, a non-ionic detergent known to permeabilise and/or lyse membranes[224].

It has been proposed that FmtB could be responsible for cell wall stability or cell wall biogenesis[223]. However, later experiments revealed that methicillin-resistance was not restored completely in a knock-out strain supplemented with FmtB, only after overexpression of FmtB or after supplementing cell wall precursors[109,225]. In support of this, *S. aureus* survives in the absence of, among other genes, FmtB, at the cost of virulence and antibiotic resistance[226]. These results suggest that FmtB is probably not directly involved in resistance against β-lactams[109,225]. Further experiments are required to elucidate the role and mechanism of FmtB and SasC.

1.6.3 B region

The B region of SasC comprises seventeen DUF1542 domains, as determined from the domain boundaries reported by Schroeder *et al*. (2009)[111]. Most (95%) DUF1542 domains are present in Firmicutes, such as staphylococci and streptococci (PFam entry PF07564)[210]. In some proteins, such as Epf in *S. pyogenes,* the number of DUF1542 repeats is variable, which might generate functional diversity[210].

Regions containing DUF1542 repeats are speculated to have a structural role[227]. This would be in agreement with the B regions of other CWA proteins[29,30,57,58]. Electron microscopy (EM) images of the B region from Epf comprising sixteen DUF1542 domains showed an extended structure with kinks of 50-60 nm in length with a diameter of 6 nm (Figure

1.10A)[228]. Structural predictions by iterative threading assembly refinement (I-TASSER)[229] of regions containing multiple DUF1542 domains yielded a globular bundle of helical bundles[210] (Figure 1.10B), where two domains are predicted to form a five-helix bundle and multiple repeats are predicted to have bundled five-helix bundles. Clearly, the predicted bundled bundle structure is in disagreement with the low-resolution structural evidence provided by Linke *et al*. (Figure 1.10A)[228]. Thus, the structure of DUF1542 domains and the size and shape of DUF1542 domains in solution are currently unclear and could provide information about the function of DUF1542 domains and DUF1542-containing proteins.

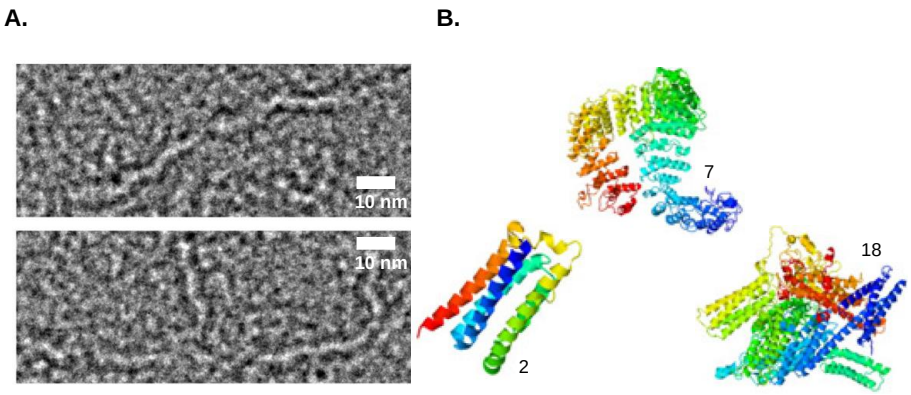


Figure 1.10: EM images and I-TASSER models of multiple DUF1542 domains. A. EM images of 16 DUF1542 domains from the CWA protein Epf from *S. pyogenes*[228]. **B.** I-TASSER models of 2, 7 and 18 DUF1542 domains, image from Lin *et al.* (2012)[210].

Bioinformatical, structural and biophysical experiments on DUF1542 domains from SasC enabled us to rename DUF1542 to DUF1542 rigid extracellular surface structural (DRESS) domains. Throughout this thesis, DUF1542 is used to refer to the old literature domain boundaries of this domain and DRESS is used for the domain boundaries proposed in this thesis.

1.7 SGO0707

1.7.1 Introduction to SGO0707

Davies *et al*. (2009)[230] identified several novel surface proteins from *S. gordonii* via the production of a *sortase*$^-$ mutant, where proteins that are typically covalently linked to the cell wall are now secreted. By electrophoresis of the secreted proteome, followed by MS, they identified a surface protein of *S. gordonii* as originating from a gene with accession number *sgo_0707* (The Institute for Genomic Research (TIGR) database). Here, the protein

encoded by this gene is referred to as SGO0707 and it comprises a Sec secretion signal peptide at the N-terminus, an A region, a repetitive B region and a wall attachment site at the C-terminus (Figure 1.11). The B region comprises highly identical tandem repeats[230], termed SGO0707 high identity repeat tandem (SHIRT) domains.

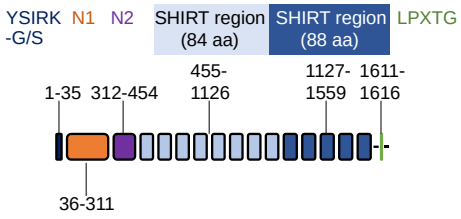


Figure 1.11: Domain organisation of SGO0707 [230,231]. Top: signal peptides and domain annotations. Middle: domain organisation with domains to relative size. Numbers: residue numbers as in accession number A8AW49.

Functional characterisation of SGO0707 revealed binding to oral keratinocytes and type I collagen[231]. The former is in line with the presence of *S. gordonii* in the oral cavity and its involvement in the formation of dental plaque. As collagen is the most abundant protein in the human body[53,54] and 80%-90% of collagen is type I[53], adherence to type I collagen proposes a role for SGO0707, where collagen is exposed, for example on a damaged heart valve, leading to infective endocarditis[17].

Structural characterisation of the A region of SGO0707[231] revealed the presence of two domains, N1 and N2, comprising β-sandwich folds and anti-parallel β-sheets (Figure 1.12A). N1 comprised a putative negatively charged binding cleft bearing a single cysteine and some potential electron density indicating binding of a metal ligand, although the crystallisation conditions did not contain metals. Thermal stability studies indicated stabilisation by Ca^{2+} and the crystal structure of a structural homologue from SspB (PDB 2WD6, a C_α root mean square deviation (RMSD) of 2.82 Å over 179 atoms, Figure 1.12B,C) showed a tightly bound calcium ion. N2 consists of two domains both forming an IgG-like fold[231].

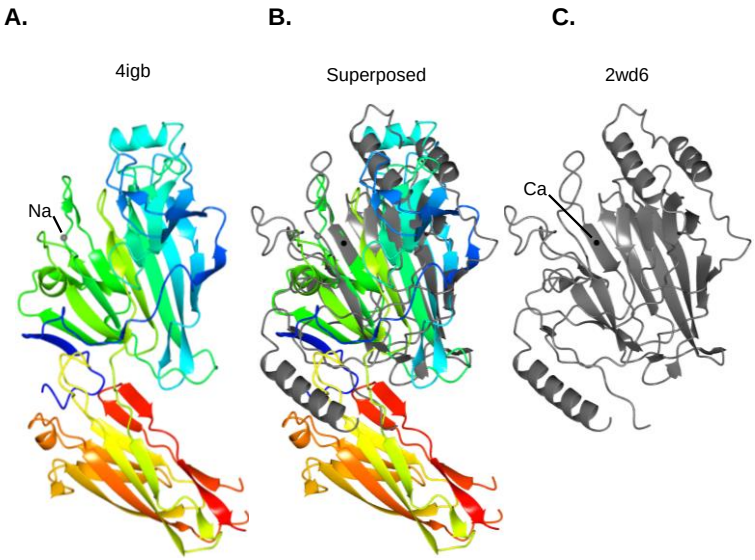


Figure 1.12: A region of SGO0707. A. N1 and N2 from SGO0707 (PDB 4igb)[231], blended from N (blue) to C (red). Sodium ion in putative binding site is indicated. **B.** Superposition of the A region from SGO0707 (blue-red)[231] and the variable domain from SspB (grey)[232]. **C.** Variable domain from SspB (PDB 2wd6)[232] with a calcium ion in the binding site as indicated. Image was created using CCP4mg.

1.7.2 B region

The repeat region of SGO0707 was hypothesised to have a structural role as a stalk for the N-terminal domain[231]. Recently, the crystal structure has been determined of single and tandem domains from different recombinant repetitive structural domains from the B region, composed of SHIRT domains (data courtesy of Dr Fiona Whelan, Dr Clement Degut, Dr James Gilburt, see Figure 5.1). SHIRT domains have a novel fold and domains from the 84 and 88 amino acid region (Figure 1.11) are structurally very similar with a backbone RMSD below 1 Å over 82 residues, indicating that the additional amino acids are located in the linker regions between SHIRT domains (see Figure 5.2).

1.8 Overall aims

CWA proteins from staphylococci and streptococci mediate biofilm formation and accumulation, processes that play important roles in infection. The putatively functional A region is quite variable between different CWA proteins. The B regions, however, from most CWA proteins are thought to have a collective stalk-like function[29,30,57,58], that might be achieved in structurally distinct ways.

For SasC, a low-resolution electron microscopy image of the B region of SasC is available (Figure 1.10A), which shows a kinked, rod-like structure[228]. However, this did not provide conclusive evidence for a stalk-like repetitive region in SasC. In this thesis, we aim to study

the B region of SasC in more detail to determine its suitability to function as a stalk. We aim to provide an atomic-resolution structure, revealing how individual domains within the B region interact with other domains to build up the putatively elongated structure. Using their thermal stability as a tool, we aim to determine if DRESS domains in the repetitive region of SasC stabilise each other, and to what extent. Furthermore, we aim to determine the end-to-end distance, size, shape and rigidity of recombinant repetitive structural domains from the DRESS region of SasC to reveal a potential stalk-like architecture. Finally, previously reported biophysical experiments on (partial) DRESS (DUF1542) domains from another CWA protein suggest an α-helix-rich topology[228], which is typically mechanically weak[233]. However, this prediction is in contrast with the hypothesis that the stalk provides a robust architecture to project the functional A region away from the cell surface. Hence, we aim to characterise for the first time the mechanical strength of recombinant repetitive structural domains from the DRESS region of SasC, to determine whether the stalk-like function fits with its biophysical mechanical properties.

For Sgo0707, the B region containing SHIRT repeats was hypothesised to form a rigid stalk[231]. Recently, in our lab, Dr F. Whelan determined the structure of tandem SHIRT domains, forming elongated, head-to-tail organised, β-sheet rich tandem repeats (Figure 5.1). However, the interface between two SHIRT repeats was open and void of inter-domain stabilising interactions, with exception of a proline-rich linker sequence. Here, we aim to determine if inter-domain interface interactions occur in solution and to what extent the proline-rich linker contributes to the observed rigidity in the X-ray structure.

Thus, this thesis aims to determine if B regions from two CWA-proteins with different structural architectures both achieve a stalk-like shape.

Chapter 2. Materials and methods

2.1 Materials

2.1.1 Bacterial strains

All bacterial strains used in this work are listed in Table 2.1. *Escherichia coli* XL1-blue cells were used for molecular biology and plasmid production. *E. coli* BL21-Gold (DE3) cells were used for the over-production of unlabelled and uniformly ^{15}N- and $^{15}N,^{13}C$- labelled proteins. *S. aureus* strain 8325-4 from the national collection of type cultures (NCTC) was used for the purification of genomic staphylococcal DNA.

Table 2.1: Bacterial strains.

Organism	Strain	Description	Supplier	Ref.
E. coli	XL1-Blue super-competent cells	*recA1 endA1 gyrA46 thi-1 hsdR17 supE44 relA1 lac* [F' *proAB lacI*^{q} *ZΔM15* Tn*10* (Tet^{r})]	Stratagene (Agilent Technologies)	[234]
E. coli	BL21-Gold (DE3)	B F^{-} *ompT hsdS(*$r_B^-m_B^-$*) dcm*$^{+}$ Tet^{r} *gal λ(*DE3*) end*A Hte	Stratagene (Agilent Technologies)	[235]
S. aureus	NCTC 8325-4	*S. aureus* 8325 strain with three prophages, ϕ11, ϕ12, ϕ13, removed by ultraviolet (UV)-treatment.	Generous gift from Prof James Moir, University of York	[236]

2.1.2 Bacterial culture media

All bacterial culture media used in this work are listed in Table 2.2. Media were supplemented with antibiotics (100 µg/mL ampicillin or 50 µg/mL kanamycin).

Table 2.2: Bacterial culture media.

Medium	Application	Description	Ref.
Lysogeny broth (LB)	Molecular biology, protein production	1% *(w/v)* tryptone, 0.5% *(w/v)* yeast extract, 0.17 mM NaCl	[237]
LB-agar	Molecular biology	1% *(w/v)* tryptone, 0.5% *(w/v)* yeast extract, 0.17 mM NaCl, 1.5% *(w/v)* agar	[237]
Super optimal broth with catabolite repression (SOC) media	Molecular biology	2% *(w/v)* tryptone, 0.5% *(w/v)* yeast extract, 2 mM glucose, 10 mM NaCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$	[237]
Auto-induction media	Protein production	1% *(w/v)* tryptone, 0.5% *(w/v)* yeast extract, 0.5% *(w/v)* glycerol, 0.05% *(w/v)* glucose, 0.2% *(w/v)* α-lactose, 0.3% *(w/v)* $(NH_4)_2SO_4$, 0.7% *(w/v)* KH_2PO_4, 0.7% *(w/v)* Na_2PO_4, 1 mM Mg_2SO_4	[238]
Tryptic soy broth (TSB)	*S. aureus* genomic DNA preparation	3% *(w/v)* tryptic soy broth (BD), 0.3% *(w/v)* glucose	[239]
Non-isotope labelled minimal media	Starter culture for isotope-labelled protein production	0.1% *(w/v)* NH_4Cl, 0.4% *(w/v)* glucose, 0.6% *(w/v)* Na_2HPO_4, 0.3% *(w/v)* KH_2PO4, 0.05% *(w/v)* NaCl, 2 mM $MgSO_4$, 0.22 mM $CaCl_2$, 50 µM $FeCl_3$, 10 µM $MnCl_2$, 10 µM $ZnSO_4$, 2 µM $CoCl_2$, 2 µM $CuCl_2$, 2 µM $NiCl_2$, 2 µM Na_2MoO_4, 2 µM Na_2SeO_3, 2 µM H_3BO_3, 1 µg/mL riboflavin, 1 µg/mL nicotinamide, 1 µg/mL pyridoxine, 1 µg/mL thiamine	[238]
^{15}N-M9 minimal media	^{15}N-labelled protein production	0.1% *(w/v)* $^{15}NH_4Cl$, 0.4% *(w/v)* glucose, 0.6% *(w/v)* Na_2HPO_4, 0.3% *(w/v)* KH_2PO4, 0.86 mM NaCl, 2 mM $MgSO_4$, 0.22 mM $CaCl_2$, 50 µM $FeCl_3$, 10 µM $MnCl_2$, 10 µM $ZnSO_4$, 2 µM $CoCl_2$, 2 µM $CuCl_2$, 2 µM $NiCl_2$, 2 µM Na_2MoO_4, 2 µM Na_2SeO_3, 2 µM H_3BO_3, 1 µg/mL riboflavin, 1 µg/mL nicotinamide, 1 µg/mL pyridoxine, 1 µg/mL thiamine	[238]
^{15}N, ^{13}C-M9 minimal media	^{15}N, ^{13}C-labelled protein production	0.1% *(w/v)* $^{15}NH_4Cl$, 0.3% *(w/v)* ^{13}C-glucose, 0.6% *(w/v)* Na_2HPO_4, 0.3% *(w/v)* KH_2PO4, 0.86 mM NaCl, 2 mM $MgSO_4$, 0.22 mM $CaCl_2$, 50 µM $FeCl_3$, 10 µM $MnCl_2$, 10 µM $ZnSO_4$, 2 µM $CoCl_2$, 2 µM $CuCl_2$, 2 µM $NiCl_2$, 2 µM Na_2MoO_4, 2 µM Na_2SeO_3, 2 µM H_3BO_3, 1 µg/mL riboflavin, 1 µg/mL nicotinamide, 1 µg/mL pyridoxine, 1 µg/mL thiamine	[238]

2.1.3 Expression vectors for recombinant proteins

All vectors used in this work are listed in Table 2.3. Variants of the plasmid for transcription by T7 RNA polymerase (pET[240]) were selected to produce recombinant proteins. Plasmids pETFPP1, pETFPP5 and pET-YSBLIC[57] encode the RNA polymerase from bacteriophage T7 with distinct promotor and terminator regions from the expression host *E. coli*. This allows

control of transcription of target RNA and production of target protein[235] (Figure 2.1A). The pBADcLIC2005[12] vector gives tight control over recombinant gene expression and over-production of proteins. In absence of arabinose, transcription is actively inhibited via expression and production of AraC. Recombinant gene expression and protein over-production is induced by addition of L-arabinose, which complexes the inhibitor protein AraC (Figure 2.1B,C)[241–243].

To insert the target genes into the expression vectors, the infusion method[244] was used. In this method, the amplification primers contained 17 (forward) or 18 (reverse) base pairs complementary to the linearised ends of the pET vector and 24-27 base pairs complementary to the target gene. The target gene was amplified by PCR, resulting in inserts featuring a 17-18 base pair non-complementary overhang at the 5' and 3' ends of the insert.

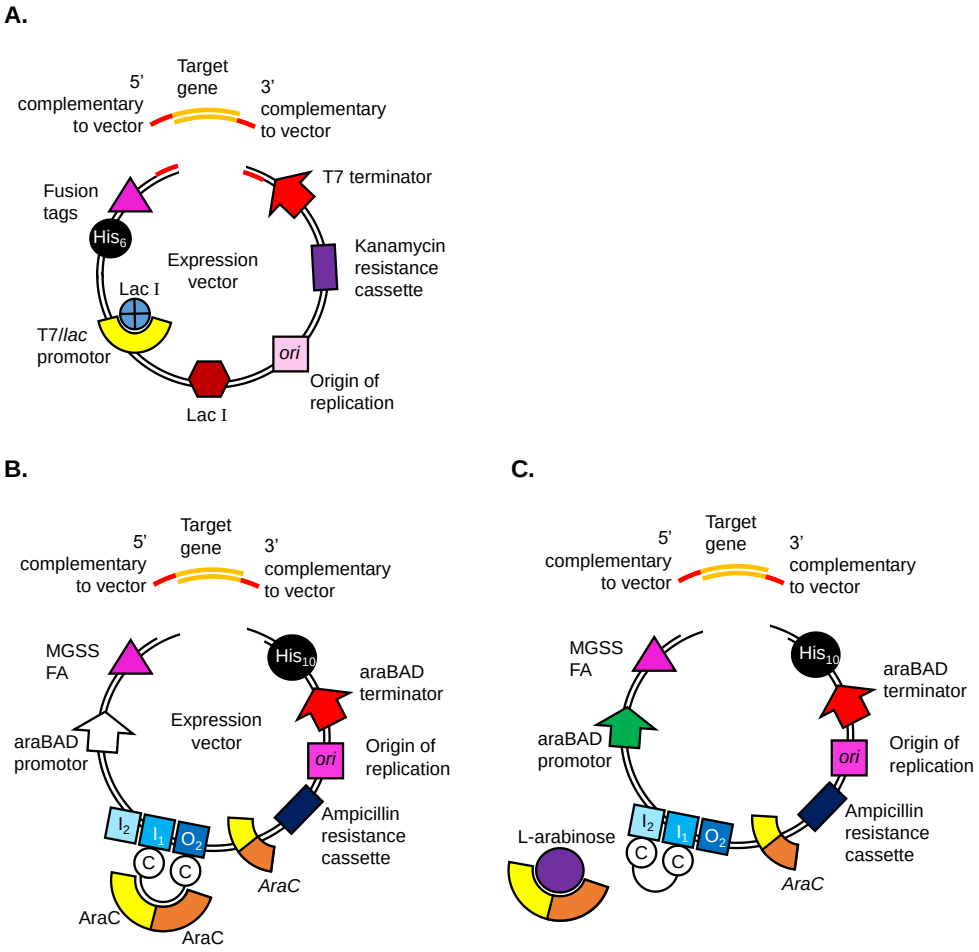


Figure 2.1: Schematics of expression vectors. Components not to scale. **A.** pET expression vector[245]. **B, C.** pBADcLIC2005 expression vector[12]. **B.** Active inhibition of transcription by production of AraC. **C.** Transcription is initiated by the addition of L-arabinose.

Table 2.3: Bacterial expression vectors. HRV: human rhinovirus. Im9: immunity protein 9. IPTG: isopropyl–β– thiogalactopyranoside.

Vector	Resistance	Recombinant gene expression for protein over-production induced by	Description	Supplier
pUC57	Kanamycin	N/A	Used by Genewiz for plasmid vector cloning encoding S0304; S0304_P704A,P706A	Genewiz
pETFPP1	Kanamycin	IPTG; metabolism of α-lactose	Recombinant expression vector featuring an N-terminal His_6 tag cleavable by HRV 3C protease. Used for D1617, D1417, D0710, SHIRT.	Generous gift from YSBL
pETFPP5	Kanamycin	IPTG; metabolism of α-lactose	Recombinant expression vector featuring an N-terminal His_6 tag and Im9 solubility tag cleavable by HRV 3C protease. Used for D17.	Generous gift from YSBL
pET-YSBLIC	Kanamycin	IPTG; metabolism of α-lactose	Recombinant expression vector featuring an N-terminal His_6 tag cleavable by HRV 3C protease and a C-terminal non-cleavable Strep tag followed by two cysteine residues. Used for D0310_scc.	Generous gift from Dr Fiona Whelan
pBADcLIC 2005	Ampicillin	Arabinose	Recombinant expression vector featuring a C-terminal non-cleavable decahistidine (His_{10})-tag. Used for D0118, D0118_2Cys.	Generous gift from Prof Gavin Thomas, University of York

2.2 Methods

2.2.1 Buffer solutions

All buffers and solutions used in the purification of DNA and proteins are listed in Table 2.4.

Table 2.4: Buffer compositions.

Buffers	Function	Composition
RFII-buffer	Molecular biology	10 mM RbCl, 10 mM 3-(N-morpholino)propanesulfonic acid (MOPS), 30 mM $CaCl_2$*2 H_2O, 15% (*w/v*) glycerol
TE-buffer	Molecular biology	10 mM Tris(hydroxymethyl)aminomethane (Tris), 1 mM ethylenediaminetetraacedic acid (EDTA), pH 8.0
Genomic DNA TE-buffer	Molecular biology	10 mM Tris, 0.5 mM EDTA, pH 9.0
TAE-buffer	Molecular biology	40 mM Tris, 20 mM acetic acid, 1 mM EDTA, pH 7.6
5x PCRBIO reaction buffer	PCR	15 mM $MgCl_2$, 5 mM 2'-deoxynucleotide 5'-triphosphate (dNTP), enhancers and stabilisers (PCRBIO)
5x HF buffer	PCR	7.5 mM $MgCl_2$ (New England Biolabs; NEB)
Lysis buffer	Protein purification	20 mM Tris, 150 mM NaCl, 20 mM imidazole, 0.2 mg/mL lysozyme, 0.02 mg/mL deoxyribonuclease (DNAse), pH 7.5
Buffer A	Protein purification	20 mM Tris, 150 mM NaCl, 20 mM imidazole, pH 7.5
Buffer B	Protein purification	20 mM Tris, 150 mM NaCl, 500 mM imidazole, pH 7.5
Strep binding buffer	Protein purification	20 mM Tris, 150 mM NaCl, 5 mM β-mercaptoethanol, pH 7.5
Strep elution buffer	Protein purification	20 mM Tris, 150 mM NaCl, 5 mM β-mercaptoethanol, 2.5 mM desthiobiotin, pH 7.5
SEC buffer	Protein purification	20 mM Tris, 150 mM NaCl, pH 7.5
Tris-Glycine running buffer (1x)	Biochemical methods	3% (*w/v*) Tris, 14% (*w/v*) glycine, 1% (*w/v*) sodium dodecyl sulfate (SDS)
SDS PAGE sample loading buffer (1x)	Biochemical methods	100 mM 1,4-dithiothreitol (DTT), 50 mM Tris, 20% (*w/v*) glycerol, 2% (*w/v*) SDS, 0.1% (*w/v*) bromophenol blue
Coomassie Brilliant Blue R dye solution	Biochemical methods	0.25% (*w/v*) Coomassie Brilliant blue R, 45% (*v/v*) ethanol, 10% (*v/v*) acetic acid

2.2.2 Preparation of chemically competent cells

BL21-Gold (DE3) competent cells were prepared as follows: cells were grown in 200 mL LB (37 °C, 220 rotations per minute (rpm)) until an optical density measured at 600 nm (OD_{600}) of 0.5 was reached. Cells were cooled on ice for 15 min, before harvesting by centrifugation at 4000 g (4 °C, 15 min). Cells were then resuspended in 60 mL ice cold 75 mM $CaCl_2$ and incubated on ice for 1 hour. After harvesting at 4000 g (4 °C, 15 min), cells were gently

resuspended in 7.5 mL RFII-buffer (Table 2.4) and incubated for 30 min. 200 µL aliquots were vitrified in liquid nitrogen and stored at -80 °C.

2.2.3 Transformation of competent cells

Cryo-preserved competent cells were thawed on ice, mixed gently with 100 ng plasmid DNA and incubated on ice for 30 min. Cells were 'heat-shocked' (42 °C, 45 s) and then placed on ice for 15 min. 120 µL SOC-medium (Table 2.2) was added and cells were incubated for 1.5 hours (37 °C, 220 rpm), before plating out on a selective LB-agar medium (see Table 2.2) and incubated at 37 °C for 18 hours.

2.2.4 Preparation of plasmid DNA

7 mL cultures of freshly transformed E. coli XL1-Blue cells were grown for 18 hours in selective LB-media (37 °C, 220 rpm). Cells were harvested by centrifugation at 4000 g (4 °C, 15 min) and plasmid DNA was isolated using the Macharey-Nagel Nucleospin Plasmid kit (Macharey-Nagel) according to the manufacturer's instructions. Briefly, precipitated proteins and genomic DNA were removed by centrifugation, while plasmid DNA was intact. Plasmid DNA was bound to a silica membrane and washed, before the pure plasmid DNA was eluted at low ionic strength with TE-buffer (Table 2.4). The DNA concentration was determined by measuring the absorbance at 260 nm. The purity of DNA was assessed by measuring the absorbance ratio A_{260}/A_{280}; DNA with a ratio of 1.6-2.0 was diluted to 100 ng/µL with MilliQ (MQ) water, before storing at -20 °C.

2.2.5 Preparation of genomic DNA from *S. aureus*

Handling of class II pathogens was performed in a class II microbiological safety cabinet only. *S. aureus* NCTC 8325-4 was plated out on non-selective LB-agar and incubated at 37 °C for 18 hours, forming small yellow colonies. 5 mL cultures of *S. aureus* NCTC 8325-4 were grown for 24 hours in non-selective TSB-media (37 °C, 220 rpm, Table 2.4). Genomic DNA was isolated using the GenElute Bacterial Genomic DNA purification kit (Sigma-Aldrich) according to the manufacturer's instructions. Briefly, cells were pelleted at 4000 g (4 °C, 15 min, Table 2.4), lysed by lysostaphin-lysozyme treatment (37 °C, 60 min) and nucleases were degraded by 0.19% (*w/v*) proteinase K treatment (55 °C, 10 min). Genomic DNA was bound to a silica membrane and washing steps were performed, before DNA was eluted with Genomic DNA TE-buffer (Table 2.4). DNA purity was assessed by measuring the

absorbance ratio A_{260}/A_{280}; stocks with a ratio of 1.6-1.9 were aliquoted and stored at -20 °C.

2.2.6 DNA insert preparation by Polymerase chain reaction (PCR)

The composition of a typical PCR is reported in Table 2.5. Primers were designed manually with a 17 base pair overlap with the target vector (Table 7.1) and ordered from Eurofins Genomics with salt-free purity. PCRs were performed on a Bio-Rad T100 Thermal Cycler. A typical cycling program is reported in Table 2.6. The annealing temperature was determined based on the approximate melting temperature (T_m) of the primer set, calculated by www.eurofinsgenomics.eu.

Table 2.5: PCR composition.

Component	Final concentration	Supplier
5x PCRBIO HiFi reaction buffer	1x	PCRBio
dNTPs	0.2 mM	Fermentas
Template DNA	Plasmid DNA: 0.2 ng/µL Genomic DNA: 1 ng/µL	
Forward primer	0.5 µM	Eurofins Genomics
Reverse primer	0.5 µM	Eurofins Genomics
PCRBIO HiFi polymerase	0.02 U/µL	PCRBio
MQ		

Table 2.6: PCR cycling program.

Cycle step	Temperature (°C)	Time	Number of cycles
Initial denaturation	95	5 min	1
Denaturation	95	30 s	35
Annealing	Varies per primer	30 s	
Extension	72	30 s/kb	
Final extension	72	10 min	1
Hold	4	∞	1

2.2.7 Agarose gel electrophoresis

1% (*w/v*) agarose gels were prepared in TAE-buffer (Table 2.4) supplemented with SYBRsafe DNA Gel Stain (Invitrogen). DNA was mixed with 6x orange loading dye (Thermo Scientific) according to the manufacturer's instructions and loaded into the gel. GeneRuler 1 kb Plus DNA ladder (Thermo Scientific) was loaded in a separate well. Separation was achieved by electrophoresis using a Bio-Rad PowerPac Basic (100 V, 60 min) in TAE-buffer. DNA was visualised by transillumination of the gels with UV light.

2.2.8 Vector linearisation by PCR

Linear vectors were prepared by PCR using linearisation primers (Appendix 7.1, Table 7.3). Typical vector linearisation compositions are reported in Table 2.5 and a typical cycling program is reported in Table 2.6. PCR products were mixed with 10x CutSmart buffer (NEB) and treated with a restriction enzyme that digests methylated DNA (DpnI; NEB) to digest the remaining template (37 °C, 2 hrs), before DnpI was inactivated (80 °C, 20 min). The reaction was mixed with 6x orange loading dye (Thermo Scientific) according to the manufacturer's instructions and analysed by agarose gel electrophoresis. Bands containing linear vector were excised and DNA was extracted using a Nucleospin gel and PCR clean-up kit (Macharey-Nagel), according to the manufacturer's instructions. Briefly, the gel was solubilised in binding buffer (50 °C) containing chaotropic salts. DNA was bound to a silica membrane and washed. Pure DNA was eluted at low ionic strength with TE-buffer (Table 2.4); DNA with a purity ratio of 1.6-2.0 was aliquoted and stored at -20 °C.

2.2.9 DNA insertion into linear vector

In an In-Fusion® reaction, the Vaccinia virus DNA polymerase with 3'-5' exonuclease activity degrades the ends of the amplified insert, which are complementary to the linearised ends of the infusion vector. Strand annealing followed by transfection of this product into *E. coli* XL1 Blue (see Table 2.1) results in DNA strand break repair to form a stable plasmid. Typically, over 50-fold of transformants contain the insert over the empty vector[244].

PCR products with overhanging bases on either end complementary to the termini of the linear vector were inserted into the linear vector using the In-Fusion® kit (Takara Bio) according to the manufacturer's instructions. Briefly, a typical reaction mix was prepared

as in Table 2.7. The insertion was performed (37 °C, 15 min), followed by inactivation of the enzyme (50 °C, 15 min). The reaction was diluted with 40 µL TE-buffer (Table 2.4) and the resulting mix was used directly for transformation.

Table 2.7: In-Fusion reaction composition.

Component	Volume	Supplier
MQ	6 µL	
In-Fusion® Enzyme premix	2 µL	Takara
Insert	1 µL (~50 ng)[244]	
Linear vector	1 µL (~20 ng)[244]	

2.2.10 Site-directed mutagenesis

Whole plasmid mutagenesis was performed to introduce base pair mutations into existing DNA constructs. The method was adapted from the QuickChange site-directed mutagenesis protocol and the primer design was as reported by Zheng *et al*. (2004)[246] and Liu & Naismith (2008)[247]. Briefly, the primers both featured the mutated region, but were offset with respect to each other, ensuring good annealing to the template DNA and minimising primer-primer hybridisation. Typically, the primers overlapped by 15 bases and had a further 25 non-overlapping bases to either side of the mutation. Primers (Table 7.2) were designed manually and ordered from Eurofins with High Purity Salt Free (HPSF) purity. A typical site-directed mutagenesis composition is reported in Table 2.8, a mutagenesis cycling program is reported in Table 2.9. 5% (*v/v*) dimethyl sulfoxide (DMSO) was used as an additive when DNA-amplification was poor. The annealing temperature gradient was guided by the T_m of the overlapped primer region.

As the ratio of template DNA (without the mutation) to amplified DNA (with the mutation) was low, no DnpI digestion step was performed. Instead, the full product of the cycling program was analysed by agarose gel electrophoresis. The amplified band was excised and DNA was extracted using a Nucleospin gel and PCR clean-up kit (Macharey-Nagel), according to the manufacturer's instruction.

Table 2.8: Site-directed mutagenesis composition.

Component	Final concentration	Supplier
5x HF reaction buffer	1x	NEB
dNTPs	0.2 mM	Fermentas
DMSO	0-5%	NEB
Template DNA	0.4 ng/µL	
Forward primer	0.3 µM	Eurofins Genomics
Reverse primer	0.3 µM	Eurofins Genomics
Phusion polymerase	0.02 U/µL	NEB
MQ		

Table 2.9: Whole plasmid site-directed mutagenesis cycling program.

Cycle step	Temperature (°C)	Time	Number of cycles
Initial denaturation	95	5 min	1
Denaturation	95	30 s	19
Annealing	60-50 gradient	30 s	
Extension	72	30 s/kb	
Final extension	72	10 min	1
Hold	4	∞	1

2.2.11 Construct validation

The correct composition of DNA inserts into the recombinant expression vectors from Table 2.3 generated in this work was verified by DNA sequencing, performed by Eurofins Genomics.

2.3 Recombinant gene expression, protein production and purification

2.3.1 Over-production of unlabelled proteins

50 mL selective LB-media were inoculated with a single colony of BL21-Gold (DE3) cells, which were freshly transformed with plasmid DNA encoding the protein of interest or were plated out from a glycerol stock, and incubated at 37 °C (220 rpm, 18 hours). Selective LB-media or auto-induction media were inoculated to an OD_{600} of 0.05 with the starter culture and incubated at 37 °C (120 rpm) until an OD_{600} was reached of 0.6. A 1 mL cell pellet was stored at -20 °C as a pre-induction reference. When using LB, isopropyl–β–thiogalactopyranoside (IPTG, Melford) or L-arabinose (Sigma-Aldrich) was added to a final concentration of 1 mM or 0.1% (*w/v*), respectively, to induce recombinant gene expression of the protein of interest, depending on the nature of the recombinant expression vector (Table 2.3). When using auto-induction media, induction of recombinant gene expression of the protein of interest starts upon exhaustion of glucose and the start of metabolic consumption of α-lactose. When an OD_{600} of 0.6 was reached, cultures were transferred to the desired temperature pre-determined from a test for optimal protein over-production conditions (typically 20 °C) and incubated for 18 hours at 120-180 rpm. Cells were pelleted by centrifugation at 6240 g (4 °C, 20 min) and stored at -20 °C. DNA and protein sequences used in this thesis are reported in section 7.8.

2.3.2 Over-production of ^{15}N and ^{15}N, ^{13}C-uniformly labelled recombinant proteins

^{15}N- or ^{15}N, ^{13}C-uniformly labelled proteins were expressed in ^{15}N-M9 or ^{15}N, ^{13}C-M9 minimal media (Table 2.2) as described in section 2.3.1. ^{15}N-labelled ammonium chloride and ^{13}C-labelled D-glucose were obtained from Cambridge Isotope Laboratories at 99% purity. Briefly, 50 mL of selective non-isotope labelled minimal medium was inoculated with a single colony of freshly transformed BL21-Gold (DE3) cells and incubated at 37 °C (220 rpm, 18 hours). Selective isotope-labelled medium was inoculated to an OD_{600} of 0.1 with starter culture to compensate for the slower cell growth in minimal media and grown (37 °C, 120 rpm) to an OD_{600} of 0.6. Recombinant gene expression and production of the target protein was induced by the addition of IPTG to a final concentration of 1 mM and cultures were incubated at 20 or 30 °C as pre-determined in a test for optimal protein over-

production conditions (18 hours, 180 rpm). Cells were pelleted by centrifugation at 6240 g (4 °C, 20 min) and stored at -20 °C. DNA and protein sequences used in this thesis are reported in section 7.8.

2.3.3 Cell lysis

Pellets stored at -20 °C were thawed on ice and resuspended in 30 mL lysis buffer (Table 2.4) per litre medium. Cells were lysed by sonication using a Sonicator 3000 (Misonix). Resuspended cells were cooled on ice prior to and during sonication. A standard sonication cycle was repeated twice and involved 60 pulses at 70 W of 3 s with 7 s recovery intervals per cycle. The soluble extract was separated from the insoluble cell debris by centrifugation at 48000 g (4 °C, 45 min).

2.3.4 Purification of His-tagged proteins

After sonication (see section 2.3.3), the soluble extract was loaded onto two HisTrap HP column (5 mL, GE Healthcare Life Sciences) at 3 mL/min on an Äkta Prime Plus (GE Healthcare Life Sciences), which was equilibrated in buffer A (Table 2.4). The flow-through of the loading step was collected and stored at 4 °C. The column was washed with buffer A, until non-bound proteins were washed off as judged from the return of the A_{280} trace to its baseline value. Bound protein was eluted using an increasing gradient of the concentration of imidazole from 20 to 500 mM (0%-100%) over 200 mL and collected into 4 mL fractions. Elution fractions were analysed by SDS poly-acrylamide gel electrophoresis (SDS PAGE; see section 2.4.1) and fractions of similar purity containing a band at approximately the expected molecular weight (MW) were pooled. Imidazole was removed by a dialysis step into 20 mM Tris, 150 mM NaCl, pH 7.5 using Spectra/Por dialysis membrane (SpectrumLabs) with a molecular weight cut-off (MWCO) of ≤ ¼ the MW of the protein of interest (4 °C, 16 hours, 5 L dialysis buffer). An approximate target protein yield was calculated using the Beer-Lambert relationship (Equation 2.1).

2.3.5 Removal of affinity tag

The N-terminal sequence of hexahistidine (His_6)-tagged proteins from the pETFPP vector family (see section 2.1.3) up to the target protein was as follows: Met – Gly – Ser – Ser – His – His – His – His – His – His – Ser – Ser – Gly – Leu – Glu – Val – Leu – Phe – Gln – Gly – Pro – Ala – Met –. The human rhinovirus (HRV) 3C protease is a cysteine protease that

recognises the underlined part of this linker sequence and cleaves between Gln and Gly, removing the His_6 tag from the target protein. Recombinant HRV 3C protease (Bioscience Technology Facility, University of York) was tagged with a non-cleavable His_6-maltose binding protein (MBP) affinity tag for purification purposes and had an approximate MW of 63 kDa.

Cleavage of a His_6-tagged protein target was carried out using a typical protease-to-target ratio of 1:150 (*w/w)* in 20 mM Tris, 150 mM NaCl, pH 7.5, 1 mM DTT at 4 °C for 18 hours. Completion of cleavage was assessed by SDS PAGE.

Separation of the His_6-affinity tag and HRV 3C protease from the protein of interest was achieved by passage over a HisTrap HP column (5 mL, GE Healthcare Life Sciences) at 3 mL/min, which was equilibrated in buffer A (Table 2.4). The protein of interest passed through the column into the flow-through, while the affinity tag and HRV 3C protease both contained a His_6-tag and bound to the column. Imidazole was removed from the flow-through containing the target protein by dialysis into 20 mM Tris, 150 mM NaCl, pH 7.5 using Spectra/Por dialysis membrane (SpectrumLabs) with a MWCO of ≤ ¼ the MW of the protein of interest (4 °C, 16 hours, 5 L dialysis buffer). The target protein was concentrated using a VivaSpin 20 concentrator (Sartorius) with a MWCO of ≤ ¼ the MW of the protein of interest at 4000 g (4 °C, 15-30 min per run).

2.3.6 Purification of proteins for AFM

Proteins for AFM were cloned into a modified pET28 vector, pET-YSBLIC (Table 2.3) with cleavable N-terminal His_6-tag and a non-cleavable C-terminal Strep-tag[248] (see later), followed by two C-terminal cysteine residues for immobilisation on a gold-coated glass square. This approach was suitable, because the protein used for AFM does not feature cysteine residues, as is common in extracellular Gram-positive proteins[249]. Typically, purification was performed as described in section 2.3.4 with buffer conditions supplemented with 5 mM β-mercaptoethanol (Sigma-Aldrich) during purification and 2 mM tris(2-carboxyethyl) phosphine hydrochloride (TCEP) during storage.

Briefly, lysis buffer was supplemented with 5 mM β-mercaptoethanol. The soluble extract containing His-tagged proteins was loaded onto a HisTrap HP column equilibrated in buffer A supplemented with 5 mM β-mercaptoethanol. The column was washed with buffer A

supplemented with 5 mM β-mercaptoethanol, until a stable baseline A_{280} trace was reached. His-tagged proteins were eluted in an increasing imidazole gradient of buffers A and B over 200 mL supplemented with 5 mM β-mercaptoethanol. Imidazole was removed by a dialysis step into 20 mM Tris, 150 mM NaCl, 5 mM β-mercaptoethanol, pH 7.5 using Spectra/Por dialysis membrane (SpectrumLabs) (4 °C, 16 hours, 2 L dialysis buffer). If appropriate, the His_6-tag was cleaved using HRV 3C protease (see section 2.3.5) in a typical protease-to-target ratio of 1:150 (*w/w)* in 20 mM Tris, 150 mM NaCl, pH 7.5, 1 mM DTT at 4 °C for 18 hours. Completion of cleavage was assessed by SDS PAGE.

Strep-tag affinity chromatography was performed to further purify the target protein, if required. The interaction of streptavidin with biotin is a well-known non-covalent high-affinity interaction[250]. Binding affinity towards streptavidin[251] was exploited for the development of a one-step affinity purification step based on the Strep-tag[248], of which the current optimised sequence is: LEVFQGP. StrepTrap HP columns (GE Healthcare Life Sciences) contain high affinity binding sites for the Strep-tag, which can be competitively eluted using desthiobiotin. The resin was regenerated by washing with 0.5 M NaOH. To keep the cysteine residues reduced, this purification was performed in the presence of 5 mM β-mercaptoethanol, however this is not required for other Strep-tagged protein purifications.

~20 mg Strep-tagged protein was loaded onto 2 StrepTrap columns at 0.2 mL/min, equilibrated in Strep binding buffer supplemented with 5 mM β-mercaptoethanol (Table 2.4). When the baseline A_{280} signal was reached, Strep-tagged proteins were competitively eluted by Strep elution buffer supplemented with 5 mM β-mercaptoethanol (Table 2.4) in gravity flow into 1.5 mL fractions in 2 column volumes (cv). Elution fractions were analysed by SDS PAGE and fractions of similar purity were pooled. Protein was concentrated by centrifugation using a VivaSpin 20 concentrator (Sartorius) with a MWCO of ≤ ¼ the MW of the protein of interest at 4000 g (4 °C, 30 min per run). Desthiobiotin was removed by dialysis (4 °C, 16 hours, 0.5 L dialysis buffer) into the final storage conditions 20 mM Tris, 150 mM NaCl, 2 mM TCEP, pH 7.5 or 25 mM MES, 150 mM NaCl, 2 mM TCEP, pH 6.0.

2.3.7 Preparation of proteins for SHRImP

2.3.7.1 Protein purification

The end-to-end distance of a multi-domain protein was assessed by measuring the inter-fluorophore distance between fluorophores bridging sixteen DRESS domains. Cysteine residues were required for the coupling of fluorophores to proteins and were introduced by site-directed mutagenesis (see section 2.2.10) into a protein construct cloned in a pBADcLic2005 expression vector (Table 2.3) with a non-cleavable C-terminal His_{10}-tag.

During purification and storage, the protein was kept in reducing conditions. Typically, purification was performed as described in section 2.3.4 with adjusted buffer conditions. Briefly, lysis buffer was supplemented with 5 mM β-mercaptoethanol. The soluble extract containing His-tagged proteins was loaded onto a HisTrap HP column (GE Healthcare Life Sciences) equilibrated in buffer A supplemented with 5 mM β-mercaptoethanol. The column was washed with buffer A supplemented with 5 mM β-mercaptoethanol, until a stable baseline A_{280} trace was reached. His-tagged proteins were eluted in an increasing gradient of the concentration of imidazole of buffers A and B over 200 mL supplemented with 5 mM β-mercaptoethanol. Imidazole was removed by a dialysis step into 20 mM Tris, 150 mM NaCl, 5 mM β-mercaptoethanol, 1 mM EDTA, pH 7.5 using Spectra/Por dialysis membrane (SpectrumLabs) (4°C, 16 hours, 5 L dialysis buffer). Size exclusion chromatography (SEC, see section 2.3.8) was employed in SEC-buffer (Table 2.4) supplemented with 5 mM β-mercaptoethanol and 1 mM EDTA, followed by concentration by centrifugation using a VivaSpin 20 concentrator (Sartorius) with a MWCO of ≤ ¼ the MW of the protein of interest at 4000 g (4 °C, 30 min per run). To prepare the protein for maleimide-coupling to a fluorophore, β-mercaptoethanol was removed through dialysis into 20 mM Tris, 150 mM NaCl, 1 mM EDTA, pH 7.0 (4°C, 2 hours, 2 L) in a Slide-A-Lyzer MIDI dialysis device (MWCO 3.5 kDa, ThermoFisher) and subsequently, the protein was dialysed into 20 mM Tris, 150 mM NaCl, 1 mM EDTA, 2 mM TCEP, pH 7.0 (4°C, 16 hours, 0.5 L).

2.3.7.2 Introduction of fluorophores

A 25-molar excess of Alexa-Fluor 488 C_5 maleimide fluorophore (final concentration; ThermoFisher; 10 mM in DMSO) to cysteine was added in three steps to protein with free cysteine residues in 20 mM Tris, 150 mM NaCl, 1 mM EDTA, 2 mM TCEP, pH 7.0. The

reaction was performed at 20 °C in the dark for 2 hours. The reaction was quenched by adding a 10-fold molar excess of DTT relative to the fluorophore. Unreacted fluorophore was removed by dialysis into 20 mM Tris, 150 mM NaCl, pH 7.0 (4°C, 16 hours, 5 L) in a Slide-A-Lyzer MIDI dialysis device (MWCO 3.5 kDa, ThermoFisher). Labelled protein was separated from any aggregates, unlabelled protein and unreacted fluorophore by SEC (see section 2.3.8) using a Superdex 200 100/300 GL (volume 24 mL, GE Healthcare Life Sciences) equilibrated in 20 mM Tris, 150 mM NaCl, 1 mM DTT, pH 7.0.

2.3.8 Size exclusion chromatography (SEC)

SEC is an affinity tag-free method of protein purification, which separates proteins by size. Larger proteins, such as aggregates, cannot enter into the pores of the matrix, which is a composite of cross-linked agarose and dextran[252] and elute in a smaller elution volume. Smaller proteins can enter these pores, leading to a longer retention time on the column and a larger elution volume. SEC was used to determine if the protein sample contained species of different oligomeric state and if required, as an additional purification step. Typically, a Superdex 75 16/600 column (volume 120 mL, GE Healthcare Life Sciences) was used for proteins with MW < 50 kDa and Superdex 200 26/600 column (volume 320 mL, GE Healthcare Life Sciences) for proteins with MW > 50 kDa. When the purpose of the SEC run was analytical, a column with a smaller volume was used.

A suitable column was equilibrated in SEC buffer on a liquid chromatography system (Äkta Purifyer Box-900) equipped with pH and conductivity measurement cell (pH/C-900), UV-detector (UV-900) and pumping system (P-900) (Amersham Biosciences). The protein was injected onto the column, followed by 1.1 cv of buffer and 1.5-4 mL fractions were collected. The elution was monitored using A_{280} (tryptophan and tyrosine residues) or A_{493} (Alexa Fluor 488 dye). Elution fractions were analysed by SDS PAGE and fractions of similar purity within the same elution peak were pooled, followed by protein concentration using a VivaSpin concentrator 6 or 20 (Sartorius) with a MWCO of ≤ ¼ the MW of the protein of interest at 4000 g (4 °C, 30 min per run). If required, the oligomeric state was determined by SEC-MALLS (see section 2.6.1).

2.3.9 Validation of protein purity and molecular mass

Protein purity was assessed by SDS PAGE (see section 2.4.1). The MW of purified proteins was determined by electrospray ionisation mass spectrometry (ESI MS), performed by the Molecular Interactions laboratory or the Metabolomics and Proteomics laboratory (Bioscience Technology Facility, Department of Biology, University of York) in 2 mM Tris pH 8.0 or 25 mM ammonium acetate pH 6.5.

2.4 Biochemical methods

2.4.1 SDS PAGE

2.4.1.1 Preparation of SDS PAGE gels

Gels for SDS PAGE analysis were prepared manually to the desired percentage of acrylamide, which determines the resolution of protein bands with different MW (Table 2.10). The resolving part of an SDS PAGE gel had the following composition: the desired concentration acrylamide (see Table 2.10; from 30% (*w/v*) of acrylamide, 0.8% (*w/*v) bis-acrylamide, National Diagnostics), 0.37 M Tris pH 8.8, 0.1% (*w/v*) SDS, 0.1% (*w/v*) ammonium persulfate (APS, Acros Organics) and 0.1% (*w/v*) N,N,N',N'-tetramethylethylenediamine (TEMED, Sigma). The stacking part of an SDS PAGE gel had as composition: 4% (*w/v*) acrylamide, 0.13 M Tris pH 6.8, 0.1% (*w/v*) SDS, 0.1% (*w/v*) APS and 0.1% (*w/v*) TEMED.

Table 2.10: Resolving range of SDS PAGE gels.

%SDS (*w/v*)	Resolving range (MW)
15	5-30
10	30-100
8	~100-200

2.4.1.2 Sample preparation and electrophoresis procedure

Protein samples for analysis by SDS PAGE were prepared as follows. To ~1 mg/mL protein, sample loading buffer was added (Table 2.4), samples were denatured (95 °C, 5-15 min depending on sample volume) and centrifuged (13000 g, 1 min, 20 °C) prior to loading 10 µL per lane (~7 µg protein). Typically, 6 µL Precision Plus Protein Standard MW marker (Bio-Rad), containing 10 protein bands over the range of 10 to 250 kDa, was used to estimate

the MW of protein bands. Electrophoresis was performed in Tris-Glycine running buffer (Table 2.4) in a Bio-Rad PowerPac Basic at 200 V for approximately 40-60 min. Proteins were visualised by staining with Coomassie Brilliant Blue R dye solution (Table 2.4, 30 min, shaking), followed by destaining (10% (*v/v*) acetic acid, 10% (*v/v*) ethanol) for 2 hours before gels were imaged. Throughout this thesis, SDS PAGE gel analyses are aided by arrows indicating protein bands and a theoretical MW is provided for the proteins putatively represented by these arrows.

2.4.2 Determination of protein concentration

Protein concentration was estimated by measuring the absorbance of a protein solution at 280 nm with the baseline absorbance determined from buffer only. The protein concentration was calculated according the Beer-Lambert law (Equation 2.1):

Equation 2.1: Beer-Lambert law

$$c = \frac{A}{\varepsilon l}$$

where c is the protein concentration in mol/L, A is the absorbance at 280 nm, ε is the molar extinction coefficient in L mol^{-1} cm^{-1} and l is the path length of the absorbance measurement in cm. ε was determined from the protein sequence using the Expert Protein Analysis System (ExPASy) Bioinformatics Resource Portal tool ProtParam (Equation 2.2)[253]:

Equation 2.2: Molar extinction coefficient ε

$$\varepsilon = N_{Tyr}\varepsilon_{Tyr} + N_{Trp}\varepsilon_{Trp} + N_{Cys}\varepsilon_{Cys}$$

where N_{Tyr}, N_{trp}, N_{Cys} is the total number of tyrosine, tryptophan and disulfide-bonded cysteine residues and ε_{Tyr}, ε_{Trp}, ε_{cys} is the average absorption value of Trp (5500 L mol^{-1} cm^{-1}), Tyr (1490 L mol^{-1} cm^{-1}) and disulfide-bonded cysteine residues (125 L mol^{-1} cm^{-1}) as reported in Pace *et al*. (1995)[254]. Any cysteine residues present in proteins in this work were assumed to be not disulfide-bonded and as such do not contribute to the ε of the protein. Most proteins in this work had few tyrosine and tryptophan residues, which may cause >10% error on concentration estimations. Therefore, absorption measurements were performed in triplicate and the average value was used for protein concentration estimation.

2.5 Bioinformatics methods

2.5.1 Secondary structure predictions

The protein secondary structure prediction servers Jpred4[255] and PSI-PRED[256] were used to analyse the predicted secondary structure in SasC. Briefly, a PSI-BLAST[257] multiple sequence alignment is created based on the single protein sequence provided. This sequence alignment is then used as input for trained neural networks, which have been trained on 480 protein sequences and protein structures from non-redundant protein families[258]. The prediction servers provide a predicted secondary structure and a confidence score.

2.5.2 Sequence alignments

Multiple sequence alignments (MSAs) of domains from the same family were aligned using Clustal Omega[259,260]. Low-homology sequence alignments of domains from the same superfamily as defined by PFam were aligned using Multiple Alignment using Fast Fourier Transform (MAFFT)[260,261] and the Jones, Taylor and Thornton accepted point mutation scoring matrix (JTT PAM) 100[262]. 211 sequences from the seed sequences of the GA-module, sequences of GA-modules from Uniprot entry PF01468, all sequences of the GA-like and SpA domains from PFam[185,188] and DRESS domain sequences from SasC were aligned. The alignment was visualised in Jalview[263] and sequences containing inserts occurring in only 1-3 sequences were removed from the alignment, resulting in 206 aligned sequences. Part of this alignment is shown (Figure 3.3).

2.6 Biophysical methods

2.6.1 Size exclusion chromatography with multi-angle laser light scattering (SEC-MALLS)

2.6.1.1 Theory

SEC-MALLS enables the determination of molecular mass and size of molecules in solution by light scattering. Molecules are first analysed by SEC (see section 2.3.8), separating molecules based on size. Following SEC, the elution is analysed by multiple detectors, connected in series, such as for MALLS, quasi-elastic light scattering (QELS), UV light absorbance and the refractive index.

MALLS detectors measure the excess Rayleigh ratio R_θ, which is defined as the excess absolute scattering of light from a dilute solution of molecules compared to the absolute scattering of pure solvent, divided by the intensity of incident light. At each angle, the light scattering is proportional to the MW and the concentration (c) of molecules in solution. When molecules are smaller than the wavelength of incident light, the scattering intensity does not depend on the scattering angle. For larger molecules, the angular dependence of the scattering intensity depends on the radius of gyration, R_g, of the molecules. The angular dependence together with the MW can be used to obtain information about particle shape in solution[264].

The Rayleigh-Debye-Gans equation (Equation 2.3) describes the relationship between the Rayleigh ratio, R_θ, and the angular dependence of the scattering[265]:

Equation 2.3: A. Rayleigh-Debye-Gans light scattering[265]. B. The angular dependence of scattered light.

A.

$$\frac{K'c}{R(\vartheta)} = \frac{1}{MW \cdot P(\theta)} + 2A_2c$$

B.

$$P(\theta) = 1 - \frac{1}{3}\left(\frac{4\pi n}{\lambda_0}\right)^2 \sin\left(\frac{\theta}{2}\right)^2 < r_g^2 >$$

where K' is a physical optical constant equal to $\frac{4\pi^2 n^2 \left(\frac{dn}{dc}\right)^2}{\lambda_0^4 N_A}$, n is the refractive index of the solvent, $\frac{dn}{dc}$ is the refractive index increment, λ_0 is the wavelength of incident light in vacuum, N_A is the number of Avogadro, MW is the average MW of the molecule, $P(\theta)$ is the angular dependence of the scattered light related to the R_g independent of the shape of the molecule, $< r_g^2 >$ is the mean square of the R_g, A_2 is the second virial coefficient assumed zero for low concentration and c is the concentration in g/mL. Equation 2.4, a reciprocal form of Equation 2.3, is linear with respect to $\sin\left(\frac{\theta}{2}\right)^2$ and allows for the calculation of MW and $<r_g^2>$. For simplification, only the second virial coefficient is included as this is sufficient expansion for $P(\theta)$ to accurately describe dilute molecules with r_g < 30 nm[264].

Equation 2.4: Inverse Zimm plot for MALLS[264].

$$\frac{R(\theta)}{K'c} = MW\left[1 - \frac{1}{3}\left(\frac{4\pi n}{\lambda_0}\right)^2 \sin\left(\frac{\theta}{2}\right)^2 < r_g^2 >\right] + \frac{1}{2A_2c}$$

Equation 2.4 can be plotted as $\frac{R(\theta)}{K'c}$ against $\sin\left(\frac{\theta}{2}\right)^2$ and allows for the determination of MW from the intercept with the y axis and $< r_g^2 >$ from the slope extrapolated at zero θ.

QELS detects fluctuations of scattered light. The speed of fluctuating particles depends on their diffusion coefficient and inherently, their size in solution. The decay of the autocorrelation function is related to the translational diffusion coefficient[266]. The radius of hydration, R_h, is defined as the equivalent radius of a sphere which diffuses at the same diffusion coefficient as the measured molecule[267] and is obtained from Equation 2.5.

Equation 2.5: Calculation of the R_h by QELS[267].

$$R_h = \frac{k_B T}{6\pi\eta D_t}$$

where k_B is the Boltzmann constant in J/K, T is the temperature in K, η is the viscosity of the solvent in Pa·s and D_t is the translational diffusion coefficient in m^2/s. The change of the solution refractive index with respect to a change in concentration of molecular species is measured by the differential refractive index (DRI) detector (Equation 2.6):

Equation 2.6: Calculation of protein concentration from DRI[264].

$$\Delta c = \frac{n_s - n_r}{\frac{dn}{dc}}$$

where n_s is the refractive index of the dilute solution containing molecules, n_r is the refractive index of pure solvent and $\frac{dn}{dc}$ is the refractive index increment as a function of concentration of molecules[264]. The similar nature of proteins allows for a known $\frac{dn}{dc}$ value for dilute proteins in aqueous buffer of 0.186 mL/g[268]. Here, the experimental value of $\frac{dn}{dc}$ is fine-tuned with an external standard, bovine serum albumin (BSA) with a known monomeric MW of 66.4 kDa.

2.6.1.2 Data acquisition and processing

SEC-MALLS experiments were performed using an analytical Superdex 75 10/300 GL column, Superdex 200 10/300 GL column (24 mL column volume, GE Healthcare Life Sciences) or Superose 6 column, a Shimadzu LC-20AD Prominence HPLC system, a Dawn HELEOS-II light scattering detector with 18 parallel detectors at different angles (Wyatt Technologies), a SPD-20A UV/Vis detector (Shimadzu), an Optilab rEX refractive index monitor (Wyatt Technologies) and the analysis program Astra (Wyatt Technologies). Typically, 100 µL protein at 1-8 mg/mL in a suitable buffer was injected onto a SEC column equilibrated in matching buffer conditions (flow rate 0.5 mL/min). The A_{280}, static light scattering, QELS and refractive index were recorded in series and Astra software was used to analyse the elution signals in parallel. The refractive index signal was calibrated to eliminate systematic errors by correcting the increment of the refractive index $\frac{dn}{dc}$ using an external standard of BSA (2.5 mg/mL).

2.6.2 Circular dichroism (CD)

2.6.2.1 Theory

CD detects the difference in absorption of left- and right-handed circularly polarised light that arises due to the electronic transitions of atoms in a protein[269]. The intensity of energy absorbed is dependent on the geometric bond angles between atoms, which are inherent to the secondary structure in a protein[270]. The resulting spectrum can be interpreted by performing interpolations to spectra of proteins of known secondary structure, resulting in an estimated amount of secondary structure present in the sample[269] (Figure 2.2). The Dichroweb server[271,272] was used for data analysis. The interpretation algorithm which fit the experimental data with the lowest error was the algorithm based on a general constrained regularisation method with local linearisation (CONTINLL)[273–275], which uses a linear combination of reference spectra from mostly α-helical proteins.

CD is an excellent technique to study the effect of pH and temperature on proteins and is complementary to nano-Differential Scanning Fluorimetry (nano-DSF), because it monitors the loss of secondary structure rather than the changing environment of tryptophan and tyrosine residues. CD cannot be used to study ionic strength effects as Cl^- ions affect the absorption properties of the buffer, leading to low protein absorption[269].

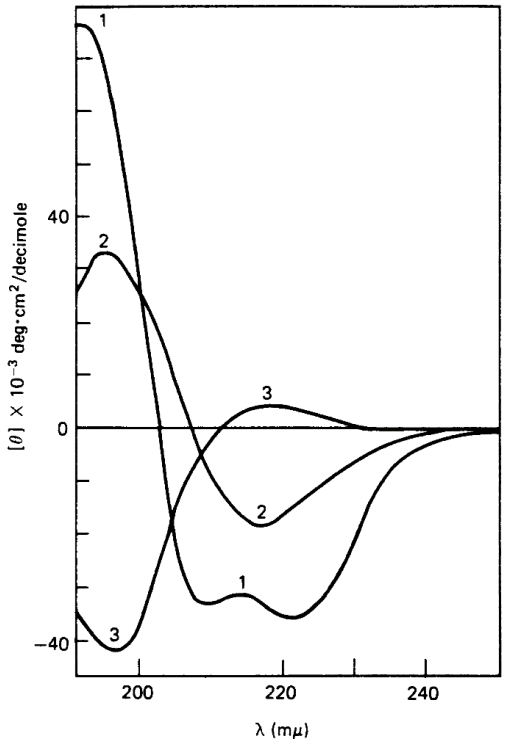


Figure 2.2: Typical CD spectra for different types of protein secondary structure. 1, 100% helix; 2. 100% β-sheet; 3. 100% random coil. Plotted as function of λ in nm[270].

2.6.2.2 Data acquisition

CD was used to determine the approximate secondary structure content and T_m (°C) of proteins. Experiments were performed on a Jasco J-810 CD spectropolarimeter with Jasco Peltier temperature control unit and Biologic SFM300 stop-flow accessory. Typically, 250 µL 0.2 mg/mL (5-25 µM) protein was dialysed into 20 mM phosphate buffer pH 5.0-7.0 and measured in a quartz cuvette of path length 0.1 cm. Dialysis buffer spectra were recorded as buffer blanks. Individual spectra were recorded at 20 °C from 190-260 nm with a sensitivity of 100 millidegrees (mdeg), a data pitch of 0.5 nm, a continuous scanning mode, a response time of 2 s, a band width of 2 nm and 5 scans, which were averaged. Two channels were recorded, CD-signal and HT (voltage), where the voltage channel was used as a control channel: scans with a voltage over 600 V contained a disproportionate amount of noise and were not used for analysis. Thermal denaturation and refolding was measured using CD at temperatures from 20 to 95 °C or 95 to 20 °C from 190-260 nm with a data pitch of 5 °C per temperature step and a temperature gradient of 2 °C/min. A delay time before recording scans was implemented of 30 s, other settings were as in the individual scan (see above).

2.6.2.3 Data processing

CD data was processed using the online analysis tool Dichroweb[271] using the analysis program CONTINLL[276,274], which analyses the spectrum as a linear combination of reference CD-spectra and performs best in estimating α-helical fractions[275], with reference set 4[275]. Unless stated otherwise, CD data was converted from mdeg into molar residual ellipticity (MRE; Equation 2.7), to allow comparison between proteins of different MW.

Equation 2.7: Molar residual ellipticity

$$[MRE] = \frac{0.1 \cdot \theta \cdot MW}{(a-1) \cdot l \cdot c}$$

where MRE is in deg cm^2 dmol^{-1} res^{-1}, θ in mdeg, MW in in g/mol, a is the total number of amino acids in the protein, l *is* the path length in cm and c is the protein concentration in mg/mL. Thermal denaturation at 222 nm in MRE was normalised and converted to % unfolded as follows:

Equation 2.8: Normalisation of thermal denaturation CD signal measured at 222 nm.

$$\%unfolded_{T=\tau} = \frac{CD_{T=\tau} - CD_{min}}{CD_{max} - CD_{min}} * (1-x) + x$$

where $\%unfolded_{T=\tau}$ is the percentage of folded protein at temperature τ, $CD_{T=\tau}$ is the CD-signal at 222 nm at temperature τ, CD_{min} is the minimum CD-signal at 222 nm over the full curve, CD_{max} is the maximum CD-signal at 222 nm over the full curve and x is the percentage of unfolded material at the start of the experiment (0.2 for D17, 0 for all other proteins; determined from CD).

2.6.3 Nano differential scanning fluorimetry (nano-DSF)

2.6.3.1 Theory

Nano-DSF employs the intrinsic fluorescence of tryptophan[277,278] and tyrosine residues[278,279,280], whose emission wavelength changes upon a change in environment. As a protein unfolds, apolar residues in the hydrophobic core of a protein become exposed to water. The fluorescence signal of Tyr and Trp is detected at 350 nm and 330 nm as a function of an increase in temperature from 15 to 95 °C. The ratio F350/F330 reports on a change in fluorophore environment[281]. The NanoTemper Prometheus NT.48 also monitors

protein aggregation via light scattering. Light is passed through the sample twice using a mirror. Any light not reflected back is a measure for protein aggregation[282].

Nano-DSF in combination with aggregation detection is measured in disposable capillaries, hence any buffer condition can be measured. This makes nano-DSF an excellent technique to study protein stability and aggregation as a function of salt and pH.

2.6.3.2 Data acquisition

Experiments were performed on a Prometheus NT.48 Nano-DSF system (Nanotemper Technologies) using ThermControl v2.1.2 software typically in high-sensitivity capillaries (Nanotemper), as the intrinsic fluorescence of many proteins measured in this work originated from weakly fluorescent tyrosine residues rather than more strongly fluorescent tryptophan residues. Clearly detectable changes in fluorescence intensity ratio were observed from proteins with one tyrosine (see Figure 3.14).

Typically, 10 µL 1 mg/mL protein (30-100 µM) in appropriate buffer conditions was loaded into a high-sensitivity capillary by capillary action. Experiments were performed in duplicate. An unfolding and refolding temperature gradient of 15-95 °C was performed with a temperature gradient of 1.3 °C/min at a laser excitation power of 100%. The T_m was determined by calculating the inflection point from the first derivative of the fluorescence ratio. For proteins with a broad unfolding transition, an additional Savitsky-Golay smoothing function averaging 75-155 points (full trace 900-1300 points) was applied to find the true inflection point.

2.7 Nuclear Magnetic Resonance (NMR)-spectroscopy

2.7.1 Theory

Nuclei have a quantum-mechanical property ‘spin’, which is determined by the arrangement of protons and neutrons. Certain isotopes of nuclei have spin ½; these act as magnetic dipoles and can interact with electromagnetic fields, such as ^{1}H, ^{13}C and ^{15}N[283]. Nuclei with magnetic dipoles are oriented randomly in absence of an external magnetic field, resulting in net zero magnetisation. Upon application of an external magnetic field, dipoles have a preferred orientation and align (net) to the external magnetic field; the energy of the parallel orientation is only slightly lower than that of the anti-parallel

orientation. The magnetic spin (dipole) precesses around the external magnetic field at the Larmor frequency (Equation 2.9):

Equation 2.9: Larmor frequency[284].

$$\omega = \gamma B_0$$

where ω is the Larmor frequency in MHz, γ is the gyromagnetic ratio of the nucleus in MHz/T and B_0 is the magnitude of the external magnetic field in T. NMR-experiments manipulate the net (rotating) dipole orientation by radiofrequency (r.f.) pulses, applied at the Larmor frequency (see Table 2.11), and detect the resulting magnetisation in the *xy* plane. Net magnetisation of dipoles always returns to the equilibrium *z* axis, due to relaxation processes involving a loss of coherence (transverse relaxation) and a return to equilibrium (longitudinal relaxation, see later). An NMR-spectrum is generated by a Fourier transform of the detected decaying magnetisation.

Table 2.11: NMR Frequency table for different isotopes, Bruker.

Nucleus	Natural abundance (%)	γ (MHz/T)	B_0 (T)	|ω (MHz)|
^{1}H	99.9885	42.58	16.4442	700
^{2}H	0.0115	6.54	16.4442	107
^{15}N	0.364	-4.32	16.4442	71
^{13}C	1.07	10.71	16.4442	176

2.7.2 Intramolecular effects result in local fluctuations of the magnetic field

If all ^{1}H nuclei in an NMR-sample were equally affected by the external magnetic field, the resulting 1D NMR-spectrum after an r.f. pulse would show a single resonance. Instead, the magnetic field experienced by nuclei depends on the environment of these nuclei within the protein structure. Several factors contribute to the local environment of nuclei.

The chemical shift effect (δ, in parts per million (ppm) shift to a reference frequency) is influenced by an additional local magnetic field generated by the electrons in the local chemical environment of a particular spin. The chemical shift effect is proportional to the external magnetic field and is dependent on the local configuration of the electrons, which

creates a local magnetic field. For example, nitrogen atoms withdraw electrons from the N-H bond and thus, these protons experience the external magnetic field more strongly (deshielding)[284].

Another contribution to the local magnetic field involves the coupling of dipoles. Very briefly, in direct coupling through space, the coupling depends on the relative orientation and magnitude of the magnetic dipoles, the dielectric constant and the distance between the dipoles. The dipole coupling through covalent bonds, termed J-coupling, indirect dipole-dipole coupling or peak splitting, depends on the electron configuration that gives rise to local magnetic fields, which average out over longer timescales, but fluctuate on shorter ones[285,286].

1D ^{1}H NMR is a great tool to determine if a protein is folded. In a folded protein, the local environment around nuclei is unique, resulting in an "NMR fingerprint" of the molecule. A folded protein shows a backbone amide proton signal dispersal around 7.5-10 ppm. In contrast, in an unfolded protein, there is fast exchange between many different local conformations. Only the average chemical shift of these conformations is observed, and thus result in a spectrum characteristic of an unfolded protein. The chemical shift for backbone amide protons in a random coil conformation is around 8-8.5 ppm[287].

2.7.3 Relaxation effects

Proteins tumble in solution and the tumbling rate is determined by the MW of the protein, where larger proteins tumble more slowly. Protein tumbling is on the ~ns timescale and determines the timescale of relaxation processes, together with the intramolecular effects reported in section 2.7.2. These fluctuations lead to longitudinal and transverse relaxation effects, which can be described by exponentially decaying functions with different time constants; T1 and T2, respectively (Figure 2.3). In practice, the relaxation time constants are often measured for ^{15}N, because H-N bonds are present in each amino acid (except proline residues) and give a good read-out of dynamics along the protein backbone. Proton-proton interactions are much more abundant in proteins, complicating interpretation of the data and the relaxation processes.

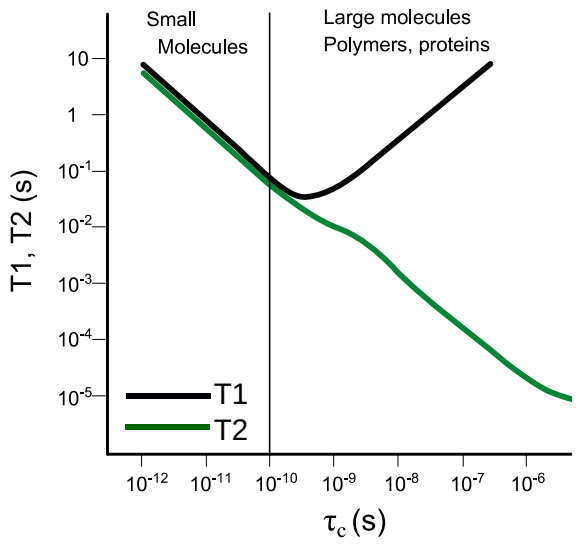


Figure 2.3: T1, T2 relaxation constants as a function of τ_C. Adapted from Reich (2017)[288].

.

2.7.3.1 ^{15}N-T1

Longitudinal or spin-lattice relaxation is initiated by all magnetic field fluctuations close to the Larmor frequency, both in the external field and in the local field[289]. They act as small r.f. pulses in all directions, returning the magnetisation to the equilibrium *z* axis. This limits the time between pulses, as first the net magnetisation has to return to the Z-axis before a new pulse can be initiated. ^{15}N-T_1 relaxation is the time constant related to the process where ^{15}N magnetisation returns to equilibrium at the Z-axis after a time delay τ[290]. Small molecules and very large proteins generate fluctuations at different frequencies than the Larmor frequency, and are therefore less affected by T1 relaxation (Figure 2.3). Small-medium proteins have motions around the Larmor frequency and therefore, they have larger ^{15}N-T_1 relaxation constants[289].

2.7.3.2 ^{15}N-T2

Transverse or spin-spin relaxation is initiated by spin-spin and dipole-dipole coupling at any frequency[289]. This slightly changes the external magnetic field that each spin experiences, leading to a loss of coherence while magnetisation precesses in the *xy* plane. This is measured as decaying net magnetisation in the *xy* plane. The rate of transverse relaxation is strongly influenced by the correlation time of the motions, where very fast motions in small molecules tend to cancel each other out, leading to slow transverse relaxation. Larger molecules have slower motions, which lead to a faster decay of transverse relaxation[289] (Figure 2.3). ^{15}N-T_2 relaxation is the time constant of the coherence loss of ^{15}N

magnetisation in the transverse plane[290]. For small proteins, transverse relaxation is shorter than longitudinal relaxation and the T2 time constant determines the theoretical linewidth of an NMR-signal:

Equation 2.10: Theoretical linewidth of an NMR signal.

$$\nu_{1/2} = \frac{1}{T_2}$$

where $\nu_{1/2}$ in Hz is the half-width at half-height of an NMR-signal and T2 is the time constant of transverse relaxation in s[283]. The real linewidth is further affected by inhomogeneities in the external magnetic field across the sample.

2.7.3.3 The Nuclear Overhauser Effect

The Nuclear Overhauser Effect (NOE) is an interaction of coupled dipoles that are close together (≤ 5 Å)[283]. In general, two coupled dipoles will show cross-relaxation of magnetisation as a function of the distance between the nuclei as r^{-6}. Therefore, NOE restraints can be used in determining the solution structure of a protein[284].

Here, the application of steady-state NOE analysis on heteronuclei in a $^{1}H–^{15}N$ bond is used. Very briefly, the proton magnetisation is saturated by the application of many r.f. pulses with a small time delay and the effect on the magnetisation of the ^{15}N nucleus is determined[291]. ^{1}H-^{15}N bonds with motions (ps-ns) faster than the molecular tumbling (ns) show a decreased hnNOE intensity ratio compared to less flexible parts of the protein backbone[290] (Figure 2.4).

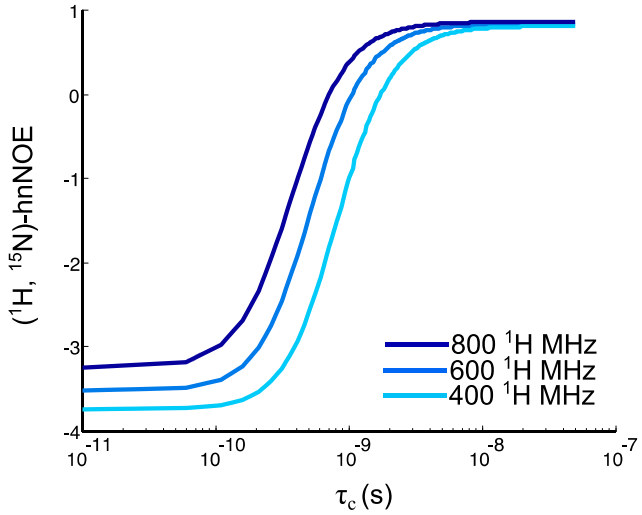


Figure 2.4: (^{1}H, ^{15}N)-hnNOE values for different field strengths[291].

2.7.4 Sample preparation

Proteins with the desired isotopic label were dialysed in 20 mM sodium phosphate buffer in pH 6.0. A typical NMR-sample (0.6 mL) contained 10% (*v/v*) D_2O (99.9%, Cambridge Isotope Laboratories). The protein concentration was adjusted to the purpose of the experiment: 0.2 mM was used for one-dimensional (1D) ^{1}H NMR-spectra and (^{1}H,^{15}N)-heteronuclear single quantum coherence (HSQC) spectra of ^{15}N-labelled proteins, while for relaxation experiments and backbone assignments ~1.0 mM ^{15}N- or ^{15}N, ^{13}C-labelled protein was used.

2.7.5 Data acquisition

NMR spectra and relaxation data were acquired on a Bruker Avance II ^{1}H 700 MHz spectrometer with room temperature triple-resonance probe or a Bruker Avance Neo with N_2-cooled triple resonance probe (Centre for Magnetic Resonance, Department of Chemistry, University of York). Relaxation data was further recorded on a Bruker Avance III HD ^{1}H 800 MHz spectrometer with He-cooled triple-resonance probe (Medical Research Council Biomedical NMR Centre, Francis Crick Institute). Solvent settings were 90% (*v/v*) H_2O, 10% (*v/v*) D_2O. Experiments were performed at 20 °C, unless stated otherwise. The experimental parameters were adjusted as in Table 2.13.

2.7.6 Data processing and referencing

Spectra were processed in TopSpin 4.0.2 (Bruker). A zero order phase correction was applied in the direct dimension. Typically, a sine squared apodisation function was applied to the free induction decay (FID) with a sine bell shift (SSB) of 2 to improve lineshape. Data was referenced to the internal lock signal of ^{2}H.

2.7.7 Triple resonance assignment

Triple resonance assignment experiments were recorded using non-uniform sampling (NUS)[292], in which the time constant in one dimension is changed non-linearly in a semi-random way. The multi-dimensional decomposition (MDD) algorithm was used to replenish missing data points in the full matrix, followed by processing as described above. Processed spectra from Topspin were imported in CcpNmr Analysis, version 2.4.2. Sequence-specific triple-resonance backbone assignments (^{1}H, ^{15}N, $^{13}C_{\alpha}$, $^{13}C_{\beta}$) were performed based on CBCACONH/CBCANH experiments. Other triple resonance assignment experiments were recorded for completeness. A detailed description of the assignment process is reported in 5.3.5. All assigned resonances are listed in Appendices 7.5, 7.6 and 7.7.

Table 2.12: T1, T2 relaxation experiment time delays. Values were randomised during measurement to generate unbiased results.

	Relaxation time delays (s)
T1	0.1, 0.1, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0, 1.5, 2.0,
T2	1, 1, 1, 2, 3, 4, 5, 6, 7, 8 (1 delay block 16.96 ms)

2.7.8 Relaxation

NMR relaxation studies provide insight into the internal dynamics of proteins on different timescales[293], in particular those of ^{15}N nuclei. Kay *et al*. (1989)[290] devised a method to measure the longitudinal (T_1) and transverse (T2) relaxation of insensitive nuclei indirectly by using two-dimensional pulse schemes. These transfer ^{1}H-magnetisation onto the insensitive nucleus, allowing relaxation processes to happen during a time delay, followed by transferring the remaining magnetisation back to the ^{1}H nuclei for sensitive detection.

2.7.8.1 ^{15}N-T1

^{15}N-T_1 relaxation experiments were recorded with ten time delays in a randomised order. The data was processed in Topspin and imported in CcpNmr Analysis. The time delays used in the experiment were imported (Table 2.12). The ^{15}N-T_1 relaxation constant was determined by fitting an exponential decay function to the decay of the signal intensity (height):

Equation 2.11: ^{15}N-T1 relaxation.

$$I(t) = I_0 e^{-R_1 t} = I_0 e^{-t/T_1}$$

where $I(t)$ is the intensity at timepoint t, I_0 is the intensity at timepoint zero, R_1 is the rate constant of ^{15}N-T_1 relaxation (s^{-1}) and t is time (s). The time constant of ^{15}N-T_1 relaxation is the inverse of the rate constant of ^{15}N-T_1 relaxation.

Table 2.13: NMR data acquisition and pulse programs for NMR experiments. [a]T2_offset experiments were recorded for S03.

Experiment	Pulse program	Scans	Number of points					%NUS, NUS points	Spectral width (ppm)			Offset (ppm)		
			^{1}H	^{15}N	^{13}C	Time points	Recycle delay (s)		^{1}H	^{15}N	^{13}C	^{1}H	^{15}N	^{13}C
^{1}H	zgesgp	4	8192				1		16.02			4.700		
(^{1}H,^{15}N)-HSQC	hsqcetf3gpsi	4	2048	512			1		16.23	30.00		4.700	118.0	
(^{1}H,^{13}C)-HSQC	hsqcctetgpsp	2	2048		512				16.23		80.00	4.700		40.00
T1	hsqct1etf3gpsi3d	4	2048	128		10	5		15.87	28.00		4.700	118.0	
T2	hsqct2etf3gpsi3d	4	2048	128		10	5		15.87	30.00		4.700	118.0	
[a]T2_offset	hsqct2etf3gpsi3d	2	2048	128		10	5		15.87	30.00		4.700	110.5	
[a]T2_offset	hsqct2etf3gpsi3d	2	2048	128		10	5		15.87	30.00		4.700	125.5	
hnNOE	hsqcnoef3gpsi	4	2048	256			5		15.87	30.00		4.700	118.0	
CBCANH	cbcanhgpwg3d	8	2048	96	160		1	40%, 1536	16.23	30.00	80.00	4.700	118.0	40.00
CBCACONH	cbcaconhgpwg3d	4	2048	96	160		1	40%, 1536	16.23	30.00	80.00	4.700	118.0	40.00
HNCO	hncogpwg3d	8	2048	96	64		1	33%, 507	16.23	30.00	16.00	4.700	118.0	172.0
HNCANNH	hncannhgpwg3d	16	2048	96, 128			1	33%, 1014	16.23	30.00, 30.00		4.700	118.0, 118.0	
HBHACONH	hbhaconhgpwg3d	4	2048, 192	96			1	33%, 1521	16.23, 16.32	30.00		4.700, 4.700	118.0	

2.7.8.2 ^{15}N-T_2

^{15}N-T_2 relaxation experiments were processed in Topspin, before they were imported in CcpNmr Analysis. As the T2 relaxation rate is influenced by the ^{15}N ppm distance from the ^{15}N offset, two T2 experiments were recorded for S03 with a ^{15}N spectral width of 30 ppm and an offset of 110.5 and 125.5 ppm, respectively (Table 2.13). The T2 values for resonances with ^{15}N ppm < 118 ppm were taken from the T2 experiment with a ^{15}N offset of 110.5 and for resonances with ^{15}N ppm ≥ 118 ppm were taken from the experiment with a ^{15}N offset of 125.5 ppm. The delay times used in the experiment were imported as in Table 2.12. ^{15}N-T_2 relaxation constants for signals were calculated by fitting the following exponential decay function to the decay of the signal intensity (height):

Equation 2.12: ^{15}N-T2 relaxation.

$$I(t) = I_0 e^{-R_2 t} = I_0 e^{-t/T_2}$$

where $I(t)$is the intensity at timepoint t, I_0 is the intensity at timepoint zero, R_2 is the rate constant of ^{15}N-T_2 relaxation (s^{-1}) and t is time (s). The time constant of ^{15}N-T_2 relaxation is the inverse of the rate constant of ^{15}N-T_2 relaxation.

2.7.8.3 Rotational correlation time

The isotropic rotational correlation time of a backbone amide bond, τ_C, is defined as the time it takes for that bond to rotate through one radian[294]. τ_C is calculated from the relaxation time constants of spin-lattice (^{15}N-T1) and spin-spin (^{15}N-T2) relaxation (Equation 2.13). The equation is valid under the assumptions of isotropic rotational Brownian diffusion and globular, rigid proteins with a MW under 25 kDa[295]. The value of τ_C is a function of the local T1 and T2 and thus is likely to be different for more rigid and more mobile parts of a protein. A global τ_c can be estimated from the average of the individual τ_c values of all the backbone amide bonds.

Equation 2.13: A. Calculation of the rotational correlation time[295]. B. error of τ_C.

A.

$$\tau_C = \frac{1}{4\pi\nu_N}\sqrt{6\frac{T1}{T2} - 7}$$

B.

$$\delta\tau_C = \frac{6}{8\pi\nu_N} \cdot \left|\frac{T1}{T2}\right| \frac{1}{\sqrt{6\frac{T1}{T2} - 7}} \cdot \sqrt{\left(\frac{\delta T1}{|T1|}\right)^2 + \left(\frac{\delta T2}{|T2|}\right)^2}$$

where τ_C is the rotational correlation time in s, ν_N is the frequency of ^{15}N at 70.971 $\cdot$ 10^6 Hz, $T1$ is the spin-lattice relaxation constant in s, $T2$ is the spin-spin relaxation constant in s and $\delta\tau_C$ is the error of parameter τ_C.

2.7.8.4 (^{1}H, ^{15}N)-heteronuclear Nuclear Overhauser Effect (hnNOE)

(^{1}H, ^{15}N)-hnNOE experiments provide insight into dynamics of the amide bond on a ps-ns timescale[293]. It is measured as a ratio between signal intensities in presence and absence of dipolar coupling by proton saturation[296], with a value over 0.5 for amide bonds in regions with limited flexibility and < 0.5 for amide bonds in regions with more flexibility[297]. (^{1}H, ^{15}N)-hnNOE experiments were split into (^{1}H,^{15}N)-HSQC spectra in presence (NOE) and absence (NONOE) of ^{1}H saturation in Topspin using the split command. Further processing was performed as described in section 2.7.6. The split (^{1}H,^{15}N)-HSQC spectra were imported in CcpNmr Analysis and (^{1}H, ^{15}N)-hnNOE-ratios of backbone amide signals were calculated as follows:

Equation 2.14: ^{1}H, ^{15}N-hnNOE-ratios[297].

$$hnNOE^Y = \frac{V_{NOE}^Y}{V_{NONOE}^Y}$$

where hnNOE^Y is the (^{1}H, ^{15}N)-hnNOE-ratio of signal Y, V_{NOE}^Y is the volume of signal Y in presence of ^{1}H saturation and V_{NONOE}^Y is the volume of signal Y in absence of ^{1}H saturation.

2.8 Crystallography

2.8.1 Theory

Protein crystallography contributes to understanding biological molecules (and processes) at the molecular level[298]. Protein structures are determined from the scattering of X-rays

on electrons ordered in a crystal lattice[299]. The crystalline state is required to enhance the scattering signal, because the scattering intensity from single molecules is inherently weak[300]. The scattering pattern is a result of the size and symmetry of the unit cell, defined as a parallelogram of four lattice points and enclosing a repeating unit in a crystal lattice[298,299], and the location of atoms in the lattice[300]. The scattering pattern and intensity are related to the electron density map by the inverse Fourier transfer as follows, containing an amplitude and a phase component[301]:

Equation 2.15: Calculation of the electron density map[301] .

$$\rho_{xyz} = \frac{1}{V}\sum |F_{hkl}|\, e^{ia_{hkl}} e^{-2\pi i hx+ky+lz}$$

where ρ_{xyz} is the electron density at position xyz, V is the volume of the unit cell, $|F_{hkl}|$ is the structure-factor amplitude and a_{hkl} is the associated phase of planes hkl. The amplitude component is directly measured from the intensity, but phase information is missing. The techniques discussed here to solve phase problem are experimental phasing by single- or multiple-wavelength anomalous dispersion (SAD,MAD) using heavy atoms[302] or halides[303] and molecular replacement (MR)[304,305].

In experimental phasing, the location of a heavy atom is identified using anomalous scattering, from which initial phases can be estimated[302]. Anomalous scattering occurs when particular atoms absorb X-rays at a specific wavelength (Figure 2.5), the 'absorption edge', which changes the atomic scattering factor f (Equation 2.16)[298]:

Equation 2.16: Atomic scattering factor[298].

$$f = f_0 + f'(\lambda) + if''(\lambda)$$

where f_0 is scattering resulting from the protein and $f'(\lambda)$, $f''(\lambda)$ are the real and imaginary components of the anomalous scattering as a function of the wavelength λ, respectively. Briefly, Friedel's law states that reflections related by 180° have equal amplitudes and phases with equal value but opposite sign[299]. However, Friedel's law is broken for anomalously scattering atoms[298]. Typically, the wavelengths λ that are selected for MAD (see Figure 2.5), are λ_1 (peak of f'), λ_2 (minimum of f' and inflection point of f'') and λ_3 (remote λ). The anomalous contribution allows for an approximation of the phases of the anomalously scattering atoms and its application to calculate an initial electron

density map, solving the phase problem[299]. Traditionally, metal ions were used as they are heavier and thus result in more anomalous scattering[306]. However, the halide ions iodine and bromine are also suitable as scattering atoms[303].

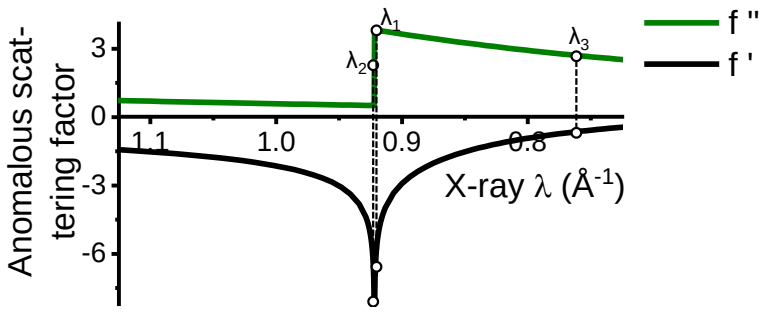


Figure 2.5: Anomalous scattering plot of bromine with real (f') and imaginary (f'') components with an absorption edge of 0.922 Å^{-1}. Data from Brennan and Cowan (1992)[307].

MR solves the phase problem by the use of homologous models[302]. A search algorithm in Phaser[304] first rotates and then translates homologous models and selects the orientation where the predicted diffraction best matches the observed diffraction. The phases are calculated from the model and combined with the measured amplitudes, to calculate an initial map[305]. The quality of the fit is reported by the Log Likelihood Gain (LLG) and Translation Function Z-score (TFZ). The LLG is a measure of the accuracy and completeness of the model with LLG > 40 likely indicating a correct solution[308]. The TFZ relates the fit of a model to the signal-to-noise and a score above 7 indicates the structure is probably solved[308].

The quality of data collection is assessed to determine which dataset is of the highest quality. R_{merge} (Equation 2.17A) represents the spread of intensities for independent measurements compared to their average intensity and traditionally, data were excluded when R_{merge} exceeded 0.6-0.8. However, the value of R_{merge} rises with multiplicity and using R_{merge} as a data cut-off discards useful reflections[309,310]. Therefore, the precision-indicating merging R factor R_{pim} (Equation 2.17B) can be used as a modified version of R_{merge}. It corrects for the multiplicity and predicts the measurement precision of the reflections[310].

Equation 2.17: A. R_{merge}[49] and B. R_{pim} [311].

A. $$R_{merge} = \frac{\sum_{hkl}\sum_{i=1}^{n}|I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl}\sum_{i=1}^{n} I_i(hkl)}$$

B. $$R_{pim} = \sum_{hkl}\frac{1}{(n-1)^{1/2}}\frac{\sum_{i=1}^{n}|I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl}\sum_{i=1}^{n} I_i(hkl)}$$

where $I_i(hkl)$ is the intensity of reflection i as a function of hkl, $\bar{I}(hkl)$ is the average reflection as a function of hkl and n is the number of independent observations of a reflection. Furthermore, the correlation coefficients (CC) between intensity estimates from half datasets ($CC_{1/2}$; Equation 2.18B) reports on the agreement between the two halves of a dataset. CC_{work} and CC_{free} are defined as the standard and cross-validated correlations of the experimental with the calculated intensities[310]. $CC_{1/2}$ approaches 1 for low-resolution data and decreases at higher resolution. A cut-off value of 0.3 is used here to select an appropriate resolution cut-off[310]. The CC for averaged data, CC* (Equation 2.18C), estimates the true CC of the data. When CC* < CC_{work}, this indicates that the crystallographic model overfits the data[312].

Equation 2.18: A. R-factor[312], **B.** $CC_{1/2}$ factor[312] and **C.** CC*[310].

A. $$R = \frac{\sum_{hkl}|F_{obs}(hkl) - F_{calc}(hkl)|}{\sum_{hkl}|F_{obs}(hkl)|}$$

B. $$CC_{1/2} = \frac{< I^2 > - < I >^2}{< I^2 > - < I >^2 + \sigma_e^2}$$

C. $$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}$$

where R is the crystallographic R-factor, $F_{obs}(hkl)$ and $F_{calc}(hkl)$ are the observed and calculated structure-factor amplitudes, $< I >$ is the average intensity and σ_e^2 is the mean error within a half-dataset. R_{work} (Equation 2.18A) describes the agreement between the structure factors calculated from the model and the experimental data. R_{free} (Equation 2.18A) is calculated using 5% of the experimental data not used during refinement and is below 0.35[313]-0.40[314] for correct models. Model validity is assessed using the gap between R_{work} and R_{free}, temperature factors or B-factors, geometry parameters (for example, Ramachandran outliers[315] and root mean square displacement (RMSD) bond angles and lengths[313,316-317].

2.8.2 Protein crystallisation

Conditions for protein crystal formation were screened using a concentrated protein solution (15-60 mg/mL) in 20 mM Tris, 150 mM NaCl, pH 7.5 and commercial crystallisation screens, typically Index (Hampton Research), the Joint Centre for Structural Genomics (JCSG)-*Plus HT-96* screen (Molecular Dimensions)[318], the pH, Anion, Cation Testing (PACT)-*Premier* HT-96 screen (Molecular Dimensions) [318], a screen with the additive 2-metyl-2,4-pentandiol (MPD)[319] (Qiagen) and the $AmSO_4$ suite (Qiagen). 54 µL screening solution was dispensed into 96-well sitting drop trays (MRC96T-UVP, SwissCi) using a Hydra96 (Robbins Scientific), after which sitting-drop crystallisation conditions were dispensed using the Mosquito LCP robot (TTP Labtech). Protein was typically dispensed from a 96-well V-bottom plate (Greiner Bio-One). Usually, two drop ratios were screened per plate: 150 nL protein solution plus 150 nL well solution (ratio 1:1) and 150 nL protein solution plus 300 nL well solution (ratio 1:2). Plates were sealed with ClearVue sheets (Molecular Dimensions) and incubated at 6 °C and 20 °C (Rigaku CrystalMation™ with two Gallery HT incubators at 6 °C and 20 °C and Minstrel™ Hi Res UV imaging system) and imaged using a custom imaging scheme, following a Fibonacci series, up to 16 weeks.

2.8.3 Cryoprotection and data collection

All crystals were harvested using nylon loops (Hampton Research) matching the crystal size and soaked for 30 s-2 min in mother liquor, supplemented with 25% ethylene glycol (EG) as cryo-protectant if required. Crystals were vitrified in liquid nitrogen and screened 'in-house' for protein diffraction on a MicroMax-007 HF generator with R-Axis IV++ image plate detector, VariMax HF confocal X-ray optics, Actor robotic sample changer (all Rigaku) and 700 series cryostream (Oxford Cryosystems). Promising conditions were optimised by screening the pH and precipitant concentration in a hanging-drop setting to obtain larger crystals, still relying on the principle of vapour diffusion[320]. Typically, 24-well trays (CellStar, Greiner) were prepared manually with 1 mL reservoir solution. 1 µL protein solution was mixed with 1 µL reservoir solution on siliconised glass slides, sealed with high vacuum grease (Dow Corning) and equilibrated against the reservoir in a hanging-drop setting. For experimental phasing, crystals were soaked in NaBr and NaI and tested for X-ray diffraction on a MicroMax-007 HF X-ray generator with a Mar345 image plate detector (MarResearch), VariMax HF confocal X-ray optics (Rigaku) and a 700 series cryostream (Oxford Cryosystems).

Datasets were collected at Diamond Light Source on the tuneable beamlines I03 and I04 with oscillations of 0.1° over a total rotation range of 360° with an exposure time of 0.0375 s per oscillation. If required, crystals were washed with liquid nitrogen to remove surface ice, before data collection. For experimental phasing, the wavelength of the 'absorption edge' of Br determined by a fluorescence scan was 0.9202 Å. For native data, the X-ray wavelength was 0.9795 Å. Spot finding and integration were performed using the '3dii' pipeline in *xia2*[321], the integrated reflections were imported in CCP4i[322], followed by scaling and merging in Aimless[323].

2.8.4 Crystallisation of D1617

D1617 was crystallised at 36 mg/mL in JCSG-*Plus HT-96* (Molecular Dimensions) in 0.2 M lithium sulfate, 50% (*v/v*) polyethylene glycol (PEG) 400, 0.1 M sodium acetate, pH 4.5 (well A1). The crystallisation conditions were optimised in a hanging drop 24-well setting to 0.2 M lithium sulfate, 50-52.5% (*v/v*) PEG 400, 0.1 M sodium acetate, pH 4.9. Crystals of ~0.5 mm size formed within 24 hours at 4°C and 20°C. No additional cryoprotectant was required for vitrification.

Protein crystals from the optimised crystallisation condition in 0.2 M lithium sulfate, 50% (*v/v*) PEG 400, 0.1 M sodium acetate pH 4.9 at 4°C were used for halide soaks. As no suitable MR model was available, a halide soak was attempted for 1 min in 2.0 M NaBr prior to vitrification to obtain phase information. Protein diffraction was confirmed 'in-house' prior to X-ray diffraction data collection at the Diamond Light Source (see section 2.8.3), where a dataset was obtained with a maximum resolution of 1.62 Å.

2.8.5 Structure determination and refinement of D1617

First, experimental phasing was attempted as no suitable homologous model for MR was available in the PDB. Anomalous scattering was detected by SHELXC[324] in the dataset recorded at the Br 'absorption edge' up to a resolution of 2.9 Å (Figure 2.6A), however SHELXD[325] failed to correctly place Br ions in the unit cell, as the CC_{all} was below 30 for all Br ions placed[326] (Figure 2.6B). Crank2[327] and AutoSol[328] also failed to correctly place Br ions in the unit cell (data not shown).

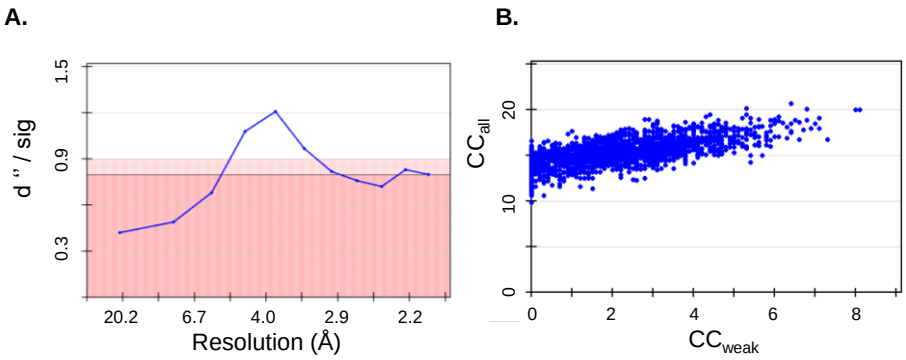


Figure 2.6: SHELX analysis of D1617 SAD data. A. Output from SHELXC[324] with the anomalous signal-to-noise shown as d''/sig as a function of resolution. Red: anomalous signal is not significant. **B.** Output from SHELXD[325]; CC_{all} is the CC between anomalous differences calculated from the placement of Br ions and experimental data, and CC_{weak} is the CC for 30% of data not used in this process.

MR was performed using 1-6 ideal α-helices of length 6-14 amino acids in Phaser[304], as *in silico* analyses and CD experiments showed that D1617 contained a high proportion of α-helices (see section 3.3.7.2). Models containing ideal α-helices are expected to be highly accurate but very incomplete. To distinguish correct solutions, no packing errors were allowed, a TFZ-score cut-off for a 'solved' solution of at least 0.75 was required, and the best 100 solutions were output. Dr Huw Jenkins is acknowledged for ideal α-helix models and advice on Phaser settings. Importantly, translational non-crystallographic symmetry (tNCS) settings were used to exclude packing clashes in the translation function. Only MR attempts using ideal α-helices of 13 amino acids in length succeeded in placing helices in the electron density and avoiding packing clashes, resulting in a TFZ of 7.76 and LLG of 81.

This MR solution was extended from 39 residues into a polyalanine model containing 151 residues (94% of the expected total amino acids) using SHELXE[329]. The resulting CC between the native intensities and those from the model increased to 44%, where a CC-value over 25% indicates that the structure is probably solved. Model building was continued using Buccaneer[330] and ARP/wARP[331]. The model was refined to a resolution of 1.62 Å by alternate cycles of manual corrections in Coot[332], followed by refinement using REFMAC5[333] and phenix.refine[334,335] using anisotropic B-factors for individual atoms. The final model comprises residues 1811 to 1962. Figures were prepared using CCP4MG[336].

2.8.6 *In silico* analysis of D1617

2.8.6.1 Properties of the interface of tandem domains

The buried surface area (BSSA) represents the area buried by the formation of the interface. The relative surface area (RSA) is the ratio of the percentage of the solvent

accessibility of residue X in the structure, compared to an extended tripeptide Ala-X-Ala. These values are calculated as follows:

Equation 2.19: Surface area calculations. A. BSSA. **B.** RSA.

A.

$$BSSA = \frac{ASA^{D16} + ASA^{D17} - ASA^{D1617}}{2}$$

B.

$$RSA = \frac{\sum_x ASA_x^p}{\sum_x ASA_x^c} * 100\%$$

where ASA^Y is the solvent accessible surface area of protein (domain) Y, ASA_x^p is the ASA for amino acid X in the protein structure and ASA_x^c is the ASA for amino acid X in an extended tripeptide Ala-X-Ala. The solvent accessible surface areas (ASAs) were calculated by Naccess[337].

2.8.6.2 Determination of key residues in the DRESS interface

PISA calculates the contribution of each residue to the solvation energy and interface area upon formation of the interface, taking into account energetic contributions from van der Waals interactions, hydrogen bonds and salt bridges[338]. 'Hotspots' are conserved residues in interfaces with a very low ASA due to the formation of the interface. They are spatially close to interacting residues and therefore exclude water molecules from the binding interface. They make significant energetic contributions to the stability of the interface[339,340]. HotSprint[340] uses the increase in BSSA upon formation of the interface, the sequence conservation at a position and the propensity of amino acids to occur in hotspots to predict which residues energetically contribute strongly to the formation of an interface, based on the O-ring theory[341,342].

2.9 Small angle X-ray scattering (SAXS)

2.9.1 Theory

SAXS of proteins in solution provides information on their shape, size and flexibility[114]. X-ray scattering is radially averaged as proteins are assumed to be in random orientations in solution and plotted one-dimensionally as scattering intensity I(q) as a function of the momentum transfer vector q = $4\pi \frac{\sin\theta}{\lambda}$ (Figure 2.7B), where 2θ is the scattering angle and λ is the radiation wavelength (nm)[343]. The distance distribution function P(*r*) describes the probability distribution of the distance *r* between two particles in the protein as a function

of *r* (Figure 2.7C). The shape of the logarithmic scattering I(q) and P(*r*) as a function of q and *r* is indicative of the particle shape in solution[344] (Figure 2.7).

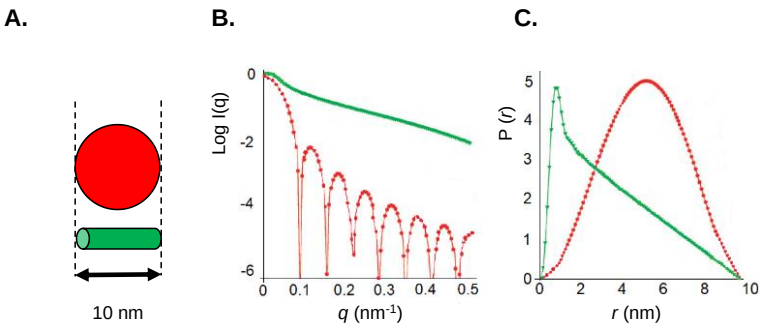


Figure 2.7: Scattering intensities and distance distribution functions of geometrical bodies (adapted from Svergun and Koch, 2003)[344]. **A.** Particle shape. **B.** Shape of logarithmic I(q). **C.** Shape of P(r).

The scattering intensity is described by Equation 2.20[345]:

Equation 2.20: Scattering function I(q)[345].

$$I(q) = 4\pi \int_0^\infty p(r) \frac{\sin qr}{qr} dr$$

where I(q) is the scattering intensity and $\int_0^\infty p(r) \frac{\sin qr}{qr} dr$ is the spherically averaged electron density distribution. At very low scattering angles, the scattering intensity can be estimated using the Guinier approximation[343,345] as follows:

Equation 2.21: Guinier approximation[343,345]:

$$\mathrm{Ln}\, I(q) = \mathrm{Ln}\, I(0) - \frac{q^2 R_g^2}{3}$$

where $I(0)$ is the extrapolated scattering intensity at zero scattering angle and R_g is the radius of gyration. This approximation is valid for rod-like particles for $qR_g \leq 1.1$[345], while globular proteins typically have a valid Guinier approximation for $qR_g < 1.3$[346]. In the Guinier plot, $\mathrm{Ln}\, I(q)$ is plotted as a function of q^2, allowing to determine $I(0)$ from the intercept and R_g from the slope. The modified Guinier approximation (Equation 2.22) is valid for rod-like particles and allows for the determination of the R_g of the cross-section of the rod[344,345]:

Equation 2.22: Modified Guinier approximation[344,345]:

$$\mathrm{Ln}\,[qI(q)] = \mathrm{Ln}\,I_c(0) - \frac{q^2 R_{c,g}^2}{2}$$

where $I_c(q)$ is the cross-sectional intensity at q, $I_c(0)$ is the extrapolated cross-sectional scattering intensity at zero scattering angle and $R_{c,g}$ is the radius of gyration of the cross-section of the rod. Plotting $\mathrm{Ln}\,[qI(q)]$ as a function of q^2, the $R_{c,g}$ can be determined from the slope of the linear proportion of the curve. The Porod exponent can be determined from the Porod plot, where Log $I(q)$ is plotted as a function of Log q. Here, the Porod exponent is defined as the negative slope of the mid-q scattering region with the following relation to the shape of the scattering molecule[347–349] (Table 2.14):

Table 2.14: Porod exponent and shape of molecule[347–349].

Porod exponent	Molecular shape
1	Rigid rod
2	Random coil, thin circular disk
3	Collapsed polymer chain, globular three-dimensional protein with rough surface
4	Well-folded three-dimensional protein with smooth surface

In a Kratky plot, $q^2 I(q)$ is plotted as a function of q and the shape of the plot provides a qualitative indication of folding of particles in solution (Figure 2.8)[343,350]. Well-folded proteins typically display a bell-shaped curve, converging to the baseline at higher values for q. Unfolded or flexible proteins do not converge back to baseline.

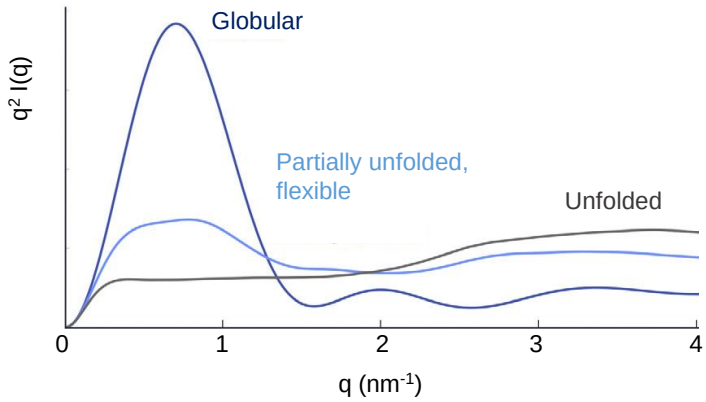


Figure 2.8: Kratky plot with qualitative indications of protein folding. Adapted from Kikhney and Svergun, (2015)[343].

P(r) is obtained by Fourier inversion of the scattering intensity function, as follows:

Equation 2.23: Distance distribution function P(r)[345].

$$P(r) = \frac{1}{2\pi^2}\int_0^\infty I(q)\cdot qr\cdot \sin qr\cdot dq \text{ for } 0 < r < D_{max}$$

where $\int_0^\infty I(q)\cdot qr\cdot \sin qr\cdot dq$ is the spherically averaged scattering function, r is the distance between two particles and D_{max} is the largest dimension of the particle.

2.9.2 Sample preparation

SAXS analysis was performed on D0710 and D0118 in SEC-SAXS mode. 60 µL D0710 at 7 mg/mL and D0118 at 0.36 mg/mL in 20 mM Tris, 150 mM NaCl, 1 mM EDTA, pH 7.5 were vitrified in liquid nitrogen. SEC-SAXS analysis of D0710 was performed on a Superdex increase 3.2/300 column (cv 2.4 mL) and of D0118 on a Shodex KW404-4F column (cv 4.6 mL) at Diamond Light Source beamline B21.

2.9.3 Data collection and processing

All data was collected at Diamond Light Source beamline B21 on a Pilatus SM detector over a range of 0.0037 – 0.50 $Å^{-1}$ at a wavelength of 1 nm. Data was analysed using the ATSAS[351] software suite (European Molecular Biology Laboratory (EMBL)). The SAXS scattering data as function of the elution frame was loaded into CHROMIXS. Elution of protein was detected by measuring the A_{280} signal. Coupled 18-angle static light scattering detectors (Wyatt) estimated the R_g. Buffer subtraction was performed by subtracting buffer frames from peak elution frames. The resulting scattering intensity was restricted to the range containing a sufficient signal-to-noise ratio by SHANUM[352].

The R_g was determined from a Guinier plot (Equation 2.21)[345,346]. The cross-sectional $R_{c,g}$ of rod-like proteins was determined from the negative slope of a modified Guinier plot of the product (Equation 2.22). The MW was estimated based on a relative intensity scale without shape assumptions from an integral of the Kratky plot, where q^2*(I(q)) is plotted as a function of q, using the online tool provided by Fischer *et al*. from www.saxs.ifsc.usp.br[353]. The Porod exponent was obtained from the Porod plot, ^{10}Log I(q) as a function of ^{10}Log q, and informs about the shape of the molecule (Table 2.14)[347].

P(r)[354] is calculated using the software tool GNOM[355] in ATSAS (EMBL)[351]. The diameter of the scattering molecule was estimated from the inflection point of P(r)[345]. The maximum dimension of the particle, D_{max}, was obtained from the maximum r of P(r).

2.9.4 *Ab initio* modelling

To quantitatively determine if the elongated rod-like molecules were flexible, ensemble optimisation modelling (EOM) was employed[356]. Rigid bodies comprising the molecule (see section 2.9.4.1) were provided, which were translated and rotated into 10,000-100,000 possible ensembles. The theoretical scattering intensity of these ensembles was then calculated, where χ^2 is the discrepancy between the scattering curves of the models and experimental data. Subsets of ensembles were selected from the pool to best represent the data and avoid redundant model incorporation. Typically, an ensemble consists of 1-5 models for rigid molecules and 10-25 models for flexible molecules.

EOM 2.0 was downloaded as part of the ATSAS[351] software suite (EMBL) and runs for D0710 were performed assuming no core symmetry (P1), a random coil-like C_α distribution, no fixed subunits, an initial ensemble of 100,000 models, a harmonics setting of 50 optimised for large molecules, q_{max} = 0.35 $Å^{-1}$, 100 points, using the genetic algorithm.

EOM runs on D0118 required too much computer power to be run on the local desktop computer, due to the ensemble comprising eighteen rigid-body models and generating a large number of models. Therefore, the online EOM server was employed, where the harmonics setting was 15 (out of 10-50) by default, which is non-optimal for rod-like proteins, and the number of models is limited to ten.

2.9.4.1 Input generation for EOM on D0710

EOM input sequences and rigid bodies are required to match. As no crystal structure was available for D0710, assumptions had to be made regarding the crystal structure of D0710. D0710 consists of four DRESS domains. Therefore, EOM was run with various rigid body inputs, in an effort to best estimate the crystal structure of D0710 (Table 2.15). The rigid body input 'D17,D17,D17,D17' was omitted, because the number of residues in the 'D17' part of the 'D1617' crystal structures is lower than in the 'D16' part, leading to a model containing too few amino acids in rigid bodies to accurately represent D0710. The sequence files were adjusted to match the provided rigid body content.

Table 2.15: EOM input models.

Number of DRESS domains	Rigid body crystal structures	Number of amino acids in rigid body crystal structures	Number of amino acids in D0710
4	D16, D16, D16, D16	313	313
4	D16, D1617, D17	313	313
4	D1617, D1617	304	313

2.9.4.2 Assessment of conformational variety in EOM pools

EOM provides a flexibility measure of the selected ensemble structures and the ensemble structures in the pool, R_{flex}[356]. This value approaches 0% for a fully rigid system and 100% for a fully flexible system. The structures in the pool sample a wide conformational space and have R_{flex} values of 85-90%. For the selections representing D0710, the R_{flex} is significantly lower than the pool; thus, these models have a narrow conformational space, suggesting limited flexibility[357] (see Table 4.5).

The limited flexibility is supplemented by the population distribution of D_{max} and R_g for different EOM inputs (Figure 2.9). The distribution of the selections of EOM runs is narrower compared to the distribution of the pools and feature only one maximum, indicative of a shared characteristic for the model/ models in the selections[358]. If D0710 contained a significant proportion of particles in, for example, a bent conformation, another peak in the D_{max} and R_g distribution would have been expected[357].

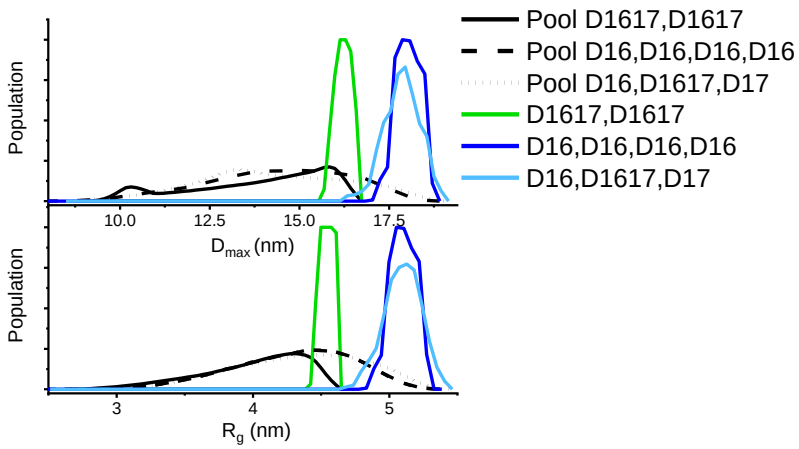


Figure 2.9: Distribution of D_{max} (nm) and R_g (nm) of models in the pool and selection of EOM runs containing different rigid body crystal structures.

2.9.4.3 *Ab initio* modelling of D0118 using the AllosMod-FoXS server

Models of DRESS domains generated by SwissModel were provided as input to the AllosMod-FoXS server[359], https://modbase.compbio.ucsf.edu/allosmod-foxs/#. The scattering profile was supplied and a model was generated. The server only generated output comprising eighteen DRESS domains when these were manually assembled into a single PDB-file. Based on the experimental scattering data, domain orientations and linkers were optimised.

2.10 Single-Molecule High-Resolution Imaging with Photobleaching (SHRImP)

2.10.1 Theory

Total internal reflection fluorescence (TIRF) microscopy is used to create an evanescent electromagnetic field at the interface between two media when the refractive index of the second medium is lower than that of the first medium and the incident angle is larger than the critical angle Θ_{crit}[360] (Equation 2.24A). This enables background fluorescence suppression of fluorophores in solution due to optical sectioning[360]. The penetration depth d_{ef} of the evanescent field is calculated from Equation 2.24B:

Equation 2.24: Critical angle (A) and penetration depth (B) for TIRF.

A.

$$\Theta_{crit} = \sin^{-1}\left(\frac{n_2}{n_1}\right)$$

B.

$$d_{ef} = \frac{\lambda}{4\pi}(n_1^2 sin^2(\Theta) - n_2^2)^{-\frac{1}{2}}$$

where Θ_{crit} is the critical angle, n_1 is the refractive index of quartz (1.46), n_2 is the refractive index of water (1.33) and λ is the wavelength of incident laser light. Θ_{crit} for an evanescent field is 66 degrees. With an incident wavelength of laser light of 488 nm to excite an Alexa Fluor 488 (A488) dye and an angle of 67 degrees, the penetration depth is 201 nm. However, it should be noted that the intensity of the evanescent field decays exponentially with distance from the interface where total internal reflection occurs.

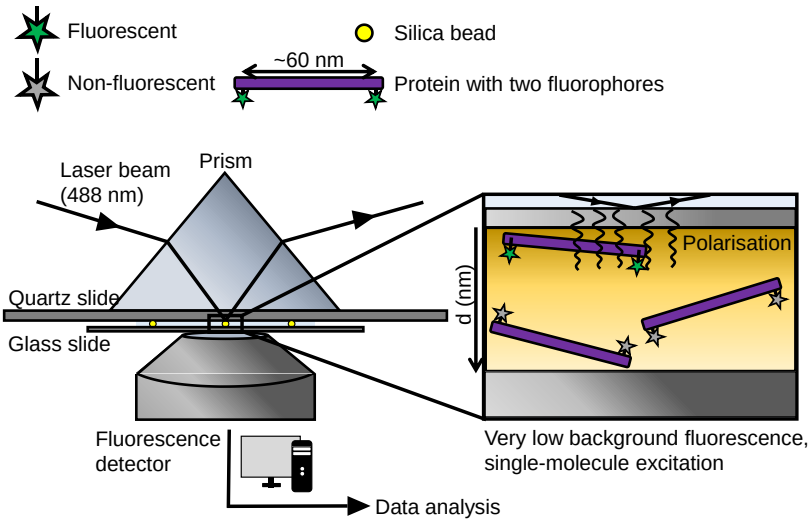


Figure 2.10: Schematic of the setup for SHRImP-TIRF microscopy. Top: legend, where stars represent A488 fluorophores (green: fluorescent; grey: non-fluorescent).

TIRF microscopy (Figure 2.10) is used to determine the inter-fluorophore distance between two A488 dyes by SHRImP[361]. SHRImP allows the measurement of distances larger than those detected by fluorescence resonance energy transfer[362,363] (FRET, up to 10 nm) and smaller than those detected by conventional light microscopy, which has the diffraction limit *d* (Equation 2.25A), here estimated to be a minimum of 174 nm. In our application of the technique, two fluorophores emit fluorescence intensity within a distance of 60 nm.

Equation 2.25: Diffraction limit.

$$d = \frac{\lambda}{2NA}$$

where λ is the wavelength of incident light, here 488 nm and *NA* is the numerical aperture, here 1.4.

SHRImP is a super-resolution imaging technique, which relies on the sequential photobleaching of fluorophores[361]. The diffraction limit describes that the distance between fluorophores with an overlapping PSF cannot be resolved[364]. In contrast, individual fluorophores can be localised with high precision by fitting a Gaussian distribution to their fluorescence intensity pattern of PSF[365]. In SHRImP, the individual PSFs of pairs of fluorophores are determined from photobleaching. Pairwise photobleaching events are characterised by two steep drops or step-wise decreases in fluorescence intensity (Figure 2.11A). The amplitudes of these drops can vary, based on the evenness of the sample illumination with the laser, the dipole-dipole orientation between the

fluorophores, the local environment of the fluorophores and 'blinking' (see later). The intensity pattern of the second-to-bleach fluorophore is obtained in a straight-forward manner, while the intensity pattern of the first-to-bleach fluorophore is obtained by subtraction of the PSF for the second-to-bleach fluorophore from the PSF for both fluorophores (Figure 2.11B). Both PSFs are fitted with a Gaussian distribution to localise the fluorophores, from which the inter-fluorophore distance is calculated[361]. Each fluorophore pair is expected to have a PSF which is circular in cross-section, because the pixel size is larger than the inter-fluorophore distance and the fluorophores have an emission dipole that rotates on a much shorter timescale than the imaging[366]. Therefore, pairwise bleaching events with an initial fluorescence eccentricity ratio outside 0.85-1.15 are rejected.

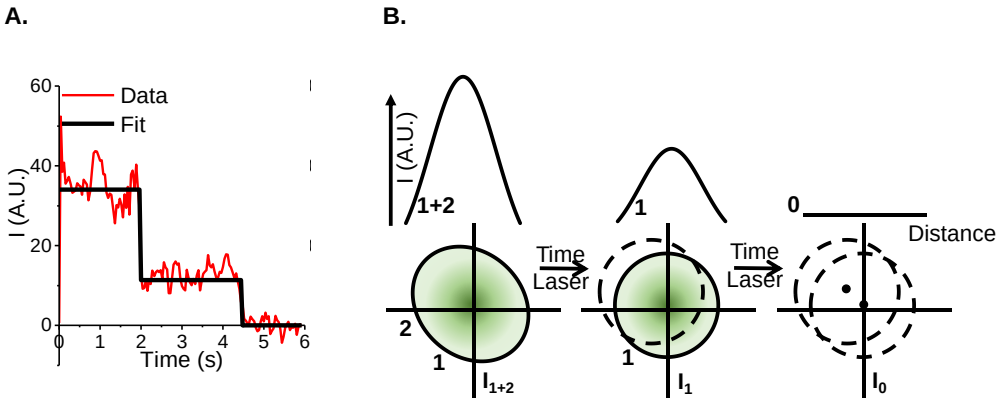


Figure 2.11: Schematic of SHRImP measurements. A. Pairwise photobleaching event. Red: raw data. Black: fit of photobleaching steps. **B.** Schematic of experimental method. For clarity, double intensity fluorophore shown as an ellipse rather than a circle.

Occasionally, fluorophores transition into a non-fluorescent state without being bleached by laser excitation (Figure 2.12A). This process is called 'blinking' and complicates analysis, as the on-off switching does not lead to the expected double stepwise decay in fluorescence. To prevent blinking, Trolox is added to the imaging buffer. Trolox (Figure 2.12B)[367] is an analogue of vitamin E, which quenches the triplet state via electron transfer (Figure 2.12A), thus preventing entry into a "dark" state. Trolox can also recover the singlet state from the "dark" state (Figure 2.12A).

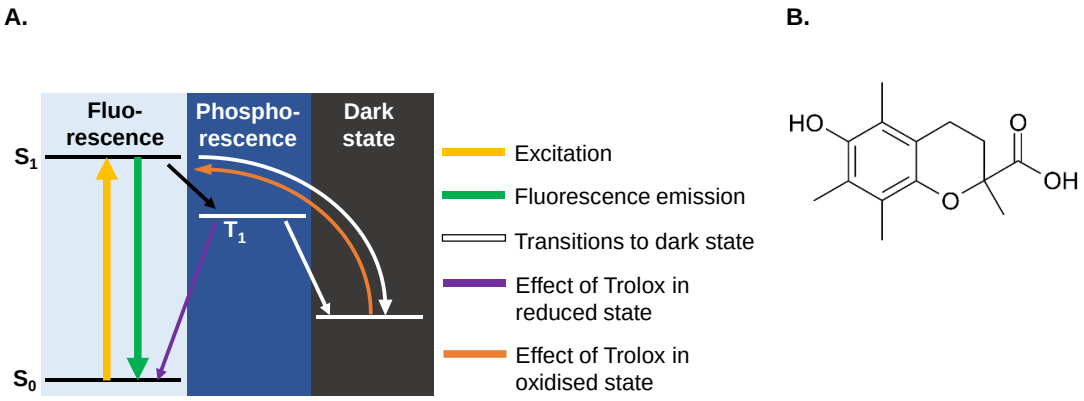


Figure 2.12: Trolox mechanism. A. S_0: singlet ground state. S_1: singlet excited state. T_1: triplet state. Adapted from Cordes *et al*.[367] **B.** Molecular structure of Trolox[368].

2.10.2 Sample preparation

Imaging buffer was prepared to a final concentration of 10 mM N-(2-hydroxyethyl)piperazine-N'-(2-ethanesulfonic acid) (HEPES), 10 mM NaCl, 1 mM Trolox, 5 mM β-mercaptoethanol, pH 7.0 or 10 mM 2-(N-morpholino)ethanesulfonic acid (MES), 10 mM NaCl, 1 mM Trolox, 5 mM β-mercaptoethanol, pH 6.5. A fresh 100 mM Trolox solution was prepared prior to each measurement as follows: 25 mg Trolox (Acros Organics) was dissolved in 150 µL methanol by vortexing and sonication. In a separate 1.5 mL microfuge tube, 50 µL 5 M NaOH was diluted with 800 µL MQ. The Trolox solution was added to the microfuge tube and mixed by vortexing.

A protein sample was prepared using 1 µL fluorescently labelled protein stock of 1 nM, 10 nM or 100 nM (depending on the desired concentration, for production please refer to section 2.3.7), 2 µL well-mixed 5 µm diameter silica bead slurry for an even sample chamber height and 47 µL imaging buffer. 30 µL protein sample was pipetted dropwise across the length of a quartz slide functionalised with 2 or 20 µg/mL poly-*D*-lysine and covered with a glass coverslip (no. 1). The short edges were sealed with transparent nail varnish and while drying, the long edges were prevented from drying out by application of some imaging buffer. This created a flow cell, which was washed with 0.5-1 mL imaging buffer to remove unbound protein and lower the background fluorescence. The long edges were sealed with transparent nail varnish. When dry, a drop of glycerol was applied on the middle of the quartz slide, on which the quartz prism was placed, to match the refractive index of the quartz. The sample was mounted in the microscope with a drop of immersion

oil (Zeiss) on the glass coverslip surface and brought into focus using the silica beads as focussing aid.

2.10.3 Data collection

The protein sample was scanned using a high electron multiplying charge-coupled device (emCCD) with gain 50 and a short exposure time (33 ms) to minimise initial photobleaching. A sample area was selected where fluorophores were well-dispersed and this area was brought into focus, after which the laser illumination of the sample was shuttered to prevent photobleaching of fluorophores before video data acquisition was started. The imaging settings were changed to a low emCCD gain (10-20), an exposure time of 200 ms and the video length was set to 150 frames. A video was recorded in which fluorophores were irreversibly bleached, after which the shutter was closed to prevent unnecessary laser illumination of the sample. Imaging settings were returned to high gain/ short exposure time to find another imaging area. Typically, up to ~100 videos could be recorded per sample.

2.10.4 Data processing

2.10.4.1 Selection of pairwise photobleaching events

A .tiff image stack file was opened in ImageJ and exported as a video in .gmv format using the ImageJ[369] plug-in "Export as AVI and GMV". In GMimPro[370], the .gmv video was opened, the full width at half maximum was increased to 400 nm and the temporal filtering parameter TempFilter was set to 3 to reduce the noise. The single fluorophore detection algorithm (SFDA)[370] was run to detect fluorophores, subtract the background fluorescence and track the fluorophores over the course of the video. The fluorescence intensity as a function of time was manually inspected for each set of *x,y*-coordinates detected by SFDA and the coordinates of events in which two fluorophores bleached were saved in a .txt file. After all photobleaching events were inspected, individual 10 x 10 pixel^2 videos were exported for all pairwise photobleaching events using the plug-in "Export particles" in ImageJ.

2.10.4.2 Determination of inter-fluorophore distance

A Matlab script kindly provided by Dr Herman Fung[57] was used to determine the inter-fluorophore distance. Briefly, it analyses the videos as follows. First, it identifies the decay

steps in fluorescence using a step-preserving Chung-Kennedy filter. Then, the intensity within each step is averaged over the frames within the step. A Gaussian distribution was fit to the resulting PSF for each fluorophore. The position of each fluorophore was determined from the maximum of the Gaussian fit using the first derivative of the Gaussian intensity. Finally, the inter-fluorophore distance was calculated as follows:

Equation 2.26: Calculation of inter-fluorophore parameters. A. Inter-fluorophore distance. **B.** Error. **C.** Eccentricity.

A.

$$d_{f_1,f_2} = l_{pixel}\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

B.

$$err_{d_{f_1 f_2}} = \frac{d_{f_1,f_2}}{2} * \sqrt{\left(2\sqrt{{err_{x1}}^2 + {err_{x2}}^2} * (x1 - x2)\right)^2 + \frac{\left(2\sqrt{{err_{y1}}^2 + {err_{y2}}^2} * (y1 - y2)\right)^2}{(x1 - x2)^2 + (y1 - y2)^2}}$$

C.

$$ecc_{f1,f2} = \frac{w_y}{w_x}$$

where d_{f_1,f_2} is the inter-fluorophore distance in nm; l_{pixel} is the size of a pixel in the magnified video image, here 97.7 nm; x_n, y_n are the *x*- and *y*-coordinate of fluorophore *n*; $err_{d_{f_1 f_2}}$ is the error of the inter-fluorophore distance in nm; err_{x_n}, err_{y_n} are the errors on the *x*- and *y*-coordinates of fluorophore *n*; $ecc_{f1,f2}$ is the eccentricity of the fluorescence intensity resulting from a fluorophore pair and w_x, w_y are the widths of the Gaussian fits in the *x*- and *y*-direction.

A filter was applied to the calculated inter-fluorophore distances based on the eccentricity of the PSF for a fluorophore pair[366]. A Matlab script kindly provided by Dr James Gilburt was applied to calculate the eccentricity of the fluorescence intensity of two fluorophores before photobleaching. The pixel size for the magnified image is 97.7 nm, thus two fluorophores attached to a single protein at an expected intramolecular distance of 60 nm will result in a circular fluorescence intensity. If the fluorescence intensity is eccentric, this suggests that the two fluorophores are on different proteins which could be adjacent to each other or cross-linked by disulfide bonds, or within aggregates. Removal of double photobleaching events with an eccentricity outside 0.85-1.15 should result in the selection

of photobleaching resulting from two fluorophores bound to the same protein within 100 nm of each other.

2.10.4.3 Statistical analysis

Inter-fluorophore distances with eccentricity values within 0.85-1-15 collected at the same poly-*D*-lysine concentration and buffer conditions were plotted as histograms with unbiased bin sizes calculated using the Freedman-Diaconis rule[371] (Equation 2.27). The histograms were fit with a unimodal Gaussian distribution (Equation 2.28). The average inter-fluorophore distance was calculated from this fit and the standard error of the fit was used as the error of the distribution (Equation 2.29). R^2 indicates the quality of the fit and represents the percentage of the variability in the data explained by the fit[372]. The standard error is preferred over the standard deviation, because n is a limiting factor in single-molecule experiments and x_i comes from a normal distribution[372].

Equation 2.27: Freedman-Diaconis rule[371].

$$bin\ size = \frac{2(Q_3 - Q_1)}{n^{-1/3}}$$

Equation 2.28: Gaussian fit to inter-fluorophore distance histograms.

$$y = y_0 + \frac{A}{w\sqrt{\frac{\pi}{2}}} e^{-2\frac{(x-x_m)^2}{w^2}}$$

Equation 2.29: Calculation of average inter-fluorophore distance. A. Mean. **B.** Standard error[372].

A.

$$x_m = \frac{\sum_1^i x_i}{n}$$

B.

$$s.e. = \frac{SD}{\sqrt{n}}$$

where $bin\ size$ is in nm; Q_1, Q_3 are the first and third quartiles (with the first quartile defined as the middle between the smallest number and the bottom half of the median and the third quartile defined as the middle between the median and the top half median); n is the number of measurements; y_0 is the offset height; A is the area; w is the width, 2 SD; x_m *is* the mean, x_i *is* a single measurement, SD is the standard deviation and $s.e.$ is the standard error.

2.11 Atomic force microscopy (AFM)

2.11.1 Sample preparation

Proteins were purified as described in section 2.3.6 and stored in 25 mM MES, 150 mM NaCl, pH 6.25, 2 mM TCEP at 0.4 mM. Prior to an experiment, 50 µL of 2 µM diluted protein was prepared by dilution into 25 mM MES, 150 mM NaCl, pH 6.25. At the AFM-facility, a protein sample was prepared as follows. A gold-coated glass square (~1 cm^2 piece of glass microscope slide) was freshly removed from the gold-coated silicon wafer (1000 Å thick gold layer with no titanium adherence layer, item no. AU.1000.SL NO TI, Platypus Technologies LLC) to minimise exposure to oxygen. The glass squares had been previously adhered to the gold-coated wafer using a two-component, thermally-cured, epoxy adhesive (EA9483, Loctite). Directly after removal, the glass square was affixed to a glass microscope slide using double-sided tape, and 10 µL of 2 µM protein solution was applied directly followed by 90 µL of 25 mM MES, 150 mM NaCl, pH 6.25. The sample was incubated for 5 minutes prior to AFM-experiments.

2.11.2 Cantilever calibration

AFM-experiments were performed on a BioScope Resolve BioAFM (Bruker). A micro-lever AFM probe with a soft silicon nitride tip (MLCT) was mounted in the fluid-cell probe holder. Probes used in this work are listed in Table 2.16. The probe holder was mounted on the AFM scan head and aligned using the EasyAlign platform. In the imaging mode setup, “PeakForce QNM mode in Fluid” under “Mechanical Properties, Quantitative Nanomechanical Mapping in Fluid” (Nanoscope v9.4), was selected and the appropriate probe type input, then the laser was aligned on the cantilever of choice (MLCT probes feature 5 cantilevers at one end). The laser deflection signal was maximised by centering on the quadrant detector. Calibration of the spring constants for these probes was performed in air, as the frequency of the thermal deflection in liquid was close to the minimum frequency detected by the system. First, the thermal deflection sensitivity was determined via the No-Touch calibration. Second, the spring constant was determined by touch calibration on a hard surface (glass coverslide, sapphire disk) using a ramp size of 800 nm and a setpoint of 0.5 V. Then, the cantilever was wetted with AFM-buffer and the laser alignment and quadrant detector signal were optimised for experiments in liquid. The protein sample was mounted and the cantilever was engaged with the sample. When using

tips with a spring constant of 0.01 N/m, this process was optimised by adjusting the Engage set-point to 0.3 V as this accounts for cantilever fluctuation in liquid. Dr Alexander Dulebo is acknowledged for AFM-support and trouble-shooting.

Table 2.16: AFM-probes.

Name	Material	Estimated spring constant (N/m)	Tip radius (nm)	Supplier
MLCT-BIO-C	Non-conductive silicon nitride	0.01	20	Bruker
MLCT-BIO-D	Non-conductive silicon nitride	0.03	20	Bruker
MLCT-C	Non-conductive silicon nitride	0.01	20	Bruker
MLCT-D	Non-conductive silicon nitride	0.03	20	Bruker

2.11.3 Data acquisition

2.11.3.1 Protein unfolding force (F) and contour length (ΔL)

The imaging mode was changed to "Mechanical properties, Fastforce Microview Contact mode in Fluid". Force experiments were performed in matrix mode, allowing a number of experiments to be recorded at different locations on the sample. This allows one to see different protein unfolding events, rather than performing pulling experiments on the same location repeatedly. The speed of the cantilever was set to 200, 800 or 1500 nm/s. The size of the ramp was set to 800 nm to ensure the cantilever was detached from any protein on the surface in between force measurements. The surface hold time was 0.1-1 s to allow the formation of a non-specific interaction between the cantilever and a protein molecule on the sample. The resolution of a force measurement was 9728 points per line. The trigger threshold was 500 pN to minimise cantilever damage. A schematic image of mechanical protein unfolding by AFM is depicted in Figure 2.13.

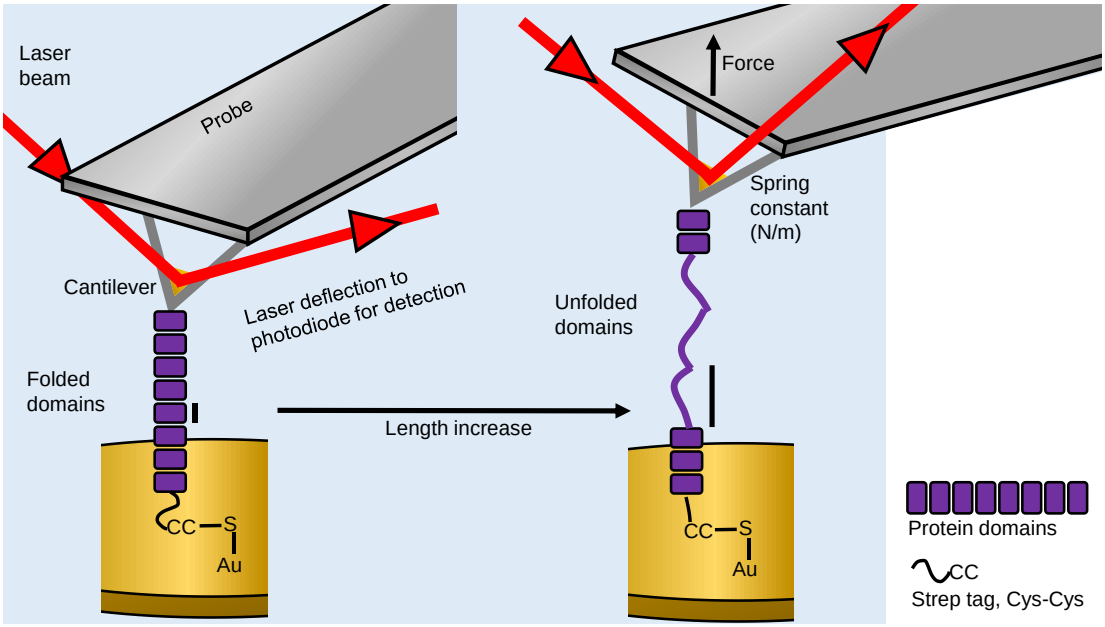


Figure 2.13: Mechanical unfolding of protein domains by AFM in solution. Proteins were immobilised on a gold-coated glass square by covalent bond formation between C-terminal cysteine residues and gold atoms. The cantilever picked up a protein via non-specific interactions. Force is applied to the probe, protein domains unfold and the increase in ΔL and F are detected.

2.11.3.2 Protein refolding

The imaging mode was changed to "Mechanical properties, Fastforce, Contact mode in Fluid Volume". Force experiments were performed in a single location, studying the repetitive unfolding and refolding of a single protein. A ramp script was optimised from the default script aimed at the mechanical unfolding and refolding of the protein domain titin[373]. The optimised ramp script parameters are shown in Table 2.17.

Table 2.17: Ramp scripting parameters.

Steps	Parameters	
Approach	Time	1 s
	Velocity	600 nm/s
	Ramp size	750 nm
	Trigger point	0.5-2 nN
Hold	Time	0.1 s
Unfold	Time	6-12 s
	Force	-35 to -50 pN
Refold	Time	2-6 s
	Force	0 pN
Retract	Time	0.6
	Velocity	1000 nm/s
	Ramp size	750 nm

2.11.4 Data processing

2.11.4.1 Protein unfolding F and ΔL

Trace selection was based on the number of unfolding events. For a protein with eight domains, up to ten unfolding events are expected: eight unfolding events for DRESS domains, an initial force signal representing pulling the protein into an extended conformation and a final force signal representing the detachment of the protein from the cantilever. Traces were selected which showed between four and ten unfolding events (between three and eight DRESS domain unfolding events), with the final event being of largest force[233]. This selection was necessary to avoid inclusion of events where multiple proteins were picked up by the cantilever simultaneously.

The analysis of a protein unfolding trace was as follows. The baselines of the approach and retract trace were corrected by first order correction using the approach baseline as a source. The approach and retract traces were then smoothed by the application of a boxcar filter, using three averaging points and a second order filter. The first and last event were not analysed, as they likely originate from sources other than individual domain unfolding. The remaining individual unfolding events were fitted using the worm-like chain (WLC)

model with a persistence length (p) of 0.4 nm. The ΔL for an unfolding event was calculated from the difference between the fits of the WLC model (Equation 2.30)[342] of two adjacent unfolding events.

Equation 2.30: WLC fit

$$F(x) = \frac{k_B T}{p}\left[\frac{1}{4}\left(1 - \frac{x}{L}\right)^{-2} - \frac{1}{4} + \frac{x}{L}\right]$$

where $F(x)$ is the force in N at an extension of x in m, k_B is the Boltzmann constant being $1.38 \cdot 10^{-23}$ m^2 kg s^{-2} K^{-1}, T is the temperature in K, p is the persistence length in m and L is the total contour length in m.

2.11.4.2 Protein refolding

Trace selection was based on the height and deflection error trajectories. Ideally, the height trajectory represents the cantilever approaching the surface over a ramp size of 700 nm, followed by height increases representing single protein domains unfolding during the Retract step. The maximum increase during the Retract step was calculated to be 224 nm, following from 8 domains with a ΔL increase of 28 nm per domain. A maximum height increase exceeding 224 nm must involve multiple proteins or a detachment of protein from the cantilever (or protein from the surface) and is therefore discarded. In the Return step, the height sensor may show a height value in the range shown in the Surface hold and Retract step. A value approaching the starting height indicates tip detachment of the protein. Finally, in the Full retract step, a sharp increase in deflection error is expected, representing the detachment of the protein from the cantilever.

The increase in ΔL per individual unfolding event was determined from the height sensor increase. The unfolding force was determined from the deflection error channel.

Chapter 3. Structural characterisation of single and tandem DRESS domains

3.1 Introduction

The repetitive region of several biofilm-forming proteins has been shown to form a rod-like region. Such proteins include SasG[59], its *S. epidermidis* homologue accumulation-associated protein (Aap[374]) and Ebh with a MW of 1.1 MDa from *S. aureus* (Uniprot accession number Q2FYJ6)[58,84]. In these examples, the rod-like structure likely has the role of projecting the functional N-terminal domain, usually involved in biofilm formation[29,30], away from the cell wall.

At the domain level, the repetitive regions employ diverse mechanisms to ensure elongation and rigidity. For example, the β-sheet rich repetitive region of SasG and Aap comprises repeats of G5- and E-domains. The highly stable interfaces between E and G5 domains ensure folding of E-domains and the formation of a stable elongated rod, while E-domains in isolation remain disordered[57,59]. In Epf, the repetitive region contains DUF1542 domains; SAXS analyses of a three-domain construct and EM analyses of a sixteen-domain construct suggested that they form an elongated structure[228], but details of inter-domain interfaces are unknown. In this chapter, DRESS (previously DUF1542) domains of the protein SasC are characterised *via in silico,* structural and biophysical methods.

3.2 Aims

In this chapter, experiments are performed to characterise the tandem DRESS domain interface and its effect on domain stability. To fulfil these goals, the following aims were set:

- To determine the correct domain boundaries of DRESS domains;
- To determine the crystal structure of DRESS domain in tandem;
- To define the linker region between DRESS domains;
- To assess the stabilising effect of the inter-domain interface.

3.3 Results

3.3.1 *In silico* analysis of the A region of SasC

3.3.1.1 Structure prediction

The secondary structure of the A region of SasC was predicted by PSIPRED[256,375]. Following the YSIRK/GXXS signal peptide (see section 1.3.2), 45% of the A region was predicted to be disordered, followed by 55% of the A region with predicted β-sheets alternated with coil regions (Figure 3.1). The coil region comprised ~230 residues and the β-sheet region 328 residues. Considering that the IgG-like domains of MSCRAMM proteins contain around 140 residues, it might be possible that SasC contains three N-terminal domains, of which N1 might be disordered and N2 and N3 might have an β-sheet-fold.

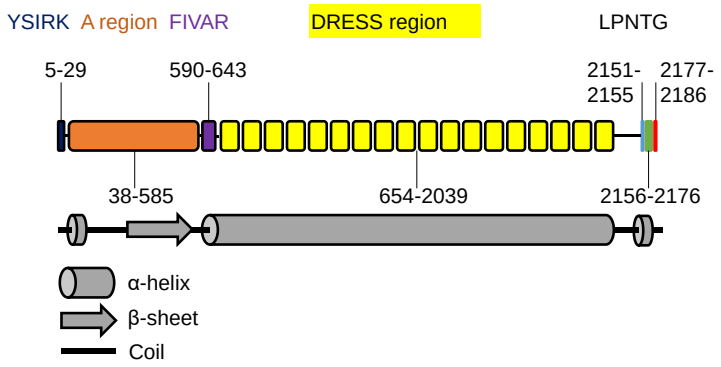


Figure 3.1: Schematic of the domain organisation and predicted secondary structure of SasC. Top: signal peptides and domain annotations. Middle: domain organisation with domains to relative size. Numbers: residue numbers as in accession number C7BUR8. Bottom: major component of the secondary structure prediction per region, calculated by PSIPRED[256,375]

A SwissModel[376,377] run on the β-sheet-rich part of the A region of SasC (residues 241-590) identified the N2 domain of ACE (PDB 2z1p), a CWA protein from *Enterococcus faecalis*, as a potentially suitable model for the putative N3 domain of SasC, with a sequence similarity of 10% (Figure 3.2). The global model quality estimation, reflecting the expected model accuracy from 0 to 1, was poor at 0.11. The QMEAN, representing the "degree of nativeness" of C_β atoms in the structure was also poor at -3.37, which is close to the cut-off of -4 for low-quality models. The local quality estimate was ~0.6, also at the cut-off for poor quality. The N1-N2 domains of ACE are similar to the collagen hug domains N1-N2 of Cna (PDB 2f68, Table 3.1). No homology was detected for residues 241-430 (189 residues) of SasC or for the part of the A region that is predicted to be disordered.

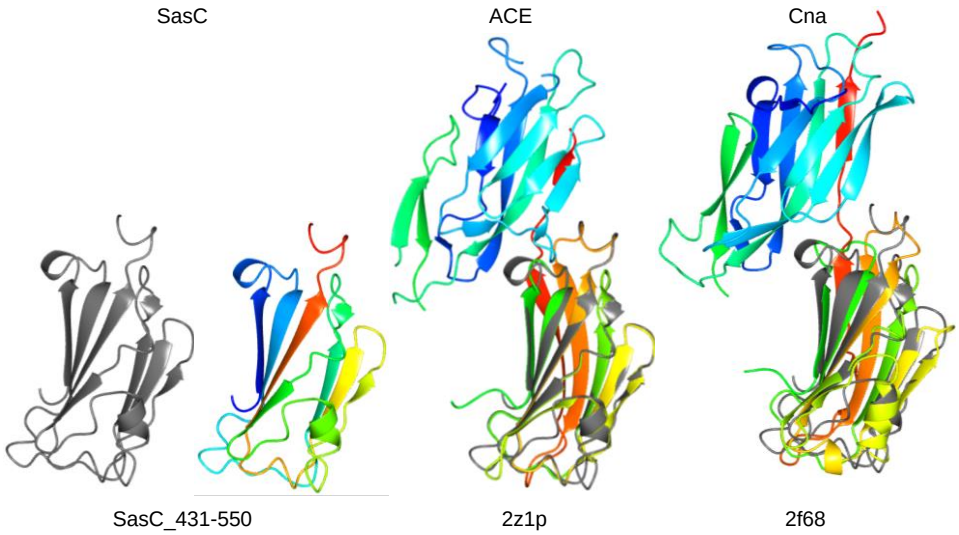


Figure 3.2: SwissModel[379] run of the A region of SasC. Image was created using CCP4mg.

Table 3.1: Structural similarity between homologues of the A regions of SasC. Shown are the *ab initio* model of SasC_431-550 and A regions of ACE (2z1p)[378] and Cna (2f68)[52].

Fixed model: 2f68[52]	RMSD (Å)	Number of res	Fixed model: 2z1p[378]	RMSD (Å)	Number of res
2z1p	3.27	255			
SasC_431-550	1.34	105	SasC_431-550	0.64	104
			2f68	3.27	255

3.3.1.2 Sequence conservation

SasC has a prevalence of 97% in clinical strains of *S. aureus*[111]. The A regions from several *S. aureus* strains were aligned and compared (see Appendix 7.2). On average, the sequence identity was 92.7%. The part of the A region that is predicted to be disordered, showed the most sequence variation and the β-sheet-rich part showed more sequence conservation.

3.3.2 *In silico* analysis of the B region of SasC

3.3.2.1 Redefining the domain boundaries of DUF1542

Sequence alignment of the seventeen reported DUF1542 domains in SasC[111] shows that sequence conservation between DUF1542 domains is low and that the number of residues comprising a DUF1542 domain is inconsistent (Figure 3.3A). Furthermore, the predicted secondary structure suggests that an α-helix connects the end of domain *n* and the start of domain *n+1*. This has been observed previously, for example in the crystal structure of tandem spectrin domains, which are connected by a contiguous helix[379], however

connecting elements of secondary structure might also indicate that the domain boundaries have been incorrectly defined.

Figure 3.3A shows that in the literature alignment, good agreement is observed prior to the start of the first DUF1542 domain (D674) and following the end of the seventeenth predicted DUF1542 domain (V1981). This might imply that redefining the domain boundaries of DUF1542 would yield another complete domain. Shifting the domain boundaries to the start of domain '0' from Figure 3.3A results in eighteen complete domains with a consistent number of 77 residues per domain, despite a slightly lower sequence conservation (28.7% instead of 30.1%, Figure 3.3B), likely due to the exclusion of the more different 'capped' domains in the literature alignment. Finally, low-homology sequence alignments of DRESS with other members from the same superfamily (see section 2.5.2) are in agreement with the new domain boundaries of DUF1542 domains, of which part is shown in Figure 3.3C.

The *in silico* observations described above allow redefinition of the domain boundaries of DUF1542. A complete domain now putatively comprises three α-helices of 20, 24 and 17 residues in length, followed by a stretch of seven residues with no predicted secondary structure, which may function as a linker. We have named the newly defined domain the DUF1542 rigid extracellular surface structural (DRESS) domain.

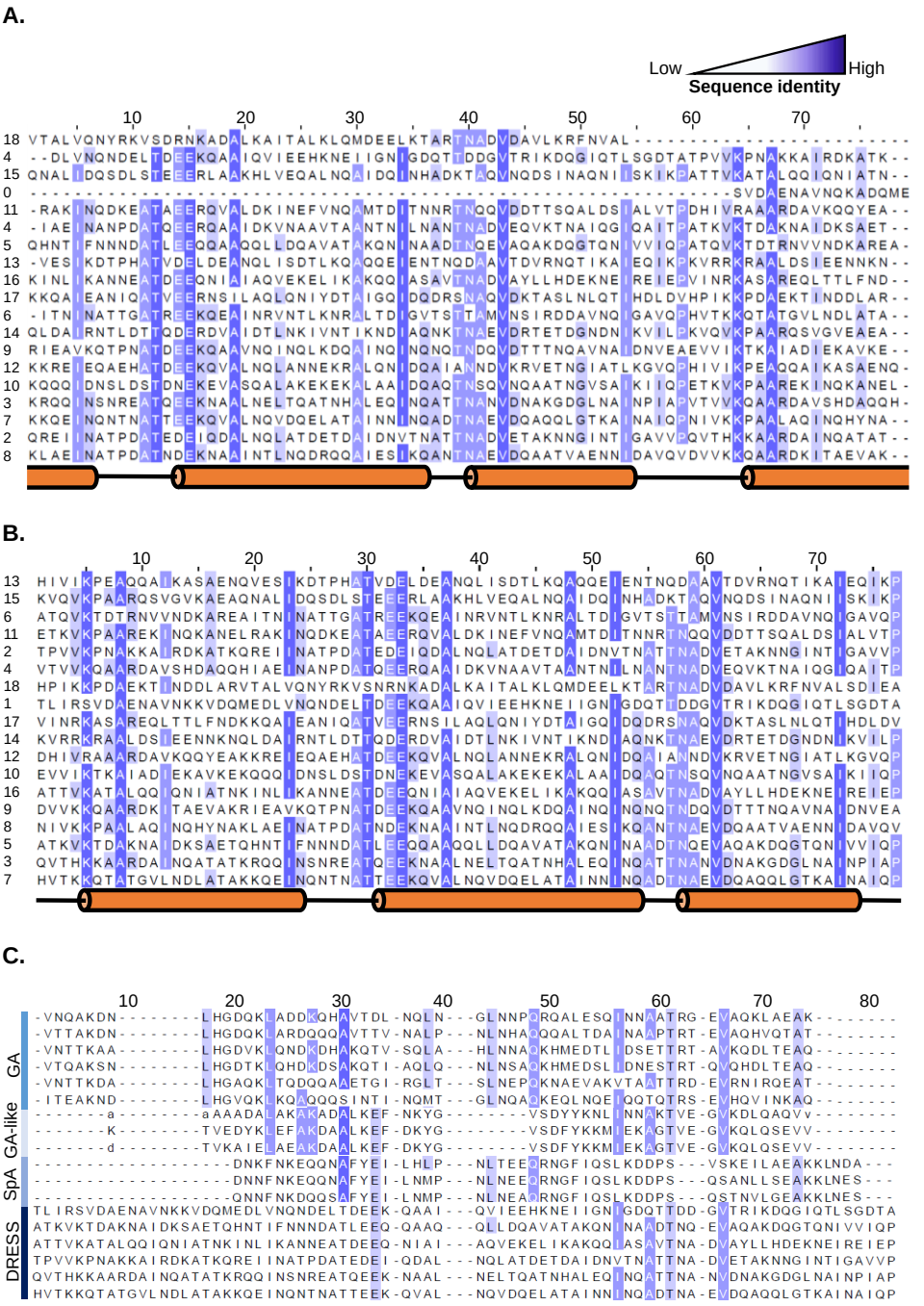


Figure 3.3: Sequence alignment of DUF1542 and DRESS domains in SasC and members of the PFam clan. **A.** MSA of DUF1542 domains using literature domain boundaries by Clustal Omega[259]. **B.** New domain boundaries. Sequences aligned by Clustal Omega[259]. **C.** Part of the sequence alignment from distant members from clan CL0598 of related triple-helical bundles aligned by MAFFT[261] (see 2.5.2).

3.3.2.2 DRESS domains in other organisms

DRESS domains are present in multiple Gram-positive bacterial species. A blastp[380,381] search was performed using eighteen DRESS domains from SasC from *S. aureus* NCTC 8325-4 to find hits in other organisms. 21-70% sequence similarities were represented in a phylogenetic tree (Figure 3.4). DRESS domains were found in both commensal and pathogenic strains of bacteria, of which 95% belonged to the phylum Firmicutes, where

bacteria are surrounded by a PG cell wall[382]. All DRESS-containing proteins contained sequences assigned, or predicted, as transmembrane regions, suggesting localisation of the DRESS-containing sequences in the extracellular environment.

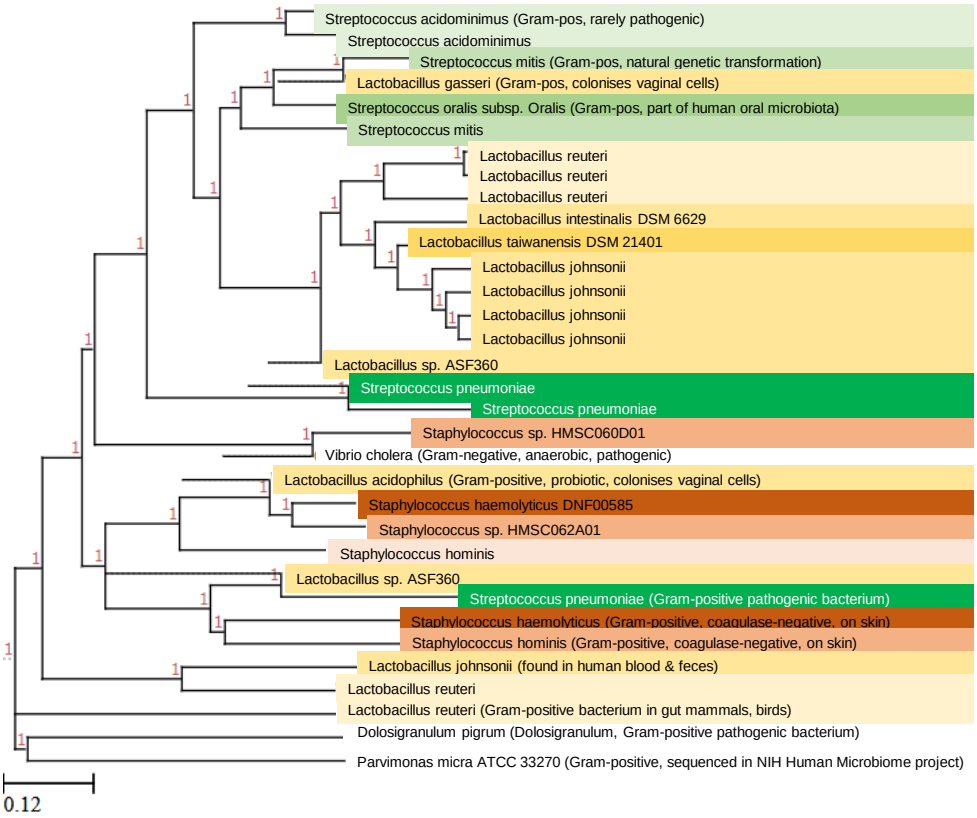


Figure 3.4: DRESS sequence similarity in other organisms. The search was based on *S. aureus* NCTC 8325-4 DRESS domains 1-18. Colour represents similarity between species. The pyhogenetic tree was created using ClustalX. The branch length indicates genetic diversity.

3.3.2.3 DRESS domains in other strains of *S. aureus*

SasC is present in 97% of clinical strains of *S. aureus*[111]. The gene encoding SasC is not essential to *S. aureus*, as strains have been identified without SasC[111], for example ATCC 29213, and the gene was not annotated as essential in a comprehensive genome screen of *S. aureus*[383]. DRESS regions from twenty *S. aureus* strains (see Appendix 7.2) were aligned in Clustal Omega[259]. All strains containing SasC contain 18 DRESS domains. The average pairwise sequence identity of the DRESS region is 97.6%.

3.3.2.4 DRESS domains in SasC

The average sequence identity between DRESS domains within SasC from strain *S. aureus* NCTC 8325-4 was 28.7%[259]. Although there are repetitive regions with very high sequence conservation (90-100% in B repeats in SasG[59], 82-100% in SHIRT repeats in SGO0707 (see

section 5.1.3)), there are other B regions with a similar sequence conservation (~24% in FnBPA)[47]. The domains with lowest sequence identity to other DRESS domains are 1 and 18 with only 11.7% average sequence identity; as the most N- and C-terminal domains, they function as 'capped' and are likely to make different connections than the middle DRESS domains. This is common in other tandemly arrayed regions[214]. The cladogram of DRESS domains (Figure 3.5) does not show conclusive evidence for formation of the repetitive region of SasC from a direct gene duplication event; rather, it might be formed by gene duplication within the ancestral gene.

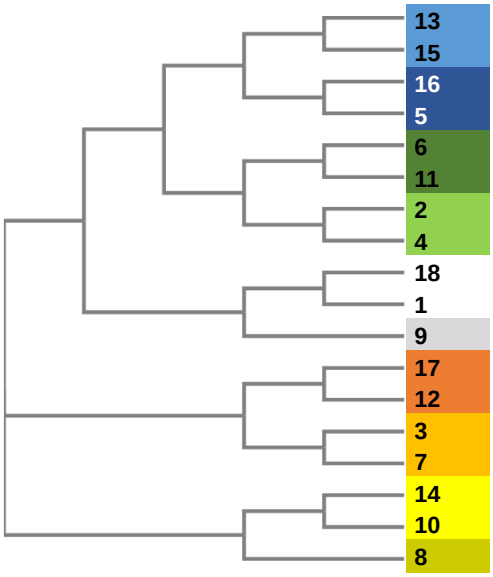


Figure 3.5 Cladogram of DRESS domains in SasC constructed by Clustal Omega. Matching colours indicate most similar domain pairs as found in the cladogram.

3.3.3 Selection of DRESS domains from SasC

DRESS domains were selected for over-production, purification and biophysical characterisation from the repetitive region of SasC (Figure 3.6). The requirement for selection was to be representative of other DRESS domains, both within SasC (Figure 3.5) and across strains of *S. aureus* (data not shown). Domains 16 and 17 showed the highest pairwise percentage sequence identity in SasC in different strains of *S. aureus* (data not shown). Domain 17 was selected for the characterisation of a DRESS domain in isolation, as the sequence identity of D17 was higher than in D16 across SasC proteins in other strains of *S. aureus* (data not shown). D17 and D1617 are at the C-terminal end of the repetitive region of SasC and thus would be located close to the cell wall (Figure 3.6).

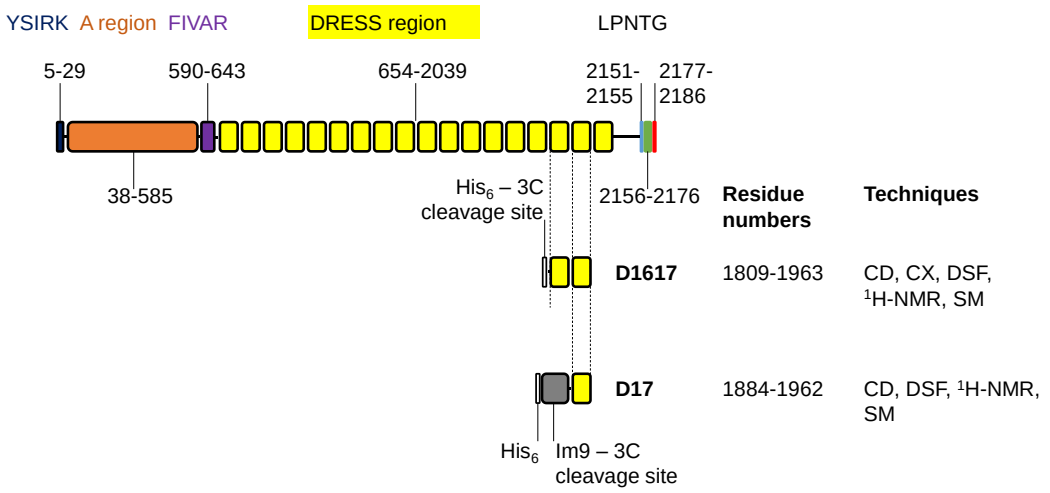


Figure 3.6: Schematic of SasC with regions indicated with residue numbers. Protein targets used in this chapter are shown. Domains are scaled to relative size. Techniques: CD: circular dichroism, CX: crystallography, DSF: nano-DSF, SM: SEC-MALLS

3.3.4 Molecular biology, recombinant gene expression, protein over-production and purification of single and tandem DRESS domains

The DNA sequences encoding single and tandem DRESS domains were amplified from genomic DNA purified from the lab strain *S. aureus* NCTC 8325-4 by PCR (see section 2.2.6) and inserted into linear pET plasmids (see section 2.2.9). Recombinant protein production conditions were screened for optimal production of single and tandem DRESS domains in BL21-Gold (DE3) cells. Typically, cells were grown at 37 °C (120 rpm) in auto-induction media, until an OD_{600} of 0.6 was reached, after which the temperature was lowered to 20 °C (180 rpm) for 16-24 hours. Successful over-production of D17 (see below) required the use of the fusion tag immunity protein 9 (Im9)[384]. All recombinant proteins produced in this chapter were recombinantly expressed, produced and purified using a His_6-tag that was subsequently proteolytically removed (see section 2.3.5). An example of a test for optimal protein over-production conditions and subsequent purification is described below for His_6-Im9-D17.

A test for the optimal recombinant gene expression and protein production conditions for D17 was performed in LB medium with induction with 1 mM IPTG at an OD_{600} of 0.6 or in auto-induction medium at 37 °C or 15-20 °C. The fusion tags glutathione S-transferase (GST)[385] or Im9[384] linked to an N-terminal His_6-tag were tested (Figure 3.7). A band of approximately the expected size for His_6-GST-D17 (theoretical MW of 35.2 kDa) appeared most clearly at 20 °C. The over-production of the target protein was most pronounced in

auto-induction medium, which also yielded higher cell densities after incubation for 16 hours (data not shown). A band of approximately 3 kDa larger than expected for His_6-Im9-D17 (theoretical MW of 23.1 kDa) was observed at both 15-20 °C and 37 °C with the most over-produced target protein in auto-induction medium. Hence, the test for the optimal recombinant gene expression and protein production conditions for D17 yielded the Im9 fusion tag in auto-induction medium with incubation at 20 °C for 16 hours as optimal conditions.

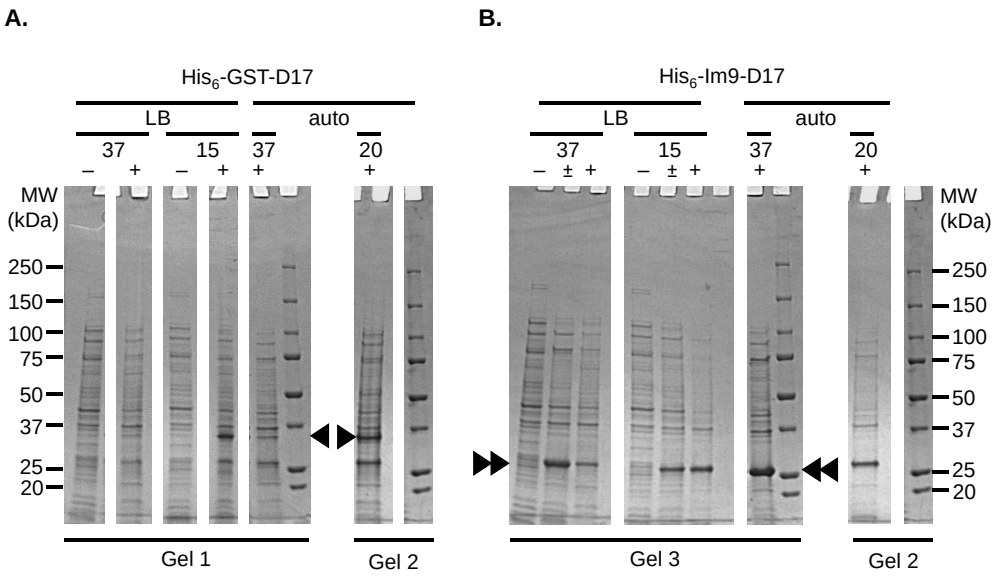


Figure 3.7: Test for optimal conditions of recombinant over-production of D17. Run on 12% (*w/v*) polyacrylamide gels (Bio-Rad). Non-assembled, raw gel images are available in Appendix 7.3. Brightness of gel 1 was adjusted by +20% post-acquisition. Total fractions are shown. Conditions are shown at the top of the gels. Media: LB or auto-induction. Temperatures: 37, 20 or 15 °C. -, ±, +: 0, 4 and 16 hours after IPTG-induction or OD_{600} = 0.6. Resuspension volume was diluted according to OD_{600} except for ±. **A.** His_6-GST-D17 (theoretical MW of 35.2 kDa, single arrow) and **B.** His_6-Im9-D17 (theoretical MW of 23.1 kDa, double arrow).

His_6-Im9-D17 was purified from the soluble cell extract by immobilised metal affinity chromatography (IMAC) using chelated nickel ions and was competitively eluted using an increasing gradient of Elution Buffer (see Table 2.4, section 2.3.4) containing an increasing imidazole concentration ranging from 20 mM (0%) to 500 mM (100%; Figure 3.8).

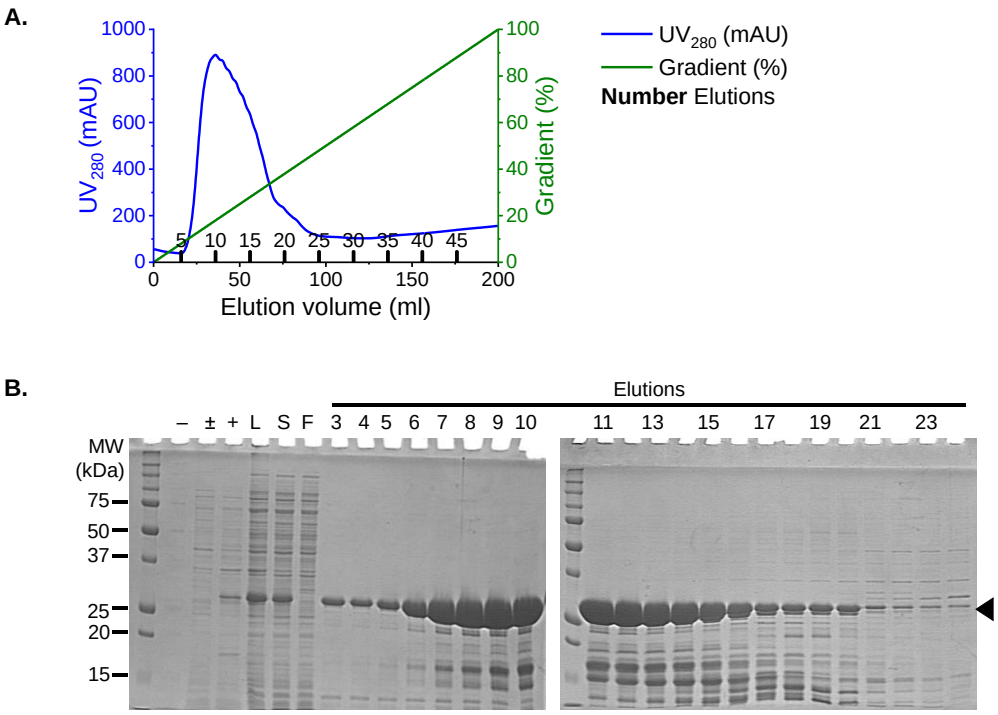


Figure 3.8: Purification of His_6-Im9-D17. A. IMAC chromatogram as monitored by A_{280} (blue). 4 mL fractions were collected (numbers shown in black). **B.** SDS PAGE analysis of purification of His_6-Im9-D17 (theoretical MW of 23.1 kDa, arrow) on 12% (*w/v*) polyacrylamide gels (Bio-Rad). Brightness adjusted by +40% post-acquisition. -: total fraction at OD_{600} = 0.05 (1 hour), ±: total fraction at OD_{600} = 0.4 (3 hours), +: total fraction at OD_{600} = 16 (18 hours). L: lysate. S: supernatant after centrifugation. F: flow-through. Numbers correspond to A.

The His_6-Im9 fusion tag was cleaved from D17 using HRV 3C protease (see section 2.3.5). The concentration of enzyme required for efficient cleavage was determined by testing different mass ratios of protease to target protein over a time course (data not shown). A mass ratio of protease to target protein of 1:150 was selected and His_6-Im9-D17 was incubated with HRV 3C protease for 18 hours at 4 °C (Figure 3.9A). The His_6-Im9 fusion tag was separated from D17 by a second round of IMAC (Figure 3.9B); D17 was in the flow-through and the His_6-Im9 tag and HRV 3C protease were competitively eluted by a stepwise increase in the concentration of imidazole ranging from 20 mM (0%) to ~260 mM (50%).

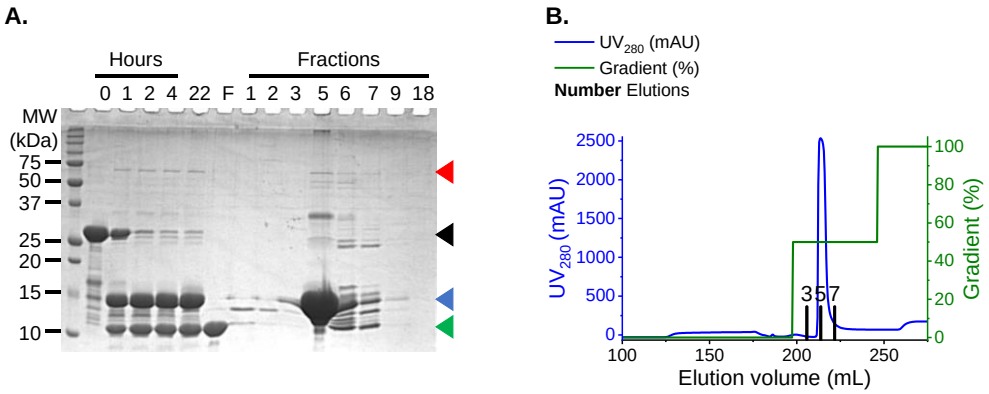


Figure 3.9: Separation of the His_6-Im9 fusion tag and HRV 3C protease from target protein D17. Analysed on 15% (*w/v*) polyacrylamide gels. Brightness was adjusted by +20% post-acquisition. **A.** SDS PAGE analysis of HRV 3C protease cleavage of His_6-Im9-D17 (black arrow, theoretical MW of 23.1 kDa). Mass ratio HRV 3C protease (red arrow, with non-cleavable His_6-MBP tag, theoretical MW of 64 kDa) to His_6-Im9-D17 1:150. D17 (green arrow, theoretical MW of 9.1 kDa) cleaved from His_6-Im9 tag (blue arrow, theoretical MW of 13.7 kDa) during 4 °C incubation for 26 hours. **B.** D17 was separated from the His_6-Im9 tag and HRV 3C protease by IMAC as monitored by A_{280}. 4 mL fractions were collected during a stepwise elution (numbered).

After the second round of IMAC purification, the target protein was approximately 90% pure as estimated from SDS PAGE analysis (Figure 3.9A; lane F). For crystallisation purposes, D17 was further purified by SEC on a Superdex 75 16/600 column (GE Healthcare Life Sciences (Figure 3.10)). At an elution volume of approximately 8-10 mL, before the void volume around 40 mL, a species absorbing at 280 nm eluted. The nature of this species is unknown, but it does not affect the purification of D17. Around 75 mL, D17 was eluted as determined from an observation of a single band of the correct approximate MW by SDS PAGE analysis and fractions 24-25 were concentrated by spin filtration (MWCO 3 KDa, Sartorius).

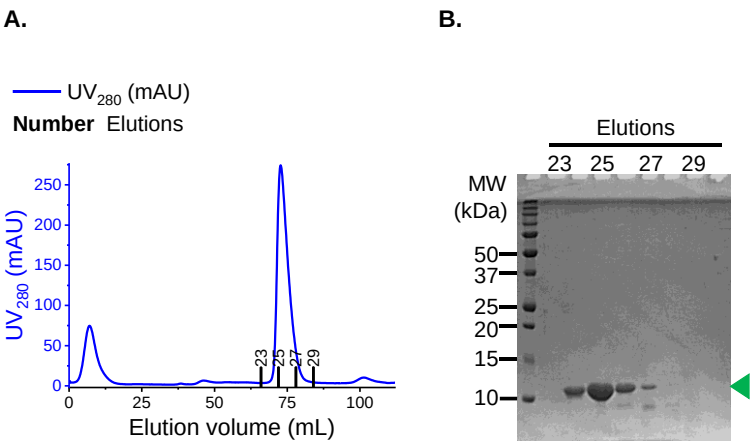


Figure 3.10: Purification of D17 by SEC. A. SEC chromatogram of D17 on a Superdex 75 column as measured by A_{280}. **B.** SDS PAGE analysis of SEC purification of D17 (green arrow) on 15% (*w/v*) polyacrylamide gel.

D1617 and D1617 with mutations across the interface (see section 3.4.3) were expressed and purified (Figure 3.11) in the same manner. The final yield of the recombinant proteins is shown in Table 3.2. The correct MW of recombinant proteins was confirmed by ESI MS (see Table 3.3).

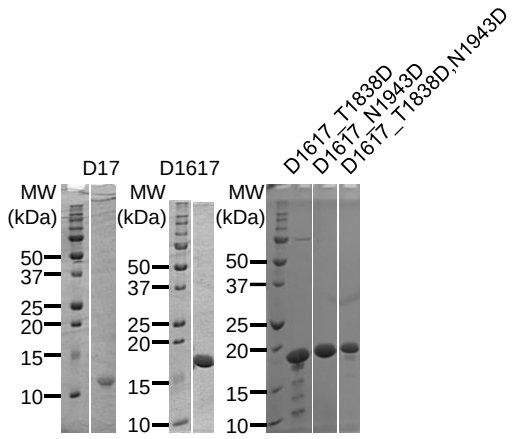


Figure 3.11: SDS PAGE analysis of purified recombinant proteins used in this chapter from assembled images of 15% (*w/v*) polyacrylamide gels. Brightness of right gel was adjusted by +20% post-acquisition. Raw images available in Appendix 7.3.

Table 3.2: Final yields of recombinant proteins used in this chapter. *: accuracy of mass determination was low due to low extinction coefficient. Yield is displayed in mg of purified target protein per litre medium. All proteins were produced in auto-induction media (see Table 2.4). IMAC 2 refers to the second round of IMAC to remove the His_6-tag and HRV 3C protease from the target protein. Isoelectric point (pI) and extinction coefficient (ε) were determined by ExPASy ProtParam[253].

Protein	Yield (mg L^{-1})	Last purification step	pI	ε (M^{-1} cm^{-1})
D17	24*	SEC	4.72	1490
D1617	76	SEC	5.08	2980
D1617_ T1838D	87	IMAC 2	4.97	2980
D1617_ N1943D	88	IMAC 2	4.97	2980
D1617_T1838D, N1943D	27	SEC	4.87	2980

3.3.5 Determination of the oligomeric state of single and tandem DRESS domains

Repetitive regions of some biofilm-mediating proteins are suggested to dimerise under specific conditions as a means for biofilm formation or accumulation. For example, parts of the repetitive region of SasG[61] and Aap[386,387] dimerise in the presence of Zn^{2+} and the B region of SdrC contributes to cell-cell interactions[147]. Here, the oligomeric state of single and tandem DRESS domains in solution was investigated by SEC-MALLS (for more information regarding the technique, please refer to section 2.6.1).

Briefly, DRESS domains were eluted over a Superdex 75 16/600 column (GE Healthcare Life Sciences) gel filtration column at 7-10 mg/mL in 25 mM MES, 150 mM NaCl, pH 6.0; except D17, which was analysed at 5 mg/mL in 20 mM Tris, 150 mM NaCl, pH 7.5. The elution was monitored by detectors for static light scattering, QELS, UV absorbance and refractive index increments. The MW was calculated from the static light scattering according to Equation 2.3. The R_h was estimated from QELS according to Equation 2.4.

All single and tandem wild-type DRESS domains tested are monomeric, based on the calculation of the molar masses (Figure 3.12). Around 0.6% of D1617_N1943D elutes at a dimeric molar mass, as estimated from the normalised DRI signal.

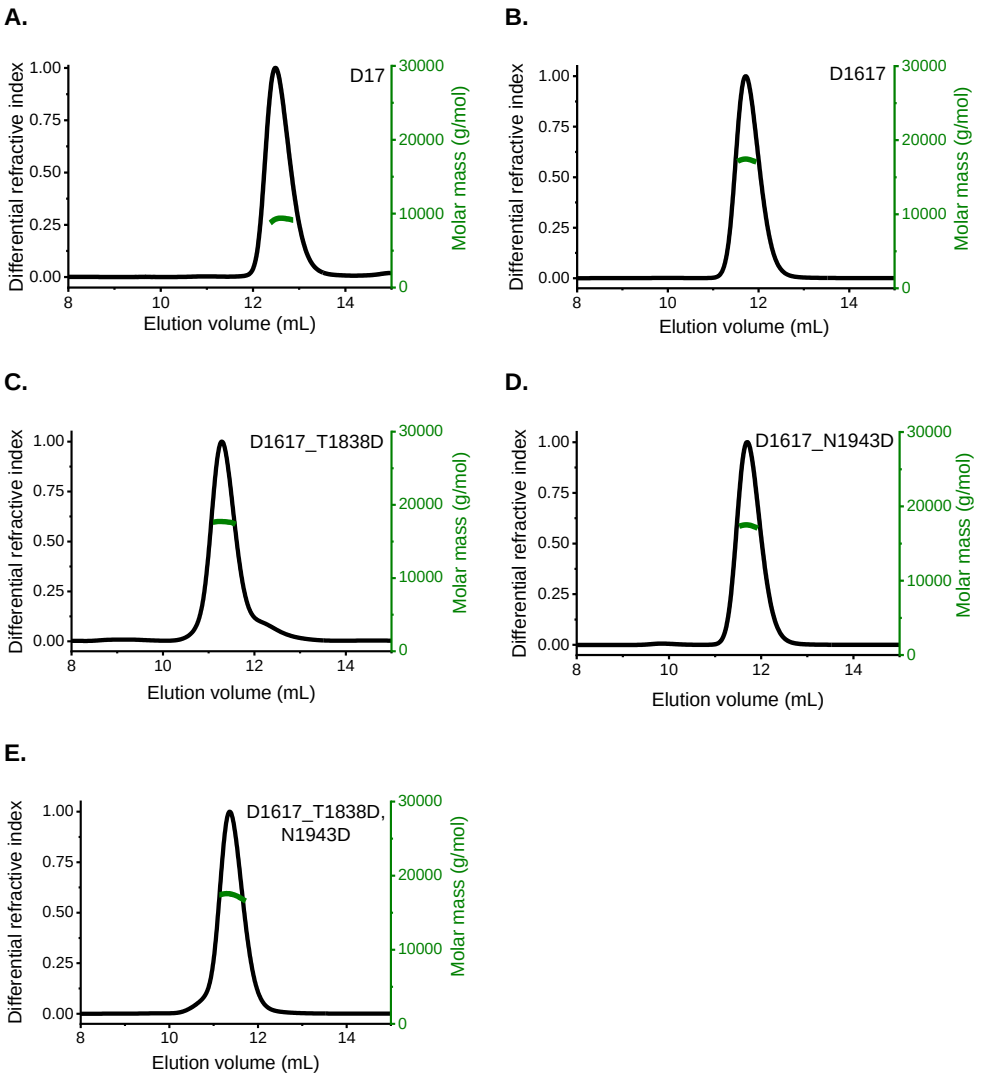


Figure 3.12: SEC-MALLS analysis of single and tandem DRESS domains on a Superdex 75 column. Elution profile in DRI (normalised) is shown in black and the molar mass estimate (g/mol) in green. The theoretical MWs are listed in **Table 3.3.** Data is shown for **A.** D17 at 3.24 mg/mL, **B.** D1617 at 9.38 mg/mL, **C.** D1617_T1838D at 4.99 mg/mL, **D.** D1617_N1943D at 5.41 mg/mL), **E.** D1617_T1838D,N1943D at 7.5 mg/mL.

The hydrodynamic radii are shown in Table 3.3. The determination of the R_h of D17 is inaccurate as shown by the large standard deviation, likely because D17 is only transiently folded at 20 °C (see section 3.3.6).

Table 3.3: Molar masses of single and tandem DRESS domains as determined from ESI MS and SEC-MALLS. Monoisotopic masses are shown unless stated otherwise. SEC-MALLS was performed in 25 mM MES, 150 mM NaCl, pH 6.0 unless stated otherwise. [a]Average theoretical MW. [b]In 20 mM Tris, 150 mM NaCl, pH 7.5.

Protein	**MW (Da)**			**R_h (nm)**
	Theoretical	MS	SEC-MALLS	
D17	[a]9106.1	[a]9105.9	[b]9164 ± 55	[b]3.0 ± 1.5
D1617	17810.3	17809.3	[b]17900 ± 72 17730 ± 17	[b]2.6 ± 0.1 2.5 ± 0.1
D1617_T1838D	17824.3	17823.2	17660 ± 35	2.8 ± 0.1
D1617_N1943D	17811.3	17810.3	17390 ± 17	2.4 ± 0.1
D1617_T1838D, N1943D	17825.3	17824.3	17350 ± 35	2.8 ± 0.1

3.3.6 Degree of folding of single and tandem DRESS domains

The degree of folding of single and tandem DRESS domains was compared by CD (see section 2.6.2) and ^{1}H-NMR spectroscopy (see section 2.7). The CD spectra of D17 and D1617 reveal typical spectra of α-helical proteins, as expected, with a characteristic positive band at 193 nm and the double negative bands at 208 and 222 nm[388] (Figure 3.13A). The spectra were fitted to spectra of proteins with known secondary structure to calculate the percentage of α-helical secondary structure in both samples using the Dichroweb[271] server and the CONTINLL algorithm[275] (see section 2.6.2.3). D1617 contains 93% α-helical secondary structure and D17 80%.

The degree of folding of D17 and D1617 was subsequently analysed by ^{1}H-NMR spectroscopy (Figure 3.13B). The "methyl region" of the spectrum of D1617 (<0.5 ppm) contains shielded signals that are shifted towards lower δ. Furthermore, the amide proton region of D1617 has larger dispersion, indicating that more amide protons are in a unique environment than the amide protons in D17. Altogether, this suggests that D1617 has a more stable fold than D17.

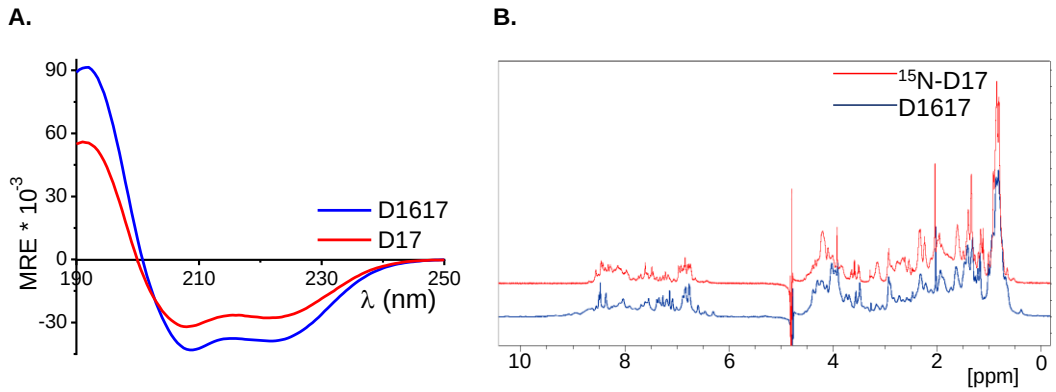


Figure 3.13: Comparison of the degree of folding of D17 and D1617. A. CD spectra of D17 (red) and D1617 (blue) at 22 µM in 20 mM sodium phosphate buffer, pH 5.5 at 20 °C. The signal is reported in MRE in degrees $cm^2\,dmol^{-1}\,res^{-1}$. **B.** ^{1}H-NMR spectra of ^{15}N-D17 (red) and D1617 (blue). Spectra were offset for clarity. Protein concentrations were 0.18 mM in 20 mM sodium phosphate buffer, pH 6.0. Recorded at 20 °C at 700 MHz.

3.3.7 Effect of temperature on the stability of DRESS domains

3.3.7.1 Nano-DSF

D17 and D1617 were thermally denatured from 15 °C to 95 °C at 1.3 °C/min. Protein unfolding was measured by monitoring the fluorescence emission ratio F350 nm/330 nm and simultaneously, aggregation was measured by static light scattering (Figure 3.14A, for more information regarding nano-DSF please refer to section 2.6.3).

D17 is only transiently folded at 15 °C, as indicated by the missing baseline of the unfolding transition, and has an estimated T_m of 34 °C (Figure 3.14A). D1617 is fully folded, as indicated by the stable baseline at temperatures below 40 °C, and has a T_m of 53 °C (Figure 3.14A). Furthermore, no aggregation was detected upon unfolding of DRESS domains. For clarity, a positive control for aggregation is shown (Figure 3.14B).

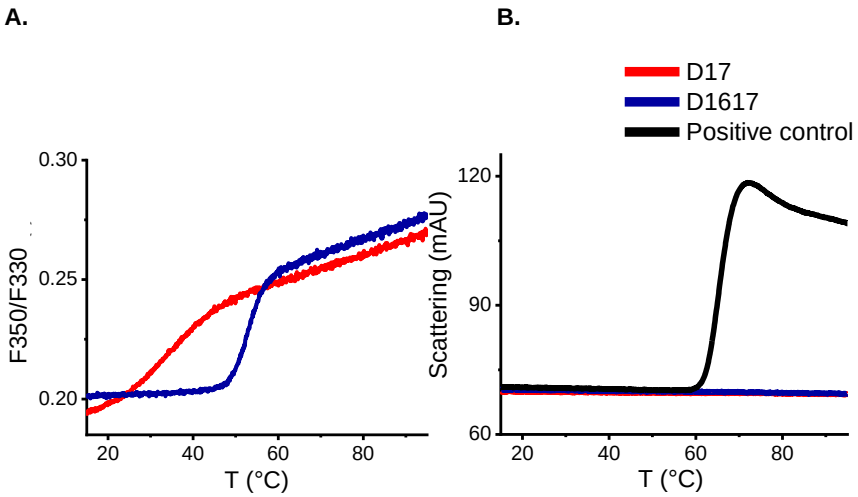


Figure 3.14: Thermal denaturation and aggregation of D17 and D1617 by nano-DSF monitored by **A.** fluorescence emission ratio and **B.** static light scattering. Measured on 1 mg/mL protein in 25 mM MES, 150 mM NaCl, pH 6.0 at a gradient of 1.3 °C/min. Positive control for aggregation is D1617 in 25 mM MES, 2 M NaCl, pH 6.0.

3.3.7.2 CD

The effect of thermal denaturation and renaturation on the secondary structure content of single and tandem DRESS domains was monitored by CD at 222 nm (Figure 3.15A,B); the signal at this wavelength is proportional to the α-helical content[389].

At 20 °C, D17 is 80% α-helical as determined using the Dichroweb[271] server and the CONTINLL algorithm[275] (see section 2.6.2.3) and unfolds between 20 °C and 40 °C. The absence of a flat baseline at low temperatures shows that D17 is only partly folded at 20 °C with an estimated T_m of 30 °C. In contrast, D1617 is 93% α-helical at 20 °C and shows a flat baseline up to 40 °C, followed by a steep unfolding transition with a T_m of 52 °C. The reversibility of thermal denaturation was around 90% for single and tandem DRESS domains (Figure 3.15C).

The T_m values of D17 and D1617 as determined by nano-DSF agree approximately with those determined by CD. The sigmoidal thermal denaturation curve, sharp unfolding transition and the presence of an isochromatic point in CD spectra of D1617 suggest a two-state cooperative unfolding pathway[390,391].

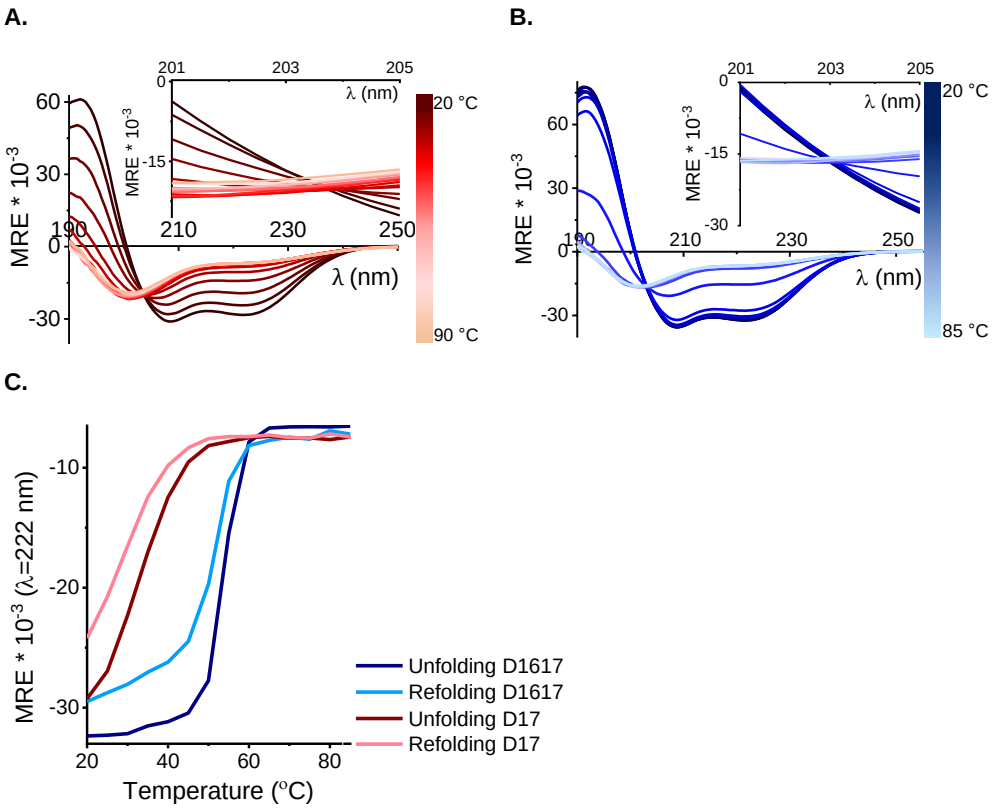


Figure 3.15: Thermal denaturation and renaturation of D17 and D1617 by CD. Superposed CD spectra of **A.** D17 at 22 µM and **B.** D1617 at 11 µM in 20 mM sodium phosphate buffer, pH 5.5 for temperatures between 20 °C and 95 ° at 5 °C steps. The insets show the intercepts of the denaturation curves. **C.** CD signal measured in MRE at 222 nm.

3.3.8 Effect of pH on the stability of DRESS domains

CD spectra were recorded of D17 and D1617 to determine the effect of pH on their stability (Figure 3.16A,B). Both D17 and D1617 show the highest percentage of α-helical secondary structure at pH 5.5, 88% in D17 and 97% in D1617. This drops at pH 7.0 to 61% and 88% helical content for D17 and D1617, respectively. The observation that DRESS domains are most stable at an acidic pH, matches with the pH of a biofilm matrix of ~5.0[392–394], putatively the physiological environment of DRESS domains on the surface of *S. aureus*.

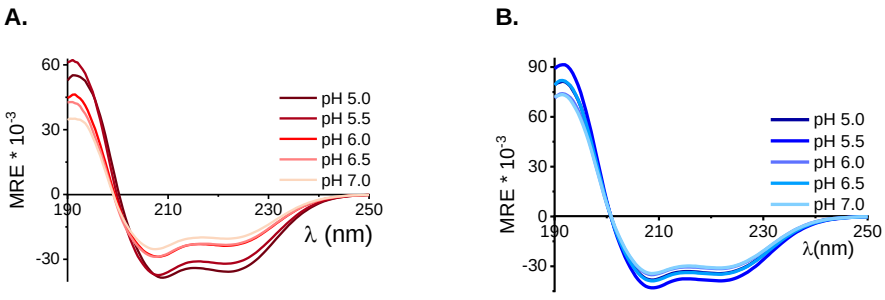


Figure 3.16: Effect of pH on D17 and D1617. A. D17 at 22 µM in 20 mM sodium phosphate, pH 5-7. **B.** D1617 at 11 µM in 20 mM sodium phosphate, pH 5-7. Spectra recorded at 20 °C.

3.3.9 Crystallography

3.3.9.1 Crystallisation of DRESS domains and structure solution of D1617

Crystallisation conditions were screened for D17 and D1617 (see section 2.8.2). D17 is more stably folded at 6 °C than at 20 °C (data not shown); therefore, crystallisation conditions were only screened at 6 °C, while D1617 was stable at both temperatures. Furthermore, the pH at which D17 and D1617 were most stable was pH 5.5 (see section 3.3.8), close to the putative pH of biofilms[392–394]. This might suggest that an acidic pH could be beneficial for crystallisation. However, good crystallisation conditions are difficult to predict and therefore, a wide range of conditions was screened (see section 2.8.2).

Crystals were obtained for both constructs (Figure 3.17A, B). D17 formed rectangular crystals with rounded edges within 24 hours at 6 °C in the PACT screen (Molecular Dimensions, pH 7.5, condition G11 in a 1:1 protein:reservoir drop). These crystals diffracted to ~2 Å, but the spots were elliptical and showed streaking (Figure 3.17C), implying the presence of multiple crystal lattices. A dataset of D17 was collected, but could not be solved by MR with ideal helices or a partial model of D1617 (see below).

D1617 formed large diamond-shaped diffracting crystals within 3 days at 6 °C in a sitting-drop setting in the JCSG-Plus screen (Molecular Dimensions, pH 4.5, condition A1 in 1:2 protein:reservoir drop). The initial crystallisation conditions for D1617 were 0.2 M lithium sulfate, 50% (*v/v*) PEG 400, 0.1 M sodium acetate pH 4.5 yielding crystals diffracting to 2.1 Å. Optimisation of precipitant and pH was performed in a hanging-drop setting, yielding a maximum resolution dataset of 1.62 Å from 0.2 M lithium sulfate, 50-52% (*v/v*) PEG 400, 0.1 M sodium acetate pH 4.9. The diffraction data of D1617 (Figure 3.17D) were anisotropic with an estimated resolution of 1.6 Å along *c* and 2.0 Å along *a* and *b*[395].

A bromide soak was performed on D1617 crystals in order to solve the phase problem by SAD or MAD (see section 2.8.1), but bromide ions were not stably incorporated into the crystal lattice. Then, MR using the program Phaser[304] was attempted to solve the phase problem using ideal helices kindly provided by Dr Huw Jenkins as search models. One to three ideal helices ranging in length from 6 to 15 residues were tested with settings as described in section 2.8.5, which were proposed by Dr Huw Jenkins; ideal helices of 13 residues were successful as a MR model to determine the crystal structure of D1617. This was followed by chain extension to generate a polyalanine model, together with

refinement of the phases using the polyalanine model in SHELXE[329] and Buccaneer[330]. Finally, model building was completed using the automated model building program ARP/wARP[331,396], followed by manual model completion in Coot[332], alternated with rounds of refinement using REFMAC5[333] and Phenix.refine[334,335]. The final model had an R_{factor}/R_{free} of 0.21/ 0.26 (Table 3.4).

D1617 crystallised with one molecule in the asymmetric unit of the unit cell. The ligands present in the asymmetric unit were 6 acetate ions and four molecules of PEG that were modelled by truncated versions of tetraethylene glycol. The presence of these ligands was justified by a reduction of R_{free} upon introduction of these ligands.

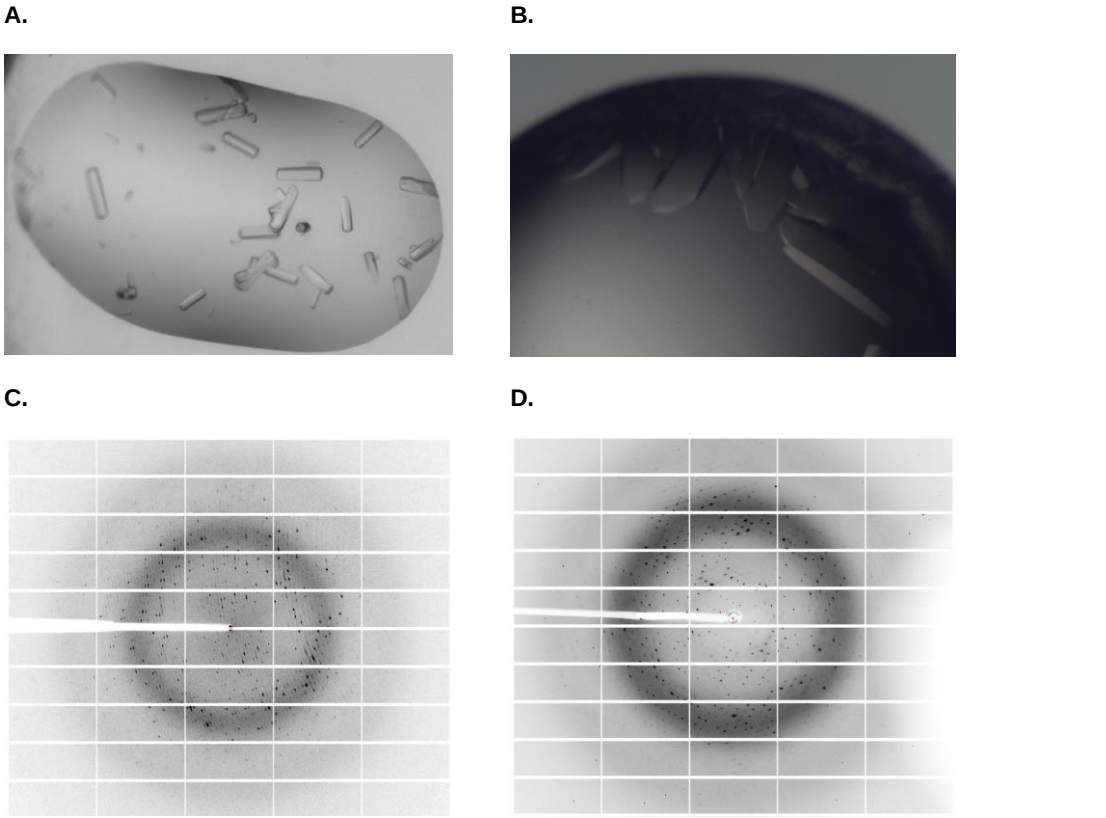


Figure 3.17: Crystals of DRESS domains. A. D17 in 0.2 M tri-sodium citrate, 20% (w/v) PEG 3350, 0.1 M Bis-Tris propane, pH 7.5, grown at 6 °C in sitting-drops (PACT, condition G11). **B.** D1617 in 50% w/v PEG 400, 0.1 M lithium sulfate, 0.1 M sodium acetate, pH 4.9, grown at 4 °C in hanging-drops. **C.** Diffraction pattern from a D17 crystal using 25% EG as cryo-protectant. **D.** Diffraction pattern from a D1617 crystal (no cryo-protectant required).

Table 3.4: Data collection and refinement statistics for D1617. Values in parentheses correspond to the highest resolution shell. RMSD: root mean square deviation. CC: correlation coefficient. For the formulas of R_{merge}, R_{pim}, R_{work}, R_{free}, $CC_{1/2}$, CC* please refer to section 2.8.1.

Parameters and statistics	**Data**
Data collection	
Beamline	I04, Diamond Light Source
Wavelength (Å)	0.97950
Resolution (Å)	53.16-1.62 (1.68-1.62)
Cell dimensions	
a, b, c, (Å)	53.89, 53.89, 323.78
α, β, γ, (°)	90 90 90
Space group	$I4_122$
Unique reflections	30492 (1415)
Completeness (%)	97.5 (93)
Multiplicity	24.9 (24.2)
I/σ(I)	21.9 (0.4)
R_{merge} (%)	0.059 (6.849)
R_{pim} (%)	0.012 (1.408)
$CC_{1/2}$	1.000 (0.34)
Wilson B-factor	33.744
Refinement	
Reflections used in refinement	30113 (2539)
Reflections used in R_{free}	1534 (135)
R_{work}	0.206 (0.569)
R_{free}	0.262 (0.647)
CC*	1.000 (0.797)
CC_{work}	0.957 (0.713)
CC_{free}	0.901 (0.576)
Number of atoms	
Macromolecules	1212
Ligands	64
Solvent	82
Protein residues	152
RMSD bonds (Å)	0.004
RMSD angles (°)	0.048
Average B-factor ($Å^2$)	73.88
Average B-factor of macromolecules ($Å^2$)	71.55
Average B-factor of solvent ($Å^2$)	88.21

Average B-factor of ligands (Å^2)	99.66

Ramachandran plot (%)	
Favoured	98.67
Allowed	1.33
Outliers	0

3.3.9.2 Structure of tandem DRESS domains

Tandem DRESS domains (D1617, Figure 3.18A) form two triple-helical bundles with an individual domain length of 4.2 nm and a diameter of 2.0 nm. The DRESS domains are organised in tandem in a head-to-tail arrangement, forming an elongated, rod-like structure. Each DRESS domain comprises three helices that are numbered 1 through 3 and the subscript a and b refer to DRESS domains 16 and 17, respectively (Figure 3.18B). At the N-terminus of the crystallisation construct, residues remaining from the fusion tag (GPAM) and the two N-terminal residues from D16 (AT) were not visible. Furthermore, the C-terminal residue (H) was not visible in the electron density map.

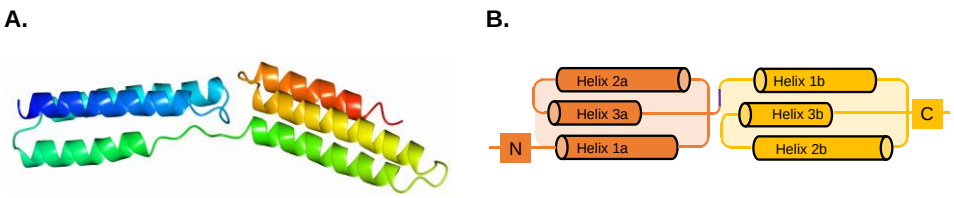


Figure 3.18: Crystal structure of D1617. A. Ribbon diagram of D1617, rainbow-coloured from N- to C-terminus in blue through red. Image was created using CCP4mg. **B.** Topology diagram of D1617 with D16 in orange, D17 in yellow and the linker in purple.

3.3.9.3 Superposition of D16 and D17

The average pairwise sequence identity between D16 and D17 is 29%. However, the C_α RMSD between D16 and D17 is only 1.19 Å over 77 C_α atoms (Figure 3.19); thus, as expected, the DRESS domains are structural repeats.

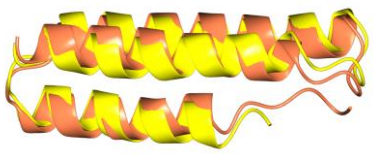


Figure 3.19: Superposition of D16 (orange) with D17 (yellow). Image was created using CCP4mg.

3.3.9.4 Tilt and twist angles between DRESS domains

The twist (Figure 3.20A,B) angle between tandemly arrayed domains was estimated using the secondary structure matching (SSM) tools in Coot[332] and CCP4mg[336], where the polar κ angle of the transformation matrix of the superposition in Coot was in agreement with the twist angle reported by CCP4mg. As it was difficult to find a quantitative, reproducible tool to calculate the tilt angles between tandemly arrayed domains, they are currently estimated manually by empirical determination (see Figure 3.20A) and quantitative determination via a custom script, courtesy of Emanuele Paci, University of Leeds, is in progress.

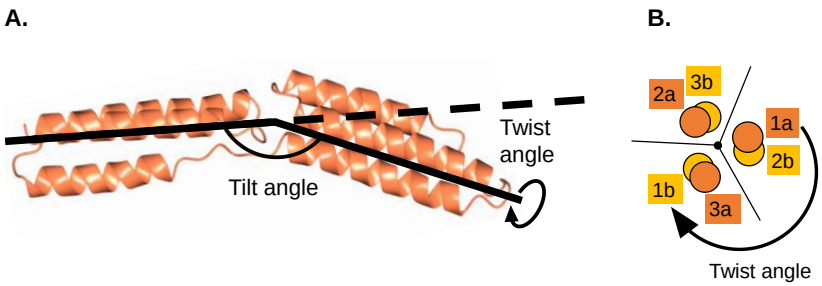


Figure 3.20: Relative rotations within D1617. A. Tilt and twist angle. Image was created using CCP4mg. **B.** Schematic of the twist angle, viewed from the side of D1617.

The twist angle between D16 and D17 was 139.2° and the tilt angle was 154°. In absence of solution studies, it is currently unclear if the tilt angle between D16 and D17 is physiological or if it is imposed by packing of D1617 molecules into a crystal lattice. Since the linker between DRESS domains is only three residues (see section 3.3.9.5), it seems likely that there would be insufficient space for a 180° tilt angle.

The twist angle is 19° larger than a three-fold symmetry axis (360°/3 = 120°, see Figure 3.20B). Assuming that the twist angle between other DRESS domains is also 139°, the B region of SasC comprising eighteen DRESS domains would twist round ~7 times. The overall topology of a repetitive region with a tilt and twist angle between adjacent domains might feature a spring-like architecture with increased extension and rigidity, as in the "twisted rope" model[387,397,398] (see Discussion in section 6.1.4.2).

3.3.9.5 Linker between DRESS domains

DRESS domains are connected via a short non-polar linker sequence, Pro1885-Ile1887 (Figure 3.21). The potential for a stabilising effect from the linker is two-fold. Firstly, the

linker includes Pro1885, which shows more than 75% sequence conservation across the 18 DRESS domains of SasC and lies two residues proximal of Ile1887 in 16 of 18 DRESS domains. Proline residues provide structural rigidity[399] and this is likely to restrict the positional flexibility of adjacent DRESS domains. Secondly, the linker residues are non-polar and form stabilising van der Waals interactions with non-polar atoms in the loop between helices 1 and 2 and within helix 2. Both effects are likely to promote and stabilise the desired linker orientation and lock DRESS domains in an elongated head-to-tail arrangement.

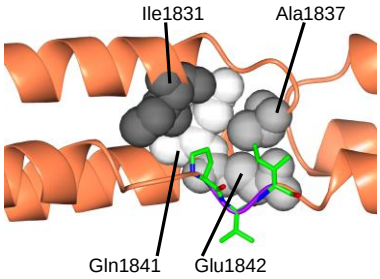


Figure 3.21: Details of the linker region of DRESS domains (Pro1885-Ile1887, backbone in purple and side chains in green) between D16 and D17. Non-polar atoms from Ile1831, Ala1837, Glu1841, Gln1842 in grey (from dark to light); they interact with linker residues. Image was created using CCP4mg.

3.3.9.6 The hydrophobic core

DRESS domains have two to three conserved hydrophobic residues per helix. Their placement along the helix ensures that the non-polar side chains point towards the core of the triple helical DRESS bundle, forming a network of van der Waals interactions. These conserved non-polar side chain residues pack along the length of the bundle (Figure 3.22).

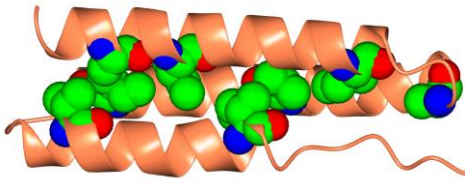


Figure 3.22: Details of the core DRESS domains containing van der Waals interactions between conserved residues. Image was created using CCP4mg.

3.4 The interface between DRESS domains

3.4.1 The size of the DRESS domain interface

The area buried between DRESS domains (BSSA) is 503 Å^2 per DRESS domain, or ~10% of the ASA of a DRESS domain (see section 2.8.6.1). Furthermore, the DRESS interface

contains twelve hydrogen bonds (donor-acceptor distance 2.5-3.7 Å, Figure 3.23A,B) and three salt bridges (2.8-3.5 Å, Figure 3.23C). Typically, domain interfaces in multi-domain proteins have one hydrogen bond per 100 $Å^2$, which extrapolates to a predicted number of 5 hydrogen bonds across the DRESS interface[400]. The BSSA for a selection of tandem repetitive domains (Table 3.5) is 450 $Å^2$ on average or ~6% of the ASA. Thus, compared to a small selection of other tandem domains, DRESS domains seem to have a relatively large interface, bury a large proportion of their ASA and have a greater-than-average number of interfacial hydrogen bonds.

57% of atoms containing the DRESS domain interface between D16 and D17 are polar in nature, in contrast to 33% in a selection of tandem repetitive domains (Table 3.5) and 38% in other multi-domain proteins[400]. The RSA of polar residues in tandem DRESS domains (D1617) is lower than that of the individual DRESS domains (D16 and D17), suggesting preferential burial of polar residues in the interface between two DRESS domains. Furthermore, the polar BSSA compared to the total or polar ASA of DRESS domains is larger than that of other tandem domains (Table 3.5). Thus, by all these measures, the polarity of the DRESS domain interface between D16 and D17 is higher than the average for selected tandem domain interfaces.

Despite their polar nature, only one fully buried, structured water molecule is observed in the interface region between DRESS domains D16 and D17. This might be consistent with the interface between DRESS domains being formed by water depletion[401] (Figure 3.23D).

Table 3.5: Properties of the interfaces of a selection of tandem domains. Analysed with Naccess[337]. BSSA: buried surface area (Equation 2.17A). RSA: relative surface area (Equation 2.17B). ASA: solvent accessible surface area.

PDB ID	Name	Residues	BSSA (Å²)	ASA (Å²)	% BSSA/ ASA	$RSA^{non-polar}/RSA^{all}$	RSA (polar)	RSA (polar)	% Polar BSSA	% Polar BSSA/ total ASA	% Polar BSSA/ polar ASA
		per domain				tandem	tandem	per domain	per domain		
Tandem repetitive domains											
DRESS (this thesis)		77	503	5168	10	1.17	29	34	57	3.1	7.2
1quu	Spectrin	124	422	8269	5	1.17	30	32	33	2.7	2.1
3pgk	Yeast phospho-glycerate	208	1222	11627	11	1.09	26	29	32	1.9	4.7
1fnf	Fibronectin type 3	93	312	5338	6	1.08	33	36	18	0.55	1.2
SHIRT (structure courtesy of Dr F. Whelan)		83	85	5186	2	1.09	40	41	43	0.36	0.8
Tandem alternating domains											
2dgj	GA	69	252	4408	6	1.19	30	34	47	1.6	3.6
	FIVAR	53		3403	7			32			
4wve	G5	83	353	6435	5	1.09	46	46	40	1.5	3.7
	E	48		3730	9			49			

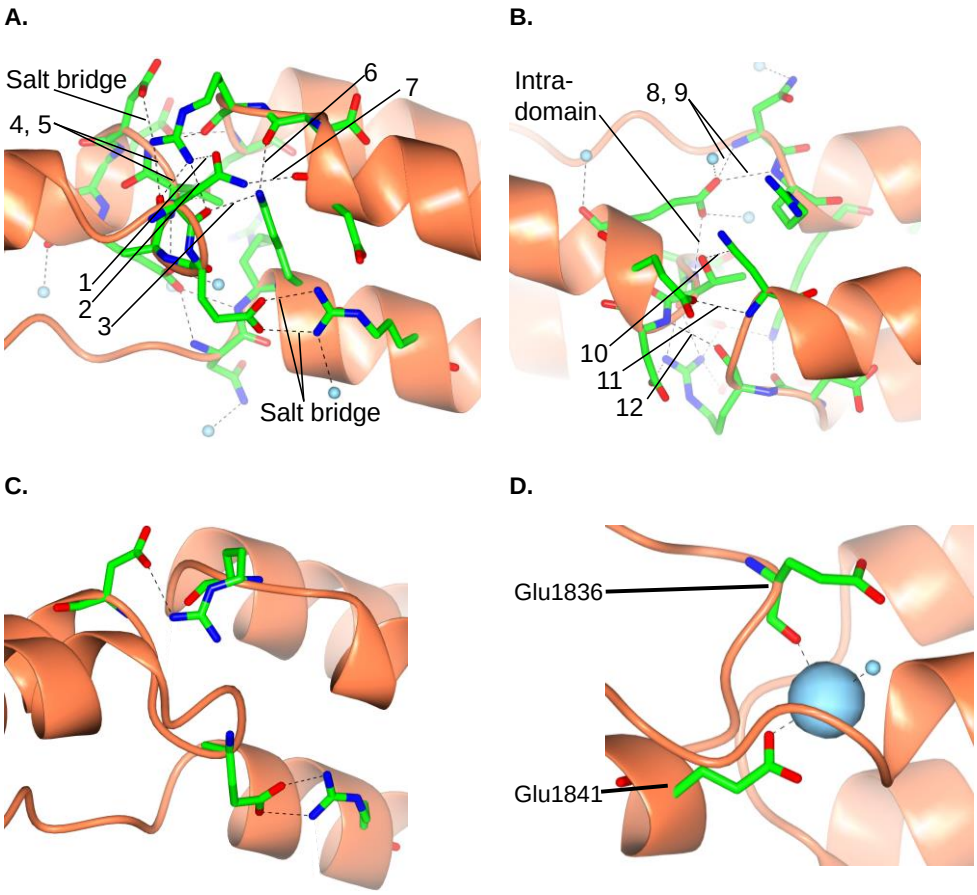


Figure 3.23: Details of the polar DRESS domain interface. A, B. Hydrogen bonds in the DRESS domain interface. **C.** Salt bridges in the DRESS domain interface. **D.** Structured water molecule (blue large sphere) buried in the DRESS domain interface. Small blue sphere: exposed water molecule. Image was created using CCP4mg.

3.4.2 *In silico* identification of key residues in the DRESS interface

Two software tools were employed to identify key residues contributing to the stability of DRESS inter-domain interfaces, PISA[338] and HotSprint[340] (see section 2.8.6.2). PISA predicted Thr1838 to be the most stabilising residue, followed by Ile1887 and Arg1941. The hydroxyl group in the side chain of Thr1838 donates a hydrogen bond to the carbonyl oxygen in the side chain of Asn1943 (Figure 3.24A) and is 94% conserved. A hydrophobic residue is conserved at the position of Ile1887 in 55% of DRESS domains in SasC. Arg1941, whose side chain makes hydrogen bonds with the backbone of Ala1837, is not conserved.

HotSprint determined that Val1946 (pair potential 21.37) was a hotspot residue and thus important for the stability of the interface. This valine is fully conserved across the DRESS domains of SasC and makes a Van der Waals interaction with the methyl group of Thr1838 (Figure 3.24B-D). Thr1838 was not solvent-accessible upon formation of the interface (Figure 3.24E), but the pair potential was too low (9.78) to be designated as a hot spot

residue. This might be because Val1946 has more atoms in close contact with interface residues than Thr1838.

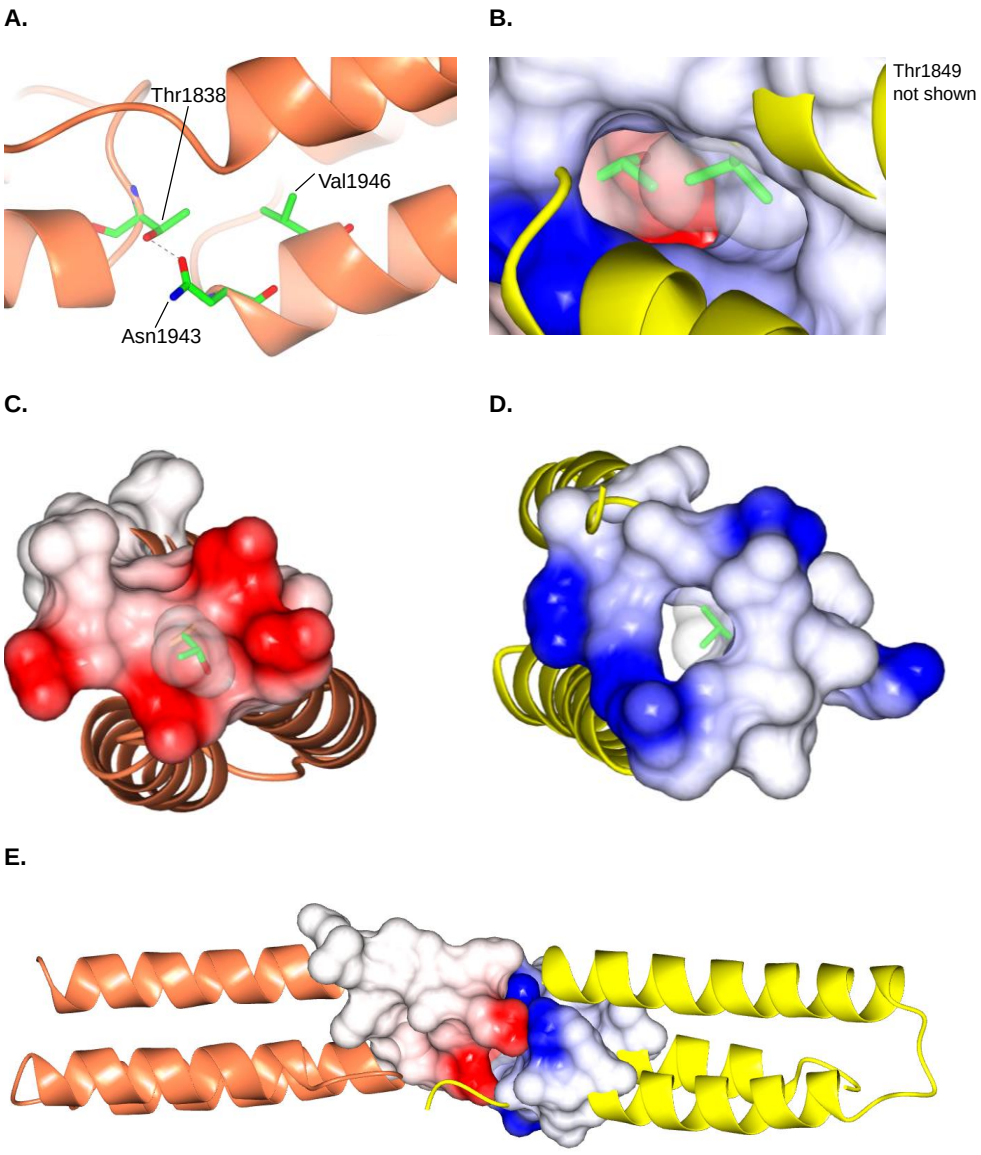


Figure 3.24: Stabilising interactions across the DRESS interface. A. The predicted most stabilising hydrogen bond between Thr1838 and Asn1943. Also shown is Val1946. **B.** Close-up of hydrophobic hotspot interaction between Thr1838 and Val1946. For clarity, surface representation of Thr1849 is not shown. **C.** Hotspot Thr1838, surrounded by the hydrophilic, negatively charged ring of D16. **D.** Hydrophobic hotspot Val1946, surrounded by a hydrophilic, positively charged ring in D17. **E.** D1617 with electrostatic surface representation of residues near the inter-domain interface. Image was created using CCP4mg.

3.4.3 Sequence conservation

The conserved residues in the MSA of DRESS domains are mapped onto the crystal structure of D1617 (Figure 3.25A). All conserved polar residues are located at the DRESS domain interface. This leads to a conserved negatively charged C-terminus (Figure 3.25B) and a positively charged N-terminus (Figure 3.25C), which aligns with the dipole moments

of helices 1 and 3 and creates an attractive electrostatic interaction between DRESS domains.

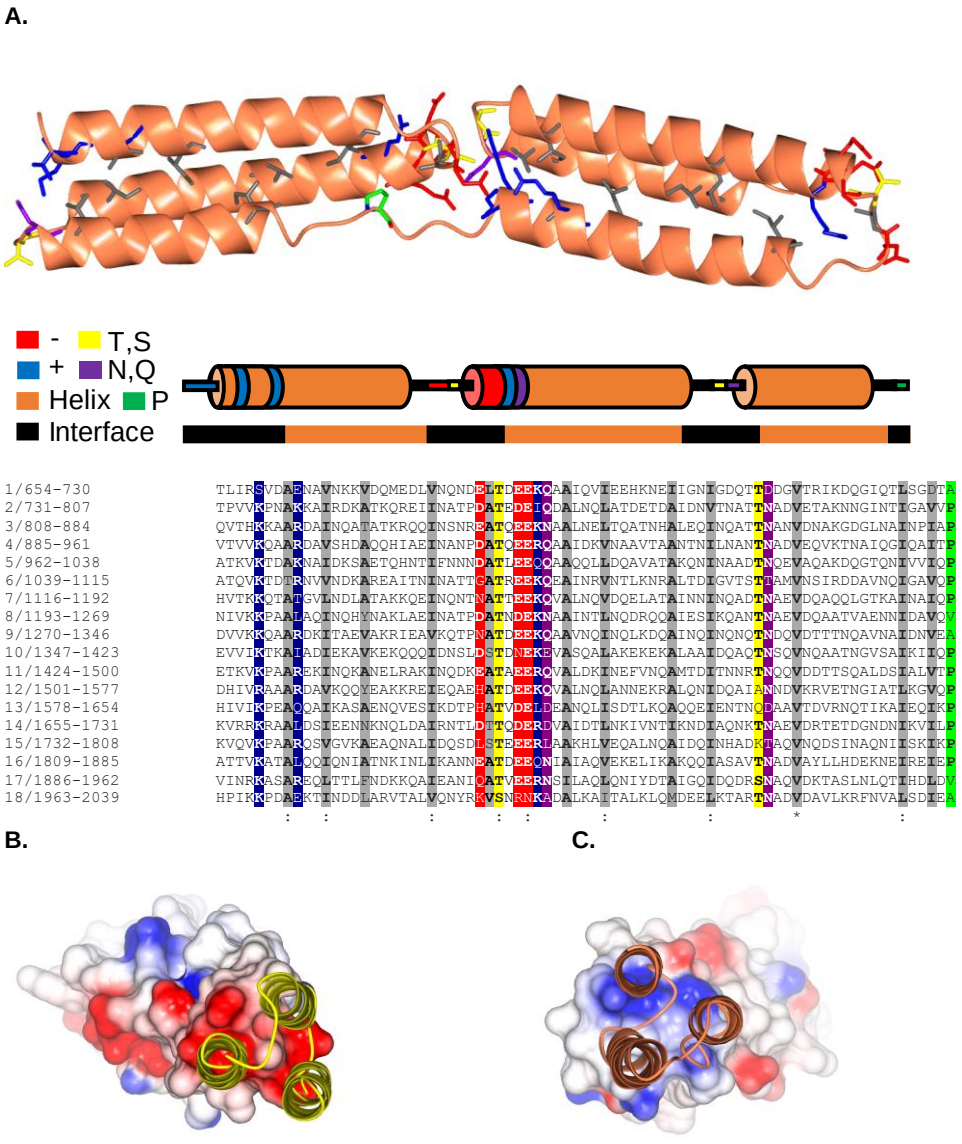


Figure 3.25: Features of the crystal structure of D1617.
A. Sequence conservation displayed on the crystal structure and the MSA of DRESS domains in SasC. Sequence similarity was calculated using Clustal Omega[259] for domains 1-18. Residues with more than 50% sequence conservation are highlighted and residues with conserved properties are shown in bold. Colour legend: blue: positively charged; red: negatively charged; yellow: nucleophilic purple: Asn, Asp; grey: hydrophobic. **B.** DRESS domain interface with a surface representation of D16 coloured by electrostatic potential; D17 shown as a yellow ribbon model. **C.** DRESS domain interface with a surface representation of D17 coloured by electrostatic potential; D16 shown as an orange ribbon model. Ribbon models and surface representations were created using CCP4mg.

3.4.4 Design of disruptive mutations in the interface between tandem DRESS domains

Based on the *in silico* analyses described in section 3.4.2, the hydrogen bond pair Thr1838 and Asn1943 was selected for mutation to Asp (see section 2.2.10). The Asn1943Asp mutation is predicted to still be capable of accepting a hydrogen bond from Thr1838, but the dipole moment of the hydrogen bonding will be larger. The more drastic Thr1838Asp mutation introduces a sterically larger side chain in the densely packed interface, removes the hydrogen bond donor and introduces a negative charge. D16717_T1838D, D1617_N1943D and D1617_T1838D,N1943D were expressed and purified as described in section 3.3.4. Their oligomeric state was determined by SEC-MALLS. D1617_T1838D and D1617_T1838D,N1943D were monomeric and D1617_N1943D was monomeric with 0.6% dimer (see section 3.3.5). Furthermore, SEC-MALLS analysis shows that D1617_T1838D and D1617_T1838D,N1943D have a larger R_h than D1617 or D1617_N1943D (Table 3.3), consistent with disruptions in the DRESS inter-domain interface, resulting in a less compactly folded tandem repeat.

3.4.5 Effect of disrupting the interface on DRESS stability

3.4.5.1 Temperature

Previously, a large T_m difference was observed between single and tandem DRESS domains (Figure 3.14A), attributed to the formation of a stabilising interface. Here, the T_m values of constructs with the mutations described above were obtained by nano-DSF (Figure 3.26, for the method please refer to section 2.6.3). The Thr1838Asp mutation and the double mutation containing the Thr1838Asp mutation had a significant effect on the T_m; it clearly reduces the T_m of tandem domains to that of a single domain. The number of replicates in which this experiment was performed was insufficient for statistical analysis.

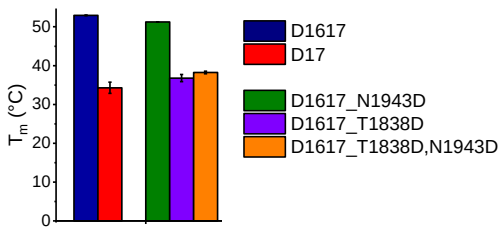


Figure 3.26: Effect of disruption of DRESS interface on domain stability. T_m of proteins in duplicate at 1 mg/mL in 25 mM MES, 150 mM NaCl, pH 6.0. Error bars show the error of the mean.

3.4.5.2 Ionic strength

The ionic strength affects the stability[402] and solubility[403] of proteins. Here, the effect of ionic strength on the stability of single and tandem DRESS domains and on tandem DRESS domains with disruptive mutations across the interface was tested using nano-DSF (see section 2.6.3; Figure 3.27).

At lower ionic strength, the thermal stability of D17 decreased with increasing ionic strength; however, the stability was recovered at higher ionic strength (Figure 3.27A). On the contrary, the thermal stability of D1617 increased with increasing ionic strength (Figure 3.27B). The conservative Asn1943Asp mutation follows the trend of D1617, while tandem DRESS domains with the disruptive Thr1838Asp mutation follow the trend of D17 (Figure 3.27C). This suggests that the inter-domain interface is influential in these experiments.

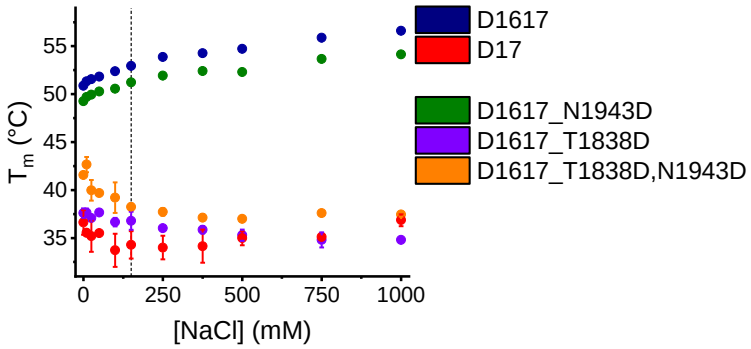


Figure 3.27: Effect of ionic strength on the thermal stability of single and tandem DRESS domains and DRESS domains with mutations across the interface. Dashed line is at 150 mM NaCl. Measurements were performed in duplicate at pH 6 on 1 mg/mL protein in 25 mM MES, error bars represent error of the mean.

3.5 Conclusions for this chapter

This chapter addresses the redefinition of the domain boundaries of DUF1542 domains, here renamed to DRESS domains, and their characterisation. *In silico* analyses allowed for a prediction of the correct domain boundaries by MSAs and secondary structure predictions. The determination of the crystal structure of tandem DRESS domains at 1.62 Å resolution confirmed the new domain boundaries. Furthermore, the crystal structure revealed a highly connected domain interface between two DRESS domains, that was shown to significantly increase the thermal stability of DRESS domains by CD and nano-DSF.

Although the *in silico* analyses are merely predictions regarding the accuracy of the domain boundaries, the crystal structure of DRESS domains provides definite proof of the redefined domain boundaries and the fold of DRESS domains. The resolution of the structure is sufficient to provide highly detailed information about, for example, the formation of the inter-domain interface. Furthermore, it informs that DRESS domains are structural repeats with a low C_α RMSD despite of their low sequence identity. This part of this chapter is ready for publication and especially the structure of DRESS domains can provide a helpful resource to others researching surface proteins of *S. aureus* or α-helical repetitive domains.

Thermal stability assays of single and tandem DRESS domains strongly suggest a stabilising effect from the highly connected interface on adjacent DRESS domains. This is probed by mutational studies, which showed that a T1838D mutation decreased the T_m of tandem DRESS domains to approximately that of a single DRESS domain; implying that the formation of the connective interface might be responsible for the higher thermal stability of tandem DRESS domains. Further experimental replicates of the thermal stability experiments are required to allow statistical analysis; furthermore, it is unclear what specific interactions might be formed by the introduction of a T1838D mutation within a highly polar interface. Therefore, this part of this chapter requires further work in the form of more thermal stability experiments and perhaps additional mutational studies; please refer to section 6.4.1. The effect of ionic strength on single and tandem DRESS domains is not fundamentally understood and is therefore not to be included in a publication. Finally, the hypothesis that DRESS domains are organised in a “twisted role” model is not tested and requires further investigation, preferably with the physiological repetitive region of SasC.

Chapter 4. Biophysical characterisation of the DRESS region

4.1 Introduction

S. aureus employs various CWA proteins to adhere to other bacteria, host cells or the surfaces of in-dwelling medical devices[30], in addition to the use of other factors such as extracellular DNA[160] and the secretion of a PNAG matrix[153]. Many of these CWA proteins contain a potentially rod-like repetitive region, which projects the putatively functional N-terminal domain(s) away from the cell wall[29,30].

For example, the repetitive region of SasG is highly extended through cooperative interactions across domain interfaces[57] and is predicted to have a length of 88 nm. An example of a CWA protein with an α-helical rod-like region is Ebh, whose repetitive region contains 52 repeats of FIVAR-GA modules[58]. Two of these repeats formed a rod-like structure with some flexibility and the length of the full repetitive region in extended state is estimated to be 320 nm, however full extension is unlikely due to the observed flexibility[58].

The N-terminal domain of SasC is involved in cell-cell interactions and contributes to biofilm accumulation[111], highlighting the need to understand the structure and function of SasC. So far, no function has been proposed for the repetitive region containing DRESS (new domain boundaries) or DUF1542 (old domain boundaries) domains. This chapter describes the biophysical characterisation of the putative rod-like DRESS region in SasC.

4.2 Aims

This chapter aims to characterise the biophysical properties of repetitive structural domains from the DRESS region of SasC. This information contributes to understanding the role of the DRESS region and may link structure to function. To fulfil these goals, the following aims were set:

- To produce and purify repetitive structural domains from the DRESS region in SasC;
- To determine the size and shape of recombinant repetitive structural domains from the DRESS region;

- To assess the rigidity of recombinant repetitive structural domains from the DRESS region;
- To infer the end-to-end distance of the entire DRESS region;
- To determine the force required to mechanically unfold individual DRESS domains and assess their refolding ability.

4.3 Results

4.3.1 *In silico* analysis of protein sequences containing tandem DRESS domains

A single DRESS domain is only transiently folded at 20 °C, while DRESS domains in tandem are stably folded at the physiological temperature of 37 °C (see section 3.3.7). This implies that, to form a rigid, rod-like structure, DRESS domains need to be organised in tandem. Here, *in silico* analysis is performed on protein sequences containing DUF1542 domains in the InterPro protein sequence database[404].

77% of all DRESS domain-containing protein sequences have a tandem domain organisation for DRESS domains (data not shown). Of proteins with an N-terminal YSIRK/GXXS signal peptide, SasC/Mrp/FmtB A region and C-terminal LPXTG wall attachment site, 100% of DRESS domains are found in tandem (Figure 4.1). These results suggest that DRESS domains in SasC/Mrp/FmtB might form a highly elongated, rod-like region to project the functional A region away from the cell wall.

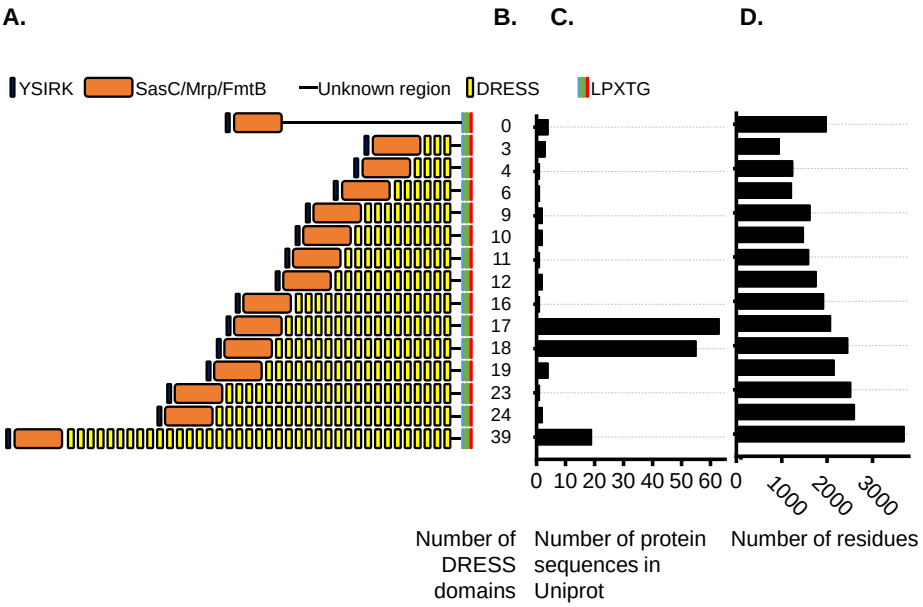


Figure 4.1: Number of DRESS domains in protein architectures in Uniprot containing the SasC/FmtB/Mrp A region. A. Schematic of domain architectures. **B.** Number of DRESS domains. **C.** Number of protein sequences in Uniprot per protein architecture. **D.** Number of residues per protein architecture.

4.3.2 Over-production and purification of recombinant repetitive structural domains from the DRESS region of SasC

Proteins with multiple DRESS domains from the repetitive region of SasC were expressed, over-produced and purified for biophysical characterisation of their size, shape, length and mechanical properties (Figure 4.2).

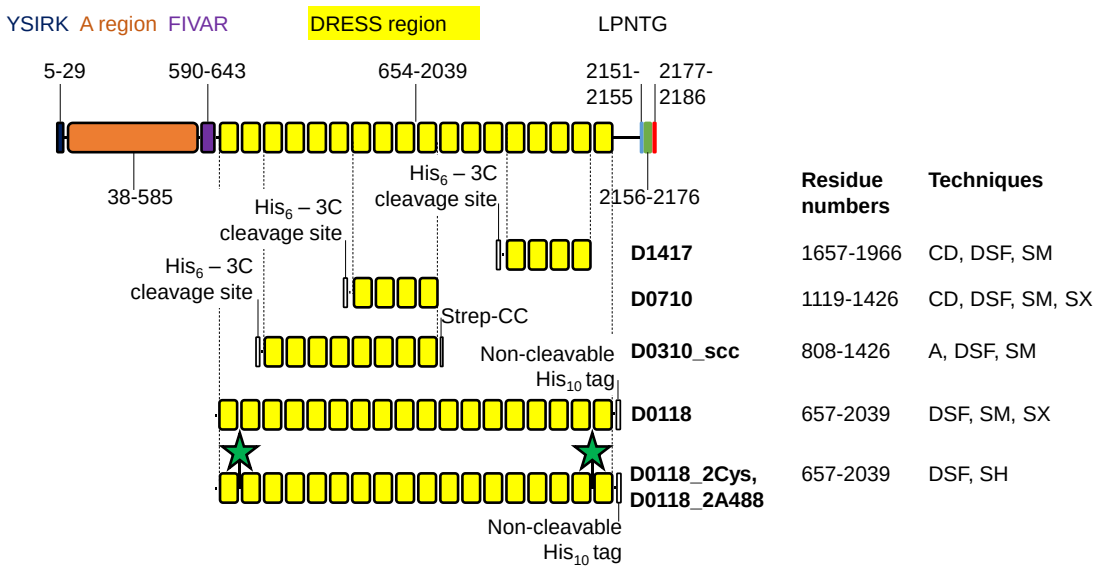


Figure 4.2: Schematic of SasC with regions indicated with residue numbers. D0310 is fused to a Strep tag[251] and two C-terminal cysteine residues. D0118_2Cys contains two cysteine residues for the covalent incorporation of A488 fluorophores using maleimide-based chemical modification (stars). Techniques: A: AFM; CD: circular dichroism; DSF: nano-DSF; SH: SHRImP-TIRFm; SM: SEC-MALLS; SX: SEC-SAXS.

4.3.2.1 D1417, D0710

Two four-domain DRESS protein constructs, D1417 and D0710, were produced with the purpose of determining their oligomeric state and studying their size and shape by SAXS. D0710 originates from the middle of the DRESS region and D1417 is located at the C-terminus of the DRESS region. These regions were selected for over-production and purification because of high sequence conservation in SasC between different *S. aureus* strains containing 18 DRESS domains (data not shown). In the same manner as described in Chapter 3, D0710 and D1417 were expressed in the pETFPP1 vector featuring an N-terminal His_6-tag, followed by a HRV 3C protease cleavage site.

Briefly, the recombinant proteins were over-produced in BL21-Gold (DE3) cells, cells were lysed and the soluble material was separated from the insoluble material by centrifugation and His_6-tagged proteins were purified by IMAC. The His_6-3C tag was removed by proteolytic cleavage, then the target proteins were separated from the His_6-tagged protease and purification tag by a second round of IMAC and finally purified by SEC (Figure 4.3). The yield of D1417 was lower than of D0710 and, in addition, D1417 eluted from the SEC column in two peaks of similar signal intensity, suggesting oligomerisation (Figure 4.3A). Therefore, fractions 16-19 and 20-22 were concentrated separately and the oligomerisation state was studied by SEC-MALLS (see sections 2.6.1, 4.3.3). The protein concentration of D0710 was underestimated due to a low ε (see Table 4.1), resulting in overloading of the SEC column (Figure 4.3C). Here, likely aggregated species eluted around fraction 32 and a single resolved peak with a shoulder eluted at fractions 50-70 (Figure 4.3D), of which fractions 49-57 were concentrated.

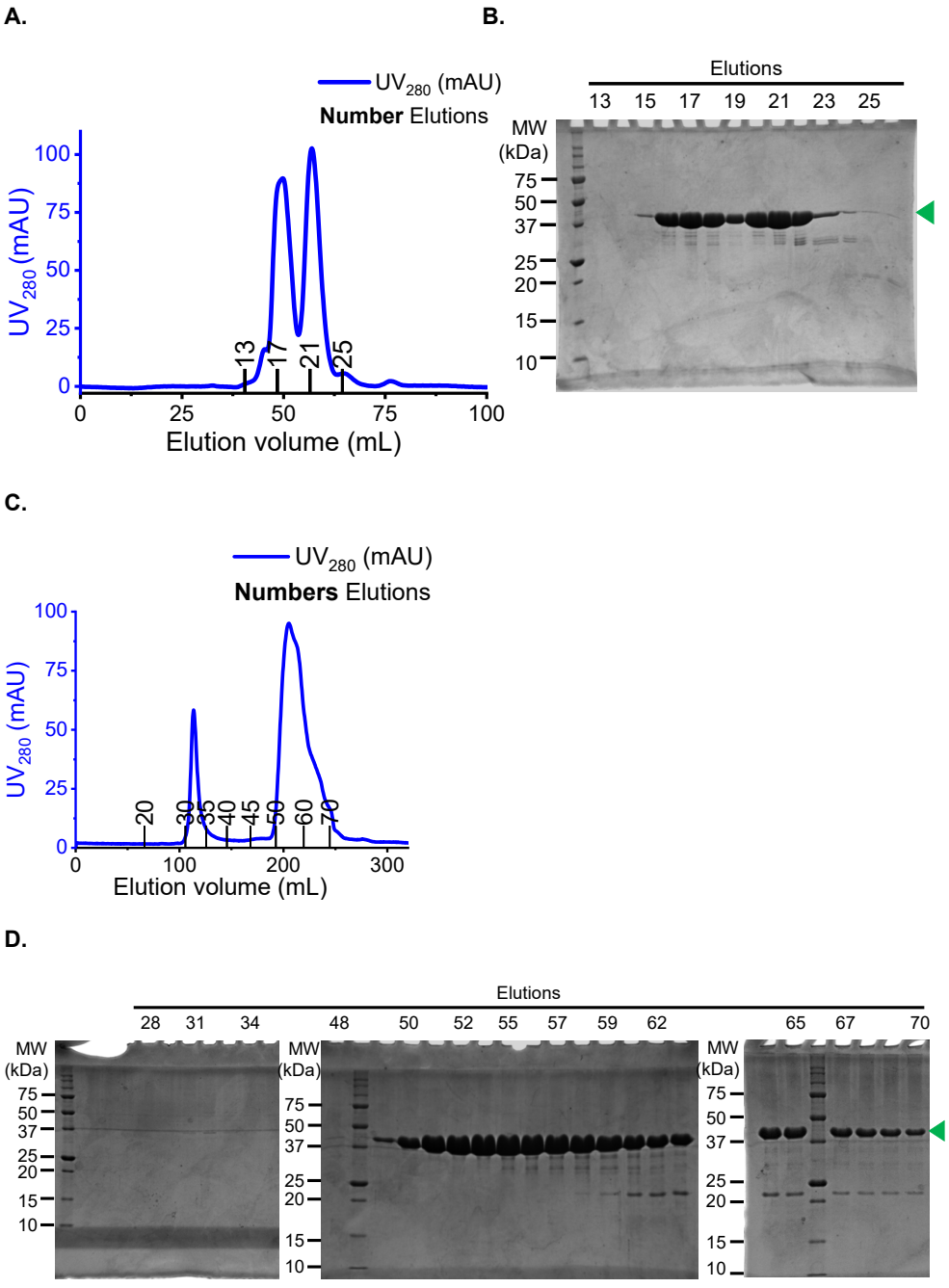


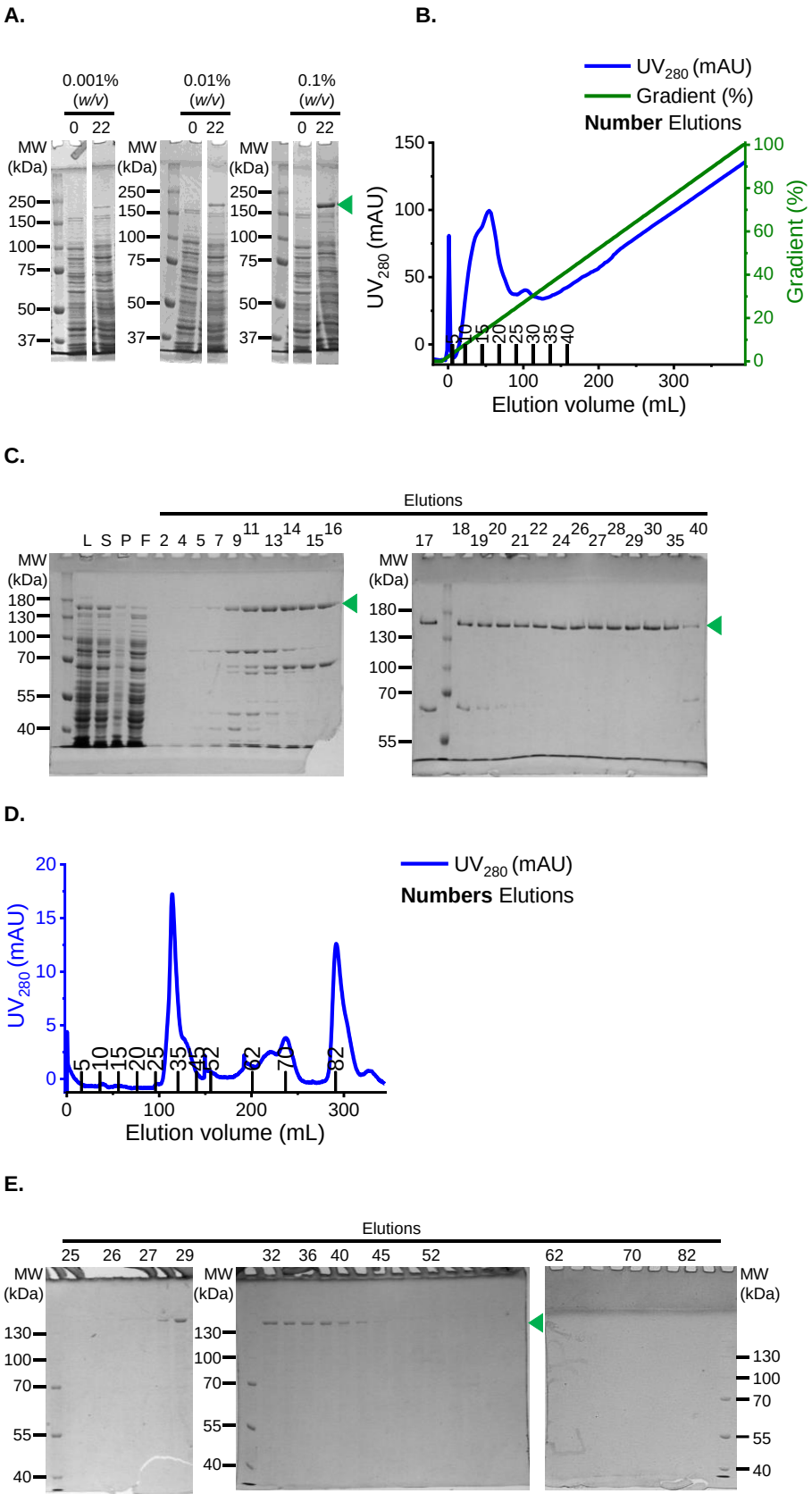
Figure 4.3: SEC purification of D1417 and D0710 analysed on 15% (*w/v*) polyacrylamide gels. **A.** SEC chromatogram of D1417 on a Superdex 75 16/600 column as measured by A_{280}. **B.** SDS PAGE analysis of SEC purification of D1417 (theoretical MW of 34.8 kDa, green arrow). Fractions correspond to A. **C.** SEC chromatogram of D0710 on Superdex 200 26/600 column as measured by A_{280}. **D.** SDS PAGE analysis of SEC purification of D0710 (theoretical MW of 33.7 kDa, green arrow). Fractions correspond to C. Brightness of left gel was adjusted post-acquisition by +20% and right gel by +40%. Raw gel images are available in Appendix 7.3.

4.3.1.1 D0118

The full physiological length of the repetitive region of SasC was produced with the aim of determining the oligomeric state and to study its size and shape by SAXS (for more detail, please refer to section 2.9). This protein construct, comprising eighteen DRESS domains,

has 1383 residues and a theoretical MW of 153.5 kDa. This MW is at the upper size limit for recombinant expression and protein production in *E. coli*, which is estimated to be 150 kDa[405,406]. A non-cleavable C-terminal His_{10}-tag was selected to allow purification of intact material. This tag was available in a pBADcLIC2005 vector, giving tight control over recombinant gene expression and protein over-production (see section 2.1.3).

Briefly, BL21-Gold (DE3) cells were grown at 37 °C (120 rpm) until an OD_{600} of 0.7 was achieved, after which recombinant gene expression and protein over-production was initiated by the addition of 0.1% (*w/v*) L-arabinose, followed by incubation at 37 °C (180 rpm). This concentration was found to be ideal during optimisation of the protein production conditions (Figure 4.4A). Cells were lysed and D0118 was purified from the soluble material by IMAC using an increasing gradient of buffer B (see Table 2.4) containing an imidazole concentration ranging from 20 mM (0%) to 500 mM (100%; Figure 4.4B,C). D0118 was further purified by SEC (Figure 4.4D,E) and fractions 28-43 were concentrated by spin filtration (MWCO 10 kDa) in portions 17-21 and 22-35. The A_{280} trace from the SEC chromatogram (Figure 4.4D) suggested a poor recombinant protein yield and some impurities in the fractions, for which the concentration was too low to be observed by SDS PAGE analysis (Figure 4.4E).



(Figure legend on next page)

Figure 4.4: Purification of D0118 on 8% (*w/v*) polyacrylamide gels with Precision Plus Protein marker (Bio-Rad, A) and PageRuler marker (ThermoScientific, C, E). Raw gel images are available in Appendix 7.3. **A.** SDS PAGE analysis of the optimisation of the concentration of L-arabinose used for induction of recombinant protein over-production. Soluble fractions are shown for protein production at 20 °C at 0 and 22 hours post-induction (green arrow, band likely corresponding to D0118 with a theoretical MW of 153.6 kDa). **B.** IMAC chromatogram as monitored by A_{280} (blue). 4 mL fractions were collected during protein elution (numbers). **C.** SDS PAGE analysis of IMAC purification of D0118. L: total lysate, S: soluble material, P: insoluble material, F: flow-through. Fractions correspond to B. Brightness adjusted by +40% and contrast by +20% of both gels post-acquisition. **D.** SEC chromatogram of D0118 on Superdex 200 26-600 (GE Healthcare Life Sciences) as measured by A_{280}. **E.** SDS PAGE analysis of SEC purification of D0118 of elutions shown in D. Contrast of all gels was adjusted by +20%; brightness of left and right gels was adjusted by +20% and of middle gel by +40% post-acquisition. Raw images available in Appendix 7.3.

4.3.2.3 D0118_2Cys, D0118_2A488

The end-to-end distance of a recombinant protein construct containing eighteen DRESS domains was studied by incorporating a fluorophore at both ends of the rod. The A488 fluorophores are linked to a maleimide functional group, which can be specifically coupled to the thiol group of a cysteine residue. Equivalent positions T731 and H1963 in the coil regions between domains 1-2 and 17-18 (Figure 4.2A) were selected for mutation to cysteine residues and fluorophore labelling, based on the following criteria:

- Minimal steric disruption: for this reason, residues with low sequence conservation and predicted to be solvent-accessible based on the crystal structure of D1617 (see section 3.4.3) were selected.
- Minimal terminal flexibility: residues between DRESS domains were selected, rather than the N- and C-termini of the protein. This lowered the inter-fluorophore distance, but also decreased the possibility of flexibility at the termini of the rods, which could complicate data analysis.
- Fluorophores need to be spatially separated from residues known to be able to quench fluorescence. Tryptophan and tyrosine are strong quenchers and histidine and methionine are weak quenchers[407].

Briefly, T731C and H1963C mutations were incorporated into the D0118 DNA sequence in the pBADcLIC2005 expression vector by site-directed mutagenesis (see section 2.2.10 and Appendix 7.1, Table 7.2), creating the gene encoding D0118_2Cys. The resulting vector was transformed into BL21-Gold (DE3) cells for recombinant protein over-production. Cells were grown in LB at 37 °C (120 rpm) to an OD_{600} of 0.6, followed by induction of recombinant gene expression and protein over-production by the addition of 0.1% (*w/v*) L-arabinose and incubation at 20 °C for 22 hours (180 rpm). D0118_2Cys was purified in

the same way as D0118 using IMAC in an increasing gradient of imidazole concentration ranging from 20 mM (0%) to 500 mM (100%). Fractions 25-35 were concentrated by spin filtration (MWCO 50 kDa) and purified by SEC on a Superdex 200 26/600 column (GE Healthcare Life Sciences) in buffers supplemented with 5 mM β-mercaptoethanol (see section 2.3.7, Figure 4.5A-D). The A_{280} trace from the SEC chromatogram suggested a poor recombinant protein yield of D0118_2Cys and the presence of many impurities, most at a concentration that was too low to be visualised by SDS PAGE analysis.

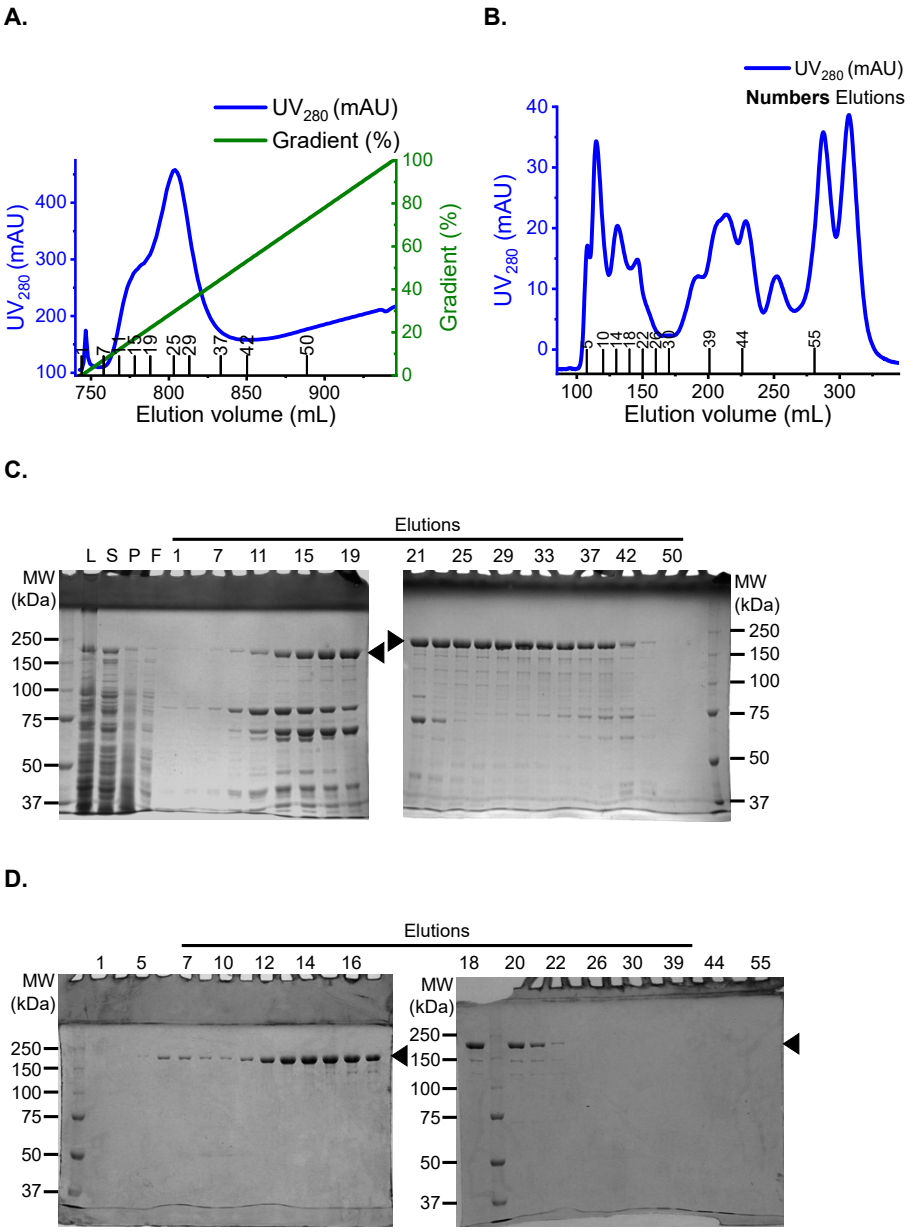


Figure 4.5: Purification of D0118_2Cys. SDS PAGE analyses on 8% (*w/v*) polyacrylamide gels. Raw gel images are available in Appendix 7.3. **A.** IMAC chromatogram as monitored by A_{280} (blue). 4 mL fractions were collected during protein elution. **B.** SEC chromatogram of D0118_2Cys as measured by A_{280}. **C.** SDS PAGE analysis of the IMAC purification of D0118_2Cys (black arrow, a theoretical MW of 153.5 kDa). L: total lysate, S: soluble material, P: insoluble material, F: flow-through. Fraction numbers correspond to A. Brightness of right gel was adjusted by +40% post-acquisition. **D.** SDS PAGE analysis of SEC purification of D0118_2Cys. Fraction numbers correspond to B. Brightness of right gel was adjusted by +20% post-acquisition.

Fractions 12-13 from SEC were concentrated by spin filtration (MWCO 50 kDa) and β-mercaptoethanol was removed by dialysis, directly followed by a stepwise labelling reaction with 25 molar excess (final amount) of A488-maleimide relative to the cysteine residues (see section 2.3.7.2, Figure 4.6A-B). After a dialysis step, D0118_2A488 was purified by SEC to remove aggregates, unlabelled protein and unreacted fluorophore

(Figure 4.6C-D), fractions 7-8 were concentrated by spin filtration (MWCO 10 kDa) and the labelling efficiency E (%) was calculated (Equation 4.1). A high labelling efficiency is preferred (where 100% labelling indicates that both cysteine residues are labelled), as only proteins with two fluorophores can be used to determine the end-to-end distance. A high proportion of proteins with a single fluorophore may introduce anomalous measurements in the end-to-end distance histograms if two surface-immobilised proteins with a single fluorophore are spatially close in the SHRImP-TIRFm experiments.

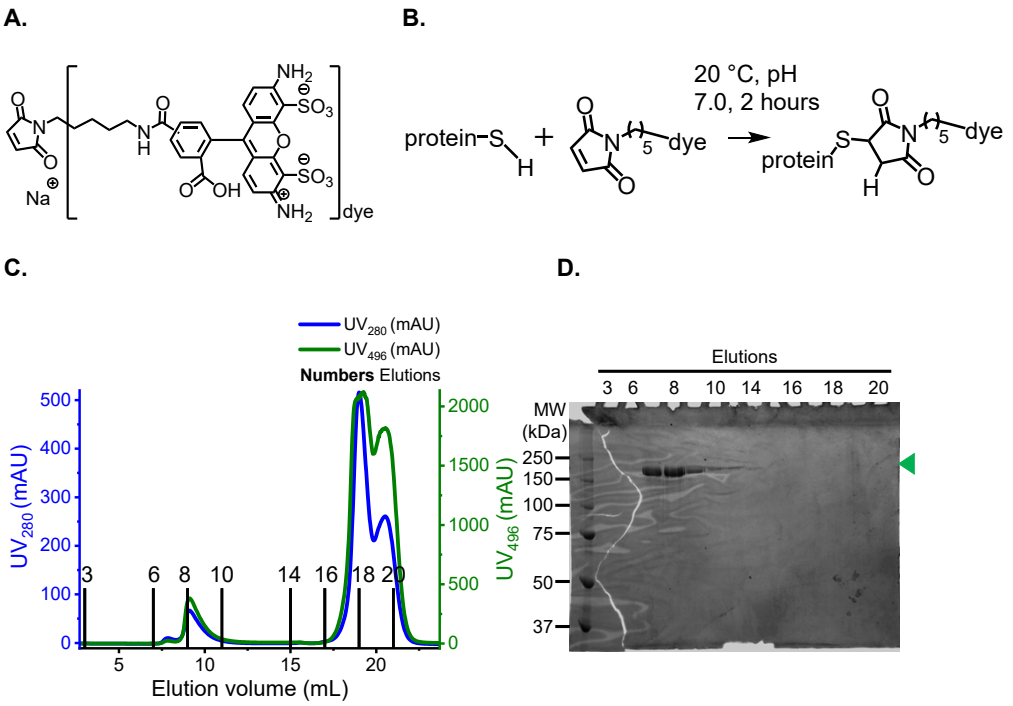


Figure 4.6: Labelling of D0118_2Cys and purification of D0118_2A488. A. Molecular structure of A488-C_5-maleimide[408]. **B.** Schematic of the labelling reaction[409]; for clarity only one fluorophore shown. **C.** Purification of D0118_2A488 from excess fluorophore by SEC on a Superdex 200 column as monitored by A_{280} (protein, blue) and A_{496} (A488, green). **D.** SDS PAGE analysis on 8% (*w/v*) polyacrylamide gel of the purification of D0118_2A488 by SEC (green arrow, theoretical MW of 155.0 Da).

Equation 4.1: Labelling efficiency, adapted from the Beer-Lambert law (see section 2.4.2).

A.

$$c_{protein} = \frac{A_{280} - 0.11 * A_{493}}{\varepsilon_{protein}}$$

B.

$$c_{A488} = \frac{A_{493}}{\varepsilon_{fluorophore}}$$

C.

$$E = \frac{c_{A488}}{2c_{protein}} * 100$$

where $\varepsilon_{protein}$ is the extinction coefficient of the protein with free cysteines (8940 M^{-1} cm^{-1}) and $\varepsilon_{fluorophore}$ the extinction coefficient of the fluorophore A488 (72000 M^{-1} cm^{-1}). The labelling efficiency of D0118_2Cys with two A488 fluorophores was 51%. Typically, a yield of 70-90% is obtained[410]. Nevertheless, this yield might suggest that one of the labelling sites was sterically unavailable or that only one cysteine was present in the protein. The latter hypothesis was tested by in-gel trypsin digestion of D0118_2Cys and identification of the resulting peptides, mapped on the sequence of D0118_2Cys. 79% sequence coverage of D0118_2Cys was obtained, clearly showing the incorporation of both cysteine residues in the protein (data not shown). This strategy was preferred over MS of the intact protein, because the MW of D0118_2Cys is close to the mass limit for successful ESI MS and the mass difference of 2 Da for a Thr730Cys mutation might be close to the experimental error of ESI MS. These MS results indicate that the poor labelling yield was not caused by the absence of a cysteine residue.

4.3.2.4 D0310_scc

An eight-domain DRESS protein was produced containing DRESS domains 3 to 10 with the purpose of studying the mechanical unfolding behaviour by AFM. This region was selected for over-production and purification, because the sequence conservation of these DRESS domains in SasC is relatively high, when compared to other strains of *S. aureus* (data not shown). The pETFPP1-modified pET-YSBLIC vector as reported in Gruszka *et al*. (2015)[57] was selected for recombinant gene expression and protein production, with an N-terminal His_6-tag cleavable by HRV 3C protease, a C-terminal non-cleavable Strep tag[251] and two C-terminal cysteine residues for covalent immobilisation on a gold-coated glass square[411].

Over-production and purification of D0310_scc was performed in the same manner as D0710 and D1417 using buffers supplemented with 5 mM β-mercaptoethanol (see section 4.3.2.1). Briefly, D0310_scc was over-produced in BL21-Gold (DE3) cells, cells were lysed and the soluble material was separated from the insoluble material by centrifugation. His_6-tagged D0310_scc was purified by IMAC using an increasing gradient of imidazole concentrations ranging from 20 mM (0%) to 500 mM (100%, Figure 4.7A,B). Fractions 12-15 were pooled and the His_6-3C tag was removed by proteolytic cleavage with HRV 3C protease in a mass ratio of protease to target protein of 1:150 (see section 2.3.5), followed by removal of the protease and purification tag by a second round of IMAC. This resulted in approximately 90% purity for the recombinant protein (Figure 4.7C, lane -). In hindsight,

this purification step was redundant, as Strep affinity chromatography would also have efficiently removed the His_6-3C tag and His_6-tagged HRV 3C protease. D0310_scc was further purified by loading on a StrepTrap column in Strep binding buffer (Table 2.4) and competitively eluted by gravity flow with Strep elution buffer (Table 2.4) containing 2.5 mM desthiobiotin, resulting in >95% purity of the eluted recombinant protein (Figure 4.7C, lane 3-4).

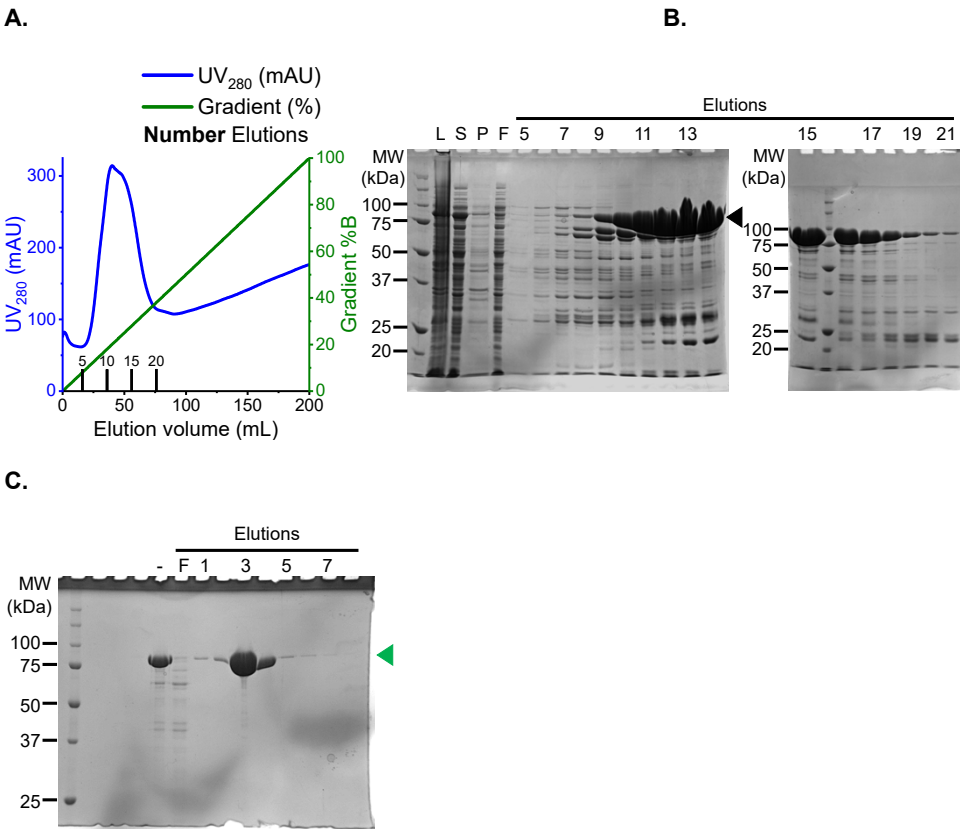


Figure 4.7: Purification of D0310_scc. SDS PAGE analyses run on 10% (*w/v*) polyacrylamide gels. **A.** IMAC chromatogram as monitored by A_{280} (blue) of His_6-3C-D0310_scc. 4 mL fractions were collected during protein elution (numbers shown in black). **B.** SDS PAGE analysis of IMAC purification of His_6-3C-D0310_scc (black arrow, theoretical MW of 69.9 kDa) with L: lysis, S: soluble material, P: insoluble material, F: flow-through. Numbers: fractions. Brightness of right gel was adjusted by 20% post-acquisition. **C.** SDS PAGE analysis of Strep-tag purification. -: D0310_scc (a theoretical MW of 68.1 kDa, green arrow) before StrepTrap chromatography. F: flow-through. Numbers: fractions, 1 mL. Brightness was adjusted by +20% post-acquisition, raw images are available in Appendix 7.3.

The purity of recombinant proteins produced in chapter 5 is shown by SDS PAGE analysis in Figure 4.8 and the yields of all proteins are reported in Table 4.1.

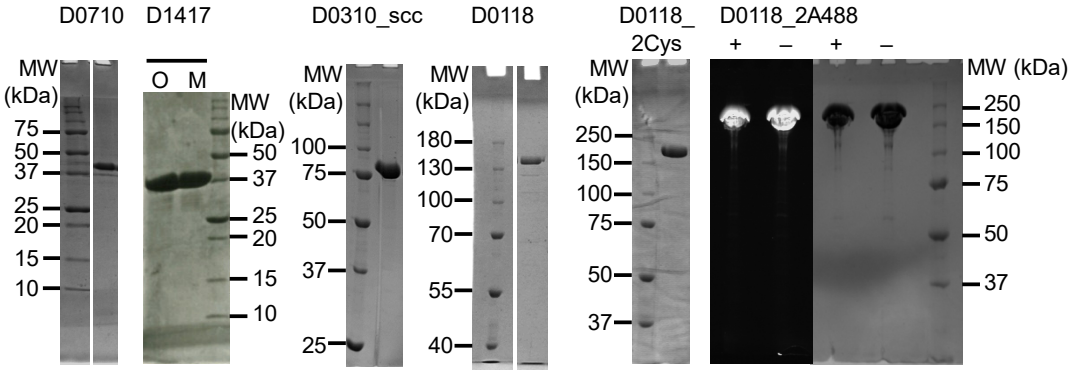


Figure 4.8: SDS PAGE analysis of purified recombinant proteins in this chapter on 10% and 8% (D0118) (*w/v*) polyacrylamide gels. D1417 purified into an oligomeric (O) and monomeric (M) fraction by SEC. The left lanes of D0118_2A488 were illuminated by UV-radiation, showing the fluorescence of A488 fluorophores, where + and – indicates reducing and non-reducing conditions, respectively. The brightness of the gels of D0310_scc and D0118_2A488 was adjusted by +20% post-acquisition and the brightness and contrast of the gel of D0118 was adjusted by +20% post-acquisition. Raw images of gels are available in Appendix 7.3.

Table 4.1: Final yields of recombinant proteins used in this chapter. *: low accuracy of mass determination due to low extinction coefficient. The yield is displayed in mg purified target protein per litre medium. Media are abbreviated as follows: A, auto-induction media; L, LB. pI and ε were determined by ExPASy ProtParam[253].

Protein	Yield (mg L^{-1})	Media	Last purification step	pI	ε (M^{-1} cm^{-1})
D0710	66.1*	A	SEC	4.76	1490
D1417	8.3	A	SEC	5.15	2980
D0310_scc	24.5	A	StrepTrap	5.05	6690
D0118	0.6	L	SEC	5.06	8940
D0118_2A488	~0.02	L	SEC	5.05	8940 (A_{280}) 72000 (per fluorophore, A_{496})

4.3.2 Determination of the oligomeric state of DRESS domain-containing constructs

D1417 eluted from a SEC column in two peaks with approximately a 1:1 ratio (Figure 4.3A). Therefore, the oligomeric state of proteins containing multiple DRESS domains was investigated by SEC-MALLS (Figure 4.9). DRESS domains were eluted over a suitable SEC column and the elution was monitored by detectors for static light scattering, QELS, UV absorbance and refractive index (see section 2.6.1.2). The MW was calculated from the static light scattering and the R_h was estimated from QELS (Table 4.2; see section 2.6.1.1).

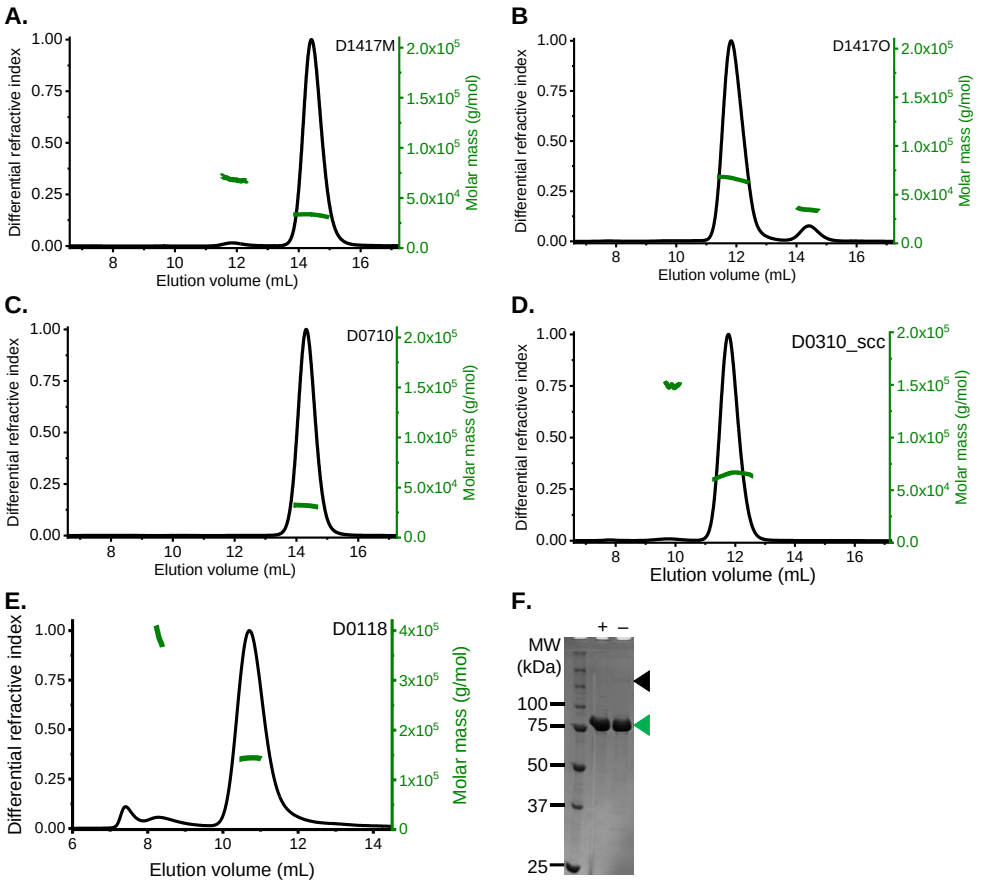


Figure 4.9: SEC-MALLS analysis of DRESS proteins on a Superdex 200 column, except for D0118. Elution profile in normalised DRI units are shown in black and the molar mass estimate (g/mol) in green. The theoretical MWs are listed in Table 4.2. Buffer conditions were 20 mM Tris, 150 mM NaCl, pH 7.5 unless otherwise stated. **A.** D1417M at 1.1 mg/mL, **B.** D1417O at 2.3 mg/mL, **C.** D0710 at 7.1 mg/mL, **D.** D0310_scc at 6.6 mg/mL in 25 mM MES, 150 mM NaCl, 1 mM TCEP, pH 6.0. **E.** D0118 at 0.92 mg/mL in 20 mM Tris, 150 mM NaCl, 1 mM EDTA, pH 7.5 on a Superose 6 column. **F.** SDS PAGE analysis of D0310_scc (theoretical MW of 68.1 kDa, green arrow) and putative D0310_scc dimer (theoretical MW of 136.2 kDa, black arrow) in reducing (+) and non-reducing (-) conditions on SDS PAGE gel (10% (*w/v*) polyacrylamide). Brightness and contrast were adjusted by +20% post-acquisition; for raw gel image, see Appendix 7.3.

Table 4.2: Molar masses and R_h of DRESS domain-containing proteins by MS and SEC-MALLS. [a]Experimental error was 1.5 Da as determined by calibration with an external standard, myoglobin. [b]79% sequence coverage was obtained by in-gel trypsin digestion.

Protein	MW (Da)				R_h (nm)
	Theoretical	MS	SEC-MALLS		
D0710	33722.2	33722.3	32070 ± 64		4.2 ± 1.8
D1417	34785.8	[a]34787.3	Monomer	33580 ± 34	4.1 ± 1.7
			Dimer	67130 ± 67	5.8 ± 0.4
D0310_scc	68076.2	68077.1 (24%), 68238.6 (65%), 68400.3 (11%)	Monomer	65800 ± 66	5.7 ± 0.1
			Dimer	149500 ± 2990	7.3 ± 2.7
D0118	153572.6	N.D.	Monomer	143500 ± 144	10.0 ± 0.03
			Oligomer	386900 ± 3482	56.8 ± 5.1
D0118_2Cys	153540.6	[b]N.D.	N.D.		N.D.

The four-domain DRESS protein D1417 from the C-terminal end of the repetitive region in SasC forms non-exchanging monomeric and dimeric states in solution (Figure 4.9A, B), as shown by the calculation of the MW (Table 4.2). In contrast, the four-domain DRESS protein D0710 from the middle of the SasC repetitive region was monomeric (Figure 4.9C) up to 70 mg/mL (data not shown). Furthermore, D0310_scc from the N-terminal region of SasC was mostly monomeric, with a low percentage of dimer present. Here, dimer formation is likely attributed to disulfide formation between C-terminal cysteine residues, which is confirmed by the observation of a minor additional band by SDS PAGE analysis in non-reducing conditions (Figure 4.9F).

To investigate the significance of these results for the intact protein, the oligomeric state of D0118 was determined, which is the full physiological length of the repetitive region in SasC containing eighteen domains. The calculated MW of D0118 is 6.6% smaller than the MW expected for a monomer of D0118, indicating that D0118 is a monomer in solution. Two higher MW species were observed, of which the first species was aggregated protein eluting in the void volume. The second higher MW species represents 5.9% of the total DRI signal and has a calculated MW of around 2.8 times that of D0118, a R_h five times that of D0118 and a globular shape as determined by SEC-SAXS (see section 4.3.6.4). Hence, it is

hypothesised that this species represents a collapsed, unfolded polymer chain of multiple D0118 molecules. Thus, the dimerisation observed for D1417 is likely non-physiological.

4.3.4 Determining the MW of DRESS domain-containing constructs by ESI MS

The MWs of proteins in this chapter were verified by ESI MS and compared to those expected and observed from SEC-MALLS (Table 4.2). The MWs of D1417 and D0710 were confirmed within the experimental error, compared to the theoretical MW. Masses of D0118 and D0118_2Cys could not be determined by ESI MS due to limiting sample amounts; a trypsin digest of D0118_2Cys confirmed the correct sequence with 79% coverage (see section 4.3.2.3). ESI MS of D0310_scc showed three peaks, of which 24% contained the correct MW species, in addition to two species which were +162.13 Da (65%) and 2x +162.13 Da (11%) heavier than the target protein. The size of the mass additions would be consistent with glycosylation and this protein differs from other recombinant protein constructs in that it contains two C-terminal cysteine residues, which might be involved in the formation of the mass adducts. Both hypotheses were tested experimentally (Bioscience Technology Facility, University of York).

The observed masses did not change upon reduction of the protein. Terminal sequencing by matrix-assisted laser desorption/ionisation with in-source delay (MALDI-ISD)[412] did not obtain results for the C-terminus of the protein, but the N-terminus was reconstructed correctly. Deglycosylation treatment with 25% (*v/v*) ammonium hydroxide for 16 hours at 45 °C[413] changed the +162.13 Da adduct back to the MW of an unmodified protein, but had no effect on the 2x +162.13 Da adduct. Finally, cysteine alkylation by iodoacetamide[414] was possible two times on the unmodified species, one time on the +162.13 Da adduct, and was not possible on the 2x +162.13 Da adduct. Thus, attempts to identify or localise the mass additions were inconclusive (Bioscience Technology Facility, University of York, see section 6.1.2).

4.3.5 Thermal stability of regions from DRESS region

A large difference in thermal stability was observed between single and tandem DRESS domains (see section 3.3.7), suggesting that the DRESS domain interface contributes strongly to the stability of individual DRESS domains. Therefore, the thermal stability is

hypothesised to increase with the number of DRESS domains. Here, the thermal stability of recombinant proteins containing multiple DRESS domains is assessed by a combination of CD and nano-DSF (Figure 4.10A,B; see sections 2.6.2 and 2.6.3) and reported in Table 4.3.

A mix of dimeric and monomeric D1417 had a double unfolding transition by CD with an equal ratio between the transitions and a single unfolding transition by nano-DSF. The higher T_m is very similar to the T_m of D1617. D0710 had a single transition in both techniques and 9.4 °C difference between T_m values. D0310_scc and D0118_2Cys were not measured by CD and their thermal stability did not increase further. The trace of D0118 shows multiple transitions, which need to be investigated further to determine if D0118 unfolds all-in-one or not. No aggregation was observed during thermal denaturation of recombinant proteins containing multiple DRESS domains (Figure 4.10C).

Table 4.3: T_m of proteins containing multiple DRESS domains by nano-DSF and CD.

Protein	T_m (°C) by nano-DSF	T_m (°C) by CD
D17	27.0	30
D1617	48.5	52
D0710	62.6	72
D1417	O 50.1 M 50.1	O,M 46, 60
D0310_scc	62.5	N.D.
D0118_2Cys	61.9	N.D.

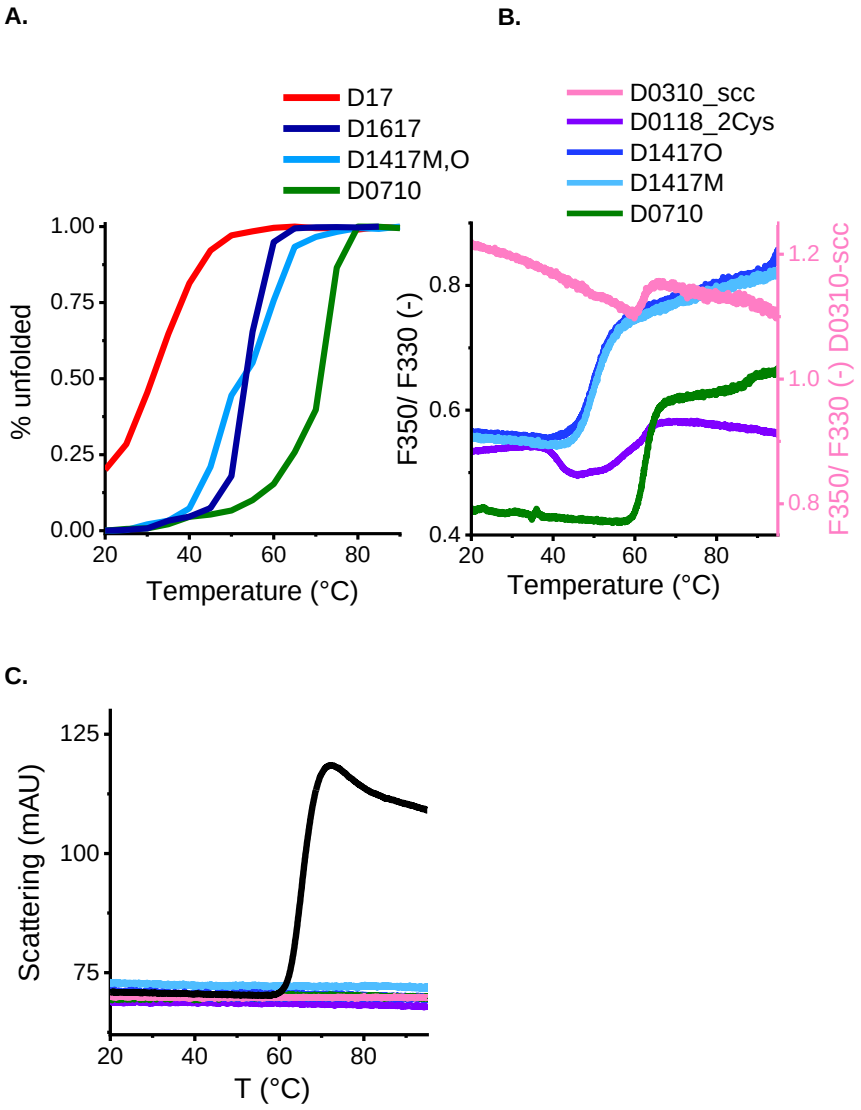


Figure 4.10: Assessing T_m and aggregation of proteins containing multiple DRESS domains by CD, nano-DSF and static light scattering. A. Thermal denaturation by CD measured at 222 nm, converted to percent unfolded (see Equation 2.8). D17 at 22 µM and D1617 at 11 µM in 20 mM sodium phosphate, pH 5.5 (data reproduced from Figure 3.15). D0710 at 5.9 µM in 20 mM sodium phosphate, pH 4.9. D1417M,O at 5.7 µM in 20 mM potassium phosphate, pH 4.6. **B.** Thermal denaturation by nano-DSF. Proteins were measured at 1 mg/mL in 20 mM Tris, 150 mM NaCl, pH 7.5; except D0118_2Cys in 20 mM Tris, 150 mM NaCl, 1 mM EDTA, 5 mM β-mercaptoethanol, pH 7.0 and D0310_scc, in 20 mM Tris, 150 mM NaCl, 1 mM TCEP, pH 7.5. All nano-DSF experiments were performed in duplicate. **C.** Static light scattering signal with conditions as in B. A positive aggregation control is shown for clarity (D1617 in 25 mM MES, 2 M NaCl, pH 6.0).

4.3.6 Elongation of DRESS domains in solution

4.3.6.1 Log-Log plot

Repetitive regions in several CWA proteins, such as SasG[57,59] and Ebh[58] have an elongated nature with the apparent functional role of projecting an N-terminal domain away from the bacterial cell surface[29]. Previously, tandem DRESS domains were found to have an elongated shape and a head-to-tail domain organisation (see section 3.3.9.2). Here, the

hypothesis that proteins containing four or more DRESS domains also have an elongated head-to-tail arrangement is investigated.

Smilgies and Folta-Stogniew (2015)[415] reported a double logarithmic correlation between MW and R_g, which can be used as a tool to assess a deviation from a globular shape. Typically, elongated proteins have a larger R_g with respect to their MW, resulting in a deviation from predicted globular behaviour in a double logarithmic plot. Here, this correlation is used to determine if proteins containing multiple DRESS domains are elongated in solution.

Using SEC-MALLS, the MW was calculated by static light scattering and R_h was calculated by QELS (see sections 2.6.1 and 4.3.3). R_g was calculated from R_h using Equation 4.2, assuming a globular shape. The resulting R_g was plotted as a function of MW on a double logarithmic plot and the trend, as reported for a range of globular proteins, is shown (Figure 4.11). Furthermore, the R_g value obtained by SAXS (see sections 4.3.6.2 and 4.3.6.4) was plotted as a function of MW. Proteins containing DRESS domains do not correlate with the trend for globular proteins. Instead, their R_g is larger, suggesting they have an elongated shape. This is less apparent for a single DRESS domain, where R_h could only be determined with low accuracy because the domain is likely only transiently folded (see section 3.3.6), and most apparent for the protein construct with eighteen DRESS domains, which is expected to display the highest anisotropy.

Equation 4.2: R_g from R_h for spherical objects[415].

$$R_g = R_h\sqrt{\frac{3}{5}}$$

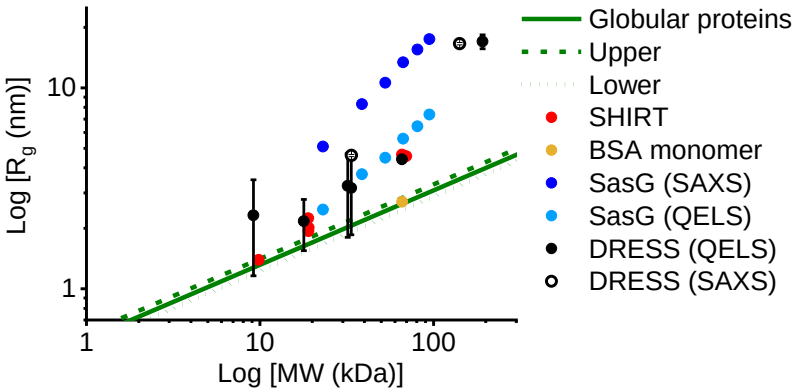


Figure 4.11: Log-Log correlation between R_g and MW[415]. Data is shown for monomeric proteins only. Green: globular trend[415]. DRESS proteins (QELS) contained (from left to right) 1, 2, 4, 4, 8 and 18 domains. DRESS proteins (SAXS) contained (from left to right) 4 (D0710) and 18 domains (see sections 4.3.6.2 and 4.3.6.4). SHIRT proteins (QELS) contained (from left to right) 1, 2, 2, 7 domains; data courtesy of Dr J. Gilburt. SasG proteins contained 2-7 E-G5 domains; data adapted from Gruszka *et al*. (2015)[57] with R_g determined by SAXS (dark blue) and calculated from R_h by QELS (light blue). BSA monomer is shown as external control.

Compared to other elongated rod-like proteins, proteins containing DRESS domains show a similar trend in the double logarithmic plot distinctly different from that of globular proteins, such as BSA[416]. Furthermore, the comparison of SasG and DRESS data with R_g obtained via SAXS and QELS validates the use of Equation 4.2. Although this equation was derived for spherical objects, it still yields a trend distinct from that for globular proteins. In summary, the double logarithmic plot of R_g as a function of MW is consistent with recombinant proteins constructs containing multiple DRESS domains behaving as elongated proteins, like other repetitive multi-domain proteins known to form elongated, rod-like structures.

4.3.6.2 SAXS on D0710

The double logarithmic correlation between MW and R_g in section 4.3.6.1 provided some insight into the shape of DRESS domain-containing constructs in solution. Here, the size and shape of D0710 is studied in more detail by SEC-SAXS (Figure 4.12A; for more information regarding the technique please refer to section 2.9).

The shape of the logarithmic scattering plot (Figure 4.12B) is suggestive of an elongated shape in solution[344], however the lack of distinct features in the scattering plot might also be interpreted as disorder or flexibility. The possibility of disorder is assessed in the Kratky plot (see later) and flexibility is assessed with the Porod plot and EOM modelling (see later).

From the Guinier approximation (Figure 4.12C), the R_g was calculated as 4.61 ± 0.06 nm. This R_g agrees well with the value determined from static light scattering (Figure 4.12A),

which was 4.52 ± 0.08 nm. From R_g (SAXS) and R_h (QELS), the shape factor $\frac{R_g}{R_h}$ can be determined. Globular proteins have a shape factor of 0.70[417] and this value is ≥1 for elongated proteins. The shape factor of D0710 is 1.1, which indicates that D0710 is elongated in solution. Furthermore, the linearity of the Guinier plot indicates that D0710 is not aggregated.

The R_g of the cross-section as determined from the modified Guinier approximation for rod-like particles is 0.651 ± 0.005 nm (Figure 4.12D). This can be converted into the diameter of the rod as follows (Equation 4.3[60]), yielding an estimated diameter of 1.84 ± 0.01 nm. From the crystal structure of D1617, the diameter of a DRESS domain was estimated to be 2.0 nm. These values are in reasonable agreement, considering they were determined for protein structures in solution and in a crystalline state, respectively.

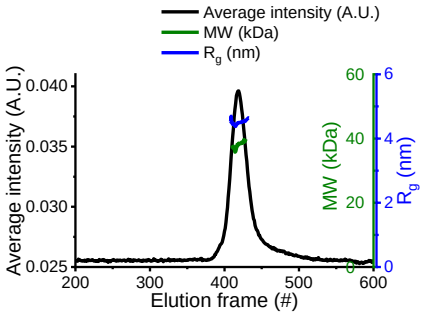
Equation 4.3: Diameter of a rod from $R_{c,g}$[60].

$$d = 2R = 2\,R_{c,g}\sqrt{2}$$

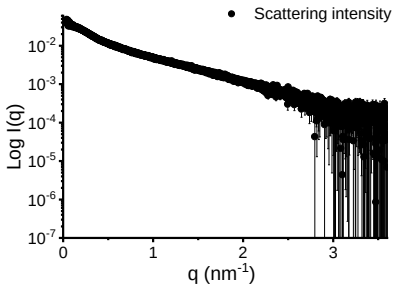
The Porod exponent, calculated from the negative slope of the mid-q scattering region of the Porod plot, informs about the shape of the particle in solution[347–349]. Here, a Porod exponent of 1.074 was calculated (Figure 4.12E), which is indicative of a rigid rod. Furthermore, the relative conformational flexibility of the particles in solution is assessed from the Kratky plot (Figure 4.12F). The curve displays a typical bell-shape, with a small tail towards high values of *q*, where the data is less accurate due to a lower signal-to-noise level. This suggests that D0710 is well folded in solution and exhibits some limited flexibility.

Finally, the shape of P(r) is typical of elongated particles (Figure 4.12G)[344]. The inflection point following the maximum in P(r) represents the diameter of the rod and has a value of 2.0 nm. This is in good agreement with the cross-sectional R_g and in excellent agreement with the expected diameter of the rod. The intersection of P(r) with the x-axis at 19.3 nm represents the largest diameter of the particle, D_{max}. Here, this corresponds to the average length of the particle in solution.

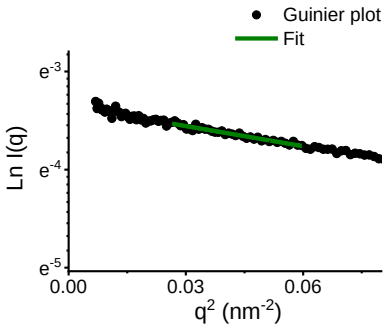
A.



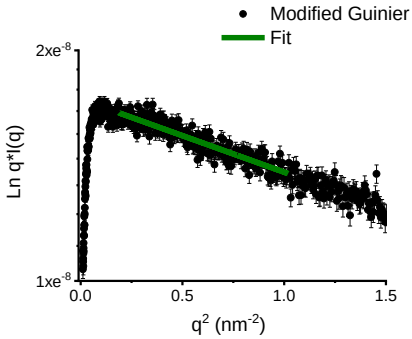
B.



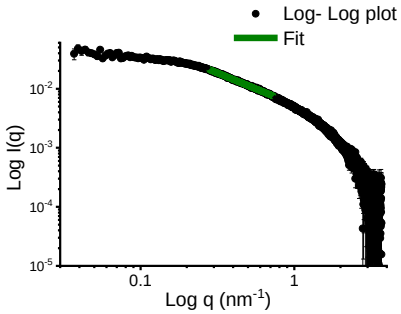
C.



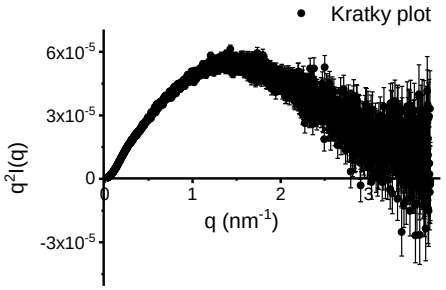
D.



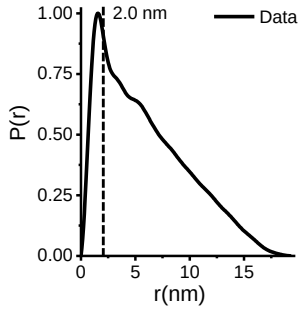
E.



F.



G.



(Figure legend on next page)

Figure 4.12: SEC-SAXS analysis of D0710. A. SEC chromatogram as monitored by A_{280}, SLS and QELS of D0710 on a Superdex 200 Increase 3.2 column. **B.** Logarithmic intensity (I) in arbitrary units as a function of the momentum transfer vector q in nm^{-1}. **C.** Fit of the Guinier approximation for $qR_g \leq 1.1$. **D.** Fit of the modified Guinier approximation for a rod. **E.** Fit of the central region of Log-Log correlation plot between I(q) and q. **F.** Kratky plot. **G.** P(r) obtained by GNOM[418], normalised.

The MW was determined from SAXS scattering on a relative scale using the method reported by Fischer *et al.*[353], which is applicable to monodisperse proteins in a dilute solution (Table 4.4). The calculated MW (33.7 kDa) was in excellent agreement with the MW obtained by ESI MS (33.7 kDa). Furthermore, a solvent envelope of D0710 was obtained by *ab initio* modelling using Gasbor[419]. The fit of the solvent envelope models to the EOM models (see below) is reported as normalised spatial discrepancy (NSD), where a value below 1 indicates that the models can be treated as identical and the discrepancy between the theoretical scattering of the solvent envelope and the experimental scattering is reported as a χ^2 value (Table 4.4).

Table 4.4: SAXS structural parameters and molecular mass determination. NSD is calculated by Damsup[420].

Structural parameters		**MW determination**	
I(0)	0.035	MW from Fischer *et al.*[353]	33.7
Guinier qR_g range	0.46-1.10	MW from SEC-MALLS (kDa)	32.1
R_g from Guinier (nm)	4.61	MW from ESI MS (kDa)	33.7
R_g from SLS (nm)	4.52	MW from sequence (kDa)	33.7
R_h from QELS (nm)	4.2		
Shape factor (R_g/R_h)	1.1	Average NSD between three representative Gasbor models	1.23
$R_{c,g}$ from modified Guinier (nm)	0.65	NSD between representative Gasbor model and D1617,D1617	1.20
D_{max} (nm)	19.3	NSD between representative Gasbor model and D16,D1617,D17	1.15
Porod exponent	1.07	NSD between representative Gasbor model and D16,D16,D16,D16	1.09
Gasbor model χ^2 fit	1.61		

4.3.6.3 *Ab initio* modelling of D0710

To investigate the possibility of flexibility in D0710 in solution further, the SAXS data was modelled using the ensemble optimisation (EOM) method[356]. Briefly, the structure is defined as an assembly of rigid bodies, of which 10,000-100,000 possible ensembles are

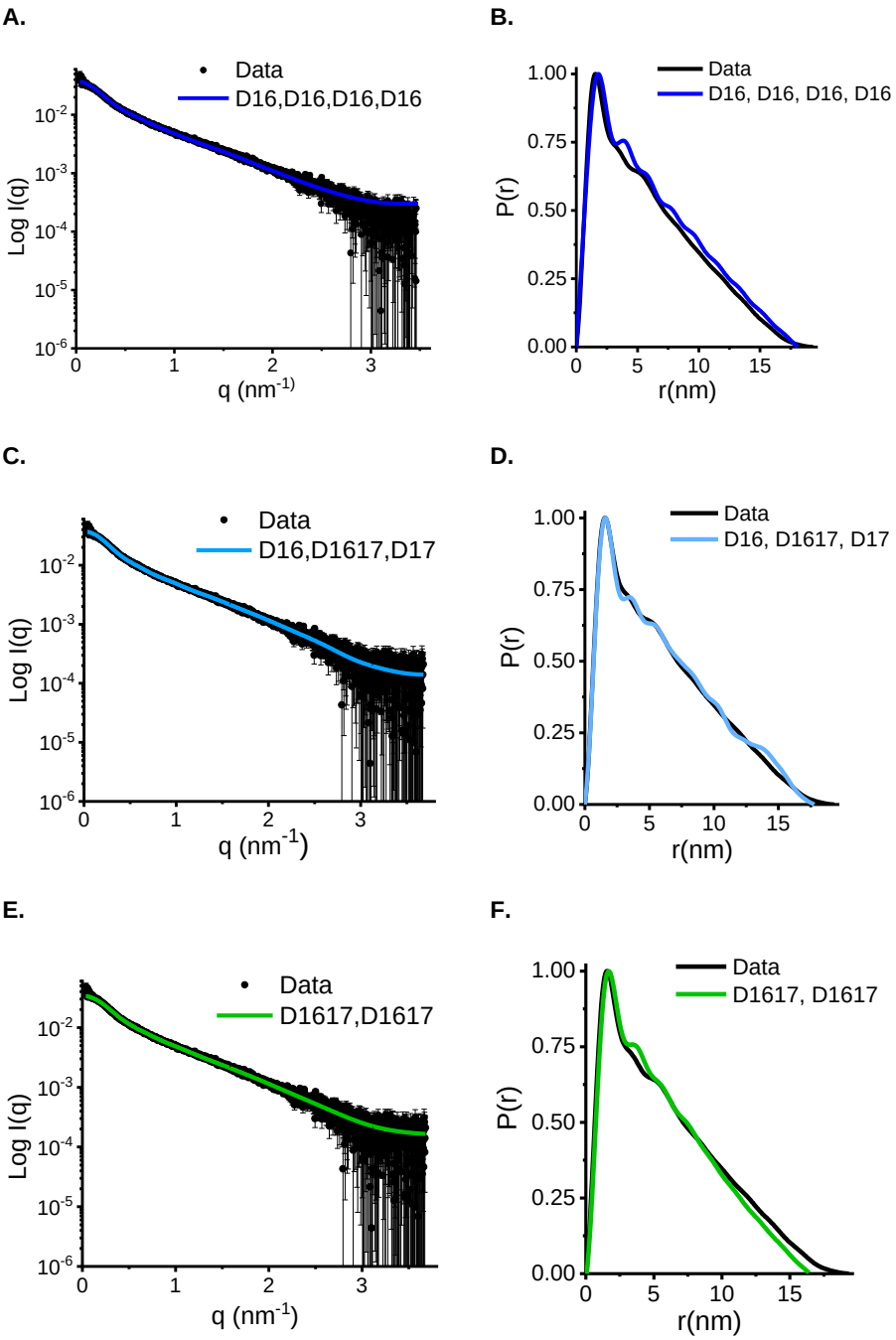
created. These are deposited in a pool and the theoretical scattering of the pool is compared with the experimental scattering, then an algorithm selects an ensemble best representing the data. Rigid structures are best represented by 1-5 conformations and flexible proteins typically contain 10-20 different conformations. As no crystallographic model is available for D0710, the crystal structure of D1617 was employed to generate three rigid body models equivalent to four DRESS domains with as close as possible to the same number of residues as in the D0710 construct (see section 2.9.4.1). They comprise four DRESS domains in tandem (D1617,D1617), two DRESS domains in tandem plus two single DRESS domains (D16,D1617,D17), or four single DRESS domains (D16,D16,D16,D16). This probes the effect of linker connectivity on how well the rigid body models fit the experimental scattering.

All three theoretical rigid body models comprising four DRESS domains with different connectivities were fit best by an ensemble of maximally two conformations. The theoretical I(q) and P(r) functions of the conformations found with highest confidence (92%-100% of total ensemble) are fit to the experimental I(q) and P(r) (Figure 4.13A-F). The rigid body model containing D16,D1617,D17 fits the data best as judged from the χ^2 value (Table 4.5).

Table 4.5: Parameters of SAXS experimental data and EOM models. The R_g value for experimental data was obtained from the Guinier approximation and for EOM models from the Guinier approximation of the theoretical scattering of the selected models. The D_{max} for experimental data was obtained from P(r), and for EOM models from theoretical scattering data converted to P(r) by GNOM[418]. The χ^2 value represents the error of the fit to the theoretical scattering of the selected ensemble to the experimental data[421]. R_{flex} (%) represents the flexibility of rigid body models in the selected ensemble and in all ensembles present in the pool.

Rigid body models	**Number of EOM structures**	**R_h (nm)**	**R_g (nm)**		**D_{max} (nm)**	**χ^2**	**R_{flex} (%)**	
		QELS	Guinier	EOM			Selection	Pool
Experimental data		4.52	4.61		19.34			
D16,D16,D16, D16	10,000			5.13	18.14	1.98	48.30	87.51
D16,D1617, D17	100,000			5.13	17.77	1.10	57.03	89.06
D1617,D1617	100,000			4.60	16.42	1.37	39.21	90.89

Furthermore, the highest confidence conformations fit into the solvent envelope of the experimental data, as calculated by *ab initio* modelling in Gasbor[419] (Figure 4.13G). This confirms that the structures comprised of different sets of DRESS domains all represent the experimental data well. The parameter which shows the largest discrepancy is D_{max} (Table 4.5). The model containing two tandem DRESS domains features the shortest D_{max} and the model with four individual DRESS domains features the longest D_{max}. The difference between these models can be explained by the different connectivities between DRESS domains. The lack of connectivity between DRESS domains allows for more flexibility with regard to adjacent domain orientations, optimising the structure guided by the observed scattering signal.



(Figure continues on next page)

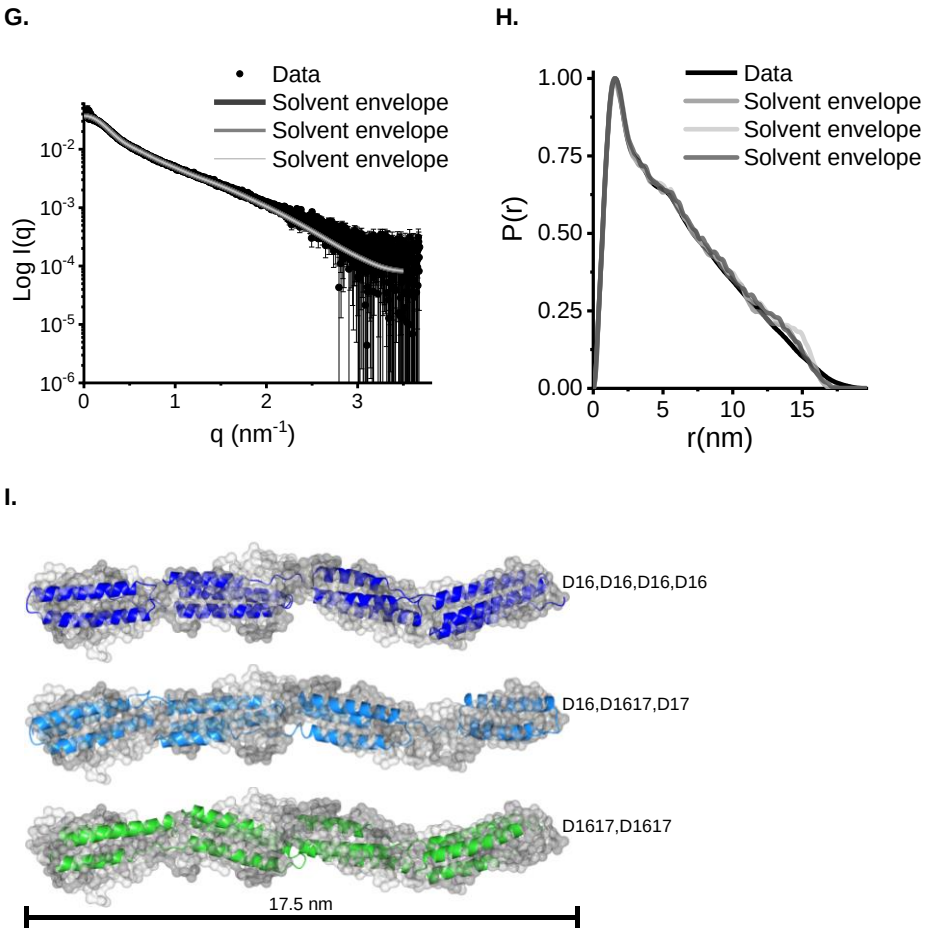


Figure 4.13: EOM models of D0710. The fit of the scattering intensity and P(r) for three different rigid body ensembles and a solvent envelope model[419]. **A, B.** D16,D16,D16,D16. **C, D.** D16,D1617,D17. **E, F.** D1617,D1617. **G,H**. Three representative solvent envelope models. **I.** Best-fits obtained by EOM for three different rigid body models shown overlayed on three representative *ab initio* solvent envelope models[419]. Image was created using CCP4mg.

4.3.6.4 SAXS on D0118

D0710, containing four DRESS domains, behaves as a rigid rod in solution (see section 4.3.6.2). Here, the size and shape of D0118 representing the full, physiological repetitive region in SasC is determined by SEC-SAXS. Briefly, D0118 was analysed on a Shodex KW404-4F column and the eluant was monitored by SLS and QELS (Figure 4.14A). Three species were detected in the A_{280} trace, but the signal-to-noise of the third peak (at elution frame 500) was too low to be analysed.

Scattering data for the minor peak (at elution frame 280) was restricted to below 1.7 nm^{-1} (data collected to 4.9 nm^{-1}) by SHANUM[352] due to a poor signal-to-noise level for this species. The logarithmic scattering intensity (Figure 4.14B) is noisy and featureless, suggesting either disorder or an extended state. The Porod exponent for the mid-q

scattering region of the double logarithmic plot (Figure 4.14C) was 3.3, which can be indicative of a collapsed polymer chain[348]. The Kratky plot (Figure 4.14D) did not converge, suggesting that the particle is disordered. Therefore, no further analysis was performed on this species.

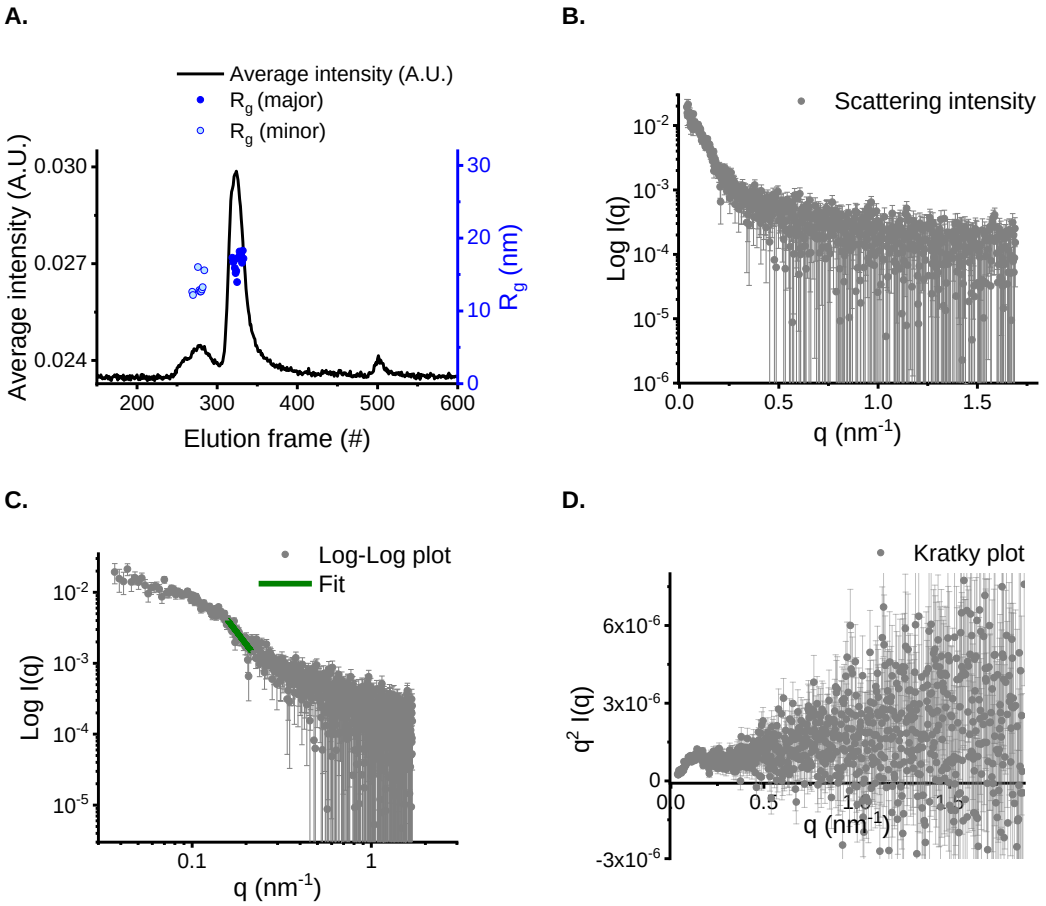


Figure 4.14. SEC-SAXS analysis of the large MW minor species in purified D0118. A. SEC chromatogram as monitored by A_{280}, SLS and QELS of D0118 on a Shodex KW-404 column (volume 4.6 mL). **B.** Logarithmic intensity (I) in arbitrary units as a function of q in nm^{-1}. **C.** Fit of the central region in the Log-Log correlation plot between I(q) and q. **D.** Kratky plot.

The shape of the logarithmic scattering plot for the major species (Figure 4.15A) suggested either disorder or an extended state[344]. The Kratky plot converged (Figure 4.15B), suggesting the particle was folded, and the Porod exponent from the Log-Log plot (Figure 4.15C) was 1.05, indicating the particle behaves as a rigid rod in solution. Hence, further analysis of this species is valid.

The R_g was estimated to be 16.6 nm from the Guinier plot (Figure 4.15D) and this value corresponds well to the R_g value obtained by SLS (16.8 nm, Table 4.6). A shape factor of 1.0 was estimated from R_g, suggesting D0118 is elongated in solution. The $R_{g,c}$ was 0.73 as

determined from the modified Guinier plot (Figure 4.15E), which provided an estimated rod diameter of 2.1 nm ± 0.03. This diameter is in good agreement with the values obtained for D0710 (1.84 nm) and from the crystal structure of D1617 (2.0 nm).

Finally, P(r) was obtained by GNOM[418] (Figure 4.15F). The P(r) indicates an elongated particle shape[344] with a maximum dimension of 66.3 nm and an average diameter of 2.5 nm. The maximum dimension represents the total length for a rod-like particle and this value is slightly smaller than the length estimated from the crystal structure for tandem domains (see section 3.3.9.2; 74 nm). The diameter of 2.5 nm is in reasonable agreement with the diameter obtained from the modified Guinier plot and the crystal structure. Possibly, a larger $R_{g,c}$ value for D0118 compared to D0710 implies that the D0118 rod may bend more than the D0710 rod, as might be expected from the estimated length.

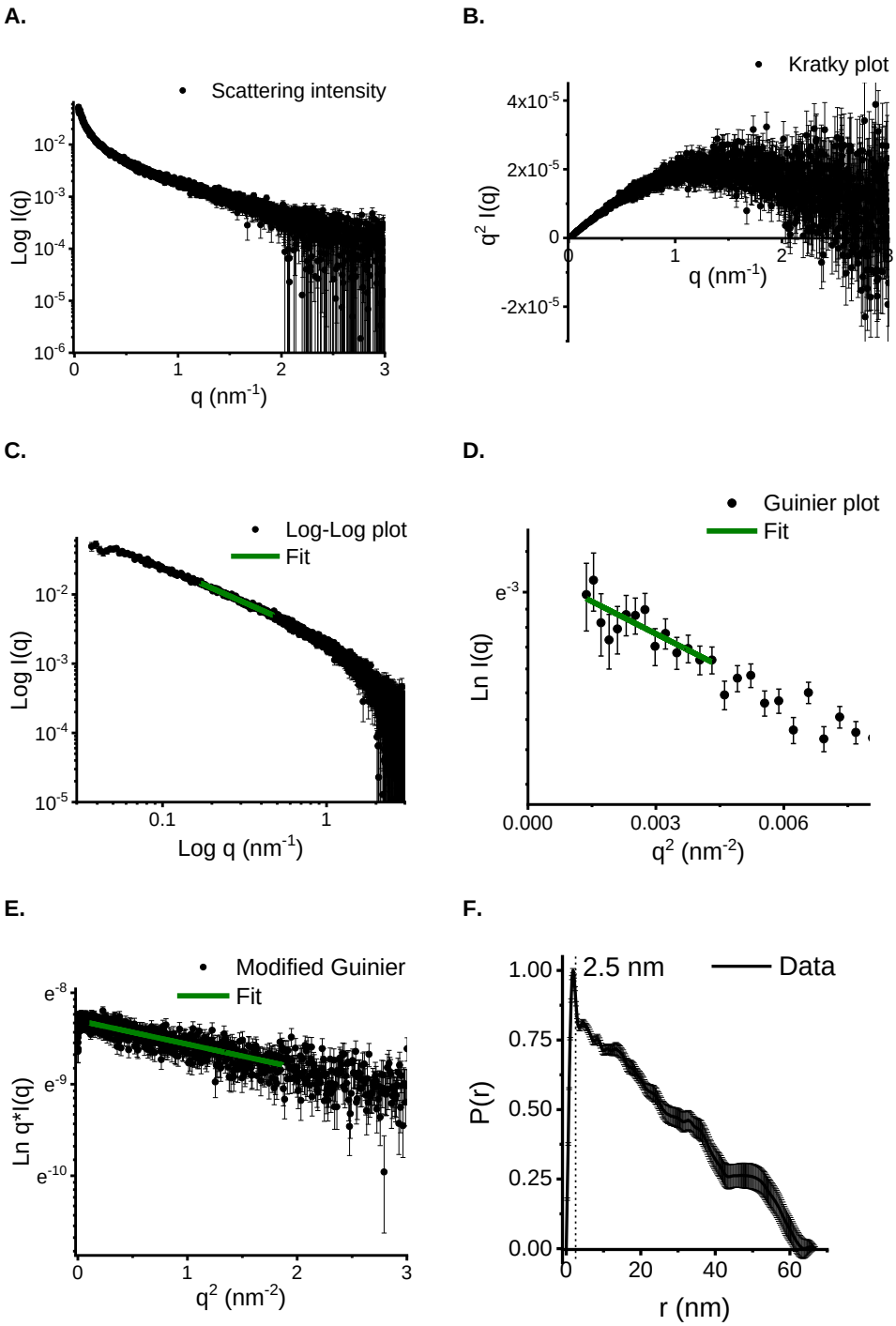


Figure 4.15: SEC-SAXS analysis of the major species in purified D0118. A. Logarithmic intensity (I) in arbitrary units as a function of q in nm^{-1}. **B.** Kratky plot. **C.** Fit of the central region in the Log-Log correlation plot between I(q) and q. **D.** Fit of the Guinier approximation for $qR_g \leq 1.1$. **E.** Fit of the modified Guinier approximation for a rod. **F.** Normalised P(r) obtained by GNOM[418].

The MW of D0118 was estimated to be 141.3 kDa (Table 4.6)[353]. This method has reported errors of ~10% for elongated molecules. The MW estimated here is 92% of the theoretical MW and 74% of the MW observed by QELS, indicating that the major species of D0118 is likely monomeric. The MW of the minor species was estimated to be ~430 kDa, however

this analysis has a poor signal-to-noise level and thus is only to be used as an approximation. Nevertheless, this MW is ~3 times that of a monomer and this in conjunction with the observed disorder, suggests that the minor species represents some 'collapsed', possibly unfolded, oligomeric state of D0118.

Table 4.6: SAXS structural parameters and molecular mass determination.

Structural parameters	Minor species	Major species
I(0)	0.00011	0.055
Guinier qR_g range	0.56-1.06	0.62-1.09
R_g from Guinier (nm)	15.1 ± 4.8	16.6 ± 0.23
R_g from SLS (nm)	13.2 ± 1.4	16.8 ± 1.2
R_h from QELS (nm)	56.8 ± 5.1	10.0 ± 0.03
Shape factor (R_g/R_h)	0.23	1.7
$R_{c,g}$ from modified Guinier (nm)	6.6 ± 0.24	0.73 ± 0.009
D_{max} (nm)	41.0	66.3
Porod exponent	3.29	1.05
Gasbor model χ^2 fit	N/A	1.24
MW determination		
MW from Fischer *et al*.[353]	~430	141.3
MW from SEC-MALLS (kDa)	~387 ± 4	143.5 ± 0.1
MW from ESI MS (kDa)	N.D.	
Theoretical MW from sequence (kDa)	153.6	

4.3.6.5 *Ab initio* modelling of D0118

The *ab initio* model of D0118 was created to determine if the rod is highly extended in solution. To do so, models were created of all DRESS domains in D0118 by SwissModel[377], based on the X-ray crystal structure of D1617 (see section 3.3.9.2). EOM modelling was attempted using the same parameters as for D0710, but due to limited computing power of the local desktop PC it was not possible to generate the necessary 10,000 models from the eighteen different rigid-body models. Use of the online EOM model server was also attempted, where by default the harmonics were set to 15, while a maximum value of 50 is optimal for large models and the maximum number of input models was ten. This sub-

optimal setting generated an ensemble of three models with an overall D_{max} of 35 nm and a R_g of 12.9 nm (Figure 4.16F). The theoretical scattering of this ensemble approximated the experimental data with a χ^2 of 2.623, indicating a poor fit (Fit B; Figure 4.16C). The normalised P(r) function suggests the presence of a rod with a larger D_{max} (Figure 4.16D).

Instead, the AllosMod-FoXS server[359,422] was employed to generate an *ab initio* model of D0118 based on the scattering data and the models generated by SwissModel (see section 2.9.4.3)[377]. Here, a single model was generated (Figure 4.16E), where the theoretical scattering intensity approximated the data with a χ^2 of 1.07 (Fit A; Figure 4.16A). The R_g of the model is 20.4 nm, close to the experimental R_g value of 16.6 nm obtained from the Guinier approximation, and the D_{max} is 69 nm, close to the D_{max} of the scattering data (66.3 nm). Flexible linkers between domains did not improve the fit to the data, suggesting D0118 is a highly elongated, rigid rod in solution.

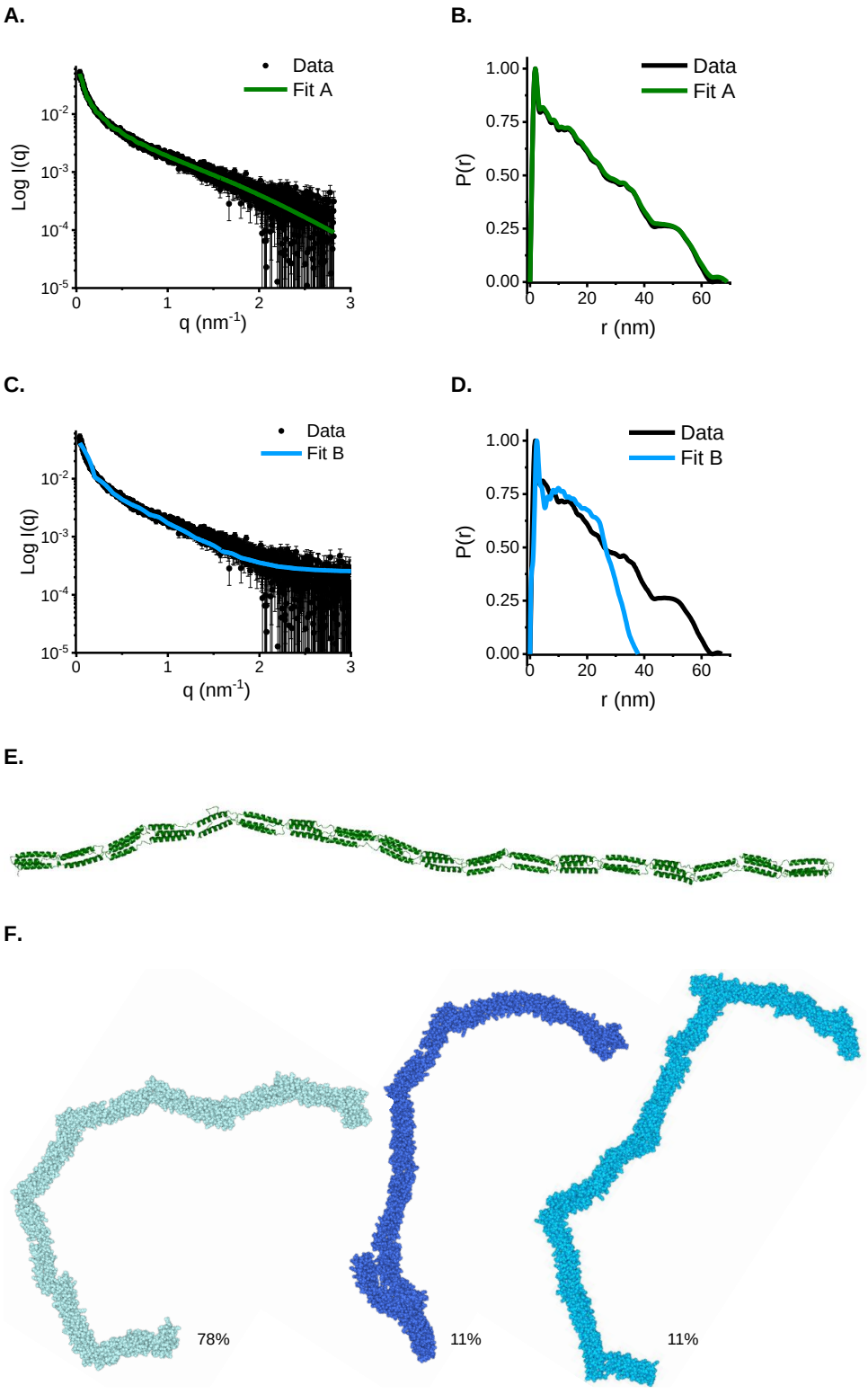


Figure 4.16: *Ab initio* models of D0118. A. Theoretical scattering intensity of AllosMod-FoXS model. **B.** Normalised P(r) function of AllesMod-FoXS model, determined by GNOM[418]. **C.** Theoretical scattering intensity of EOM ensemble. **D.** Normalised P(r) function of EOM ensemble. **E.** AllosMod-FoXS model. **F.** EOM ensemble. Image was created using CCP4mg.

4.3.7 Elongation of the entire, intact DRESS region from SasC measured on a surface

The repetitive region of SasC is hypothesised to form an elongated, rod-like conformation, where the predicted functional role is to project an N-terminal region away from the bacterial cell surface[29]. The elongation of protein constructs representing truncated and full portions of the DRESS region from SasC has been demonstrated by solution biophysical techniques. Here, the end-to-end distance of an A488-labelled recombinant protein construct containing eighteen DRESS domains (D0118_2A488) is determined as it represents the entire, intact repetitive DRESS region in SasC. This distance was determined by measuring the inter-fluorophore distance of two A488 fluorophores coupled to D0118_2Cys, where the fluorophore-labelled protein construct is immobilised on a 2 µg/mL poly-*D*-lysine coated quartz slide via electrostatic interactions. Previously, SHRImP-TIRF microscopy has been used to determine the end-to-end distances of protein constructs representing the entire repetitive region from SasG[57]. For further details about SHRImP-TIRF, please refer to section 2.10.

The predicted inter-fluorophore distance is 65 nm based on the size of single and tandem DRESS domains in the X-ray crystal structure (Figure 4.17A, B; see Figure 3.18). This is in agreement with the inter-fluorophore distance of 60 ± 2.4 nm (mean ± s.e.) at pH 7.0 in HEPES imaging buffer (Figure 4.17C, D; Table 4.7), as determined from SHRImP-TIRF. At pH 6.5 in MES imaging buffer, the observed inter-fluorophore distance is 44 ± 1.9 nm (mean ± s.e.; Figure 4.17E, F; Table 4.7). However, this dataset should be interpreted with caution as it was collected from a single experimental replicate. In addition, the Gaussian fit is shown for inter-fluorophore distance histograms without (Figure 4.17C, E) and with (Figure 4.17D, F) an eccentricity filter applied to individual, detected fluorescent spots. The standard deviation of the Gaussian fit becomes smaller upon applying the eccentricity filter, thus increasing the precision of the mean inter-fluorophore distance.

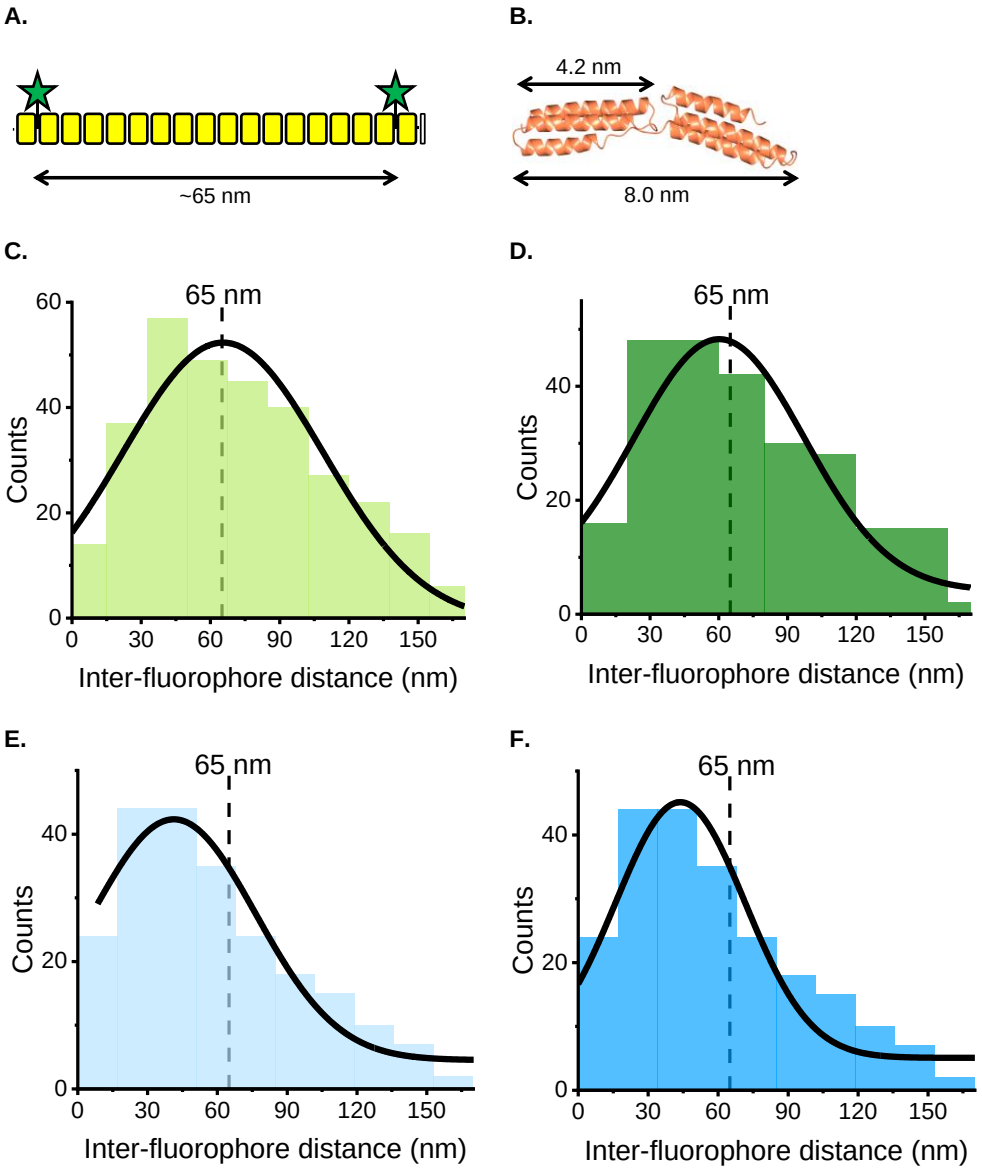


Figure 4.17: Inter-fluorophore distances on a 2 µg/mL poly-D-lysine coated quartz surface. Dashed line indicates the expected size. Solid line represents the Gaussian fit to inter-fluorophore distances <160 nm. **A.** Schematic inter-fluorophore distance of D0118_2A488. **B.** Schematic length of single and tandem DRESS domains; image was created using CCP4mg. **C, D.** 20-2000 pM D0118_2A488 in 10 mM HEPES, 10 mM NaCl, 1 mM Trolox, 5 mM β-mercaptoethanol, pH 7.0, without (C) and with (D) eccentricity filter applied to fluorescent events. **E, F.** 100 pM D0118_2A488 in 10 mM MES, 10 mM NaCl, 1 mM Trolox, 5 mM β-mercaptoethanol, pH 6.5, without (E) and with (F) eccentricity filter applied to fluorescent events; this dataset was collected from a single experimental replicate.

Table 4.7: Parameters for inter-fluorophore distance histograms and their Gaussian fit. Number of slides represents the number of independently prepared samples for imaging. [a]Eccentricity ratio between 0.85 and 1.15. n: number of measured inter-fluorophore distances. R^2 represents the goodness of fit of the Gaussian curves to the histograms.

Gaussian fit from Figure 4.17	**C.**	**D.**	**E.**	**F.**
pH, buffer	7.0, HEPES	7.0, HEPES	6.5, MES	6.5, MES
Number of slides imaged	5	5	1	1
Eccentricity filter applied[a]	No	Yes	No	Yes
n	313	244	273	223
Mean + s.e. (nm)	65 ± 2.4	60 ± 2.4	41 ± 2.1	44 ± 1.9
SD (nm)	43	37	35	28
R^2	0.89	0.82	0.94	0.88

4.3.8 Mechanical unfolding and refolding of DRESS domains

4.3.8.1 Introduction to mechanical unfolding of proteins using AFM

Other α-helical repetitive domains with a mechano-sensitive functional role, such as spectrin in erythrocytes[342] or talin in cell-matrix connections[423], unfold at an applied elongational force of tens of piconewtons[233]. Furthermore, they refold[342,424,425] and the unfolding/refolding pathways for various α-helical repetitive domains include both cooperative and non-cooperative folding/unfolding transitions[423,426,427]. Based on this literature, it was hypothesised that DRESS domains would unfold at applied mechanical forces on the order of tens of piconewtons. In addition to the previously observed refolding ability after thermal denaturation (data not shown), it was hypothesised that DRESS domains would also refold when the applied force is reduced to near zero. Finally, a cooperative unfolding pathway was observed for thermal denaturation (see section 3.3.7); here, it is assessed if DRESS domains also unfold cooperatively in mechanical pulling experiments.

Briefly, a MLCT probe (Bruker) was calibrated as described in section 2.11.2. D0310_scc was deposited on a freshly exposed gold-coated glass square in a final concentration of

0.2-2 µM and incubated for five minutes at 20 °C to allow for the formation of thiol-gold bonds (average strength ~700 pN at pH 7.4[411]).

4.3.8.2 Unfolding of DRESS domains under mechanical load

The mechanical strength (F) required to unfold individual DRESS domains and the increase in contour length per unfolding event (ΔL) were measured by AFM. The fully extended, folded protein D0310_scc under at low applied force has an expected length of 44 nm, based on the crystal structure of D1617 and extended linker residues. The expected ΔL per unfolded domain is 28 nm, as calculated from the freely jointed chain model for rigid chains[428] (Equation 4.4), where the contour length of a single amino acid in the extended state is assumed to be 0.42 nm[429,430]. The total contour length of a fully extended, unfolded D0310_scc molecule is estimated to be 266 nm.

Equation 4.4: ΔL from freely jointed chain model[428].

$$\Delta L = L_{ext} - L_{folded} = N_{aa} \cdot b_{aa} - L_{folded}$$

where ΔL is the increase in contour length per domain, L_{ext} is the freely jointed chain length for a fully extended polypeptide representing one domain, L_{folded} is the length of a single folded domain and is obtained from the crystal structure of D1617, N_{aa} is the number of amino acids in the polypeptide representing one domain and b_{aa} is the contour length of a single amino acid in the extended state.

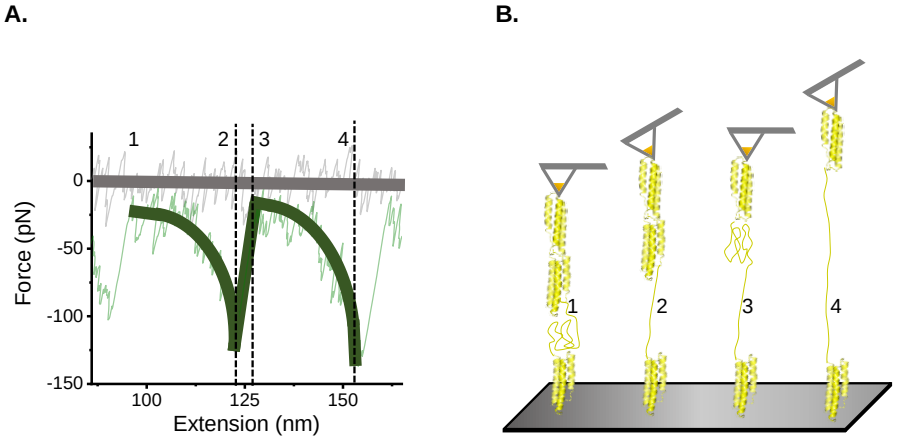


Figure 4.18: Cartoon representation of the mechanical unfolding of DRESS domains. Only four domains and two unfolding events are shown for clarity. **A.** Schematic saw-tooth force-extension approach (grey) and retract (green) traces superposed on data from Figure 4.19A. The schematic, bold, green retract trace is a WLC fit to the unfolding curve. **B.** Cartoon representation of the sequential unfolding of DRESS domains under constant applied force. Numbers correspond to sequential unfolding events in A.

Typically, the probe approaches the surface (Figure 4.18B; grey) and the force increases to a threshold force of 0.5-2 nN when the probe touches the surface (not shown). At the surface, the probe occasionally picks up a protein via non-specific interactions at a random position along the eight-domain D0310_scc protein[233]. As the probe lifts off from the surface, an initial peak of differing size is observed in the retract trace (green; not shown), which represents the protein in a random orientation lifting off from the surface while tethered between the probe and the surface. This is followed by the sequential unfolding of DRESS domains (Figure 4.18A, B). A single saw-tooth represents the elastic extension of an unfolded domain with increasing force (Figure 4.18; steps 1, 3), followed by a drop in force upon unfolding of the next domain (Figure 4.18; steps 2, 4)[233,431]. In the final peak, the non-specific, non-covalent interactions between the immobilised protein and the probe are disrupted and the protein is detached from the tip, usually requiring a force much larger (>200 pN) than that of the unfolding of DRESS domains. Figure 4.19 shows typical saw-tooth force-extension traces obtained at different constant pulling speeds.

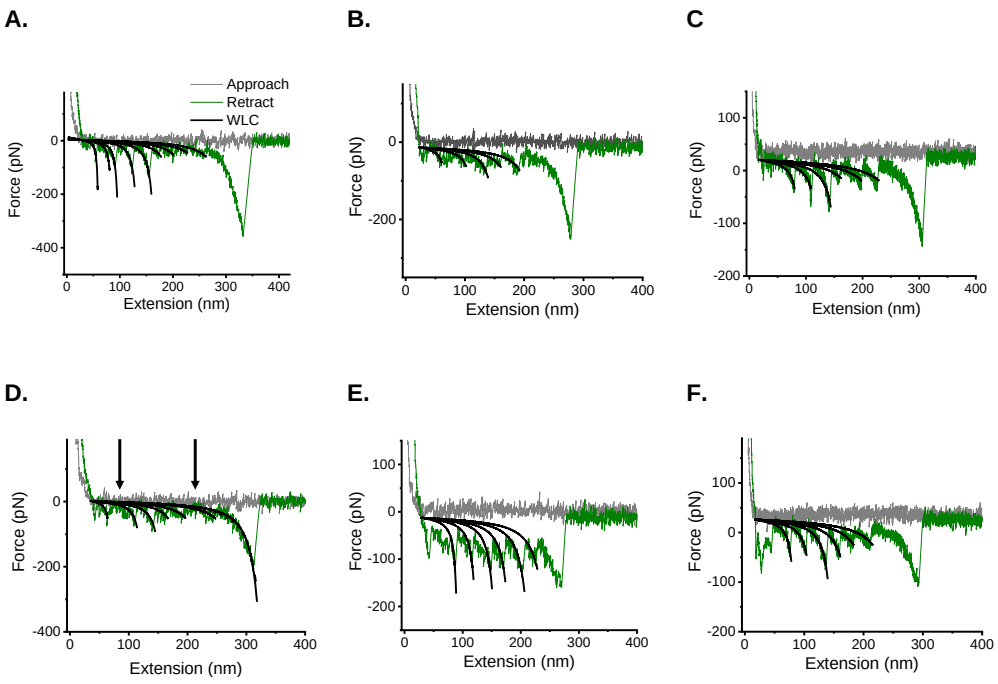


Figure 4.19: Unfolding of D0310_scc. D0310_scc at 0.2-2 µM in 25 mM MES, 150 mM NaCl, pH 6.25 was unfolded using a MLCT-C or MLCT-D cantilever (see Table 2.16) at 21 °C. Unfolding speeds were **A, D.** 1500 nm/s, where arrows indicate unfolding events with a putative contour length of 2ΔL; **B, E.** 800 nm/s; **C, F.** 200 nm/s. Grey: probe approaching the surface; green: probe retracting from the surface; black solid line: WLC fit with p=0.40 nm.

Saw-tooth force-extension traces with at least three DRESS domain unfolding events were fitted to the WLC model (Equation 2.30) to obtain ΔL and F. Histograms of observable parameters were fitted with a Gaussian distribution to obtain the mean and SD (Figure

4.20). DRESS domains unfolded at 109 pN ± 40 pN at a constant speed of 1500 nm/s (Figure 4.20A); this value was obtained from analysis of a single data set, more experimental replicates are available but not yet incorporated in this average number. The ΔL was 34 nm ± 6.5 nm for unfolding events recorded at 200 nm/s and 1500 nm/s as determined from multiple experimental replicates (no difference in ΔL between unfolding speeds; Figure 4.20B). The difference between the expected and the observed ΔL per DRESS domain (6 nm) might suggest a more compact folded state than anticipated, possibly due to the formation of a superhelical spring (see section 6.1.4.2).

In most saw-tooth force-extension traces, unfolding events are observed, where the force does not return to baseline before the next observed unfolding event. This suggests that the next DRESS domain to unfold is mechanically less resilient than the DRESS domain that has just unfolded. This suggests that the domains are stabilised by the presence of adjacent folded domains. Most domains unfold individually, as opposed to cooperatively. Some traces show evidence of tandem domain unfolding, where ΔL approaches the expected increase of contour length for two presumably adjacent domains unfolding simultaneously (see arrows in Figure 4.19D).

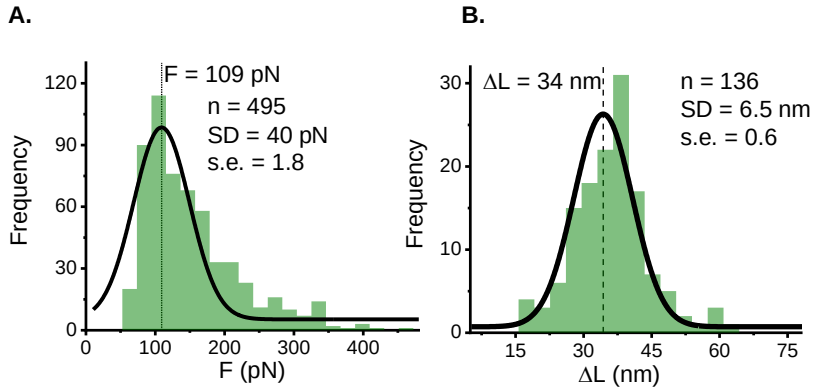


Figure 4.20: Histograms of F and ΔL for mechanical unfolding of individual DRESS domains. A. Unfolding force (F) for 495 unfolding events recorded at a pulling speed of 1500 nm/s (this dataset was analysed from a single experimental replicate). The final protein-detachment event was usually much greater and is therefore not included in this analysis. Unfolding event detection and data analysis was kindly performed by William Rochira. **B.** Increase in contour length per domain (ΔL) for 136 unfolding events recorded at pulling speeds of 200 and 1500 nm/s.

4.3.8.3 Refolding of DRESS domains under mechanical load

The refolding ability of DRESS domains was assessed by first mechanically unfolding the construct, followed by relaxation of the force and refolding was monitored by AFM (Figure 4.21; see section 2.11.4.2). Briefly, the probe approaches the surface in the 'Approach' step

until it touches the surface, where it holds for 0.1 s. During the 'Unfold' step, a constant force is applied of 35 to 50 pN for 6-12 s. At 50 pN, DRESS domains unfold frequently, followed by a flat baseline at constant force (Figure 4.21A,B) or detachment of the protein from the tip (Figure 4.21C). At 35 pN, unfolding followed by refolding under force is observed (Figure 4.21D), but generally, unfolding events are rare at this force. During the 'Refold' step, the constant force is reduced to zero and this is held for 6 to 12 seconds. Refolding happened directly after relaxation of the force within a second. Finally, the protein is detached from the probe in the 'Retract' step and this is characterised by the large negative force required to break the non-specific, non-covalent interaction between the probe and the protein.

Immediately after the 'Hold' step, an initial height increase is observed of ~35-70 nm, which is never fully recovered during the 'Refold' step. Considering that the expected length of the fully folded, extended protein is 44 nm, this height hysteresis might refer to 'pulling' the folded protein vertically, depending on the location of the probe along the rod and the initial folding state of the DRESS domains. When the protein is detached from the probe during the 'Unfold' step, this requirement is broken, as is reflected in Figure 4.21C, where the height of the probe after return to the surface is lower than the initial height increase. Here, the probe does return to the surface, however the rate at which the probe returns to the surface is slower, implying that no pulling force is acting on the probe.

During the 'Unfold' step, force spikes are observed, which usually decrease in amplitude with an increasing number of unfolding events. A similar trend was present in Figure 4.19 and supports the previous suggestion that unfolding of the first DRESS domain weakens adjacent DRESS domains. The increase in height per transition after 'pulling' the protein vertically varies from ~4 to 50 nm. This may reflect unfolding events ranging from individual helices in a single DRESS domain to the simultaneous unfolding of multiple DRESS domains. The fact that ΔL does not always represent a complete DRESS domain, implies that some domains might unfold partially. Finally, in Figure 4.21B, unfolding and refolding of part of a DRESS domain with a ΔL of ~5 nm was observed in quick succession and under constant force.

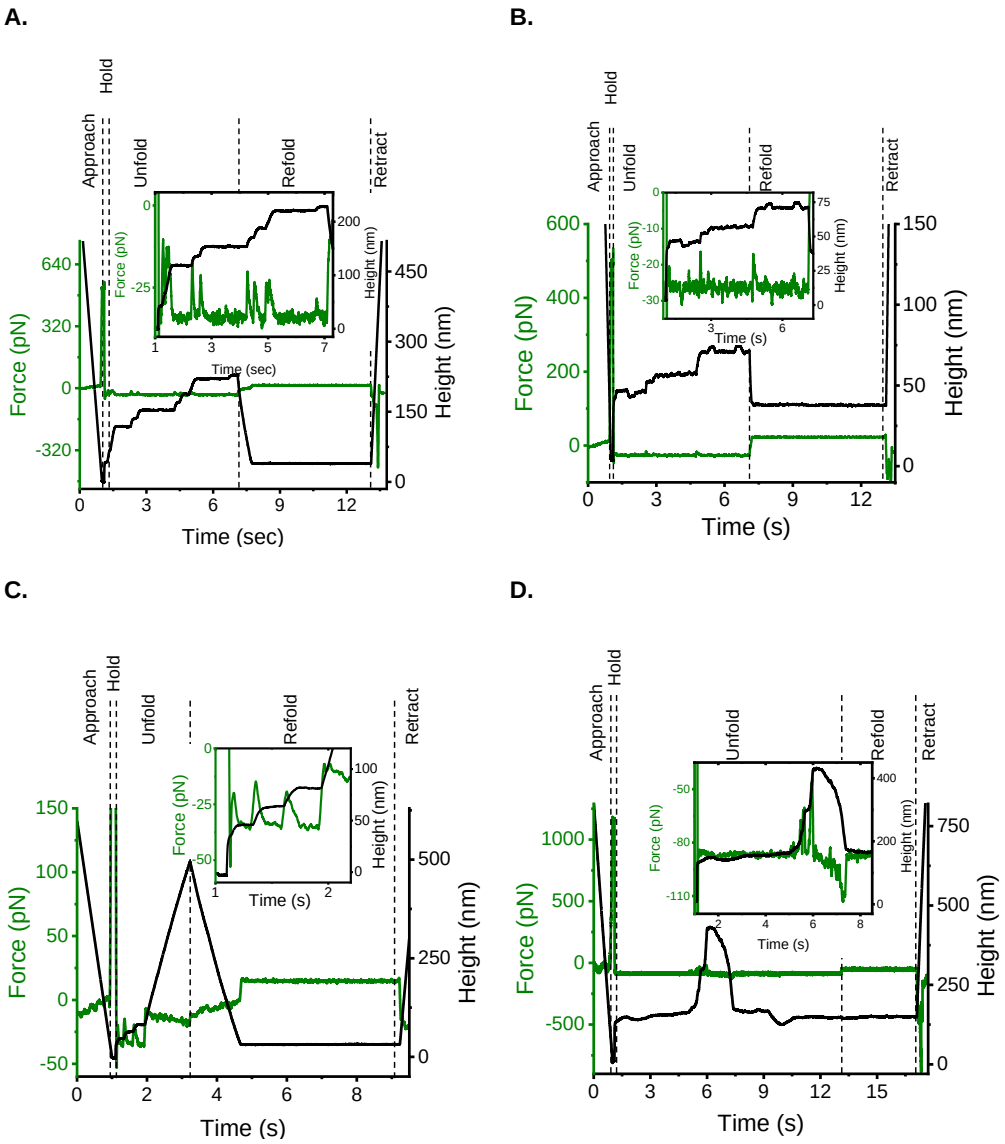


Figure 4.21: Refolding of D0310_scc after unfolding at a constant applied force. A 75-point mean smoothing filter was applied to the force data. **A, B.** Typical refolding behaviour of D0310_scc construct at constant applied force of 50 pN. **C.** Protein detachment from the AFM tip occurs during the unfolding step (applied force of 50 pN), supported by the absence of a large detachment peak. **D.** Unfolding and refolding of D0310_scc construct during constant applied force of 35 pN.

4.4 Conclusions for this chapter

This chapter addresses the end-to-end distance, shape, rigidity, thermal stability and mechanical strength of recombinant repetitive structural domains from the repetitive region of SasC.

The thermal stability studies of recombinant repetitive structural domains suggest that at least some domains unfold simultaneously, suggesting some cooperativity. However, the quality of the thermal denaturation curve for D0118 is insufficient to make conclusions

about the number of transitions and to assign these to parts of the DRESS region of SasC. Follow-up thermal denaturation experiments by both CD and nano-DSF of smaller constructs covering the entire length of the DRESS region would provide more insight into the degree of cooperativity during thermal denaturation; thus, this data is interesting but not yet ready for publication.

The size, shape and an estimation of the end-to-end distance were obtained for four DRESS domains and the physiological repetitive region of SasC by SEC-SAXS. The quality of the data of D0710 was excellent and the quality of the major species of purified D0118 was sufficient for accurate analysis, as is observed from the data plots shown in this chapter. Both datasets conclusively show that recombinant repetitive protein constructs containing multiple DRESS domains are elongated close to the maximum elongation that is expected from the crystal structure of tandem DRESS domains, that the proteins behave as rod-like particles in solution and that *ab initio* modelling of the particles as rigid rods fits best to the experimental data. Thus, this data is of sufficient quality for publication and no further experiments are required to support the statement that these recombinant protein constructs containing DRESS domains from SasC form elongated, rigid, rod-like particles in solution with approximately the expected dimensions in terms of end-to-end distance and diameter.

The estimation of the end-to-end distance of the physiological repetitive region of SasC was further obtained by SHRImP-TIRF microscopy, where the recombinant protein was electrostatically immobilised on a poly-*D*-lysine coated quartz slide. The data presented in this thesis is of sufficient quality to conclude that at pH 7.0 at 2 µg/mL poly-*D*-lysine, the inter-fluorophore distance of D0118_2A488 suggests maximum elongation and implies rigidity. The same was observed at pH 6.5, however the experimental number of replicates is currently insufficient for publication. Although in this super-resolution technique, the standard deviation of the Gaussian fit to the histogram of inter-fluorophore distances seems large, this fit is of sufficient confidence for publication as was previously shown for a similar set-up in Gruszka *et al.*[57].

The mechanical strength and refolding ability of DRESS domains from SasC was studied using AFM. The veracity of unfolding events classified as DRESS domains was confirmed by the correct approximate increase in contour length per unfolding event, determined using

the WLC model with a p appropriate for a single polypeptide chain, and the selection criteria that were used for unfolding curves. Thus, the traces presented are assumed to correctly represent individual DRESS domain unfolding events. The values reported for the average unfolding force are currently obtained from a single dataset and therefore not yet ready for publication. A sufficient number of datasets has been recorded at 200 nm/s and 1500 nm/s; obtaining a publication-ready average unfolding force and contour length for DRESS domains at these unfolding speeds is currently in progress. The refolding ability of DRESS domains shown in this thesis is to be used as preliminary evidence of mechanical refolding. Analysis of the increase in contour length per unfolding/refolding step and an attempt at repeated unfolding/refolding of a recombinant protein containing multiple DRESS domains is required prior to publication.

Chapter 5. Dynamic studies of SHIRT domains

5.1 Introduction

5.1.1 Background of SGO0707

S. gordonii initiates the formation of dental plaque[432] (see section 1.4.4) by adhering to the tooth surface[18] and mediating cell-cell interactions[433], both via CWA proteins[433,434]. However, *S. gordonii* can become opportunistic[18] and spread to non-oral sites and cause, for example, infective endocarditis[435].

S. gordonii expresses various surface adhesin proteins with functions in adhesion and cell-cell accumulation[433,434]. A CWA protein encoded by a gene with accession number *sgo_0707*[230], here termed SGO0707, adheres to oral keratinocytes and type I collagen putatively via the A region[231], but is not involved in cell-cell accumulation[231]. Interestingly, although type I collagen is the most abundant protein in the human body[54], it is usually not readily available in the oral cavity[436,437], suggesting that SGO0707 may play a role in pathogenesis[231] by attaching to type I collagen at other sites, for example.

The A region of SGO0707 is hypothesised to be involved in collagen binding[231]. The B region comprises thirteen SGO0707 high-identity repeat tandem (SHIRT) domains, of which the function is currently unknown. The repetitive region of SGO0707 is hypothesised to form an elongated stalk[231], and this hypothesis is confirmed by preliminary experiments (F. Whelan and G. Gilburt *et al.*, manuscript in preparation).

5.1.2 Domain boundaries of SHIRT domains

The repetitive region of SGO0707 comprises SHIRT domains. The domain boundaries of SHIRT domains were determined by Dr Fiona Whelan by crystallography. The structure of SHIRT domain S02_offset was determined, in which the correct domain boundaries appeared offset by five residues (Figure 5.1A). This allowed the correct domain boundaries to be proposed, which were validated by the structure of SHIRT domains 3 and 4 (S0304, Figure 5.1B). S0304 forms an elongated structure with SHIRT domains in a head-to-tail conformation.

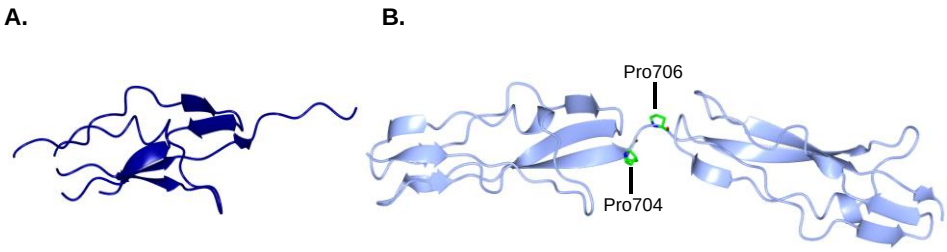


Figure 5.1: Crystal structure of SHIRT domains in SGO0707 (data courtesy of Dr F. Whelan, manuscript in preparation). **A.** Crystal structure of S02_offset. **B.** Crystal structure of S0304 with Pro704 and Pro706 (green). Image was created using CCP4mg.

The distance between domains S03 and S04 (>7 Å) is too large for stabilising interactions to occur (Figure 5.1B). However, the tandem domains are in an extended conformation in the crystal structure, linked by Pro704-Ala705-Pro706. Proline residues impose structural rigidity on a polypeptide chain[438,439]. However, it is unclear whether the observed extension of tandem domains is due to the proline-rich linker sequence or due to the packing of S0304 in the crystal lattice.

5.1.3 Sequence alignment of SHIRT domains

The average pairwise protein sequence identity between SHIRT domains is 88% (Figure 5.2). S03 and S04 have a pairwise identity of 98%. Met631 in S03 is equivalent to Val715 in S04 and Ala673 in S03 is equivalent to Thr757 in S04. Here, experimental work is focused on S0304 with the linker sequence Pro704-Ala705-Pro706.

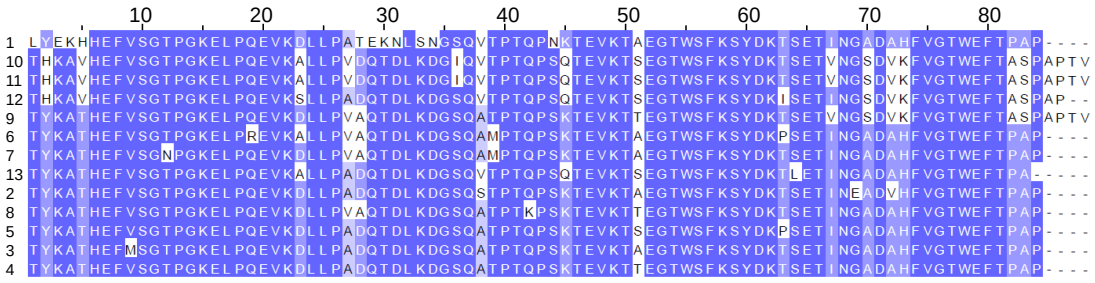


Figure 5.2: MSA of SHIRT domains by Clustal Omega[259] and coloured by percent identity in Jalview[263].

5.1.4 Relevance of studying repetitive domains

CWA proteins with highly identical tandemly arrayed repeats in their B region are an exception in the well-regarded view that a lower sequence identity avoids domain aggregation[211]; thus, it is expected that highly identical tandem arrays confer another function. Indeed, changing the number of repeats was suggested to be an immune evasion strategy for a surface antigen protein of group B streptococci, alpha C[440]. Rib (resistance to

proteases, immunity, group B) is part of the alpha-like family[441] and comprises a variable number of repeats with a sequence conservation up to 100% in its B region with a lower sequence conservation for the 'capped' domains[441]. Proteins with SHIRT domains also have a variable number of SHIRT repeats (Interpro entry IPR041030; analysis performed by Dr A. Bateman; F. Whelan *et al.*, manuscript in preparation).

However, molecular biology techniques and biophysical characterisation of high-identity tandem repeats can be challenging. Thus, it is important to generate tools that allow the study of tandemly arrayed, high-identity repetitive domains.

5.2 Aims

This chapter aims to characterise the flexibility/rigidity in SHIRT domains and in the linker between tandem SHIRT domains. To fulfil this goal, the following aims were set:

- To introduce mutations in the linkers between tandem SHIRT domains to assess the effect of the substitution of the proline residues;
- To assign the (^{1}H,^{15}N)-HSQC spectrum of SHIRT and tandem SHIRT domains;
- To assess the flexibility of the linker between SHIRT domains through experiments that probe the dynamics of the protein backbone.

5.3 Results

5.3.1 Molecular biology for SHIRT domains

SHIRT domains 03 and 04 were selected for recombinant gene expression and protein production, purification and biophysical characterisation from the repetitive region of SGO0707, because structural information is available to identify the residues in the loop region (Figure 5.3).

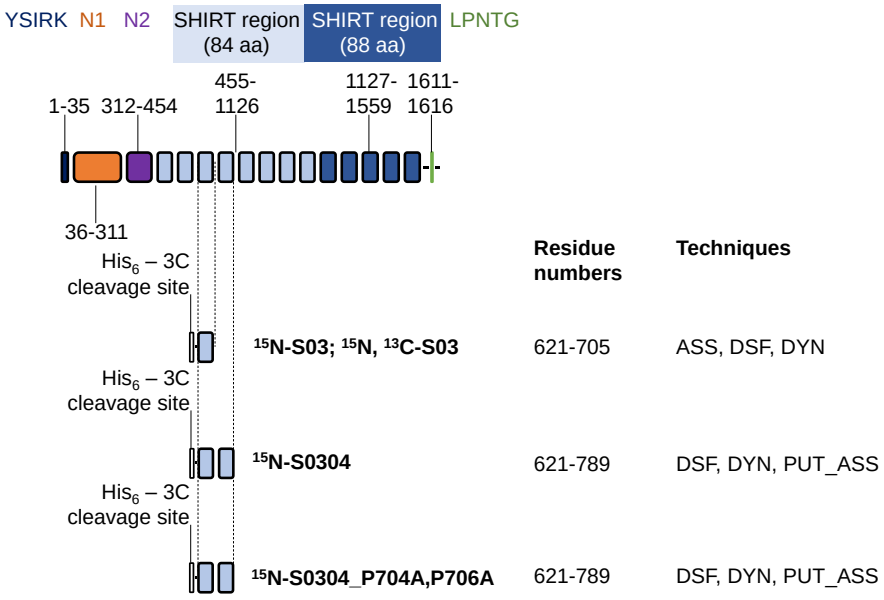


Figure 5.3: Schematic of SGO0707 with protein targets for this chapter. Domains scaled to relative size (residue numbers shown). Techniques: ASS: triple-resonance backbone assignment; DSF: nano-DSF; DYN: backbone dynamics experiments (T1, T2, (^{1}H, ^{15}N)-hnNOE); PUT_ASS: putative backbone assignment.

To assess their role in restricting the flexibility of the loop region, the proline residues were substituted for alanine residues. Alanine was chosen because of a higher-than-average occurrence in native linkers[442,443], a more comparable hydrophobic nature to Pro than Ser/Thr[442] and a limited contribution to intrinsic flexibility, as opposed to Gly[444].

The DNA encoding S03 and S0304 was available in the pETFPP expression vector (courtesy of Dr Fiona Whelan). The DNA encoding S0304_P704A,P706A was ordered. To lower the DNA sequence identity between S03 and S04 to 80%; codons in redundant regions were mutated to different codons of high occurrence in *E. coli*. This facilitated the design of non-redundant In-Fusion primers. The resulting DNA-sequence was ordered from Genewiz (see Appendix 7.4).

5.3.2 Over-production and purification of ^{15}N, ^{13}C- and ^{15}N-labelled SHIRT domains

Recombinant gene expression and protein production conditions were screened for optimal production of uniformly isotopically labelled single and tandem SHIRT domains in BL21-Gold (DE3) cells. Typically, cells were grown at 37 °C in minimal media (see Table 2.4) supplemented with the appropriate isotope/isotopes, until an OD_{600} of 0.6 was reached. Recombinant gene expression and protein production was induced by the addition of IPTG

to a final concentration of 1 mM and cells were then grown at 20 °C for 18-22 hours. An example purification is shown for ^{15}N-S0304.

Briefly, cells were lysed, ^{15}N-His_6-3C-S0304 was purified from the soluble cell extract by IMAC and was competitively eluted using an increasing gradient of elution buffer (see Table 2.4; Figure 5.4A, B) containing an increasing concentration of imidazole ranging from 20 mM (0%) to 500 mM (100%). Fractions 6-10 and 11-23 were pooled and the His_6-3C tag was cleaved from ^{15}N-S0304 using HRV 3C protease in a mass ratio of protease to target protein of 1:150 for 16-20 hours at 4 °C (Figure 5.4C, shown for fractions 6-10). A diffuse band around 17 kDa (grey arrow) that appeared after cleavage of the His_6-3C tag, and HRV 3C protease were removed from ^{15}N-S0304 by a second round of IMAC (Figure 5.4D,E; shown for fractions 6-10 from Figure 5.4A). ^{15}N-S0304 eluted in the flow-through and the His_6-3C tag and protease were competitively eluted by a stepwise increase in the concentration of imidazole from 20 mM (0%) to ~260 mM (50%). ^{15}N-S0304 was concentrated by spin filtration and dialysed into 20 mM sodium phosphate, pH 6.0.

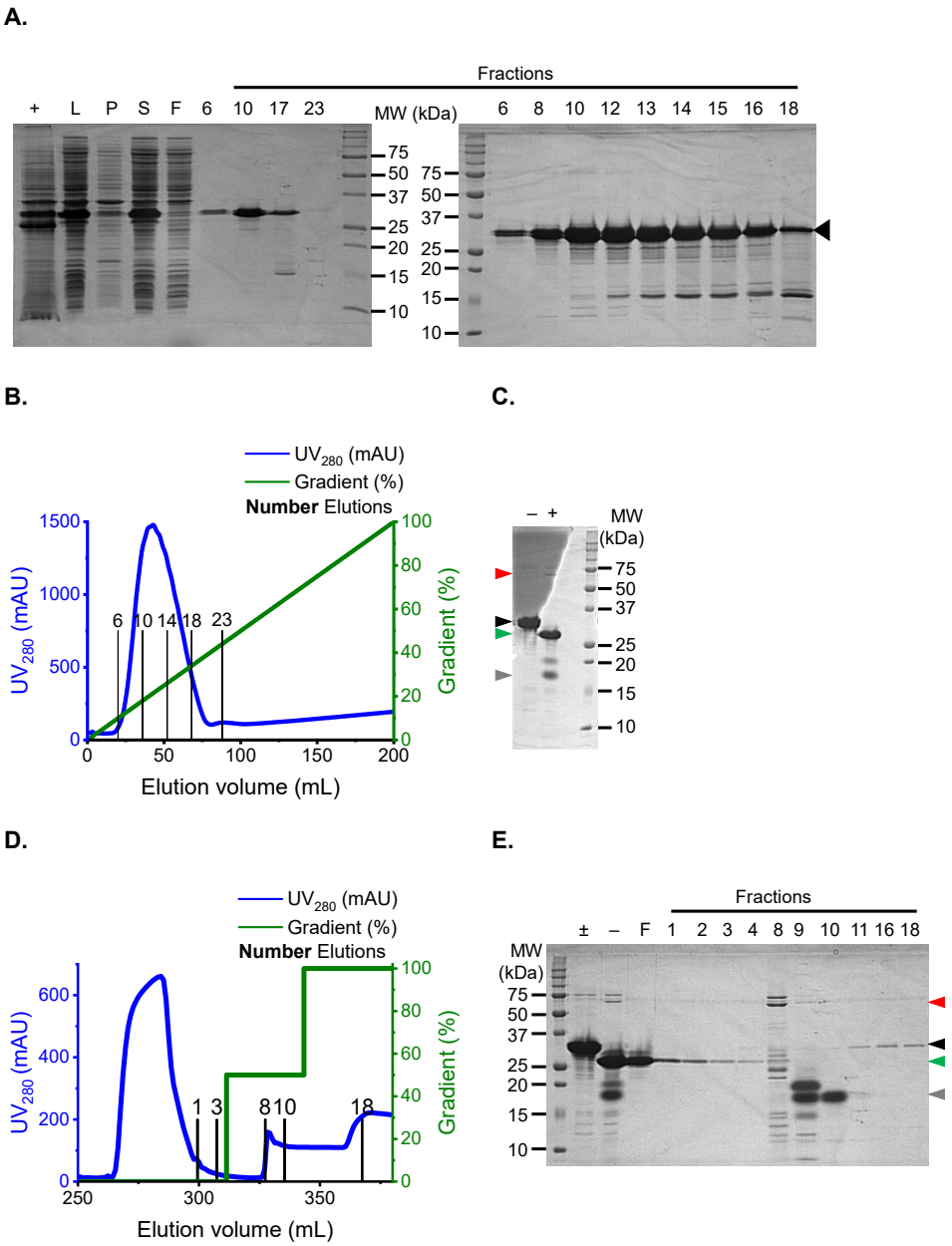


Figure 5.4: Purification of ^{15}N-S0304. SDS PAGE analyses on 15% (*w/v*) polyacrylamide gels. **A.** SDS PAGE analysis of IMAC of ^{15}N-His$_6$-3C-S0304 (theoretical MW 20.9 kDa, black arrow). +: total fraction after 22 hours. L: lysate; P: insoluble material; S: soluble material; F: flow-through. Samples in left gel were diluted six times compared to right gel. **B.** IMAC chromatogram as monitored by A_{280}. 4 mL fractions were collected, numbers correspond to A. **C.** SDS PAGE analysis of ^{15}N-His$_6$-3C-S0304 (black arrow) before (–) and after (+) HRV 3C protease (red arrow) cleavage into ^{15}N-S0304 (theoretical MW 19.0 kDa, green arrow) and His$_6$-3C tag (putatively assigned by grey arrow). Brightness was adjusted by +20% post-acquisition. Raw gel image is available in Appendix 7.3. **D.** ^{15}N-S0304 was separated from the His$_6$-3C tag and HRV 3C protease by IMAC as monitored by A_{280}. **E.** SDS PAGE analysis of IMAC monitored by A_{280}. ±: before HRV 3C cleavage. –: before IMAC2. F: flow-through. Arrows correspond to C and fractions correspond to D.

^{15}N-S03; ^{15}N-, ^{13}C-S03 and ^{15}N-S03_P704A,P706A were purified in the same manner as ^{15}N-S0304. The purity of recombinant proteins was over 95%, as estimated from SDS PAGE analysis (Figure 5.5), the yield is reported in Table 5.1 and the correct mass and uniform labelling of recombinant proteins were confirmed by MS (Equation 5.1; Table 5.2).

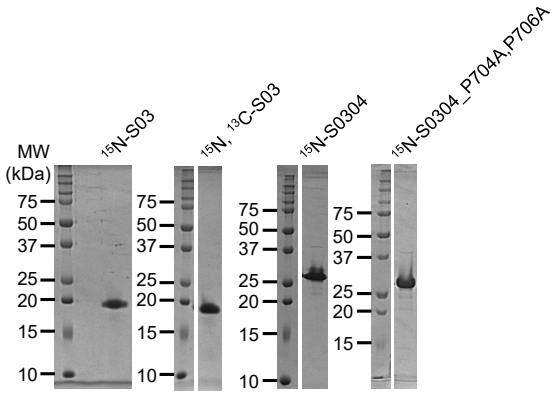


Figure 5.5: SDS PAGE analysis of purified recombinant proteins from this chapter on 15% (*w/v*) polyacrylamide gels. Brightness of ^{15}N-S0304 was adjusted by +20% post-acquisition. Raw, non-assembled gel images are shown in Appendix 7.3.

Table 5.1: Final yields of recombinant proteins used in this chapter. Yield is displayed in mg purified target protein per litre medium. M: minimal media (see Table 2.2). pI and ε were determined by ExPASy ProtParam[253].

Protein	Yield (mg L^{-1})	Media	Last purification step	pI	ε (M^{-1} cm^{-1})
^{15}N-S03	60	M	IMAC 2	4.80	13980
^{15}N, ^{13}C-S03	26	M	IMAC 2	4.80	13980
^{15}N-S0304	94	M	IMAC 2	4.79	27960
^{15}N-S0304_P704A, P706A	33	M	IMAC 2	4.79	27960

Equation 5.1: Calculation of labelling efficiency.

$$\%\ labelling = 100\% - \frac{\Delta m}{\#isotopes}$$

where Δm is the mass difference between the observed and theoretical MW and $\#isotopes$ is the expected number of isotope atoms (see Table 5.2).

Table 5.2: Molar masses of single and tandem SHIRT domains as determined from ESI MS. [a]Average labelling across ^{15}N and ^{13}C isotopes.

Protein	Expected number of isotope atoms		MW (Da)		Labelling efficiency	
	^{15}N	^{13}C	Theoretical	MS	Δm (Da)	% labelling
^{15}N-S03	108	N/A	9712.6	9710.0	-2.6	97.6
^{15}N, ^{13}C-S03	108	425	10137.6	10128.4	-9.2	[a]98.3
^{15}N-S0304	211	N/A	18972.6	18970.3	-2.3	98.9
^{15}N-S0304_P704A, P706A	211	N/A	18920.5	18917.9	-2.6	98.8

5.3.3 Thermal stability of SHIRT domains

The thermal stability of SHIRT domains was determined by nano-DSF (see section 2.6.3, Figure 5.6A). The T_m values of ^{15}N-S03 and ^{15}N-S0304 are very similar (Table 5.3) and do not imply a significant inter-domain stabilisation. The T_m of ^{15}N-S0304_P704A,P706A is approximately one degree lower than that of S0304 (Table 5.3), suggesting that the proline-to-alanine mutations were not disruptive. SHIRT domains do not show signs of aggregation during thermal denaturation (Figure 5.6B).

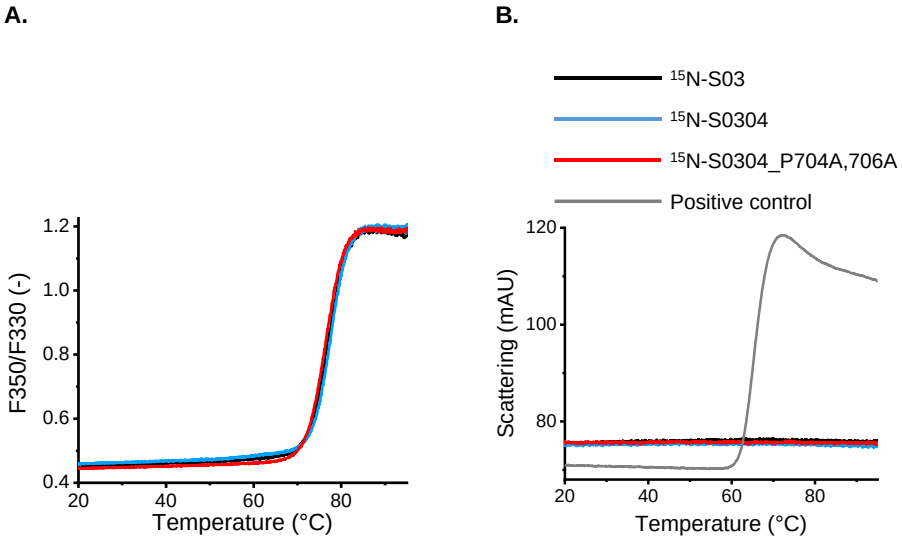


Figure 5.6: Thermal denaturation and aggregation of ^{15}N-S03 (black), ^{15}N-S0304 (blue) and ^{15}N-S0304_P704A,P706A (red) monitored by **A.** fluorescence ratio and **B.** static light scattering. Measured on 1 mg/mL protein in 20 mM sodium phosphate buffer, pH 6.0 (^{15}N-S03) or pH 6.5 (^{15}N-S0304, ^{15}N-S0304_P704A,P706A). Positive aggregation control in B (in grey) is D1617 in 25 mM MES, 2 M NaCl, pH 6.0.

Table 5.3: T_m values of SHIRT domains.

Protein	T_m (°C)
^{15}N-S03	77.1
^{15}N-S0304	77.6
^{15}N-S0304_P704A,P706A	76.4

5.3.4 (^{1}H,^{15}N)-HSQC-spectra of SHIRT domains

The two-dimensional (^{1}H,^{15}N)-HSQC spectrum shows the correlation between the chemical shifts of directly bound ^{1}H and ^{15}N nuclei. The dispersion of ^{1}H correlations provides information about the fold of the protein in solution[445]. Therefore, two-dimensional (^{1}H,^{15}N)-HSQC-spectra of S03, S0304 and S0304_P704A,P706A were acquired (Figure 5.7). (^{1}H,^{15}N)-HSQC-spectra were recorded in 20 mM sodium phosphate buffer at pH-values 5.5, 6.0 and 6.5 to verify that SHIRT domains were stably folded in this pH range (data not shown for pH 5.5 and 6.5). All three proteins showed a wide dispersion in ^{1}H correlations in all three conditions, indicating that they have a stable fold in solution.

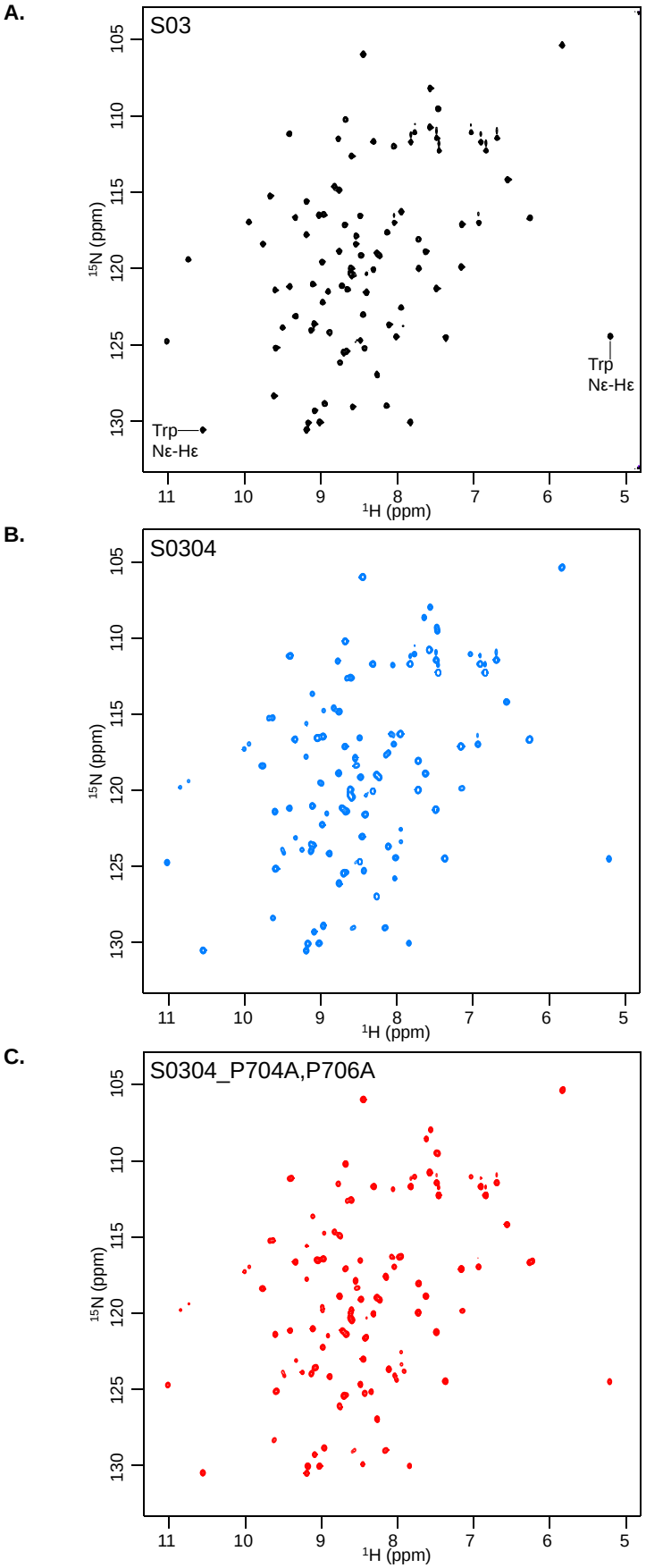


Figure 5.7: (^{1}H,^{15}N)-HSQC spectra in 20 mM sodium phosphate, pH 6.0 of **A.** S03, with Trp Nε-Hε side chains indicated, **B.** S0304 and **C.** S0304_P704A,P706A.

(^{1}H,^{15}N)-HSQC-spectra at pH 6.0 typically show ^{1}H, ^{15}N resonances from backbone amide N-H bonds, Trp side-chain Nε-Hε bonds and bonds in $–NH_2$ side chain groups from Arg, Asn and Gln. All amino acids in proteins except proline residues and the N-terminal residue are expected to give rise to backbone amide resonances. The resonances of Nε-Hε bonds in Trp side chains were identified by selective unlabelling of Trp by the addition of excess unlabelled Trp to the growth medium (data courtesy of Dr Michael Plevin, positions indicated in Figure 5.7A). The position for the Trp Nε-Hε resonance at 5.2 ppm resides three SDs from the average ^{1}H ppm shift at ~10 ppm[446]. Here, the resonance has likely shifted due to ring current effects from proximal aromatic residues (F. Whelan *et al.*, manuscript in preparation).

The positions of $–NH_2$ groups from Asn and Gln side chains are identified by the paired correlations with identical ^{15}N chemical shift but different ^{1}H chemical shift. S03 contains one Asn residue and four Gln residues and this matches the observed number of paired resonances (Figure 5.7A). S03 contains no Arg residues. Thus, the expected number of backbone amide resonances was observed in the (^{1}H,^{15}N)-HSQC-spectrum of S03 (Table 5.4).

SHIRT domains are tandem repeats with high sequence identity (see section 5.1.3). This is corroborated by the crystal structures of S0304, whose SHIRT domains have a backbone RMSD of 0.14 Å for the superposition of backbone atoms from 82 residues. Thus, SHIRT domains are very similar in sequence and structure. Therefore, it is expected that equivalent backbone amide bonds in S03 and S04 reside in a very similar chemical shift environment and give rise to near-identical (^{1}H, ^{15}N) resonances. The (^{1}H,^{15}N)-HSQC-spectrum of S03 shows the expected number of backbone amide resonances (Table 5.4), but only 58% and 60% of expected backbone amide resonances are observed for S0304 and S0304_P704A,P706A, respectively (Table 5.4); suggesting significant overlap of equivalent resonances in S03 and S04. Furthermore, the shift in only a small number of resonances in S0304/ S0304_P704A,P706A compared to S03 suggests that the chemical shift of backbone resonances is only influenced to a small extent by the presence of an adjacent SHIRT domain and inter-domain interface. Hence, S03 and S04 appear to have a minimal inter-domain interface, as expected from the crystal structure (Figure 5.1B) and similar T_m values (Table 5.3).

Table 5.4: Expected number of backbone amide resonances for S03, S0304, S0304_P704A,P706A.

Protein	Residue number	Number of proline residues	Expected number of backbone amide resonances	Observed number of backbone amide resonances (% of total expected)
S03	88	8	79	79 (100%)
S0304	173	15	157	91 (58%)
S0304_P704A, P706A	173	13	159	96 (60%)

To study the ^{15}N relaxation properties of the backbone amide resonances in SHIRT domains, assignment is required. Overlapping resonances cannot be assigned to a specific domain due to ambiguity. However, for most non-overlapping resonances in the (^{1}H,^{15}N)-HSQC-spectra of S0304 and S0304_P704A,P706A it is clear to which resonance in S03 they are equivalent. Therefore, the backbone amide resonances of S03 are assigned.

5.3.5 Assignment of backbone resonances of S03

The assignment of the backbone amide resonances of S03 were required to study the relaxation properties of backbone amide residues in the linker between SHIRT domains. To this end, S03 was uniformly labelled with ^{13}C and ^{15}N isotopes and three-dimensional triple-resonance NMR spectra were acquired of 1.1 mM ^{15}N-, ^{13}C-S03 in 20 mM sodium phosphate, pH 6.0, 10% (*v/v*) D_2O at 700 MHz, 25 °C (see Table 2.13).

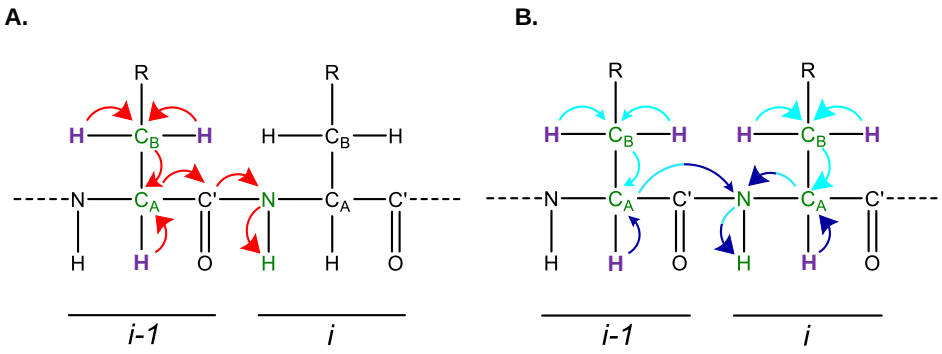


Figure 5.8: Schematics of three-dimensional experiments used in the assignment of backbone amide resonances. Detected nuclei are shown in green and nuclei through which magnetisation is transferred are shown in bold in purple. The J-couplings via which the magnetisation is transferred are shown in red for **A.** CBCA(CO)NH and in blue for **B.** CBCANH, where the phase for the magnetisation detected on C_B is negative (light blue) and on C_A is positive (dark blue) and the magnetisation from residues from *i-1* are weaker (smaller arrows) than from residue *i* (larger arrows).

The C_A, C_B, $^1H^N$ and $^{15}N^H$ resonances of ^{13}C, ^{15}N-S03 were assigned based on CBCA(CO)NH and CBCANH spectra. In the CBCA(CO)NH experiment (Figure 5.8A), the $^{15}N^H$ and $^1H^{15}$ resonances of residue *i* are correlated with the $^{13}C_A$ and $^{13}C_B$ resonances from residue *i-1*[447]. In the CBCANH experiment (Figure 5.8B), the $^{15}N^H$ and $^1H^{15}$ resonances of residue *i* are correlated with the $^{13}C_A$ and $^{13}C_B$ resonances from residues *i* and *i-1*[448]. $^{13}C_B$ resonances are opposite in sign to $^{13}C_A$ resonances in CBCANH spectra, due to phase inversion, which facilitates the distinction between C_A and C_B resonances within strips (Figure 5.9B)[448]. The sensitivity of the CBCANH experiment for $^{13}C_A$ and $^{13}C_B$ resonances from residue *i-1* is lower than from residue *i.* Therefore, CBCA(CO)NH and CBCANH experiments complement each other to unambiguously perform sequential backbone amide assignment of ^{13}C, ^{15}N-labelled proteins[447,448]. An example of the resonance assignment from S03 is shown in Figure 5.9.

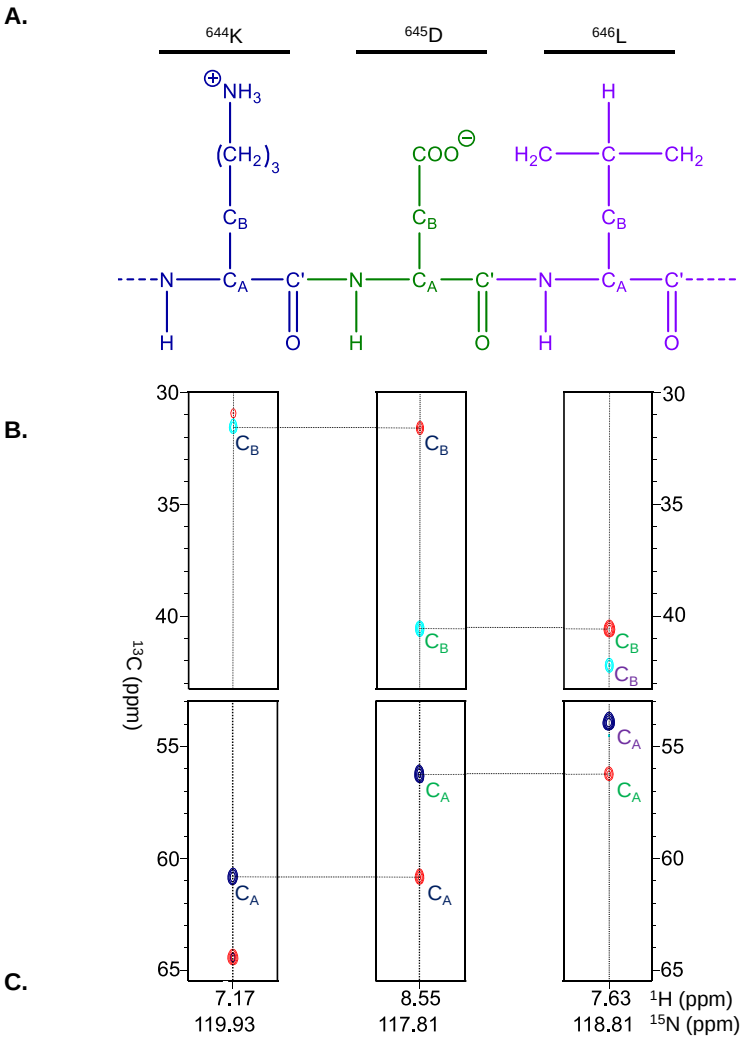


Figure 5.9: Example of backbone amide assignment of S03 using CBCANH and CBCA(CO)NH experiments. A. Schematic representation of ^{644}Lys-^{645}Asp-^{646}Leu tripeptide from S03. **B.** Strips of CBCA(CO)NH and CBCANH spectra for each residue of the tripeptide. CBCA(CO)NH resonances are shown in red, positive CBCANH resonances in dark blue and negative CBCANH resonances in light blue. Horizontal lines indicate the sequential assignment process. **C.** ^{1}H and ^{15}N resonances of the strips.

Sequential backbone amide ordering results in chains of $^{15}N^{H}$ resonances with corresponding $^{13}C_A$ and $^{13}C_B$ resonances. As proline residues lack backbone amide protons, they separate the sequence of S03 into chains ranging from two to 48 residues. The strips of backbone amide resonances are then matched with their respective residues in the sequence of the protein. Generally, the $^{13}C_A$ shift (57 ppm, on average[446]) is larger than the $^{13}C_B$ shift (41 ppm, on average[446]), except for Ser ($^{13}C_B$ 64 ppm, on average[446]) and Thr ($^{13}C_B$ 69 ppm, on average[446]). Furthermore, the $^{13}C_B$ shift for Ala at $^{13}C_B$ of 19 ppm, on average[446], is characteristically small. Finally, glycine residues lack $^{13}C_B$ chemical shifts and their $^{13}C_A$ chemical shift is negative in sign. These properties facilitate matching chains of $^{15}N^{H}$

resonances to the S03 sequence. The resulting backbone assignment of S03 is shown in Figure 5.10. The list of assigned resonances for S03 is shown in Appendix 7.5.

A.

```
  621 625  630  635  640  645  650  655  660
GPAMAPTYKATHEFMSGTPGKELPQEVKDLLPADQTDLKDGSQA

  665  670  675  680  685  690  695  700  705
TPTQPSKTEVKTAEGTWSFKSYDKTSETINGADAHFVGTWEFTPA
```

B.

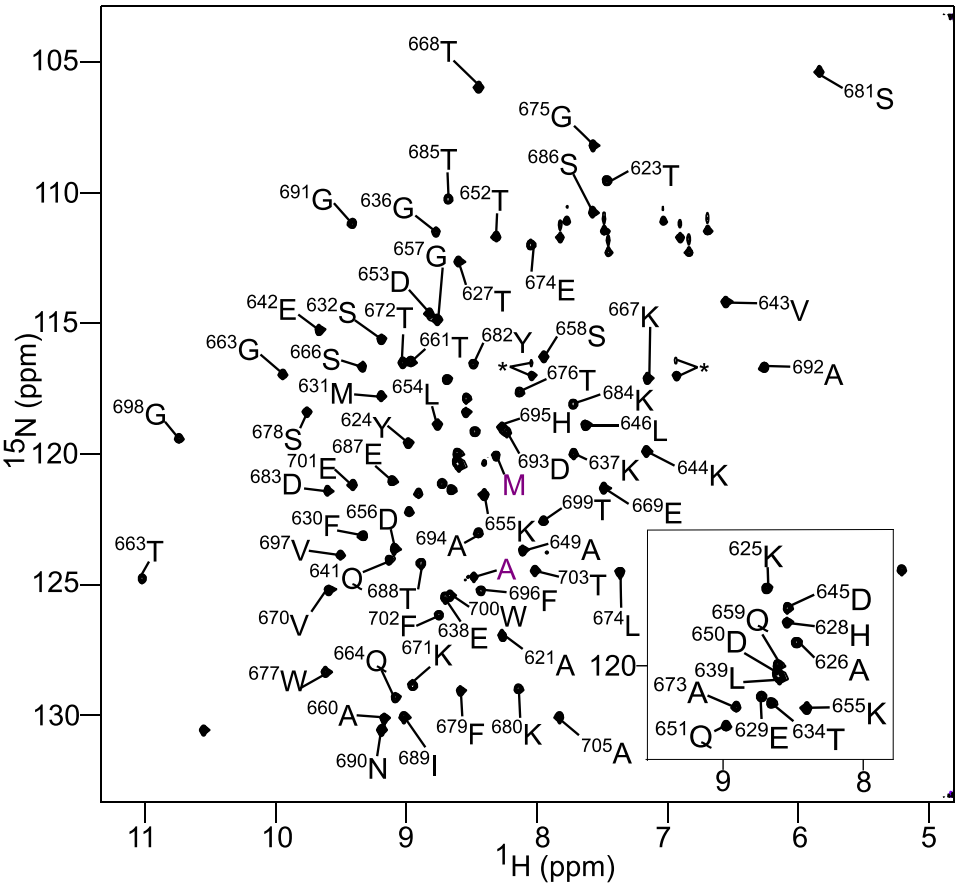


Figure 5.10: (^{1}H,^{15}N)-HSQC spectrum of S03. A. Sequence of S03. **B.** Assigned (^{1}H,^{15}N)-HSQC spectrum of S03. Inset: zoom of assigned (^{1}H,^{15}N)-HSQC spectrum. The residue numbering is as in SGO0707. Residues remaining from fusion tag are shown in purple. * represent the side chain (^{1}H,^{15}N) resonances of ^{641}Q.

5.3.6 Putative assignment of S0304 and S0304_P704A,P706A

To investigate the ^{15}N relaxation properties of tandem SHIRT domains, the non-overlapping backbone amide resonances in the (^{1}H,^{15}N)-HSQC spectra of S0304 and S0304_P704A,P706A need to be assigned. As sequential backbone assignment for tandem domains with high sequence identity and many overlapping resonances leads to

ambiguous sequential assignment, the backbone assignment of S03 (Figure 5.10) was transferred to spectra of S0304 and S0304_P704A,P706A. This approach is regularly used in the interpretation of (^{1}H,^{15}N)-HSQC-spectra of point mutants of wild-type proteins[449–451] or the (marginal) shift of resonances in the presence of different binding partners[452], when the backbone assignment of the reference protein is available.

Here, the transfer of backbone assignments from S03 to tandem domains is valid, because of the large overlap of resonances between single and tandem SHIRT domains (Figure 5.11). 15 resonances in S0304 (8.7% of residues) and 17 resonances in S0304_P704A,P706A (9.8% of residues) do not fully overlap with resonances in S03. In most cases, one of the putatively equivalent resonances in tandem SHIRT domains overlaps with a resonance in S03, while another resonance in its proximity does not overlap with resonances in S03.

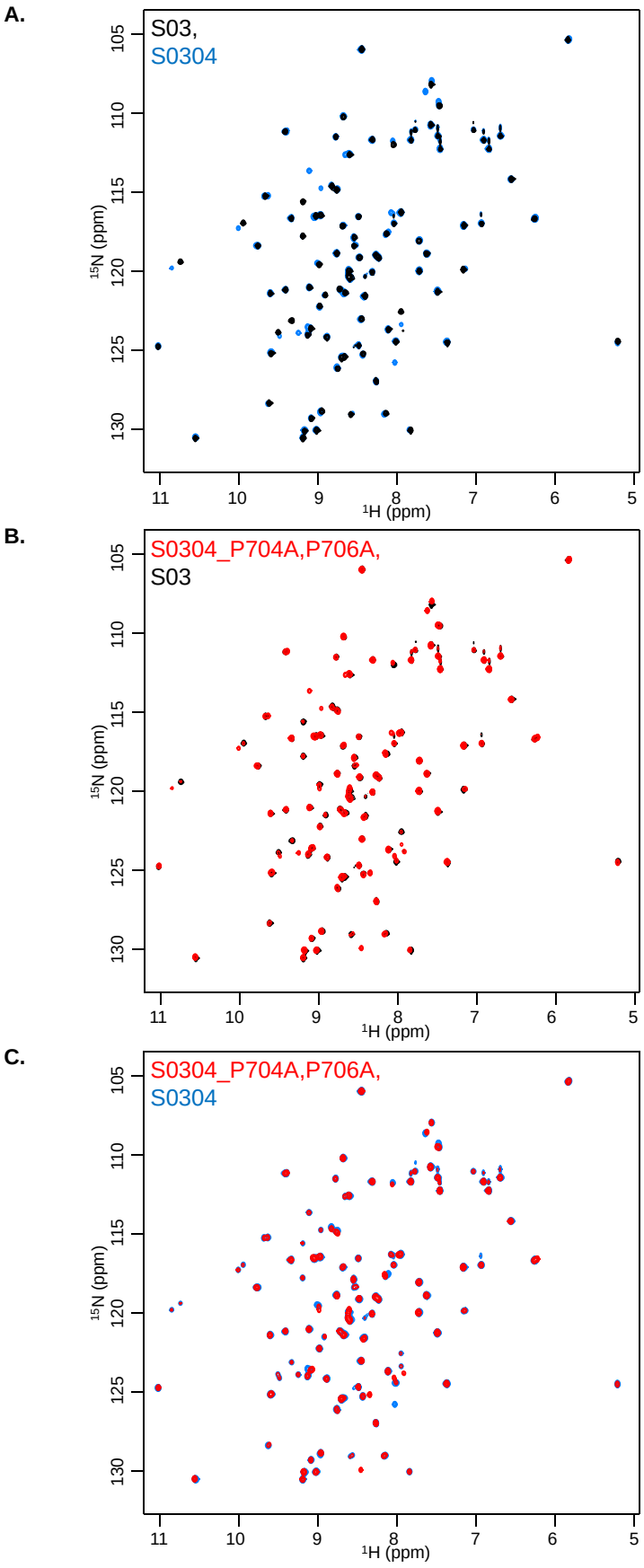


Figure 5.11: Overlays of (^{1}H,^{15}N)-HSQC spectra of **A.** S03 (black) and S0304 (blue), **B.** S03 (black) and S0304_P704A,P706A (red) and **C.** S0304 (blue) and S0304_P704A,P706A (red).

To determine which non-overlapping equivalent resonance belongs to which SHIRT domain in S0304 or S0304P_704A,P706A, the crystal structure of S0304 is used (Figure 5.12). The overlapping resonance is assigned to the equivalent residue furthest away from the domain interface, as likely this chemical environment is nearly identical to that in S03. Conversely, the proximal non-overlapping resonance in S0304 or S0304_P704A,P706A is assigned to the equivalent residue closest to the domain interface, as this chemical environment is likely different from that in S03. In some cases, such as for ^{697}V/ ^{781}V, the position of equivalent residues in the crystal structure is not sufficient to assign the non-overlapping residues of tandem SHIRT domains. Here, the residue context is taken into account. The adjacent equivalent residues ^{698}G/ ^{782}G are also non-overlapping and ^{698}G is close to the linker, while ^{782}G is in an environment similar to S03 (data not shown). Therefore, ^{781}V and ^{782}G are assigned to the resonances superposed with that of S03 and ^{697}V and ^{698}G are assigned to the non-overlapping residue.

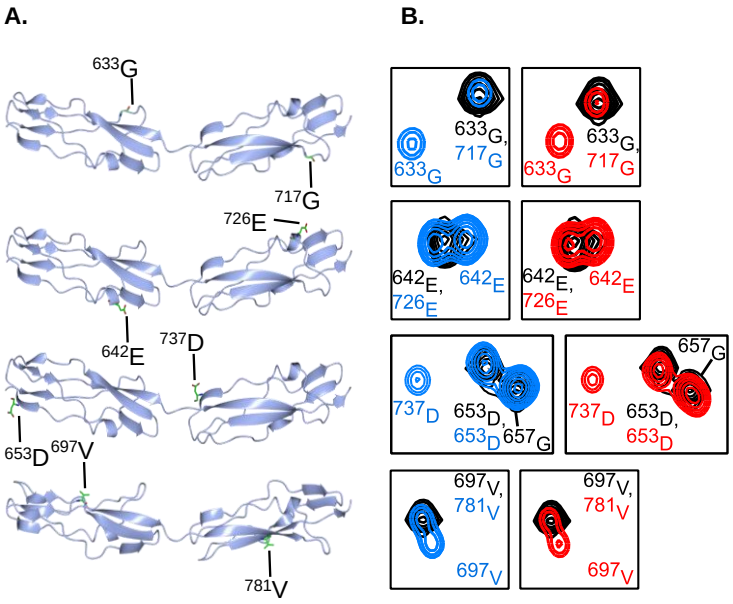


Figure 5.12: Examples of backbone amide assignment strategy of S0304 and S0304_P704A,P706A. A. Locations of non-overlapping residues in the crystal structure of S0304. Image was created using CCP4mg. **B.** Putatively assigned non-overlapping resonances from (^{1}H,^{15}N)-HSQC spectra of S0304 (blue) and S0304P&704A,P706A (red) superposed with the matching assigned resonance in S03 (black).

The putative backbone amide assignment procedure was repeated for all non-overlapping resonances and the resulting putative backbone amide assignments are shown for S0304 (Figure 5.13) and S0304_P704A,P706A (Figure 5.14). The lists of assigned resonances for S0304 and S0304_P704A,P706 are shown in Appendices 7.6 and 7.7, respectively.

A.

```
          621  625   630   635   640   645   650   655   660
S03  GPAMAPTYKATHEFMSGTPGKELPQEVKDLLPADQTDLKDGSQA
               710   715   720   725   730   735   740
S04        PTYKATHEFVSGTPGKELPQEVKDLLPADQTDLKDGSQA

          665   670   675   680   685   690   695   700   705
S03  TPTQPSKTEVKTAEGTWSFKSYDKTSETINGADAHFVGTWEFTPA
           750   755   760   765   770   775   780   785
S04  TPTQPSKTEVKTTEGTWSFKSYDKTSETINGADAHFVGTWEFTPA
```

B.

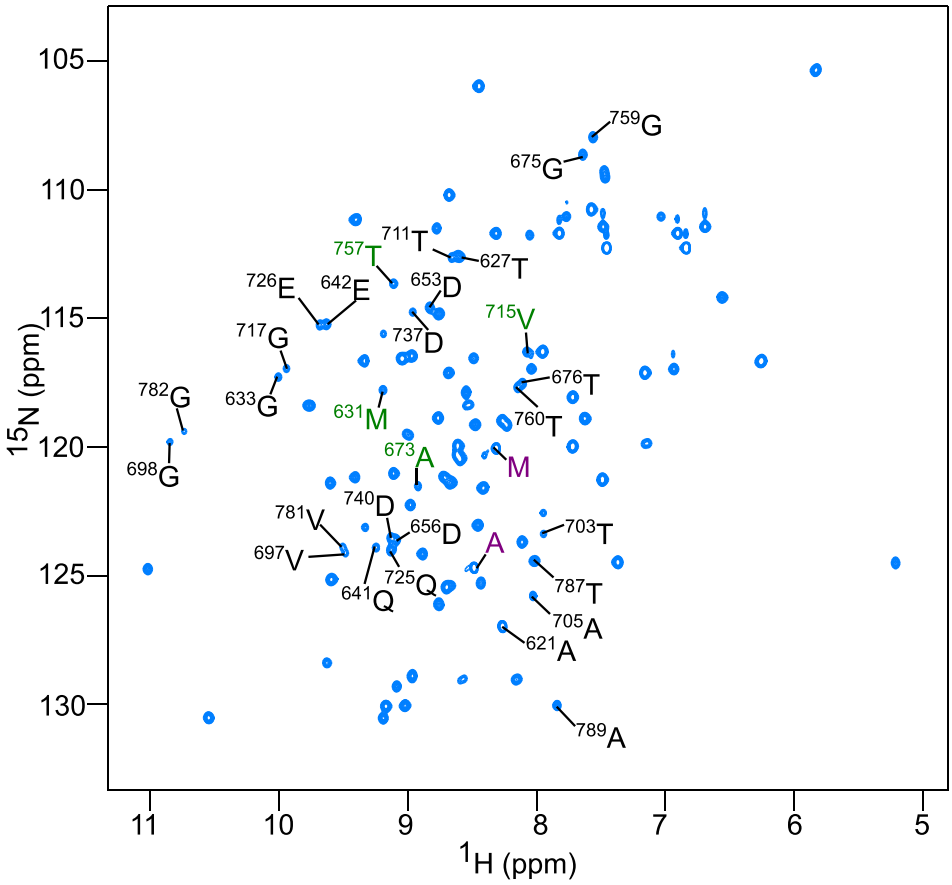


Figure 5.13: Putatively assigned (^{1}H,^{15}N)-HSQC spectrum of S0304. Assigned non-overlapping resonances are shown in bold black. Resonances from equivalent non-identical residues shown in green. Residues remaining from fusion tag are shown in purple. **A.** Sequence alignment of S03 and S04 with residue numbering as in SGO0707. **B.** (^{1}H,^{15}N)-HSQC-spectrum of S0304.

A.

```
        621  625   630   635   640   645   650   655   660
S03 GPAMAPTYKATHEFMSGTPGKELPQEVKDLLPADQTDLKDGSQA
             710   715   720   725   730   735   740
S04       ATYKATHEFVSGTPGKELPQEVKDLLPADQTDLKDGSQA

        665   670   675   680   685   690   695   700   705
S03 TPTQPSKTEVKTAEGTWSFKSYDKTSETINGADAHFVGTWEFTAA
          750   755   760   765   770   775   780   785
S04 TPTQPSKTEVKTTEGTWSFKSYDKTSETINGADAHFVGTWEFTPA
```

B.

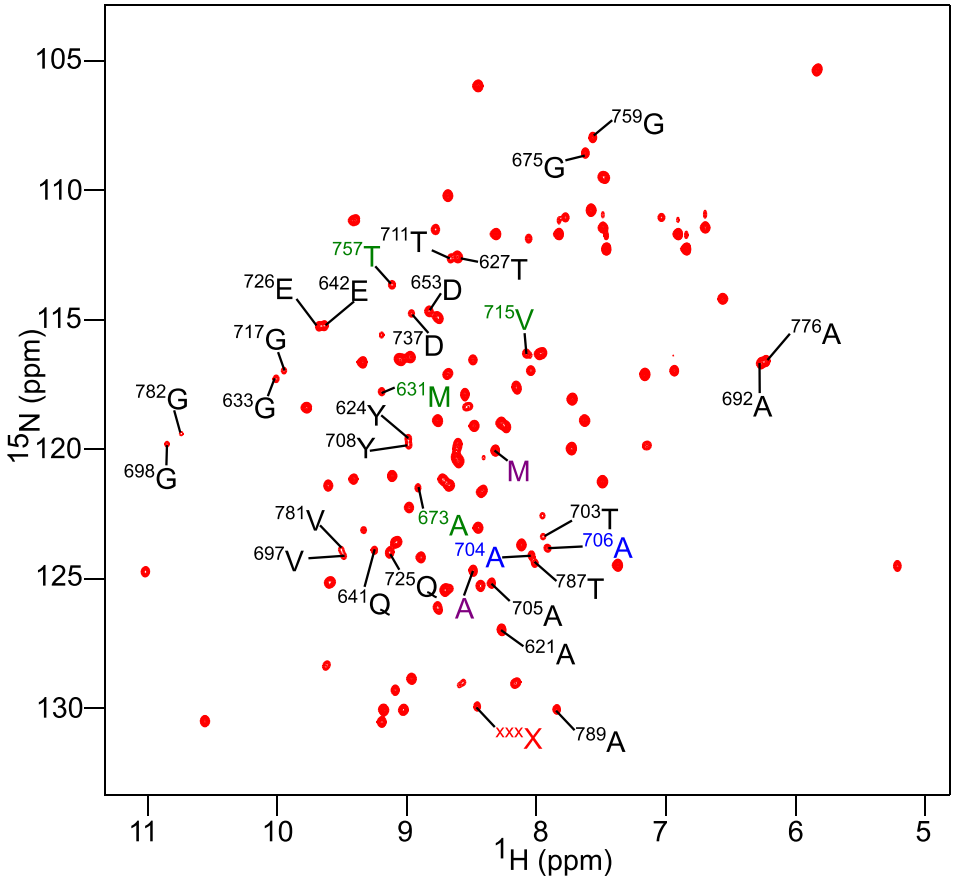


Figure 5.14: Putatively assigned (^{1}H,^{15}N)-HSQC spectrum of S0304_P704A,P705A. Resonances from equivalent non-identical residues are shown in green. Residues in blue cannot be unambiguously assigned. Unassigned resonance ^{xxx}X is shown in red. **A.** Sequence alignment of S03 and S04 with residue numbering as in SGO0707. **B.** (^{1}H,^{15}N)-HSQC-spectrum of S0304_P704A,P706A. Residues remaining from fusion tag are shown in purple. Assigned non-overlapping resonances are shown in bold black.

5.3.7 Backbone dynamics of SHIRT domains

The dynamics of backbone amide residues were studied using different relaxation parameters. The steady-state (^{1}H, ^{15}N)-hnNOE ratio reports on the flexibility of backbone amide resonances on the picosecond to nanosecond timescale[453]. The (^{1}H, ^{15}N)-hnNOE ratio is determined from the ratio of the intensities of resonances in the presence and absence of ^{1}H saturation (Equation 2.14, see sections 2.7.3.3 and 2.7.8.4)[297]. Although there is no hard distinction between rigidity and flexibility based on (^{1}H, ^{15}N)-hnNOE ratio alone; here, it is assumed that a value above 0.5 suggests limited flexibility and a value below 0.5 indicates that the (^{1}H, ^{15}N)-bond is in a more flexible environment.

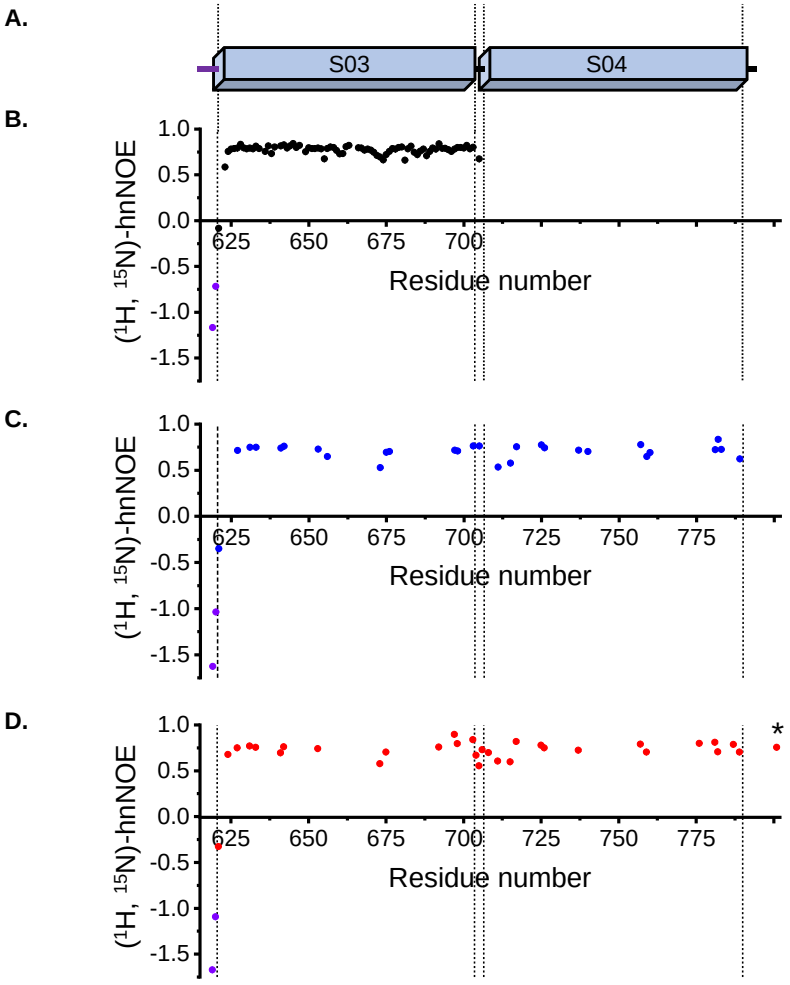


Figure 5.15: (^{1}H, ^{15}N)-hnNOE ratios of SHIRT domains. Residues remaining from fusion tag are shown in purple. **A.** Schematic overview of tandem SHIRT domains with domain boundaries indicated by dashed lines. **B-D.** (^{1}H, ^{15}N)-hnNOE ratio for **B.** S03, **C.** S0304 and **D.** S0304_P704A,P706A with * corresponding to the (^{1}H, ^{15}N)-hnNOE ratio for the remaining unassigned resonance.

(^{1}H, ^{15}N)-hnNOE ratios for backbone amide resonances in single and tandem SHIRT domains are shown in Figure 5.15. The residues remaining from the fusion tag (purple) are

hypothesised to be unstructured and this is confirmed by the observation of (^{1}H, ^{15}N)-hnNOE ratio below 0. The backbone amide bond from Ala621 is also flexible, with a (^{1}H, ^{15}N)-hnNOE ratio below 0. The average (^{1}H, ^{15}N)-hnNOE ratio of backbone amide bonds in S03 and S04 are very similar and generally have (^{1}H, ^{15}N)-hnNOE ratios above 0.70 (Table 5.5). This indicates that all backbone amide residues in SHIRT domains are in an environment with limited flexibility on the ps-ns timescale[453].

The proline residues in the linker region of S0304 are hypothesised to provide rigidity, minimising the flexibility in the linker region. The (^{1}H, ^{15}N)-hnNOE ratio of the putatively assigned backbone amide resonances Ala705 in the linker of S0304 is 0.76 ± 0.01 and the average (^{1}H, ^{15}N)-hnNOE ratio of resonances putatively assigned to the Ala704-Ala705-Ala706 region in the linker of S0304_P704A,P706A is 0.65 ± 0.01. This suggests that mutation of proline residues in the linker increases dynamics on the ps-ns timescale.

Furthermore, the rigidity of proline residues in the linker is hypothesised to link the rotational correlation of tandem SHIRT domains. In absence of proline residues, the increased flexibility in the linker might allow tandem SHIRT domains to rotate with a time constant more appropriate to the MW for their individual domains. This is probed by the calculation of the rotational correlation time, τ_C, which is defined as the time it takes for a molecule to rotate through one radian[294] (Equation 2.13). The relaxation time constants T1 and T2 of backbone amide resonances in single and tandem SHIRT domains are determined (Figure 5.16, Figure 5.17) and the τ_C values are estimated (Figure 5.18).

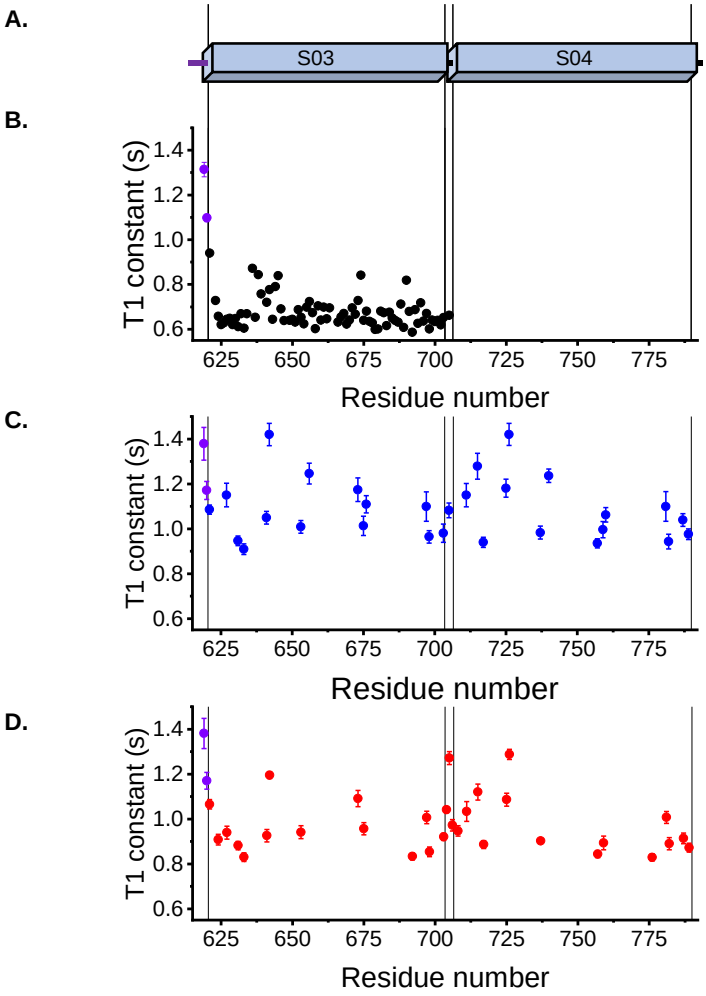


Figure 5.16: T1 constants of backbone amide resonances in SHIRT domains. Residues remaining from fusion tag are shown in purple. **A.** Schematic overview of tandem SHIRT domains with domain boundaries indicated by lines. **B-D.** T1 values for **B.** S03, **C.** S0304 and **D.** S0304_P704A,P706A.

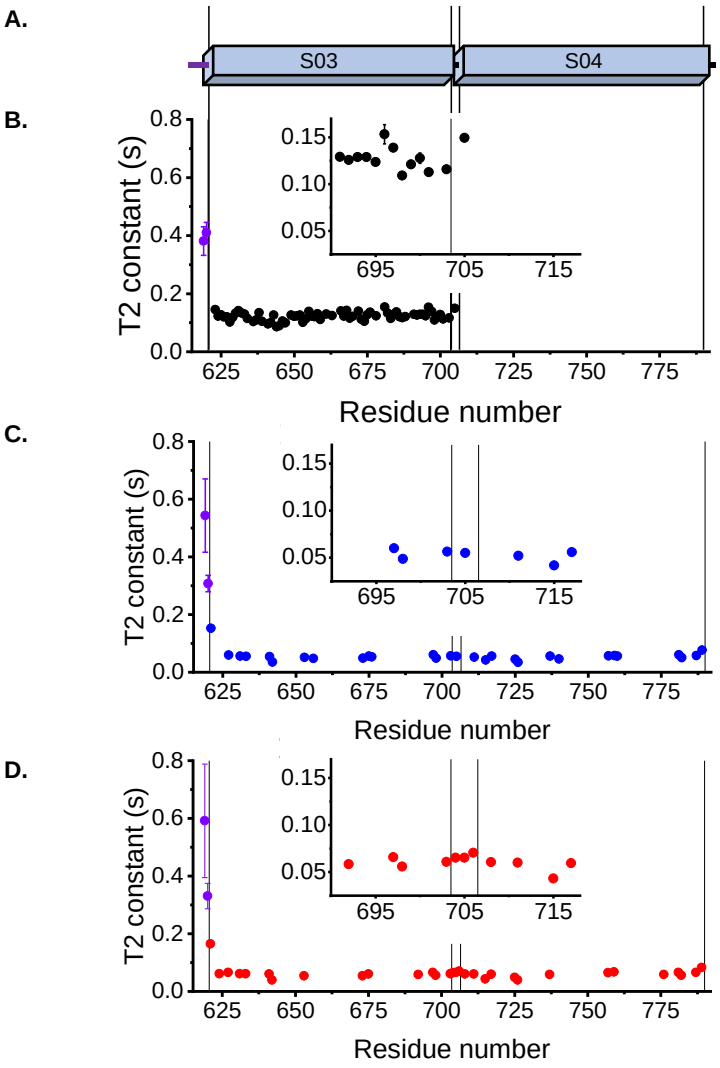


Figure 5.17: T2 values of backbone amide resonances in SHIRT domains. Residues remaining from fusion tag are shown in purple. Insets: close-up of T2 values of the linker region. **A.** Schematic overview of tandem SHIRT domains with domain boundaries indicated by lines. **B-D.** T2 values for **B.** S03, **C.** S0304 and **D.** S0304_P704A,P706A.

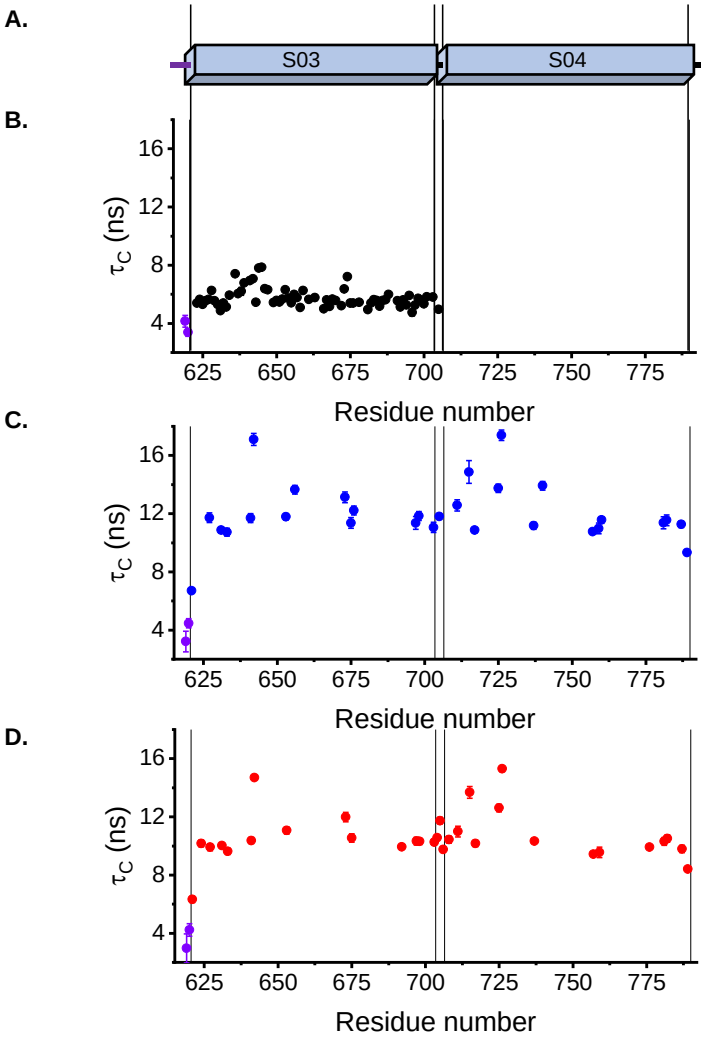


Figure 5.18: τ_C of backbone amide resonances in SHIRT domains. Residues remaining from fusion tag are shown in purple. **A.** Schematic overview of tandem SHIRT domains with domain boundaries indicated by lines. **B-D.** τ_C values for **B.** S03, **C.** S0304 and **D.** S0304_P704A,P706A.

S03 has an average τ_C of 5.76 ± 0.08 ns, while the average τ_C of S03 in the tandem SHIRT domain S0304 is 11.80 ± 0.31 ns (Table 5.5). An increased τ_C for a S0304 compared to S03 is expected, because the higher MW and more anisotropic shape limit the rotation along the long axis of the molecule. The proline-rich linker connects both SHIRT domains rigidly, leading to a rotational correlation time appropriate for a larger molecule. The P704A,P706A mutation between SHIRT domains lowers the average τ_C of S03 to 10.40 ± 0.22 ns, suggesting a small increase in flexibility due to the absence of proline residues in the linker. However, the average τ_C of SHIRT domains in S0304_P704A,P706A does not approach the τ_C of a single SHIRT domain in solution, indicating that the short three-linker does not allow the domains to rotate independently.

This might be expected, as linker flexibility not only depends on the residue types in the linker, but also on the length of the linker. Wriggers *et al*. (2005)[454] reported the distance at which a polypeptide chain comprised of one type of amino acid, changes direction. At three residues, hardly any difference was observed between a tripeptide of proline and alanine.

Table 5.5: Average values of the relaxation analyses on S03, S0304 and S0304_P704A,P706A. Relaxation data was recorded at ^{1}H 700 MHz. τ_C was calculated from Equation 2.13. Residues remaining from fusion tag were excluded from analysis. Ala621 was excluded from analysis.

Protein	Region	(^{1}H, ^{15}N)-hnNOE (-)	^{15}N T1 (s)	^{15}N T2 (ms)	τ_C (ns)
S03	S03	0.77 ± 0.003	0.67 ± 0.004	12.2 ± 0.25	5.76 ± 0.08
	Linker	0.67 ± 0.02	0.66 ± 0.003	14.9 ± 0.18	4.96 ± 0.04
	S04	N/A	N/A	N/A	N/A
S0304	S03	0.71 ± 0.01	1.08 ± 0.04	5.97 ± 0.21	11.80 ± 0.31
	Linker	0.76 ± 0.01	1.08 ± 0.03	5.51 ± 0.06	11.81 ± 0.2
	S04	0.71 ± 0.01	1.09 ± 0.04	5.33 ± 0.17	12.24 ± 0.32
S0304_P704A, P706A	S03	0.75 ± 0.01	0.95 ± 0.10	6.60 ± 0.22	10.40 ± 0.22
	Linker	0.65 ± 0.01	1.09 ± 0.16	6.71 ± 0.19	10.69 ± 0.21
	S04	0.73 ± 0.01	0.97 ± 0.13	5.95 ± 0.19	10.82 ± 0.24

5.3.8 Qualitative comparison of relaxation parameters

As SHIRT domains 03 and 04 have 98% sequence conservation and a backbone C_α RMSD of 0.14 Å, it was of interest to determine if their relaxation properties were also similar. To this end, estimated τ_C values and (^{1}H, ^{15}N)-hnNOE ratios were plotted such, that values for equivalent residues were superposed along the *x*-axis for S0304 (Figure 5.19) and S0304_P704A,P706A (Figure 5.20).

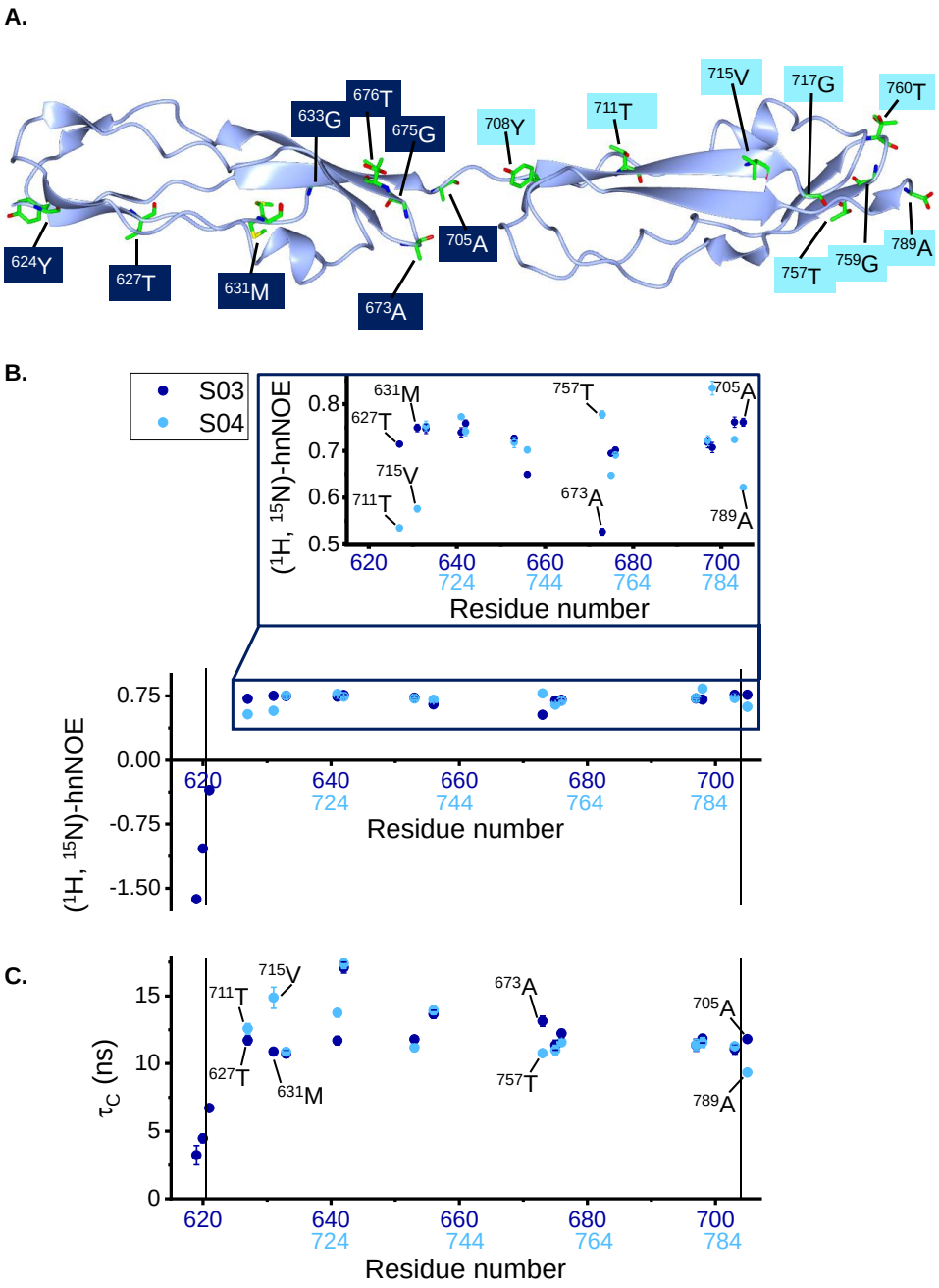


Figure 5.19: Comparison of relaxation parameters for equivalent residues in S0304. A. Crystal structure of S0304 (courtesy of Dr F. Whelan) with residues in equivalent positions indicated. Image was created using CCP4mg. **B.** (^{1}H, ^{15}N)-hnNOE values for equivalent residues in S0304. Box: zoom of (^{1}H, ^{15}N)-hnNOE values from 0.5 to 0.86. **C.** τ_C values for equivalent residues in S0304.

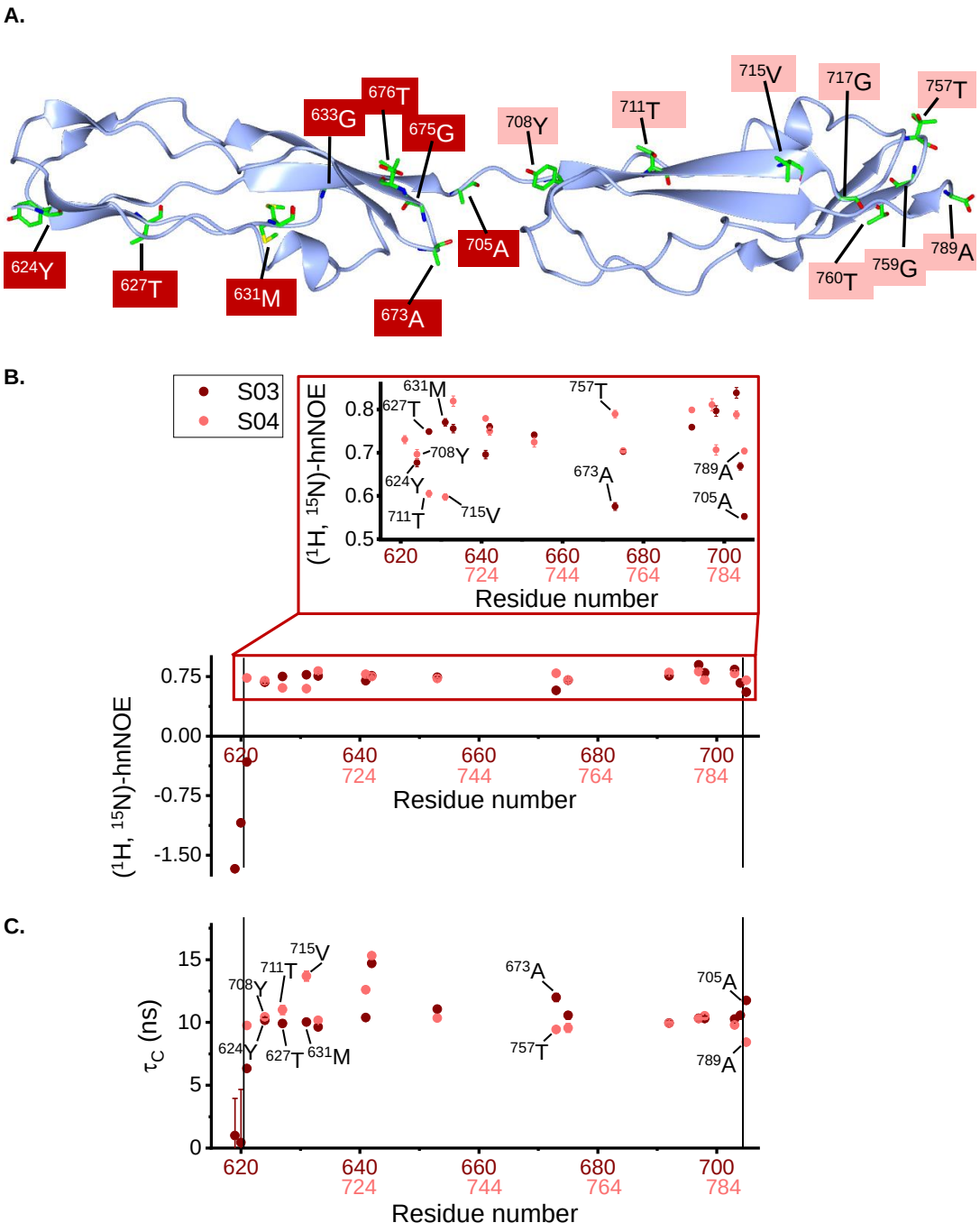


Figure 5.20: Comparison of relaxation parameters for equivalent residues in S0304_P704A,P706A. A. Crystal structure of S0304 (courtesy of Dr F. Whelan) with residues in equivalent positions indicated. Image was created using CCP4mg. **B.** (^{1}H, ^{15}N)-hnNOE ratios for equivalent residues in S0304_P704A,P706A. Box: zoom of (^{1}H, ^{15}N)-hnNOE ratios from 0.5 to 0.86. **C.** τ_C values for equivalent residues in S0304_P704A,P706A.

Generally, a good agreement was obtained for the relaxation properties of equivalent residues (Figure 5.19, Figure 5.20). In some cases, differences were observed for (^{1}H, ^{15}N)-hnNOE ratios and τ_C values between equivalent residues. For example, it was expected that the (minimal) SHIRT interface might stabilise the adjacent SHIRT domains. Rather, the first β-sheet of the SHIRT fold near the SHIRT interface (containing Thr627/Thr711 and

Met731/Val715) and not the adjacent loops in domain S04 of both S0304 and S0304_P704A,P706A (Tyr624/Tyr708, Gly633/Gly717) had significantly different relaxation properties (see Table 5.6). Perhaps, the first strand in S04 experiences motions on the ps-ns timescale, detected by the (^{1}H, ^{15}N)-hnNOE ratio, which are not accompanied by increased motions detected by T1 and T2 relaxation experiments; or the effect of the interface is different than what was expected.

In S03, the loop containing Ala673 was observed to bear some flexibility with a (^{1}H, ^{15}N)-hnNOE ratio of <0.8, with a minimum for Glu674 of 0.66 (Figure 5.15B). Generally, loop regions might be expected to be more flexible than other elements of secondary structure and inter-domain interfaces likely stabilise adjacent domains. In tandem SHIRT domains, the opposite was observed, where the loop residue facing the SHIRT inter-domain interface (Ala673) has a significantly lower (^{1}H, ^{15}N)-hnNOE ratio and higher τ_C value than Thr757 (Table 5.6).

Finally, it was expected that the flexibility of Ala705 was reduced upon the introduction of a rigid linker in S0304 and reduced to a lesser extent in S0304_P704A,P706A. This was in agreement with the observed relaxation properties for Ala705 and Ala789 (Table 5.6).

Table 5.6: Relaxation values of residues in equivalent positions in S0304 and S0304_P704A,P706A.

Protein	S03			S04		
	Res.	**(^{1}H, ^{15}N)-hnNOE ratio**	**τ_C**	**Res.**	**(^{1}H, ^{15}N)-hnNOE ratio**	**τ_C**
S0304_PA	624Y	0.75 ± 0.01	9.9 ± 0.2	708Y	0.61 ± 0.01	10.4 ± 0.2
S0304	627T	0.714 ± 0.004	11.72 ± 0.3	711T	0.54 ± 0.01	12.6 ± 0.4
S0304_PA	627T	0.73 ± 0.01	9.9 ± 0.2	711T	0.61 ± 0.01	11.0 ± 0.4
S0304	631M	0.73 ± 0.01	10.9 ± 0.2	715V	0.58 ± 0.01	14.9 ± 0.8
S0304_PA	631M	0.77 ± 0.01	10.0 ± 0.2	715V	0.60 ± 0.01	13.7 ± 0.4
S0304	633G	0.75 ± 0.01	10.7 ± 0.3	717G	0.75 ± 0.01	10.9 ± 0.2
S0304_PA	633G	0.76 ± 0.01	9.6 ± 0.2	717G	0.81 ± 0.01	10.2 ± 0.2
S0304	673A	0.53 ± 0.01	13.1 ± 0.4	757T	0.78 ± 0.01	10.7 ± 0.2
S0304_PA	673A	0.60 ± 0.01	12.0 ± 0.4	757T	0.79 ± 0.01	9.4 ± 0.1
S0304	675G	0.70 ± 0.01	11.4 ± 0.4	759G	0.647 ± 0.004	12.0 ± 0.4
S0304_PA	675G	0.70 ± 0.01	10.6 ± 0.3	759G	0.71 ± 0.01	9.6 ± 0.4
S0304	676T	0.70 ± 0.01	12.2 ± 0.3	760T	0.69 ± 0.01	11.6 ± 0.2
S03	705A	0.67 ± 0.02	4.9 ± 0.04	N/A		
S0304	705A	0.76 ± 0.01	11.8 ± 0.2	789A	0.62 ± 0.01	9.3 ± 0.2
S0304_PA	705A	0.55 ± 0.01	11.7 ± 0.2	789A	0.70 ± 0.01	8.4 ± 0.2

5.4 Conclusions of this chapter

This chapter reports on the dynamic studies performed on SHIRT domains 03 and 04 and on the dynamics in the linker region between these domains. A backbone assignment was performed on S03; this data is ready for publication. Putative backbone assignments of S0304 and S0304_P704A,P706A were based on the backbone assignment of S03 and the crystal structure of S0304. An analogous approach is routinely used in the assignment of very similar domains, for example for the backbone assignment of point mutants of wild-type proteins[449–451]. Here, this approach is considered reliable due to the high overlap of resonances between single and tandem SHIRT domains and thus may be used for publication.

Full backbone dynamics datasets were recorded at 700 and 800 MHz; here, the analysis is shown for the data recorded at 700 MHz. This data conclusively shows that SHIRT domains lack flexible loops. The backbone dynamics of residues putatively assigned to the linker region reveal that the proline residues in the linker limit the backbone flexibility of Ala705.

However, there is scope for both sets of relaxation data (recorded at 700 and 800 MHz) to be analysed using model-free analysis to separate the (likely anisotropic) global correlation time of the molecules from the faster internal motions. This approach allows a more detailed look into the anisotropic dynamics of SHIRT domains and this level of detail would be expected for publication purposes. Potentially, a third dataset at e.g. 500 or 900 MHz might be required as the frequencies of the current datasets are quite close together.

Chapter 6. Discussion, conclusions and future directions

6.1 Discussion for DRESS domains

6.1.1 DRESS domains; DUF1542 domain boundaries redefined

Secondary structure predictions and MSAs of DUF1542 domains suggested that the current domain boundaries reported in the literature[111] are incorrect and should be shifted by ~20 residues (Figure 6.1). Using the resulting domain boundaries, the crystal structure of tandemly arrayed DRESS domains was solved. The structure comprised two complete domains and therefore verified the new domain boundaries.

The redefinition of the domain boundaries of DUF1542 domains and the biophysical characterisation of single and tandem DUF1542 domains enable us to rename DUF1542 as DRESS domains.

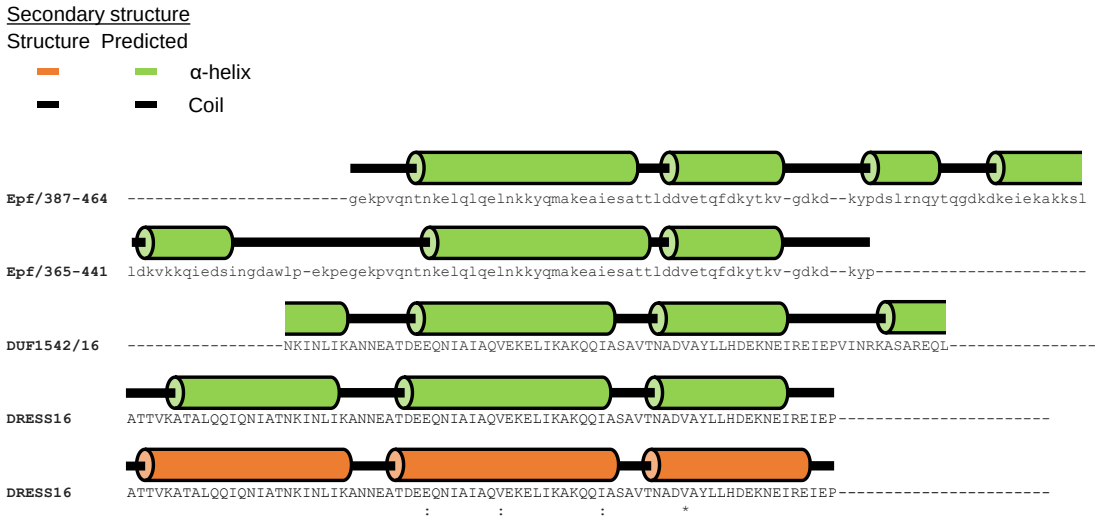


Figure 6.1: Domain boundaries and secondary structure prediction of DUF1542 domain 1 in Epf and DUF1542 and DRESS domain 16 in SasC. Epf/387-464: domain boundaries of DUF1542 1[227]. Epf/365-441: domain boundaries shifted by 24 residues. SasC/DUF1542/16: domain boundaries from Schroeder *et al*. (2009)[111]. MSA was performed with 18 DRESS domains in Clustal Omega[259], here DRESS domain 16 is shown. The secondary structure prediction was performed by PSIPRED[375] with α-helices shown in green. The α-helical residues in the crystal structure of D1617 are shown in orange. Black lines represent coil regions.

6.1.1.1 Recombinant production of DRESS domains

Previously, multiple DUF1542 domains were produced from the CWA protein Epf from *S. pyogenes*[227] or from SasC[111]. Attempts to produce one DUF1542 domain from Epf with domain boundaries as in Figure 6.1 failed[227]. This is consistent with preliminary

experiments performed for this thesis, where a single DUF1542 domain with literature domain boundaries was degraded during over-production of the recombinant protein (data not shown). Genes encoding four or sixteen DUF1542 domains of Epf, putatively containing five and seventeen complete, folded DRESS domains with unstructured termini of maximally 23 residues, could be recombinantly expressed and produced[228,227]; while genes encoding eight DUF1542 domains of Epf, putatively comprising nine folded DRESS domains and ~35 putatively unstructured residues at either end, could not be expressed and produced[227]. Schroeder and co-workers[111] could successfully produce eight DUF1542 domains, putatively comprising seven folded DRESS domains with predicted unstructured termini of ~18 residues, from SasC. In this work, redefinition of the domain boundaries enabled the successful over-production and purification of single, tandem and multiple DRESS domain constructs.

6.1.2 Evidence of adducts in some SasC constructs

The MW of proteins produced in this thesis was assessed by MS. In most cases, a MW was found that was within 1 Da of the expected MW. For D1617, D1617_T1838D and D1617_T1838D,N1843D; a consequent mass difference of 1 Da was observed that remained unaccounted for. For D0310_scc, three species were observed, of which 24% contained the correct MW species, in addition to two species that were +162.13 Da (65%) and 2x +162.13 Da (11%) heavier than the target protein. The size of the mass adducts is consistent with glycosylation to incorporate a monosaccharide. The location of the modification seems consistent with a modification of the cysteine residues via a bond that is not affected by reduction.

S-linked glycosylation is a rare post-translational modification of free cysteine residues and is reported *in vivo* in the antimicrobial peptides sublancin[455] and glycocin F[456], which contain five cysteine residues of which four are involved in the formation of disulfide bonds and the remaining in S-glycosylation. In glycocin F, secreted by the Gram-positive bacterium *Lactobacillus plantarum* strain KW30, the S-glycosylated cysteine residue is the C-terminal residue of the peptide comprising 43 residues, which is separated from the core of the disulfide-bonded peptide via a predicted flexible, unstructured linker with the sequence HHSSGSSSYHC[456]. So far, the mechanism behind the secretion of glycocin F remains elusive[457].

Here, the nature and location of the modification on D0310_scc remains speculative. A doubly S-glycosylated protein would not bind to the gold surface. However, at least 65% of D0310_scc still featured a free cysteine residue, which is functional for immobilisation on a gold-coated glass square; AFM immobilisation experiments confirmed this. Therefore, it is assumed that AFM experiments were not affected by the presence of adducts on D0310_scc.

6.1.3 The oligomeric state of DRESS domain-containing proteins

Different oligomeric states were observed for different parts of the repetitive region of SasC. Four domains from the C-terminal end of the DRESS region, D1417, showed a dimeric state as determined from the MW calculated by SLS. On the contrary, four domains from the middle of the DRESS region, D0710, were monomeric. The entire repetitive region of SasC, D0118, that contains both D0710 and D1417; was 94% monomeric as determined from the MW calculated by SLS (see sections 4.3.3 and 4.3.4). Finally, four DRESS domains from the N-terminal end of the repetitive region in Epf were monomeric in solution[228]. As the intact repetitive region of SasC is a monomer in solution, the observed dimerisation of D1417 might be non-physiological.

The presence of fusion tags has been reported in the literature to be able to cause non-physiological dimerisation[458,459]. However, the tag has been efficiently removed from D1417 as observed by SDS PAGE analysis (data not shown) and a correct MW as determined by MS (see Table 4.2).

6.1.4 Crystal structure of D1617

The crystal structure of tandem DRESS domains (Figure 3.18A) verified the proposed domain boundaries for DRESS domains. D1617 consists of two triple-helical bundles. The domains are arranged in a head-to-tail conformation with N- and C-termini at opposite ends, forming an elongated rod-like structure (Figure 3.18B). Other tandem domains in repetitive regions also adopt a head-to-tail conformation, such as repeats in the repetitive region of SasG[57] or alternate GA- and FIVAR modules in Ebh[58]. Many CWA proteins contain a repetitive region between the (putative) N-terminal adhesin and the C-terminal covalent linkage to the cell wall[29,30], and for SasG, extension in the repetitive region effectively

projects the N-terminal adhesin away from the cell wall[59]. Here, the structure of tandem DRESS domains is discussed.

6.1.4.1 Sequence conservation mapped on DRESS domains

The average pairwise sequence identity between DRESS domains in SasC is 28.7% (calculated by Clustal Omega[259]). However, the sequence conservation of residues in the interface of DRESS domains was significantly higher (see Figure 3.25). For example, Thr1838 is conserved in DRESS domains 1 to 17 and is located at the C-terminus of a DRESS repeat. This conservation is lost in DRESS domain 18, which does not C-terminally interact with another DRESS domain. On the contrary, no such conservation is observed for polar residues within the α-helical secondary structure. All conserved polar residues were located in domain interface regions. These interfaces are essential for tandem DRESS domain stability, as was shown by the disruptive mutation of the interface, T1838D, which changed the tandem DRESS response to thermal denaturation or ionic strength to that of a single domain.

Generally, individual domains in multi-domain proteins with sequence identities over 30-40% maintain their interface geometry[200,460]. A pairwise sequence identity below 40% is considered advantageous for adjacent domains to avoid misfolding and aggregation[196]. Here, the residues not involved in the formation of inter-domain interfaces lie below this cut-off and might help to avoid misfolding. The residues that are essential for the formation of interfaces are much more conserved and thereby are likely to mediate very similar inter-domain geometries.

6.1.4.2 Tilt and twist angles between tandem domains

Two DRESS domains form a tandem, elongated structure containing head-to-tail organised tandem domains. The tilt angle between D16 and D17 was 154° and the twist angle was 139.2° (see Figure 3.20). The tilt and twist angles of selected tandemly arrayed domains were estimated as previously described (see section 3.3.9.4) and compared (Table 6.1). With the exception of spectrin domains, other tandemly arrayed domains had a high twist angle between adjacent tandem repeats. In B regions containing repeating units of multiple domains (alternating tandemly arrayed domains; Table 6.1), a high twist angle was observed for every other domain. The estimated tilt angles range from 135° (talin) to 174° (E-G5[2]).

A high twist/high tilt angle between multiple repeats would result in a highly extended and twisted repetitive region that bears a resemblance to the topology of a twisted rope. This might increase the overall strength of the twisted region[387,397,398]. A high twist/intermediate tilt angle between multiple repeats might still be capable of forming the twisted rope-like structure; in addition, the lower tilt angle might introduce a superhelical-like shape of the B region that might have a spring-like function.

Table 6.1: Twist and tilt angles of head-to-tail, tandemly arrayed repetitive domains. The twist angle is obtained from the average of the κ angle from superposition in Coot[332] and from the superposition angle in CCP4mg. The tilt angle is obtained from the manually calculated average between the long axis of domain 1 with the long axis of domain 2. [a]WinCoot superposition failed. [b]Rib: Resistance to proteases, immunity, group B[441].

Protein	PDB	Domains	Number of domains in B region	Twist (°)	Tilt (°)	Reference
Tandemly arrayed						
SasC	N/A	D1617	18	139.2	154	**This work**
α-Actinin	1quu	Spectrin repeats 2-3	4	[a]57.0	164	[379]
Plakin	5j1g	Spectrin domains 7-8	9	58.6	161	[398]
Talin	3dyj	IBS2-A,B	12-13	177.4	135	[461]
Titin	3b43	I67-I68	31	176.1	Variable	[462]
[b]Rib	N/A	2 Rib domains	12	170.6	164	Crystal structure by Dr F. Whelan (F. Whelan *et al.*, manuscript in preparation)
Alternating tandemly arrayed domains						
Ebh	2dgj	FIVAR-GA	52	141.7	169	[58]
	2dgj	GA-FIVAR		24.0	137	[58]
SasG	3tiq	$G5^1$-E	9	[a]12.9	160	[59]
	3tiq	E-$G5^2$		111.4	174	[59]

The observation that many different consecutive domains in bacterial extracellular repetitive regions display large twist/tilt angles, might suggest that this is a successful approach to create a higher tensile strength. This might replace disulfide bonds, that are not usually present in the extracellular environment of Gram-positive bacteria[249] or isopeptide bonds (covalent bonds between side chains; not detected in the crystal

structure of D1617)[463]. This is supported by the fact that for all cases described in Table 6.1, an extended rod-like repetitive region is suggested (Rib (F. Whelan *et al*., manuscript in preparation), plakin[398], α-actinin[379], talin[461], titin[203], SasG[59], Ebh[58]).

6.1.4.3 Comparison of tandem DRESS domains to other known structures

Staphylococcal proteins feature several known triple-helical domains, such as the triple-helical domains in SpA[44], the GA-module[464] and the FIVAR motif[58]. Three-helix bundles are a very common structural motif with a plethora of functions[465], ranging from forming a part of a human heat shock protein (PDB 3lof) to the nucleocapsid-binding domain from mumps virus[466] (PDB 3bbz). Here, the structural similarity between DRESS domains and other triple-helical protein domains is assessed.

Structures homologous to DRESS domains in the PDB[467] were identified using the DALI server[468,469] and aligned using the all-against-all SSM functionality of DALI[468,469] (Figure 6.2A-E). Over 100 hits were found for D16 with significant structural similarity (Z score >2), consistent with helical bundles being common protein structures[465].

The non-tandem, triple-helical domain Rich in Charged Residues (RICH) from the N-terminal region of Choline binding protein A (CbpA) most closely resembled the structure of a DRESS domain (Table 6.2, Figure 6.2). RICH interacts with its binding partner complement factor 9[470]. No conservation is observed between the binding residues of RICH and equivalent residues in DRESS domains, which is consistent with different functions for RICH and DRESS.

Table 6.2: SSM parameters of selected three-helix bundles, aligned to D16.

Domain	PDB	RMSD (Å)	Pairwise sequence identity	Number of aligned residues
RICH	4k12	1.6	7%	82
D17	N/A	1.1	29%	73

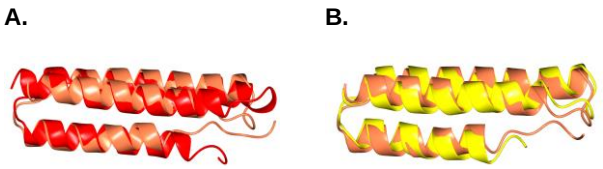


Figure 6.2: Structural homologues of DRESS domain D16 (orange). All domains have the N-terminus on the left. Superposition performed by DALI[471] with reference to D16. Images were created using CCP4mg. **A.** RICH domain[470] (red; PDB 4k12), **B.** DRESS domain 17 (yellow).

6.1.4.4 Linkers in multi-domain proteins

Linkers between protein domains are essential for inter-domain communication, inter-domain flexibility and/or maintaining end-to-end distances between domains[454]. The rigidity in linkers is generally achieved by contiguous helical secondary structure, such as in spectrin[472], or by the incorporation of proline residues into non-helical linkers. For example, triple-helical domains linked by six residues in SpA require flexibility to mediate binding to different binding partners[56]. On the contrary, repetitive regions forming an elongated rod such as FIVAR-GA domains in Ebh[58] or E-G5 domains in SasG[59] are consecutive without residues assigned to a linker function, leading to restricted inter-domain flexibility and rigid repetitive regions. Typically, linkers are 5-10 residues in length[399,454], however head-to-tail oriented domains tend to have shorter linkers[473]. Although the classical description of a linker involves more hydrophilic, flexible, non-conserved regions[474], linkers are, on average, 42% buried upon interface formation[439].

DRESS domains are connected by a Pro1885-Ile1887 linker, which, on average, is 37% buried upon interface formation as determined by PISA[338]. Proline residues are common in linkers[439] and small hydrophobic residues such as Val and Ile have increased propensities to occur in short linkers[439]. Here, this short linker is likely to minimise inter-domain flexibility with the proline residue providing structural rigidity[399]. Hence, based on the linker length and amino acid content, the DRESS region might form a rigid structure.

6.1.5 The DRESS region forms a rod-like region

6.1.5.1 Solution techniques

SAXS

Solution-based structural information from SAXS has been used in studying the overall size and shape of repetitive regions from other multi-domain proteins, such as DUF1542 domains in Epf[228], E-G5 repeats in SasG[57], three-helix bundles in SpA[45] and spectrin repeats in plectin and desmoplakin[398]. Here, the size of DRESS domains in multi-domain proteins was estimated by SAXS.

The Epf SAXS construct, EpfN_DUF1-3[228], contains an N-terminal region comprising two sub-domains of known structure and three putative DRESS domains. A MSA to DRESS domains in SasC reveals the presence of three complete DRESS domains, plus sixteen residues of DRESS domain 4 in Epf (Figure 6.3C) that likely remain disordered. *Ab initio* models of flexible regions require multiple conformations[343], therefore it is likely that the sixteen putatively disordered residues in EpfN_DUF1-3 are not visible in an *ab initio* model with a single conformation in Figure 6.3B. The D_{max} of EpfN_DUF1-3 is 20 nm with 8 attributed to the N-terminal region, leaving 12 nm for three DRESS domains, estimated to be 4.0 nm per DRESS domain.

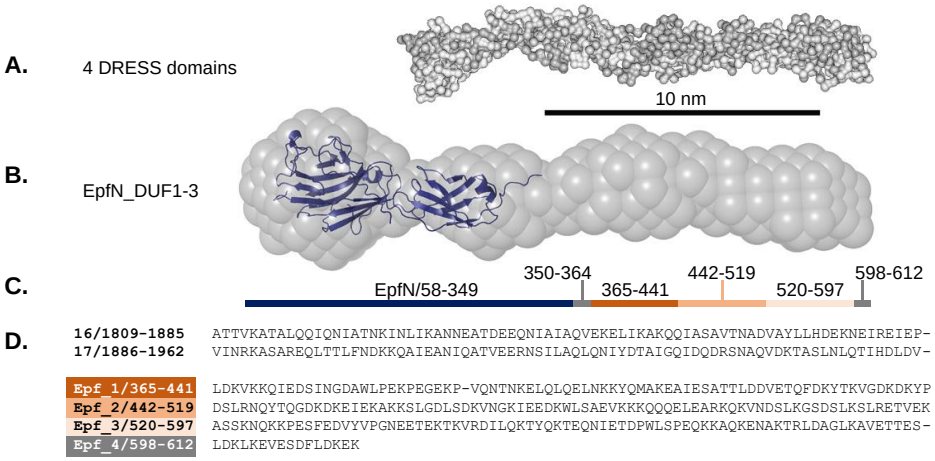


Figure 6.3: Analysis of the SAXS model of EpfN_DUF1-3 (residues 58-612)[228]. **A.** Three *ab initio* modesl of 4 DRESS domains from SasC (D0710) generated in Gasbor[419]. **B.** *Ab initio* model of EpfN_DUF1-3 (grey)[228]. Scale bar represents 10 nm in both models. **C.** Residue numbers in the EpfN_DUF1-3 SAXS model, colours correspond to MSA of DRESS domains in D. **D.** MSA of DRESS domains from SasC (D16 and D17 shown; top) and Epf with proposed domain boundaries for DRESS domains within the EpfN_DUF1-3 construct.

Here, the D_{max} of four DRESS domains from SasC was 17.5 nm as determined by SEC-SAXS (Figure 6.3A), implying one DRESS domain has a length of 4.4 nm. This is close to the length of one DRESS domain as obtained from the X-ray crystal structure of 4.2 nm. Eighteen DRESS domains in D0118 had a D_{max} of 66.3 nm, implying one DRESS domain has a length of 3.7 nm. However, the diameter of the D0118 rod was larger than the D0710 rod, indicating that D0118 might bend more than D0170.

Porod exponent

The "dimensionality" of particles in solution was analysed using the Porod exponent, calculated from the negative slope of the mid-q region of a logarithmic plot for q versus I(q)[348]; see sections 4.3.6.2 and 4.3.6.4. Globular proteins have a Porod exponent of 3-4; particle shapes with two dimensions, such as lamellae, have a Porod exponent of 2; and stiff, rigid rods or thin cylinders have a Porod exponent of 1[348,349].

The Porod exponents of different lengths of the repetitive region of SasG were studied by SAXS[57]. Values were found to range from 1.04-1.10, indicating rod-like behaviour in solution. For SasC, the Porod exponents of D0710 and the major species of purified D0118 were 1.07 and 1.05, respectively, confirming elongation and rigidity in solution. The Porod exponent of the minor species in purified D0118 was 3.3, suggesting a 'collapsed' polymer chain. Together with the observation of disorder from the Kratky plot (see Figure 4.14C), the very large $R_{g,c}$, the higher oligomeric state from MW analysis (see Table 4.6) and its elution from a SEC column close to the void volume (see Figure 4.13A), this species might represent unfolded, aggregated D0118.

The existence of flexibility within the rods formed by the D0710 and D0118 protein constructs was investigated by *ab initio* modelling (see sections 4.3.6.3 and 4.3.6.5). Both protein constructs were best represented by only 1-2 models, whereas flexible proteins usually require 10-20 models[356]. In summary, SEC-SAXS analysis and *ab initio* modelling of D0710 and D0118 show that DRESS regions form highly elongated, rigid rods in solution.

6.1.5.2 Elongation of the entire, intact DRESS region from SasC measured on a surface

SHRImP-TIRF microscopy was employed to determine the end-to-end distance between fluorophores in the D0118_2A488 construct (see section 4.3.7). A distance of 65 nm was

expected, based on the end-to-end distance of single and tandem DRESS domains in the X-ray crystal structure (see Figure 3.18). Here, an end-to-end distance of 60 ± 2.4 nm (mean ± s.e.) was determined in HEPES imaging buffer, pH 7.0 (see Table 2.4) from 244 measurements on 5 independently prepared imaging samples. An end-to-end distance of 44 ± 1.9 nm was determined in MES imaging buffer, pH 6.5 (see Table 2.4) from 223 measurements on one imaging sample. By reducing the pH of the imaging buffer to near the average pI of a DRESS domain (= 5.5), the overall charge on the D0118_2A488 construct should be reduced (predicted charge[475] at pH 7.0: -47.5, at pH 6.5: -39.4). It was hypothesised that decreasing the surface charge would reduce the local electrostatic repulsion between adjacent DRESS domains. In turn, this would lead to an increase in the flexibility of the elongated rod and a decrease in the observed end-to-end distance. The lower surface charge repulsion at pH 6.5 may also lead to the observed higher thermal stability of D0118_2Cys (used to prepare D0118_2A488) in the MES imaging buffer compared to the HEPES imaging buffer (+0.6 °C, determined by nano-DSF).

The large reduction in the end-to-end distance (16 nm, 25% of the expected distance) between the inter-fluorophore distances recorded at pH 7.0 and 6.5 is consistent with this prediction. The end-to-end distances recorded at pH 6.5 were obtained from just one imaging sample. This observation should be confirmed by repeating this experiment to obtain inter-fluorophore distances from independent samples prepared and imaged on different days. It should be noted that electrostatic immobilisation of the D0118_2A488 construct allows for conformational equilibration on the imaging surface rather than kinetic trapping of the solution conformations[476]. A further reduction in the surface charge of the protein construct will facilitate this equilibration process, thus the conformation of the surface immobilised protein may be more dynamic at pH 6.5 compared to pH 7.0.

6.1.5.3 Comparison of the end-to-end distances

The end-to-end distances of recombinant protein constructs from the DRESS region from SasC were estimated using complementary biophysical techniques (Figure 6.4A) and compared with predictions of the end-to-end distance of the repetitive region of SasG (Figure 6.4B). SAXS and SHRImP-TIRF measurements on the entire repetitive region of SasC are both expected to represent the solution state well; the predicted end-to-end distances are in close agreement (Figure 6.4A).

The end-to-end distances predicted by X-ray crystallography were the longest. The number of repetitive domains from which the predicted end-to-end distance was determined, was the lowest for this technique, making it prone to a larger error. Furthermore, DRESS SAXS data was recorded at pH 7.5, while the crystal structure was determined at pH 4.9. This may affect the inter-domain distances and the overall length of the repetitive region. A significant pH-dependent decrease in the end-to-end distance as measured by SHRImP-TIRF has already been observed. Finally, inter-domain distances in a crystal lattice are not necessarily identical in solution conditions[477,478]. Therefore, X-ray and SAXS data complement each other and provide structural information on different resolutions[479].

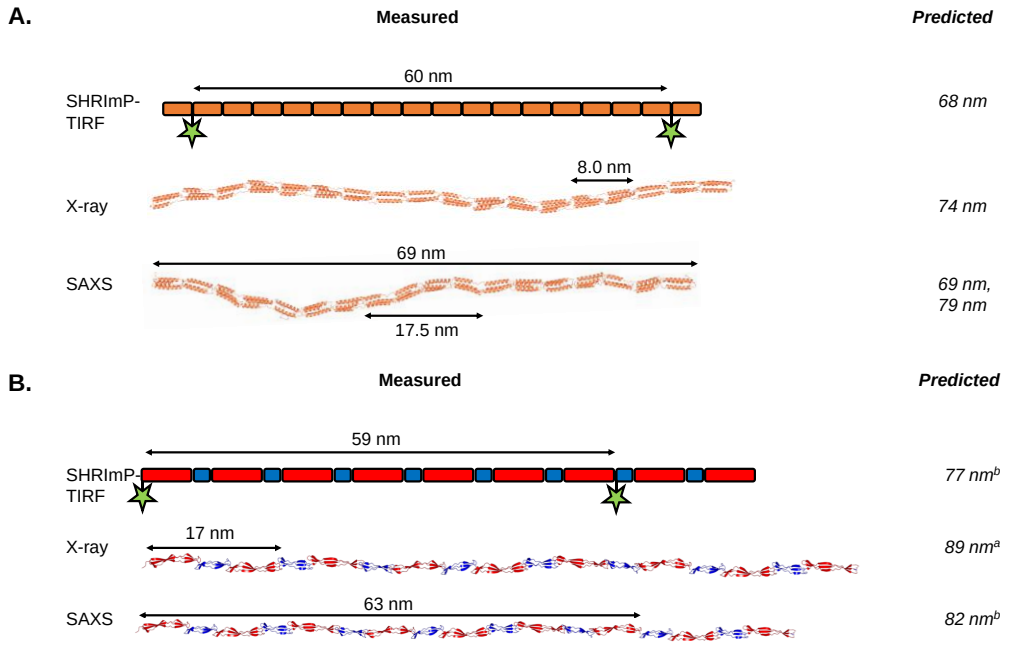


Figure 6.4: Observed and predicted end-to-end distances of repetitive regions from **A.** SasC and **B.** SasG. Structures of repetitive regions are drawn to relative size. Arrows represent measured part of a repetitive region. Size in italics represents predicted, extended, entire end-to-end distance. Ribbon model images were created with CCP4mg. **A.** Models of the DRESS region from SasC. The SHRImP-TIRF model is shown at pH 7.0. The full SAXS structure was modelled using the AllosMod-FOxS server[359,422] and the distance of the four-domain model originates from the *ab initio* modelling of D16,D1617,D17. **B.** Schematic and X-ray model of the repetitive E-G5 region from SasG based on data in Gruzska *et al*. (2012)[a,59] and Gruzska *et al*. (2015)[b,57].

6.1.5.4 SasC, an elongated stalk on the surface of *S. aureus*

The repetitive regions of SasC and SasG are both highly extended and display limited conformational flexibility. The end-to-end distance of the B region in SasG with nine B repeats is estimated to be ~10 nm longer than the DRESS region. The diameter of an *S. aureus* cell measures approximately 1.2 µm[480,481] and its cell wall is approximately 20-40 nm thick[90,92]. The PG layer is constantly remodelled[94,95], where the most outer layers are

degraded and new layers are thought to be formed closest to the cell membrane[96,97], although this is subject to debate[482].

The expression of CWA proteins is different in different environmental conditions[99,100] and strain types[101]; and might occur at any time during bacterial growth. Their extended nature (SasG[57], Ebh[58], SdrD[66], ClfB[102], Cna[103], Bap[104]) might ensure that they can protrude through the pores in the PG layer; even if they are attached closest to the cell membrane. Other components on the cell wall are WTAs and LTA, which are embedded in the PG layer or the phospholipid cell membrane, respectively. They are negatively charged and have a flexible nature[483]. The length of LTA is estimated to be <20 nm, suggesting that they might remain buried in the PG layer[484]. The length of WTA is estimated to be 40[485,486]-60[487] repeating units of ribitol 5-phosphate and is expected to reach well beyond the PG layer[487]. The coverage of WTA molecules on the surface of *S. aureus* is estimated to be one WTA unit per nine molecules of the PG disaccharide Gnc*N*Ac-Mur*N*Ac[487], however this is subject to strain variation[485].

Previously, experimental evidence was provided for SasG that showed that five B repeats[158], estimated to be 51 nm using the reported crystal structure for B repeats from the repetitive region of SasG[59,57], were required to mediate biofilm accumulation[158]. Corrigan *et al*. (2007)[158] showed by EM that SasG fibres with an estimated repetitive region length of 88 nm[57] were highly extended on the surface of a *S. aureus* cell (Figure 6.5C). Comparing the predicted length of WTA, the predicted length of SasC and SasG, and considering the potential attachment sites in the cell wall structure (Figure 6.5A, B), it is likely that the A regions of SasC and SasG will easily be projected out of the PG layer and might also be projected away further than WTAs (Figure 6.5A). As a comparison, the A region of SasG comprising five B repeats would only be available if it were attached to the most outer PG layer (image not shown). For SasC, the B region always comprises eighteen DRESS domains (see section 4.3.1) and is thus expected to always project its A region further than WTAs.

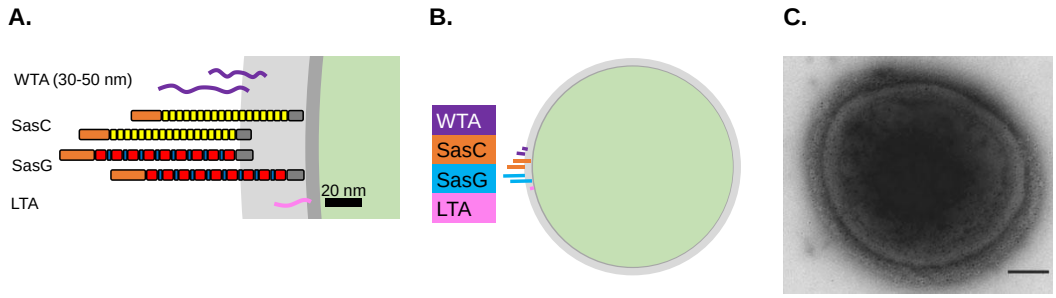


Figure 6.5: Length comparison of SasC, SasG, WTA and LTA on an *S. aureus* cell. Relative lengths drawn to scale. **A.** Scaled for the relative comparison of components on the *S. aureus* cell. Orange: A region of CWA proteins; grey: PG anchoring region of CWA proteins. *S. aureus* cell, from left to right: PG, cell membrane and cytosol. **B.** Scaled for relative comparison with an *S. aureus* cell. **C.** Negative stain image of S. aureus cell over-producing SasG labelled with antibodies against the SasG A region. Reproduced from Corrigan *et al*. (2007)[158]. Scale bar represents 100 nm.

6.1.6 Mechanical strength of DRESS domains

6.1.6.1 Comparison of mechanical strength of DRESS domains with other α-helical domains

Generally, multi-domain proteins containing α-helical domains are expected to be mechanically weak due to the alignment of hydrogen bonds that stabilise the secondary structure with the direction of the force[342,488]. Here, the F of DRESS domains is determined to be 109 pN ± 40 pN at a constant unfolding speed of 1500 nm/s (from a single dataset, more datasets recorded, see section 4.3.8.2). The F at lower constant unfolding speeds is analysed but not yet determined. The F of other tandem α-helical domains is lower than is observed for DRESS domains. For example, spectrin is a structural component of erythrocytes and contributes to their mechanical stability. Spectrins exist in head-to-tail organised triple helical bundles, which are connected by contiguous helices (Figure 6.6A). They unfold at 25-35 pN at a constant applied pulling speed of 800 nm/s per single or tandem domain[342]. Cooperative unfolding of tandem spectrin domains under force was suggested by Rief (1999)[342] and others[489,490], but Randles *et al*. (2007)[426] claim that cooperative unfolding only happens in thermal or chemical denaturation.

Talin is a structural adaptor protein with a mechano-sensitive function (Figure 6.6C). Talin domains are in a head-to-tail conformation and unfold at 5-25 pN. Cooperative unfolding has been observed for repeat 8, which is protected from force by repeat 7. Upon mechanical unfolding of repeat 7, the more mechanically compliant repeat 8 unfolds in a cooperative manner, leading to an increase in ΔL of approximately two talin domains[423].

ANK repeats comprise two anti-parallel α-helices present as side-by-side stacked tandem repeats in cytoskeletal adaptor proteins[491] (Figure 6.6B). They feature an extensive hydrophobic interface and hydrogen bonding network between repeats, which contributes to these repeats stacking to form a full helical turn for every 24 repeats, possessing a spring-like function[492] and refolding rapidly after mechanical unfolding[425]. Their F is ~50 pN at a constant applied pulling speed of 400 nm/s. Both cooperative and individual domain unfolding can be observed within a single force-extension curve[427].

β-catenin performs a mechano-transduction function in managing cell-to-cell and cell-to-extracellular matrix adhesion[493]. It contains twelve tandem armadillo repeats, comprising three α-helices that stack side-by-side (Figure 6.6D). They display a multimodal ΔL distribution with a monomodal force distribution upon mechanical unfolding, suggestive of multiple unfolding pathways, which may include cooperative transitions. The average unfolding force is 44 pN[493].

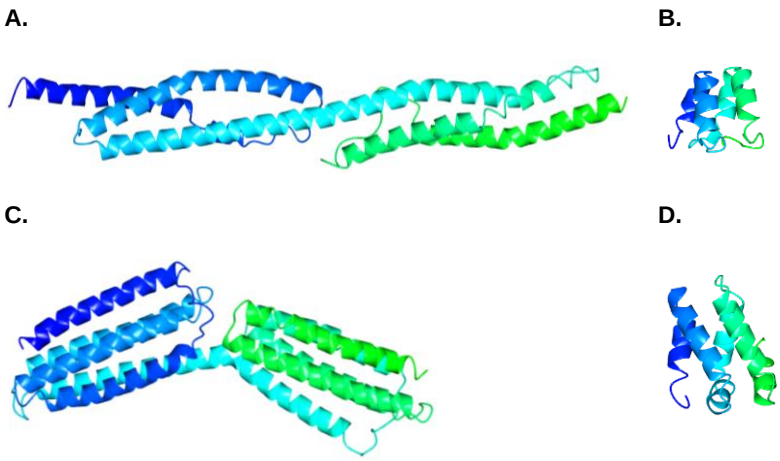


Figure 6.6: Crystal structure from tandem α-helical repeats blended from N to C in blue to green. A. spectrin (PDB 1hc1)[379]; **B.** ANK (PDB 1n11)[491]; **C.** talin (PDB 3dyj)[461]; **D.** armadillo (PDB 2z6h)[493]. Image was created using CCP4mg.

Spectrin and talin contain α-helical domains in a head-to-tail conformation, like DRESS domains. On the contrary, α-helical ANK and armadillo repeats stack side-by-side. Head-to-tail organised tandem domains have hydrogen bonds in α-helices aligned to the unfolding force, while the hydrogen bonds in α-helices of side-by-side organised tandem domains are approximately perpendicular to the externally applied force. This comparison suggests that head-to-tail organised domains tend to have a lower mechanical unfolding force than side-by-side organised domains, partly due to the relative orientation of the hydrogen bonding, among other inter-domain stabilising forces, such as hydrophobic

packing between large interfaces. Thus, with an average unfolding force of 109 pN ± 40 pN at 1500 nm/s as determined from a single dataset, DRESS domains are among the most mechanically strong α-helical tandem head-to-tail organised domains, to be reported.

6.1.6.2 Refolding of DRESS domains under mechanical load

Other tandem α-helical domains from repetitive protein regions have varying abilities to refold after the relaxation of an applied mechanical load. In armadillo repeats, sub-sections of individual domains were observed to fluctuate between a folded and unfolded state during unfolding and refolding cycles[425]. This is thought to represent refolding intermediates and transient folding/unfolding events, which allow the protein to fine-tune its length and tension in response to a mechanical load. In contrast, talin domains unfold and refold with a characteristic jump in ΔL corresponding to an entire domain[423], as does ANK[425]. For spectrin, refolding has only been shown qualitatively[342].

Here, refolding of a DRESS domains occurs within a second as an apparently single transition after relaxation of the applied mechanical load, as observed for spectrin and armadillo repeats. However, sub-sections of a DRESS domain can also refold under constant applied force. These observations might imply multiple unfolding/refolding pathways, with some transitions influenced by the stabilising effect from an adjacent, folded domain. A similar mixed unfolding/refolding pathway was observed for armadillo repeats[425].

6.1.6.3 Multi-domain unfolding

It is helpful to consider the potential unfolding order of individual domains in a multi-domain protein. In AFM studies of DRESS domains from SasC, typically six unfolding events were observed; of which three were mechanically strong (170 pN; see Figure 6.9) and three were mechanically weaker (~100 pN; see Figure 6.9).

Detailed unfolding studies in proteins have been performed, for example for a single domain of titin, as studied by AFM, mutational studies and molecular dynamics simulations[494]. Sequential unfolding pathways of multi-domain proteins have been studied to a lesser extent. For phosphoglycerate kinase that consists of two domains, two unfolding pathways were detected by AFM and coarse-grained simulations, where unfolding starting at the N-terminus encompassed a transient intermediate and unfolding of the C-terminus

proceeded without detectable intermediates[495]. Another interesting example is the sequential unfolding of β-helical protein regions in the extracellular part of the bacterial protein TpsA[496]. Using molecular dynamics simulations of mechanical unfolding, they observed a sequential unfolding pathway, where the core fold was sequentially unfolded from both ends. Sikora and Cieplak (2011)[497] discussed mechanical clamps in multi-domain proteins with rich contacts between domains, that increased their mechanical stability. Among these clamps is the shear between anti-parallel β-sheets[497], a clamp that might also exist between anti-parallel α-helices. Apart from anti-parallel packing between α-helices, no molecular clamp architecture is present in tandem DRESS domains D1617.

The order in which DRESS domains unfold under mechanical force, is currently unknown. Here, the unfolding pathway is speculated to involve sequential domain unfolding, based on the previously presented information obtained for DRESS domains. To allow this, the following assumptions were made about DRESS domains: that two DRESS domains (with one inter-domain interface) bear some mechanical stability, that each DRESS domain flanked by multiple rod-like DRESS regions on either side has an identical unfolding force, that unfolding of the final domain coincides with detachment from the probe and that the eight-domain protein D0310_scc is non-specifically attached to the probe at the far N-terminus. These assumptions are required[498] to propose some sort of unfolding pathway in the absence of experiments probing the unfolding trajectory, such as molecular dynamics simulations or mechanical unfolding of a protein containing FRET pairs.

The multi-domain protein D0310_scc comprises eight DRESS domains and is anchored to a surface at the C-terminus. The protein is non-specifically picked up by the probe and attachment can occur anywhere along the rod, where up to eight unfolding events might be observed in the statistically rare event of probe attachment at the far N-terminus. The position of the first unfolding event is currently not known. Previous experiments revealed that DRESS domains are more thermally stable in longer arrays and mechanical unfolding events showed that generally, the first unfolding event was the strongest and the unfolding force decreased with the sequential unfolding of the protein (see section 6.1.7.3).

If mechanical unfolding of DRESS domains occurred via a non-adjacent pathway (Figure 6.7A) maximally two “strong” unfolding events would take place, before the remaining domain pairs would unfold at lower force. In experimental traces, usually more strong

unfolding events were observed (Figure 6.9); hence this unfolding pathway is probably not correct.

Another mechanical unfolding pathway could involve sequential domain unfolding. Here, the longest, most stable rod would be preserved and might provide a long-range stabilising effect to intact DRESS domains. In order to sustain the highest stability via the longest multi-domain protein, mechanical unfolding might start at a terminus (Figure 6.7B) and work its way sequentially along the length of the rod. This pathway might incorporate up to 6 “strong” unfolding events. However, this is in disagreement with the observed number of strong unfolding events in Figure 6.9.

In Figure 6.9, three “strong” unfolding events are followed by weaker unfolding events, before the protein detaches from the probe at a typically large force. This might be in best agreement with a sequential unfolding trajectory, initiated at a random position in the rod (Figure 6.7C). When unfolding starts at domain 3 or 6, three strong unfolding events are expected. When unfolding starts at domain 4 or 5, four strong events might be expected (not shown).

Further work might involve tracking the unfolding of individual domains e.g. by fluorescence, to further study the mechanical unfolding trajectory of DRESS domains. Furthermore, it would be very interesting to mechanically unfold the full DRESS region, comprising eighteen DRESS domains. Possibly, more “strong” unfolding events might be observed and the force required to unfold the first domain might be even higher. Finally, molecular dynamics simulations could reveal further details about the mechanical unfolding pathway of DRESS domains.

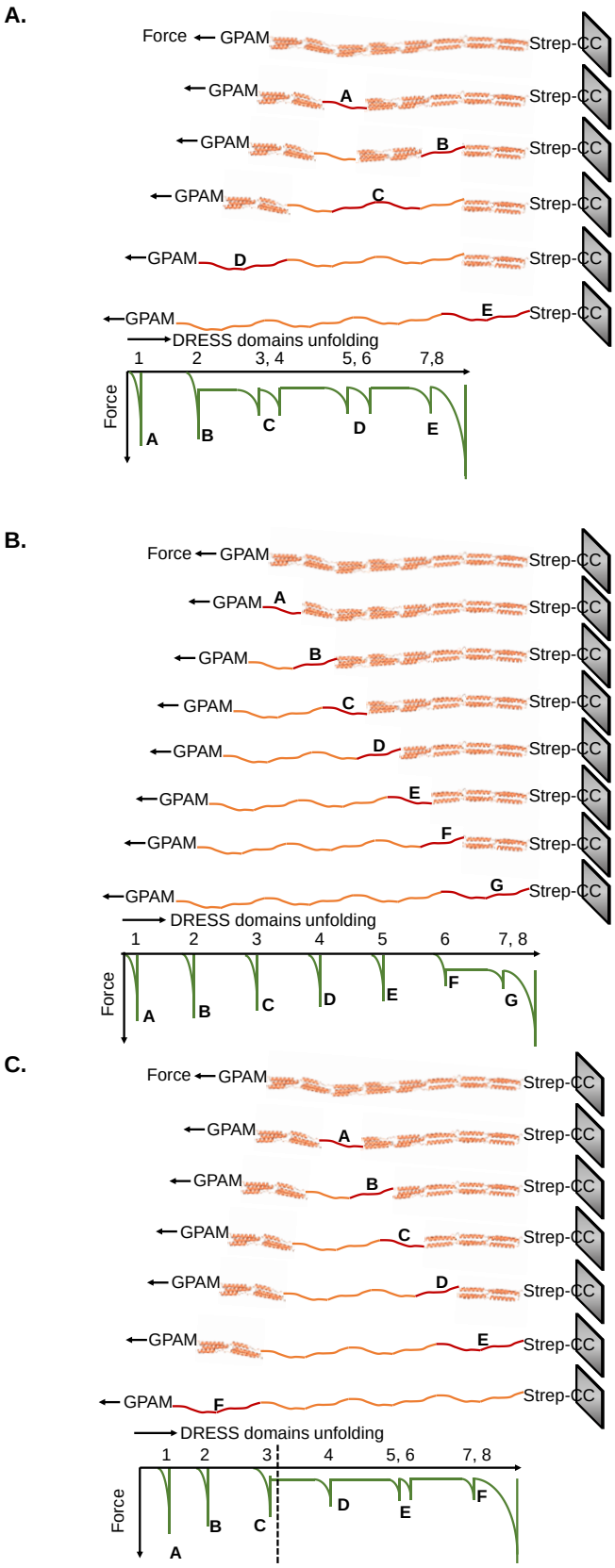


Figure 6.7: Mechanical unfolding pathways and a schematic of a predicted force-extension curve of DRESS domains. A. Non-adjacent domains unfold first. **B.** Sequential domains unfold from a terminus. **C.** Sequential domains unfold from random seed. Dashed line separates “strong” unfolding events from “weaker” events.

6.1.6.4 Forces in biofilms

Physiological forces in biofilms

Bacteria are able to form biofilms in environments within the host, where they are subject to shear stresses, such as in saliva or in blood. The shear stresses in these environments are estimated to be 0.36 pN in saliva and 4.5 pN in blood on a single cell[480,481,499]. When biofilms of *S. aureus* were grown under flow (4.5-45 pN per cell), their stiffness increased threefold compared to growth in a static environment[500] and extracellular proteins were partly responsible for the integrity of the biofilm, as biofilm growth in the presence of proteinase K limited the rigidity of the resulting biofilm[500]. On the contrary, biofilm growth in the presence of DNAse did not have an effect on the rigidity of the mature biofilm, suggesting that CWA and secreted proteins play a key role[500], although it is unclear what the contribution of extracellular DNA is in biofilms that have been grown for 5-6 hours[501].

Mechanical strength of biofilm-mediating proteins

Proteins that are regularly exposed to mechanical tension, typically display topological solutions to resist this force[502]. CWA proteins on the surface of *S. aureus* are exposed to a variety of forces in different environments. *S. aureus* can initiate infections on wound tissue via adherence to collagen, employing Cna[503,504]. This interaction has a F of ~1200 pN and is complemented by the stiff spring-like properties of the B repeats, which are mechanically resilient and function as a stalk for the Cna A domains[503]. Infections on in-dwelling medical devices can also be initiated by SdrC. An *S. aureus* cell over-producing SdrC adheres to plastic surfaces with a force of >5000 pN through unfolding of multiple SdrC molecules to expose hydrophobic patches[147]. In comparison with the expected physiological forces exerted by shear stresses on single cells, the mechanical stability of a single copy of Cna or SdrC would be more than sufficient to ensure secure attachment in these conditions.

Interactions between cells can be promoted through the homophilic association of B regions of SasG on opposing cells in the presence of Zn^{2+}, which is hypothesised to smoothen exposed WTA, thereby exposing the B regions of SasG[61]. The F of specific homophilic interactions between B regions of SasG on the surface of an *S. aureus* cell is ~414 pN[61]. Determination of the F for individual domains within the repetitive region of SasG revealed that mechanically strong G5 domains (421 pN) are 'protected' from

unfolding by a force buffer, provided by mechanically weaker E domains (250 pN)[57]. This mechanism involves non-cooperative unfolding of domains in the repetitive region, allowing SasG to tune its length and flexibility to the local environmental conditions. SdrC also mediates weak cell-cell adhesion through homophilic interactions, which yield under forces of ~40 pN and strengthen over time to ~280 pN by the formation of multiple parallel interactions. The forces involved in cell-cell interactions seem to be weaker than forces that are putatively responsible for surface attachment; nevertheless, the F of cell-cell interactions is estimated to be <10 times stronger than the estimated shear force on a single cell in blood.

The mechanical force required to unfold DRESS domains is ~109 pN ± 40 pN at 1500 nm/s (analysed from a single dataset), much weaker than the forces observed for proteins involved in initial adherence to other cells or surfaces, such as Cna or SdrC. Rather, this force is of the same order of magnitude as the pairwise homophilic interactions between SdrC molecules on opposing cells in cell-cell interactions, important in the accumulation phase of biofilm formation. This might suggest that SasC contributes more to biofilm accumulation than to biofilm formation, which would be in agreement with functional studies on SasC by Schroeder *et al*. (2009)[111]. However, currently this hypothesis is speculative and further work is required to test it experimentally. For a persistent pathogen such as *S. aureus*, it is advantageous to retain redundant means to achieve biofilm formation in order to succeed in many niche-specific environmental conditions[174], including those in which stronger shear forces might be present.

6.1.7 The DRESS domain interface mediates stability and cooperativity

6.1.7.1 Cooperativity in DRESS domains

DRESS domains D16 and D17 bury a large percentage of their surface area in inter-domain interfaces with adjacent DRESS domains (see Table 3.5). Furthermore, the presence of an inter-domain interface significantly increases the thermal stability of adjacent DRESS domains, as shown from the large difference in T_m between single and tandem DRESS domains (see section 3.3.7), and from the steric and electrostatic disruption of the tandem domain interface (see Figure 3.26).

The presence of a two-state sigmoidal unfolding transition for tandem DRESS domains implies energetic coupling of structural elements, referred to as cooperativity[505].

Cooperativity is observed in domains with short rigid linkers or connecting elements of secondary structure. For example, spectrin domains are connected tightly by contiguous helices[472] and unfold cooperatively in a single transition, as shown by thermal denaturation, urea denaturation[205] and putatively[426] by AFM[490,506]. Another example is the cooperativity of the repetitive region of SasG, where the interfaces between G5 and E domains mediate the stability of the entire repetitive region, resulting in a single thermal unfolding transition[57]. On the other hand, triple-helical domains in SpA are connected by a long and flexible linker; this decouples the cooperative unfolding of adjacent domains[56].

Here, cooperativity within the DRESS region is proposed to be mediated by the domain interfaces. This is probed in thermal and mechanical assays. In the future, this might be extended to interpret the association of DRESS domains into a rod-like structure using the 1D Ising model[507].

6.1.7.2 Thermal stability of DRESS domains

The thermal stability of DRESS domains was probed by CD and nano-DSF (see section 3.3.7). The T_m increases with the number of DRESS domains up to a length of 4-8 domains, after which the T_m as determined from nano-DSF remains approximately constant up to eighteen domains (Figure 6.8). This supports a medium-range cooperativity effect.

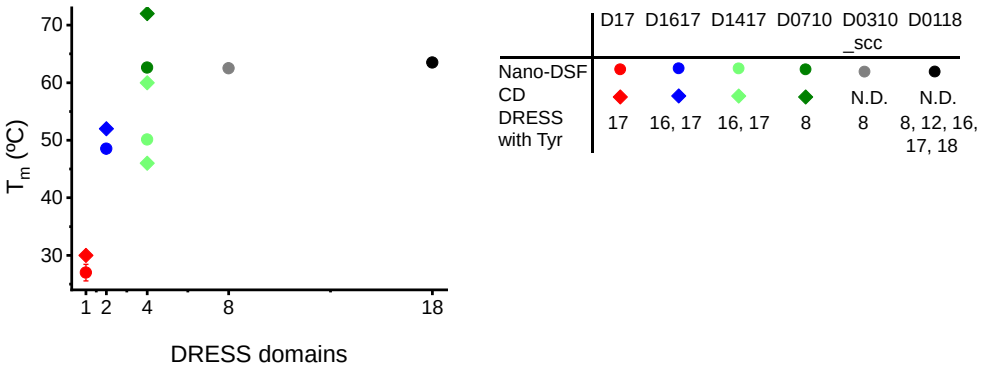


Figure 6.8: Thermal stability by nano-DSF and CD as a function of the number of DRESS domains in the protein construct (see section 3.3.7). T_m values measured by nano-DSF were determined in duplicate in 20 mM Tris, 150 mM NaCl, pH 7.5 at 1 mg/mL protein; except D0310_scc, which was in 20 mM Tris, 150 mM NaCl, 1 mM TCEP, pH 7.5. All T_m values measured by CD were determined in 20 mM sodium phosphate at 0.2 mg/mL protein, except D1417 which was determined on a mixture of oligomeric and monomeric recombinant protein (see section 4.3.3). The pH was adjusted to the most stable pH of each construct, as determined from CD measurements. D17 and D1617 were measured at pH 5.5 (see Figure 3.15), D0710 at pH 4.9 and D1417 at pH 4.6.

Between the two different techniques, an additional thermal denaturation transition was observed by CD for D1417. This could have been due to the mixture of oligomeric and

monomeric protein; however, the T_m measured by nano-DSF was identical for these fractions (see Figure 4.10). Another possibility would be that domains 14 and 15 unfold separately from domains 16 and 17, as nano-DSF can only detect the unfolding transition of domains 16 and 17 due to the absence of tyrosine residues in domains 14 and 15. If correct, this would be in disagreement with the observed cooperative thermal unfolding transitions for other constructs containing multiple DRESS domains. For D0710, a higher T_m was detected by CD and in both techniques, a single transition was detected, with the note that the unfolding transition detected for D0710 by CD was broader than for D1617 (see Figure 4.10). This might suggest that the tyrosine residue in DRESS domain 8 is exposed to solvent prior to the loss of α-helical secondary structure. This is in agreement with the partly solvent exposed orientation of equivalent residues Ala1824 and Phe1901 in the crystal structure of D1617. For the two four-domain recombinant DRESS constructs, quite different T_m values were obtained. This might suggest a different thermal stability across the DRESS region. More experiments are required to follow this up.

6.1.7.3 Mechanical stability of DRESS domains

An inverse trend was observed between the number of DRESS domains unfolded and the force required to unfold the next DRESS domain (Figure 6.9). Thus, DRESS domains benefit from a long-range stabilising effect on their mechanical stability. Yet, they unfold independently via a non-cooperative pathway, as observed for individual unfolding events (and the occasional tandem domain unfolding event).

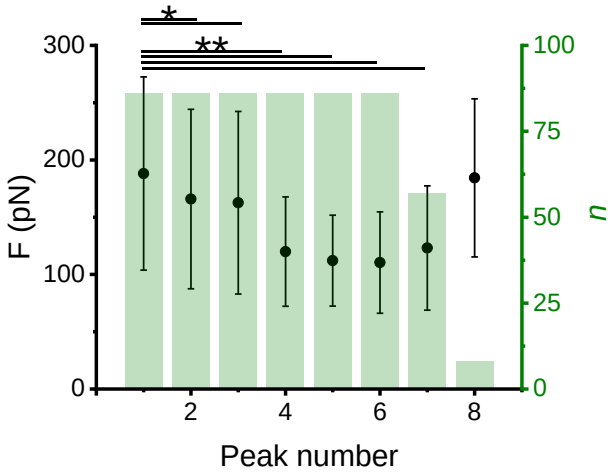


Figure 6.9: Mechanical stability as a function of the number of unfolding events (*n*). Unfolding events were automatically recognised from traces using a custom algorithm (courtesy of William Rochira) and a WLC model was fit to traces with six to ten unfolding events, excluding the first and last event. Unfolding events were grouped to sequential number and the mean and SD are presented. Here, data was analysed from traces recorded at 1500 nm/s. *: $p < 0.05$. **: $p < 0.01$. Data analysis was kindly performed by William Rochira.

The long-range mechanical stabilising effect yields an inverse force-extension curve relative to what is usually observed. Typically, weaker domains unfold first, followed by the mechanically stronger domains. An example is the selective unfolding of mechanically weaker E-domains, before the unfolding of stronger G5 domains in the repetitive region of SasG[57]. This independence between location in a multi-domain protein and the unfolding force of a domain is exploited in the use of titin as a molecular ruler[508]. Here however, it is hypothesised that all eight DRESS domains are equally strong when all domains are in the folded state. The force required to unfold the 'first' DRESS domain is the highest. Currently, the 'first' is defined as a random location in the rod (see section 6.1.6.3), but future experiments involving fluorescent labels may be able to determine which domain unfolds first.

6.1.8 Cooperativity of the DRESS region and the Ising model

The folding pathway of linear, tandemly arrayed domains, such as ANK repeats[215] and TPR repeats[213] (see section 1.5.4) can be interpreted using the 1D Ising model[507], or nearest-neighbour model[214]. Here, distant repeats do not interact, yet the tandemly arrayed domains as a whole show increased stability, where a partially unfolded state may occur during unfolding.

ANK repeats are, among other proteins, present in the Notch receptor of *Drosophila* and are responsible for signal transduction[509]. Although their sequence conservation is low

with 17%, their structural similarity is high with an average C_α RMSD of 0.17 Å[509]. Tandemly arrayed ANK domains bury 27% (750 Å^2) of their ASA in hydrophobic inter-domain interfaces (determined by PISA[338] on PDB entry 1n11, repeats 15-16)[509] and their unfolding pathway is highly cooperative[510], yielding a rigid overall tertiary structure for the full ANK region[215]. Individual domains were intrinsically unstable, but the favourable interfacial energy compensated for the energetic cost of folding[215]. Thus, the nearest-neighbour model accounted for the observed cooperativity between distant ANK domains[215]. Another well-described example where local interactions contribute to long-range stability, is in TPR repeats[220].

DRESS domains are 77 residues in size, larger than other tandemly arrayed domains[511]. The sequence conservation is around 29%, close to the 25% observed for other domains in repetitive regions[512] and well below the reported cut-off of 40% to avoid aggregation[196]; and they are structurally very similar with a backbone RMSD of 1.1 Å. DRESS domains bury ~10% of their ASA, or ~500 Å^2, in inter-domain interfaces, which is a relatively large proportion compared to other head-to-tail organised, tandemly arrayed helical bundles (see Table 3.5). They unfold cooperatively in thermal denaturation (see Figure 3.15) and independently when unfolding by mechanical force (see Figure 4.19). However, the force required to unfold DRESS domains decreases with the number of unfolding events, suggesting the presence of long-range stabilising contributions (see Figure 6.9). The inter-domain interface is essential for domain stability, as a single electrostatic and steric mutation in the interface completely disrupts the stabilising effect on tandem DRESS domains (see Figure 3.26) and the DRESS region is elongated and rigid in solution (see sections 4.3.6 and 4.3.7). Based on the structural and biophysical information presented in this thesis, it is proposed that individual DRESS domains in the repetitive DRESS region of SasC behave according to a 1D Ising model, where inter-domain interfaces stabilise both adjacent and non-adjacent DRESS domains, leading to the formation of a rigid, extended, tandemly arrayed, multi-domain rod. This hypothesis is currently under study in collaboration with Prof Doug Barrick and Mark Petersen at John Hopkins University.

6.2 Discussion for the dynamics of SHIRT domains

6.2.1 Choice of mutation

The effect of proline residues in the linker between SHIRT domains was assessed via proline-to-alanine mutations. These yield a linker sequence of three alanine residues between SHIRT domains. Literature research was carried out to ensure that a triple-alanine sequence did not generate any unfavourable effects. While sequences containing over 11 alanine residues have a decreased flexibility due to rigidity in the chain[444], no such effect has been observed for 3 alanine residues[444,513]. Furthermore, George and Heringa (2003)[439] analysed a larger dataset and noted that alanine residues were less favoured in non-helical linkers, but there was a slight preference for alanine over other amino acids in three-residue linkers. Thus, it is expected that the inclusion of the amino acid sequence Ala-Ala-Ala does not reduce flexibility and the effects can be attributed to the loss of proline residues.

6.2.2 The interface between SHIRT domains

The crystal structure of S0304 (courtesy of Dr F. Whelan) showed that the size of the interface is very small with 85 Å^2 per SHIRT domain (see Table 3.5). This is confirmed by thermal denaturation studies and solution studies by NMR. The T_m of tandem SHIRT domains was only 0.5 °C higher than that of a single SHIRT domain (see section 5.3.3). The (^{1}H,^{15}N)-HSQC-spectrum of S0304 showed only few non-overlapping resonances (5.3.4); indicating that the chemical environment of only a few residues was changed by the introduction of another domain.

Generally, multi-domain proteins with small inter-domain interfaces tend to contain domains that are thermodynamically decoupled, have longer linkers and behave according to the beads-on-a-string architecture[196]. For example, such behaviour was reported for the tandem repeats of titin[204]. Although they have short linker sequences, they display minimal inter-domain interactions[204] with some bending and twisting along the extended conformation[203]. SHIRT domains also have shorter linkers and are expected to have a thermodynamically decoupled stability of domains, as observed from the minimal increase in thermal stability between S03 and S0304.

6.2.3 Flexibility of backbone amides in SHIRT domains

The flexibility of backbone amide resonances as determined from the (^{1}H, ^{15}N)-hnNOE ratios in a single SHIRT domain was very minimal and only displayed one loop with marginally increased flexibility between strands three and four (see section 5.3.8). The relaxation properties of backbone amide resonances putatively assigned to the linker residues between the tandem SHIRT domains S0304 and S0304P_704A,P706A indicate that mutation of proline residues to alanine increases the flexibility of the linker on the ps-ns timescale. This is in agreement with the structural rigidity that is provided by proline residues[439].

In order to separate the putatively anisotropic global correlation time from the intramolecular motions, model-free analysis could be performed. This might provide further in-depth information on the relaxation properties of backbone amide residues on different timescales[514,515].

6.2.4 Comparison of relaxation parameters with other proteins

The relaxation properties of tandem SHIRT domains presented in chapter 5 suggest limited flexibility of the linker region between SHIRT domains. This approach is common in studying relaxation parameters of multi-domain proteins.

Barbato *et al*. (1992)[296] showed that the central helix between the N- and C-terminal domains of calmodulin bears flexibility based on the (^{1}H, ^{15}N)-hnNOE ratios of the linker residues. Furthermore, calmodulin domains rotate independent of each other, as observed from the two standard deviation difference in τ_C between calmodulin domains.

The amino-terminal fragment (ATF) of urokinase-type plasminogen activator (u-PA) contains a kringle domain (larger MW) and a growth factor (GF, smaller MW) domain. The GF domain was best modelled using anisotropic analysis, while the kringle domain was best modelled using isotropic analysis[516]. This fits with the solution structure of the domains, where the GF domain is highly anisotropic and the kringle domain is more globular[517]. Despite the independent motions between the beads-on-a-string domains, the amides in the linker region did not show decreased (^{1}H, ^{15}N)-hnNOE ratio or a lower order parameter in model-free analysis[514].

The large difference in individual correlation times between Src homology (SH) 3 and SH2 domains in FynSH32 suggests independent inter-domain motions. This is partly supported by a low (^{1}H, ^{15}N)-hnNOE ratio for the backbone amide of Ile144, located in the linker between SH3 and SH2 domains. Two residues following Ile144 have (^{1}H, ^{15}N)-hnNOE ratios above 0.5 and form a short 3_{10} helix, maintaining the relative orientations of SH3 and SH2 domains[453].

Finally, the relative mobility between Vaccinia virus complement control protein (VCP) modules was studied by Henderson *et al*. (2001)[518]. The average (^{1}H, ^{15}N)-hnNOE ratios and the ^{15}N T1 and T2 relaxation constants are different for VCP modules 2 and 3. Furthermore, the backbone (^{1}H, ^{15}N)-hnNOE ratios of the linker residues between VCP 2 and 3 remain above 0.5 and decrease gradually from module 2 to module 3. The MW of VCP2 is larger than that of VCP3 and the additional residues form a more elongated structure, explaining the observed difference in relaxation parameters. The limited flexibility of the linker residues was supported by the observation of NOE correlations between nuclei in the linker with both adjacent VCP domains.

The τ_C of the individual and tandem domains described above are shown in Table 6.3. Furthermore, the τ_C of each tandem domain is estimated from the MW according to Equation 6.1[295]. Generally, the τ_C values of individual domains approach more closely the estimated τ_C of the tandem domain when the linker is rigid, because the rotation of domains is tightly linked. This effect is supplemented by the more anisotropic shape of multi-domain proteins with rigid linkers.

Equation 6.1: Estimation of rotational correlation time[295].

$$\tau_C \approx 0.6 * MW$$

In this chapter, the hypothesised rigidity in the linker between SHIRT domains was assessed. Similarly as for the linker between tandem SHIRT domains, high (^{1}H, ^{15}N)-hnNOE ratios were observed for residues between the kringle and GF domain in the ATF fragment of u-PA[516], between VCP domains 2 and 3[518] and partly between FynSH32 domains[453]. Thus, a less flexible linker region between tandem domains is not very uncommon.

Table 6.3: Isotropic and anisotropic correlation times for multi-domain proteins with rigid or flexible linkers. Relaxation obtained from ^{1}H, ^{15}N resonances. [a]Rotational correlation time was estimated from the MW, assuming a globular protein[295]. [c]The rotational correlation time of these multi-domain proteins was calculated from the weighted average from the appropriate τ_C from the individual domains and their number of residues. [d]Values from Table 5.5.

Protein		MW (kDa)	τ_C (ns)		[a]τ_C (ns) estimated	Linker nature	No. linker res.	Reference
			Isotropic	Anisotropic				
Calmodulin	Total	16.7	[c]6.7		10.0	Flexible	6	[296]
	N	7.4	7.2 ± 0.4		4.47			
	C	8.4	6.2 ± 0.5		5.01			
ATF	Total	15.0	[c]7.7		9.0	Rigid	4	[516]
	Kringle	9.7	7.3 ± 0.2		5.8	Isotropic behaviour		
	GF	4.6		10.5 (τ1), 7.0 (τ2)	2.7	Anisotropic behaviour		
FynSH3-SH2	Total	19.1	[c]9.4		11.5	Flexible, while maintaining a relative domain orientation	10	[453]
	SH3	7.0	8.8 ± 0.1		4.2			
	SH2	11.4	9.7 ± 0.2		6.8			
VCP2-3	Total	12.8	[c]7.1		7.7	Fairly rigid	4	[518]
	VCP2	6.2		7.9 ± 0.3	3.7	A_x symmetric model was best		
	VCP3	6.1	6.6 ± 0.3		3.6	Isotropic model was best		
S03		9.7	5.8 ± 0.1		5.8	N/A	N/A	N/A
S0304	Total	19.0	[c,d]11.8		11.4	Rigid	3	N/A
	S03	9.6	[d]11.8 ± 0.3		5.9			
	S04	9.2	[d]12.2 ± 0.3		5.5			
S0304_ P704A, P706A	Total	18.9	[c]10.6		11.4	More flexible than S0304	3	N/A
	S03	9.6	[d]10.4 ± 0.2		5.7			
	S04	9.1	[d]10.8 ± 0.2		5.5			

6.2.5 Towards resolving spectral overlap

Due to the high spectral overlap between S03 and S0304, only 15 out of 157 backbone amide resonances can be putatively assigned. If resonances were non-overlapped, more information might be extracted about the backbone dynamics within SHIRT domains. This section discusses ways in which the observed high spectral overlap can be circumvented.

The most straightforward option would be to select SHIRT domains with the lowest sequence identity. That way, more resonances will be in non-identical positions, despite the overall fold likely being very similar. SHIRT domains 1 and 10/11 (these are identical) are the most different with 73% sequence identity. However, production of tandem SHIRT domains comprising domains 1 and 10 leads to the formation of a non-native interface, which is undesirable as it might not represent the native state[211]. Adjacent SHIRT domains with the lowest sequence identity are domains 1 and 2 (83% sequence identity) and 12-13 (88% sequence identity), however 'capped' terminal domains 1 and 13 are no suitable representatives of the repetitive region, because of their domains make contacts with other parts than repetitive domains[214]. Thus, this is not a good option to reduce the spectral overlap in tandem SHIRT domains.

Mutation of key residues in one SHIRT domain of a tandem protein construct resolves spectral overlap well (data courtesy of Dr Michael Plevin). For example, this has been shown for mutation of Val643, a key residue in forming the stabilising core of SHIRT domains (Figure 6.10C), into Ala (Figure 6.10A) or Thr (Figure 6.10B). Resonances affected by the mutation do not overlap with those of the wild type (Figure 6.10, Glu642), while other resonances overlap (Tyr624). If a mutation were found, where the (thermal, chemical) stability of SHIRT domains is very marginally affected by the mutation; then this approach might be fruitful in resolving spectral overlap.

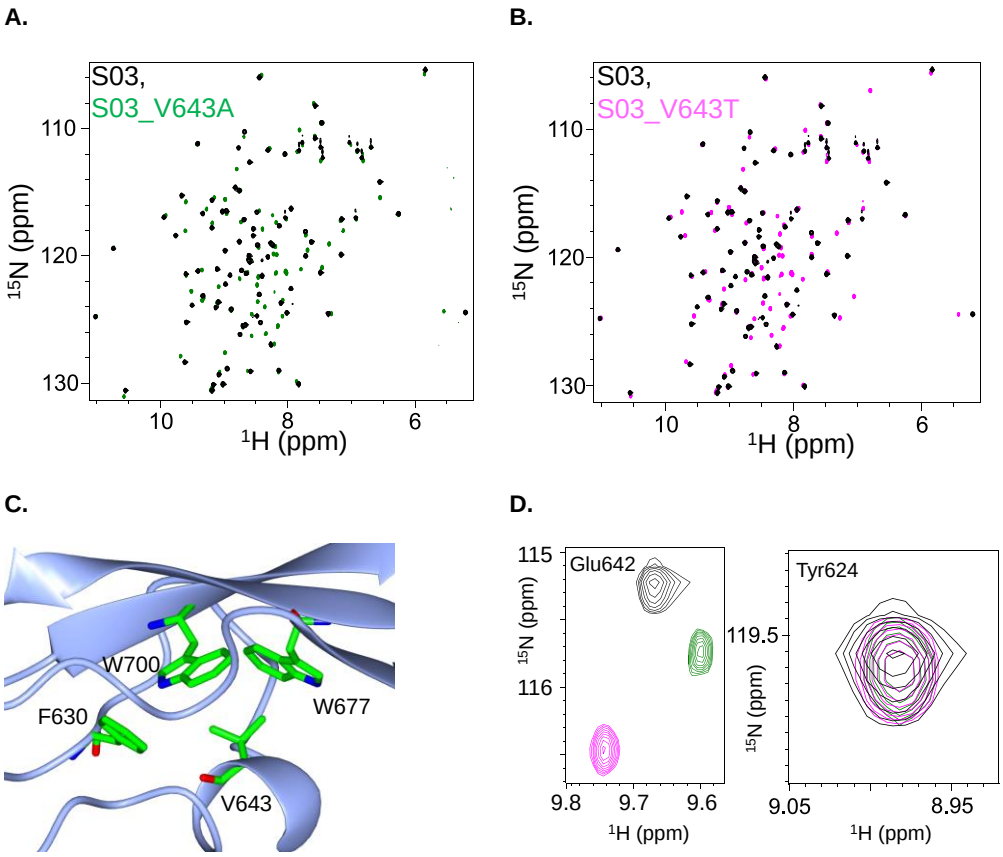


Figure 6.10: (^{1}H,^{15}N)-HSQC spectra of S03 (black) with mutations of Val643. NMR data courtesy of Dr M. Plevin. **A.** S03_V643A (green). **B.** S03_V643T (magenta). **C.** Position of V643 in the S0304 crystal structure (courtesy of Dr F. Whelan). Image was created using CCP4mg. **D.** Zoom of non-overlapping residue, Glu642, and overlapping residue, Tyr624 of S03 (black), S03_V643A (green), S03_V643T (magenta).

Another approach, which resolves spectral overlap and does not introduce destabilising mutations within domains, is segmental labelling[519]. Here, two domains are produced separately, where the desired domain can be isotopically labelled. Following purification, the domains are ligated to form the tandem domain structure. A disadvantage might involve any additional residues introduced due to the ligation method, such as native chemical ligation (NCL), where a C-terminal thioester and an N-terminal cysteine residue are ligated to form a connecting cysteine residue[520]. This ligation technique is not useful here, where the linker residues are of key interest.

Thus, the putative assignment strategy presented in section 5.3.6 is one of the strategies, in line with those discussed above, to perform biophysical characterisation by NMR of highly identical tandemly arrayed repeats.

6.3 Conclusions

6.3.1 Background

S. aureus and *S. gordonii* are opportunistic Gram-positive bacteria that can mediate biofilm formation via multiple redundant mechanisms[8], including via adhesion to in-dwelling medical devices or damaged tissue using CWA proteins[21,30]. Furthermore, they have mechanisms, including CWA proteins, for cell-cell aggregation into a mature biofilm[19,30]. Bacteria organised into a biofilm pose a threat for healthcare[12,13], because of their increased antibiotic resistance[138]; for example, they can cause infective endocarditis, with a mortality rate of approximately 30%[6]. CWA proteins bear a structurally similar domain architecture[29], where the B region is often hypothesised or shown to form an elongated, rigid rod[29,30,57,58]. In this thesis, parts of the B regions of two CWA proteins are structurally and biophysically characterised.

Prior to this work, SasC was identified as a prevalent CWA protein of *S. aureus* involved in predominantly cell-cell interactions[111] and its B region comprises DUF1542 repeats, although the structure and function of this region in SasC remained uncharacterised. Another B region containing DUF1542 domains formed an elongated structure with kinks, as determined by EM[228]. A CWA protein from *S. gordonii*, SGO0707, is important in the formation of dental plaque and adherence to type I collagen[231]. Its B region comprises high-identity tandem repeats, of which the crystal structure was determined recently by Dr Fiona Whelan (F. Whelan *et al.*, manuscript in preparation). Tandem SHIRT domains formed a highly extended, rigid rod, despite the absence of a clear interface between domains.

6.3.2 Structural and biophysical characterisation of parts of the B region of SasC

The domain boundaries of the repetitive units in the B region of SasC, DUF1542 repeats, were redefined *in silico* and the resulting domains were characterised biophysically, allowing us to rename DUF1542 as DRESS domains. The crystal structure of tandem DRESS domains was determined and consisted of head-to-tail organised, rod-like, tandemly arrayed triple-helical bundles, which confirmed the redefinition of the domain boundaries and revealed a highly connected inter-domain interface. The importance of this interface for tandem domain stability was highlighted by biophysical characterisation and

mutational studies and the inter-domain interface mediated cooperative thermal unfolding behaviour. Extension to four, eight and the physiological length of the B region of SasC, eighteen tandemly arrayed DRESS domains, revealed a highly extended and rigid repetitive region as determined from solution and surface studies.

At ~109 pN on average (determined from a single dataset), DRESS domains are currently among the most mechanically stable α-helical folds[342,423,492,493] and DRESS domains were shown to refold. This opens up the possibility for the B region of SasC to function as a spring, together with the high twist-intermediate tilt angle between two DRESS domains that was observed in the crystal structure. The mechanical unfolding pathway was non-cooperative, yet the observed force for unfolding events decreased sequentially with an increasing number of events. This mechanical long-range stability and the observed thermal long-range stabilisation are likely mediated by the inter-domain interfaces, which form a highly connected, rigid, strong rod-like region. This may enable the display of the putatively functional N-terminal region of SasC on its elongated, rigid stalk containing DRESS domains. The predicted total length of the B region of SasC was comparable to the length of the B region of SasG[57], however for SasC, the observed extension and rigidity of the stalk is constructed solely from mechanically weak α-helices, rather than the mechanically stronger β-sheets, present in the B region of SasG. This makes SasC a so far unique member of the extracellular proteome of *S. aureus*.

6.3.3 Characterisation of the flexibility of SHIRT domains and their connecting linker

The flexibility of SHIRT domains S03 and S04 and their linker region, Pro704-Ala705-Pro706, was studied by ^{15}N backbone dynamics to determine if the proline residues in the linker were responsible for the tandem domain elongation observed in the crystal structure. To this end, proline residues were mutated to alanine residues and the backbone amide resonances of domain S03 were assigned by triple-resonance assignment procedures. In S0304, many of the backbone amide resonances of S04 overlapped with S03, as expected; indicating the absence of a highly connected domain interface. After mutation of proline residues in the linker to alanine, only two more sets of equivalent resonances became putatively non-overlapped, indicating that also in absence of proline residues in the linker, S03 and S04 did not have a significant domain interface.

^{15}N backbone amide dynamics were measured for S03. Remarkably, little flexibility was detected in SHIRT domains, with the exception of the remaining non-native N-terminal residues from the fusion tag. This highlights that SHIRT domains have a rigid fold and that little flexibility is observed, even in most linker regions.

For the non-overlapping resonances, the putative backbone amide resonance assignments of tandem SHIRT domains were inferred from the triple-resonance assignment of S03 and the position of the equivalent residues in the crystal structure of S0304. Subsequently, their ^{15}N backbone amide dynamics were studied. The (^{1}H, ^{15}N)-hnNOE ratios of resonances putatively assigned to the linker region between SHIRT domains were higher in the linker featuring proline residues than in the mutated linker sequence, suggesting that the presence of proline residues contributes to increased rigidity on the ps-ns timescale, as expected. Furthermore, the estimated effective isotropic correlation time of S0304_P704A,P706A was smaller than that of S0304, suggesting that removing the proline residues from the linker sequence allows tandem SHIRT domains to rotate faster with a smaller apparent hydrodynamic radius. Taken together, solution studies of tandem SHIRT domains confirmed the lack of an extensive inter-domain interface and the proline residues increased the rigidity of the linker region on the ps-ns timescale.

6.3.4 Concluding remarks

Based on the structural and biophysical characterisation of DRESS domains from the B region of SasC, tandem DRESS domains are proposed to form an elongated, rigid, rod-like stalk on the surface of an *S. aureus* cell able to project the putatively functional N-terminal region away from the cell wall, maximising its ability to mediate cell-cell interactions. Remarkably, this rigid rod-like structure is composed of three-helix bundles with highly connected inter-domain interfaces, rather than the more commonly observed β-sheet-rich folds.

Comparison of backbone amide resonances of single and tandem SHIRT domains revealed that the interface between tandem SHIRT domains is marginal in solution, which is in agreement with the observed extension in the crystal structure of tandem SHIRT domains. The ^{15}N backbone dynamics of SHIRT domains from the B region of SGO0707 revealed that this novel fold is nearly deficient of flexibility, even in the linker regions between SHIRT

domains, and that the proline residues in the linker are partly responsible for the observed rigidity.

Thus, nature has come up with distinctly different architectures for B regions of CWA proteins, that can both lead to extension and rigidity. Mechanically weaker α-helical bundles with highly connected inter-domain interfaces form a rod-like, highly extended structure; as do mechanically stronger β-sheet-rich folds in absence of such interfaces.

6.4 Future directions

6.4.1 DRESS domains of SasC

This work laid the foundations to allow further, in-depth study of DRESS domains. One interesting question awaiting an answer is the study of the thermodynamic stability of DRESS domains, which is related to their folding and the association of inter-domain interfaces. The observation of significant inter-domain stabilisation by the formation of an interface and cooperative thermal denaturation might suggest that the Ising model could be applied to the folding and association of DRESS domains. This work is currently ongoing in collaboration with Prof Doug Barrick and Mark Peterson (John Hopkins University, USA).

In this work, a steric and electrostatic disruption of the inter-domain interface showed that the interface couples the adjacent DRESS domains together, resulting in an increased stability. However, the molecular detail for this interaction is poorly understood. The introduction of probing mutations across the interface might reveal which interaction, if not multiple, is key for the observed stabilisation. Proposed mutations would be Thr1838Ser to study the effect of the van der Waals interaction between Thr1838 and Val1946 and Thr1838Val to study the effect of the hydrogen bond between Thr1838 and Asn1943.

Furthermore, the stabilising effect of the inter-domain interface is thought to contribute to the rigidity and rod-like behaviour of the DRESS region and possibly, to the remarkable mechanical stability. Introducing mutations in positions equivalent to Thr1838 in consecutive DRESS repeats might help to understand if the highly connected interface is also responsible for the observed extension, rigidity and mechanical stability for constructs containing multiple DRESS domains.

It would be interesting to extend the number of data points in the T_m vs number of domains plot in Figure 6.8 with T_m values for constructs containing different number of DRESS domains and for constructs obtained from different parts of the repetitive region. In this work, a discrepancy was observed between the T_m values of constructs comprising four DRESS domains from the middle (D0710) and C-terminal end (D1417) from the repetitive region and between different techniques. Currently, it is unclear if this discrepancy is caused by a different intrinsic stability for DRESS domains and if this is linked to their position along the rod of SasC.

The elongated state of an entire, intact DRESS domain-containing region from SasC was probed using surface immobilised recombinant protein and super-resolution microscopy. Surface immobilisation was accomplished through electrostatic interactions between the negatively charged protein and a poly-*D*-lysine coated quartz slide. For SasG, inter-fluorophore distances of 53 nm and 69 nm were observed at 20 and 500 µg/mL poly-*D*-lysine, respectively[57]. This suggests that at a lower surface density of poly-*D*-lysine, more conformational equilibration may occur on the imaging surface, leading to an apparent shortening of the elongated rod (see section 6.1.5.2). More control experiments have to be carried out using different concentrations of poly-*D*-lysine to determine how this phenomenon might affect the observed inter-fluorophore distance.

Mechanical unfolding experiments were carried out using a construct comprising eight DRESS domains. A decreasing unfolding force was observed with an increase in unfolding events (see section 4.3.8), suggesting that long-range interactions increase the mechanical stability of tandemly arrayed DRESS domains. It would be very interesting to determine the F of DRESS domains in a construct containing the physiological number of DRESS domains.

6.4.2 N-terminal region of SasC

Expression and display of SasC on the surface of *S. aureus* has been implicated with an increased ability to mediate biofilm accumulation[111]. However, there is a complete lack in structural and functional understanding of the N-terminal domain(s) of SasC. Based on preliminary *in silico* analyses reported in section 3.3.1.1, it is proposed in this work that the N-terminal region of SasC comprises three domains/regions and with these, might mediate adhesion to different moieties. A good starting point might be the recombinant gene expression and protein production and purification of SasC_241-550, as this was predicted

to show structural homology to the collagen-binding N-terminal regions of CWA proteins Cna from *S. aureus* and ACE from *E. faecalis* (see section 3.3.1.1).

Furthermore, functional characterisation is required to elucidate the molecular mechanism of biofilm accumulation and adherence. Schroeder *et al*. (2009)[111] carried out preliminary binding studies and found no interaction between SasC and the ECM proteins fibrinogen, the von Willebrand factor or platelets. In relation to the predicted structural homology, it might be interesting to test binding of SasC to collagen, among other components present in the biofilm ECM. Binding to collagen might propose that SasC can also mediate biofilm formation, aside of its better characterised cell-cell aggregation ability[111].

6.4.3 SHIRT domains of SGO0707

The ^{15}N relaxation analysis of backbone amide resonances in SHIRT domains could be interpreted in greater depth by separating the global correlation time contributions from the intramolecular motions of the (^{1}H, ^{15}N) bonds in model-free analysis[514]. To this end, relaxation data has been recorded at 800 ^{1}H MHz. Briefly, in model-free analysis[514], several sets of relaxation data are used to obtain time constants for (fast and slow)[538] intramolecular motions and the flexibility/rigidity for each of these motions, aside from a global estimate of the correlation time. Furthermore, the diffusion tensor of the macromolecule in solution can be determined[521,522]. Based on the crystal structure of S0304 and the observed limited flexibility for the linker residues, the z-axis of S0304 is ~7 times longer than the x- and y-axis; therefore, it might be expected that the diffusion along the z-axis is slower than in the x- and y-directions.

Furthermore, relaxation data for tandem SHIRT domains is limited by the overlapping backbone resonances for many equivalent residues in domains S03 and S04. Applying one of the strategies suggested (see section 6.2.5) for non-overlapping backbone amide assignment would enable us to obtain more detailed relaxation information across tandem SHIRT domains.

Chapter 7. Appendices

7.1 Primer tables

Table 7.1: Primers for amplification of target genes.

Construct	Primer direction	Sequence
D17	Forward	TCCAGGGACCAGCAATGGAACCTGTTATTAACAGAAAGGC
	Reverse	TGAGGAGAAGGCGCGTCAATCTAAATCATGTATTGTTTGTAG
D17_wrong	Forward	TCCAGGGACCAGCAATGAACGATAAAAAACAAGCAATTGAAGC
	Reverse	TGAGGAGAAGGCGCGTCATTGCACTAAAGCAGTGACG
D1617	Forward	TCCAGGGACCAGCAATGAAACCAGCGACAACAGTTAAAGC
	Reverse	TGAGGAGAAGGCGCGTCAATGTACATCTAAATCATGTATTGTTTG
D1417	Forward	TCCAGGGACCAGCAATGAGACGTAAACGAGCTGCGCTT
	Reverse	TGAGGAGAAGGCGCGTCATTTAATAGGATGTACATCTAAATCATGTATTGTTT
D1017	Forward	ATGGGTGGTGGATTTGCTGAAGTAGTAATTAAAACAAAGGC
	Reverse	TTGGAAGTATAAATTTTCATGTACATCTAAATCATGTATTGTTTGTAG
D1017_strep_CC	Forward	TCCAGGGACCAGCAATGGAAGTAGTAATTAAAACAAAGGCAATTGC
	Reverse	CAAATTGAGGATGAGACCATTTAATAGGATGTACATCTAAATCATGTATTGTTTG
D0710	Forward	TCCAGGGACCAGCAATGACGAAGAAACAAACTGCTACA
	Reverse	TGAGGAGAAGGCGCGTCATTTTGTTTCAGGTTGAATAATTTTAATCG
D0310_strep_CC	Forward	TCCAGGGACCAGCAATGCAAGTAACTCATAAAAAAGCTG
	Reverse	CAAATTGAGGATGAGACCATTTTGTTTCAGGTTGAATAATTTTAATC
D1-18	Forward	ATGGGTGGTGGATTTGCTCGAAGTGTTGATGCTGAAAAT
	Reverse	TTGGAAGTATAAATTTTCTGCTTCTATATCGCTTAATGCAACATTAAATCGTTTTAAAAC
S0304	Forward	TCCAGGGACCAGCAATGGCGCCGACCTAC
	Reverse	CACCCCGGCATAACGCGCCTTCTCCTCA

Table 7.2: Primers for mutagenesis

Construct	Mutation	Primer direction	Sequence
D1c-c18	T731C	Forward	ATACTGCATGTCCGGTTGTTAAACCAAATGCTAAAAAA GCAATACG
		Reverse	ACCGGACATGCAGTATCCCCACTTAAGGTCTGTATAC
D1c-c18	H1963C	Forward	CATGATTTAGATGTATGTCCTATTAAAAAG
		Reverse	GGCTTTTTAATAGGACATACATCTAAATC
D1617	T1838D	Forward	CGAAGCGGATGATGAAGAACAAAATATTGCAATAGCAC AAG
		Reverse	CTTCATCATCCGCTTCGTTATTTGCTTTAATTAAATTAAT TTTATTTGTAG
	N1943D	Forward	GATCGTAGCGATGCACAAGTTGATAAAACAGCATCATT AAATC
		Reverse	TGTGCATCGCTACGATCTTGATCAATTTGTCCAATAGC AG

Table 7.3: Primers for vector linearisation.

Name	Construct	Primer direction	Sequence
CleF	pETFPP	Forward	CGCGCCTTCTCCTCACATATGGCTAGC
CleR		Reverse	TTGCTGGTCCCTGGAACAGAACTTCC
CleF_scc	pETFPP_strep_CC	Forward	TGGTCTCATCCTCAATTTGAAAAATGCT GCTAAC
CleR		Reverse	TTGCTGGTCCCTGGAACAGAACTTCC

7.2 Strains of *S. aureus* used in MSA of the DRESS region

Table 7.4: Accession details for hypothetical protein sequences of SasC from different strains of *S. aureus*. NCBI: National centre for biotechnology information.

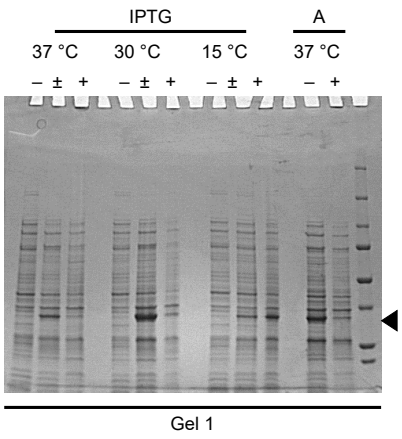
Name	Number of residues	Accession details	Accession website
EMRSA16	2189	Accession number NZ_GG770513.1, WP_001050520.1	NCBI Genome
MRSA252	2189	NC_002952.2, WP_001050520.1,	NCBI Genome
SA40TW	2185	Accession number NZ_CP013182.1, WP_001050576.1	NCBI Genome

N315	2186	Accession number NC_002745.2, WP_001050566.1	NCBI Genome
13-001	2186	Assembly GCA_001680925.1, WP_001050538.1	NCBI Genome
4074	2186	Schroeder *et al.* (2009)[111]	
M1	2186	CCW22089, BN843_17590	Kegg
ST20130941	2186	Accession number NZ_CP012978.1, WP_001641584.1	NCBI Genome
Newman	2186	NCBI Reference Sequence WP_001050548.1	NCBI Genome
Col	2186	NCBI Reference Sequence WP_001050554.1	NCBI Genome
USA300_SUR10	2186	Accession number NZ_CP014397.1, WP_077442718.1	NCBI Genome
NCTC 8325-4	2186	Assemble GCF_000013425.1, WP_001050546.1	NCBI Genome
NCTC 8325-4_CP000254	2186	CP000253	NCBI Genome
RN4220	2186	Assembly GCA_001996505.1, WP_001050546.1	NCBI Genome
V541	2186	Accession number NZ_CP013957.1, WP_001050548.1	NCBI Genome
NCTC-13454	2186	NCBI Reference Sequence, WP_001050536.1	NCBI Genome
MW2	2186	Accession NZ_JYAU01000001.1, WP_001050541.1	NCBI Genome
MSSA476	2186	Accession NZ_JXZG01000001.1, WP_001050543.1	NCBI Genome
Mu50	No SasC, contains DUF1542 domain-containing protein with predicted neuromodulin_N domains	Genome assembly 299275	NCBI Genome
ATCC 29213	No SasC, contains gene for Ebh	Accession number LHUS02000001.1,	NCBI Genome
pSA-CC022-1, pSA-CC022-2	No SasC, contains partial gene for Ebh	Accession number NZ_CM003520.1	NCBI Genome

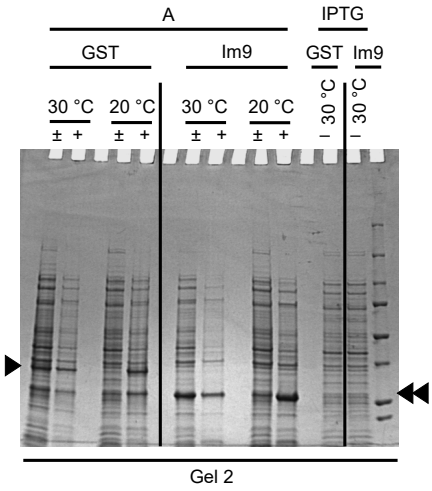
7.3 Raw gel images

Arrows above the gels indicate the relevant lanes. Arrows on the side of the gels indicate a band likely representative of the protein of interest.

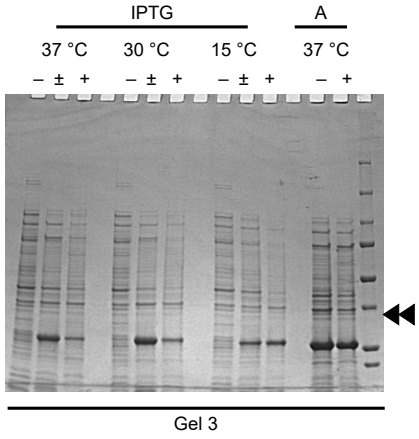
His_6-GST-D17 test for optimal expression and protein over-production conditions (Figure 3.7A)



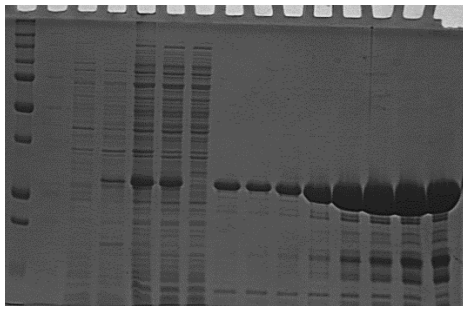
His_6-GST-D17, His_6-Im9-D17 test for optimal expression and protein over-production conditions (Figure 3.7A, B)



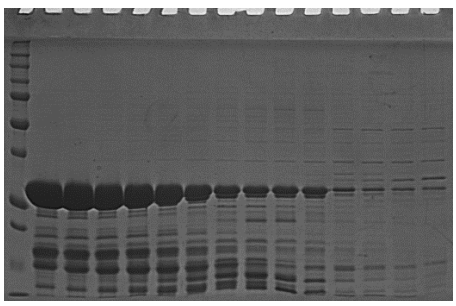
His_6-Im9-D17 test for optimal expression and protein over-production conditions (Figure 3.7B)



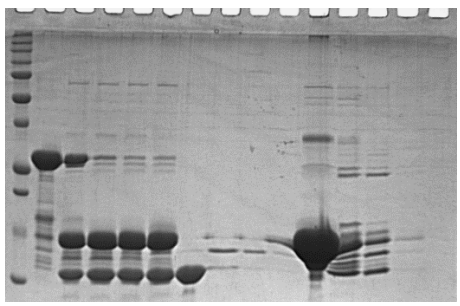
His_6-Im9-D17 IMAC 1 gel 1 (Figure 3.8B)



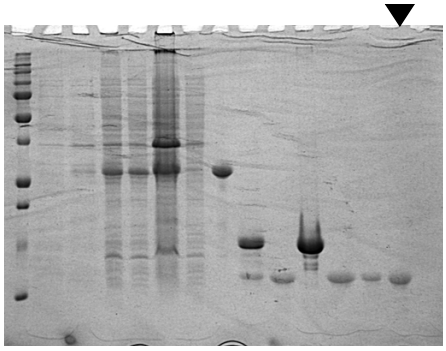
His_6-Im9-D17 IMAC 1 gel 2 (Figure 3.8B)



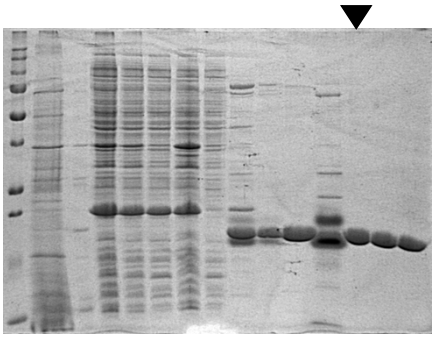
His_6-Im9-D17 HRV 3C protease cleavage and IMAC 2 (Figure 3.9A)



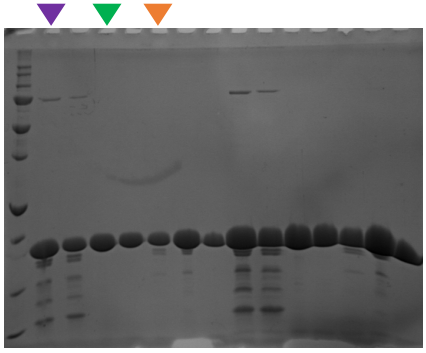
D17 (Figure 3.11)



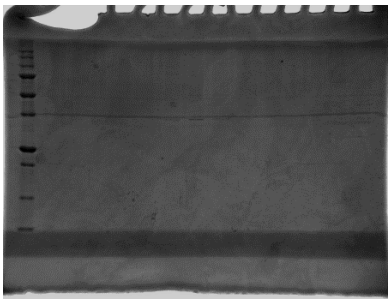
D1617 (Figure 3.11)



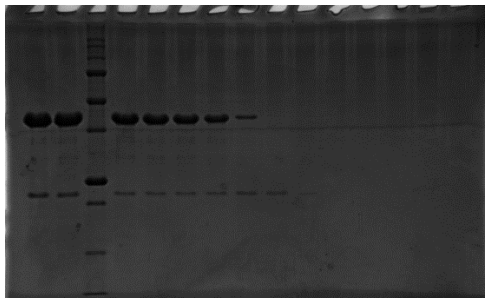
D1617_T1838D (purple), D1617_N1943D (green), D1617_T1838D,N1943D (orange, Figure 3.11)



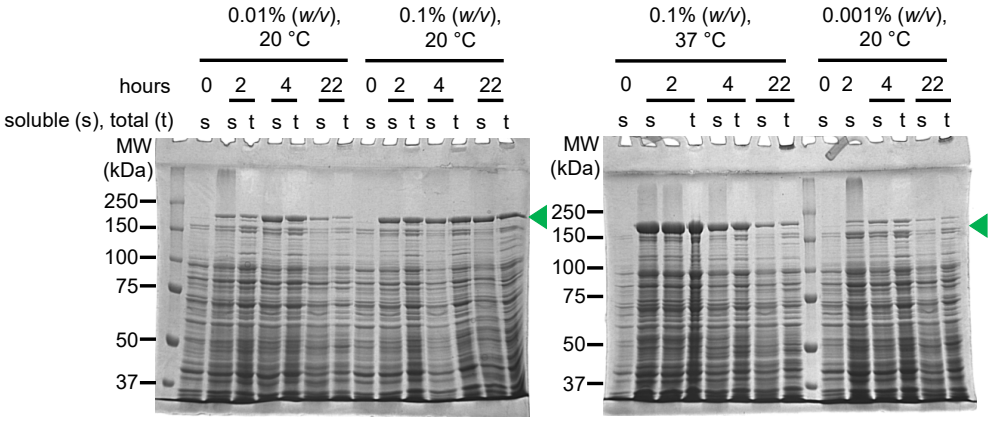
D0710 SEC gel 1 (Figure 4.3D)



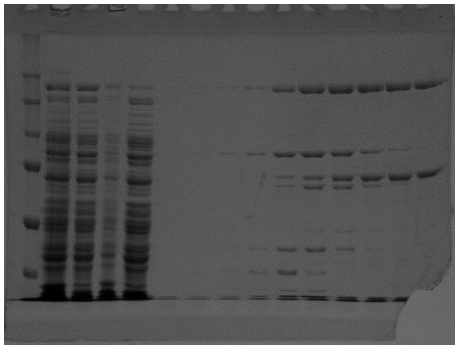
D0710 SEC 3 (Figure 4.3D)



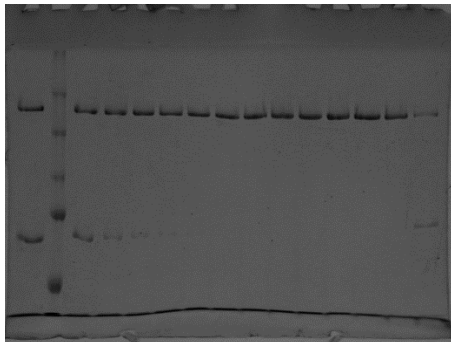
Test for optimal over-production conditions of D0118 (Figure 4.4A)



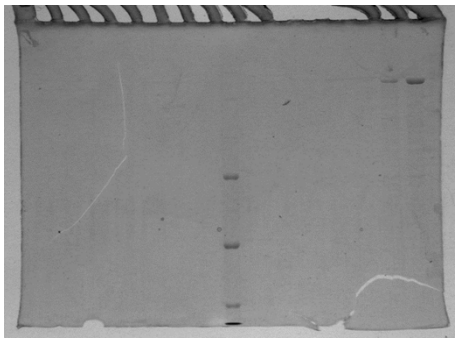
D0118 IMAC gel 1 (Figure 4.4C)



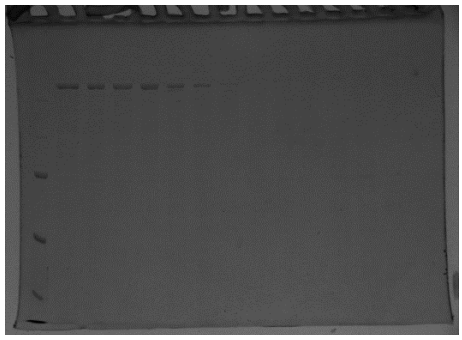
D0118 IMAC gel 2 (Figure 4.4C)



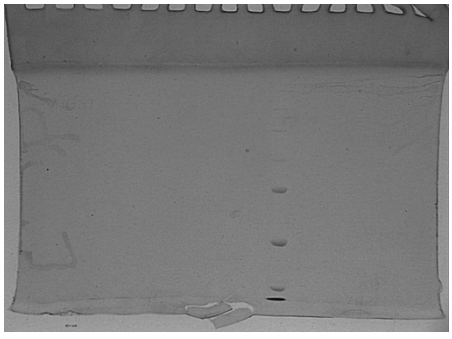
D0118 SEC left gel (Figure 4.4E)



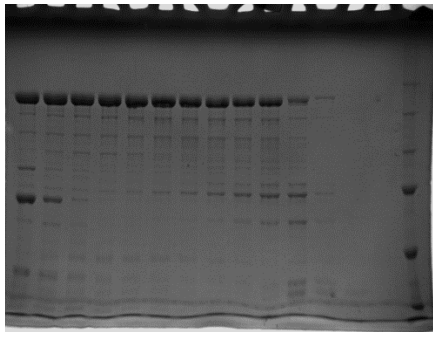
D0118 SEC middle gel (Figure 4.4E)



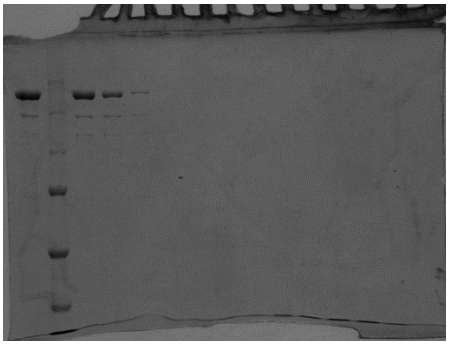
D0118 SEC right gel (Figure 4.4E)



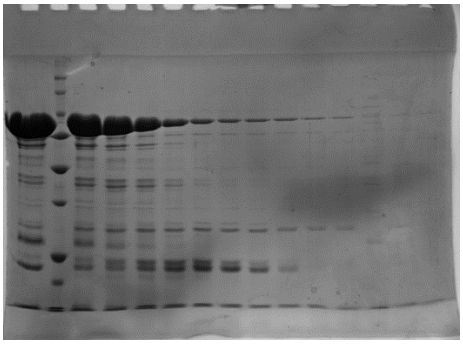
D0118_2Cys IMAC right gel (Figure 4.5C)



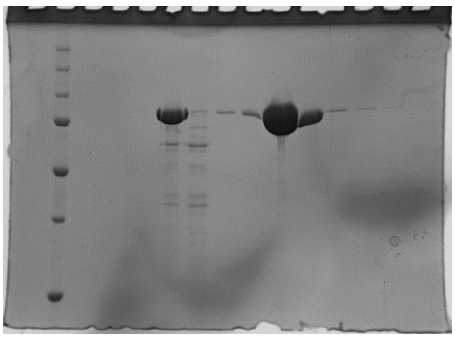
D0118_2Cys SEC right gel (Figure 4.5D)



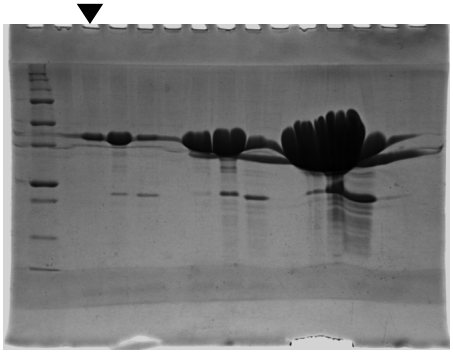
D0310_scc IMAC 1 gel 2 (Figure 4.7B)



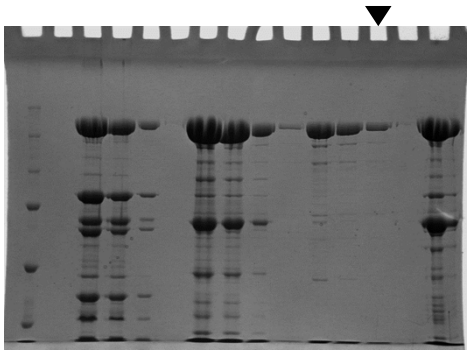
D0310_scc Strep purification (Figure 4.7C)



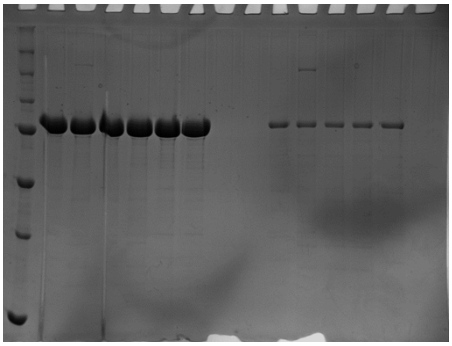
D0710 (Figure 4.8)



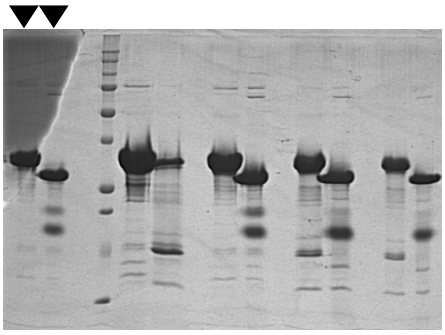
D0118 (Figure 4.8)



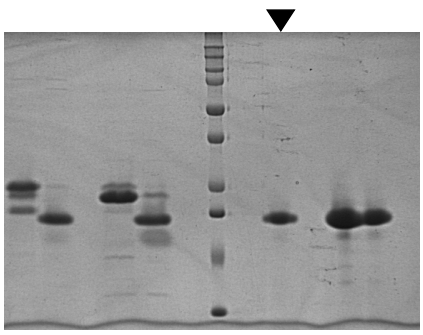
D0310_scc (Figure 4.2B and Figure 4.9F)



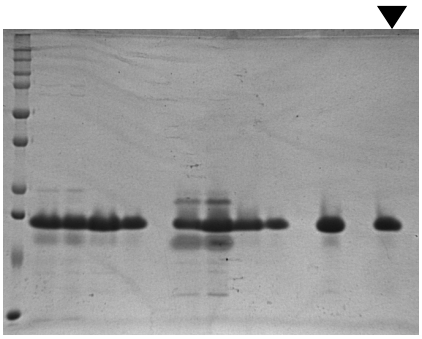
^{15}N-S0304 HRV 3C cleavage (Figure 5.4C)



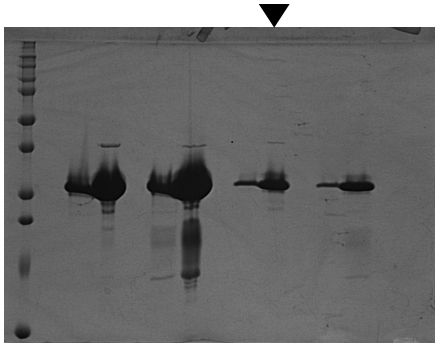
^{15}N-S03 concentration (Figure 5.5)



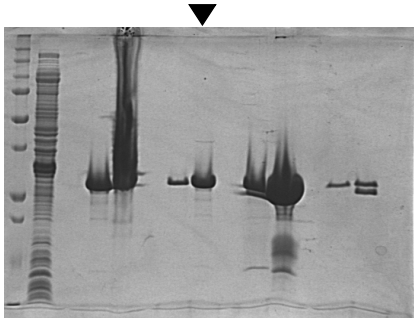
^{15}N, ^{13}C-S03 concentration (Figure 5.5)



^{15}N-S0304 concentration (Figure 5.5)



^{15}N-S0304_P704A,P706A concentration (Figure 5.5)



7.4 Non-redundant DNA-sequence of S0304_P704A,P706A

Pairwise sequence alignment by EMBOSS Water[260] of the DNA-sequence of S0304 and S0304_P704A,P706A which is manually made less redundant.

```
S0304              1 GCCCCGACCTACAAAGCCACCCATGAGTTCATGAGCGGCACACCGGGTAA     50
                     ||.|||||||||||||||.||||||||||||||||||||||||||||||||
P704A_P706A        1 GCGCCGACCTACAAAGCGACCCATGAGTTCATGAGCGGCACACCGGGTAA     50

S0304             51 AGAACTGCCTCAAGAGGTGAAGGATCTGCTGCCTGCAGATCAAACCGACC    100
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
P704A_P706A       51 AGAACTGCCTCAAGAGGTGAAGGATCTGCTGCCTGCAGATCAAACCGACC    100

S0304            101 TGAAGGATGGCAGTCAAGCCACCCCGACACAGCCGAGCAAAACAGAAGTG    150
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||.
P704A_P706A      101 TGAAGGATGGCAGTCAAGCCACCCCGACACAGCCGAGCAAAACAGAAGTT    150

S0304            151 AAGACAGCCGAGGGCACCTGGAGCTTCAAAAGCTATGACAAAACCAGCGA    200
                     ||||||||||||||.||||||||||.||||||||||||||||||||||||||
P704A_P706A      151 AAGACAGCCGAGGGTACCTGGAGCTTTAAAAGCTATGACAAAACCAGCGA    200

S0304            201 GACCATTAACGGCGCAGATGCCCACTTTGTGGGCACCTGGGAATTTACCC    250
                     ||||||||||||||||||||||||||||||||||||||||||||||||||..
P704A_P706A      201 GACCATTAACGGCGCAGATGCCCACTTTGTGGGCACCTGGGAATTTACTG    250

S0304            251 CGGCCCCGACATACAAGGCCACCCACGAGTTTGTGAGCGGTACACCGGGC    300
                     |||||.||||||||||||||||||||||||||||||||||||||||||||||
P704A_P706A      251 CGGCCGCGACATACAAGGCCACCCACGAGTTTGTGAGCGGTACACCGGGC    300
```

```
S0304           301 AAAGAACTGCCGCAGGAAGTTAAAGATCTGCTGCCGGCCGACCAGACCGA    350
                    ||||||||||||||||||||||||||.||||||||||||||||||||||||
P704A_P706A     301 AAAGAACTGCCGCAGGAAGTTAAAGACCTGCTGCCGGCCGACCAGACCGA    350

S0304           351 CCTGAAAGATGGTAGCCAGGCAACCCCGACCCAACCGAGCAAGACAGAAG    400
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
P704A_P706A     351 CCTGAAAGATGGTAGCCAGGCAACCCCGACCCAACCGAGCAAGACAGAAG    400

S0304           401 TGAAAACCACCGAAGGCACCTGGAGCTTCAAGAGTTATGATAAGACCAGC    450
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
P704A_P706A     401 TGAAAACCACCGAAGGCACCTGGAGCTTCAAGAGTTATGATAAGACCAGC    450

S0304           451 GAAACCATTAATGGCGCCGATGCCCATTTTGTTGGTACCTGGGAATTCAC    500
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
P704A_P706A     451 GAAACCATTAATGGCGCCGATGCCCATTTTGTTGGTACCTGGGAATTCAC    500

S0304           501 CCCGGCATAA    510
                    ||||||||||
P704A_P706A     501 CCCGGCATAA    510
```

Pairwise sequence alignment by EMBOSS Water[260] of the protein sequences of S0304 and S0304_P704A,P706A.

```
S0304             1 APTYKATHEFMSGTPGKELPQEVKDLLPADQTDLKDGSQATPTQPSKTEV     50
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
P704A_P706A       1 APTYKATHEFMSGTPGKELPQEVKDLLPADQTDLKDGSQATPTQPSKTEV     50

S0304            51 KTAEGTWSFKSYDKTSETINGADAHFVGTWEFTPAPTYKATHEFVSGTPG    100
                    ||||||||||||||||||||||||||||||||.|.|||||||||||||||
P704A_P706A      51 KTAEGTWSFKSYDKTSETINGADAHFVGTWEFTAAATYKATHEFVSGTPG    100

S0304           101 KELPQEVKDLLPADQTDLKDGSQATPTQPSKTEVKTTEGTWSFKSYDKTS    150
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||
P704A_P706A     101 KELPQEVKDLLPADQTDLKDGSQATPTQPSKTEVKTTEGTWSFKSYDKTS    150

S0304           151 ETINGADAHFVGTWEFTPA    169
                    ||||||||||||||||||
P704A_P706A     151 ETINGADAHFVGTWEFTPA    169
```

7.5 Assigned resonances for S03

Table 7.5: List of assigned resonances for S03. [a]Residues remaining from fusion tag.

Number	Residue	$^{1}H^{N}$ (ppm)	$^{15}N^{H}$ (ppm)	$^{13}C_{A}$ (ppm)	$^{13}C_{B}$ (ppm)
[a]618	Pro	-	-	62.8	32.30
[a]619	Ala	8.487	124.63	52.5	19.05
[a]620	Met	8.314	119.99	54.8	33.19
621	Ala	8.265	126.91	50.4	18.28
622	Pro	-	-	62.9	32.29
623	Thr	7.47	109.45	59.2	71.93
624	Tyr	8.989	119.50	57.9	42.37
625	Lys	8.694	117.07	53.8	37.25
626	Ala	8.48	119.07	50.0	21.06
627	Thr	8.608	112.56	59.0	71.96
628	His	8.55	118.32	55.7	31.83
629	Glu	8.731	121.07	54.9	34.84
630	Phe	9.329	123.07	51.3	42.28
631	Met	9.197	117.72	54.2	37.24
632	Ser	9.195	115.53	56.7	64.29

633	Gly	9.949	116.90	44.9	-
634	Thr	8.662	121.31	59.8	71.26
635	Pro	-	-	63.7	32.03
636	Gly	8.776	111.43	45.4	-
637	Lys	7.723	119.93	53.4	33.63
638	Glu	8.707	125.42	54.4	30.89
639	Leu	8.601	120.40	52.3	41.69
640	Pro	-	-	61.1	32.01
641	Gln	9.138	123.97	58.6	28.20
642	Glu	9.675	115.18	60.2	29.56
643	Val	6.557	114.10	64.4	30.87
644	Lys	7.168	119.83	60.8	31.54
645	Asp	8.548	117.81	56.2	40.58
646	Leu	7.635	118.81	54.0	42.21
647	Leu	7.367	124.44	53.8	42.59
648	Pro	-	-	62.4	32.43
649	Ala	8.115	123.61	51.3	19.54
650	Asp	8.625	120.22	55.0	40.25
651	Gln	8.983	122.16	54.8	30.08
652	Thr	8.312	111.62	60.4	71.13
653	Asp	8.825	114.55	55.2	39.45
654	Leu	8.764	118.79	54.5	41.83
655	Lys	8.406	121.50	55.5	33.56
656	Asp	9.098	123.56	56.1	40.66
657	Gly	8.767	114.79	45.3	-
658	Ser	7.951	116.21	59.0	64.55
659	Gln	8.616	119.93	55.2	29.74
660	Ala	9.175	130.03	50.4	20.56
661	Thr	8.968	116.40	58.1	70.33
662	Pro	-	-	61.2	29.94
663	Thr	11.028	124.69	62.4	68.67
664	Gln	9.09	129.25	54.2	27.68
665	Pro	-	-	62.1	31.88
666	Ser	9.335	116.59	60.7	62.86
667	Lys	7.16	117.04	55.2	36.13
668	Thr	8.449	105.91	61.0	68.87
669	Glu	7.493	121.25	55.7	33.63
670	Val	9.6	125.12	61.7	35.76
671	Lys	8.956	128.80	56.1	33.60
672	Thr	9.033	116.44	59.3	71.86
673	Ala	8.909	121.45	54.8	18.38
674	Glu	8.053	111.93	56.5	31.29
675	Gly	7.577	108.10	46.6	-
676	Thr	8.14	117.57	62.0	71.66
677	Trp	9.625	128.30	56.1	31.28
678	Ser	9.763	118.32	56.6	65.14

679	Phe	8.593	129.00	58.0	37.94
680	Lys	8.147	128.94	55.8	31.82
681	Ser	5.847	105.31	57.8	63.76
682	Tyr	8.491	116.48	59.0	40.66
683	Asp	9.61	121.36	55.7	39.50
684	Lys	7.726	118.01	54.4	35.55
685	Thr	8.685	110.17	62.7	69.09
686	Ser	7.58	110.69	57.5	64.83
687	Glu	9.115	120.97	56.4	34.26
688	Thr	8.888	124.11	62.4	69.09
689	Ile	9.023	129.98	59.5	35.75
690	Asn	9.196	130.48	50.9	38.37
691	Gly	9.415	111.12	46.4	-
692	Ala	6.262	116.61	51.0	21.45
693	Asp	8.234	119.09	54.9	41.41
694	Ala	8.452	122.96	50.2	21.30
695	His	8.27	118.91	53.9	31.22
696	Phe	8.43	125.15	56.5	42.91
697	Val	9.507	123.80	61.3	33.50
698	Gly	10.744	119.34	45.7	-
699	Thr	7.953	122.51	62.9	70.34
700	Trp	8.67	125.33	55.6	32.08
701	Glu	9.41	121.12	54.8	34.16
702	Phe	8.753	126.09	56.1	40.47
703	Thr	8.02	124.38	58.5	70.87
704	Pro	-	-	62.7	32.10
705	Ala	7.829	129.99	53.9	19.46

7.6 Putatively assigned residues for S0304

Table 7.6: List of putatively assigned resonances for S0304. [a]Residues remaining from fusion tag.

Number	Residue	$^{1}H^{N}$ (ppm)	$^{15}N^{H}$ (ppm)
[a]619	Ala	8.49	124.69
[a]620	Met	8.31	120.06
621	Ala	8.26	126.98
627	Thr	8.61	112.59
631	Met	9.19	117.79
633	Gly	10.01	117.28
641	Gln	9.25	123.90
642	Glu	9.63	115.22
653	Asp	8.83	114.58
656	Asp	9.10	123.60
673	Ala	8.92	121.52
675	Gly	7.64	108.64

676	Thr	8.11	117.54
697	Val	9.49	124.12
698	Gly	10.85	119.80
703	Thr	7.95	123.37
705	Ala	8.03	125.79
711	Thr	8.66	112.63
715	Val	8.08	116.30
717	Gly	9.94	116.95
725	Gln	9.13	124.00
726	Glu	9.68	115.25
737	Asp	8.96	114.76
740	Asp	9.13	123.54
757	Thr	9.11	113.65
759	Gly	7.56	107.96
760	Thr	8.14	117.67
781	Val	9.50	123.90
782	Gly	10.74	119.39
787	Thr	8.02	124.43
789	Ala	7.84	130.05

7.7 Putatively assigned residues for S0304_P704A,P706A

Table 7.7: List of putatively assigned resonances for S0304_P704A,P706A. [a]Residues remaining from fusion tag.

Number	Residue	$^{1}H^{N}$ (ppm)	$^{15}N^{H}$ (ppm)
[a]619	Ala	8.49	124.69
[a]620	Met	8.32	120.06
621	Ala	8.27	126.98
624	Tyr	8.99	119.59
627	Thr	8.61	112.57
631	Met	9.19	117.78
633	Gly	10.01	117.28
641	Gln	9.25	123.90
642	Glu	9.64	115.23
653	Asp	8.83	114.66
673	Ala	8.91	121.49
675	Gly	7.62	108.56
692	Ala	6.26	116.68
697	Val	9.49	124.12
698	Gly	10.85	119.81
703	Thr	7.95	123.37
704	Ala	8.04	124.11
705	Ala	8.34	125.17

706	Ala	7.91	123.83
708	Tyr	8.99	119.83
711	Thr	8.66	112.64
715	Val	8.08	116.30
717	Gly	9.95	116.96
725	Gln	9.13	123.99
726	Glu	9.68	115.25
737	Asp	8.96	114.76
757	Thr	9.11	113.65
759	Gly	7.57	107.96
776	Ala	6.23	116.59
781	Val	9.51	123.89
782	Gly	10.74	119.40
787	Thr	8.01	124.41
789	Ala	7.84	130.05

7.8 DNA and protein sequences of recombinant protein constructs used in this thesis

D17

```
atgggcagcagccatcatcatcatcacagcagcggcctggaagttctgttccagggaccaggaccagcaatg
gaacctgttattaacagaaaggcgtctgctcgagaacaattgacaacattattcaacgataaaaaacaagcaatt
gaagcgaatattcaagcaacggtagaagaaagaaatagtatattagcacagttacaaaatatttatgacactgct
attggacaaattgatcaagatcgtagcaatgcacaagttgataaaacagcatcattaaatctacaaacaatacat
gatttagat

MGSSHHHHHHSSGLEVLFQGPAMEPVINRKASAREQLTTLFNDKKQAIEANIQATVEERN
SILAQLQNIYDTAIGQIDQDRSNAQVDKTASLNLQTIHDLD
```

D1617

```
atgggcagcagccatcatcatcatcacagcagcggcctggaagttctgttccagggaccagcaatgaaacca
gcgacaacagttaaagcaacagcattacaacaaattcaaaatatcgctacaaataaaattaatttaattaaagca
aataacgaagcgacagatgaagaacaaaatattgcaatagcacaagttgaaaaagagttaattaaagctaaacaa
caaattgctagtgcagtgactaatgcagatgtggcatatttattgcatgatgagaaaaacgaaattcgtgaaatc
gaacctgttattaacagaaaggcgtctgctcgagaacaattgacaacattattcaacgataaaaaacaagcaatt
gaagcgaatattcaagcaacggtagaagaaagaaatagtatattagcacagttacaaaatatttatgacactgct
attggacaaattgatcaagatcgtagcaatgcacaagttgataaaacagcatcattaaatctacaaacaatacat
gatttagatgtacat

MGSSHHHHHHSSGLEVLFQGPAMKPATTVKATALQQIQNIATNKINLIKANNEATDEEQN
IAIAQVEKELIKAKQQIASAVTNADVAYLLHDEKNEIREIEPVINRKASAREQLTTLFND
KKQAIEANIQATVEERNSILAQLQNIYDTAIGQIDQDRSNAQVDKTASLNLQTIHDLDVH
```

D1617_T1838D

```
atgggcagcagccatcatcatcatcacagcagcggcctggaagttctgttccagggaccagcaatgaaacca
gcgacaacagttaaagcaacagcattacaacaaattcaaaatatcgctacaaataaaattaatttaattaaagca
aataacgaagcggatgatgaagaacaaaatattgcaatagcacaagttgaaaaagagttaattaaagctaaacaa
caaattgctagtgcagtgactaatgcagatgtggcatatttattgcatgatgagaaaaacgaaattcgtgaaatc
gaacctgttattaacagaaaggcgtctgctcgagaacaattgacaacattattcaacgataaaaaacaagcaatt
gaagcgaatattcaagcaacggtagaagaaagaaatagtatattagcacagttacaaaatatttatgacactgct
attggacaaattgatcaagatcgtagcaatgcacaagttgataaaacagcatcattaaatctacaaacaatacat
gatttagatgtacat
```

MGSSHHHHHHSSGLEVLFQGPAMKPATTVKATALQQIQNIATNKINLIKANNEADDEEQN
IAIAQVEKELIKAKQQIASAVTNADVAYLLHDEKNEIREIEPVINRKASAREQLTTLFND
KKQAIEANIQATVEERNSILAQLQNIYDTAIGQIDQDRSNAQVDKTASLNLQTIHDLDVH

D1617_N1943D

atgggcagcagccatcatcatcatcatcacagcagcggcctggaagttctgttccagggaccagcaatgaaacca
gcgacaacagttaaagcaacagcattacaacaaattcaaaatatcgctacaaataaaattaatttaattaaagca
aataacgaagcgacagatgaagaacaaaatattgcaatagcacaagttgaaaaagagttaattaaagctaaacaa
caaattgctagtgcagtgactaatgcagatgtggcatatttattgcatgatgagaaaaacgaaattcgtgaaatc
gaacctgttattaacagaaaggcgtctgctcgagaacaattgacaacattattcaacgataaaaaacaagcaatt
gaagcgaatattcaagcaacggtagaagaaagaaatagtatattagcacagttacaaaatatttatgacactgct
attggacaaattgatcaagatcgtagcgatgcacaagttgataaaacagcatcattaaatctacaaacaatacat
gatttagatgtacat

MGSSHHHHHHSSGLEVLFQGPAMKPATTVKATALQQIQNIATNKINLIKANNEATDEEQN
IAIAQVEKELIKAKQQIASAVTNADVAYLLHDEKNEIREIEPVINRKASAREQLTTLFND
KKQAIEANIQATVEERNSILAQLQNIYDTAIGQIDQDRSDAQVDKTASLNLQTIHDLDVH

D1617_T1838D,N1943D

atgggcagcagccatcatcatcatcatcacagcagcggcctggaagttctgttccagggaccagcaatgaaacca
gcgacaacagttaaagcaacagcattacaacaaattcaaaatatcgctacaaataaaattaatttaattaaagca
aataacgaagcggatgatgaagaacaaaatattgcaatagcacaagttgaaaaagagttaattaaagctaaacaa
caaattgctagtgcagtgactaatgcagatgtggcatatttattgcatgatgagaaaaacgaaattcgtgaaatc
gaacctgttattaacagaaaggcgtctgctcgagaacaattgacaacattattcaacgataaaaaacaagcaatt
gaagcgaatattcaagcaacggtagaagaaagaaatagtatattagcacagttacaaaatatttatgacactgct
attggacaaattgatcaagatcgtagcgatgcacaagttgataaaacagcatcattaaatctacaaacaatacat
gatttagatgtacat

MGSSHHHHHHSSGLEVLFQGPAMKPATTVKATALQQIQNIATNKINLIKANNEADDEEQN
IAIAQVEKELIKAKQQIASAVTNADVAYLLHDEKNEIREIEPVINRKASAREQLTTLFND
KKQAIEANIQATVEERNSILAQLQNIYDTAIGQIDQDRSDAQVDKTASLNLQTIHDLDVH

D1417

atgggcagcagccatcatcatcatcatcacagcagcggcctggaagttctgttccagggaccagcaatg
agacgtaaacgagctgcgcttgatagcattgaagaaaataataaaaatcaactcgatgcaatccgaaatacgttg
gatactactcaagatgaaagagatgttgctattgatactttaaataaaattgtaaatacaattaaaaatgacatt
gcacaaaacaaaacgaatgcagaagtggatcgaactgagactgatggcaacgacaacatcaaagtgattttacct
aaagttcaagttaaaccagcagcgcgtcaatctgttggtgtaaaagccgaagctcaaaatgcactaatcgatcaa
agcgatttatcaactgaagaagaaagactagctgctaaacatttagtagaacaagcacttaatcaggctattgat
cagatcaatcatgcagataagactgcccaagttaatcaagatagtataaatgctcaaaatattatttcaaaaatt
aaaccagcgacaacagttaaagcaacagcattacaacaaattcaaaatatcgctacaaataaaattaatttaatt
aaagcaaataacgaagcgacagatgaagaacaaaatattgcaatagcacaagttgaaaaagagttaattaaagct
aaacaacaaattgctagtgcagtgactaatgcagatgtggcatatttattgcatgatgagaaaaacgaaattcgt
gaaatcgaacctgttattaacagaaaggcgtctgctcgagaacaattgacaacattattcaacgataaaaaacaa
gcaattgaagcgaatattcaagcaacggtagaagaaagaaatagtatattagcacagttacaaaatatttatgac
actgctattggacaaattgatcaagatcgtagcaatgcacaagttgataaaacagcatcattaaatctacaaaca
atacatgatttagatgtacatcctattaaa

MGSSHHHHHHSSGLEVLFQGPAMRRKRAALDSIEENNKNQLDAIRNTLDTTQDERDVAID
TLNKIVNTIKNDIAQNKTNAEVDRTETDGNDNIKVILPKVQVKPAARQSVGVKAEAQNAL
IDQSDLSTEEERLAAKHLVEQALNQAIDQINHADKTAQVNQDSINAQNIISKIKPATTVK
ATALQQIQNIATNKINLIKANNEATDEEQNIAIAQVEKELIKAKQQIASAVTNADVAYLL
HDEKNEIREIEPVINRKASAREQLTTLFNDKKQAIEANIQATVEERNSILAQLQNIYDTA
IGQIDQDRSNAQVDKTASLNLQTIHDLDVHPIK

D0710

atgggcagcagccatcatcatcatcatcacagcagcggcctggaagttctgttccagggaccagcaatgacgaag
aaacaaactgctacaggtgtattaaatgatttagcaactgctaaaaagcaagaaattaatcaaaacacaaatgca
acaactgaagaaaagcaagtggctttaaatcaagtggatcaagagttagcaacggcaattaataatataaatcaa
gctgatacaaatgcggaagtagatcaagcgcaacaattaggtacaaaagcaattaatgcgattcagccaaatatt

gttaaaaaacctgcagcattagcacaaatcaatcagcattataatgctaaattagctgaaatcaatgctacacca
gatgcaacgaatgatgagaaaaatgctgcgatcaatactttaaatcaagacagacaacaagctattgaaagtatt
aaacaagctaacacaaatgcagaagtagaccaagctgcgacagtagcagagaataatatcgatgctgttcaagtt
gatgtagtaaaaaaacaagcagcgcgagataaaatcactgctgaagtggcgaagcgtattgaagcggttaaacaa
acacctaatgcaactgacgaagaaaagcaggctgctgttaatcaaatcaatcaacttaaagatcaagcaattaat
caaattaatcaaaaccaaacaaatgatcaggtagacacaactacaaatcaagcggtaaatgctatagataatgtt
gaagctgaagtagtaattaaaacaaaggcaattgcagatattgaaaaagctgttaaagaaaagcaacagcaaatt
gataatagtcttgattcaacagataatgagaaagaagttgcttcacaagcattagctaaagaaaaagaaaaagca
cttgcagctattgaccaagctcaaacgaatagtcaggtgaatcaagcagcaacaaatggtgtatcagcgattaaa
attattcaacctgaaacaaaatga

MGSSHHHHHHSSGLEVLFQGPAMTKKQTATGVLNDLATAKKQEINQNTNATTEEKQVALN
QVDQELATAINNINQADTNAEVDQAQQLGTKAINAIQPNIVKKPAALAQINQHYNAKLAE
INATPDATNDEKNAAINTLNQDRQQAIESIKQANTNAEVDQAATVAENNIDAVQVDVVKK
QAARDKITAEVAKRIEAVKQTPNATDEEKQAAVNQINQLKDQAINQINQNQTNDQVDTTT
NQAVNAIDNVEAEVVIKTKAIADIEKAVKEKQQQIDNSLDSTDNEKEVASQALAKEKEKA
LAAIDQAQTNSQVNQAATNGVSAIKIIQPETK

D0310_scc

atgggcagcagccatcatcatcatcatcacagcagcggcctggaagttctgttccagggaccagcaatgcaagta
actcataaaaaagctgcaagagatgcaattaaccaagcaacagcaacgaaaagacaacaaataaatagtaataga
gaagcaactcaggaagagaaaaatgcagcattgaacgaattaactcaagcaaccaaccatgctttagaacaaatc
aatcaagcaacaacaaatgctaatgttgataacgccaaaggagatggtctaaatgccattaatccaattgctcct
gtaactgttgttaagcaagctgcaagggatgccgtatcacatgatgcacaacaacatatcgcagagatcaatgct
aatcctgatgcgactcaagaagaaagacaagcagcaattgacaaagtgaatgctgctgtaactgcagcaaacaca
aacattttaaacgctaataccaatgctgatgttgaacaagtaaagacaaatgcgattcaaggaatacaagcaatt
acaccagctacaaaagtaaaaacagatgcaaaaaatgccatcgataaaagtgcggaaacgcaacataatacgata
tttaataataatgatgcgacgctcgaagaacaacaagcagcacaacaattacttgatcaagctgtagccacagcg
aagcaaaatattaatgcagcagatacgaatcaagaagttgcacaagcaaaagatcagggcacacaaaatatagta
gtgattcaaccggcaacacaagttaaaacggatactcgcaatgttgtaaatgataaagcgcgagaggcgataaca
aatatcaatgctacaactggcgcgactcgagaagagaaacaagaagcgataaatcgtgtcaatacacttaaaaat
agagcattaactgatattggtgtgacgtctactactgcgatggtcaatagtattagagacgatgcagtcaatcaa
atcggcgcagttcaaccgcatgtaacgaagaaacaaactgctacaggtgtattaaatgatttagcaactgctaaa
aagcaagaaattaatcaaaacacaaatgcaacaactgaagaaaagcaagtggctttaaatcaagtggatcaagag
ttagcaacggcaattaataatataaatcaagctgatacaaatgcggaagtagatcaagcgcaacaattaggtaca
aaagcaattaatgcgattcagccaaatattgttaaaaaacctgcagcattagcacaaatcaatcagcattataat
gctaaattagctgaaatcaatgctacaccagatgcaacgaatgatgagaaaaatgctgcgatcaatactttaaat
caagacagacaacaagctattgaaagtattaaacaagctaacacaaatgcagaagtagaccaagctgcgacagta
gcagagaataatatcgatgctgttcaagttgatgtagtaaaaaaacaagcagcgcgagataaaatcactgctgaa
gtggcgaagcgtattgaagcggttaaacaaacacctaatgcaactgacgaagaaaagcaggctgctgttaatcaa
atcaatcaacttaaagatcaagcaattaatcaaattaatcaaaaccaaacaaatgatcaggtagacacaactaca
aatcaagcggtaaatgctatagataatgttgaagctgaagtagtaattaaaacaaaggcaattgcagatattgaa
aaagctgttaaagaaaagcaacagcaaattgataatagtcttgattcaacagataatgagaaagaagttgcttca
caagcattagctaaagaaaaagaaaaagcacttgcagctattgaccaagctcaaacgaatagtcaggtgaatcaa
gcagcaacaaatggtgtatcagcgattaaaattattcaacctgaaacaaaatggtctcatcctcaatttgaaaaa
tgctgctaac

MGSSHHHHHHSSGLEVLFQGPAMQVTHKKAARDAINQATATKRQQINSNREATQEEKNAA
LNELTQATNHALEQINQATTNANVDNAKGDGLNAINPIAPVTVVKQAARDAVSHDAQQHI
AEINANPDATQEERQAAIDKVNAAVTAANTNILNANTNADVEQVKTNAIQGIQAITPATK
VKTDAKNAIDKSAETQHNTIFNNNDATLEEQQAAQQLLDQAVATAKQNINAADTNQEVAQ
AKDQGTQNIVVIQPATQVKTDTRNVVNDKAREAITNINATTGATREEKQEAINRVNTLKN
RALTDIGVTSTTAMVNSIRDDAVNQIGAVQPHVTKKQTATGVLNDLATAKKQEINQNTNA
TTEEKQVALNQVDQELATAINNINQADTNAEVDQAQQLGTKAINAIQPNIVKKPAALAQI
NQHYNAKLAEINATPDATNDEKNAAINTLNQDRQQAIESIKQANTNAEVDQAATVAENNI
DAVQVDVVKKQAARDKITAEVAKRIEAVKQTPNATDEEKQAAVNQINQLKDQAINQINQN
QTNDQVDTTTNQAVNAIDNVEAEVVIKTKAIADIEKAVKEKQQQIDNSLDSTDNEKEVAS
QALAKEKEKALAAIDQAQTNSQVNQAATNGVSAIKIIQPETKWSHPQFEKCC

D0118

atgggtggtggatttgctcgaagtgttgatgctgaaaatgcagttaataaaaaagttgaccaaatggaagattta
gttaatcaaaatgatgaattgacagatgaagaaaaacaagcagcaatacaagttatcgaggaacataaaaatgaa
ataattggtaatattggtgaccaaacgactgatgatggcgttactagaatcaaagatcaaggtatacagacctta
agtggggatactgcaacaccggttgttaaaccaaatgctaaaaaagcaatacgtgataaagcaacgaaacaaagg

gaaattatcaatgcaacaccagatgctactgaagacgagattcaagatgcactaaatcaattagctacggatgaa
acagatgctattgataatgttacgaatgctactacaaatgctgacgttgaaacagctaaaaataatggcatcaat
actattggagcagttgttcctcaagtaactcataaaaaagctgcaagagatgcaattaaccaagcaacagcaacg
aaaagacaacaaataaatagtaatagagaagcaactcaggaagagaaaaatgcagcattgaacgaattaactcaa
gcaaccaaccatgctttagaacaaatcaatcaagcaacaacaaatgctaatgttgataacgccaaaggagatggt
ctaaatgccattaatccaattgctcctgtaactgttgttaagcaagctgcaagggatgccgtatcacatgatgca
caacaacatatcgcagagatcaatgctaatcctgatgcgactcaagaagaaagacaagcagcaattgacaaagtg
aatgctgctgtaactgcagcaaacacaaacattttaaacgctaataccaatgctgatgttgaacaagtaaagaca
aatgcgattcaaggaatacaagcaattacaccagctacaaaagtaaaaacagatgcaaaaaatgccatcgataaa
agtgcggaaacgcaacataatacgatatttaataataatgatgcgacgctcgaagaacaacaagcagcacaacaa
ttacttgatcaagctgtagccacagcgaagcaaaatattaatgcagcagatacgaatcaagaagttgcacaagca
aaagatcagggcacacaaaatatagtagtgattcaaccggcaacacaagttaaaacggatactcgcaatgttgta
aatgataaagcgcgagaggcgataacaaatatcaatgctacaactggcgcgactcgagaagagaaacaagaagcg
ataaatcgtgtcaatacacttaaaaatagagcattaactgatattggtgtgacgtctactactgcgatggtcaat
agtattagagacgatgcagtcaatcaaatcggcgcagttcaaccgcatgtaacgaagaaacaaactgctacaggt
gtattaaatgatttagcaactgctaaaaagcaagaaattaatcaaaacacaaatgcaacaactgaagaaaagcaa
gtggctttaaatcaagtggatcaagagttagcaacggcaattaataatataaatcaagctgatacaaatgcggaa
gtagatcaagcgcaacaattaggtacaaaagcaattaatgcgattcagccaaatattgttaaaaaacctgcagca
ttagcacaaatcaatcagcattataatgctaaattagctgaaatcaatgctacaccagatgcaacgaatgatgag
aaaaatgctgcgatcaatactttaaatcaagacagacaacaagctattgaaagtattaaacaagctaacacaaat
gcagaagtagaccaagctgcgacagtagcagagaataatatcgatgctgttcaagttgatgtagtaaaaaaacaa
gcagcgcgagataaaatcactgctgaagtggcgaagcgtattgaagcggttaaacaaacacctaatgcaactgac
gaagaaaagcaggctgctgttaatcaaatcaatcaacttaaagatcaagcaattaatcaaattaatcaaaaccaa
acaaatgatcaggtagacacaactacaaatcaagcggtaaatgctatagataatgttgaagctgaagtagtaatt
aaaacaaaggcaattgcagatattgaaaaagctgttaaagaaaagcaacagcaaattgataatagtcttgattca
acagataatgagaaagaagttgcttcacaagcattagctaaagaaaaagaaaaagcacttgcagctattgaccaa
gctcaaacgaatagtcaggtgaatcaagcagcaacaaatggtgtatcagcgattaaaattattcaacctgaaaca
aaagttaaaccagctgcacgtgaaaaaatcaatcaaaaagcgaatgaattacgtgctaagattaatcaggataaa
gaagcaacagcagaagaaagacaagtagcactagataaaatcaatgaatttgtaaatcaagccatgacagatatt
acgaataatagaacaaatcaacaagttgatgatacaacgagtcaagcgcttgatagcattgctttagtgacgcct
gaccatattgttagagcagctgctagagatgcagttaagcaacaatatgaagctaaaaagcgcgaaattgagcaa
gcggaacatgcgactgatgaagaaaaacaagttgctttaaatcaattagcgaataatgaaaaacgtgcattacaa
aacatcgatcaagcaatagcgaataatgatgtgaaacgtgttgaaacaaatggcattgctacactaaaaggtgta
caacctcatattgtaattaagcctgaagcacaacaagcaataaaagcaagtgcagaaaatcaagtagaatcaata
aaagatacaccacatgcaacagttgatgaattagatgaagcgaatcaattaattagcgacacactcaaacaagcg
caacaagaaatagaaaatacaaatcaagatgctgctgttactgatgttagaaatcaaacaatcaaggcaatagag
caaataaaacctaaagtaagacgtaaacgagctgcgcttgatagcattgaagaaaataataaaaatcaactcgat
gcaatccgaaatacgttggatactactcaagatgaaagagatgttgctattgatactttaaataaaattgtaaat
acaattaaaaatgacattgcacaaaacaaaacgaatgcagaagtggatcgaactgagactgatggcaacgacaac
atcaaagtgattttacctaaagttcaagttaaaccagcagcgcgtcaatctgttggtgtaaaagccgaagctcaa
aatgcactaatcgatcaaagcgatttatcaactgaagaagaaagactagctgctaaacatttagtagaacaagca
cttaatcaggctattgatcagatcaatcatgcagataagactgcccaagttaatcaagatagtataaatgctcaa
aatattatttcaaaaattaaaccagcgacaacagttaaagcaacagcattacaacaaattcaaaatatcgctaca
aataaaattaatttaattaaagcaaataacgaagcgacagatgaagaacaaaatattgcaatagcacaagttgaa
aaagagttaattaaagctaaacaacaaattgctagtgcagtgactaatgcagatgtggcatatttattgcatgat
gagaaaaacgaaattcgtgaaatcgaacctgttattaacagaaaggcgtctgctcgagaacaattgacaacatta
ttcaacgataaaaaacaagcaattgaagcgaatattcaagcaacggtagaagaaagaaatagtatattagcacag
ttacaaaatatttatgacactgctattggacaaattgatcaagatcgtagcaatgcacaagttgataaaacagca
tcattaaatctacaaacaatacatgatttagatgtacatcctattaaaaagccagatgctgaaaaaacgattaat
gatgatcttgcacgcgtcactgctttagtgcaaaattatcgaaaagtaagtaatcgtaataaggctgatgcatta
aaagctataactgctttaaaattacaaatggatgaagaattaaaaacagcacgcactaatgctgatgttgatgca
gttttaaaacgatttaatgttgcattaagcgatatagaagcagaaaatttatacttccaaggtcatcatcaccat
catcaccatcaccatcat

MGGGFARSVDAENAVNKKVDQMEDLVNQNDELTDEEKQAAIQVIEEHKNEIIGNIGDQTT
DDGVTRIKDQGIQTLSGDTATPVVKPNAKKAIRDKATKQREIINATPDATEDEIQDALNQ
LATDETDAIDNVTNATTNADVETAKNNGINTIGAVVPQVTHKKAARDAINQATATKRQQI
NSNREATQEEKNAALNELTQATNHALEQINQATTNANVDNAKGDGLNAINPIAPVTVVKQ
AARDAVSHDAQQHIAEINANPDATQEERQAAIDKVNAAVTAANTNILNANTNADVEQVKT
NAIQGIQAITPATKVKTDAKNAIDKSAETQHNTIFNNNDATLEEQQAAQQLLDQAVATAK
QNINAADTNQEVAQAKDQGTQNIVVIQPATQVKTDTRNVVNDKAREAITNINATTGATRE
EKQEAINRVNTLKNRALTDIGVTSTTAMVNSIRDDAVNQIGAVQPHVTKKQTATGVLNDL
ATAKKQEINQNTNATTEEKQVALNQVDQELATAINNINQADTNAEVDQAQQLGTKAINAI
QPNIVKKPAALAQINQHYNAKLAEINATPDATNDEKNAAINTLNQDRQQAIESIKQANTN
AEVDQAATVAENNIDAVQVDVVKKQAARDKITAEVAKRIEAVKQTPNATDEEKQAAVNQI
NQLKDQAINQINQNQTNDQVDTTTNQAVNAIDNVEAEVVIKTKAIADIEKAVKEKQQQID
NSLDSTDNEKEVASQALAKEKEKALAAIDQAQTNSQVNQAATNGVSAIKIIQPETKVKPA
AREKINQKANELRAKINQDKEATAEERQVALDKINEFVNQAMTDITNNRTNQQVDDTTSQ

ALDSIALVTPDHIVRAAARDAVKQQYEAKKREIEQAEHATDEEKQVALNQLANNEKRALQ
NIDQAIANNDVKRVETNGIATLKGVQPHIVIKPEAQQAIKASAENQVESIKDTPHATVDE
LDEANQLISDTLKQAQQEIENTNQDAAVTDVRNQTIKAIEQIKPKVRRKRAALDSIEENN
KNQLDAIRNTLDTTQDERDVAIDTLNKIVNTIKNDIAQNKTNAEVDRTETDGNDNIKVIL
PKVQVKPAARQSVGVKAEAQNALIDQSDLSTEEERLAAKHLVEQALNQAIDQINHADKTA
QVNQDSINAQNIISKIKPATTVKATALQQIQNIATNKINLIKANNEATDEEQNIAIAQVE
KELIKAKQQIASAVTNADVAYLLHDEKNEIREIEPVINRKASAREQLTTLFNDKKQAIEA
NIQATVEERNSILAQLQNIYDTAIGQIDQDRSNAQVDKTASLNLQTIHDLDVHPIKKPDA
EKTINDDLARVTALVQNYRKVSNRNKADALKAITALKLQMDEELKTARTNADVDAVLKRF
NVALSDIEAENLYFQGHHHHHHHHHH

D0118_2Cys

cgaagtgttgatgctgaaaatgcagttaataaaaaagttgaccaaatggaagatttagttaatcaaaatgatgaa
ttgacagatgaagaaaaacaagcagcaatacaagttatcgaggaacataaaaatgaaataattggtaatattggt
gaccaaacgactgatgatggcgttactagaatcaaagatcaaggtatacagaccttaagtggggatactgcatgt
ccggttgttaaaccaaatgctaaaaaagcaatacgtgataaagcaacgaaacaaagggaaattatcaatgcaaca
ccagatgctactgaagacgagattcaagatgcactaaatcaattagctacggatgaaacagatgctattgataat
gttacgaatgctactacaaatgctgacgttgaaacagctaaaaataatggcatcaatactattggagcagttgtt
cctcaagtaactcataaaaaagctgcaagagatgcaattaaccaagcaacagcaacgaaaagacaacaaataaat
agtaatagagaagcaactcaggaagagaaaaatgcagcattgaacgaattaactcaagcaaccaaccatgcttta
gaacaaatcaatcaagcaacaacaaatgctaatgttgataacgccaaaggagatggtctaaatgccattaatcca
attgctcctgtaactgttgttaagcaagctgcaagggatgccgtatcacatgatgcacaacaacatatcgcagag
atcaatgctaatcctgatgcgactcaagaagaaagacaagcagcaattgacaaagtgaatgctgctgtaactgca
gcaaacacaaacattttaaacgctaataccaatgctgatgttgaacaagtaaagacaaatgcgattcaaggaata
caagcaattacaccagctacaaaagtaaaaacagatgcaaaaaatgccatcgataaaagtgcggaaacgcaacat
aatacgatatttaataataatgatgcgacgctcgaagaacaacaagcagcacaacaattacttgatcaagctgta
gccacagcgaagcaaaatattaatgcagcagatacgaatcaagaagttgcacaagcaaaagatcagggcacacaa
aatatagtagtgattcaaccggcaacacaagttaaaacggatactcgcaatgttgtaaatgataaagcgcgagag
gcgataacaaatatcaatgctacaactggcgcgactcgagaagagaaacaagaagcgataaatcgtgtcaataca
cttaaaaatagagcattaactgatattggtgtgacgtctactactgcgatggtcaatagtattagagacgatgca
gtcaatcaaatcggcgcagttcaaccgcatgtaacgaagaaacaaactgctacaggtgtattaaatgatttagca
actgctaaaaagcaagaaattaatcaaaacacaaatgcaacaactgaagaaaagcaagtggctttaaatcaagtg
gatcaagagttagcaacggcaattaataatataaatcaagctgatacaaatgcggaagtagatcaagcgcaacaa
ttaggtacaaaagcaattaatgcgattcagccaaatattgttaaaaaacctgcagcattagcacaaatcaatcag
cattataatgctaaattagctgaaatcaatgctacaccagatgcaacgaatgatgagaaaaatgctgcgatcaat
actttaaatcaagacagacaacaagctattgaaagtattaaacaagctaacacaaatgcagaagtagaccaagct
gcgacagtagcagagaataatatcgatgctgttcaagttgatgtagtaaaaaaacaagcagcgcgagataaaatc
actgctgaagtggcgaagcgtattgaagcggttaaacaaacacctaatgcaactgacgaagaaaagcaggctgct
gttaatcaaatcaatcaacttaaagatcaagcaattaatcaaattaatcaaaaccaaacaaatgatcaggtagac
acaactacaaatcaagcggtaaatgctatagataatgttgaagctgaagtagtaattaaaacaaaggcaattgca
gatattgaaaaagctgttaaagaaaagcaacagcaaattgataatagtcttgattcaacagataatgagaaagaa
gttgcttcacaagcattagctaaagaaaaagaaaaagcacttgcagctattgaccaagctcaaacgaatagtcag
gtgaatcaagcagcaacaaatggtgtatcagcgattaaaattattcaacctgaaacaaaagttaaaccagctgca
cgtgaaaaaatcaatcaaaaagcgaatgaattacgtgctaagattaatcaggataaagaagcaacagcagaagaa
agacaagtagcactagataaaatcaatgaatttgtaaatcaagccatgacagatattacgaataatagaacaaat
caacaagttgatgatacaacaagtcaagcgcttgatagcattgctttagtgacgcctgaccatattgttagagca
gctgctagagatgcagttaagcaacaatatgaagctaaaaagcgcgaaattgagcaagcggaacatgcgactgat
gaagaaaaacaagttgctttaaatcaattagcgaataatgaaaaacgtgcattacaaaacatcgatcaagcaata
gcgaataatgatgtgaaacgtgttgaaacaaatggcattgctacactaaaaggtgtacaacctcatattgtaatt
aagcctgaagcacaacaagcaataaaagcaagtgcagaaaatcaagtagaatcaataaaagatacaccacatgca
acagttgatgaattagatgaagcgaatcaattaattagcgacacactcaaacaagcgcaacaagaaatagaaaat
acaaatcaagatgctgctgttactgatgttagaaatcaaacaatcaaggcaatagagcaaataaaacctaaagta
agacgtaaacgagctgcgcttgatagcattgaagaaaataataaaaatcaactcgatgcaatccgaaatacgttg
gatactactcaagatgaaagagatgttgctattgatactttaaataaaattgtaaatacaattaaaaatgacatt
gcacaaaacaaaacgaatgcagaagtggatcgaactgagactgatggcaacgacaacatcaaagtgattttacct
aaagttcaagttaaaccagcagcgcgtcaatctgttggtgtaaaagccgaagctcaaaatgcactaatcgatcaa
agcgatttatcaactgaagaagaaagactagctgctaaacatttagtagaacaagcacttaatcaggctattgat
cagatcaatcatgcagataagactgcccaagttaatcaagatagtataaatgctcaaaatattatttcaaaaatt
aaaccagcgacaacagttaaagcaacagcattacaacaaattcaaaatatcgctacaaataaaattaatttaatt
aaagcaaataacgaagcgacagatgaagaacaaaatattgcaatagcacaagttgaaaaagagttaattaaagct
aaacaacaaattgctagtgcagtgactaatgcagatgtggcatatttattgcatgatgagaaaaacgaaattcgt
gaaatcgaacctgttattaacagaaaggcgtctgctcgagaacaattgacaacattattcaacgataaaaaacaa
gcaattgaagcgaatattcaagcaacggtagaagaaagaaatagtatattagcacagttacaaaatatttatgac
actgctattggacaaattgatcaagatcgtagcaatgcacaagttgataaaacagcatcattaaatctacaaaca
atacatgatttagatgtatgtcctattaaaaagccagatgctgaaaaaacgattaatgatgatcttgcacgcgtc
actgctttagtgcaaaattatcgaaaagtaagtaatcgtaataaggctgatgcattaaaagctataactgcttta

aaattacaaatggatgaagaattaaaaacagcacgcactaatgctgatgttgatgcagttttaaaacgatttaat
gttgcattaagcgatatagaagca

MGGGFARSVDAENAVNKKVDQMEDLVNQNDELTDEEKQAAIQVIEEHKNEIIGNIGDQTT
DDGVTRIKDQGIQTLSGDTACPVVKPNAKKAIRDKATKQREIINATPDATEDEIQDALNQ
LATDETDAIDNVTNATTNADVETAKNNGINTIGAVVPQVTHKKAARDAINQATATKRQQI
NSNREATQEEKNAALNELTQATNHALEQINQATTNANVDNAKGDGLNAINPIAPVTVVKQ
AARDAVSHDAQQHIAEINANPDATQEERQAAIDKVNAAVTAANTNILNANTNADVEQVKT
NAIQGIQAITPATKVKTDAKNAIDKSAETQHNTIFNNNDATLEEQQAAQQLLDQAVATAK
QNINAADTNQEVAQAKDQGTQNIVVIQPATQVKTDTRNVVNDKAREAITNINATTGATRE
EKQEAINRVNTLKNRALTDIGVTSTTAMVNSIRDDAVNQIGAVQPHVTKKQTATGVLNDL
ATAKKQEINQNTNATTEEKQVALNQVDQELATAINNINQADTNAEVDQAQQLGTKAINAI
QPNIVKKPAALAQINQHYNAKLAEINATPDATNDEKNAAINTLNQDRQQAIESIKQANTN
AEVDQAATVAENNIDAVQVDVVKKQAARDKITAEVAKRIEAVKQTPNATDEEKQAAVNQI
NQLKDQAINQINQNQTNDQVDTTTNQAVNAIDNVEAEVVIKTKAIADIEKAVKEKQQQID
NSLDSTDNEKEVASQALAKEKEKALAAIDQAQTNSQVNQAATNGVSAIKIIQPETKVKPA
AREKINQKANELRAKINQDKEATAEERQVALDKINEFVNQAMTDITNNRTNQQVDDTTSQ
ALDSIALVTPDHIVRAAARDAVKQQYEAKKREIEQAEHATDEEKQVALNQLANNEKRALQ
NIDQAIANNDVKRVETNGIATLKGVQPHIVIKPEAQQAIKASAENQVESIKDTPHATVDE
LDEANQLISDTLKQAQQEIENTNQDAAVTDVRNQTIKAIEQIKPKVRRKRAALDSIEENN
KNQLDAIRNTLDTTQDERDVAIDTLNKIVNTIKNDIAQNKTNAEVDRTETDGNDNIKVIL
PKVQVKPAARQSVGVKAEAQNALIDQSDLSTEEERLAAKHLVEQALNQAIDQINHADKTA
QVNQDSINAQNIISKIKPATTVKATALQQIQNIATNKINLIKANNEATDEEQNIAIAQVE
KELIKAKQQIASAVTNADVAYLLHDEKNEIREIEPVINRKASAREQLTTLFNDKKQAIEA
NIQATVEERNSILAQLQNIYDTAIGQIDQDRSNAQVDKTASLNLQTIHDLDVCPIKKPDA
EKTINDDLARVTALVQNYRKVSNRNKADALKAITALKLQMDEELKTARTNADVDAVLKRF
NVALSDIEAENLYFQGHHHHHHHHHH

S03

atgggcagcagccatcatcatcatcatcacagcagcggcctggaagttctgttccagggaccagcaatggccccg
acctacaaagccacccatgagttcatgagcggcacaccgggtaaagaactgcctcaagaggtgaaggatctgctg
cctgcagatcaaaccgacctgaaggatggcagtcaagccaccccgacacagccgagcaaaacagaagtgaagaca
gccgagggcacctggagcttcaaaagctatgacaaaaccagcgagaccattaacggcgcagatgcccactttgtg
ggcacctgggaatttaccccggcctaa

MGSSHHHHHHSSGLEVLFQGPAMAPTYKATHEFMSGTPGKELPQEVKDLLPADQTDLKDG
SQATPTQPSKTEVKTAEGTWSFKSYDKTSETINGADAHFVGTWEFTPA

S0304

atgggcagcagccatcatcatcatcatcacagcagcggcctggaagttctgttccagggaccagcaatggccccg
acctacaaagccacccatgagttcatgagcggcacaccgggtaaagaactgcctcaagaggtgaaggatctgctg
cctgcagatcaaaccgacctgaaggatggcagtcaagccaccccgacacagccgagcaaaacagaagtgaagaca
gccgagggcacctggagcttcaaaagctatgacaaaaccagcgagaccattaacggcgcagatgcccactttgtg
ggcacctgggaatttaccccggccccgacatacaaggccacccacgagtttgtgagcggtacaccgggcaaagaa
ctgccgcaggaagttaaagatctgctgccggccgaccagaccgacctgaaagatggtagccaggcaaccccgacc
caaccgagcaagacagaagtgaaaaccaccgaaggcacctggagcttcaagagttatgataagaccagcgaaacc
attaatggcgccgatgcccattttgttggtacctgggaattcaccccggcataa

MGSSHHHHHHSSGLEVLFQGPAMAPTYKATHEFMSGTPGKELPQEVKDLLPADQTDLKDG
SQATPTQPSKTEVKTAEGTWSFKSYDKTSETINGADAHFVGTWEFTPAPTYKATHEFVSG
TPGKELPQEVKDLLPADQTDLKDGSQATPTQPSKTEVKTTEGTWSFKSYDKTSETINGAD
AHFVGTWEFTPA

S0304_P704A,P706A

atgggcagcagccatcatcatcatcatcacagcagcggcctggaagttctgttccagggaccagcaatggcgccg
acctacaaagcgacccatgagttcatgagcggcacaccgggtaaagaactgcctcaagaggtgaaggatctgctg
cctgcagatcaaaccgacctgaaggatggcagtcaagccaccccgacacagccgagcaaaacagaagttaagaca
gccgagggtacctggagctttaaaagctatgacaaaaccagcgagaccattaacggcgcagatgcccactttgtg
ggcacctgggaatttactgcggccgcgacatacaaggccacccacgagtttgtgagcggtacaccgggcaaagaa
ctgccgcaggaagttaaagacctgctgccggccgaccagaccgacctgaaagatggtagccaggcaaccccgacc
caaccgagcaagacagaagtgaaaaccaccgaaggcacctggagcttcaagagttatgataagaccagcgaaacc
attaatggcgccgatgcccattttgttggtacctgggaattcaccccggcataa

MGSSHHHHHHSSGLEVLFQGPAMAPTYKATHEFMSGTPGKELPQEVKDLLPADQTDLKDG
SQATPTQPSKTEVKTAEGTWSFKSYDKTSETINGADAHFVGTWEFTAAATYKATHEFVSG
TPGKELPQEVKDLLPADQTDLKDGSQATPTQPSKTEVKTTEGTWSFKSYDKTSETINGAD
AHFVGTWEFTPA

Chapter 8. Bibliography

1. Mizuno, S. *et al.* Comparison of national strategies to reduce meticillin-resistant Staphylococcus aureus infections in Japan and England. *J. Hosp. Infect.* **100**, 280–298 (2018).

2. Grayson, M. L. *et al.* Effects of the Australian National Hand Hygiene Initiative after 8 years on infection control practices, health-care worker education, and clinical outcomes: a longitudinal study. *Lancet Infect. Dis.* **18**, 1269–1277 (2018).

3. McNeil, J. C. & Fritz, S. A. Prevention Strategies for Recurrent Community-Associated Staphylococcus aureus Skin and Soft Tissue Infections. *Curr. Infect. Dis. Rep.* **21**, (2019).

4. Tande, A. J. & Patel, R. Prosthetic joint infection. *Clin. Microbiol. Rev.* **27**, 302–345 (2014).

5. Saeed, K. *et al.* An update on Staphylococcus aureus infective endocarditis from the International Society of Antimicrobial Chemotherapy (ISAC). *Int. J. Antimicrob. Agents* **53**, 9–15 (2019).

6. Hoerr, V. *et al.* S. aureus endocarditis: Clinical aspects and experimental approaches. *Int. J. Med. Microbiol.* (2018). doi:10.1016/J.IJMM.2018.02.004

7. Abdallah, L. *et al.* Long-term prognosis of left-sided native-valve Staphylococcus aureus endocarditis. *Arch. Cardiovasc. Dis.* **109**, 260–267 (2016).

8. Hall-Stoodley, L., Costerton, J. W. & Stoodley, P. Bacterial biofilms: from the natural environment to infectious diseases. *Nat. Rev. Microbiol.* **2**, 95–108 (2004).

9. WILLIAMS, R. E. Healthy carriage of Staphylococcus aureus: its prevalence and importance. *Bacteriol. Rev.* **27**, 56–71 (1963).

10. Sakr, A., Brégeon, F., Mège, J. L., Rolain, J. M. & Blin, O. Staphylococcus aureus

nasal colonization: An update on mechanisms, epidemiology, risk factors, and subsequent infections. *Front. Microbiol.* **9**, 1–15 (2018).

11. Brown, A. F., Leech, J. M., Rogers, T. R. & McLoughlin, R. M. Staphylococcus aureus colonization: Modulation of host immune response and impact on human vaccine design. *Front. Immunol.* **4**, 1–20 (2014).

12. Klein, E. Y. *et al.* National Costs Associated With Methicillin-Susceptible and Methicillin-Resistant Staphylococcus aureus Hospitalizations in the United States, 2010-2014. *Clin. Infect. Dis.* **68**, 22–28 (2019).

13. Hernández, A. *et al.* Evolution of the Incidence, Mortality, and Cost of Infective Endocarditis in Spain Between 1997 and 2014. *J. Gen. Intern. Med.* **33**, 1610–1613 (2018).

14. Fritschi, B. Z., Albert-Kiszely, A. & Persson, G. R. Staphylococcus aureus and Other Bacteria in Untreated Periodontitis. *J. Dent. Res.* **87**, 589–593 (2008).

15. Doern, C. D. & Burnham, C. A. D. It's not easy being green: The viridans group streptococci, with a focus on pediatric clinical manifestations. *J. Clin. Microbiol.* **48**, 3829–3835 (2010).

16. Aas, J. a, Paster, B. J., Stokes, L. N., Olsen, I. & Dewhirst, F. E. Defining the Normal Bacterial Flora of the Oral Cavity Defining the Normal Bacterial Flora of the Oral Cavity. *J. Clin. Microbiol.* **43**, 5721–5732 (2005).

17. Loo, C. Y., Corliss, D. a & Ganeshkumar, N. Streptococcus gordonii Biofilm Formation : Identification of Genes that Code for Biofilm Phenotypes Streptococcus gordonii Biofilm Formation : Identification of Genes that Code for Biofilm Phenotypes. *J. Bacteriol.* **182**, 1374–1382 (2000).

18. Marsh, P. D., Head, D. A. & Devine, D. A. Dental plaque as a biofilm and a microbial community - Implications for treatment. *J. Oral Biosci.* **57**, 185–191 (2015).

19. Jakubovics, N. S. Intermicrobial Interactions as a Driver for Community Composition and Stratification of Oral Biofilms. *J. Mol. Biol.* **427**, 3662–3675 (2015).

20. Cook, G. S., Costerton, J. W. & Lamont, R. J. Biofilm formation by Porphyromonas gingivalis and Streptococcus gordonii. *J. Periodontal Res.* **33**, 323–327 (1998).

21. Nobbs, A. H., Jenkinson, H. F. & Jakubovics, N. S. Stick to your gums: Mechanisms of oral microbial adherence. *J. Dent. Res.* **90**, 1271–1278 (2011).

22. Douglas, C. W. I., Heath, J., Hampton, K. K. & Preston, F. E. Identity of viridans streptococci isolated from cases of infective endocarditis. *J. Med. Microbiol.* **39**, 179–182 (1993).

23. Wells, V. D., Munro, C. L., Sulavik, M. C., Clewell, D. B. & Macrina, F. L. Infectivity of a glucan synthesis-defective mutant of Streptococcus gordonii (Challis) in a rat endocarditis model. *Microbiol. Immunol.* **112**, 301–305 (1993).

24. Slipczuk, L. *et al.* Infective endocarditis epidemiology over five decades: A systematic review. *PLoS One* **8**, (2013).

25. Schneewind, O., Fowler, A. & Faull, K. F. Structure of the cell wall anchor of surface proteins in Staphylococcus aureus. *Science (80-.).* **268**, 103–106 (1995).

26. Marraffini, L. A. & Schneewind, O. Anchor Structure of Staphylococcal Surface Proteins. *J. Biol. Chem.* **280**, 16263–16271 (2005).

27. Hendrickx, A. P. A., Van Wamel, W. J. B., Posthuma, G., Bonten, M. J. M. & Willems, R. J. L. Five genes encoding surface-exposed LPXTG proteins are enriched in hospital-adapted Enterococcus faecium clonal complex 17 isolates. *J. Bacteriol.* **189**, 8321–8332 (2007).

28. Chhatwal, G. S. Anchorless adhesins and invasins of Gram-positive bacteria: a new class of virulence factors. *Trends Microbiol.* **10**, 205–208 (2002).

29. Foster, T. J., Geoghegan, J. A., Ganesh, V. K. & Höök, M. Adhesion, invasion and evasion: the many functions of the surface proteins of Staphylococcus aureus. *Nat.Rev.Microbiol.* **12**, 49–62 (2014).

30. Speziale, P. *et al.* Protein-based biofilm matrices in Staphylococci. *Front. Cell. Infect. Microbiol.* **4:171**, 1–10 (2014).

31. Jan-Roblero, J., García-Gómez, E., Rodríguez-Martínez, S., Cancino-Diaz, M. E. & Cancino-Diaz, J. *The Rise of Virulence and Antibiotic Resistance in S. aureus. Chapter 10. Surface proteins of Staphylococcus aureus*. (2017).

32. Bowden, C. F. M. *et al.* Structure-function analyses reveal key features in Staphylococcus aureus IsdB-associated unfolding of the heme-binding pocket of human hemoglobin. *J. Biol. Chem.* **293**, 177–190 (2018).

33. Gaudin, C. F. M., Grigg, J. C., Arrieta, A. L. & Murphy, M. E. P. Unique heme-iron coordination by the hemoglobin receptor IsdB of staphylococcus aureus. *Biochemistry* **50**, 5443–5452 (2011).

34. Hammer, N. D. & Skaar, E. P. Molecular Mechanisms of Staphylococcus aureus Iron Acquisition . *Annu. Rev. Microbiol.* **65**, 129–147 (2010).

35. Bowden, C. F. M., Verstraete, M. M., Eltis, L. D. & Murphy, M. E. P. Hemoglobin binding and catalytic heme extraction by IsdB near iron transporter domains. *Biochemistry* **53**, 2286–2294 (2014).

36. Grigg, J. C., Ukpabi, G., Gaudin, C. F. M. & Murphy, M. E. P. Structural biology of heme binding in the Staphylococcus aureus Isd system. *J. Inorg. Biochem.* **104**, 341–348 (2010).

37. Zhang, Y. *et al. Staphylococcus aureus* SdrE captures complement factor H's C-terminus via a novel 'close, dock, lock and latch'' mechanism for complement evasion'. *Biochem. J.* **474**, 1619–1631 (2017).

38. Otto, M. Staphylococcal biofilms. *Curr. Top. Microbiol. Immunol.* **322**, 207–28

(2008).

39. Foster, T. J. Immune evasion by staphylococci. *Nat. Rev. Microbiol.* **3**, 948–958 (2005).

40. Li, M. *et al.* MRSA epidemic linked to a quickly spreading colonization and virulence determinant. *Nat. Med.* **18**, 816–819 (2012).

41. Jansson, B., Uhlén, M. & Nygren, P. Å. All individual domains of staphylococcal protein A show Fab binding. *FEMS Immunol. Med. Microbiol.* **20**, 69–78 (1998).

42. Sjoquist, J., Movits, J., Johansson, I.-B. & Hjelm, H. Localization of Protein A in the Bacteria. *Eur. J. Biochem.* **30**, 190–194 (1972).

43. Sjödahl, J. Repetitive Sequences in Protein A from Staphylococcus aureus. *Eur J Biochem* **73**, 343–351 (1977).

44. Graille, M. *et al.* Crystal structure of a Staphylococcus aureus protein A domain complexed with the Fab fragment of a human IgM antibody: structural basis for recognition of B-cell receptors and superantigen activity. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5399–404 (2000).

45. Capp, J. A., Hagarman, A., Richardson, D. C. & Oas, T. G. The statistical conformation of a highly flexible protein: Small-angle x-ray scattering of S. aureus protein A. *Structure* **22**, 1184–1195 (2014).

46. Stemberk, V. *et al.* Evidence for steric regulation of fibrinogen binding to staphylococcus aureus fibronectin-binding protein a(FnBPA). *J. Biol. Chem.* **289**, 12842–12851 (2014).

47. Bingham, R. J. *et al.* Crystal structures of fibronectin-binding sites from Staphylococcus aureus FnBPA in complex with fibronectin domains. *Proc. Natl. Acad. Sci.* **105**, 12254–12258 (2008).

48. Sibbald, M. K. *et al.* Synthetic effects of secG and secY2 mutations on

exoproteome biogenesis in Staphylococcus aureus. *J. Bacterio* **192**, 3788–3800 (2010).

49. Hartleib, J. *et al.* Protein A is the von Willebrand factor binding protein on Staphylococcus aureus Protein A is the von Willebrand factor binding protein on Staphylococcus aureus. *Blood* **96**, 2149–2156 (2000).

50. Askarian, F. *et al.* The interaction between Staphylococcus aureus SdrD and desmoglein 1 is important for adhesion to host cells. *Sci. Rep.* **6**, 1–11 (2016).

51. Formosa-Dague, C. *et al.* Molecular interactions and inhibition of the staphylococcal biofilm-forming protein SdrC. *Proc. Natl. Acad. Sci.* **114**, 3738–3743 (2017).

52. Zong, Y. *et al.* A 'Collagen Hug' model for Staphylococcus aureus CNA binding to collagen. *EMBO J.* **24**, 4224–36 (2005).

53. Phang, J. M., Liu, W., Hancock, C. N. & Fischer, J. W. Proline metabolism and cancer : emerging links to glutamine and collagen. (2015). doi:10.1097/MCO.0000000000000121

54. S, M. L. & Rodr, L. G. Collagen : A review on its sources and potential cosmetic applications. 20–26 (2018). doi:10.1111/jocd.12450

55. Luo, M. *et al.* Crystal Structure of an Invasivity-Associated Domain of SdrE in S. aureus. *PLoS One* **12**, e0168814 (2017).

56. Deis, L. N. *et al.* Multiscale conformational heterogeneity in staphylococcal protein a: Possible determinant of functional plasticity. *Structure* **22**, 1467–1477 (2014).

57. Gruszka, D. T. *et al.* Cooperative folding of intrinsically disordered domains drives assembly of a strong elongated protein. *Nat. Commun.* **6**, 7271 (2015).

58. Tanaka, Y. *et al.* A Helical String of Alternately Connected Three-Helix Bundles

for the Cell Wall-Associated Adhesion Protein Ebh from Staphylococcus aureus. *Structure* **16**, 488–496 (2008).

59. Gruszka, D. T. *et al.* Staphylococcal biofilm-forming protein has a contiguous rod-like structure. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E1011-8 (2012).

60. Rohde, H. *et al.* Polysaccharide intercellular adhesin or protein factors in biofilm accumulation of Staphylococcus epidermidis and Staphylococcus aureus isolated from prosthetic hip and knee joint infections. *Biomaterials* **28**, 1711–1720 (2007).

61. Formosa-Dague, C., Speziale, P., Foster, T. J., Geoghegan, J. A. & Dufrêne, Y. F. Zinc-dependent mechanical properties of Staphylococcus aureus biofilm-forming surface protein SasG. *Proc. Natl. Acad. Sci.* **113**, 410–415 (2016).

62. Benigar, E. *et al.* Evaluating SAXS results on aqueous solutions of various bacterial Levan utilizing the string-of-beads model. *Acta Chim. Slov.* **62**, 509–517 (2015).

63. McDevitt, D., Francois, P., Vaudaux, P. & Foster, T. J. Molecular characterization of the clumping factor (fibrinogen receptor) of Staphylococcus aureus. *Mol. Microbiol.* **11**, 237–248 (1994).

64. Josefsson, E. *et al.* Three new members of the serine-aspartate repeat protein multigene family of Staphylococcus aureus. *Microbiology* **144**, 3387–3395 (1998).

65. Josefsson, E., O'Connell, D., Foster, T. J., Durussel, I. & Cox, J. A. The binding of calcium to the B-repeat segment of SdrD, a cell surface protein of Staphylococcus aureus. *J. Biol. Chem.* **273**, 31145–31152 (1998).

66. Wang, X., Ge, J., Liu, B., Hu, Y. & Yang, M. Structures of SdrD from Staphylococcus aureus reveal the molecular mechanism of how the cell surface receptors recognize their ligands. *Protein Cell* **4**, 277–285 (2013).

67. Schneewind, O. & Missiakas, D. M. Protein secretion and surface display in Gram-positive bacteria. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 1123–1139 (2012).

68. Tsirigotaki, A., De Geyter, J., Šoštarić, N., Economou, A. & Karamanou, S. Protein export through the bacterial Sec pathway. *Nat. Rev. Microbiol.* **15**, 21–36 (2017).

69. Dalbey, R. E., Wang, P. & van Dijl, J. M. Membrane Proteases in the Bacterial Protein Secretion and Quality Control Pathway. *Microbiol. Mol. Biol. Rev.* **76**, 311–330 (2012).

70. Yu, W., Missiakas, D. & Schneewind, O. Septal secretion of protein A in Staphylococcus aureus requires SecA and lipoteichoic acid synthesis. *Elife* **7**, 1–27 (2018).

71. Van Roosmalen, M. L. *et al.* Type I signal peptidases of Gram-positive bacteria. *Biochim. Biophys. Acta - Mol. Cell Res.* **1694**, 279–297 (2004).

72. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

73. Edgar, R. C. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 1–19 (2004).

74. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).

75. Rosch, J. & Caparon, M. A Microdomain for Protein Secretion in Gram-Positive Bacteria. *Science (80-.).* **304**, 1513–1515 (2004).

76. DeDent, A., Bae, T., Missiakas, D. M. & Schneewind, O. Signal peptides direct surface proteins to two distinct envelope locations of Staphylococcus aureus. *EMBO J.* **27**, 2656–68 (2008).

77. Freudl, R. Signal peptides for recombinant protein secretion in bacterial expression systems. *Microb. Cell Fact.* **17**, 1–10 (2018).

78. Akopian, D., Shen, K., Zhang, X. & Shan, S. *Signal Recognition Particle: An Essential Protein-Targeting Machine*. *Annual Review of Biochemistry* **82**, (2013).

79. Pechmann, S., Chartron, J. W. & Frydman, J. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. *Nat. Struct. Mol. Biol.* **21**, 1100–1105 (2014).

80. Schneewind, O. & Missiakas, D. Sortases, Surface Proteins, and Their Roles in Staphylococcus aureus Disease and Vaccine Development. *Microbiol. Spectr.* **7**, 1–13 (2019).

81. Monteiro, J. M. *et al.* Peptidoglycan synthesis drives an FtsZ-treadmilling-independent step of cytokinesis. *Nature* **554**, 528–532 (2018).

82. Mazmanian, S. K., Liu, G., Jensen, E. R., Lenoy, E. & Schneewind, O. Staphylococcus aureus sortase mutants defective in the display of surface proteins and in the pathogenesis of animal infections. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5510–5 (2000).

83. Pluym, M., Vermeiren, C. L., Mack, J., Heinrichs, D. E. & Stillman, M. J. Heme binding properties of Staphylococcus aureus IsdE. *Biochemistry* **46**, 12777–12787 (2007).

84. Clarke, S. R., Harris, L. G., Richards, R. G. & Foster, S. J. Protein of Staphylococcus aureus Analysis of Ebh , a 1 . 1-Megadalton Cell Wall-Associated Fibronectin-Binding Protein of Staphylococcus aureus. **70**, 6680–6687 (2002).

85. Lei, M. G., Gupta, R. K. & Lee, C. Y. Proteomics of Staphylococcus aureus biofilm matrix in a rat model of orthopedic implant-associated infection. *PLoS One* **12**, 1–17 (2017).

86. Zong, Y., Bice, T. W., Ton-That, H., Schneewind, O. & Narayana, S. V. L. Crystal structures of Staphylococcus aureus Sortase A and its substrate complex. *J. Biol. Chem.* **279**, 31383–31389 (2004).

87. Ton-That, H. *et al.* Anchoring of Surface Proteins to the Cell Wall of Staphylococcus aureus. *J. Biol. Chem.* **275**, 9876–9881 (2000).

88. Navarre, W. W. & Schneewind, O. Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiol. Mol. Biol. Rev.* **63**, 174–229 (1999).

89. Perry, A. M., Ton-That, H., Mazmanian, S. K. & Schneewind, O. Anchoring of surface proteins to the cell wall of Staphylococcus aureus. III. Lipid II is an in vivo peptidoglycan substrate for sortase-catalyzed surface protein anchoring. *J. Biol. Chem.* **277**, 16241–16248 (2002).

90. Egan, A. J. F., Cleverley, R. M., Peters, K., Lewis, R. J. & Vollmer, W. Regulation of bacterial cell wall growth. *FEBS J.* **284**, 851–867 (2017).

91. Vollmer, W., Blanot, D. & De Pedro, M. A. Peptidoglycan structure and architecture. *FEMS Microbiol. Rev.* **32**, 149–167 (2008).

92. Dmitriev, B. A., Toukach, F. V, Holst, O., Rietschel, E. T. & Ehlers, S. Tertiary Structure of *Staphylococcus aureus* Cell Wall Murein. *J. Bacteriol.* **186**, 7141–7148 (2004).

93. Kim, S. J., Chang, J., Rimal, B., Yang, H. & Schaefer, J. Surface proteins and the formation of biofilms by Staphylococcus aureus. *Biochim. Biophys. Acta - Biomembr.* **1860**, 749–756 (2018).

94. Vermassen, A. *et al.* Cell Wall Hydrolases in Bacteria: Insight on the Diversity of Cell Wall Amidases, Glycosidases and Peptidases Toward Peptidoglycan. *Front. Microbiol.* **10**, (2019).

95. Johnson, J. W., Fisher, J. F. & Mobashery, S. Bacterial cell-wall recycling. *Ann. N. Y. Acad. Sci.* **1277**, 54–75 (2013).

96. Koch, A. L. & Doyle, R. J. Inside-to-outside growth and turnover of the wall of gram-positive rods. *J. Theor. Biol.* **117**, 137–157 (1985).

97. Scheffers, D. J. & Pinho, M. G. Bacterial Cell Wall Synthesis: New Insights from Localization Studies. *Microbiol. Mol. Biol. Rev.* **69**, 585–607 (2005).

98. Borisova, M. *et al.* Peptidoglycan Recycling in Gram-Positive Bacteria Is Crucial for. *MBio* **7**, 1–10 (2016).

99. Resch, A. *et al.* Comparative proteome analysis of Staphylococcus aureus biofilm and planktonic cells and correlation with transcriptome profiling. *Proteomics* **6**, 1867–1877 (2006).

100. Sitkiewicz, I., Babiak, I. & Hryniewicz, W. Characterization of transcription within sdr region of Staphylococcus aureus. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.* **99**, 409–416 (2011).

101. Jenkins, A. *et al.* Differential Expression and Roles of S. aureus Cirulence Determinants during Colonization and Disease. *MBio* **6**, 1–10 (2015).

102. Perkins, S. *et al.* Structural Organization of the Fibrinogen-binding Region of the Clumping Factor B MSCRAMM of Staphylococcus aureus. *J. Biol. Chem.* **276**, 44721–44728 (2001).

103. Deivanayagam, C. C. S. *et al.* Novel fold and assembly of the repetitive B region of the Staphylococcus aureus collagen-binding surface protein. *Structure* **8**, 67–78 (2000).

104. Cucarella, C. *et al.* Expression of the biofilm-associated protein interferes with host protein receptors of Staphylococcus aureus and alters the infective process. *Infect. Immun.* **70**, 3180–3186 (2002).

105. Typas, A., Banzhaf, M., Gross, C. A. & Vollmer, W. From the regulation of peptidoglycan synthesis to bacterial growth and morphology. *Nat. Rev. Microbiol.* **10**, 123–136 (2012).

106. Kukita, K. *et al.* Staphylococcus aureus SasA is responsible for binding to the salivary agglutinin gp340, derived from human saliva. *Infect. Immun.* **81**, 1870–

1879 (2013).

107. Roche, F. M. *et al.* Characterization of novel LPXTG-containing proteins of Staphylococcus aureus identified from genome sequences. *Microbiology* **149**, 643–654 (2003).

108. Siboo, I. R., Chambers, H. F. & Sullam, P. M. Role of SraP, a Serine-Rich Surface Protein of Staphylococcus aureus, in Binding to Human Platelets. *Infect. Immun.* **73**, 2273–2280 (2005).

109. Komatsuzawa, H. *et al.* Tn551-mediated insertional inactivation of the fmtB gene encoding a cell wall-associated protein abolishes methicillin resistance in Staphylococcus aureus. *J. Antimicrob. Chemother.* **45**, 421–431 (2000).

110. Wu, S. W. & De Lencastre, H. Mrp—A New Auxiliary Gene Essential for Optimal Expression of Methicillin Resistance in Staphylococcus aureus. *Microb. Drug Reistance* **5**, (2009).

111. Schroeder, K. *et al.* Molecular characterization of a novel Staphylococcus aureus surface protein (SasC) involved in cell aggregation and biofilm accumulation. *PLoS One* **4**, (2009).

112. Paharik, A. E. & Horswill, A. R. The Staphylococcal Biofilm: Adhesins, Regulation, and Host Response. *Virulence Mech. Bact. Pathog. Fifth Ed.* 529–566 (2016). doi:10.1128/microbiolspec.vmbf-0022-2015

113. Taylor, J. M. & Heinrichs, D. E. Transferrin binding in Staphylococcus aureus: Involvement of a cell wall-anchored protein. *Mol. Microbiol.* **43**, 1603–1614 (2002).

114. Geoghegan, J. A. *et al.* Role of surface protein SasG in biofilm formation by Staphylococcus aureus. *J. Bacteriol.* **192**, 5663–5673 (2010).

115. Thammavongsa, V., Kern, J. W., Missiakas, D. M. & Schneewind, O. Staphylococcus aureus synthesizes adenosine to escape host immune responses

. *J. Exp. Med.* **206**, 2417–2427 (2009).

116. Valle, J. *et al.* Bap, a Biofilm Matrix Protein of Staphylococcus aureus Prevents Cellular Internalization through Binding to GP96 Host Receptor. *PLoS Pathog.* **8**, (2012).

117. Cucarella, C., Solano, C. & Valle, J. Bap, a Staphylococcus aureus surface protein involved in biofilm formation. *J. Bacteriol.* **183**, 2888–2896 (2001).

118. Eldhin, D. N. *et al.* Clumping factor B (ClfB), a new surface-located fibrinogen-binding adhesin of Staphylococcus aureus. *Mol. Microbiol.* **30**, 245–257 (1998).

119. Barbu, E. M., Mackenzie, C., Foster, T. J. & Höök, M. SdrC induces staphylococcal biofilm formation through a homophilic interaction. *Mol. Microbiol.* **94**, 172–185 (2014).

120. O'Brien, L. *et al.* Multiple mechanisms for the activation of human platelet aggregation by Staphylococcus aureus: Roles for the clumping factors ClfA and ClfB, the serine-aspartate repeat protein SdrE and protein A. *Mol. Microbiol.* **44**, 1033–1044 (2002).

121. Herman-bausier, P., Formosa-dague, C., Feuillie, C., Valotteau, C. & Dufrêne, Y. F. Forces guiding staphylococcal adhesion. *J. Struct. Biol.* **197**, 65–69 (2017).

122. Clarke, S. R., Wiltshire, M. D. & Foster, S. J. IsdA of Staphylococcus aureus is a broad spectrum, iron-regulated adhesin. *Mol. Microbiol.* **51**, 1509–1519 (2004).

123. Wann, E. R., Gurusiddappa, S. & Höök, M. The fibronectin-binding MSCRAMM FnbpA of Staphylococcus aureus is a bifunctional protein that also binds to fibrinogen. *J. Biol. Chem.* **275**, 13863–13871 (2000).

124. O'Neill, E. *et al.* A novel Staphylococcus aureus biofilm phenotype mediated by the fibronectin-binding proteins, FnBPA and FnBPB. *J. Bacteriol.* **190**, 3835–3850 (2008).

125. Foster, T. J. The remarkably multifunctional fibronectin binding proteins of Staphylococcus aureus. *Eur. J. Clin. Microbiol. Infect. Dis.* **35**, 1923–1931 (2016).

126. Plata, K., Rosato, A. E. & Wegrzyn, G. Staphylococcus aureus as an infectious agent: Overview of biochemistry and molecular genetics of its pathogenicity. *Acta Biochim. Pol.* **56**, 597–612 (2009).

127. Savolainen, K. *et al.* Expression of pls, a gene closely associated with the mecA gene of methicillin-resistant Staphylococcus aureus, prevents bacterial adhesion in vitro. *Infect. Immun.* **69**, 3013–3020 (2001).

128. Elias, S. & Banin, E. Multi-species biofilms: Living with friendly neighbors. *FEMS Microbiol. Rev.* **36**, 990–1004 (2012).

129. Makovcova, J. *et al.* Dynamics of mono- and dual-species biofilm formation and interactions between Staphylococcus aureus and Gram-negative bacteria. *Microb. Biotechnol.* **10**, 819–832 (2017).

130. Iwase, T. *et al.* Staphylococcus epidermidis Esp inhibits Staphylococcus aureus biofilm formation and nasal colonization. *Nature* **465**, 346–349 (2010).

131. Orazi, G. & O'Toole, G. A. Pseudomonas aeruginosa Alters Staphylococcus aureus Sensitivity to Vancomycin in a Biofilm Model of Cystic Fibrosis Infection . *MBio* **8**, 1–17 (2017).

132. Weigel, L. M. *et al.* High-Level Vancomycin-Resistant Staphylococcus aureus Isolates Associated with a Polymicrobial Biofilm. *Antimicrob. Agents Chemother.* **51**, 231–238 (2006).

133. Jiang, J.-H. *et al.* Antibiotic resistance and host immune evasion in Staphylococcus aureus mediated by a metabolic adaptation . *Proc. Natl. Acad. Sci.* **116**, 3722–3727 (2019).

134. Haaber, J., Penadés, J. R. & Ingmer, H. Transfer of Antibiotic Resistance in Staphylococcus aureus. *Trends Microbiol.* **xx**, 1–13 (2017).

135. Langhanki, L. *et al.* In vivo competition and horizontal gene transfer among distinct Staphylococcus aureus lineages as major drivers for adaptational changes during long-term persistence in humans. *BMC Microbiol.* **18**, 1–12 (2018).

136. King, A. N. *et al.* Guanine limitation results in CodY-dependent and -independent alteration of S. aureus physiology and gene expression. *J. Bacteriol.* **200**, e00136-18 (2018).

137. Mah, T. F. C. & O'Toole, G. A. Mechanisms of biofilm resistance to antimicrobial agents. *Trends Microbiol.* **9**, 34–39 (2001).

138. Barber, K. E., Werth, B. J., McRoberts, J. P. & Rybak, M. J. A novel approach utilizing biofilm time-kill curves to assess the bactericidal activity of ceftaroline combinations against biofilm- producing methicillin-resistant staphylococcus aureus. *Antimicrob. Agents Chemother.* **58**, 2989–2992 (2014).

139. Yan, J., Nadell, C. D. & Bassler, B. L. Environmental fluctuation governs selection for plasticity in biofilm production. *ISME J.* **11**, 1569–1577 (2017).

140. Singh, R., Sahore, S., Kaur, P., Rani, A. & Ray, P. Penetration barrier contributes to bacterial biofilm-associated resistance against only select antibiotics, and exhibits genus-, strain- and antibiotic-specific differences. *Pathog. Dis.* **74**, 1–6 (2016).

141. Hoyle, B. D., Alcantara, J. & Costerton, J. W. Pseudomonas aeruginosa Biofilm as a Diffusion Barrier to Piperacillin. *Antimicrob. Agents Chemother.* **36**, 2054–2056 (1992).

142. Stewart, P. S. Mechanisms of antibiotic resistance in bacterial biofilms. *Int. J. Med. Microbiol.* **292**, 107–113 (2002).

143. Cornforth, D. M. & Foster, K. R. Competition sensing: The social side of bacterial stress responses. *Nat. Rev. Microbiol.* **11**, 285–293 (2013).

144. Lister, J. L. & Horswill, A. R. Staphylococcus aureus biofilms: recent developments in biofilm dispersal. *Front. Cell. Infect. Microbiol.* **4**, 1–9 (2014).

145. Francois, P. *et al.* Identification of plasma proteins adsorbed on hemodialysis tubing that promote Staphylococcus aureus adhesion. *J. Lab. Clin. Med.* **135**, 32–42 (2000).

146. Vaudaux, P. *et al.* Host Factors Selectively Increase Staphylococcal Adherence on Inserted Catheters: A Role for Fibronectin and Fibrinogen or Fibrin. *J. Infect. Dis.* **160**, 865–875 (1989).

147. Feuillie, C. *et al.* Molecular interactions and inhibition of the staphylococcal biofilm-forming protein SdrC. *Proc. Natl. Acad. Sci.* **114**, 3738–3743 (2017).

148. Gross, M., Cramton, S. E., Götz, F. & Peschel, A. Key Role of Teichoic Acid Net Charge in Staphylococcus aureus Colonization of Artificial Surfaces Key Role of Teichoic Acid Net Charge in Staphylococcus aureus Colonization of Artificial Surfaces. *Infect. Immun.* **69**, 3423–2426 (2001).

149. Baier, R. E. The organization of blood components near interfaces. *Ann. N. Y. Acad. Sci.* **283**, 17–36 (1977).

150. Otto, M. Staphylococcal Infections: Mechanisms of Biofilm Maturation and Detachment as Critical Determinants of Pathogenicity. *Annu. Rev. Med.* **64**, 175–188 (2012).

151. Von Eiff, C., Heilmann, C., Herrmann, M. & Peters, G. Basic aspects of the pathogenesis of staphylococcal polymer-associated infections. *Infection* **27**, 7–10 (1999).

152. Mack, D., Haeder, M. & Siemssen, N. Association of Biofilm Production of Coagulase-Negative Staphylococci with Expression of a Specific Polysaccharide Intercellular Adhesin Author (s): Dietrich Mack , Michael Haeder , Nicolaus Siemssen and Rainer Laufs Published by : Oxford University Pre. **174**, 881–884

(2016).

153. Maira-litrán, T. *et al.* Immunochemical Properties of the Staphylococcal Poly- N -Acetylglucosamine Surface Polysaccharide Immunochemical Properties of the Staphylococcal Poly- N -Acetylglucosamine Surface Polysaccharide. *Infect. Immun.* **70**, 4433–4440 (2002).

154. Fitzpatrick, F., Humphreys, H. & O'Gara, J. P. Evidence for icaADBC-independent biofilm development mechanism in methicillin-resistant Staphylococcus aureus clinical isolates. *J. Clin. Microbiol.* **43**, 1973–1976 (2005).

155. Merino, N. *et al.* Protein A-mediated multicellular behavior in Staphylococcus aureus. *J. Bacteriol.* **191**, 832–843 (2009).

156. Geoghegan, J. A., Monk, I. R., O'Gara, J. P. & Foster, T. J. Subdomains N2N3 of fibronectin binding protein a mediate staphylococcus aureus biofilm formation and adherence to fibrinogen using distinct mechanisms. *J. Bacteriol.* **195**, 2675–2683 (2013).

157. Abraham, N. M. & Jefferson, K. K. Staphylococcus aureus clumping factor B mediates biofilm formation in the absence of calcium. *Microbiol. (United Kingdom)* **158**, 1504–1512 (2012).

158. Corrigan, R. M., Rigby, D., Handley, P. & Foster, T. J. The role of Staphylococcus aureus surface protein SasG in adherence and biofilm formation. *Microbiology* **153**, 2435–2446 (2007).

159. Irina, S., Evgueny, V., Sigrid, F., Grigorij, K. & Saïd, J. Extracellular carbohydrate-containing polymers of a model biofilm-producing strain, Staphylococcus epidermidis RP62A. *Infect. Immun.* **73**, 3007–3017 (2005).

160. Sugimoto, S. *et al.* Broad impact of extracellular DNA on biofilm formation by clinically isolated Methicillin-resistant and -sensitive strains of Staphylococcus aureus. *Sci. Rep.* 1–11 (2018). doi:10.1038/s41598-018-20485-z

161. Flemming, H. & Wingender, J. The biofilm matrix. *Nat. Rev. Microbiol.* **8**, 623–33 (2010).

162. Zhang, W. *et al.* Extracellular matrix-associated proteins form an integral and dynamic system during Pseudomonas aeruginosa biofilm development. *Front. Cell. Infect. Microbiol.* **5**, 1–10 (2015).

163. Arciola, C. R., Campoccia, D. & Montanaro, L. Implant infections: Adhesion, biofilm formation and immune evasion. *Nat. Rev. Microbiol.* **16**, 397–409 (2018).

164. Götz, F. Staphylococcus and biofilms. *Mol. Microbiol.* **43**, 1367–1378 (2002).

165. Kretschmer, D. *et al.* Human formyl peptide receptor 2 senses highly pathogenic Staphylococcus aureus. *Cell Host Microbe* **7**, 463–473 (2010).

166. Wang, R. *et al.* Identification of novel cytolytic peptides as key virulence determinants for community-associated MRSA. *Nat. Med.* **13**, 1510–1514 (2007).

167. McGavin, M. J., Zahradka, C., Rice, K. & Scott, J. E. Modification of the Staphylococcus aureus fibronectin binding phenotype by V8 protease. *Infect. Immun.* **65**, 2621–2628 (1997).

168. Lauderdale, K. J., Malone, C. L., Boles, B. R., Morcuende, J. & Horswill, A. R. Biofilm dispersal of community-associated methicillin-resistant Staphylococcus aureus on orthopedic implant material. *J. Orthop. Res.* **28**, 55–61 (2010).

169. Martí, M. *et al.* Extracellular proteases inhibit protein-dependent biofilm formation in Staphylococcus aureus. *Microbes Infect.* **12**, 55–64 (2010).

170. Kiedrowski, M. R. *et al.* Nuclease modulates biofilm formation in community-associated methicillin-resistant staphylococcus aureus. *PLoS One* **6**, (2011).

171. Kiedrowski, M. R. *et al.* Staphylococcus aureus Nuc2 is a functional, surface-attached extracellular nuclease. *PLoS One* **9**, (2014).

172. Donelli, G. *et al.* Synergistic activity of dispersin B and cefamandole nafate in inhibition of staphylococcal biofilm growth on polyurethanes. *Antimicrob. Agents Chemother.* **51**, 2733–2740 (2007).

173. Periasamy, S. *et al.* How Staphylococcus aureus biofilms develop their characteristic structure. *Proc. Natl. Acad. Sci.* **109**, 1281–1286 (2012).

174. Zapotoczna, M., O'Neill, E. & O'Gara, J. P. Untangling the Diverse and Redundant Mechanisms of Staphylococcus aureus Biofilm Formation. *PLOS Pathog.* **12**, 1–6 (2016).

175. Zhang, Y., Lei, Y., Nobbs, A., Khammanivong, A. & Herzberg, M. C. Inactivation of Streptococcus gordonii SspAB alters expression of multiple adhesin genes. *Infect. Immun.* **73**, 3351–3357 (2005).

176. Takamatsu, D., Bensing, B. A., Prakobphol, A., Fisher, S. J. & Sullam, P. M. Binding of the Streptococcal Surface Glycoproteins GspB and Hsa to Human Salivary Proteins. *Infect. Immun.* **74**, 1933–1940 (2006).

177. Koo, H., Xiao, J., Klein, M. I. & Jeon, J. G. Exopolysaccharides produced by Streptococcus mutans glucosyltransferases modulate the establishment of microcolonies within multispecies biofilms. *J. Bacteriol.* **192**, 3024–3032 (2010).

178. McNab, R. *et al.* LuxS-Based Signaling in *Streptococcus gordonii*: Autoinducer 2 Controls Carbohydrate Metabolism and Biofilm Formation with *Porphyromonas gingivalis*. *J. Bacteriol.* **185**, 274–284 (2003).

179. Back, C. R., Douglas, S. K., Emerson, J. E., Nobbs, A. H. & Jenkinson, H. F. Streptococcus gordonii DL1 adhesin SspB V-region mediates coaggregation via receptor polysaccharide of Actinomyces oris T14V. *Mol. Oral Microbiol.* **30**, 411–424 (2015).

180. Mutha, N. V. R. *et al.* Transcriptional responses of Streptococcus gordonii and Fusobacterium nucleatum to coaggregation. *Mol. Oral Microbiol.* **33**, 450–464

(2018).

181. Rabin, N. *et al.* Biofilm formation mechanisms and targets for developing antibiofilm agents. *Future Med. Chem.* **7**, 493–512 (2015).

182. Rossmann, M. G. & Argos, P. Protein Folding. *Annu. Rev. Biochem.* **50**, 497–532 (1981).

183. Basu, M. K., Poliakov, E. & Rogozin, I. B. Domain mobility in proteins: Functional and evolutionary implications. *Brief. Bioinform.* **10**, 205–216 (2009).

184. Majumdar, I., Kinch, L. N. & Grishin, N. V. A database of domain definitions for proteins with complex interdomain geometry. *PLoS One* **4**, (2009).

185. Bateman, A. *et al.* The Pfam Protein Families Database. *Nucleic Acids Res.* **28**, 263–266 (1999).

186. Janin, J. & Chothia, C. Domains in proteins: Definitions, location, and structural principles. *Methods Enzymol.* **115**, 420–430 (1985).

187. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *J Mol Biol* **247**, 536–540 (1995).

188. Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**, D247–D251 (2005).

189. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

190. Levitt, M. Structural patterns in globular proteins. *Nature* **261**, 552–558 (1976).

191. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).

192. Berezin, C. *et al.* ConSeq: The identification of functionally and structurally

important residues in protein sequences. *Bioinformatics* **20**, 1322–1324 (2004).

193. Dobson CM. Protein Folding and Misfolding. *Nature* **426**, 884–90 (2003).

194. Jacobs, D. J. & Dallakyan, S. Elucidating protein thermodynamics from the three-dimensional structure of the native state using network rigidity. *Biophys. J.* **88**, 903–915 (2005).

195. Dill, K. A. Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155 (1990).

196. Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A. & Clarke, J. The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* **8**, 319–330 (2007).

197. Pasek, S., Risler, J. L. & Brézellec, P. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* **22**, 1418–1423 (2006).

198. Vogel, C. *et al.* Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* **14**, 208–216 (2004).

199. Apic, G., Huber, W. & Teichmann, S. A. Multi-domain protein families and domain pairs: Comparison with known structures and a random model of domain recombination. *J. Struct. Funct. Genomics* **4**, 67–78 (2003).

200. Aloy, P., Ceulemans, H., Stark, A. & Russell, R. B. The Relationship Between Sequence and Interaction Divergence in Proteins. 989–998 (2003). doi:10.1016/j.jmb.2003.07.006

201. Larsen, T. a, Olson, A. J. & Goodsell, D. S. Morphology of protein – protein interfaces. *Structure* **6**, 421–427 (1998).

202. Arviv, O. & Levy, Y. Folding of multidomain proteins: Biophysical consequences of tethering even in apparently independent folding. *Proteins Struct. Funct. Bioinforma.* **80**, 2780–2798 (2012).

203. Improta, S. *et al.* The Assembly of Immunoglobulin-like Modules in Titin :

Implications for Muscle Elasticity. (1998).

204. Scott, K. A., Steward, A., Fowler, S. B. & Clarke, J. Titin ; a Multidomain Protein that Behaves as the Sum of its Parts. *JMB* 819–829 (2002). doi:10.1006/jmbi.2001.5260

205. MacDonald, R. I. & Pozharski, E. V. Free energies of urea and of thermal unfolding show that two tandem repeats of spectrin are thermodynamically more stable than a single repeat. *Biochemistry* **40**, 3974–3984 (2001).

206. MacDonald, R. I. & Cummings, J. A. Stabilities of folding of clustered, two-repeat fragments of spectrin reveal a potential hinge in the human erythroid spectrin tetramer. *Proc. Natl. Acad. Sci.* **101**, 1502–1507 (2004).

207. Hunter, C. A. & Anderson, H. L. What is cooperativity? *Angew. Chemie - Int. Ed.* **48**, 7488–7499 (2009).

208. Heringa, J. & Taylor, W. R. Three-dimensional domain duplication , swapping and stealing. *Curr. Opin. Struct. Biol.* **7**, 416–421 (1997).

209. Kusunoki, H., MacDonald, R. I. & Mondragón, A. Structural insights into the stability and flexibility of unusual erythroid spectrin repeats. *Structure* **12**, 645–656 (2004).

210. Lin, I. H., Hsu, M. T. & Chang, C. H. Protein domain repetition is enriched in Streptococcal cell-surface proteins. *Genomics* **100**, 370–379 (2012).

211. Wright, C. F., Teichmann, S. A., Clarke, J. & Dobson, C. M. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature* **438**, 878–881 (2005).

212. Borgia, M. B. *et al.* Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* **474**, 662–665 (2011).

213. Kajander, T., Cortajarena, A. L., Main, E. R. G., Mochrie, S. G. J. & Regan, L. A new

folding paradigm for repeat proteins. *J. Am. Chem. Soc.* **127**, 10188–10190 (2005).

214. Aksel, T. & Barrick, D. Chapter 4. Analysis of Repeat-Protein Folding Using Nearest-Neighbor Statistical Mechanical Models. *Methods Enzymol.* **455**, 95–125 (2009).

215. Mello, C. C. & Barrick, D. An experimentally determined protein folding energy landscape. *PNAS* **101**, 14102–14107 (2004).

216. Zweifel, M. E. & Barrick, D. Studies of the ankyrin repeats of the Drosophila melanogaster Notch receptor. 2. Solution stability and cooperativity of unfolding. *Biochemistry* **40**, 14357–14367 (2001).

217. Zweifel, M. E. & Barrick, D. Studies of the ankyrin repeats of the Drosophila melanogaster Notch receptor. 1. Solution conformational and hydrodynamic properties. *Biochemistry* **40**, 14344–14356 (2001).

218. D'Andrea, L. D. & Regan, L. TPR proteins: The versatile helix. *Trends Biochem. Sci.* **28**, 655–662 (2003).

219. Main, E. R. G., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. Design of stable α-helical arrays from an idealized TPR motif. *Structure* **11**, 497–508 (2003).

220. Main, E. R. G., Stott, K., Jackson, S. E. & Regan, L. Local and long-range stability in tandemly arrayed tetratricopeptide repeats. *Proc. Natl. Acad. Sci.* **102**, 5721–5726 (2005).

221. Mazmanian, S. K., Ton-That, H. & Schneewind, O. Sortase-catalysed anchoring of surface proteins to the cell wall of Staphylococcus aureus. *Mol. Microbiol.* **40**, 1049–1057 (2001).

222. Robinson, D. A. & Enright, M. C. Evolutionary models of the emergence of methicillin-resistant Staphylococcus aureus. *Antimicrob.Agents Chemother.* **47**, 3926–3934 (2003).

223. Komatsuzawa, H. *et al.* Cloning and characterization of the fmt gene which affects the methicillin resistance level and autolysis in the presence of triton X-100 in methicillin-resistant Staphylococcus aureus. *Antimicrob. Agents Chemother.* **41**, 2355–61 (1997).

224. Koley, D. & Bard, A. J. Triton X-100 concentration effects on membrane permeability of a single HeLa cell by scanning electrochemical microscopy (SECM). *Proc. Natl. Acad. Sci.* **107**, 16783–16787 (2010).

225. Stapleton, P. D. & Taylor, P. W. Methicillin resistance in Staphylococcus aureus : Methicillin resistance. *Sci. Prog.* **85**, 57–72 (2002).

226. Reed, P. *et al.* Staphylococcus aureus Survives with a Minimal Peptidoglycan Synthesis Machine but Sacrifices Virulence and Antibiotic Resistance. *PLoS Pathog.* **11**, 1–19 (2015).

227. Linke, C., Siemens, N., Middleditch, M. J., Kreikemeyer, B. & Baker, E. N. Purification, crystallization and preliminary crystallographic analysis of the adhesion domain of Epf from Streptococcus pyogenes. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **68**, 793–797 (2012).

228. Linke, C. *et al.* The extracellular protein factor Epf from Streptococcus pyogenes is a cell surface adhesin that binds to cells through an N-terminal domain containing a carbohydrate-binding module. *J. Biol. Chem.* **287**, 38178–38189 (2012).

229. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40 (2008).

230. Davies, J. R., Svensäter, G. & Herzberg, M. C. Identification of novel LPXTG-linked surface proteins from Streptococcus gordonii. *Microbiology* **155**, 1977–1988 (2009).

231. Nylander, Å. *et al.* Structural and Functional Analysis of the N-terminal Domain

of the Streptococcus gordonii Adhesin Sgo0707. *PLoS One* **8**, (2013).

232. Forsgren, N., Lamont, R. J. & Persson, K. Crystal structure of the variable domain of the Streptococcus gordonii surface protein SspB. *Protein Sci.* **18**, 1896–1905 (2009).

233. Best, R. B. & Clarke, J. What can atomic force microscopy tell us about protein folding ? 183–192 (2002).

234. Shareef, M. M. *et al.* A noncommercial polymerase chain reaction-based method to approach one hundred percent recombinant clone selection efficiency. *Anal. Biochem.* **382**, 75–76 (2008).

235. Studier, F. W. & Moffatt, B. A. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189**, 113–130 (1986).

236. O'Neill, A. J. Staphylococcus aureus SH1000 and 8325-4: Comparative genome sequences of key laboratory strains in staphylococcal research. *Lett. Appl. Microbiol.* **51**, 358–361 (2010).

237. Sambrook, J. & Russell, D. W. *Molecular Cloning, A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, 2001).

238. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).

239. BD. *BBL ™ Trypticase ™ Soy Broth*. (2015).

240. Rosenberg, A. H. *et al.* Vectors for selective expression of cloned DNAs by T7 RNA polymerase. *Gene* **56**, 125–135 (1987).

241. Guzman, L.-M., Belin, D., Carson, M. J. & Beckwith, J. Tight Regulation, Modulation, and High-Level Expression by Vectors Containing the Arabinose P BAD Promoter. *J. Bacteriol.* **177**, 4121–4130 (1995).

242. Guzman, L.-M., Belin, D., Cartee, R. T., J., M. & Beckwith, J. Tight Regulation, Modulation, and High-Level Expression by Vectors containing the Arabinose Pbad promoter. *J. Bacteriol.* **177**, 4121–4130 (2000).

243. Invitrogen Life Technologies. *Tightly regulated Bacterial Protein Expression - pBAD expression system*. (2002).

244. Clontech (Takara). In-Fusion® HD Cloning Kit User Manual. *In-Fusion Cloning* **1**, 1–15 (2012).

245. Sørensen, H. P. & Mortensen, K. K. Advanced genetic strategies for recombinant protein expression in Escherichia coli. *J. Biotechnol.* **115**, 113–128 (2005).

246. Zheng, L., Baumann, U. & Reymond, J. L. An efficient one-step site-directed and site-saturation mutagenesis protocol. *Nucleic Acids Res.* **32**, e115–e115 (2004).

247. Liu, H. & Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol.* **8**, 91 (2008).

248. Schmidt, T. G. M. & Skerra, A. The random peptide library-assisted engineering of a C-terminal affinity peptide, useful for the detection and purification of a functional Ig Fv fragment. *Protein Eng. Des. Sel.* **6**, 109–122 (1993).

249. Daniels, R. *et al.* Disulfide bond formation and cysteine exclusion in gram-positive bacteria. *J. Biol. Chem.* **285**, 3300–3309 (2010).

250. Wilchek, M. & Bayer, E. A. Introduction to avidin-biotin technology. *Methods Enzymol.* **184**, 5–13 (1990).

251. Schmidt, T. G. M. & Skerra, A. One-step affinity purification of bacterially produced proteins by means of the 'Strep tag' and immobilized recombinant core streptavidin. *J. Chromatogr. A* **676**, 337–345 (1994).

252. GE Healthcare Life Sciences. Instructions HiLoad 16/600 and 26/600 Superdex

30 prep grade, HiLoad 16/600 and 26/600 Superdex 75 prep grade, HiLoad 16/600 and 26/600 Superdex 200 prep grade. *Gen. Elecric Healthc.* **28**-**9920**–**17**, 26–31 (2011).

253. Gasteiger, E. *et al. Chapter 52: Protein Identification and Analysis Tools on the ExPASy Server*. *The Proteomics Protocols Handbook* (Humana Press, 2005). doi:10.1385/1592598900

254. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4**, 2411–2423 (1995).

255. Drozdetskly, A., Cole, C., Proctor, J. & Barton, G. J. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, W389–W394 (2015).

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

257. Altschul, S. F. *et al.* Gapped BLAST and PS I-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402 (1997).

258. Cuff, J. A. & Barton, G. J. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins Struct. Funct. Genet.* **40**, 502–511 (2000).

259. Sievers, F. *et al.* Fast , scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 1–6 (2011).

260. Chojnacki, S., Cowley, A., Lee, J., Foix, A. & Lopez, R. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Res.* **45**, W550–W553 (2017).

261. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7 : Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

262. Dessimoz, C., Gil, M., Schneider, A. & Gonnet, G. H. Fast estimation of the difference between two PAM / JTT evolutionary distances in triplets of homologous sequences. *BMC Bioinformatics* **7**, 529 (2006).

263. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

264. Wyatt, P. J. Light scattering and the absolute characterization of macromolecules. *Anal. Chim. Acta* **272**, 1–40 (1993).

265. Zimm, B. H. The scattering of light and the radial distribution function of high polymer solutions. *J. Chem. Phys.* **16**, 1093–1099 (1948).

266. Koppel, D. E. Analysis of macromolecular polydispersity in intensity correlation spectroscopy: The method of cumulants. *J. Chem. Phys.* **57**, 4814–4820 (1972).

267. Kato, T., Nakamura, K., Kawaguchi, M. & Takahashi, A. Quasielastic Light Scattering Measurements of Polystyrene Latices and Conformation of Poly(oxyethylene) Adsorbed on the Latices. *Polymer Journal* **13**, 1037–1043 (1981).

268. Ye, H. Simultaneous determination of protein aggregation, degradation, and absolute molecular weight by size exclusion chromatography-multiangle laser light scattering. *Anal. Biochem.* **356**, 76–85 (2006).

269. Whitmore, L. & Wallace, B. A. Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases. *Biopolymers* **89**, 392–400 (2008).

270. Greenfield, N. J. & Fasman, G. D. Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry* **8**, 4108–4116 (1969).

271. Lobley, A., Whitmore, L. & Wallace, B. A. DICHROWEB: An interactive website for the analysis of protein secondary structure from circular dichroism spectra.

Bioinformatics **18**, 211–212 (2002).

272. Whitmore, L. & Wallace, B. A. Dichroweb. *Online analysis for protein circular dichroism spectra* (2016). Available at: http://dichroweb.cryst.bbk.ac.uk/html/home.shtml. (Accessed: 9th June 2016)

273. Provencher, S. W. CONTIN: A General Purpose Constrained Regularization Program for Inverting Noisy Linear Algebraic and Integral Equations. *Comput. Phys. Commun.* **27**, 229–242 (1982).

274. van Stokkum, I. H. M., Spoelder, H. J. W., Bloemendal, M., van Grondelle, R. & Groen, F. C. A. Estimation of protein secondary structure and error analysis from circular dichroism spectra. *Anal. Biochem.* **191**, 110–118 (1990).

275. Sreerama, N. & Woody, R. W. Estimation of protein secondary structure from circular dichroism spectra: Comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal. Biochem.* **287**, 252–260 (2000).

276. Provencher, S. W. & Glöckner, J. Estimation of Globular Protein Secondary Structure from Circular Dichroism. *Biochemistry* **20**, 33–37 (1981).

277. Hospes, M., Hendriks, J. & Hellingwerf, K. J. Tryptophan fluorescence as a reporter for structural changes in photoactive yellow protein elicited by photo-activation. *Photochem. Photobiol. Sci.* **12**, 479–488 (2013).

278. Joshi, D., Kumar, D., Maini, A. K. & Sharma, R. C. Detection of biological warfare agents using ultra violet-laser induced fluorescence LIDAR. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **112**, 446–456 (2013).

279. Cowgill, R. W. Fluorescence and protein structure 4. Tyrosine fluorescence in helical muscle proteins. *Biochim. Biophys. Acta* **168**, 417–430 (1968).

280. Satoh, A. & Mihashi, K. Thermal Modification of Structure of Tropomyosin. *J. Biochem.* **71**, 597–605 (1972).

281. Maschberger, M., Resch, H. M., Duhr, S. & Breitsprecher, D. *Exploring Protein Stability by nanoDSF Prometheus NT . 48 – The Stability Expert*. (2015).

282. Söltl, F., Derix, J., Blech, M. & Breitsprecher, D. *Unfolding and Aggregation of mAbsAnalysis of formulation-dependent colloidal and conformational stability of monoclonal antibodies. NanoTemper Technologies*. *Application Note NT-PR-005* **Applicatio**, (2015).

283. van Holde, K. E., Johnson, W. C. & Ho, P. S. *Principles of Physical Biochemistry. Chapter 12.* (Prentice-Hall, Inc, 1998).

284. Sheenan, D. *Physical Biochemistry: Principles and Applications*. (2008).

285. Levitt, M. H. *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. (John Wiley & Sons, Ltd, 2000).

286. Reson8. *NMR Training Week Notes*. (2017).

287. Wüthrich, K. NMR studies of structure and function of biological macromolecules (Nobel Lecture). *Angew. Chemie - Int. Ed.* **42**, 3340–3363 (2003).

288. Reich, H. *8.1 Relaxation in NMR Spectroscopy*. (2017).

289. Rule, G. S. & Hitchens, T. K. Chapter 19: Nuclear spin relaxation and molecular dynamics. in *Fundamentals of Protein NMR Spectroscopy* (2006).

290. Kay, L. E., Torchia, D. A. & Bax, A. Backbone Dynamics of Proteins As Studied by 15N Inverse Detected Heteronuclear NMR Spectroscopy: Application to Staphylococcal Nuclease? *Biochemistry* **28**, 8972–8979 (1989).

291. Grzesiek, S. *Notes on relaxation and dynamics*. *EMBO Course* (2003).

292. Jaravine, V., Ibraghimov, I. & Orekhov, V. Y. Removal of a time barrier for high-resolution multidimensional NMR spectroscopy. *Nat. Methods* **3**, 605–607 (2006).

293. Kleckner, I. R. & Foster, M. P. An introduction to NMR-based approaches for measuring protein dynamics. *Biochim. Biophys. Acta - Proteins Proteomics* **1814**, 942–968 (2011).

294. Cavanagh, J., Fairbrother, W. J., Palmer III, A. G., Rance, M. & Skelton, N. J. Classical NMR spectroscopy. in *Protein NMR Spectroscopy: principles and practice* 17–21 (Elsevier Academic Press, 2007).

295. Rossi, P. *et al.* A microscale protein NMR sample screening pipeline. *J. Biomol. NMR* **46**, 11–22 (2010).

296. Barbato, G., Ikura, M., Kay, L. E., Pastor, R. W. & Bax, A. Backbone Dynamics of Calmodulin Studied by 15N Relaxation Using Inverse Detected Two-Dimensional NMR Spectroscopy: The Central Helix Is Flexible. *Biochemistry* **31**, 5269–5278 (1992).

297. Farrow, N. A. *et al.* Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation. *Biochemistry* **33**, 5984–6003 (1994).

298. Helliwell, J. R. *Synchrotron X-radiation protein crystallography: Instrumentation, methods and applications*. *Reports on Progress in Physics* **47**, (1984).

299. Clegg, W. *X-ray crystallography*. (Oxford Univeristy Press, 2015).

300. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.* **275**, 1–21 (2008).

301. Taylor, G. The phase problem. *Acta Crystallogr. D. Biol. Crystallogr.* **59**, 1881–90 (2003).

302. Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *FEBS J.* **280**, 5705–5736 (2013).

303. Dauter, M. & Dauter, Z. Phase determination using halide ions. *Methods Mol. Biol.* **364**, 149–158 (2007).

304. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

305. Evans, P. & McCoy, A. An introduction to molecular replacement. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **64**, 1–10 (2007).

306. Dodson, E. Is it jolly SAD? *Acta Crystallogr. - Sect. D Biol. Crystallogr.* **59**, 1958–1965 (2003).

307. Brennan, S. & Cowan, P. L. A suite of programs for calculating x-ray absorption, reflection, and diffraction performance for a variety of materials at arbitrary wavelengths. *Rev. Sci. Instrum.* **63**, 850–853 (1992).

308. McCoy, A. J. Phaser wiki - Molecular replacement. (2018). Available at: http://www.phaser.cimr.cam.ac.uk/index.php/Molecular_Replacement.

309. Weiss, M. S. & Hilgenfeld, R. On the use of the merging R factor as a quality indicator for X-ray data. *J. Appl. Crystallogr.* **30**, 203–205 (1997).

310. Karplus, P. A. & Diederichs, K. Linking Crystallographic Model and Data Quality. *Science (80-.).* **336**, 1030–1034 (2012).

311. Weiss, M. S. Global indicators of X-ray data quality. *J. Appl. Crystallogr.* **34**, 130–135 (2001).

312. Karplus, P. A. & Diederichs, K. Assessing and maximizing data quality in macromolecular crystallography. *Curr. Opin. Struct. Biol.* **34**, 60–68 (2015).

313. Kleywegt, G. J. & Jones, T. A. Model building and refinement practice. *Methods Enzymol.* **277**, 208–230 (1997).

314. Brunger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475 (1992).

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

316. Acharya, K. R. & Lloyd, M. D. The advantages and limitations of protein crystal structures. *Trends Pharmacol. Sci.* **26**, 10–14 (2005).

317. Kleywegt, G. J. Validation of protein crystal structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **56**, 249–265 (2000).

318. Newman, J. *et al.* Towards rationalization of crystallization screening for small-To medium-sized academic laboratories: The PACT/JCSG+ strategy. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **61**, 1426–1431 (2005).

319. Anand, K., Pal, D. & Hilgenfeld, R. An overview on 2-methyl-2,4-pentanediol in crystallization and in crystals of biological macromolecules. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58**, 1722–1728 (2002).

320. McPherson, A. & Gavira, J. A. Introduction to protein crystallization. *Acta Crystallogr. Sect. FStructural Biol. Commun.* **70**, 2–20 (2014).

321. Winter, G., Lobley, C. M. C. & Prince, S. M. Decision making in xia2. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **69**, 1260–1273 (2013).

322. Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. A graphical user interface to the CCP4 program suite. *Acta Crystallogr. - Sect. D Biol. Crystallogr.* **59**, 1131–1137 (2003).

323. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. Sect. D Biol. Crystallogr.* **69**, 1204–1214 (2013).

324. Sheldrick, G. M. Experimental phasing with SHELXC / D / E : combining chain tracing with density modification research papers. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **D66**, 479–485 (2009).

325. Schneider, T. R. & Sheldrick, G. M. Substructure solution with SHELXD. *Acta*

Crystallogr. - Sect. D Biol. Crystallogr. **D58**, 1772–1779 (2002).

326. CCP4 Wiki. SHELX C/D/E. (2018). Available at: https://strucbio.biologie.uni-konstanz.de/ccp4wiki/index.php?title=SHELX_C/D/E#SHELXD. (Accessed: 14th May 2019)

327. Skubák, P. & Pannu, N. S. Automatic protein structure solution from weak X-ray data. *Nat. Commun.* **4**, 1–6 (2013).

328. Terwilliger, T. C. *et al.* Decision-making in structure solution using Bayesian estimates of map quality: The PHENIX AutoSol wizard. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **65**, 582–601 (2009).

329. Thorn, A. & Sheldrick, G. M. Extending molecular-replacement solutions with SHELXE. *Acta Crystallogr. Sect. D* 2251–2256 (2013). doi:10.1107/S0907444913027534

330. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **62**, 1002–1011 (2006).

331. Morris, R. J. *et al.* Breaking good resolutions with ARP/wARP. *J. Synchrotron Radiat.* **11**, 56–59 (2004).

332. Emsley, P. & Cowtan, K. Coot: Model-building tools for molecular graphics. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

333. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 355–367 (2011).

334. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221 (2010).

335. Afonine, P. V. *et al.* Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. Sect. D Struct. Biol.* **74**, 531–544 (2018).

336. McNicholas, S., Potterton, E., Wilson, K. S. & Noble, M. E. M. Presenting your structures: The CCP4mg molecular-graphics software. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 386–394 (2011).

337. Hubbard, S. & Thornton, J. Naccess V2.1.1 - Solvent accessible area calculations. (1993). Available at: http://www.bioinf.manchester.ac.uk/naccess/. (Accessed: 24th August 2018)

338. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797 (2007).

339. Li, J. & Liu, Q. 'Double water exclusion': A hypothesis refining the O-ring theory for the hot spots at protein interfaces. *Bioinformatics* **25**, 743–750 (2009).

340. Guney, E., Tuncbag, N., Keskin, O. & Gursoy, A. HotSprint: Database of computational hot spots in protein interfaces. *Nucleic Acids Res.* **36**, 662–666 (2008).

341. Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Hot spots - A Review of the protein-protein interface determinant amino-acid residues. *Proteins* **68**, 803–812 (2007).

342. Rief, M., Pascual, J., Saraste, M. & Gaub, H. E. Single Molecule Force Spectroscopy of Spectrin Repeats : Low Unfolding Forces in Helix Bundles. (1999).

343. Kikhney, A. G. & Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* **589**, 2570–2577 (2015).

344. Svergun, D. I. & Koch, M. H. J. Small-angle scattering studies of biological macromolecules in solution. *Reports Prog. Phys.* **66**, 1735–1782 (2003).

345. Glatter, O. *et al. Small Angle X-ray Scattering*. (Academic Press Inc. (London) Ltd., 1982).

346. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. *PRIMUS* : a Windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.* **36**, 1277–1282 (2003).

347. Hammouda, B. *Chapter 22: Standard Plots*. *National Institute of Standards and Technology, Center for Neutron Research, Distance Learning Course* (2016).

348. Hammouda, B. A new Guinier-Porod model. *J. Appl. Crystallogr.* **43**, 716–719 (2010).

349. Molodenskiy, D. *et al.* Thermally induced conformational changes and protein-protein interactions of bovine serum albumin in aqueous solution under different pH and ionic strengths as revealed by SAXS measurements. *Phys. Chem. Chem. Phys.* **19**, 17143–17155 (2017).

350. Rambo, R. P. & Tainer, J. A. Super-Resolution in Solution X-Ray Scattering and Its Applications to Structural Systems Biology. *Annu. Rev. Biophys.* **42**, 415–441 (2013).

351. Franke, D. *et al.* ATSAS 2.8: A comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystallogr.* **50**, 1212–1225 (2017).

352. Konarev, P. V & Svergun, D. I. A posteriori determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems. *IUCrJ* **2**, 352–360 (2015).

353. Fischer, H., De Oliveira Neto, M., Napolitano, H. B., Polikarpov, I. & Craievich, A. F. Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. *J. Appl. Crystallogr.* **43**, 101–109 (2009).

354. Glatter, O. The interpretation of real-space information from small-angle scattering experiments. *J. Appl. Crystallogr.* **12**, 166–175 (1979).

355. Semenyuk, A. V. & Svergun, D. I. GNOM. A program package for small-angle scattering data processing. *J. Appl. Crystallogr.* **24**, 537–540 (1991).

356. Tria, G., Mertens, H. D. T., Kachala, M. & Svergun, D. I. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* **2**, 207–217 (2015).

357. Koenigsberg, A. L., Heldwein, E. E. & Heldwein, E. E. Biochemical and structural characterization of PRV UL37. (2018). doi:10.1074/jbc.RA118.004481

358. Da Silva, V. M. *et al.* Modular hyperthermostable bacterial endo-β-1, 4-mannanase: Molecular shape, flexibility and temperature-dependent conformational changes. *PLoS One* **9**, (2014).

359. Schneidman-Duhovny, D., Hammel, M. & Sali, A. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* **38**, 540–544 (2010).

360. Schneckenburger, H. Total internal reflection fluorescence microscopy: Technical innovations and novel applications. *Curr. Opin. Biotechnol.* **16**, 13–18 (2005).

361. Gordon, M. P., Ha, T. & Selvin, P. R. Single-molecule high-resolution imaging with photobleaching. *Proc. Natl. Acad. Sci.* **101**, 6462–6465 (2004).

362. Forster, B. Y. T. H. 10Th Spiers Memorial Lecture. *Discuss. Faraday Soc.* **27**, 7–17 (1959).

363. Jares-Erijman, E. A. & Jovin, T. M. FRET imaging. *Nat. Biotechnol.* **21**, 1387–1395 (2003).

364. Rayleigh. XXXI. *Investigations in optics, with special reference to the spectroscope*. *Philos. Mag. Ser. 5* **8**, 261–274 (1879).

365. Moerner, W. E. Microscopy beyond the diffraction limit using actively controlled single molecules. *J. Microsc.* **246**, 213–220 (2012).

366. Small, A. & Stahlheber, S. Fluorophore localization algorithms for super-resolution microscopy. *Nat. Methods* **11**, 267–279 (2014).

367. Cordes, T., Vogelsang, J. & Tinnefeld, P. On the mechanism of trolox as antiblinking and antibleaching reagent. *J. Am. Chem. Soc.* **131**, 5018–5019 (2009).

368. Alberto, M. E., Russo, N., Grand, A. & Galano, A. A physicochemical examination of the free radical scavenging activity of Trolox: Mechanism, kinetics and influence of the environment. *Phys. Chem. Chem. Phys.* **15**, 4642–4650 (2013).

369. Abràmoff, M. D., Hospitals, I., Magalhães, P. J. & Abràmoff, M. <ImageJ.pdf>. doi:10.1201/9781420005615.ax4

370. Mashanov, G. I. & Molloy, J. E. Automatic detection of single fluorophores in live cells. *Biophys. J.* **92**, 2199–2211 (2007).

371. Freedman, D. & Diaconis, P. On the Histogram as a Density Estimator: L2 Theory. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* **57**, 453–476 (1981).

372. Walpole, R. E., Myers, R. H., Myers, S. L. & Ye, K. *Probability & Statistics for Engineers & Scientists*. (Pearson, 2012).

373. Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J. M. & Gaub, H. E. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science (80-.).* **276**, 1109–1112 (1997).

374. Yarawsky, A. E., English, L. R., Whitten, S. T. & Herr, A. B. The Proline/Glycine-Rich Region Of The Biofilm Adhesion Protein Aap Forms An Extended Stalk That Resists Compaction. *J. Mol. Biol.* (2016). doi:10.1016/j.jmb.2016.11.017

375. Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41**, 349–357 (2013).

376. Benkert, P., Biasini, M. & Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27**, 343–350 (2011).

377. Waterhouse, A. *et al.* SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).

378. Liu, Q. *et al.* The Enterococcus faecalis MSCRAMM ace binds its ligand by the collagen hug model. *J. Biol. Chem.* **282**, 19629–19637 (2007).

379. Ylänne, J., Scheffzek, K., Young, P. & Saraste, M. Crystal Structure of the Alpha-Actinin Rod Reveals an Extensive Torsional Twist. *Structure* **9**, 597–604 (2001).

380. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

381. Madden, T. L. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

382. Hawker, L. E. & Linton, A. H. Chapter 9. Structure, biology and classification of prokaryotic micro-organisms. in *Micro-organisms: function, form and environment* 274–351 (1972).

383. Harrison, M. *et al.* Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC Genomics* **10**, 291 (2009).

384. Hecht, O. *et al.* Self-recognition by an intrinsically disordered protein. *FEBS Lett.* **582**, 2673–2677 (2008).

385. Terpe, K. Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* **60**, 523–533 (2003).

386. Conrady, D. G. *et al.* A zinc-dependent adhesion module is responsible for

intercellular adhesion in staphylococcal biofilms. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 19456–19461 (2008).

387. Conrady, D. G., Wilson, J. J. & Herr, A. B. Structural basis for Zn2+-dependent intercellular adhesion in staphylococcal biofilms. *Proc. Natl. Acad. Sci.* **110**, E202–E211 (2013).

388. Greenfield, N. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc.* **1**, 2876–2890 (2007).

389. Benjwal, S., Verma, S., Rohm, K.-H. & Gursky, O. Monitoring protein aggregation during thermal unfolding in circular dichroism experiments. *Protein Sci.* **15**, 635–639 (2006).

390. Gursky, O. & Atkinson, D. High- and low-temperature unfolding of human high-density apolipoprotein A-2. *Protein Sci.* **5**, 1874–82 (1996).

391. Schaub, L. J., Campbell, J. C. & Whitten, S. T. Thermal unfolding of the N-terminal region of p53 monitored by circular dichroism spectroscopy. *Protein Sci.* **21**, 1682–1688 (2012).

392. Hamadi, F. *et al.* Effect of pH on distribution and adhesion of Staphylococcus aureus to glass. *J. Adhes. Sci. Technol.* **19**, 73–85 (2005).

393. Jones, E. M., Cochrane, C. A. & Percival, S. L. The Effect of pH on the Extracellular Matrix and Biofilms. *Adv. wound care* **4**, 431–439 (2015).

394. Zmantar, T., Kouidhi, B., Miladi, H., Mahdouani, K. & Bakhrouf, A. A Microtiter plate assay for staphylococcus aureus biofilm quantification at various pH levels and hydrogen peroxide supplementation. *New Microbiol.* **33**, 137–145 (2010).

395. Strong, M. *et al.* Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci.* **103**, 8060–8065 (2006).

396. Morris, B. R. J., Perrakis, A. & Lamzin, V. S. [11] ARP / wARP and Automatic Interpretation of Protein Electron Density Maps. *Methods* **374**, (2003).

397. Speicher, D. W., Weglarz, L. & DeSilva, T. M. Properties of human red cell spectrin heterodimer (side-to-side) assembly and identification of an essential nucleation site. *J. Biol. Chem.* **267**, 14775–14782 (1992).

398. Ortega, E. *et al.* The structure of the plakin domain of plectin reveals an extended rod-like shape. *J. Biol. Chem.* **291**, 18643–18662 (2016).

399. George, R. & Heringa, J. An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng.* **15**, 871–9 (2002).

400. Jones, S., Marin, a & Thornton, J. M. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.* **13**, 77–82 (2000).

401. Zhou, R., Huang, X., Margulis, C. J. & Berne, B. J. Hydrophobic Collapse in Multidomain Protein Folding. *Science (80-.).* **305**, 1605–1609 (2004).

402. Kohn, W. D., Kay, C. M. & Hodges, R. S. Salt effects on protein stability: Two-stranded ⍺-helical coiled- coils containing inter- or intrahelical ion pairs. *J. Mol. Biol.* **267**, 1039–1052 (1997).

403. Zhang, Y. & Cremer, P. S. Interactions between macromolecules and ions: the Hofmeister series. *Curr. Opin. Chem. Biol.* **10**, 658–663 (2006).

404. Mitchell, A. L. *et al.* InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).

405. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: Advances and challenges. *Front. Microbiol.* **5**, 1–17 (2014).

406. Seyit, G., Rockel, B., Baumeister, W. & Peters, J. Size matters for the

tripeptidylpeptidase II complex from drosophila: The 6-MDa spindle form stabilizes the activated state. *J. Biol. Chem.* **281**, 25723–25733 (2006).

407. Ahsan, S. S., Chen, H., Santiago-Berrios, M. B., Abruna, H. D. & Webb, W. W. Quenching of Alexa Dyes by Amino Acids. *Biophys. J.* **98**, 583a (2010).

408. ThermoFisher Scientific. Alexa Fluor™ 488 C5 Maleimide. *Product Information* Available at: https://www.thermofisher.com/order/catalog/product/A10254. (Accessed: 27th February 2019)

409. Yiamsawas, D., Wagner, M., Baier, G., Landfester, K. & Wurm, F. R. Competing and simultaneous click reactions at the interface and in solution. *RSC Adv.* **6**, 51327–51331 (2016).

410. Kim, Y. *et al.* Efficient Site-Specific Labeling of Proteins via Cysteines. *Bioconj. Chem.* **19**, 786–791 (2008).

411. Xue, Y., Li, X., Li, H. & Zhang, W. Quantifying thiol-gold interactions towards the efficient strength control. *Nat. Commun.* **5**, 1–9 (2014).

412. Reiber, D. C., Brown, R. S., Weinberger, S., Kenny, J. & Bailey, J. Unknown Peptide Sequencing Using Matrix-Assisted Laser Desorption / Ionization and In-Source Decay. *Anal. Chem.* **70**, 1214–1222 (1998).

413. Skeene, K. *et al.* One Filter, One Sample, and the N- and O-Glyco(proteo)me: Toward a System to Study Disorders of Protein Glycosylation. *Anal. Chem.* **89**, 5840–5849 (2017).

414. Sechi, S. & Chait, B. T. Modification of cysteine residues by alkylation. A tool in peptide mapping and protein identification. *Anal. Chem.* **70**, 5150–5158 (1998).

415. Smilgies, D. M. & Folta-Stogniew, E. Molecular weight-gyration radius relation of globular proteins: A comparison of light scattering, small-angle X-ray scattering and structure-based data. *J. Appl. Crystallogr.* **48**, 1604–1606 (2015).

416. Su, R., Qi, W., He, Z., Zhang, Y. & Jin, F. Multilevel structural nature and interactions of bovine serum albumin during heat-induced aggregation process. *Food Hydrocoll.* **22**, 995–1005 (2008).

417. Stetefeld, J., McKenna, S. A. & Patel, T. R. Dynamic light scattering: a practical guide and applications in biomedical sciences. *Biophys. Rev.* **8**, 409–427 (2016).

418. Svergun, D. I. Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.* **25**, 495–503 (1992).

419. Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. Determination of domain structure of proteins from x-ray solution scattering. *Biophys. J.* **80**, 2946–2953 (2001).

420. Petoukhov, M. V. *et al.* New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45**, 342–350 (2012).

421. Svergun, D., Barberato, C. & Koch, M. H. CRYSOL - A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).

422. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* **44**, W424–W429 (2016).

423. Yao, M. *et al.* The mechanical response of talin. *Nat. Commun.* **7**, 1–11 (2016).

424. Lenne, P., Raae, A. J., Altmann, S. M., Saraste, M. & Ho, J. K. H. States and transitions during forced unfolding of a single spectrin repeat. **476**, 124–128 (2000).

425. Kim, M. *et al.* Fast and forceful refolding of stretched α-helical solenoid proteins. *Biophys. J.* **98**, 3086–3092 (2010).

426. Randles, L. G., Rounsevell, R. W. S. & Clarke, J. Spectrin Domains Lose Cooperativity in Forced Unfolding. *Biophys. J.* **92**, 571–577 (2007).

427. Li, L., Wetzel, S., Plückthun, A. & Fernandez, J. M. Stepwise unfolding of ankyrin repeats in a single protein revealed by atomic force microscopy. *Biophys. J.* **90**, 30–32 (2006).

428. Bloomfield, V. A., Crothers, D. M., Tinoco Jr, I. *Size and shape of nucleic acids in solution*. (University Science Books, 2000).

429. Weast, R. C. *Handbook of Chemistry and Physics*.

430. Berg, J. M., Tymoczko, J. L. & Stryer, L. *Biochemistry*. (W. H. Freeman and Company, 2006).

431. Bujalowski, P. J. & Oberhauser, A. F. Tracking unfolding and refolding reactions of single proteins using atomic force microscopy methods. *Methods* **60**, 151–160 (2013).

432. Marsh, P. D. Dental plaque as a microbial biofilm. *Caries Res.* **38**, 204–211 (2004).

433. Sulikowski, G. A. *et al.* A Structural Model for Binding of the Serine-Rich Repeat Adhesin GspB to Host Carbohydrate Receptors. *PLoS Pathog.* **7**, e1002112 (2011).

434. McNab, R. *et al.* Cell wall-anchored csha polypeptide (259 kilodaltons) in streptococcus gordonii forms surface fibrils that confer hydrophobic and adhesive properties. *J. Bacteriol.* **181**, 3087–3095 (1999).

435. Takahashi, Y. *et al.* Contribution of sialic acid-binding adhesin to pathogenesis of experimental endocarditis caused by Streptococcus gordonii DL1. *Infect. Immun.* **74**, 740–743 (2006).

436. J., K. *et al.* Co-expression of colligin and collagen in oral submucous fibrosis: Plausible role in pathogenesis. *Oral Oncol.* **37**, 282–287 (2001).

437. Kamath, V. The nature of collagen in oral submucous fibrosis: A systematic review of the literature. *Saudi J. Oral Sci.* **1**, 57 (2014).

438. Williamson, M. P. The structure and function of proline-rich regions in proteins. *Biochem. J.* **297**, 249–260 (1994).

439. George, R. A. & Heringa, J. An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng. Des. Sel.* **15**, 871–879 (2003).

440. Madoff, L. C., Michel, J. L., Gong, E. W., Kling, D. E. & Kasper, D. L. Group B streptococci escape host immunity by deletion of tandem repeat elements of the alpha C protein. *Proc. Natl. Acad. Sci.* **93**, 4131–4136 (1996).

441. Lindahl, G., Stålhammar-Carlemalm, M. & Areschoug, T. Surface proteins of Streptococcus agalactiae and related proteins in other bacterial pathogens. *Clin. Microbiol. Rev.* **18**, 102–27 (2005).

442. Argos, P. An investigation of oligopeptides linking domains in protein tertiary structures and possible candidates for general gene fusion. *J. Mol. Biol.* **211**, 943–958 (1990).

443. Dieckmann, R., Pavela-Vrancic, M., Von Döhren, H. & Kleinkauf, H. Probing the domain structure and ligand-induced conformational changes by limited proteolysis of tyrocidine synthetase 1. *J. Mol. Biol.* **288**, 129–140 (1999).

444. Robinson, C. R. & Sauer, R. T. Optimizing the stability of single-chain proteins by linker length and composition mutagenesis. *Proc. Natl. Acad. Sci.* **95**, 5929–5934 (1998).

445. Wishart, D. S., Sykes, B. D. & Richards, F. M. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J. Mol. Biol.* **222**, 311–333 (1991).

446. Biological Magnetic Resonance Data Bank. BMRB Chemical Shift Statistics.

Statistics on chemical shifts from atoms in amino acids (2019). Available at: http://www.bmrb.wisc.edu/ref_info/. (Accessed: 12th March 2019)

447. Grzesiek, S. & Bax, A. Correlating Backbone Amide and Side Chain Resonances in Larger Proteins by Multiple Relayed Triple Resonance NMR. *J. Am. Chem. Soc.* **114**, 6291–6293 (1992).

448. Grzesiek, S. & Bax, A. An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. *J. Magn. Reson.* **99**, 201–207 (1992).

449. Banci, L. *et al.* Assignment of backbone NMR resonances and secondary structural elements of a reduced monomeric mutant of copper/zinc superoxide dismutase. *Magn. Reson. Chem.* **35**, 845–853 (1997).

450. Banci, L. *et al.* Solution Structure of Reduced Monomeric Q133M2 Copper, Zinc Superoxide Dismutase (SOD). Why Is SOD a Dimeric Enzyme? *Biochemistry* **37**, 11780–11791 (1998).

451. Banci, L. *et al.* Backbone Dynamics of Human Cu,Zn Superoxide Dismutase and of Its Monomeric F50E/G51E/E133Q Mutant: The Influence of Dimerization on Mobility and Function †. *Biochemistry* **39**, 9108–9118 (2000).

452. Arunkumar, A. I., Stauffer, M. E., Bochkareva, E., Bochkarev, A. & Chazin, W. J. Independent and Coordinated Functions of Replication Protein A Tandem High Affinity Single-stranded DNA Binding Domains. *J. Biol. Chem.* **278**, 41077–41082 (2003).

453. Arold, S. T. *et al.* The Role of the Src Homology 3-Src Homology 2 Interface in the Regulation of Src Kinases. *J. Biol. Chem.* **276**, 17199–17205 (2001).

454. Wriggers, W., Chakravarty, S. & Jennings, P. A. Control of protein functional dynamics by peptide linkers. *Curr. Trends Pept. Sci.* **80**, 736–746 (2005).

455. Oman, T. J., Boettcher, J. M., Wang, H., Okalibe, X. N. & Donk, W. A. Van Der.

Sublancin is not a lantibiotic but an S-linked glycopeptide. *Nat. Chem. Biol.* **7**, 1–3 (2011).

456. Stepper, J. *et al.* Cysteine S-glycosylation, a new post-translational modification found in glycopeptide bacteriocins. *FEBS Lett.* **585**, 645–650 (2011).

457. Todorov, S. D. Bacteriocins from Lactobacillus plantarum - production, genetic organization and mode of action. *Brazilian J. Microbiol.* **40**, 209–221 (2009).

458. Gräslund, S. *et al.* Protein production and purification. *Nat. Methods* **5**, 135–146 (2008).

459. Wu, J. & Filutowicz, M. Hexahistidine (His6)-tag dependent protein dimerization: a cautionary tale. *Acta Biochim. Pol.* **46**, 591–599 (1999).

460. Han, J., Kerrison, N., Chothia, C. & Teichmann, S. A. Divergence of Interdomain Geometry in Two-Domain Proteins. *Structure* **14**, 935–945 (2006).

461. Goult, B. T. *et al.* The structure of an interdomain complex that regulates Talin activity. *J. Biol. Chem.* **284**, 15097–15106 (2009).

462. Castelmur, E. Von *et al.* A regular pattern of Ig super-motifs defines segmental flexibility as the elastic mechanism of the titin chain. 1–6 (2008).

463. Kang, H. J. & Baker, E. N. Intramolecular isopeptide bonds: Protein crosslinks built for stress? *Trends Biochem. Sci.* **36**, 229–237 (2011).

464. Cramer, J. F., Nordberg, P. A., Hajdu, J. & Lejon, S. Crystal structure of a bacterial albumin-binding domain at 1.4 Å resolution. *FEBS Lett.* **581**, 3178–3182 (2007).

465. Schneider, J. P., Lombardi, A. & DeGrado, W. F. Analysis and design of three-stranded coiled coils and three-helix bundles. *Fold. Des.* **3**, 29–40 (1998).

466. Kingston, R. L., Gay, L. S., Baase, W. S. & Matthews, B. W. Structure of the Nucleocapsid-Binding Domain from the Mumps Virus Polymerase; an Example of Protein Folding Induced by Crystallization. *J. Mol. Biol.* **379**, 719–731 (2008).

467. Berman, H. M. *et al.* The Protein Data Bank Helen. *Nucleic Acids Res.* **28**, 235–242 (2000).

468. Holm, L. & Sander, C. Protein Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.* **233**, 123–138 (1993).

469. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res.* **44**, W351–W355 (2016).

470. Achila, D. *et al.* Structural determinants of host specificity of complement Factor H recruitment by *Streptococcus pneumoniae*. *Biochem. J.* **465**, 325–335 (2015).

471. Holm, L. & Rosenström, P. Dali server: Conservation mapping in 3D. *Nucleic Acids Res.* **38**, 545–549 (2010).

472. Djinović-Carugo, K., Young, P., Gautel, M. & Saraste, M. Structure of the α-actinin rod: Molecular basis for cross-linking of actin filaments. *Cell* **98**, 537–546 (1999).

473. Bateman, A. & Bycroft, M. The Structure of a LysM Domain from E . coli Membrane-bound Lytic Murein Transglycosylase D (MltD). 1113–1119 (2000). doi:10.1006/jmbi.2000.3778

474. Kong, L. Delineation of modular proteins: Domain boundary prediction from sequence information. *Brief. Bioinform.* **5**, 179–192 (2004).

475. Protein Calculator v3.4. (2013). Available at: http://protcalc.sourceforge.net/. (Accessed: 6th February 2019)

476. Podestà, A. *et al.* Positively Charged Surfaces Increase the Flexibility of DNA. *Biophys. J.* **89**, 2558–2563 (2005).

477. Heidorn, D. B. & Trewhell, J. Comparison of the Crystal and Solution Structures of Calmodulin and Troponin C. *Biochemistry* **27**, 909–915 (1988).

478. Gouda, H. *et al.* Three-Dimensional Solution Structure of the B Domain of Staphylococcal Protein A: Comparisons of the Solution and Crystal Structures.

Biochemistry **31**, 9665–9672 (1992).

479. Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. X-ray solution scattering (SAXS) combined with crystallography and computation: Defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* **40**, 191–285 (2007).

480. Pinho, M. G. & Errington, J. Dispersed mode of Staphylococcus aureus cell wall synthesis in the absence of the division machinery. *Mol. Microbiol.* **50**, 871–881 (2003).

481. Jorge, A. M., Hoiczyk, E., Gomes, J. P. & Pinho, M. G. EzrA contributes to the regulation of cell size in Staphylococcus aureus. *PLoS One* **6**, (2011).

482. Matias, V. R. F. & Beveridge, T. J. Native cell wall organization shown by cryo-electron microscopy confirms the existence of a periplasmic space in Staphylococcus aureus. *J. Bacteriol.* **188**, 1011–1021 (2006).

483. Xia, G., Kohler, T. & Peschel, A. The wall teichoic acid and lipoteichoic acid polymers of Staphylococcus aureus. *Int. J. Med. Microbiol.* **300**, 148–154 (2010).

484. Reichmann, N. T. & Gründling, A. Location, synthesis and function of glycolipids and polyglycerolphosphate lipoteichoic acid in Gram-positive bacteria of the phylum Firmicutes. *FEMS Microbiol. Lett.* **319**, 97–105 (2011).

485. Wanner, S. *et al.* Wall teichoic acids mediate increased virulence in Staphylococcus aureus. *Nat. Microbiol.* **2**, (2017).

486. Swoboda, J. G., Campbell, J., Meredith, T. C. & Walker, S. Wall teichoic acid function, biosynthesis, and inhibition. *ChemBioChem* **11**, 35–45 (2010).

487. Brown, S., Santa Maria, J. P. & Walker, S. Wall teichoic acids of gram-positive bacteria. *Annu. Rev. Microbiol.* **67**, 313–36 (2013).

488. Ackbarow, T., Keten, S. & Buehler, M. J. A multi-timescale strength model of

alpha-helical protein domains. *J. Phys. Condens. Matter* **21**, 1–6 (2009).

489. Menhart, N., Mitchell, T., Lusitani, D., Topouzian, N. & Fung, L. W. M. Peptides with more than one 106-amino acid sequence motif are needed to mimic the structural stability of spectrin. *J. Biol. Chem.* **271**, 30410–30416 (1996).

490. Ortiz, V., Nielsen, S. O., Klein, M. L. & Discher, D. E. Unfolding a linker between helical repeats. *J. Mol. Biol.* **349**, 638–647 (2005).

491. Michaely, P., Tomchick, D. R., Machius, M. & Anderson, R. G. W. Crystal structure of a 12 ANK repeat stack from human ankyrinR. *EMBO J.* **21**, 6387–6396 (2002).

492. Lee, G. *et al.* Nanospring behaviour of ankyrin repeats. *Nature* **440**, 246–249 (2006).

493. Valbuena, A., Vera, A. M., Oroz, J., Menéndez, M. & Carrión-Vázquez, M. Mechanical properties of β-catenin revealed by single-molecule experiments. *Biophys. J.* **103**, 1744–1752 (2012).

494. Best, R. B. *et al.* Mechanical unfolding of a titin Ig domain: structure of transition state revealed by combining atomic force microscopy, protein engineering and molecular dynamics simulations. *J. Mol. Biol.* **330**, 867–877 (2003).

495. Li, Q., Scholl, Z. N. & Marszalek, P. E. Unraveling the Mechanical Unfolding Pathways of a Multidomain Protein: Phosphoglycerate Kinase. *Biophys. J.* **115**, 46–58 (2018).

496. Alsteens, D., Martinez, N., Jamin, M. & Jacob-Dubuisson, F. Sequential Unfolding of Beta Helical Protein by Single-Molecule Atomic Force Microscopy. *PLoS One* **8**, 1–10 (2013).

497. Sikora, M., Sulkowska, J. I., Witkowski, B. S. & Cieplak, M. BSDB: The biomolecule stretching database. *Nucleic Acids Res.* **39**, (2011).

498. Marszalek, P. E. *et al.* Mechanical unfolding intermediates in titin modules.

Nature **402**, 100–103 (1999).

499. Thomas, W. E., Trintchina, E., Forero, M., Vogel, V. & Sokurenko, E. V. Bacterial adhesion to target cells enhanced by shear force. *Cell* **109**, 913–23 (2002).

500. Hart, J. W., Waigh, T. A., Lu, J. R. & Roberts, I. S. Microrheology and Spatial Heterogeneity of Staphylococcus aureus Biofilms Modulated by Hydrodynamic Shear and Biofilm-Degrading Enzymes. *Langmuir* **35**, 3553–3561 (2019).

501. Moormeier, D. E., Bose, J. L., Horswill, A. R. & Bayles, K. W. Temporal and Stochastic Control of Staphylococcus aureus Biofilm Development. *MBio* **5**, 1–12 (2014).

502. Carrion-Vazquez, M. *et al.* Mechanical design of proteins-studied by single-molecule force spectroscopy and protein engineering. *Prog Biophys Mol Biol* **74**, 63–91 (2003).

503. Herman-Bausier, P. *et al.* Mechanical Strength and Inhibition of the Staphylococcus aureus. *MBio* **7**, 1–11 (2016).

504. Madani, A., Garakani, K. & Mofrad, M. R. K. Molecular mechanics of Staphylococcus aureus adhesin, CNA, and the inhibition of bacterial adhesion by stretching collagen. *PLoS One* **12**, 1–19 (2017).

505. Aksel, T., Majumdar, A. & Barrick, D. The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. *Structure* **19**, 349–360 (2011).

506. Law, R. *et al.* Cooperativity in Forced Unfolding of Tandem Spectrin Repeats. *Biophys. J.* **84**, 533–544 (2003).

507. Brush, S. G. History of the Lenz-Ising model. *Rev. Mod. Phys.* **39**, 883–893 (1967).

508. Rief, M., Gautel, M., Schemmel, A. & Gaub, H. E. The mechanical stability of immunoglobulin and fibronectin III domains in the muscle protein titin measured

by atomic force microscopy. *Biophys. J.* **75**, 3008–3014 (1998).

509. Zweifel, M. E., Leahy, D. J., Hughson, F. M. & Barrick, D. Structure and stability of the ankyrin domain of the Drosophila Notch receptor. *Protein Sci.* **12**, 2622–2632 (2003).

510. Tang, K. S., Guralnick, B. J., Wang, W. K., Fersht, A. R. & Itzhaki, L. S. Stability and Folding of the Tumour Suppressor Protein p16. *J. Mol. Biol.* **285**, 1869–1886 (1999).

511. Main, E. R. G., Lowe, A. R., Mochrie, S. G. J., Jackson, S. E. & Regan, L. A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Curr. Opin. Struct. Biol.* **15**, 464–471 (2005).

512. Kloss, E., Courtemanche, N. & Barrick, D. Repeat-protein folding: New insights into origins of cooperativity, stability, and topology. *Arch. Biochem. Biophys.* **469**, 83–99 (2008).

513. Markelz, A. Protein Dynamical Transition Does Not Require Protein Structure. (2015). doi:10.1103/PhysRevLett.101.178103

514. Lipari, G. & Szabo, A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 2. Analysis of experimental results. *J. Am. Chem. Soc.* **104**, 4559–4570 (1982).

515. Clore, G. M., Driscoll, P. C., Wingfield, P. T. & & Gronenborn, A. Analysis of the backbone dynamics of interleukin 1-b using 2D inverse detected heteronuclear 15N-1H NMR spectroscopy. *Biochemistry* **29**, 7387–7401. (1990).

516. Hansen, A. P., Petros, A. M., Meadows, R. P. & Fesik, S. W. Backbone Dynamics of a Two-Domain Protein: I5N Relaxation Studies of the Amino-Terminal Fragment of Urokinase-Type Plasminogen Activator. *Biochemistry* **33**, 15418–15424 (1994).

517. Hansen, A. P. *et al.* Solution structure of the amino-terminal fragment of

urokinase-type plasminogen activator. *Biochemistry* **33**, 4847–4864 (1994).

518. Henderson, C. E. *et al.* Solution structure and dynamics of the central CCP module pair of a poxvirus complement control protein. *J. Mol. Biol.* **307**, 323–339 (2002).

519. Gallagher, C., Burli, F., Offer, J. & Ramos, A. A method for the unbiased and efficient segmental labelling of RNA-binding proteins for structure and biophysics. *Sci. Rep.* **7**, 1–9 (2017).

520. Dawson, P. E., Muir, T. W., Clark-Lewis, I. & Kent, S. B. Synthesis of proteins by native chemical ligation. *Science (80-.).* **266**, 776–779 (1994).

521. d'Auvergne, E. J. & Gooley, P. R. Optimisation of NMR dynamic models I. Minimisation algorithms and their performance within the model-free and Brownian rotational diffusion spaces. *J. Biomol. NMR* **40**, 107–119 (2008).

522. d'Auvergne, E. J. & Gooley, P. R. Optimisation of NMR dynamic models II. A new methodology for the dual optimisation of the model-free parameters and the Brownian rotational diffusion tensor. *J. Biomol. NMR* **40**, 121–133 (2008).