Structural studies of pathogenicity-related proteins in *Clostridium difficile*

Adam Daniel Crawshaw

Supervisors:

Dr Paula Salgado

Professor Rick Lewis

A thesis submitted for the degree of

Doctor of Philosophy

Institute for Cell and Molecular Biosciences

September 2016

Abstract

Clostridium difficile is a Gram-positive obligate anaerobic pathogen that causes debilitating infections which can ultimately be fatal. The life cycle of *C. difficile* from gut colonisation in vegetative cell form, to survival in the aerobic environment as spores is not fully understood.

Sporulation is a cell-cycle stress response that produces a daughter cell by asymmetric division. This forespore is engulfed by the mother-cell and matured before release into the environment. The SpoIIQ-SpoIIIAH complex has been proposed to enable communication between the two cells during this process. The aim of this work was to characterise the inter-sporangial domains of the complex in *C. difficile*. Localised to the forespore membrane, SpoIIQ was shown to bind a Zn^{2+} within a conserved LytM endopeptidase domain. Furthermore, it was demonstrated that Zn^{2+} is essential to form a stable interaction with SpoIIIAH, located in the mother cell membrane, a role not previously observed in homologous SpoIIQ-SpoIIIAH complexes.

Cell-surface adhesion is an important factor in gut colonisation. Type IV pili (TFP) form filamentous protein appendages that extended from the cell wall into the environment and have recently been recognised in Gram-positive bacteria. In *C. difficile*, TFP enable twitching-motility. This work aimed to determine the structures of major and minor pilins in TFP filaments in *C. difficile*. The crystal structures of the major pilin, PilA1, were determined from two strains of *C. difficile*. These exhibit similarities to pseudo-pilins from Gram-negative bacteria. Although crystals of a minor pilin, PilK, were obtained, structure determination was so far unachievable. Investigations of potential interactions between major and minor pilins were performed suggesting they may interact, presumably to form the complete pilus.

This work contributes to the understanding of proteins involved in sporulation and colonisation, two key mechanisms in *C. difficile* pathogenicity.

Acknowledgements

I would like to thank my supervisors Dr Paula Saglado and Professor Rick Lewis for their support and advice throughout my PhD. I would also like to thank our collaborators, Professor Adriano Henriques and Professor Neil Fairweather and their groups for their valuable input. Additionally, I would like to express my gratitude to Gwyndaf Evans, Anna Warren and Pierre Aller at Diamond Light Source who guided me through my internship. I am grateful to both current and past members of the Salgado, Waldron and Lewis labs for both their moral and scientific support, in particular; Marcin Dembek, Abbie Kelly, Emma Tarrant, Anna Barwinksa-Sendra, Gus Pelicioli-Riboldi, Jack Stevenson, Kevin Waldron, Vincent Rao, Rob Cleverly, Jon Marles-Wright and Lorraine Hewitt. A special thank you to Arnaud Basle for his endless support and the many pints. I am also especially grateful to Orla Dunne for the encouragement and tough love. Finally, I'd like to thank all of my friends and most importantly, I'd like to thank my Mum and my brother Richard for their endless support, encouragement and listening over the past 4 years.

Contents

List of Figures

List of Tables

Nomenclature

ABC Dimethylamine-borane complex

ASU Asymmetric unit

BCA Bicinchoninic acid

BSA Bovine serum albumin

c-di-GMP Cyclic diguanosine monophosphate

CC Correlation coefficient

CD Circular Dichroism

CDI *Clostridium difficile* infection

Cm Chloramphenicol

CSS Complex fomration significance score

DISTL Diffraction Image Screening Tool and Library

DMP SpoIID-SpoIIM-SpoIIP complex

DPFGSE Double pulsed field spin echo

EDC 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride

EM Electron microscopy

Fc1 SPR sensor chip flow cell 1

Fc2 SPR sensor chip flow cell 2

FOM Figure of merit

GMQE Global model quality estimation

HEPES 4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid, N-(2-Hydroxyethyl)piperazine-N´-(2-ethanesulfonic acid)

HMW High molecular weight

HSQC Heteronuclear single quantum coherence

IPTG Isopropyl β-D-1-thiogalactopyranoside

Kan Kanamycin

kpsi Kilo pounds per square inch

LB Lysogeny Broth

LC/MS/MS Liquid chromatography/ mass spectrometry/ mass spectrometry

MBP Maltose binding protein

mdeg millidegrees

MRE Mean residue elipticity

MW Molecular weight

MWCO Molecular weight cut-off

NHS N-hydroxysuccinimide

NMR Nuclear magnetic resonance

NOESY Nuclear overhauser spectroscopy

OD Optical density

PAGE Polyacrylamide gel-electrophoresis

PCR Polymerase chain reaction

PDB Protein Databank

PISA Protein interfaces, surfaces and assemblies server

RMSD Root mean squared deviation

rpm Revolutions per minute

RU Response unit

S200 Superdex 200

S75 Superdex 75

SAD Single anomalous dispersion

SAD Single-wavelength anomalous dispersion

sAH SpoIIIAH29-229 construct, where the N-terminal His-tag has been removed.

SDS Sodium dodecyl sulphate

SEC Size-exclusion chromatography

sQ SpoIIQ31-222 construct, where the N-terminal His-tag has been removed.

sQ^{H120S} SpoIIQ31-222 construct where Histidine 120 has been mutated to serine

TFP Type IV pili or pilins

TFZ Translation function Z-score

TIISS Type II secretion system

Tm Melting temperature

UV Ultra violet

YT Yeast tryptone

Chapter 1

Introduction

1.1 *Clostridium difficile*

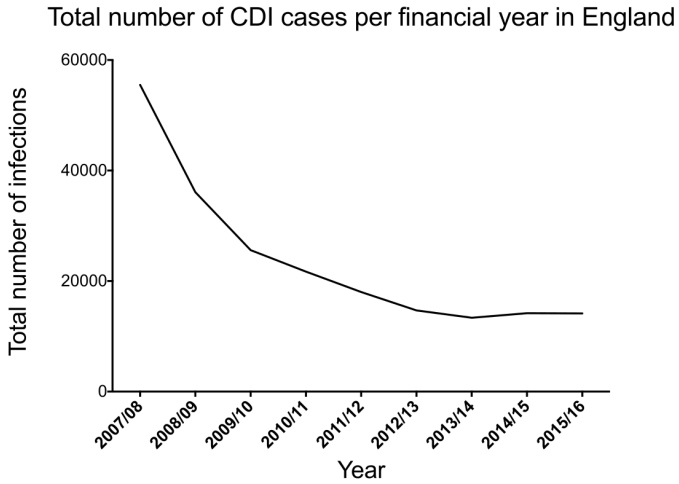


Figure 1.1.1 – Incidence of *C. difficile* infection in England. Annual number of confirmed *C. difficile* infections in English hospitals between March 2007 and March 2016. Source: Office for National Statistics, *C. difficile*: annual epidemiological commentary, July 2016.

Clostridium difficile is a Gram-positive obligate anaerobic bacterium. As a pathogen, *C. difficile* was the cause of over 14,000 hospital-acquired infections in England during 2015 (Figure 1.1.1) and continues to be a burden on Western healthcare systems (Office for National Statistics, July 2016). Since 2008, the number of *C. difficile* infections has decreased (Figure 1.1.1) due to improved hospital infection controls and cleaning protocols (Gerding *et al.,* 2008). However, the number of cases and associated deaths seems to have plateaued in recent years, indicating that control measurements are unlikely to be

able to reduce the number of infections further (Figure 1.1.1). Importantly, new strains of *C. difficile* that have greater virulence are often identified as the cause of *C. difficile* infections (CDI) (Merrigan *et al.,* 2010). A problem in CDI treatment is that *C. difficile* displays resistance to the most commonly used antibiotics (Huang *et al.,* 2009). As the last antibiotic in the cupboard is reached for, there is a pressing need for new approaches to treating and preventing CDI. Some of these new approaches include treating CDI with populations of normal gut bacteria from healthy donors via faecal transplant (Khanna *et al.,* 2016).

While CDI is most often a hospital-associated disease that is predominantly reported in elderly individuals, there is an increasing emergence of community acquired CDI (Khanna and Gupta, 2014). Although there is little data available on the epidemiology of community-acquired infection, one study determined that infection was most prevalent in individuals aged between 30-40 years (Fellmeth *et al.,* 2010). It is increasingly clear that CDI affects a more diverse demographic than had previously been acknowledged and further work is needed to understand the life style of *C. difficile* in and outside of the clinical environment.

An opportunistic pathogen, *C. difficile* colonises the gut of individuals on antibiotic therapies, in whom the normal gut microbiota has been compromised (Kachrimanidou and Malisiovas, 2011). Through poorly understood mechanisms, *C. difficile* colonises the gut, where it expresses two enterotoxins, toxin A and toxin B. These toxins damage the lining of the gut causing severe diarrhoea, a hallmark symptom of CDI (Carter *et al.,* 2012). In serious cases, toxin damage can lead to pseudomembranous colitis, mega colon and sepsis that can ultimately prove fatal.

Even when successfully treated initially, CDI can often reoccur in individuals several times during their lifetime. In some of these cases, the previous choice of treatment is no longer effective and other options must be pursued. It is in these individuals that CDI can be most debilitating. The mechanisms of CDI recurrence are even less well understood, although it is believed that persistent populations that adapt to the therapeutics used are likely to be the underlying cause (Abt *et al.,* 2016). Only by understanding the basic lifestyle of *C. difficile,* throughout its life cycle, will new therapeutic targets be discovered.

There are many strains of *C. difficile* that display considerable sequence and virulence variability (Kurka *et al.,* 2014). The most studied *C. difficile* strain is 630, which was isolated from a patient in 1982 and represents a virulent, drug resistant strain although it is

not a reliable spore forming variant (Wüst *et al.,* 1982). Used widely as a lab strain, the 630 genome, which includes 1 circular chromosome and a plasmid, has been sequenced and annotated, containing over 4 million base pairs that encode 3,762 proteins (Sebaihia *et al.,* 2006; Eijk *et al.,* 2015). In the past decade, new 'hyper-virulent' strains have emerged such as R20291 (027/BI/NAP1), a strain that has a greater resistance to antibiotic treatment and greater ability to sporulate (Akerlund *et al.,* 2008; Burns *et al.,* 2011). R20291 has been responsible for recent outbreaks in hospitals in the UK (Cartman *et al.,* 2010; Merrigan *et al.,* 2010). The genome of R20291, which is also contained on a circular chromosome and a plasmid, has also been fully sequenced and annotated. R20291 contains 3505 protein-expressing genes on its chromosome (Stabler *et al.,* 2009).

1.1.1 Life cycle of *C. difficile*

The life cycle of *C. difficile* can be divided into three stages (Figure 1.1.2): sporulation; germination; colonisation. Since the aerobic environment is toxic to *C. difficile*, these stages can be categorised into anoxic (in vegetative form) and oxic (in spore form) periods. Anaerobic environments such as the human and animal gut are ideal for colonisation by *C. difficile* (Smits *et al.,* 2016). Populations of *C. difficile* that do not cause infection can reside within many individuals and it has been recognised that a balanced gut microbiota is an important factor for this and for the prevention of CDI (Buffie, 2013). Recently, new treatments have sought to repopulate the gut of *C. difficile* infected individuals with a normal gut flora because *C. difficile* colonies only thrive in a non-competitive environment such as in the compromised gut.

C. difficile forms biofilms rich in proteins, polysaccharides and DNA (Dapa and Unnikrishnan, 2013). These biofilms are a recognised virulence factor in CDI and are associated with infection and antibiotic resistance (Dapa and Unnikrishnan, 2013). A class of proteins that has been implicated in biofilm formation are the Type IV pili (TFP), which have only recently been identified in Gram-positive bacteria and appear to represent an important virulence factor (Dapa and Unnikrishnan, 2013; Maldarelli *et al.,* 2016).

During sporulation, a highly resistant dormant cell, known as a spore, is formed, containing the *C. difficile* chromosome and the components required for germination. This ability of *C. difficile* to sporulate is critical to its survival in aerobic environments and also

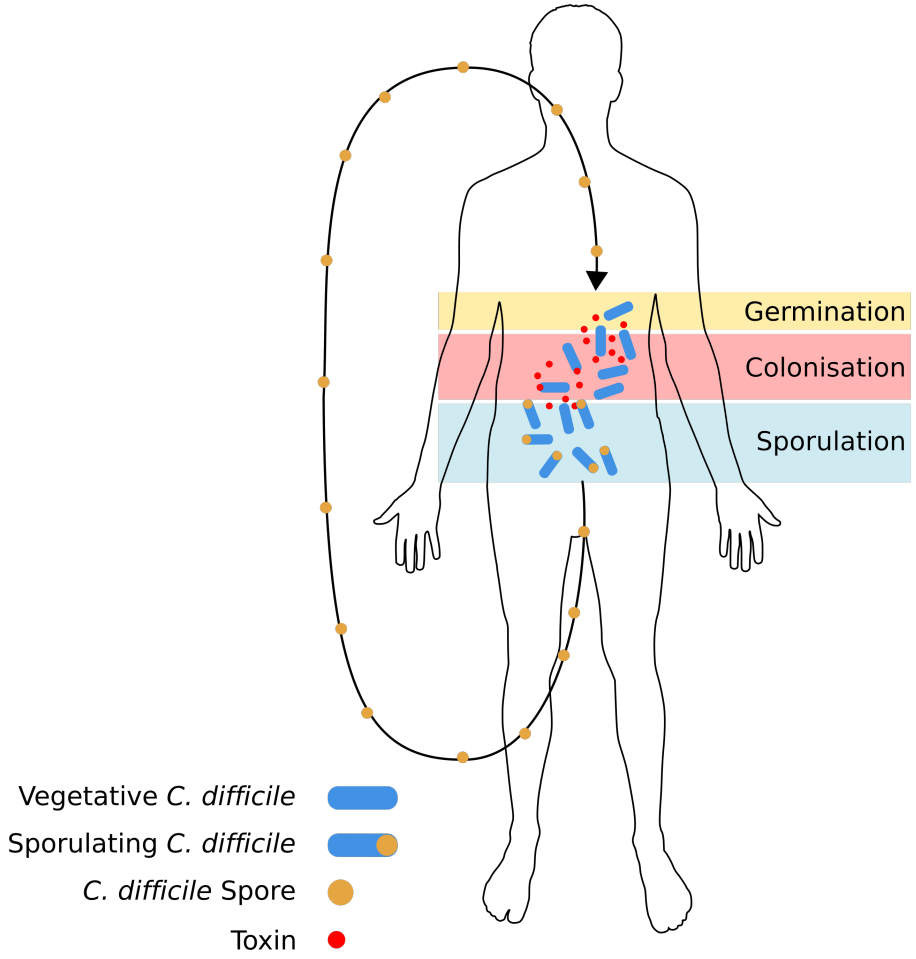


Figure 1.1.2 – Life cycle of *Clostridium difficile*. Life cycle of *C. difficile* in relation to the aerobic and anaerobic environments. Spores (yellow) are ingested and are stimulated to germinate by the anaerobic environment and other germinants such as bile salts. The vegetative cells (blue) colonise the compromised gut, forming biofilms and releasing enterotoxins (red) that disrupt the lining of the gut. Nutrient deprived or oxygen exposed cells sporulate in response to such stresses, producing highly resistant dormant spores which are released into the aerobic environment.

enables the transfer of CDI between individuals (Smits *et al.,* 2016). Sporulation can occur at any time during the life cycle of *C. difficile* and it is often recognised as stress-induced (Underwood *et al.,* 2009).

Once *C. difficile* spores are in a favourable, nutrient rich and anoxic environment, they undergo an understudied process known as germination. The product of germination is a viable vegetative cell that can replicate normally and form colonies as described above. Environment specific molecules such as bile salts or nutrients that can activate the germination process are known as germinants (Sorg and Sonenshein, 2008).

1.2 Sporulation

Sporulation is a cell cycle response to environmental stresses imposed on a cell such as nutrient deficiency or external factors such as an aerobic environment. The product of sporulation, a highly resistant dormant cell type, can tolerate heat, radiation and chemical assault (Setlow, 2006). The vegetative cell undergoes asymmetric division, producing a mother cell and a forespore, which is then engulfed by the mother cell in an endocytosis-like manner, isolating it from the external medium (Figure 1.2.1) (Hilbert and Piggot, 2004; Higgins and Dworkin, 2011). The forespore undergoes maturation, where the different spore layers are assembled. Finally, lysis of the mother cell releases the mature spore into the external environment (Errington, 2003).

Until recently, most of the understanding of sporulation was based in the Gram-positive model organism *Bacillus subtilis*. In *B. subtilis*, over 500 genes have been identified that are involved in the sporulation mechanism (Fawcett *et al.,* 2000). Gene expression must be coordinated between the mother cell and forespore, and this is performed by four sigma factors: σ^E; σ^F; σ^G; σ^K (Iber *et al.,* 2006). Sigma factors direct RNA polymerase to specific promoter sequences to direct transcription of small sub-groups of genes. Factors σ^F and σ^G are expressed and activate gene expression in the forespore whilst σ^E and σ^K are expressed and act in the mother cell (Figure 1.2.1) (Steil *et al.,* 2005). Some sigma factors are expressed in an inactive form with an N-terminal pro-peptide, which is cleaved by a protease to produce the active sigma factor (Errington, 2003). The master regulator of the sigma factor cascade is Spo0A (Fawcett *et al.,* 2000). As part of the stress re-

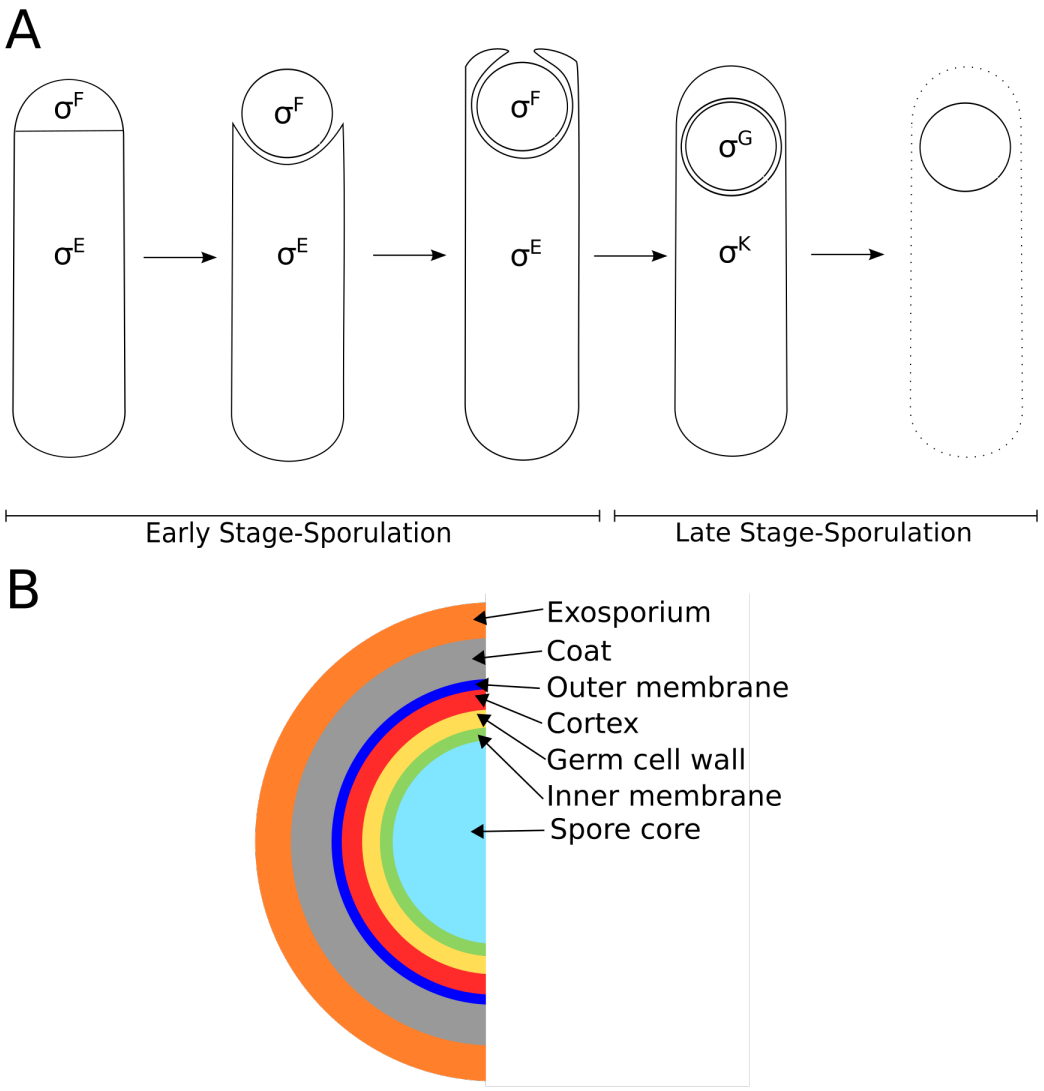


Figure 1.2.1 – Sporulation mechanism and spore layers. A: The stressed cell undergoes asymmetric division producing a daughter cell, known as a forespore, that is engulfed by the mother cell. The forespore is isolated from the external medium by a double membrane and undergoes maturation. Once the forespore is matured, the mother cell lyses releasing the mature spore into the environment. Gene expression is controlled during sporulation by compartment specific σ-factors that are divided into early stage (before completion of engulfment of the forespore) and late stage sporulation (post forespore engulfment). During early stage sporulation, gene expression is regulated by σ^F in the forespore and by σ^E in the mother cell and in late stage sporulation by σ^G in the forespore and σ^K in the mother cell. **B:** The mature spore has several layers that surround the spore core that contains the chromosome including an inner membrane, germ cell wall, cortex, outer membrane, coat layer and exosporium. (Paredes-Sabja *et al.,* 2014)

sponse, Spo0A is activated by a complex phosphorelay pathway that eventually results in the expression of 500 gene products and, the activation of the compartment-specific sigma factors, σ^F and σ^E (Eichenberger *et al.,* 2004; Camp *et al.,* 2011). Activation of σ^E results in the activation of the forespore factor, σ^G, finally leading to activation of σ^K in the mother cell, which is only activated once engulfment has taken place (Figure 1.2.2A) (Eichenberger *et al.,* 2003; Wang *et al.,* 2006). This relay between mother cell and forespore sigma factors allows the controlled division, engulfment and maturation of the forespore into a complete spore. Mutants of any of these sigma factors, particularly early stage sporulation factors (σ^F and σ^E), prevent completion of sporulation or the sporulation mechanism entirely.

As work extends to more clinically relevant bacteria such as *C. difficile* it is becoming apparent that the same rules do not fully apply. An example of this is the differing nature of the sigma-factor cascade that controls gene expression during sporulation (Pereira *et al.,* 2013; Fimlaid and Shen, 2015). Even though the sigma factors are conserved between *C. difficile* and *B. subtilis,* not all of the gene products under the control of these sigma factors are conserved (Stephenson and Lewis, 2005). It has recently been determined that there are different dependencies between the sigma factors downstream of σ^F (Figure 1.2.2B) (Pereira *et al.,* 2013; Fimlaid *et al.,* 2013). As in *B. subtilis,* but without the involvement of a complex phosphorelay pathway, Spo0A activates σ^F; σ^G is then activated by σ^F, however, unlike in *B. subtilis*, activation of σ^G is independent of whether σ^E is activated or not (Saujet *et al.,* 2013; Fimlaid *et al.,* 2013; Saujet *et al.,* 2014; Fimlaid and Shen, 2015). In *C. difficile*, σ^E activation is required for σ^K, activation which, is not strictly coupled to forespore engulfment as observed in *B. subtilis* (Saujet *et al.,* 2013; Fimlaid *et al.,* 2013; Saujet *et al.,* 2014; Fimlaid and Shen, 2015). The *C. difficile* mother cell factor, σ^K does not have an N-terminal pro-peptide element, is not found in an operon with cognate regulators, nor is its open reading frame interrupted with a phage-like element as in *B. subtilis*, and its mechanism of activation is thus unclear (Haraldsen and Sonenshein, 2003).

The differences in sigma factor regulation between *C. difficile* and *B. subtilis* result in temporal differences in gene activation. Specifically, genes expressed under the control of σ^G can be detected before engulfment is complete in *C. difficile*, but are exclusively found post-engulfment in *B. subtilis* (Pereira *et al.,* 2013; Saujet *et al.,* 2014). Importantly, it has

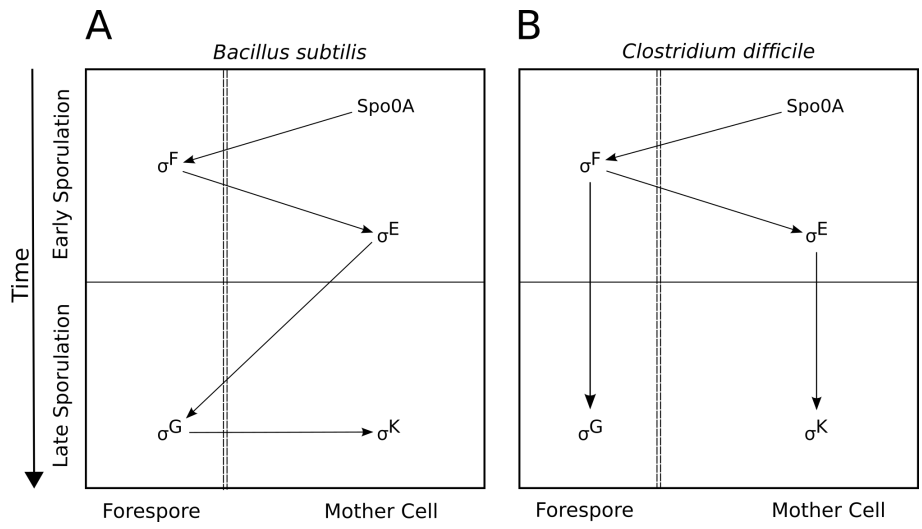


Figure 1.2.2 – Comparison of the sporulation gene cascades in *B. subtilis* and *C. difficile*. A: The sigma factors in *B. subtilis* activate in a sequential manner between the mother cell and forespore. Spo0A is the master regulator in the mother cell that activates σ^F in the forespore, in turn σ^E in the mother cell is then activated by σ^F up to and during engulfment. Post-engulfment σ^G is activated by expression of σ^E regulated genes and σ^G activation in the forespore results in the activation of σ^K in the mother cell. **B:** In *C. difficil*e, the master regulator Spo0A activates σ^F, however, σ^F activation then results in the activation of both σ^G in the forespore and σ^Ein the mother cell. Unlike in *B. subtilis*, σ^K activation is independent of σ^Gactivation in the forespore (Fimlaid *et al.,* 2013; Pereira *et al.,* 2013; Saujet *et al.,* 2014).

been shown in *C. difficile* that not only is Spo0A a master sporulation regulator but is also involved in toxin and biofilm production (Deakin *et al.,* 2012).

The mechanism of coordination of the sigma factors across the double membrane barrier between the forespore and the mother cell in either *C. difficile* or *B. subtilis* is not fully understood. However, it is proposed that a channel between the two compartments, the SpoIIQ:SpoIIIAH complex, is required for the completion of engulfment in both species and could provide a possible signalling mechanism (Camp and Losick, 2009).

1.2.1 SpoIIQ and SpoIIIAH

SpoIIIAH is part of the eight-gene *spoIIIA* operon which encodes a group of proteins localised to the mother cell membrane that are expressed under the control of σ^E and is a conserved hallmark of sporulation capability in many species (Camp and Losick, 2008). Anchored to the membrane by a 30 residue N-terminal transmembrane spanning domain, the C-terminus of SpoIIIAH is located in the inter-sporangial space (Figure 1.2.3). The intersporangial domain of SpoIIIAH shares sequence identity to the YscJ/EscJ pore form-

ing proteins found in Type III secretion systems, more commonly associated with Gram-negative bacteria, and as such is predicted to organise in a pore-like arrangement in the mother cell membrane (Camp and Losick, 2008; Meisner *et al.,* 2008).

SpoIIQ is a forespore membrane bound protein that is expressed by the *spoII* regulon, under the control of σ^F during early stage sporulation. SpoIIQ is an essential protein for forespore engulfment after asymmetric division has occurred (Illing and Errington, 1991; Londoño-Vallejo *et al.,* 1997; Sun *et al.,* 2000; Serrano *et al.,* 2015; Fimlaid *et al.,* 2015). An N-terminal transmembrane domain consisting of 30 residues also anchors SpoIIQ to the forespore membrane, with the C-terminal region of the protein located in the intersporangial space between the mother cell and forespore membranes (Meisner and Moran, 2011). The C-terminal domain of *C. difficile* SpoIIQ has 38% sequence identity with the LytM family of endopeptidases (Meisner and Moran, 2011).

It was previously reported that there was no *spoIIQ* homologue in *C. difficile* (Galperin *et al.,* 2012), however, gene CD0125 has since been identified as a *spoIIQ* homologue in *C. difficile* strain 630 and *spoIIIA* operon containing a SpoIIIAH homologue has long been annotated in the *C. difficile* 630 genome (Sebaihia *et al.,* 2006; Monot *et al.,* 2011; Fimlaid *et al.,* 2013).

SpoIIQ and SpoIIIAH are required for the completion of engulfment in both *B. subtilis* and *C. difficile* (Londoño-Vallejo *et al.,* 1997; Dembek *et al.,* 2015; Serrano *et al.,* 2015). In *B. subtilis*, SpoIIQ and SpoIIIAH were shown by SEC-MALLS to form a 1:1 heterodimer via their extracellular, C-terminal intersporangial domains (Levdikov *et al.,* 2012). The complex connects the forespore and mother cell together in a zipper-like mechanism, allowing access between the two cells, as visualised by fluorescence microscopy (Meisner *et al.,* 2008; Blaylock, 2004; Ojkic *et al.,* 2014; Rubio and Pogliano, 2004). The interaction was further illustrated by the determination of the crystal structure of the intersporangial domains of SpoIIQ:SpoIIIAH complex from *B. subtilis,* which revealed an anti-parallel β-sheet interface between SpoIIQ and SpoIIIAH (Figure 1.2.4) (Meisner *et al.,* 2012; Levdikov *et al.,* 2012).

The work presented in this thesis was part of a collaboration with Professor Adriano Henriques' lab at IQTB, Lisbon, Portugal, who performed *in vivo* studies on the SpoIIQ:SpoIIIAH complex in *C. difficile*. It was shown that the C-terminal domains of

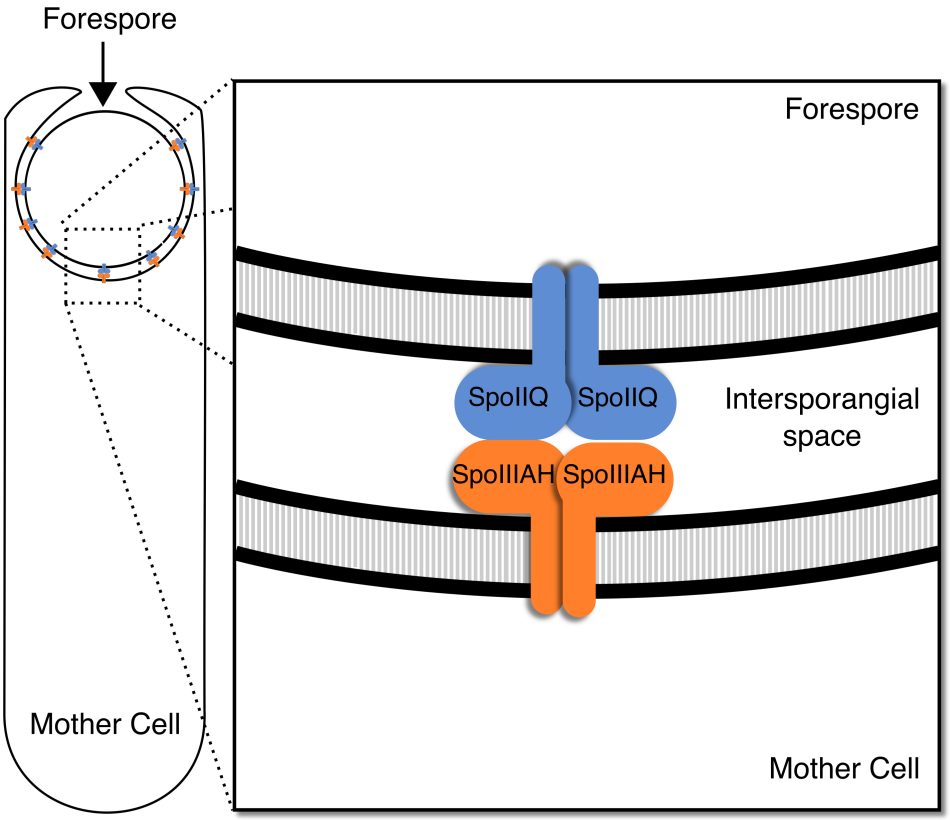


Figure 1.2.3 – Organisation of SpoIIQ:SpoIIIAH at the mother cell to forespore septum. Both SpoIIQ (blue) and SpoIIIAH (orange) have N-terminal transmembrane domains that anchor the proteins in the forespore and mother cell membranes, respectively. The globular C-terminal domains of the proteins are positioned in an extracellular space known as the intersporangial space between the these membranes. SpoIIQ and SpoIIIAH interact via their respective C-terminal domains. The complex is localised to the forespore-mother cell interface and as engulfment proceeds, clusters of SpoIIQ:SpoIIIAH complex are observed to move around the forespore with a ratchet-like motion until the forespore is completely surrounded by a double membrane punctate with SpoIIQ:SpoIIIAH complexes.

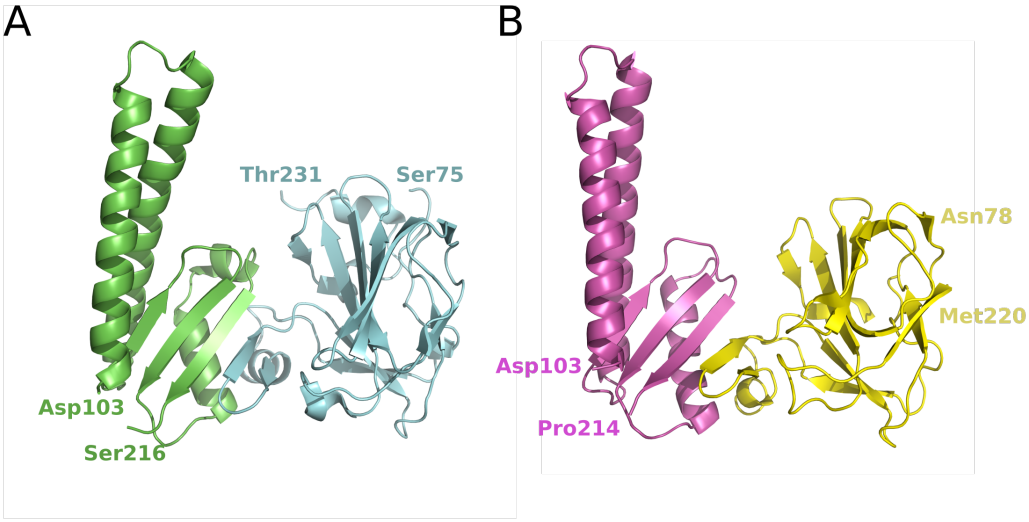


Figure 1.2.4 – *B. subtilis* SpoIIQ:SpoIIIAH crystal structures. The crystal structures of the intersporangial domains of *B. subtilis* SpoIIQ (blue, yellow) and SpoIIIAH (green, magenta) were determined by Levdikov *et al.* (A) and Meisner *et al.* (B). The terminal residues of each chain are labelled accordingly. Both of these structures reveal an anti-parallel β-sheet interface between SpoIIQ and SpoIIIAH (Levdikov *et al.,* 2012; Meisner *et al.,* 2012).

SpoIIQ and SpoIIIAH also interact in *C. difficile* across the intersporangial space (Figure 1.2.5A/B) (Serrano *et al.,* 2015). Neither *spoIIIAH* nor *spoIIQ* mutants could produce mature *C. difficile* spores as these mutants did not progress beyond spore engulfment initiation (Serrano *et al.,* 2015; Fimlaid *et al.,* 2015). In *C. difficile,* the SpoIIQ:SpoIIIAH complex is required for the late stage activity of σ^G, however, early stage σ^G activity can be detected in *spoIIQ* and *spoIIIAH* mutants (Serrano *et al.,* 2015; Fimlaid *et al.,* 2015). In *B. subtilis,* no σ^G activity is observed in mutants of either *spoIIQ* or *spoIIIAH*. We showed that a *C. difficile spoIIIAH* mutant has a reduced σ^K signal, indicating that SpoIIIAH is involved in σ^Kactivation (Serrano *et al.,* 2015). In addition, both *spoIIQ* and *spoIIIAH* mutants displayed membrane buckling or inverse septa, indicating that peptidoglycan processing was impaired (Figure 1.2.5C) (Serrano *et al.,* 2015). Such membrane buckling has not been described in *B. subtilis spoIIQ* or *spoIIIAH* mutants but has been observed in *spoIID* and *spoIIP* mutants, which could not complete forespore engulfment (Londoño-Vallejo *et al.,* 1997). SpoIID and SpoIIP are peptidoglycan peptidases that form a complex with SpoIIM (DMP machinery) and have been identified in the forespore engulfing mother cell membrane (Fredlund *et al.,* 2013; Rodrigues *et al.,* 2013). It appears that there is an association between the SpoIIQ:SpoIIIAH and DMP complexes in both *C. difficile* and *B. subtilis,*

although the function of such an interaction appears to differ between the species. In *C. difficile,* it appears that SpoIIQ and SpoIIIAH are required for the localisation of the DMP complex to the engulfing membrane. Conversely in *B. subtilis, spoIID* and *-P* mutants result in incorrect localisation of SpoIIQ, localised away from the sporulation septum of the forespore membrane, indicating that the DMP machinery is required for correct localisation of SpoIIQ (Aung *et al.,* 2007; Fredlund *et al.,* 2013). However, in *B. subtilis spoIIQ* mutants can complete engulfment under certain growth conditions and the DMP complex may instead be localised through proteins for which there are no direct homologues in *C. difficile* (Aung *et al.,* 2007; Fredlund *et al.,* 2013). In *B. subtilis,* an interaction has been observed between SpoIIQ and the phosphatase, SpoIIE (Flanagan *et al.,* 2016).

After the forespore has been engulfed by the mother cell, spore coat proteins are expressed and surround the forespore (Figure 1.2.1A) (Yutin and Galperin, 2013; Fimlaid *et al.,* 2013). However, 15% of *spoIIIAH C.* difficile mutants expressed *cotE*, a σ^K regulated protein, compared with 84% of WT cells, indicating that SpoIIIAH is required for late stage expression (Serrano *et al.,* 2015). Even though neither *spoIIQ* nor *spoIIIAH* mutants completed engulfment, these mutants displayed disorganisation of the coat proteins when compared with the wildtype (Fimlaid *et al.,* 2015). Late forespore expression defects have also been observed in *B. subtilis spoIIQ* mutants that develop malformed spore coats (McKenney and Eichenberger, 2012). In *B. subtilis,* SpoIIQ:SpoIIIAH remain in a complex until engulfment is complete and the intersporangial domain of SpoIIQ is cleaved from the transmembrane domain, releasing the complex (Chiba *et al.,* 2007). The timescale of complex degradation in *C. difficile* is currently not clear as SpoIIQ and SpoIIIAH are detectable post-engulfment (Serrano *et al.,* 2015).

The substrate of the SpoIIQ:SpoIIIAH channel is to-date unknown in many bacteria. It has been proposed that the SpoIIQ:SpoIIIAH complex functions as a feeding tube to the otherwise isolated forespore, or as a signalling channel in the temporal regulation of the sporulation σ-factor cascade (Camp and Losick, 2009; Fimlaid *et al.,* 2015; Serrano *et al.,* 2015). It is clear that SpoIIQ:SpoIIIAH is required for late σ^G expression and is involved in the regulation of σ^K and therefore the production of mature *C. difficile* spores. While the regulation of the σ-factor cascade differs between *C. difficile* and *B. subtilis,* it is possible that the function of the complex in these species is similar. The SpoIIQ:SpoIIIAH

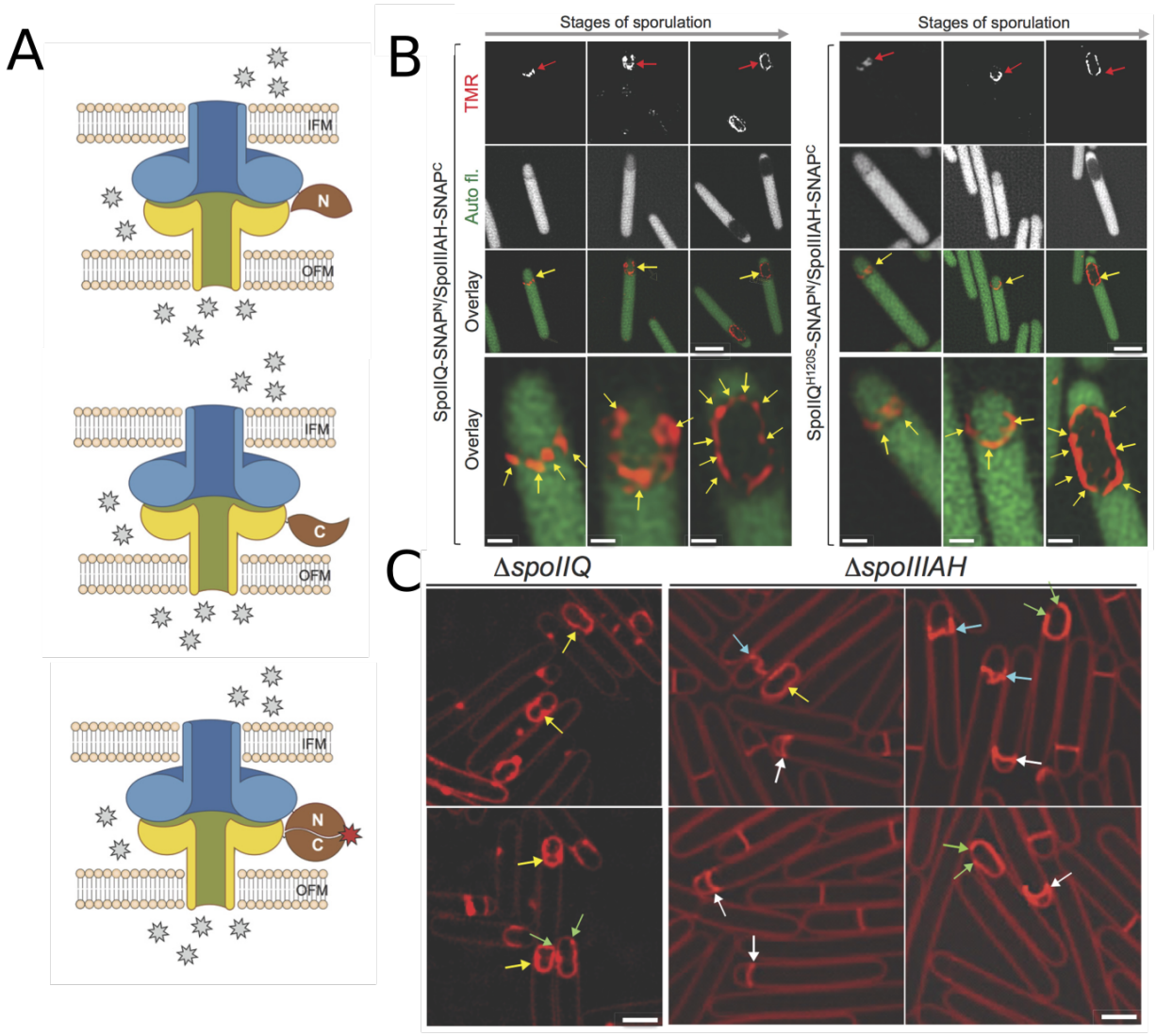


Figure 1.2.5 – Membrane defects in *spoIIQ* and *spoIIIAH* mutants and complex localisation in *C. difficile*. Adapted from Serrano *et al.,* 2015. **A:** Schematic of the split fluorescent SNAP tag (brown), the N-terminal domain of SNAP was fused to SpoIIQ and the C-terminal domain was fused to SpoIIIAH. When the two domains of the SNAP tag were in proximity the binding site for the fluorophore, TMR-Star, was formed and resulted in a fluorescence signal. **B:** The split SNAP SpoIIQ-SNAP^{N}/SpoIIIAH-SNAP^{C} showed the formation of clusters (yellow arrows) at the forespore-mother cell septum and as the forespore was engulfed these clusters were observed around the entirety of the forespore.The presence of a fluorescent signal for TM-Star showed that The C-termini of SpoIIQ and SpoIIIAH interacted in *C. difficile*. **C:** Membrane staining (FM-64) of mutants of *spoIIQ* (left) and *spoIIIAH* (right) revealed bulging (yellow arrows), buckled (blue arrows) or inverse septa (white arrows) during engulfment of the forespore which indicated improper peptidoglycan processing.

complex could transport a molecule required for the complete activation of σ^G and σ^K in the mother cell. In addition to the conveyance of a σ-factor related signal, it appears that SpoIIQ:SpoIIIAH also has a role in the recruitment and localisation of other proteins that are required for proper engulfment of the forespore, such as the DMP complex. An interaction between *B. subtilis* SpoIIQ and SpoIIE, was stabilised by Tyr28 within the transmembrane domain and its mutation resulted in improper localisation of SpoIIE and was required for maximum σ^G activity via the anti-sigma factor CsfB (Flanagan *et al.,* 2016). However, there are no Tyr residues in the transmembrane domain of *C. difficile* SpoIIQ.

1.2.1.1 SpoIIQ:SpoIIIAH channel assembly

The current model in *B. subtilis* and *C. difficile* is that SpoIIQ and SpoIIIAH form a multimeric complex, with SpoIIQ molecules interacting within the forespore membrane to form a pore like assembly and SpoIIIAH organising in a similar manner in the mother cell membrane (Figure 1.2.3). The crystal structures of the soluble, intersporangial domains of *B. subtilis* SpoIIQ:SpoIIIAH in complex have been used to model possible pore-like assemblies (Figure 1.2.6) (Levdikov *et al.,* 2012; Meisner *et al.,* 2012). These models suggest that SpoIIQ and SpoIIIAH can form 12-mer, 15-mer or 18-mer rings within their respective membranes (Figure 1.2.6) (Levdikov *et al.,* 2012; Meisner *et al.,* 2012). Such models also suggest that opening through such a system would be particularly large, with a lumen at least ~60 Å across. It is likely that other proteins, possibly from the mother cell expressed *spoIIIA* operon, are required for regulation of the SpoIIQ:SpoIIIAH channel.

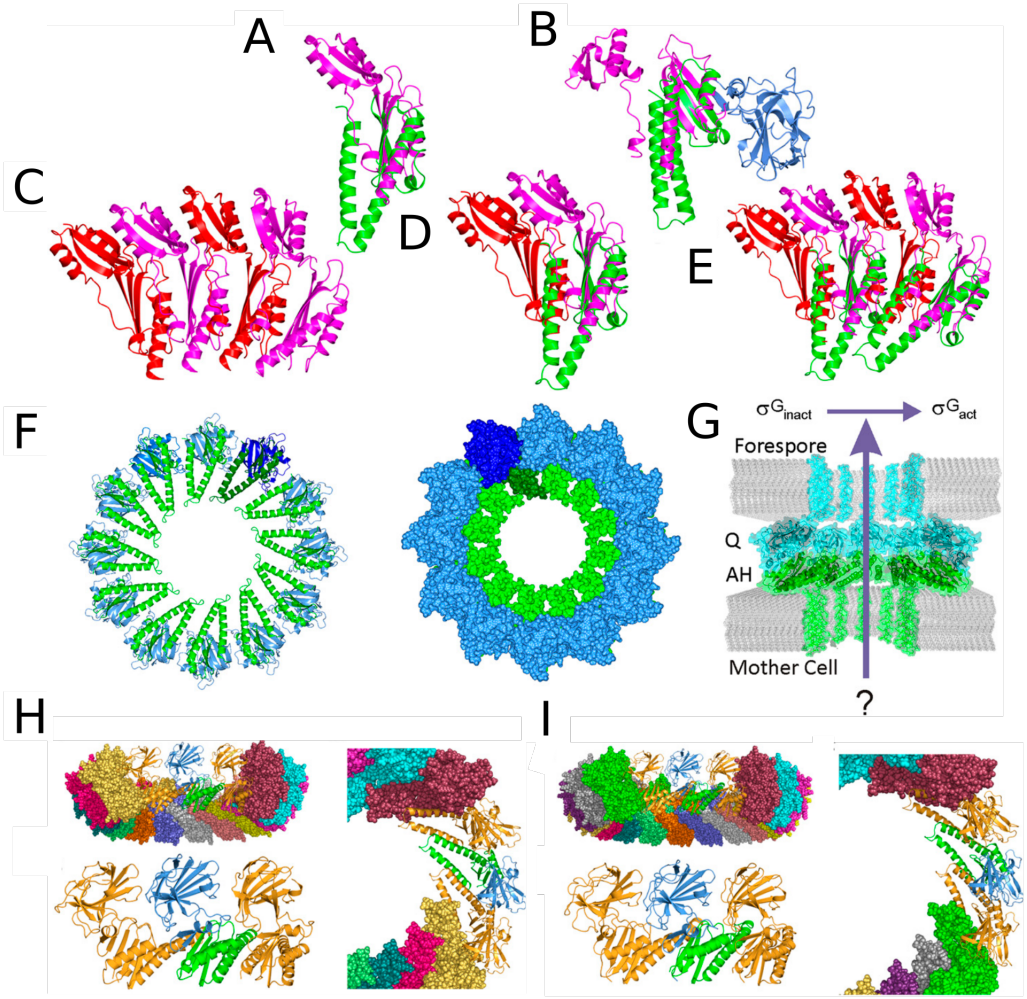


Figure 1.2.6 – Proposed models for SpoIIQ:SpoIIIAH channel formation. Adapted from Levdikov *et al.,* 2012 and Meisner *et al.,* 2012. Levdikov et al. superposed (A, D, E) SpoIIIAH (green) and the structure of EscJ (red or magenta), which crystallised in a helical manner (C) to form a pore structure. The SpoIIQ:SpoIIIAH (SpoIIQ in blue) was superimposed on via the SpoIIIAH chain with EscJ (B) to form a complete pore structure containing 12 SpoIIQ:SpoIIIAH dimers (F cartoon representation and surface). Levdikov et al. modelled the transmembrane regions of SpoIIQ and SpoIIIAH within their respective membranes to build a complete SpoIIQ:SpoIIIAH channel model. Two ring formation models were proposed by Meisner et al. formed of 15 (H) and 18 (I) SpoIIQ:SpoIIIAH dimers based upon a set of assumption calculated using the modelling package Rosetta, with a set of assumptions based on the similarities of SpoIIIAH with EscJ.

1.2.1.2 LytM endopeptidases

The C-terminal domain of SpoIIQ shares sequence homology with a group of endopeptidases known as LytM, part of the M23 family of metalloproteases that cleave the peptide cross-bridges of peptidoglycan (Firczuk and Bochtler, 2007). Two motifs, HxxxD (motif 1) and HxH (motif 2), are required for the coordination of a single catalytic Zn^{2+} metal ion. Comparison of SpoIIQ sequences from the *Clostridia* and *Bacilli* genera, the major spore forming Firmicutes, showed that motif 1 was more conserved amongst the *Clostridia* than in the *Bacilli* (Figure 1.2.7). The SpoIIQ of *C. difficile* contains a complete LytM domain that may enable it to co-ordinate a catalytic Zn^{2+}, which would be essential for endopeptidase activity (Firczuk *et al.,* 2005; Firczuk and Bochtler, 2007). While motif 1 is conserved in *C. difficile* it is degenerate in *B. subtilis,* where the His of motif 1 is substituted by a Ser (Crawshaw *et al.,* 2014).

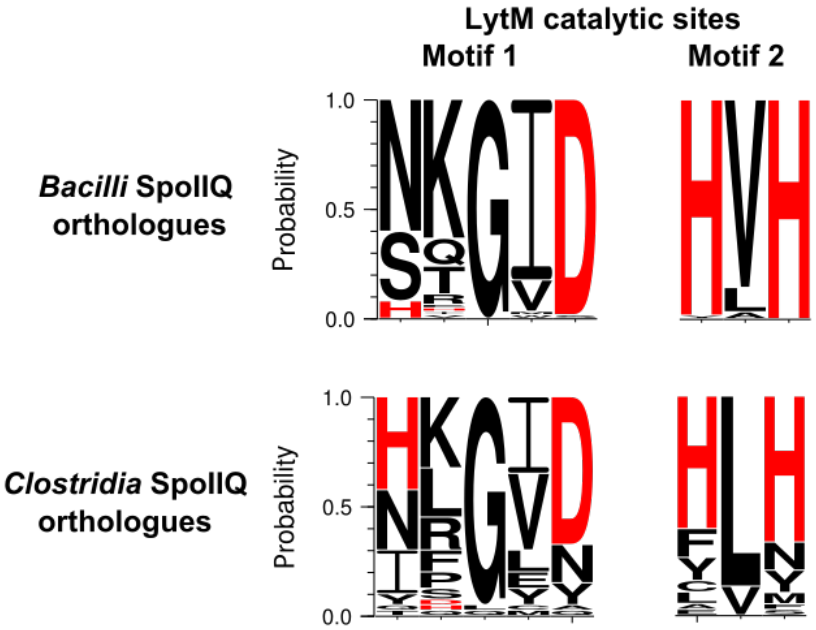


Figure 1.2.7 – **Conservation of LytM motifs in SpoIIQ of *Bacilli* and *Clostridia***. Adapted from Crawshaw et al. (Crawshaw *et al.,* 2014) Conservation of the metal coordinating residues within two motifs (HxxxD and HxH) in SpoIIQ across the *Bacilli* (top) and *Clostridia* (bottom). The complete motif is conserved in almost half of *Clostridia* and in a very small group of *Bacilli*, indicating that between these two genera SpoIIQ may have different roles in engulfment (Crawshaw *et al.,* 2014). A hidden Markov model derived SpoIIQ sequence was produced using HMMER3 (Eddy, 1998) from a selection of representative endospore formers (Hoon *et al.,* 2010) and used in BLASTp searches to identify SpoIIQ genes from the *Bacilli* and *Clostridia* genera and Weblogo (Crooks *et al.,* 2004) was used to create the sequence logo.

Structures of *Staphylococcus aureus* LytM are represented in the PDB, including a structure with a penta-glycine peptidoglycan analogue (PDB: 4ZYB) (Grabowska *et al.,* 2015), from which a catalytic mechanism has been proposed (Figure 1.2.8). The Zn^{2+} is important for the catalytic function of the endopeptidase, providing a mechanism for nucleophilic attack of the scissile peptide bond bridging peptidoglycan strands (Figure 1.2.8) (Firczuk *et al.,* 2005).

The crystal structures of *B. subtilis* SpoIIQ with its degenerate LytM domain do not contain a metal ion at the coordination site (Meisner and Moran, 2011; Levdikov *et al.,* 2012; Meisner *et al.,* 2012). The structures of *B. subtilis* SpoIIQ and *S. aureus* LytM can be superimposed with a core RMSD of 1.8 Å (Figure 1.2.8C), indicating that the fold is maintained in *B. subtilis* SpoIIQ.

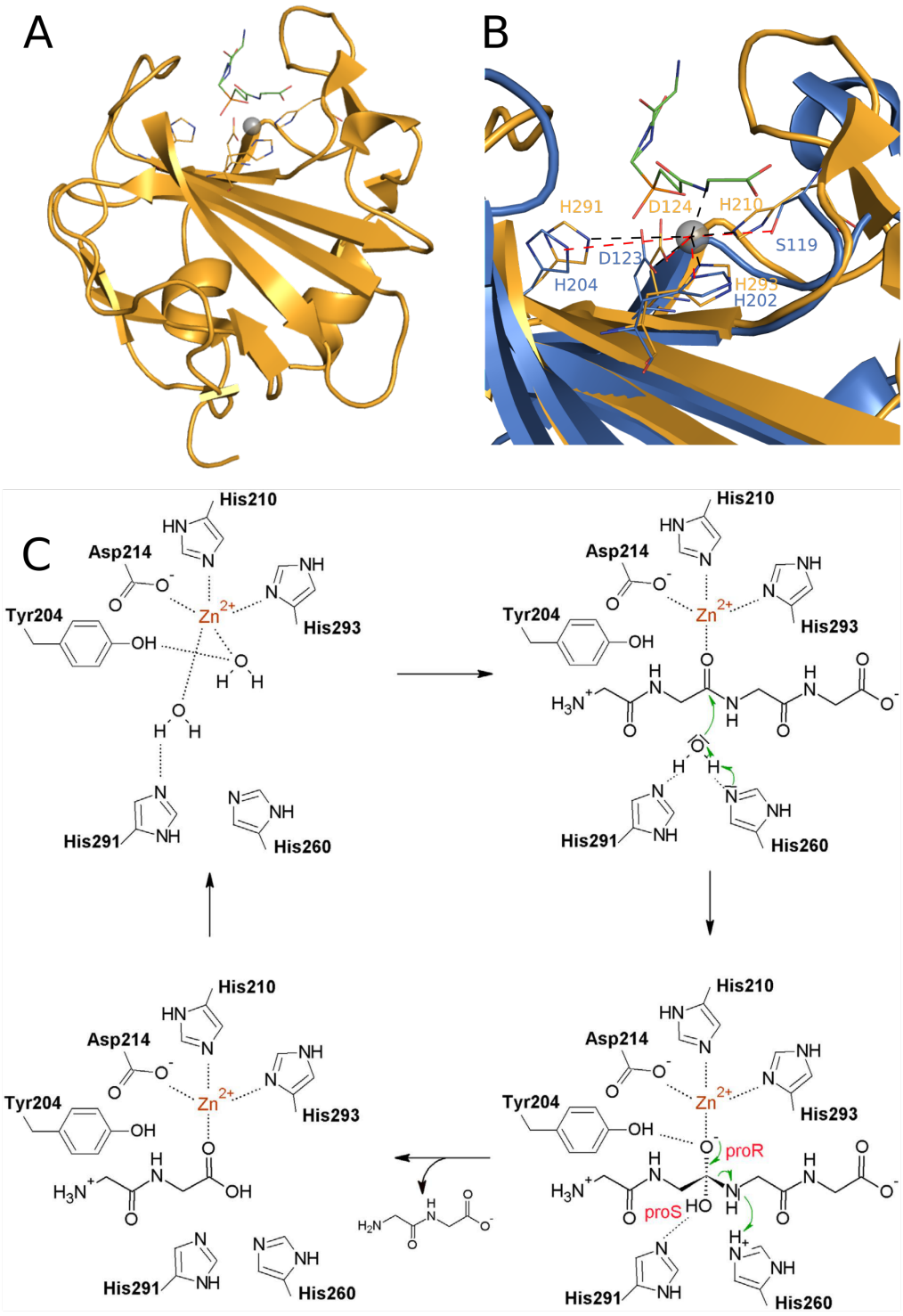


Figure 1.2.8 – Structure of *S. aureus* LytM endopeptidase and proposed mechanism. Adapted from Grabowska *et al.,* 2015. **A:** Cartoon of the structure of *Staphylococcus aureus* LytM (PDB: 4ZYB) with penta-glycine, peptidoglycan fragment mimic and Zn^{2+} bound. **B:** Secondary structure superimposition of the Zn^{2+} coordination site of *S. aureus* LytM (gold) and *B. subtilis* SpoIIQ (blue) (core RMSD: 1.8 Å). The coordinating residues are displayed in stick form and labelled. **C:** The proposed mechanism of catalysis by the *S. aureus* LytM. Lone pairs are represented by lines surrounding the oxygen. The transition state oxygen atom of the peptide backbone is labelled proR and the oxygen from the incoming water is labelled as proS.

1.3 *C. difficile colonisation*

Once a *C. difficile* spore is in a favourable environment, the process of germination is stimulated resulting in the production of a vegetative cell. In a nutrient rich, anoxic environment, the vegetative cell will replicate forming a colony. *C. difficile* colonies are able to form protective biofilms and migrate towards resources close by. To perform these functions, it has recently been discovered that *C. difficile* express Type IV pili (TFP), retractable protein filaments that can provide adhesion to host cells and other surfaces, providing the cell with gliding motility.

1.3.1 Bacterial pili

First observed in Gram-negative bacteria in the 1940s (Pirie, 1949), pili are present across many prokaryotes in a number of forms. A single pilus is a polymerised protein filament that is made up of several hundreds or thousands of a single protein unit known as a pilin (Wu and Fives-Taylor, 2001). The pilin filaments extend from the cell in a manner similar to the guide lines of a tent and can aggregate to form strong bundles. Pili have a number of functions including cell-cell adhesion, DNA uptake, biofilm formation and bacterial motility (Proft and Baker, 2009). Pilins have been most extensively studied in Gram-negative bacteria, in which four different groups have been identified: Type I pili; Type IV pili (TFP); curli pili; CS1 pilus family (Proft and Baker, 2009). These groups of have been categorised by their assembly pathways and, with the exception of TFP, are exclusively observed in Gram-negative bacteria. The Type I, curli pili and CS1 family are strongly associated with host cell adhesion. Structures of the Type I pilins have been shown to contain (Ig)-like folds such as described in FimH from uropathogenic *Escherichia coli* (Choudhury *et al.,* 1999). Type I pilins are secreted into the periplasm before being transported and processed on to the extracellular surface of the outer membrane by chaperone and usher proteins, the polymerised pilins are then secreted as a fibre (Proft and Baker, 2009). Curli pili are unusual coiled fibres associated with enteric bacteria such as *E. coli* and their highly β-stranded structures resemble, both structurally and biochemically, eukaryotic amyloid fibres as observed in Alzheimer's and some prion diseases (Proft and Baker, 2009). Finally, the CS1 family are a distinct group of pili that are associated

with enterotoxigenic *E. coli* and share little structural similarity with the other pili . The CS1 family also utilises chaperones in pilus assembly and is referred to as the alternative chaperone usher pathway (Forest, 2013).

In addition to TFP, Gram-positive pathogens such as *Streptococcus pyogenes* have also been observed to express pili known as Gram-positive pili, which have been shown to play a role in host-cell adhesion (Manetti *et al.,* 2007). Unlike the pili observed in the Gram-negative bacteria, Gram-positive pilin proteins are covalently attached to each other during formation of the filament, which is also covalently attached to the lipids of the cell membrane (Ton-That and Schneewind, 2003; Pansegrau and Bagnoli, 2016). Gram-positive pili are secreted by the Sec-dependent secretion pathway and the signal peptide is cleaved before assembly (Pansegrau and Bagnoli, 2016). However, it is not clear how assembly of Gram-positive pili filaments takes places since the pilin proteins are covalently linked to one another (Pansegrau and Bagnoli, 2016). Interestingly, a common feature to most of the pili structures that have so far been determined, regardless of the type of pili from either Gram-negative and Gram-positive bacteria, is a variable region that contains a disulphide bridge (Proft and Baker, 2009). The TFP are unique in that they are the only pili that have the ability to retract and therefore give the bacterial cell the ability to glide along surfaces in a manner that is independent of flagella (Jin *et al.,* 2011).

1.4 Type IV pili

Type IV pili were first observed in Gram-negative bacteria (Ottow, 1975) and have more recently been identified in Gram-positive genera such as the *Clostridia* (Varga *et al.,* 2006). TFP fibres have been measured at 6-8 nm wide and several microns long (Proft and Baker, 2009). TFP have several components that form the basal element of the pilin fibre, including core membrane proteins, extension and retraction ATPases and a pre-pilin peptidase (Mattick, 2002). Pilin proteins are expressed as pre-pilins that have conserved signal peptide sequences at the N-terminus, which has also enabled identification of pilins genetically (Melville and Craig, 2013). There are two distinct signal peptide sequences by which the TFP have been classified into two groups: Type IVa (TFPa) and Type IVb (TFPb). TFPa have short signal peptides of 5-7 amino acid residues that include the

conserved motif, GFxLxE, which is cleaved by a pre-pilin peptidase at the carboxyl side of the Gly residue. The first residue of the mature pilins is Phe. The Glu at position 5 has been identified as a key residue for recognition by the pre-pilin peptidase and in assembly of the pilin fibre (Proft and Baker, 2009). TFPb signal peptides are longer (~26 residues) and the N-terminal residue of the mature TFPb pilins is a hydrophobic residue other than Phe (Melville and Craig, 2013). TFPa pilins are commonly 150-160 residues in length whilst the TFPb pilins are either very short (40-50 residues) or much longer than TFPa proteins (>180 residues) (Proft and Baker, 2009). The most extensively studied TFP are from the Gram-negative *Vibrio cholerae* and the *Neisseria* pathogens, *N. gonorrhoeae* and *N. meningitidis*, members of the TFPa family. TFP are a key virulence factor for both of these *Neisseria* species as TFP deficient mutants do not cause infection (Craig *et al.,* 2004).

1.4.1 Type IV pili structure and organisation

The nomenclature of TFP proteins in many species is based upon the Pil- prefix, the major pilin unit is known as PilA with exception of the *Neisseria* and *Vibrio* genera in which the major pilin is known as PilE and TcpA, respectively. In Gram-negative bacteria, the basal element of the TFP filament is located in the inner membrane and is formed of the pre-pilin peptidase (PilD), an assembly ATPase (PilB), a retraction ATPase (PilT) and core membrane associated proteins (PilC/M/O) that are localised at the very base of the fibre (Ayers *et al.,* 2009; Chang *et al.,* 2016). In addition to these components, a secretion protein (PilQ) is present in the Gram-negative TFP systems and functions as a channel through which the pilin fibre can pass through the peptidoglycan in the periplasm and the outer membrane (Martin *et al.,* 1993; Frye *et al.,* 2006).

Recently, the structure of the TFPa basal unit of the Gram-negative *Myxococcus xanthus* has been determined using cryo-EM at a resolution of 3-4 nm (Chang *et al.,* 2016). Using systematic mutants of the TFPa components and cryo-EM imaging, Chang *et al.* were able to determine the assembly of the Gram-negative basal unit (Figure 1.4.1). The *M. xanthus* basal unit structures showed in detail how the secretion proteins are arranged within the periplasm, revealing an extensive lateral anchoring mechanism through the peptidoglycan layer (Chang *et al.,* 2016).

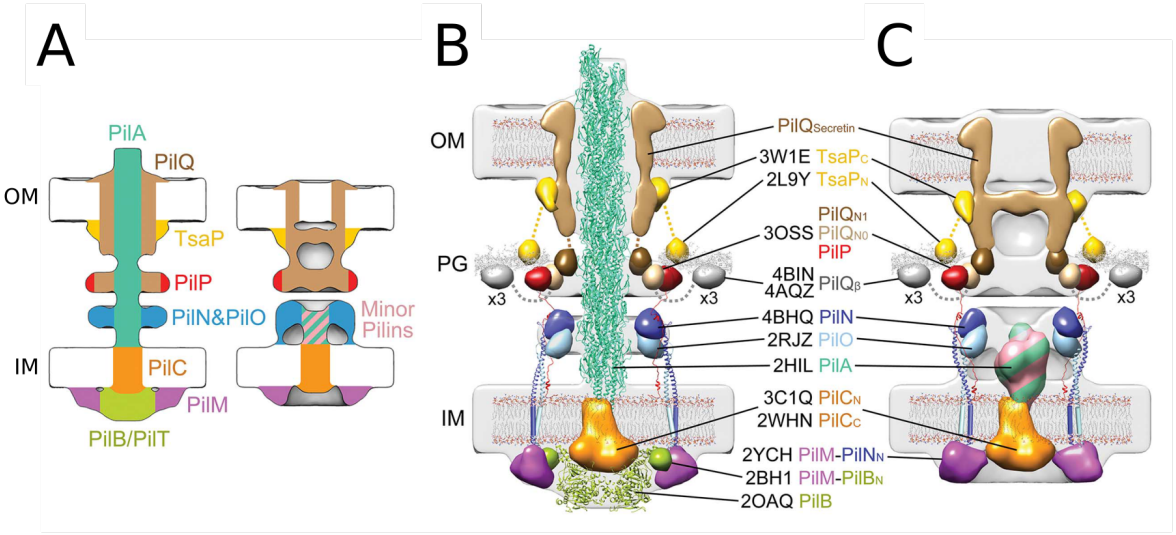


Figure 1.4.1 – **Architectural model of the *Myxococcus xanthus* TFPa basal unit.** Adapted from Chang *et al.,* 2016. **A:** Schematic of the basal unit with pilin fibre (piliated, left) and an empty basal unit without fibre (right). **B:** Available crystal structures of the components of the *M. xanthus* TFP were modelled into the electron density of a piliated basal unit determined by cryo-EM **C:** The electron density of a non-piliated basal unit with the available crystal structures of basal unit modelled in. Note that electron density was observed (green/pink striped) that appeared to plug the empty basal unit.

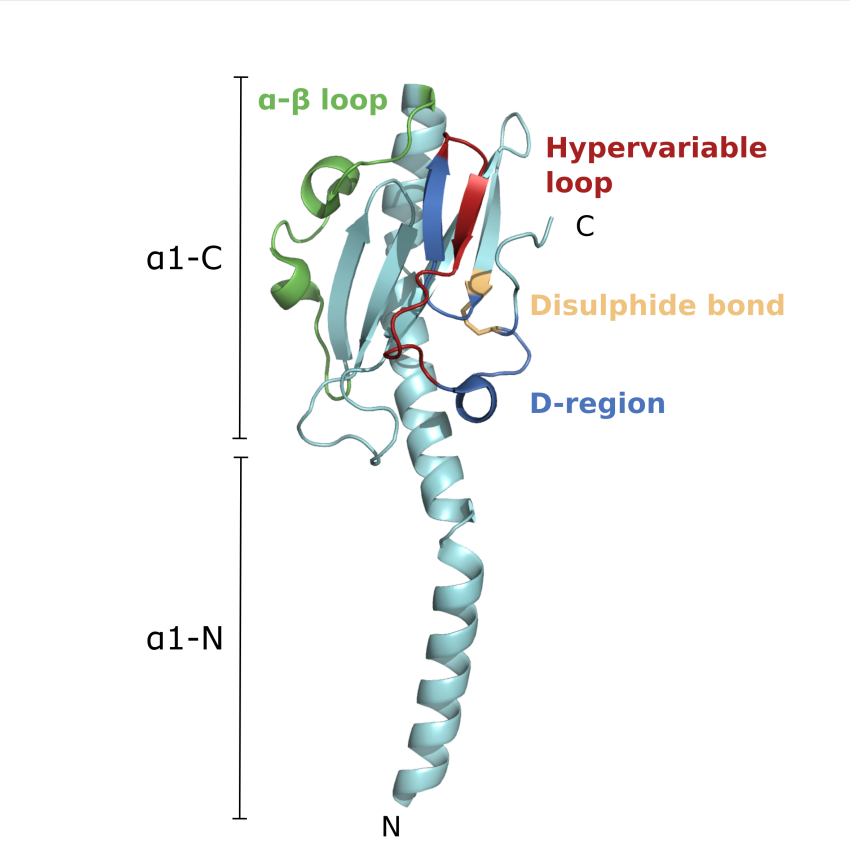


Figure 1.4.2 – Crystal structure of full-length PilE1 from *N. gonorrhoeae.* The structure of full-length PilE1 (Craig *et al.,* 2006) has distinct domains: an N-terminal hydrophobic helix (α1) responsible for pilin polymerisation, an α-β loop and D-region. The D-region is delimited by a disulphide bond that links the termini of the D-region (Cys121-Cys151, yellow) and exhibits the lowest level of sequence conservation across TFPa pili and contains a hyper-variable loop with the highest level of variability.

The predominant unit of the pilin fibres is classified as the major pilin. A number of crystal structures of major TFP pilin proteins from Gram-negative bacteria have been determined from *V. cholerae* (TcpA), *P. aeruginos*a (PAK), *N. gonorrhoea* (PilE1) and *N. meningitidis* (PilE1) (Craig *et al.,* 2003; Craig *et al.,* 2004; Lim *et al.,* 2010). The structure of full-length PilE1 from *N. gonorrhoea* is presented in Figure 1.4.2 (Craig *et al.,* 2006). The structures share a conserved pair of structural domains: an N-terminal α-helix for pilin polymerisation and C-terminal headgroup. Only *P. aeruginosa* PAK and the *Neisseria* PilE pilin structures represented full-length mature pilin proteins, including the hydrophobic N-terminal helix that extends away from the globular headgroup of the protein (α1-N, Figure 1.4.2) (Craig *et al.,* 2003; Craig *et al.,* 2006). The headgroup of the pilin proteins con-

tains two distinct regions: an α-β loop and a D-region (Craig *et al.,* 2004). The α-β loop connects the α1-helix to a β-sheet of at least 3 anti-parallel strands. The D-region is a sequence-variable region encapsulated within a highly extended loop between 2 strands of the β-sheet domain (Craig *et al.,* 2004). This region is predicted to be exposed on the surface of the pilin filament and shows high sequence variability across species and strains, important for evading immune responses from the host's immune system and making TFP an important virulence factor (Miller *et al.,* 2014). The D-region of all TFP pilin structures determined from Gram-negative species have contained a disulphide bond that links the C-terminal residues to the β-sheet containing region (Craig *et al.,* 2003; Craig *et al.,* 2006).

Electron microscopy studies have elucidated the overall organisation of the pilin subunits within the filaments. Craig *et al.,* have shown that the pilins polymerise in a helical manner, with a 45 Å pitch and modelled the TcpA and PilE1 crystal structures within electron microscopy data of their respective filaments to propose a model for the overall assembly of the pilins from *V. cholerae* and *N. gonorrhoea*, respectively (Figure 1.4.3D/F) (Craig *et al.,* 2003; Craig *et al.,* 2006). Additionally, helical packing was observed in the crystal of *V. cholerae* TcpA (Figure 1.4.3A), in such a way that the N-terminal α1 helices were positioned in the centre of the filament, with a pitch of 35.7 Å (Figure 1.4.3C) (Craig *et al.,* 2003). Although the crystallographic helix and empty bore inside the filament are larger than observed by EM, the overall models are in general agreement (Figure 1.4.3).

The TFP filaments are not exclusively formed of major pilins and are decorated with minor pilins. These pilins are described as minor since they exhibit a lower frequency of expression and incorporation into the pilin fibre (Melville and Craig, 2013). A number of functions have been determined in Gram-negative bacteria including cell-adhesion (Helaine *et al.,* 2005), filament control (Szeto *et al.,* 2011) and DNA binding (Cehovin *et al.,* 2013), which is present in biofilm matrices as eDNA (Whitchurch *et al.,* 2002). The minor pilins share sequence conservation with the major pilins at the N-terminal region, containing the conserved signal peptide sequence and ~30 hydrophobic residues that form the α1-helix, responsible for the polymerisation of the minor pilin pilins in the filament.

One of the most studied Gram-negative TFP minor pilins is PilX of *N. meningitidis*. The 2.5 Å crystal structure of PilX revealed the conserved structural scheme observed in

the major pilins, including the N-terminal α1-helix connected to a 4 stranded anti-parallel β-sheet via an α-β loop and a D-region that was delimited by a disulphide bond (Helaine *et al.,* 2007). *N. meningitidis* PilX and PilE1 structures are superimposable with a core RMSD of 3 Å (Helaine *et al.,* 2007). PilX has been predicted to be key for conformational changes in the TFP filament of *N. meningitidis*, to enable epitope exposure along the pilin during binding to endothelial cells, therefore providing a key virulence factor in *N. meningitidis* infection (Helaine *et al.,* 2005; Biais *et al.,* 2010; Brissac *et al.,* 2012). A further *N. meningitidis* minor pilin, PilV, has been shown to be important for internalisation of *N. meningitidis* cells by human endothelial and epithelial cells (Takahashi *et al.,* 2012). Overall, TFP are an important virulence factor for *N. meningitidis,* with three pilin proteins implicated in the infection mechanisms of this important pathogen making TFP a potential vaccine target (Helaine *et al.,* 2007).

The mechanisms of assembly, extension and retraction of TFP filaments is relatively poorly understood. Mature pilin units that have been processed by the prepilin peptidase, PilD, are predicted to diffuse along the membrane. The hydrophobic nature of the α-1 helix enables it to act as a membrane anchor. However, the Glu5 residue at the N-terminus of the α-1 helix is negatively charged and could result in pilin instability within the membrane. It is predicted, that as a result, the pilin is attracted to the positive charge of an extending pilin filament and is positioned in a gap at the base of the filament. The hydrolysis of ATP by the assembly ATPase is expected to cause a conformational change, anchoring the N-terminal helix of the terminal pilin unit already in the filament within the C-terminal domain of the new pilin unit. This then causes a conformational change that forces the filament away from the membrane resulting in its extension (Craig and Li, 2008).

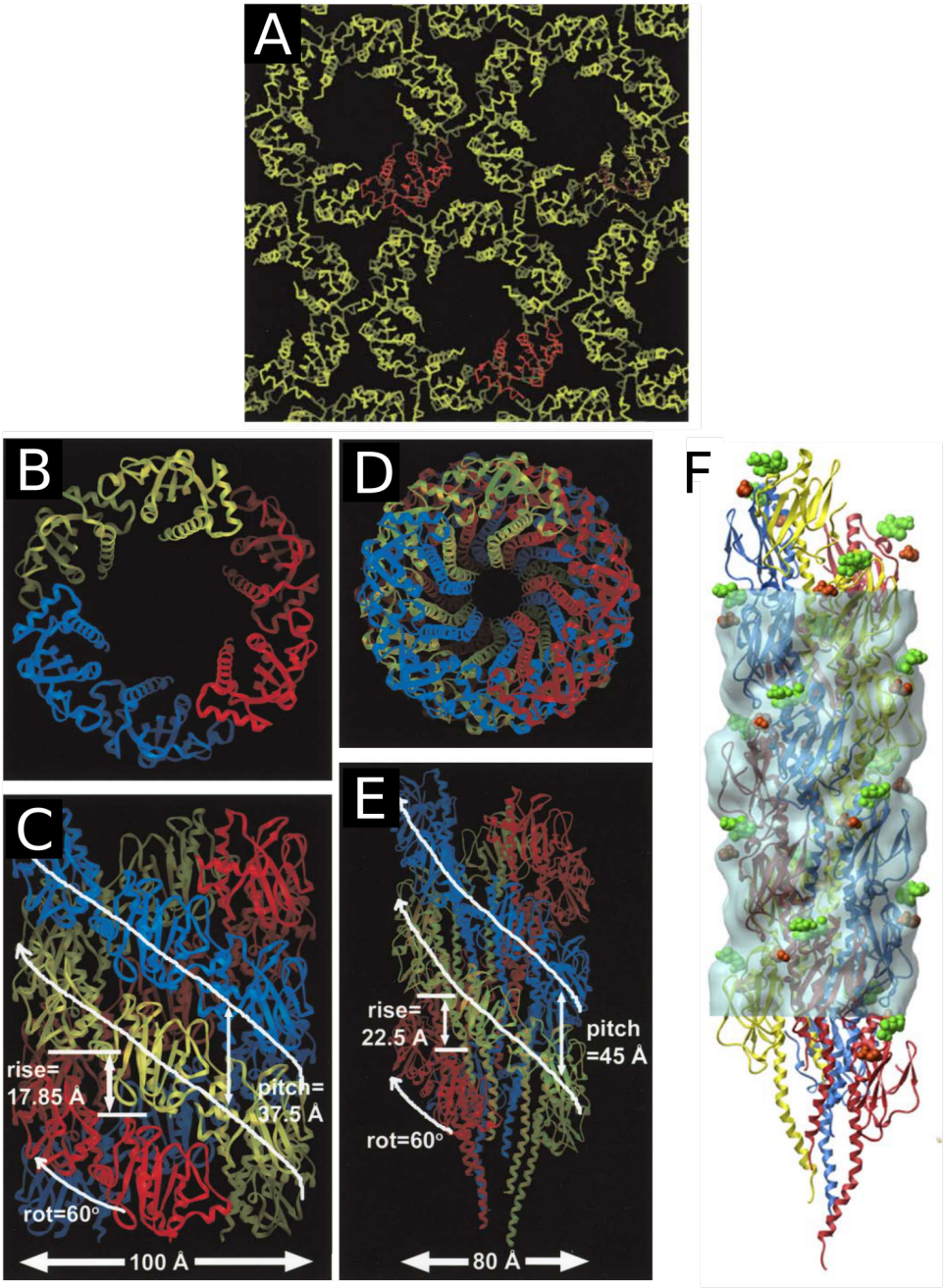


Figure 1.4.3 – **TFP filament model**s. Adapted from Craig et al. 2003 and 2006. **A:** Crystal lattice of *V. cholerae* TcpA (spacegroup $P6_3$) with molecules were arranged in a helical manner. **B:** Top view of the crystallographic filament showing that the α1 helices are positioned on the inside of the filament and have the same polarity. **C:** Side view of the crystallography filament showing the left-handed three start helix with dimensions. **D:** Top view of the electron microscopy derived filament showing a much narrower shaft within the filament in comparison with the crystallographic filament. **E:** Side view of the electron microscopy fitted TcpA filament structure that also displayed a left-handed three start helix but with differing dimensions. **F:** *N. gonorrhoea* TFP filament model built by fitting the full-length crystal structure of PilE1 into the cryo-EM density of the filament. This model also shows filament surface bound molecules (green spheres: disaccharide Gal-DADDGlc; red spheres:phosphoethanolamine). (Craig *et al.,* 2003; Craig *et al.,* 2006)

1.4.2 Gram-negative pseudopilins

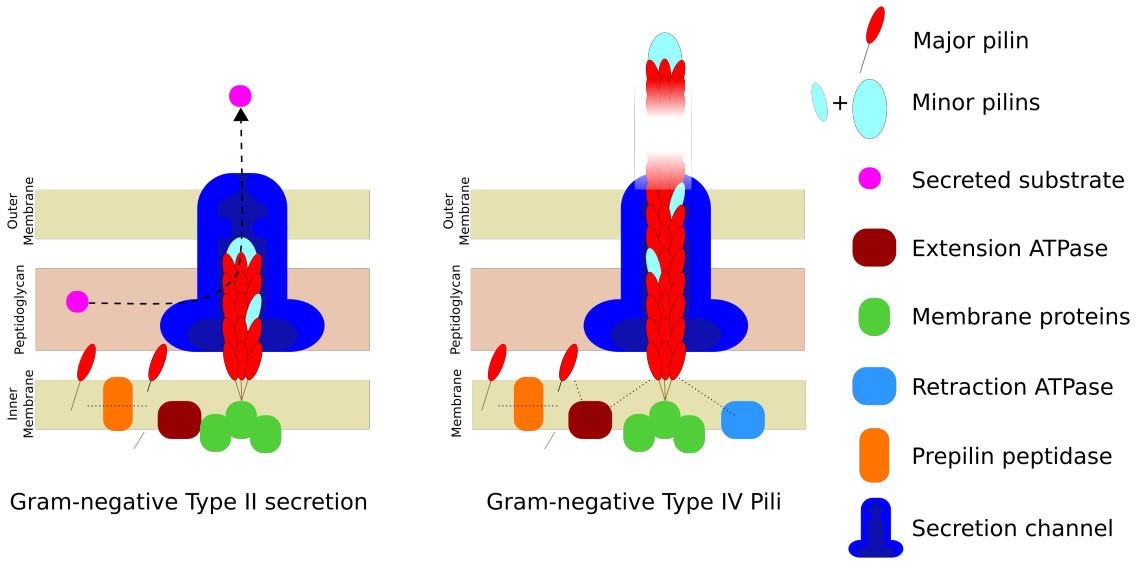


Figure 1.4.4 – Type II secretion system. The Type II secretion of Gram-negative bacteria utilises components similar to those of TFP (shown for reference): a pre-pilin peptidase that processes pseudopilins in the same manner as TFP, PilD. The pseudopilins are added to a pseudopilus by an extension ATPase and a core protein is localised to the base of the pseudopilus. A secretion channel across the periplasm and outer membrane, an the equivalent component to PilQ., enables the secretion of substrates from within the periplasm to the external medium.

Type II secretion systems (TIISS) are a protein secretion system associated with Gram-negative bacteria and thus far no TIISS have been specifically identified in the Gram-positive bacteria. TIISS share many of the components from TFP systems including the core proteins, pilin proteins (known in TIISS as pseudopilins) and ATPases (Korotkov *et al.,* 2012). TIISS allows export of proteins from the periplasm to the external media and some of these substrates have been identified as toxins and biofilm forming proteins, making TIISS an important virulence factor (Figure 1.4.4) (Korotkov *et al.,* 2012; Johnson *et al.,* 2014).

Structures of pseudopilin proteins, have been determined such as PulG from *Klebsiella pneumoniae* (Figure 1.4.5) (Köhler *et al.,* 2004). PulG shares the same structural elements observed in Gram-negative TFP, including an N-terminal α-1 helix, a 3 strand anti-parallel β-sheet connected to the α-1 helix by an α-β loop. However, unlike the Gram-negative TFP, PulG lacks the disulphide bond containing D-region (Köhler *et al.,* 2004). The pseudopilins are not expected to form filaments that extend from the cell, however,

short pseudopili have been observed within the basal unit (Figure 1.4.4) (Korotkov *et al.,* 2012).

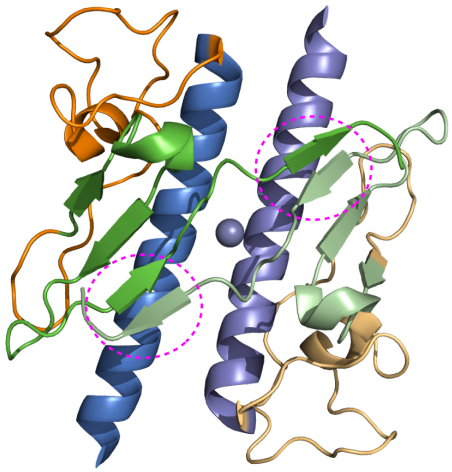


Figure 1.4.5 – Crystal structure of the pseudopilin PulG. Crystal structure of PulG from *Klebsiella pneumoniae* (Köhler *et al.,* 2004). Two molecules are shown with crystallographic domain swapped C-termini (dashed pink circle) which co-ordinate a Zn^{2+}. PulG features the α1 helix (blue), α-β loop (orange) and a 3 strand anti-parallel β-sheet (green), which are also observed in TFP pilin structures. However, the D-region and associated disulphide bridge do not feature in the structure. Dimerisation and metal binding are thought to be crystallographic artefacts.

1.4.3 TFP and the Gram-positive bacteria

More recently, TFP have been identified genetically in Gram-positive bacteria (Figure 1.4.6) including *C. perfringens* and *Ruminococcus albus*, via the highly conserved nature of the N-terminal regions and the clustering of gene components that are associated with the processing, assembly and secretion of TFP filaments (Varga *et al.,* 2006; Rakotoarivonina *et al.,* 2002). These genes were determined to enable *C. perfringens*, which does not possess flagella, to glide across agar surfaces (Varga *et al.,* 2006; Shimizu *et al.,* 2002; Myers *et al.,* 2006).

There are obvious architectural differences between the Gram-negative and Gram-positive cell walls. The TFP filaments of Gram-negative bacteria that are anchored in the inner membrane and secretion proteins, such as PilQ, are required to enable the filament

to pass through the periplasm and outer-membrane (Martin *et al.,* 1993). Meanwhile, in Gram-positive bacteria, there is only one membrane that is coated by a thick layer of peptidoglycan (Brown *et al.,* 2015). While the basal elements found at the inner membrane of Gram-negative bacteria are conserved on the Gram-positive membrane, how the filaments pass through the peptidoglycan layer is not clear. Interestingly, a number of *Clostridia* species exhibit more than one TFP locus and satellite pilin-like genes. It has been proposed that the small TFP loci identified in Gram-positive bacteria may in fact represent TIISS and not TFP and may share some of the components with the TFP loci such as the pre-pilin peptidase (Melville and Craig, 2013).

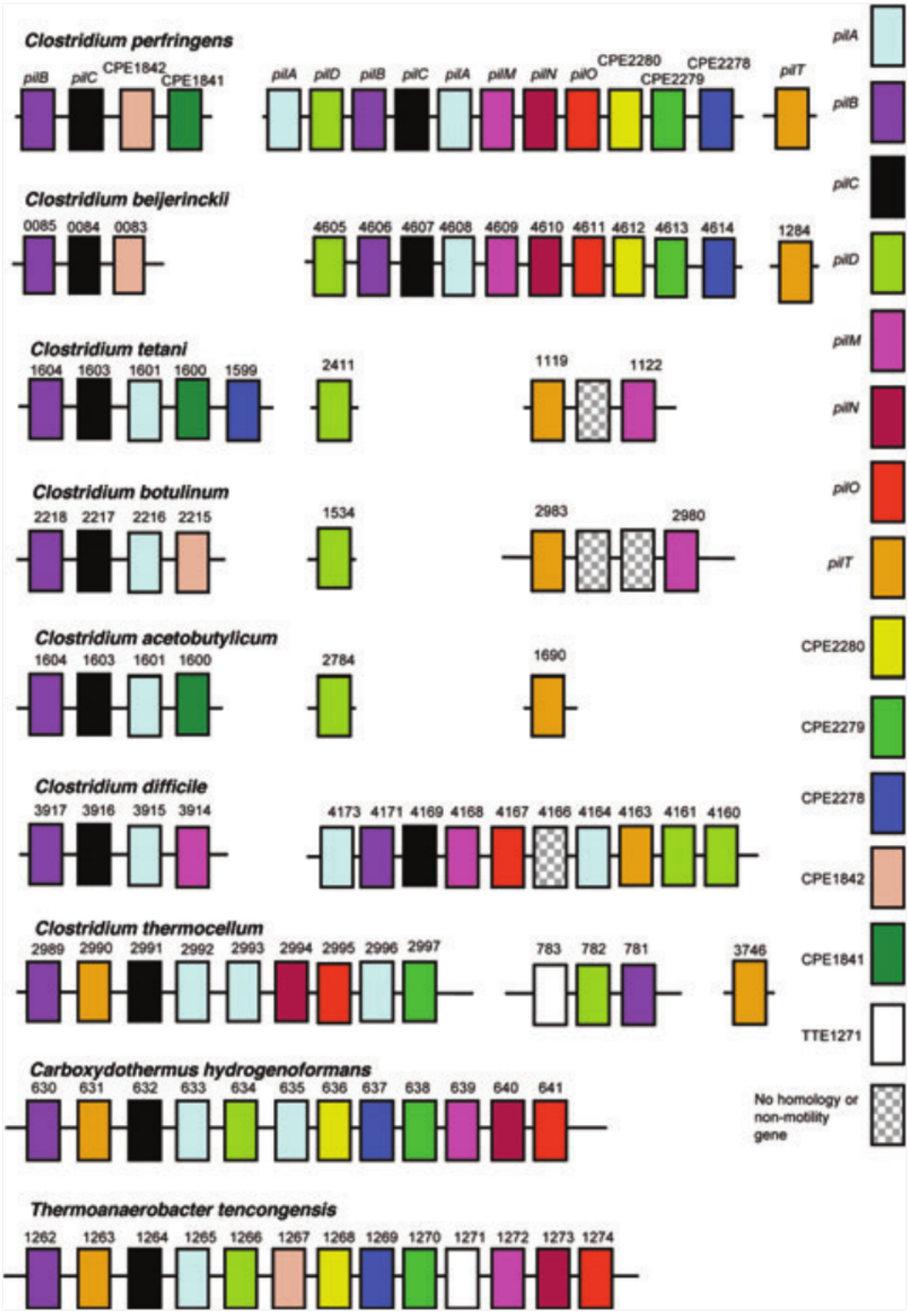


Figure 1.4.6 – **TFP ORFs in clostridial species.** Adapted from Varga et al. 2006. The genes of the *C. perfringens* TFP pilin locus were used to search for other TFP genes and identity TFP loci in other clostridial species.

1.4.4 TFP in *C. difficile*

Pili structures associated with *C. difficile* were first observed by electron microscopy of infected hamster tissue and it was determined that these filaments contained the product of a gene (CD3507) that was predicted to be part of a TFP expressing locus (Goulding *et al.,* 2009). Two loci were initially identified in the *C. difficile* strain 630, a main locus comprising 11 genes (CD3513-CD3503) and a smaller secondary locus comprising up to 6 genes (CD3296-CD3291) (Figure 1.4.7A) (Sebaihia *et al.,* 2006; Varga *et al.,* 2006). Each locus contains a predicted major pilin described as PilA1 (CD3513) and PilA2 (CD3294), extension and retraction ATPases (PilB/T) and at least one core membrane protein (PilC/M/O). Only the main locus contains two pre-pilin peptidases (PilD1/D2), as often observed in TFP loci. Three further pilin-like genes in the middle of the main cluster (CD3508-CD3506), originally classified as PilA proteins, are now described as the minor pilins PilV, PilU and PilK (Figure 1.4.7B).

Based upon sequence similarity with the TFP components in Gram-negative bacteria, it is possible to propose a role for each gene in the main locus to a role and location within the TFP basal unit (Figure 1.4.8). However, unlike the Gram-negative bacteria, there is no PilQ orthologue, even though the filament must pass through the peptidoglycan layer through an as yet unknown mechanism. *C. difficile* also has an outer protein coat on the surface of the peptidoglycan known as the S-layer, which is formed of a paracrystalline layer (Fagan and Fairweather, 2014).

Since the identification of TFP genes in *C. difficile*, pilin expression has been investigated by a number of groups. In an on going collaboration with Professor Neil Fairweather's group, Imperial College, London, it has been determined that all of the genes within the main locus are required for the presentation of TFP on the cell surface as detected via anti-PilA1 antibodies (personal communication).

Bordeleau et al. have shown that *in vivo* expression of TFP filaments is influenced by the presence of cyclic diguanosine monophosphate (c-di-GMP) and that up regulation of c-di-GMP production in *C. difficile* resulted in greater filament expression and cell aggregation (Figure 1.4.9) (Bordeleau *et al.,* 2015). Regulation of TFP expression by c-di-GMP has also been observed in the Gram-negative *M. xanthus* (Skotnicka *et al.,* 2015). C-di-GMP is a secondary messenger that has been implicated in a number of processes, such

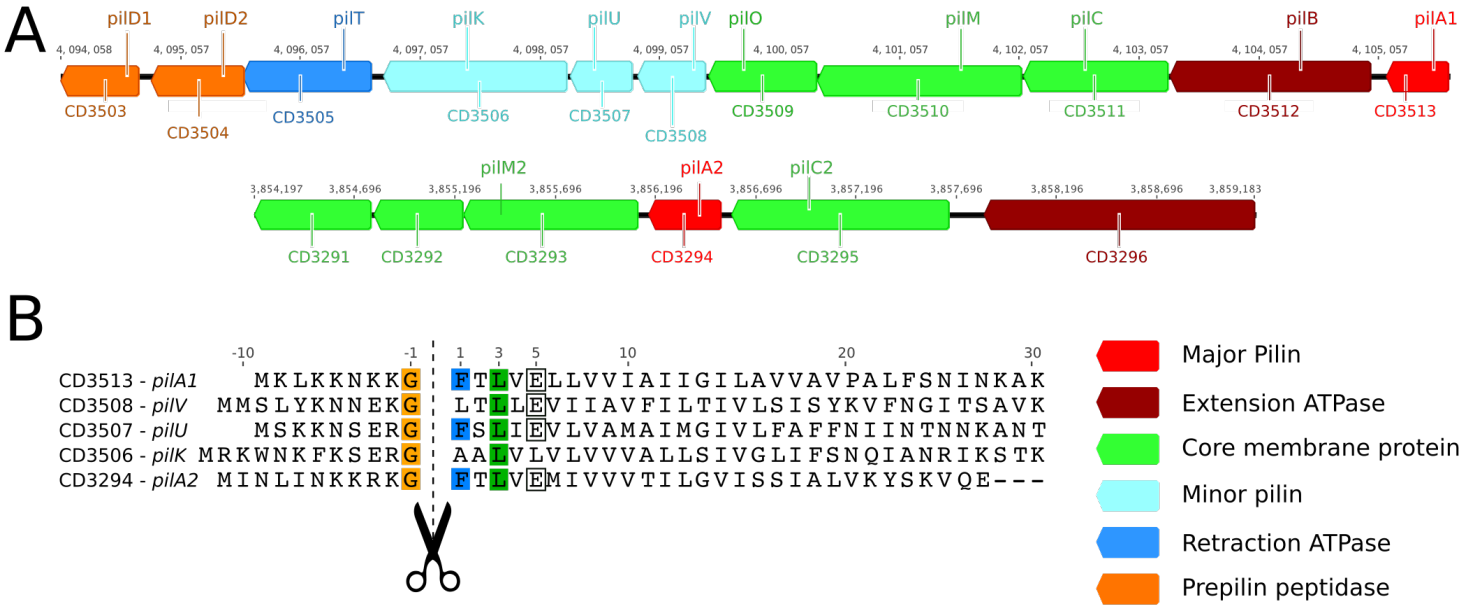


Figure 1.4.7 – :**Loci and TFP signal peptide sequence conservation in *C. difficile* 630. A:** Two loci were identified during annotation of the 630 genome. The main locus has 11 components, coloured by predicted function/localisation (see key). A secondary locus contains minimal components including a major pilin, membrane associated proteins and an extension ATPase. **B:** The N-terminal residues of the pilins from the main locus (PilA1, PilV, PilU and PilK) and secondary locus (PilA2) were aligned and show the conservation of the Gly-1 residue which is cleaved at the carboxyl side during processing by PilD. Of the mature pilins only PilA1, -A2 and -V have Phe at position 1. PilU and PilK have the hydrophobic residues Leu and Ala, respectively at position 1. Leu is completely conserved at position 3 and Glu is conserved at position 5 in all sequences except PilK.

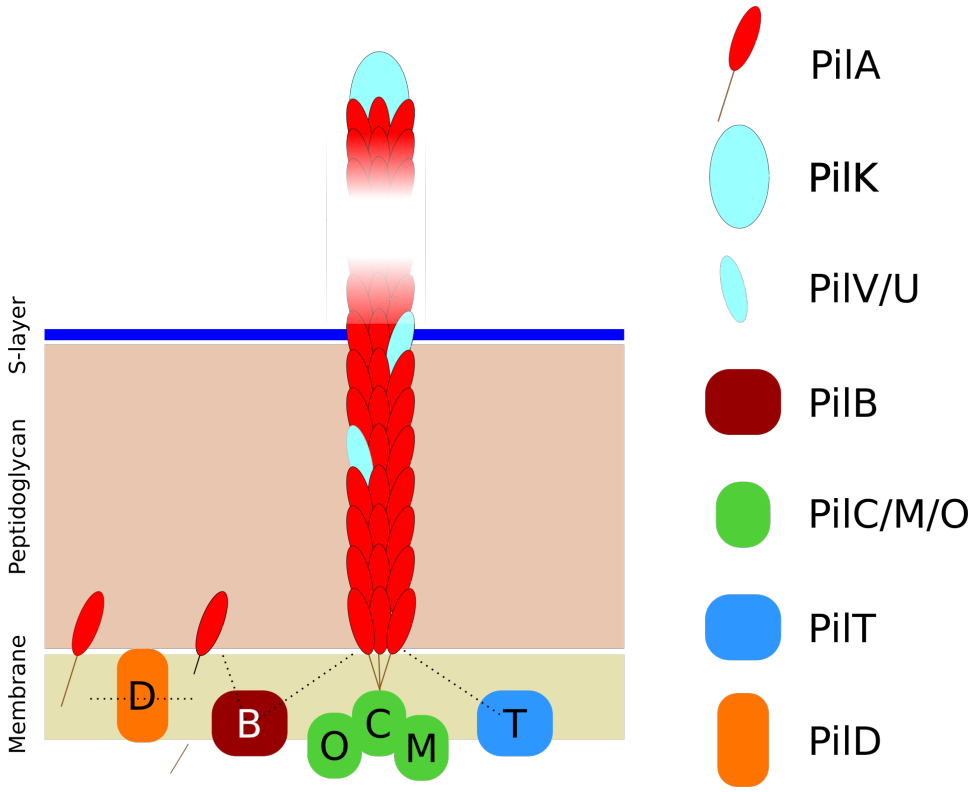


Figure 1.4.8 – **Gram-positive TFP assembly.** The predicted assembly of the Gram-positive TFP, adapted from Melville and Craig, 2013. The major pilin, PilA1 (red), is expressed on the cytosolic side of the cell membrane and translocated across the membrane and 9 pre-pilin residues are cleaved from the N-terminus. The ATPase, PilB (maroon), is predicted to enable a conformational change in the filament providing space for additional pilins and thus extend the filament. The filament base is associated with a core protein, PilC, and accessory core proteins PilM/O (green). Retraction of the filament takes place using the ATPase, PilT (blue). The filament maybe supplemented with additional minor pilins (light blue), which may play a role in surface adhesion and filament capping. The S-layer is a cell wall feature present in some Gram-positive bacteria such as *C. difficile.*

as flagella regulation, and bacterial virulence mechanisms, including toxin synthesis, in *C. difficile* (Bordeleau and Burrus, 2015). In other bacteria such as E. coli, c-di-GMP has been implicated in the expression of Type I pili and recent advances in the c-di-GMP have identified large regulatory networks that simultaneously regulate many aspects of bacterial physiology (Schirmer and Jenal, 2009; Abgottspon *et al.,* 2010; Rotem *et al.,* 2015).

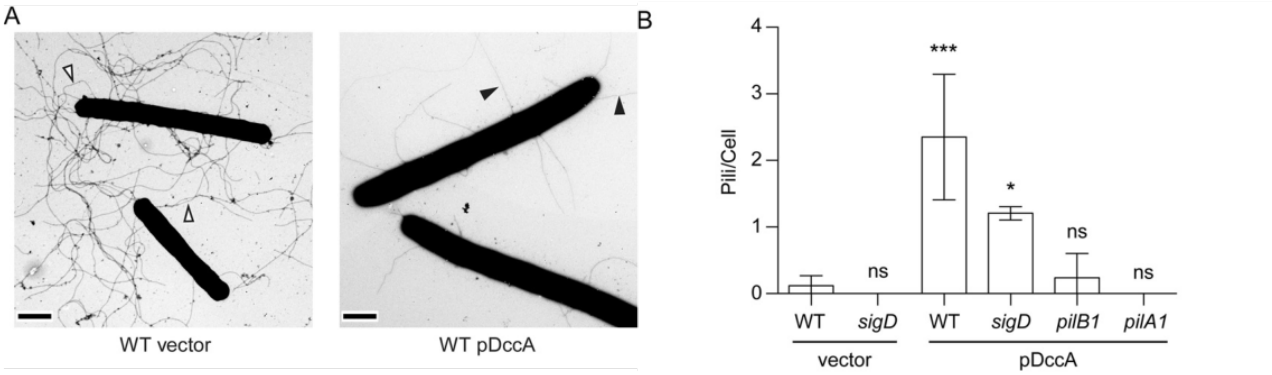


Figure 1.4.9 – **Electron micrograph of *C. difficile* TFP.** Adapted from Bordeleau et al. 2015. **A:** Transmission electron micrographs of WT *C. difficile* 630 Δ*erm* (left panel) and *C. difficile* 630 Δ*erm* with a vector containing DccA, a diguanylate cyclase protein that increases the levels of c-di-GMP in the cell (right panel). The open arrowheads highlight flagella, no TFP filaments could be identified in the wildtype. The closed arrowheads highlight TFP filaments. **B:** The mean number of pili/cell in the WT, a *sigD* mutant (that is recognised to display impaired motility) *pilB* and *pilA1* mutants where pDccA was complemented. Where the pilin components were not disrupted (WT and *sigD*) and c-di-GMP levels were increased, a greater number of TFP filaments were observed. In the *pilB* and *pilA1* mutants, aggregation was observed around the cells.

TFP have been identified as an important component for biofilm formation in *C. difficile* (Purcell *et al.,* 2015; Maldarelli *et al.,* 2016). Biofilms enable *C. difficile* to form stable colonies, an important virulence factor in CDI. *C. difficile pilA1* mutants produce thinner biofilms with less mass than wildtype colonies and TFP genes represent a large proportion of genes expressed in *C. difficile* biofilms (Maldarelli *et al.,* 2016). In addition to biofilm formation, it has been shown in the hyper virulent strain R20291 that *C. difficile* has surface-dependent motility during biofilm formation and this has been attributed to TFP (Purcell *et al.,* 2015).

In Gram-negative pathogens, such as *N. meningitidis*, the variable regions of the pilins and specific minor pilins are important for binding to target host cell surfaces and actively enable the pathogen to cause infection (Helaine *et al.,* 2005; Szeto *et al.,* 2011; Brissac *et al.,* 2012). The diversification of surface exposed regions of the pilin proteins in *C.*

difficile shows that there is antibody cross-reactivity between the minor pilins, PilV and PilU, indicating that they share similar structural features (Maldarelli *et al.,* 2014). Such results suggest that TFP pilins provide an ideal epitope in the development of vaccines against TFP producing pathogens, such as *C. difficile* (Maldarelli *et al.,* 2014).

The structure of a minor pilin, PilJ, was determined by Piepenbrink et al. 2014 (Figure 1.4.10). Expressed from a remote ORF, the *pilJ* gene does not feature in the main or secondary locus (Figure 1.4.7) although immunogold labelling of *C. difficile* TFP filaments showed that PilJ was incorporated into the filament (Piepenbrink *et al.,* 2014). The structure revealed a minor pilin with two repeating structural domains (described as N-terminal and C-terminal), each containing the typical TFP structure features: α-1 helix; α-β loop; β-sheet; D-region (C-terminal only). A Zn^{2+} ion was coordinated at the junction between the two domains by three Cys residues from the N-terminal domain and a His residue from the C-terminal domain (Figure 1.4.10). Interestingly, two of the Cys residues (81 and 111) are located close to the β-sheet, which is similar to the previously determined TFP pilins structures from Gram-negative examples that form disulphide bonds between the β-sheet and the C-terminal residues (Figure 1.4.2).

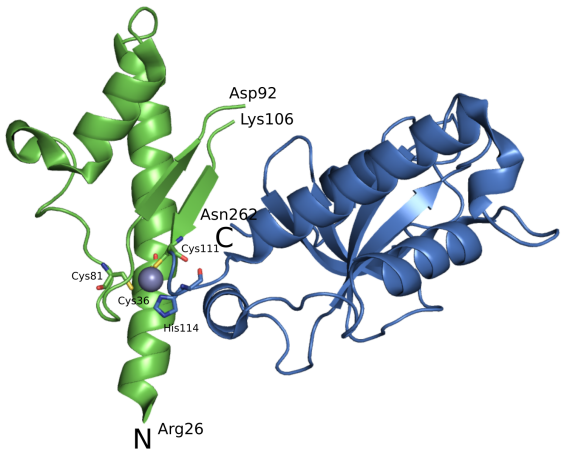


Figure 1.4.10 – *C. difficile* PilJ crystal structure. The structure of truncated *C. difficile* PilJ features an N-terminal domain (green) and C-terminal region (blue), each containing common TFP structural elements including the α-1 helix, α-β loop, β-sheet and a variable region in the C-terminal domain. 14 residues that link the two domains were not modelled. A Zn^{2+} is coordinated by three Cys residues (36, 81, 111) in the N-terminal domain and a His114 residue in the C-terminal domain and stabilises the interface between the two domains.

Piepenbrink et al. superimposed the N-terminal domain of the PilJ structure onto the structure of the major pilin TcpA from *V. cholerae* and the cryo-EM filament model proposed (Figure 1.4.3), to build a pilus model for *C. difficile* formed of PilJ pilins (Figure 1.4.11). Since PilJ is more bulky that TcpA, due to the structural repeat, a truncated PilJ using only the N-terminal region was directly modelled on the TcpA filament (Figure 1.4.11A). The truncated model was then used as the basis of the full-length PilJ filament model, resulting in a filament that had a greater diameter than the TcpA filament (Figure 1.4.11B).

Due to the presence of the structural repeat, the PilJ filament model has a greater diameter than the *V. cholera* model (Figure 1.4.11B). Importantly, this model does not include the proposed major pilin, PilA1, nor any of the other pilins present in the two TFP loci identified in *C. difficile* (Figure 1.4.7).

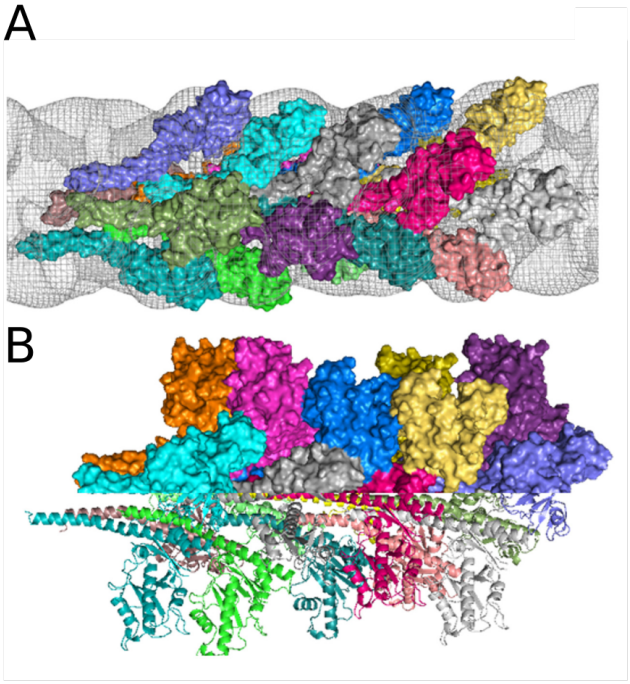


Figure 1.4.11 – Proposed *C. difficile* filament model by Piepenbrink et al. 2014. Adapted from Piepenbrink et al. 2014. The structure of the major pilin TcpA and overall TFP filament of *V. cholerae* was used to model a *C. difficile* TFP using the PilJ structure. **A:** A truncated PilJ model of only the C-terminal domain (Figure 1.4.10, green) that was superimposable with TcpA was fitted to the *V. cholerae* filament electron density with a correlation coefficient of 0.78 (grey). **B:** Full-length PilJ with a modelled N-terminal polymerisation helix was used to model a filament. This filament is bulkier than the TcpA filament (A).

1.5 Aims

The work described in this thesis aimed to improve the understanding of components of key pathogenicity related aspects of the *C. difficile* life cycle: sporulation and colonisation. Despite the importance of such mechanisms in the ability of *C. difficile* to cause and spread infections, insight at the molecular level of proteins involved in these processes is still very limited. Work presented here aimed to shed new light on our knowledge of two components: SpoIIQ:SpoIIIAH in sporulation and TFP involved in biofilm formation and colonisation.

During sporulation, coordination of the forespore and mother cell gene expression regulators is required and it has been proposed in *B. subtilis* that this takes place via a channel formed of the SpoIIQ:SpoIIIAH proteins. While the sporulation gene expression regulators are conserved between *B. subtilis* and *C. difficile,* temporal and regulatory differences have been shown between these Firmicutes upon the knowledge of *B. subtilis* SpoIIQ:SpoIIIAH two questions were posed by the *C. difficile* homologues. Does the conserved LytM domain of SpoIIQ coordinate Zn^{2+} that could be related to endopeptidase activity? Do *C. difficile* SpoIIQ and SpoIIIAH interact to form a stable 1:1 complex *in vitro*? Structural and biophysical studies have been carried out to understand possible SpoIIQ Zn^{2+} binding capability and SpoIIQ and SpoIIIAH complex formation in *C. difficile*.

Type IV pili are important structures for colonisation and biofilm formation, yet little is known about the structure of the Gram-positive TFP or how the pilin proteins may interact during pilin formation. Structural and biophysical studies of the filament major (PilA1) and minor (PilV,-U,-K) forming pilin units from *C. difficile* were undertaken to improve the understanding of the structure and function of Gram-positive TFP and *C. difficile* colonisation.

Chapter 2

Materials and Methods

2.1 Bacterial Strains

Escherichia coli DH5-α cells [New England Biolabs] were used for cloning (Section 2.3). Rosetta DE3 cells [Novagen] and BL-21 were used for protein expression (Section 2.4), the former *E. coli* strain contains a plasmid for the recognition of rare codons. The methionine auxotroph *E. coli* strain B834 [Novagen] was used for the expression of selenomethionine containing protein. A complete list of strains created in this thesis is shown in Table 2.1.1. The methods by which these strains were created are described in Section 2.3.

Strain Name	Host strain	Vector	Antibiotic resistance	Description	Acronym
42	Rosetta	NF1329	Amp + Cm	Expression strain of TEV Protease	TEV
45	DH5-α	pPSS001	Kan	Vector storage strain of pPSS001	QFL
46	DH5-α	pPSS002	Kan	Vector storage strain of pPSS002	sQ
47	DH5-α	pPSS003	Kan	Vector storage strain of pPSS003	AHFL
48	DH5-α	pPSS004	Kan	Vector storage strain of pPSS004	sAH
49	Rosetta	pPSS001	Kan + Cm	Expression strain of pPSS001	QFL
50	Rosetta	pPSS002	Kan + Cm	Expression strain of pPSS002	sQ
51	Rosetta	pPSS003	Kan + Cm	Expression strain of pPSS003	AHFL
52	Rosetta	pPSS004	Kan + Cm	Expression strain of pPSS004	sAH
55	DH5-α	pPSS007	Kan	Vector storage strain of pPSS007	sQ^{H120S}
59	Rosetta	pPSS007	Kan + Cm	Expression strain of pPSS007	sQ^{H120S}
66	DH5-α	pECC52	Kan	Vector storage strain of pECC52	PilA1-R20291
68	Rosetta	pECC52	Kan + Cm	Expression strain of pECC52	PilA1-R20291
188	Rosetta	pECC73	Kan + Cm	Expression strain of pECC73	PilV C-His
189	Rosetta	pECC74	Kan + Cm	Expression strain of pECC74	PilU C-His
190	Rosetta	pECC75	Kan + Cm	Expression strain of pECC75	PilK C-His
197	Rosetta	pJM212	Amp + Cm	Expression strain of pJM212	BS-AH
198	Rosetta	pJM243	Amp + Cm	Expression strain of pJM243	BS-Q
204	DH5-α	pADC011	Kan	Vector storage strain of pADC011	PilA1-630
205	Rosetta	pADC011	Kan + Cm	Expression strain of pADC011	PilA1-630
210	B834	pECC75	Kan	SeMet expression strain of pECC75	PilK SeMet
233	DH5-α	pADC013	Kan	Vector storage strain of pADC013	PilK N-His
234	DH5-α	pADC014	Kan	Vector storage strain of pADC014	PilU N-His
235	DH5-α	pADC015	Kan	Vector storage strain of pADC015	PilV N-His
237	BL-21	pADC012	Kan	Expression strain of pADC012	PilA1-630 N-His
238	BL-21	pADC013	Kan	Expression strain of pADC013	PilA1-630 N-His
–	BL-21	pADC015	Kan	Expression strain of pADC015	PilV N-His
–	BL-21	pADC014	Kan	Expression strain of pADC014	PilU N-His
–	Rosetta	pECC038	Kan + Cm	Expression strain of pECC038	PilA1FL-R20291

Table 2.1.1 – Summary of vector containing bacterial strains. List of vector containing bacterial strains created and used in this thesis. Production of these constructs is described in section 2.3 and the vectors are described in Table 2.3.6.

2.2 Bacterial Growth Conditions and Storage

DH5-α and Rosetta DE3 *E. coli* strains were grown in Lysogeny Broth (LB) by dissolving pellets in deionised water, under conditions that were optimised for each strain (section 2.4). Overnight cultures of every strain produced and used in this thesis were stored in a strain library: 1 ml of cell culture was mixed with 400 μl of 70% glyercol in a cryo compatible tube and stored at -80° C.

2.3 Molecular Biology

This section outlines the standard protocol for the molecular biology techniques used in this thesis for the creation of expression vectors for target proteins. Protocols are also included for the modification of these vectors to produce construct truncations or to introduce point mutations into any target. A number of constructs, used in Chapters 4 and 5 were made by Edward Couchman and Dr. Rob Fagan as part of a collaboration with Professor Neil Fairweather at Imperial College London (outlined in Table 2.3.6).

2.3.1 Expression Vector Production

All constructs used in this thesis used pET-28a or the EMBL modified pET-28a known as pET-M11 as the expression vector backbone (Dümmler *et al.,* 2005). pET-28a can be used to produce constructs encoding either N-terminal or C-terminal His-tags. Whilst the C-terminal tag is linked to the target by 2 amino acid residues (Leu and Glu), the N-terminal His-tag is linked to the target by a thrombin recognition site that enables the removal of the His-tag. The proteins encoded from a pET-M11 plasmid differ from those encoded from pET-28a in that the thrombin cleavage linker of the latter has been replaced by a Tobacco Edge Virus (TEV) protease cleavage site. The resulting cleaved protein has a Gly-Ser linker at the N-terminus of the polypeptide after cleavage.

Primers and vectors were designed using the Geneious software package (Kearse *et al.,* 2012) to ensure that restriction sites and primer T_m temperatures were appropriate and that gene inserts were in frame. All *C. difficile* gene containing vectors were designed based on published annotations of strain 630 (Sebaihia *et al.,* 2006) and strain R20291

(Stabler *et al.,* 2009). Primers were designed to be ~30 bases long so that the T_m of the primer was ~59° C and with a GC anchor at the 5‘ end. In addition to a complementary sequence with the target gene, an appropriate restriction site was included in the primer that was unique for the forward and reverse primer to allow for insertion into the expression vector in the correct orientation. Unmodified oligonucleotides were ordered from Eurofins and upon delivery were dissolved in a volume of nuclease free water [Ambion] to a final concentration of 100 pmol/μl as instructed by the supplied synthesis report. The primers used in this thesis are described in Table 2.3.1.

The reagents for the polymerase chain reaction (PCR) are outlined in Table 2.3.2. The PCR reaction was carried out using a PCR Thermocycler [Bio-Rad] and the reaction cycle is outlined in Table 2.3.3. The amplification time was modified according to the length of the fragment being amplified.

The products of the PCR reaction were examined by running 5 μl on a 1% agarose [Melford] gel containing 1X SYBR safe gel stain [Life Technologies] electrophoresed at 100V for ~45 mins. The agarose gel was visualised under ultra violet (UV) light using a UV gel viewing system [Bio-Rad]. The remainder of the reaction was cleaned up using a PCR clean up kit [Sigma] and the amplified DNA eluted in 30 μl of nuclease free water. The concentration of purified PCR product was determined using a NanoDrop spectrophotometer [Thermo]. Purified PCR products and the expression vector were digested with the appropriate restriction enzymes [Invitrogen], using reactions described in Table 2.3.4.

Digested PCR product and vector were analysed by 1% agarose gel and cleaved products of the correct size were extracted using a DNA gel extraction kit according to the manufacturer’s instructions [Sigma]. Ligation reactions were set up using T4 quick ligase [Thermo] as described in Table 2.3.5, and incubated at room temperature for 15 minutes. Calcium chloride competent *Escherichia coli* DH5-α cells [New England Biolabs] were transformed by the addition of 2 μl of ligation product to a 50 μl aliquot of cells and incubated for 20 mins on ice before a heat shock at 42° C for 30 seconds. 300 μl of LB was added to the competent cells which were incubated with shaking at 37° C for 1 hour. Transformed colonies were selected for by plating onto LB agar plates containing 50 μg/ml kanamycin (Kan), which were incubated overnight at 37° C.

Primer Name	Primer Sequence	Restriction Site	Primer Description
FWDAHFL	GATCCCATGGGCAAGTTT AATTATAAG	NcoI	Forward primer to amplify full length *SpoIIIAH* (CD1199) gene.
FWDAHSOL	GTCCCATGGGCTTAAGTAA GAAATC	NcoI	Forward primer to amplify the bases for residues 29-229 of *SpoIIIAH*.
REVAH	GATCCTCGAGTTACTTATTA CTATTATT	XhoI	Reverse primer to amplify both *SpoIIIAH* construct fragments.
FWDQFL	GATCCCATGGGCAAGAAAA AGCTGTTAG	NcoI	Forward primer to amplify the full-length *SpoIIQ* (CD0125) gene.
FWDQSOL	GATCCCATGGGCAATAATA ATGTAG	NcoI	Forward primer to amplify the bases for residues 31-222 of *SpoIIQ*.
REVQ	GACCTCGAGTTACTTAATT AGACTCATTGG	XhoI	Reverse primer to amplify both *SpoIIQ* construct fragments.
REV-PilV	GTGCTCGAGCTACCCTATT CTT	XhoI	Reverse primer to amplify *PilV* (CD3508) residues 36-188 with N-His tag.
FWD-PilV	ATATACCATGGTGGCAATG GTATAAC	NcoI	Forward primer to amplify *PilV* (CD3508) residues 36-188 with N-His tag.
REV-PilU	GTGGTGCTCGAGCTATTTA TCTTTAA	XhoI	Reverse primer to amplify *PilU* (CD3507) residues 34-176 with N-His tag.
FWD-PilU	ATATACCATGGCAATACTAA TAACAAAGC	NcoI	Forward primer to amplify *PilU* (CD3507) residues 34-176 with N-His tag.
REV-PilK	GTGCTCGAGCTAGTTTACT TTTTTG	XhoI	Reverse primer to amplify *PilK* (CD3506) residues 33-512 with N-His tag.
FWD-PilK	ATATACCATGGATCAAATCA GATAGC	NcoI	Forward primer to amplify *PilK* (CD3506) residues 33-512 with N-His tag.
REV-PilA1-630	GTGCTCGAGCTATCCTTGT TG	XhoI	*PilA1* $\triangle$1-34 CD630 N-His reverse primer
FWD-PilA1-630	GAGATATACCATGGAAGTAA TATAAACAAG	NcoI	*PilA1* $\triangle$1-34 CD630 N-His forward primer

Table 2.3.1 – **Summary of primers.** Primers used for the amplification of target genes from *Clostridium difficile* strain 630 for subsequent insertion into expression vectors. The restriction sites in these primers are also described.

	Volume	Final Concentration
Forward Primer [Eurofins-MWG]	1.5 µl	0.3 pmol
Reverse Primer [Eurofins-MWG]	1.5 µl	0.3 pmol
Template DNA	X µl	10 ng (plasmid) / 100 ng (genomic)
KOD Hot Start Master Mix (2X) [Merck Millipore]	25 µl	1X
Nuclease Free Water [Ambion]	to a final total volume of 50 µl	–

Table 2.3.2 – **PCR reaction reagents.** Components in a typical PCR reaction. Primers were stored at stock concentrations of 100 pmol and working aliquots of 10 pmol. The volume of template DNA (X) was dependent upon the concentration of the template DNA and the type of DNA. All PCR reactions had a total volume of 50 µl.

Stage	Time	Temperature (°C)
1 - Polymerase activation	2 min	95
2 - DNA Denaturation	20 secs	95
3 - Annealing	10 secs	X
4 - Extension	25 secs*	70
Steps 2-4 repeated in 30 cycles		

Table 2.3.3 – **PCR reaction cycle.** PCR reaction cycle used for KOD Hot Start polymerase [Merck-Millipore]. *Extension time was varied depending on the length of the DNA fragment that was being amplified: 15 secs for <500 kbp; 20 secs for up to 3000 kbp; 25 secs for >3000 kbp.

Reagent	Volume
XhoI	2 µl
NcoI	2 µl
10X Cutsmart buffer	5 µl
DNA	X
Nuclease free water	up to 50 µl

Table 2.3.4 – **DNA restriction digest reaction.** A typical restriction double digest reaction where NcoI and XhoI have been used to digest DNA fragments or vectors. The volume of DNA (X) was adjusted according to a maximum final concentration of 2 µg. Restriction enzymes and buffer are part of the New England Biolabs DNA restriction system.

Reagent	Volume
Insert DNA	X μl
Linear Vector DNA	Y μl
5X ligase buffer	4 μl
Ligase	1 μl
H_2O	up to 20 μl

Table 2.3.5 – Components for ligation reaction. The volume of insert DNA (X) and linearised vector DNA (Y) was dependent upon the concentrations of these fragments. Insert DNA was 3-fold the concentration of linearised vector DNA (3:1) of which a total of 50 ng was used in the ligation reaction.

Single colonies from the transformation plates were picked and grown overnight at 37° C in 10 ml LB 50 μg/ml Kan. 1 ml of overnight culture was mixed with 400 μl of 70% glycerol and added to the strain collection at -80° C. The remaining culture was pelleted and the plasmid extracted using a mini-prep kit as per the manufacturer's instructions [Sigma]. Plasmid DNA was eluted from the kit in 40 μl nuclease-free water and 20 μl sent for Sanger DNA sequencing [GATC Biotech]. After sequence verification, 2 μl of the plasmids containing the correct genes were used to transform *E. coli* Rosetta expression cells as previously described, and positive colonies were selected by growth on LB agar 50 μg/ml Kan and 30 μg/ml chloramphenicol (Cm). 10 ml cultures grown overnight in LB with Kan and Cm were added to the strain collection as described.

Construct Name	Vector Name	Vector Back-bone	Description
QFL	pPSS001	pET-M11	Contains the full CD630_0125 SpoIIQ , including N-terminal transmembrane domain and signal peptide. N-terminal His-tag cleavable using TEV protease.
sQ	pPSS002	pET-M11	CD630_0125 SpoIIQ residues 31-222 lacking N-terminal transmembrane domain and signal peptide. This region is the extracellular domain of the protein. N-terminal His-tag cleavable using TEV protease.
sQ^{H120S}	pPSS007	pET-M11	As sQ construct but with a point mutation of histidine 120 to serine as per *B. subitilis* SpoIIQ.
AHFL	pPSS003	pET-M11	Contains the full CD630_1199 SpoIIIAH gene, including N-terminal transmembrane domain and signal peptide. N-terminal His-tag cleavable using TEV protease.
sAH	pPSS004	pET-M11	Contains residues 29-229 of the CD630_1199 SpoIIIAH gene, the extracellular domain. N-terminal His-tag cleavable using TEV protease.
BS-Q	pJM243	pET-M11	Contains residues 25-218 of *B. subtilis* 068 SpoIIQ gene with thrombin cleavable N-terminal His-tag.
BS-AH	pJM212	pET-M11	Contains residues 43-283 of *B. subtilis* 068 SpoIIQ gene with thrombin cleavable N-terminal His-tag.
PilA1△1-34 R20291	pECC52	pET-28a	PilA1 (CD3355) from *C. difficile* strain R20291 lacking the N-terminal 34 residues (hydrophobic oligmerisation domain). Non-cleavable N-terminal His-tag.
PilA1△1-34 CD630 (C-His)	pADC011	pET-28a	PilA1 (CD3513) from *C. difficile* strain 630 lacking the N-terminal 34 residues (hydrophobic oligmerisation domain). Non-cleavable C-terminal His-tag.
PilA1△1-34 CD630 (N-His)	pADC012	pET-28a	PilA1 (CD3513) from *C. difficile* strain 630 lacking the N-terminal 34 residues (hydrophobic oligmerisation domain). Non-cleavable N-terminal His-tag.
PilA2△1-33 R20291	pECC038	pET-28a	PilA2 (CD3155) from *C. difficile* strain 630 lacking the N-terminal 34 residues (hydrophobic oligmerisation domain). Non-cleavable C-terminal His-tag.
PilV△1-35 (C-His)	pECC73	pET-28a	PilV from *C. difficile* strain 630 lacking the N-terminal 35 residues (hydrophobic oligmerisation domain). Non-cleavable C-terminal His-tag.
PilV△1-35 (N-His)	pADC015	pET-28a	PilV from *C. difficile* strain 630 lacking the N-terminal 35 residues (hydrophobic oligmerisation domain). Non-cleavable N-terminal His-tag.
PilU△1-33 (C-His)	pECC74	pET-28a	PilU from *C. difficile* strain 630 lacking the N-terminal 33 residues (hydrophobic oligmerisation domain). Non-cleavable C-terminal His-tag.
PilU△1-33 (N-His)	pADC014	pET-28a	PilU from *C. difficile* strain 630 lacking the N-terminal 33 residues (hydrophobic oligmerisation domain). Non-cleavable N-terminal His-tag.
PilK△1-32 (C-His)	pECC75	pET-28a	PilK from *C. difficile* strain 630 lacking the N-terminal 32 residues (hydrophobic oligmerisation domain). Non-cleavable C-terminal His-tag.
PilK△1-32 (N-His)	pADC013	pET-28a	PilK from *C. difficile* strain 630 lacking the N-terminal 32 residues (hydrophobic oligmerisation domain). Non-cleavable N-terminal His-tag.
TEV	NF1329	pET-28a	His-tagged TEV protease with S219V mutation that prevents self-cleavage.

Table 2.3.6 – Summary of construct vectors.

2.3.2 Modification of expression vectors

For the production of point mutants and the truncated constructs, inverse PCR was used. Forward and reverse primers flanking the bases to be substituted or deleted were designed based on the template expression plasmid. The primers were designed with a guanine or cytosine 3'-end and with an overall melting temperature (T_m) of ~58° C. Table 2.3.7 shows the primers used for inverse PCR. The PCR reagents and reaction used are outlined in Table 2.3.2 and Table 2.3.3, a longer extension time of 45 seconds was used to ensure amplification of the complete plasmid.

Primer Name	Sequence	Description
–	GCCAAAGGTGTAGATATTAGTTGTACTAAAG	Forward primer for inverse PCR to mutate H120S of CD0125 SpoIIQ
–	AGCAAAGGTGTAGATATTAGTTGTACTAAAG	Reverse primer for inverse PCR to mutate H120S of CD0125 SpoIIQ
185	AGTAATATAAACAAGGCTAAGGTAGC	Reverse primer for inverse PCR to truncate CD3513 in pRPF227
186	CATGGTATATCTCCTTCTTAAAGTTAAAC	Forward primer for inverse PCR to truncate CD3513 in pRPF227

Table 2.3.7 – Summary of inverse PCR primers. Primers used for inverse PCR to modify expression vectors.

2.4 Protein Expression

Single colonies of transformants or the strain glycerol stock were picked and used to inoculate 100 ml LB containing 50 μg/ml kanamycin (Kan) and 30 μg/ml chloramphenicol (Cm) and grown overnight at 37° C. 1L LB containing 50 μg/ml Kan and 30 μg/ml Cm were inoculated with 10 ml of overnight culture and cultured at 37° C until an optical density at 600 nm (OD_{600}) of 0.4-0.6 was reached. The OD_{600} was measured by aliquoting 1 ml of culture into a 1 ml cuvette [Fisherbrand] and its absorbance measured using a spectrophotometer [Biochrom] with the wavelength set to 600 nm; 1 ml of sterile LB was used as a reference sample. Cultures were induced with a final concentration of 1mM isopropyl β-D-1-thiogalactopyranoside (IPTG). Cells were incubated at 20° C or 25° C with shaking at 180 rpm for ~20 hrs (overnight). Cells were harvested by centrifugation at 3036 *x g* for 30 mins, resuspended in the appropriate lysis buffer (see Table 2.5.1) and centrifuged at 3011 *x g* for a further 10 mins and the supernatant discarded. Cell pellets were stored at -80° C.

The optimal expression conditions of new expression strains was determined by varying the growth medium, temperature and expression time (time after induction with IPTG) in small scale cultures (10 ml). Typically LB, 2x Yeast Tryptone (YT) and auto-induction [Novagen] media were probed for ideal expression conditions. The combinations of temperature and expression time using LB and YT were as follows: 37° C, 3 hrs; 30° C, 3hrs; 30° C, ~20 hours (overnight); 20° C, ~20 hrs (overnight). Cultures grown in auto-induction were inoculated from overnight starter cultures and incubated at either 30° C or 20° C for ~24 hrs. Cultures in LB and YT media were grown to an OD_{600} of 0.4-0.6, a 1ml culture aliquot was taken and frozen at -20° C (uninduced sample) before induction with a final concentration of 1 mM IPTG. Cultures were harvested by centrifugation at 4000 *x g* for 10 mins and lysed using Bugbuster lysis solution [Merck-Millipore]. The soluble and insoluble fractions of the lysate were separated by centrifugation at 17000 *x g* for 10 mins. The supernatant was removed and 10 μl was mixed with 10 μl of 2X SDS-sample buffer and 10 μl of these samples were loaded on SDS-PAGE(section 2.5.3). The uninduced samples and the insoluble pellets were resuspended in 10 μl of SDS-sample buffer and boiled at 95° C for 5 mins and 5 μl of these samples were loaded on SDS-PAGE with the soluble fractions. The conditions that produced the greatest amount of protein in the soluble fraction were

chosen as the conditions for large scale expression. All strains were cultured at 20° C with shaking at 180 rpm for ~20 hrs (overnight) apart from TEV protease which was cultured at 25° C with shaking at 180 rpm for ~20 hrs (overnight).

2.4.1 Selenomethionine Incorporation

Stock Reagent	Volume (ml)	Final Concentration
20X M9 Salts	100	2X
1 M $MgSO_4$	2	2 mM
$FeSO_4$ (12.5 mg/ml)	2	12.5 μg/ml
40% w/v D-glucose	10	0.4 %
Amino Acids I*	10	0.4 μg/ml
Amino Acids II†	10	0.4 μg/ml
1000X Vitamins‡	1	1X
Se-Met (10 mg/ml)	4	10 μg/ml
Deionised H_2O	861	–

Table 2.4.1 – **Composition of SeMet media.** 20X M9 salts contained 2% NH_4Cl, 6% KH_2PO_4 and 12% Na_2HPO_4. *Amino acids I includes 0.4 g of the following amino acids in 100 ml H_2O: Ala; Arg; Asn; Asp; Cys; Gln; Glu; Gly; His; Ile; Leu; Lys; Pro; Ser; Thr; Val. †Amino acids II includes 0.4 g in 100 ml of the following amino acids: Phe; Trp; Tyr. ‡Vitamin 1000X stock includes 1mg/ml of the following: niacinamide; pryoxidine monohydrochloride; riboflavin; thiamine.

An overnight culture of *E. coli* B834, transformed with the target expression vector, was grown as described in section 2.2. 1 ml of the overnight culture was used to inoculate 100 ml of LB containing the appropriate antibiotics and was grown at 37° C with shaking at 180 rpm, to an OD_{600} of ~0.2. The starter culture was pelleted by centrifugation at 3011 *x g* for 10 mins and the LB supernatant discarded. The pellet was resuspended in 10 ml of the SeMet medium (Table 2.4.1) containing all components including appropriate antibiotics, with exception of the selenomethionine and was again pelleted by centrifugation as before to remove any remaining LB. The pellet was again resuspended in 10 ml of SeMet medium before inoculating a 1 L flask. A 10 ml control culture was removed and placed into a 50 ml falcon to be incubated alongside the expression culture. The selenomethionine was

added to the culture before incubation at 37° C to an OD_{600} of 0.4-0.6, induced with 1 mM IPTG, and growth conditions used for native protein expression were adopted.

2.4.2 ^{15}N Incorporation

For 2D NMR ^{1}H:^{15}N HSQC experiments (Section 2.8) proteins must be labelled with the isotope ^{15}N. Expression strains producing the protein to be used in the HSQC experiments were picked from the glycerol strain library and used to inoculate a 100 ml overnight start culture as described in section 2.4. 10 ml of these starter cultures were used to inoculate M9 minimal medium supplemented with $^{15}NH_4Cl$ (Table 2.4.2) and grown to an OD_{600} of 0.4-0.6 initially at 37° C before induction with 1 mM IPTG. The minimal nature of the growth medium and the presence of $^{15}NH_4Cl$ resulted in very slow cell culture growth such that >24 hrs of incubation was required to reach an OD_{600} of 0.4-0.6. To promote growth, the incubation temperature was reduced to 30° C and the culture was supplemented with 10 ml of 10% glucose. Once an appropriate OD_{600} was reached an uninduced sample was taken, the culture was supplemented with 1 mM IPTG, cultured and harvested as previously described (section 2.4).

Stock Reagent	Volume (ml)	Final Concentration
2X M9 Salts*	500	1X
1 M $MgSO_4$	2	2 mM
20% Glucose	20	0.4 %
1000X Vitamins†	1	1X
100X Trace elements‡	10	1X
100 mg/ml $^{15}NH_4Cl$	10	1 mg/ml
H_2O	457	–

Table 2.4.2 – Composition of 15^N incorporation media. *2X M9 salts contained 0.6% KH_2PO_4 and 1.2% Na_2HPO_4. †Vitamin 1000X stock includes 1mg/ml of the following: niacinamide; pryoxidine mono-hydrochloride; riboflavin; thiamine. ‡100X trace elements: 5 mg/ml EDTA; 0.8 mg/ml $FeCl_3$; 50 μg/ml $ZnCl_2$; 10 μg/ml $CuCl_2$; 10 μg/ml $CoCl_2$; 10 μg/ml H_3BO_3; 1.6 mg/ml $MnCl_2$.

2.5 Protein Purification

2.5.1 Buffers

The buffers used in the purification and study of proteins in this thesis are outlined in Table 2.5.1. Generally, the buffer that was used for gel filtration during purification of the protein was used for crystallisation and most biochemical and biophysical assays with the exception of buffer sensitive techniques such as circular dichroism (CD) or nuclear magnetic resonance (NMR). The theoretical isoelectric point, calculated using ProtParam (Wilkins *et al.,* 1999), was considered in the selection of buffer pH, which also informed the choice of buffering compound dependent on their optimal buffering pH range. Characterisation by CD also informed the choice of buffers: scans were conducted of samples at different pH strengths to determine the pH at which the protein exhibited the highest level of foldedness. This characterisation was important in the selection of buffers for studying sQ and sAH proteins. Finally, the buffers must also be compatible with any co-factors that the proteins being studied may require, such as Zn^{2+} in sQ.

Stage of purification/ Experiment	sQ/sAH	BS-Q/BS-AH	pilA1/A2	pilV/U/K	TEV Protease
Lysis/ His-Trap Load	50 mM MES pH 6.0, 500 mM NaCl	50 mM Tris-HCl pH 7.5, 500 mM NaCl	20 mM Tris-HCl pH 8.0, 500 mM NaCl	50 mM Tris-HCl pH 8.0, 500 mM NaCl	50 mM Na_2HPO_4 pH 8.0, 150 mM NaCl, 25 mM Imidazole, 10% glyercol
His-Trap Elution	50 mM MES pH 6.0, 500 mM NaCl, 500 mM Imidazole	50 mM Tris-HCl pH 7.5, 500 mM NaCl, 500 mM Imidazole	20 mM Tris-HCl pH 8.0, 500 mM NaCl, 500 mM Imidazole	50 mM Tris-HCl pH 8.0, 500 mM NaCl, 500 mM Imidazole	50 mM Na_2HPO_4 pH 8.0, 150 mM NaCl, 800 mM Imidazole, 10% glycerol
Gel Filtration	50 mM MES pH 6.0, 250 mM NaCl	50 mM Tris-HCl pH 7.5, 150 mM NaCl	20 mM Tris-HCl pH 8.0, 150 mM NaCl	50 mM Tris-HCl pH 7.5, 250 mM NaCl	–
SEC-MALLS	50 mM MES pH 6.0, 250 mM NaCl	–	–	–	–
Circular Dichroism	50 mM Na_2HPO_4 pH 6.0, 50 mm NaF	–	50 mM Na_2HPO_4 pH 8.0, 50 mM NaF	50 mM Tris-HCl pH 7.5, 50 mM NaF	–
NMR	50 mM Na_2HPO_4 pH 6.0, 150 mm NaCl	–	–	–	–

Table 2.5.1 – **Summary of protein buffers.** A summary of the buffers used in the purification of the proteins studied in this thesis.

2.5.2 Protein Concentration Determination

Protein concentration was determined using three methods: absorbance at 280 nm; Bradford assay; and Pierce bicinchoninic acid (BCA). Absorbance at 280 nm was measured using a NanoDrop spectrophotometer [Thermo]. Using the Beer-Lambert law, $A = \varepsilon \times c \times l$, and an extinction coefficient (ε) estimated by submitting the peptide sequence to the ProtParam tool (web.expasy.org) and a path-length of 1 cm (NanoDrop normalises the path-length to 1 cm), the concentration of a protein sample could be calculated.

Bradford assays were performed by mixing 100 μl of protein sample with 100 μl Bradford reagent [Thermo] in a 96-well plate [Corning] and absorbance was measured at 595 nm in a plate-reading spectrophotometer [BioTek]. Absorbance values were normalised using the absorbance of 100 μl of buffer mixed with 100 μl of Bradford reagent. A standard curve using bovine serum albumin (BSA) with concentrations of 0 μg/ml, 1 μg/ml, 2 μg/ml, 5 μg/ml, 10 μg/ml, 15 μg/ml and 20 μg/ml was recorded. The absorbance values from the standard curve were used to calculate the concentration of the protein samples. A BCA assay kit [Thermo] was used to determine the concentration of samples using the Pierce BCA method and was performed as per the manufacturer's instructions.

2.5.3 SDS-PAGE Analysis of proteins

Denaturing sodium dodecyl sulphate (SDS) polyacrylamide gel electrophoresis (PAGE) was used throughout to identify and verify the quality of protein samples. For appropriate separation of the proteins studied in this thesis, 15% acrylamide gels were used and are described in Table 2.5.2. Samples were mixed with 2X SDS-sample buffer prior to loading on a gel (Table 2.5.2). SDS-PAGE was performed at 200V for 50-70 mins depending upon the mass of the sample.

Reagent	Resolving gel	Stacking gel	2X Sample buffer
Bis-acrylamide	15%	5%	–
Tris-HCl pH 8.8	375 mM	–	–
Tris-HCl pH 6.8	–	126 mM	100 mM
SDS	0.1%	0.1%	4%
Ammonium persulphate	0.1%	0.1%	–
TEMED	0.05%	0.1%	–
Glycerol	–	–	20%
Bromophenol Blue	–	–	0.2%
DTT	–	–	200 mM

Table 2.5.2 – **SDS-PAGE reagents.** Composition of reagents in 15% acrylamide SDS-PAGE gels and 2X sample buffer.

2.5.4 sQ/sAH Protein Purification

Cell pellets were resuspended in 30 of ml 50 mM MES pH 6.0, 500 mM NaCl. 50 μl of lysozyme (1 mg/ml), 50 μl DNase I (20 μg/ml) and 1 EDTA-free complete protease inhibitor tablet [Roche] were added to the cell suspension. Cell lysis was carried out by sonication [Bandelin] for 4 mins with 50% intervals, 75-80% sonication power, and the lysate clarified by centrifugation using a JA-25 rotor in an Avanti J-26XP centrifuge [Beckman-Coulter] for 30 mins at 32816 *x g*. Lysis supernatant was filtered through a 0.45 μm filter.

The supernatant was loaded to a 5 ml His-Trap NTA column [GE Healthcare] using an Äkta Prime [GE Healthcare]. The column was washed with 25 ml of 50 mM MES pH 6.0, 500 mM NaCl to remove unbound protein. His-tagged proteins were eluted with 50 mM MES pH 6.0, 500 mM NaCl, 500 mM imidazole using steps of 125 mM, 250 mM and 375 mM imidazole, over 40 ml, and 2 ml fractions were collected. Fractions were analysed by SDS-PAGE (Section 2.5.3) to identify protein-containing fractions. These fractions were pooled and the protein concentration determined using a NanoDrop [Thermo]. Previously purified TEV protease was added at the ratio of 1 mg of TEV protease for every 10 mg of purified protein. The digestion mixture was dialysed overnight at 4° C against 50 mM MES pH 6.0, 250 mM NaCl.

To separate cleaved, untagged protein from His-tagged protein and TEV protease the cleavage reaction was run over a 5 ml His-Trap pre-equilibrated in 50 mM MES pH 6.0, 250

mM NaCl. Unbound proteins (the flow-through) were collected in 4 ml fractions and the remaining His-tagged proteins, including TEV were eluted in 500 mM imidazole. The flow through fractions were pooled and reduced to a volume of 1 ml using a 10 kDa molecular weight cut off (MWCO) centrifugal concentrator [Amicon] at 3011 *x g*. For sQ protein, the concentration of the pooled fractions was measured before concentration and a 5-fold molar excess of $ZnCl_2$ was added to the sQ sample, except for protein intended for ICP-MS experiments. To remove any potential precipitant, samples were centrifuged at 1600 *x g* for 2 mins. Samples were then loaded into a sample loop of an Äkta Pure [GE Healthcare] and injected onto a Superdex 200 (S200) 26/600 size exclusion column [GE Healthcare] pre-equilibrated in 50 mM MES pH 6.0, 250 mM NaCl. Fractions of 2 ml were collected and peak fractions were analysed by SDS-PAGE. Fractions containing the protein were pooled and concentrated to ~10 mg/ml as previously described and flash frozen in liquid nitrogen and stored at -80° C.

For the purification of the *B. subtilis* BS-Q and BS-AH proteins the protocol was followed as for the *C. difficile* sQ and sAH with the exception that thrombin protease [Sigma-Aldrich] was used to cleave the His-tag during dialysis. These proteins were purified in Tris-HCl based buffers (Meisner and Moran, 2011).

2.5.5 Pilin Protein Purification

Cell pellets were resuspended in 10 ml of 50 mM Tris-HCl pH 8.0, 500 mM NaCl and 50 μl DNase I (20 μg/ml) and a complete EDTA-free protease inhibitor tablet [Roche] added. Cells were lysed using a one-shot cell disruptor [Constant Systems] at a pressure of 25 kpsi. The lysate was clarified by centrifugation at 32816 *x g* for 30 mins and the supernatant filtered using a 0.45 μm syringe filter [Merck].

The lysate was loaded to a 5 ml His-Trap HP column [GE Healthcare] connected to an Äkta Start [GE Healthcare] and pre-equilibrated with 50 mM Tris-HCl pH 8.0, 500 mM NaCl. His-tagged proteins were collected in 2 ml fractions during elution with 50 mM, 125 mM, 250 mM, 375 mM and 500 mM imidazole in five steps of 15 ml. Peak fractions were analysed by SDS-PAGE (Section 2.5.3), which were then pooled and the volume reduced to ~2 ml using a 3 kDa MWCO concentrator [Amicon] for PilA1, V or U constructs or a 30 kDa MWCO for PilK. The sample was then loaded to a size exclusion column equilibrated

with 50 mM Tris-HCl pH 8.0, 250 mM NaCl connected to an Äkta Pure FPLC system [GE Healthcare]. Either a S200 or Superdex 75 (S75) in the 16/600 format [GE Healthcare] was used dependent upon the size of the construct being purified. Peak fractionation was used to collect 1 ml fractions when the UV absorbance was greater than 5 mAU. These fractions were analysed by SDS-PAGE and fractions containing pure protein were pooled and concentrated to ~15 mg/ml as previously described. Concentrated protein was divided into 2 sets of aliquots, one was flash cooled in liquid nitrogen and stored at -80° C and the other stored at 4° C.

2.5.6 TEV Protease Purification

Tobacco Etch Virus (TEV) protease was produced in the lab for use in protein purification. The protease was expressed in Rosetta DE3 cells that were transformed with the NF1329 plasmid. This plasmid encodes for a His-tagged TEV protease with an S219V mutation that prevents self-cleavage of the protein. TEV was over-expressed as detailed in section 2.4. Cells were resuspended in the TEV lysis buffer (Table 2.5.1) with 50 μl of DNase I (20 μg/ml) and sonicated for 4 minutes in 50% intervals at ~75% power. The lysate was clarified by centrifugation at 32816 *x g* for 30 mins and the supernatant filtered with a 0.45 μm syringe filter before being loaded onto a 5 ml His-Trap NTA column [GE Healthcare] connected to a an Äkta Start [GE Healthcare] and prequilibrated in the TEV lysis buffer. TEV protease was eluted from the column using a gradient from 0-60% of the elution buffer (Table 2.5.1). Fractions were collected in tubes containing DTT to a final concentration of 1 mM to stabilise the TEV. The purity of peak fractions was verified by SDS-PAGE analysis, as described previously, and fractions containing TEV were pooled, and the concentration determined by absorbance at 280 nm (NanoDrop). The protein was concentrated or diluted as necessary to a final concentration of 0.5-1 mg/ml and flash cooled in liquid nitrogen in 500 μl aliquots and stored in the -80° C freezer.

2.6 Biophysical and biochemical characterisation of proteins

2.6.1 Size-exclusion chromatography for protein interactions

Size-exclusion chromatography (SEC) was used to probe potential protein:protein interactions. The individual proteins and mixtures of proteins were prepared at a final total protein concentration of 2 mg/ml in a volume of 500 μl and injected onto an S200 10/300 GL Increase [GE Healthcare] connected to an Äkta Pure [GE Healthcare] and equilibrated in the appropriate buffer for the proteins being assayed (see Table 2.5.1). Fractions of 0.5 ml were collected and 10 μl of peak fractions were mixed with 10 μl of SDS sample buffer and analysed by SDS-PAGE for co-elution of proteins indicative of complex formation. UV elution profiles were compared between individual and mixtures of proteins and with the profiles of standard proteins to determine interactions. The molecular weight standards were thyroglobulin (670 kDa), γ-globulin (158.0 kDa), ovalbumin (44 kDa), myoglobin (17.0 kDa) and Vitamin B_{12} (1.4 kDa) [Bio-rad]. The equation below was used to determine K_{av}.

$$K_{av} = \frac{V_e - V_o}{V_c - V_o}$$

Where V_o = column void volume, V_e = elution volume, V_c = geometric column volume. K_{av} for each peak was plotted versus the log of the molecular mass (LogMW) of the corresponding peak and the equation of the line of best fit was calculated. The K_{av} was calculated for each elution peak and the equation from the line of best fit used to calculate the LogMW for the elution peak and was converted to the natural number to calculate the equivalent molecular mass of the elution peak.

2.6.2 Circular Dichroism

Circular dichroism (CD) is a technique for determining the folded state of a protein in solution. Buffers that contain Tris, MES and NaCl absorb far UV-light and as a result are not ideal sample buffers for CD. Since the purification buffers used in this thesis (Table 2.5.1) contain these compounds, samples to be used in CD were buffer exchanged into a phosphate based buffer. A number of buffers were used at different pH within the buffering

pH range of the buffer (Na_2HPO_4: pH 5.8-8.0). Samples were buffer exchanged using S6 Bio-spin buffer exchange columns as per the manufacturer's instructions [Bio-rad]. The packing buffer was replaced with 4 washes of the CD buffer described in Table 2.5.1 for the particular sample being measured.

Samples were diluted to ~0.1 mg/ml, as measured by A280. 300 μl of sample was aliquoted into a quartz cuvette with a path-length of 0.2 mm [Hellma]. CD spectra were measured in a Jasco J-810 spectrophotometer with a Jasco PTC-423S temperature controller. Scans were measured at 20° C, 4° C and 95° C and the wavelength range of the scan was adjusted in accordance with detector high tension (HT). A HT value higher than 600 V is indicative of insufficient signal and no useful CD is measured. Scans were measured in accumulations of 4 and normalised by subtraction of a reference scan of the CD buffer being used. Thermostability experiments were conducted at 208 nm or 220 nm over a temperature range of 4° C to 95° C with a pitch of 1° C/min. Some of these experiments also included a CD scan from 260-~195 nm at 5° C temperature intervals.

CD scan data that were measured over a wavelength range of at least 260-190 nm were processed and interpreted using the DichroWeb CD data analysis server (Lobley *et al.,* 2002; Whitmore and Wallace, 2004). The data were input in the CD millidegrees (mdeg), CDSSTR was selected as the analysis programme and reference set number 4 used (260-190 nm). The protein concentration and the cuvette path-length values were also submitted to DichroWeb. The resulting output data included the conversion of the CD units to mean residue ellipticity ($[\theta]\,mrw,\lambda$) (degrees cm^2 $dmol^{-1}$ $residue^{-1}$) and estimates of the secondary structure content of the sample. For data that could not be processed using the Dichroweb server, CD millidegree values (machine units) were converted to mean residue ellipticity using the equation:

$$[\theta]mrw,\lambda = MRW \times \theta\lambda/10 \times c \times d$$

MRW = mean residue weight, $\theta\lambda$ = degrees at a given wavelength, c = concentration (g/ml) and d = path-length (cm). $MRW = M/(N-1)$ where M is the total molecular weight of the protein and N is the number of residues.

2.6.3 Differential scanning fluorimetry

Differential scanning fluorimetry (DSF) is a technique that measures the thermal stability of a protein using a fluorescent dye (Kohlstaedt *et al.,* 2015). Proteins were prepared in the same buffer as used in CD experiments to enable direct comparison of the denaturing temperatures of the proteins. Samples were diluted with buffer to 1 μM, 5 μM and 10 μM in a total of 300 μl and 0.5 μl of 5000X SYPRO orange fluorescent dye was added and mixed by pipetting. A baseline sample of buffer was also prepared with SYPRO orange to verify that none of components of the buffer interacted with the SYPRO orange to result in a fluorescent signal. Samples were aliquoted in 40 μl volumes into 3 PCR tubes as technical repeats and these were loaded into a rotary RT-PCR machine [Corbett]. Fluorescence/emission at 570 nm was measured during a temperature gradient from 25° C to 95° C (1° C min^{-1}). The fluorescence data were processed using the RT-PCR machine Rotor-Gene software [Corbett] to calculate the first derivative (dF/dT) and derive the T_m of the sample.

2.6.4 Inductively coupled plasma mass spectrometry

Inductively coupled plasma mass spectrometry (ICP-MS) is a technique for determining the metal content of a given sample at concentrations as low as parts per billion (ppb) (Ammann, 2007). Purified protein was incubated with a molar excess of $ZnCl_2$, 1 mM EDTA or without addition of either before running on an S200 10/300 GL Increase size-exclusion column [GE Healthcare] and 0.5 ml fractions were collected. ICP-MS samples of 3 ml were prepared by adding 300 μl of each fraction to 2.7 ml of 2.5% HNO_3 containing 20 ppb silver as the internal standard. Standards were also run of the metal ions magnesium, manganese, iron, copper, nickel, cobalt and zinc at concentrations of 0 ppb, 1 ppb, 5 ppb, 10 ppb, 25 ppb, 50 ppb, 75 ppb and 100 ppb, in solutions containing a fixed internal standard concentration of 20 ppb Ag. The ICP-MS [Thermo] collected 3 individual repeat measurements of each sample detecting for each of the metals included in the standards and the internal standard. The final molar concentration of the metal in a given sample was determined using the equation below:

$$[metal] = 10 \times (X/A)$$

Metal concentration $[metal]$ (μM), X is the value output from the ICP-MS in ppb and A is the atomic number of the metal being analysed. The molar concentration of protein in each fraction was also determined by measuring the absorbance at 280 nm using the NanoDrop and using the calculated extinction coefficient of the given protein (ProtParam) and the Beer-Lambert equation (Section 2.5.2). The ratio of metal bound to the protein in the fraction sample was determined by dividing the protein concentration by the concentration of metal present.

2.6.5 Size-exclusion chromatography multi-angle laser-light scattering

SEC-MALLS was used to determine the absolute mass of protein species in solution. After centrifugation at 17 000 *x g* for 5 minutes, 100 μl of purified proteins were injected onto a Wyatt SEC-050S5 column. These samples were either the proteins individually or proteins mixed with a potential binding partner at a total protein concentration of ~2.5 mg/ml. The column was connected to an Äkta Pure purification system [GE Healthcare] with DAWN HELEOS 8 light scattering and Optilab T-rEX refractive index detectors [Wyatt] connected downstream. Bovine serum albumin (BSA) [Sigma] was used as a control sample to calibrate and align the light scattering and refractive index detectors and setup the subsequent analysis methods. Fractions of 0.5 ml were collected for further analysis by SDS-PAGE (Section 2.5.3). Data processing was carried out using ASTRA 6 [Wyatt] SEC-MALLS software using a dn/dc value of 0.185 ml/g in the calculation of the absolute species mass.

2.6.6 Microscale thermophoresis

Microscale thermophoresis (MST) is a solution based method that can be used for measuring binding affinities between interacting molecules. MST utilises the principal of thermophoresis, where molecules move across a temperature gradient in equilibrial manner (Duhr and Braun, 2006b; Duhr and Braun, 2006a; Jerabek-Willemsen *et al.,* 2011; Wienken *et al.,* 2010). The MST experiments in this thesis were conducted using a Monolith NT.115 [NanoTemper Technologies] instrument with a Blue/Red fluorescence channel.

This instrument accepts 16 capillaries containing fluorescently labelled protein at a fixed concentration and unlabelled protein along a dilution series, the midpoint of which comprises the predicted K_d.

The protein was labelled with Cy5 fluorophore using the NHS-red labelling kit provided by NanoTemper Technologies, which involves an amine coupling reaction between the Cy5 fluorophore and the protein. The product of the labelling reaction is ~1 ml of labelled protein at a concentration of ~ 4 μM in the buffer that the protein was purified in. Optimisation of final concentration of labelled protein, the capillary type and whether additives are required in the buffer to prevent the protein sticking to the capillary were conducted using capillary scans (baseline fluorescence measurement). In all MST experiments in this thesis the premium capillary [NanoTemper Technologies] was used and Tween 20 was added to all buffers to a final concentration of 0.05 % v/v.

Unlabelled protein with 0.05 % Tween 20 was concentrated as previously described and 20 μl added to a 0.5 ml eppendorf tube. A further 15 tubes with 10 μl of buffer (0.05 % Tween 20) were prepared for a 1:1 serial dilution, or 31 tubes with 5 μl for a 2:1 dilution. The 1:1 serial dilution was performed by mixing 10 μl of the highest concentration with 10 μl of buffer in the subsequent 15 tubes. The excess 10 μl from the final dilution was discarded. In the 2:1 dilution the same procedure was followed. 10 μl of labelled protein, diluted so that the fluorescent signal was in the correct range of the detector using an appropriate LED power, was added to each of the dilutions. At each stage the samples were thoroughly mixed by pipetting. The capillary tubes were loaded by surface tension alone and placed into the Monolith.

The MST experiment parameters were adjusted, an LED power of 10 % or 15 % in experiments where proteins were well labelled and as a result, even very low protein concentrations (nanomolar) resulted in detector saturation. The MST power was set to 20% and 40%, this represents the temperature increase of the infrared beam and thus temperature gradient. The quality of the raw data was verified using MO.Affinity Analysis [NanoTemper Technologies], ensuring that fluorescence recorded in the capillary scans across the 16 capillaries were within ±10 %. The analysis software was used to fit the data using K_D(law of mass action described below) fit where the fluorophore concentration is fixed, c is the concentration of unlabelled protein, $Fluo$ is the concentration of labelled protein,

$K_d$is the dissociation constant, $f(c)$ represents the concentration of complexes. Unbound and bound are derived from the fluorescence during the thermophoresis experiment.

$$f(c) = unbound + (bound - unbound) \div 2([Fluo] + c + K_d - \sqrt{([Fluo] + c + K_d)^2 - 4 \times [Fluo] \times c})$$

Where 32 concentrations were used in the serial dilution, two separate data acquisitions were performed and the data were merged. Triplicate data were collected using new serial dilutions of unlabelled protein and the mean fluorescence values were used to fit a K_D.

2.6.7 Surface plasmon resonance

Surface plasmon resonance (SPR) is a technique for determining binding affinities and kinetics between a ligand (attached to a sensor surface) and analyte (molecule flowed across the ligand bound sensor surface). A CM5 sensor chip [GE Healthcare] was used in the SPR experiments in this thesis. This sensor chip is coated in carboxymethylated dextran and a protein ligand can be covalently attached to this surface via an amine coupling reaction. To determine the most efficient ligand surface pre-concentration and binding conditions, protein ligand was put into four immobilisation buffers of pH 4.0, 4.5, 5.0 and 5.5 and the binding to the sensor chip measured using a preprogrammed method. The 70 μl aliquots of protein ligand at a concentration of 1 mg/ml were buffer exchanged into buffers containing 10 mM Na acetate at pH 4.0, 4.5, 5.0 and 5.5 with an ion concentration of 10 mM NaCl. The CM5 sensor chip was docked into a Biacore X100 SPR machine [GE Healthcare] and the system was primed with 100 mM Na acetate pH 5.6, 100 mM NaCl running buffer that had been filtered with a 0.45 μm filter. The protein ligand aliquots in the Na acetate buffers were diluted to a final concentration of 10 μg/ml in a volume of 500 μl. The aliquots of protein ligand and a 500 μl aliquot of 100 mM NaOH were loaded into the sample holder of the X100 and the preprogrammed method run. At a flow rate of 10 μl/min, 2 minute injections of the ligand in the different pH buffers were performed. Between each ligand injection a 30 second regeneration injection of 100 mM NaOH was performed to remove any remaining protein. The sensograms of the injections of ligand

at the different pH values were compared and the pH value that resulted in the greatest response unit (RU) value was chosen for the ligand binding reaction.

The reagents for the amine coupling reactions were 0.4 M EDC 1-ethyl-3-(3-dimethylam inopropyl) carbodimide hydrochloride (EDC), 0.1 M N-hydroxysuccinimide (NHS) and 1 M ethanol

amine-HCl pH 8.5. EDC and NHS were mixed 1:1 immediately before the injection by the X100 device, these reagents activate the dextran surface. Ethanolamine is used to deactivate the surface after the ligand has been immobilised. Aliquots of 85 μl EDC, 85 μl NHS, 150 μl ethanolamine-HCl pH 8.5 and 100 μl of ligand at 10 μg/ml in the appropriate immobilisation buffer were placed into the sample rack. A preprogrammed manufacturer's method was used to perform the ligand immobilisation reaction, the method was set to achieve a ligand immobilisation of ~200 RU which would be most appropriate for the type of SPR experiment being conducted and the mass of the ligand and the analytes being probed. The equation to determine the appropriate level of ligand immobilisation is shown below:

$$analyte\,binding\,capacity\,(RU) = \frac{analyte\,MW}{ligand\,MW} \times immobilised\,ligand\,level\,(RU)$$

To determine the most appropriate injection parameters for the immobilisation of ligand to the sensor surface, the preprogrammed method conducted pre-concentration scouting injection which resulted in the desired RU values during the ligand pre-concentration steps. Ligand was immobilised to flow cell 2 (Fc2) and flow cell 1 (Fc1) was activated with EDC/NHS and deactivated with ethanolamine but no ligand was injected and immobilised to Fc1.

After ligand immobilisation the X100 system and sensor chip were primed with the buffer the proteins had been purified in, for example where PilA1 was purified in 20 mM Tris-HCl pH 8.0, 150 mM NaCl this was used as the running buffer. A series of analyte concentrations were prepared using a 1:1 dilution series in a final volume of 50 μl and a highest concentration of 500 μM. 8-10 analyte concentrations were initially probed. A preprogrammed method was used to perform the binding analysis at a flow rate of 10 μl/min and an inject time of 60 seconds. The ligand surface was regenerated by a 30

second injection of buffer with identical composition to running buffer but containing 1 M NaCl. Binding of analyte to the reference surface (Fc1) was monitored and were there was a RU value at Fc1 greater than 15 RU the data were discarded. Once an initial binding analysis run had been performed and an initial K_d value was calculated using the affinity fitting method in the Biacore evaluation software [GE Healthcare], the analyte concentrations were adjusted so that the K_d concentration was in the middle of the analyte concentration series and the binding analysis method performed again. Triplicate analyte concentration series were made and binding analysis methods performed to determine a reliable affinity between the ligand and analyte.

2.7 X-ray Crystallography

2.7.1 Protein crystallisation

For initial sparse matrix crystallisation screening purified protein was concentrated to 10-15 mg/ml, as measured by A_{280}, using a centrifugal concentrator with a molecular weight cut-off of at least half the mass of the target protein. Typically ~75 µl of each condition from sparse matrix screens were dispensed into a 96-well MRC crystallisation trays [Molecular Dimensions] with conical wells. A Mosquito dispensing robot [TTP Labtech] was used to set up sitting drops in 1:1 (100 nl: 100 nl) and 2:1 (200 nl: 100 nl) protein:reservoir volume ratios. Commercial screens used included Structure, Morpheus, JCSG+ [Molecular Dimensions], Index [Hampton Research], PACT [Qiagen] and Cryo [Jena Biosciences]. Crystallisation trays were stored in the crystallisation room at 20° C unless otherwise stated. Crystal trays were checked periodically for the presence of crystals using a 16X optical zoom light microscope [Leica], with which crystals were also photomicrographed.

2.7.1.1 Optimisation

Where an initial hit was identified using commercial sparse matrix crystallisation screens, an optimisation strategy was developed to improve the size, form or diffraction quality of the crystals. There are many variables that can be altered during crystal optimisation such as varying the concentrations of the salt or precipitant, the pH of the buffer, to changing

the vapour diffusion setup (hanging or sitting drop), drop volume or by varying the protein concentration. Initial optimisation screens searched the chemical space around the initial hit, precipitant, salt concentrations and the pH of the buffer. A 96 deep well block was divided into quarters (6 x 4), each quarter offering a different pH value and within each quarter salt concentration was varied on the Y-axis and precipitant on the X-axis (Figure 2.7.1 and Table 2.7.1). The 96 conditions were then transferred to an MRC crystallisation plate and mixed with protein using the Mosquito dispensing robot.

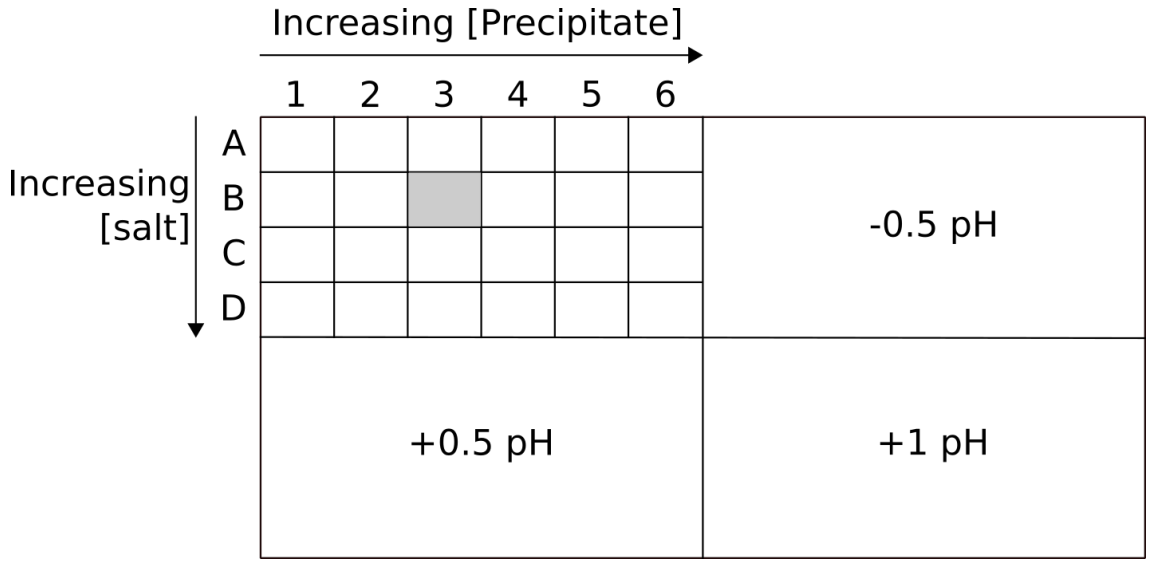


Figure 2.7.1 – **Crystallisation optimisation strategy**. Overview of typical crystallisation optimisation strategy employed in both 24-well (single buffer pH) and 96-well format (multiple buffer pH). The shaded square represents a condition in the optimisation tray that is the same as the initial hit.

Another optimisation strategy was the use of 24-well hanging drop trays. The same 24 well matrix as shown in Figure 2.7.1 and Table 2.7.1 using a single buffering pH was used. Drop ratios of 1:1 (1 μl :1 μl) and 2:1 (2 μl :1 μl) were set up on glass slides and sealed with vacuum grease.

	1	2	3	4	5	6
A	0.1 M HEPES 0.075 M $MgCl_2$ 24 % PEG 400	0.1 M HEPES 0.075 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES 0.075 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES 0.075 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES 0.075 M $MgCl_2$ 32 % PEG 400	0.1 M HEPES 0.075 M $MgCl_2$ 34 % PEG 400
B	0.1 M HEPES 0.125 M $MgCl_2$ 24 % PEG 400	0.1 M HEPES 0.125 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES 0.125 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES 0.125 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES 0.125 M $MgCl_2$ 32 % PEG 400	0.1 M HEPES 0.125 M $MgCl_2$ 34 % PEG 400
C	0.1 M HEPES 0.150 M $MgCl_2$ 24 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 32 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 34 % PEG 400
D	0.1 M HEPES 0.2 M $MgCl_2$ 24 % PEG 400	0.1 M HEPES 0.2 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES 0.2 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES 0.2 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES 0.2 M $MgCl_2$ 32 % PEG 400	0.1 M HEPES 0.2 M $MgCl_2$ 34 % PEG 400

Table 2.7.1 – PilK△1-32 crystal optimisation strategy. An example of a crystal hit optimisation strategy showing the variation of salt and precipitant concentrations. PilK Optimisation Screen #2 is based upon the Structure screen condition B12 (0.1 M HEPES pH 7.5 0.2 M $MgCl_2$ 30% PEG 400). This 24 well arrangement was repeated 4 times across a 96-well block with 0.1 M HEPES pH 7.0, 7.2, 7.5 and 8.0.

2.7.2 Crystallisation rescue strategies

2.7.2.1 Lysine methylation

Protein was diluted to a concentration of 1 mg/ml. Where the sample buffer contained Tris the protein sample was buffer-exchanged into 50 mM HEPES pH 7.5, 250 mM NaCl. 1 M dimethylamine-borane complex (ABC) [Sigma-Aldrich] was prepared in a volume of 1 ml. 20 µl 1M ABC and 40 µl 1 M formaldehyde [Sigma-Aldrich] were added to 1 ml of protein sample (1 mg/ml) in a 1.5 ml tube. The reaction was incubated at 4° C for 2 hours whilst mixing on a rotary mixer. A further 20 µl of 1M ABC and 40 µl 1 M formaldehyde were added to the reaction and incubated for a further 2 hours at 4° C whilst mixing. A final 10 µl of 1 M ABC was added to the reaction and incubated overnight at 4° C.

After overnight incubation 125 µl 1 M Tris-HCl pH 7.5 was added to the reaction tube to stop the reaction. The sample was spun at 16 000 *x g* for 5 minutes and loaded to an S200 16/600 equilibrated in the original sample buffer (Table 2.5.1), connected to an Äkta Pure [GE Healthcare]. Peak fractions of 1 ml were collected and 10 µl of each fraction analysed by SDS-PAGE as previously described. Sample containing fractions were pooled, concentrated to ~ 10-20 mg/ml and sparse matrix crystallisation trays dispensed as described (Section 2.7.1).

2.7.2.2 *In situ* proteolysis

A 1 ml aliquot of α-chymotrypsin [Sigma-Aldrich] at a concentration of 1 mg/ml was prepared by dissolving 1 mg of lyophilised protease in 1 ml of milliQ H_2O. A 1:10 dilution of the α-chymotrypsin was performed twice to a final concentration of 10 µg/ml. An aliquot of 100 µl of protein sample for crystallisation was placed into a tube and 1 µl of 10 µg/ml α-chymotrypsin was added to the sample immediately before dispensing on the dispensing robot. Crystallisation trays were dispensed as previously described (Section 2.7.1).

2.7.2.3 Additive screening

90 µl of the crystallisation solution that had produced crystals was dispensed into 96-well MRC crystallisation plate. 10 µl of additive screen conditions [Hampton Research] was then added to the reservoirs containing the crystallisation solution and mixed by pipet-

ting. The mosquito pipetting robot was then used to dispense protein and crystallisation crystallisation solution as described in section 2.7.1.

2.7.3 Crystal cryo-cooling

To reduce the affects of radiation damage in the crystal to enable the collection of the best quality diffraction data, protein crystals were cryo-cooled. Cryo-cooling is performed by harvesting crystals in an appropriate mount and cryoprotectant before flash cooling them in liquid nitrogen resulting in vitrification of the remaining liquid surrounding the crystal. Common cryoprotectant conditions include 25% PEG 400, 25% ethylene glycol and 3.5 M ammonium sulphate. The cryoprotectant chosen is dependent upon the components present in the crystallisation solution. After mounting a crystal, the loop and crystal were soaked in a drop of the crystallisation solution containing the cryoprotectant before cooling in LN_2. Some crystallisation conditions already include precipitants that provide cryo-protection (see Table 2.7.1), in these instances crystals were harvested from the drop where they formed and were cooled in LN_2. If the crystallisation condition is not immediately suitable, a cryo-protectant was mixed with an aliquot of reservoir solution prior to crystal harvesting.

2.7.4 Data collection

Test diffraction data and complete diffraction datasets used in this thesis were collected at the macromolecular crystallography (MX) beamlines at Diamond Light Source. Test images of the crystals were collected at 0° and 90° during an oscillation of 0.5° and 1s exposure, these data were input into Mosflm (Battye *et al.,* 2011). Mosflm was used to index the data to determine the spacegroup and unit cell parameters of the crystal and then to determine the most appropriate starting angle for the best diffraction data collection of the crystal.

Collection of complete datasets of native crystals were over 200° total oscillation with images recorded over 0.1° oscillations and exposure times of 0.01 seconds. For the collection of datasets during a single-wavelength anomalous dispersion (SAD) experiment, a total of 720° were collected to ensure that these data had high single anomalous com-

pleteness and redundancy.

2.7.5 Data processing

Figure 2.7.2 shows the data processing pathway that was used in this thesis to determine the PilA1 crystal structure (Chapter 4). The phase problem was overcome by using the single-wavelength anomalous dispersion method (Taylor, 2010).

Native and SAD datasets were indexed and integrated using 4 different programs or pipelines (Figure 2.7.2, steps in black). These included the stand alone programs, Mosflm (Battye *et al.,* 2011), XDS (Kabsch, 2010), DIALS (Gildea *et al.,* 2014) and the pipeline Xia2 (Winter *et al.,* 2013) which invokes the use of XDS. The resulting reflection files were input into Aimless for scaling, reduction and to determine the point group of the crystal's spacegroup (Evans and Murshudov, 2013). The indexing and integration strategies applied to each dataset were ranked based on the resolution of the dataset, a low R_{merge} value for the given resolution and the signal to noise ratio ($I/\sigma I$) of which a lower limit of 1.5 was imposed. The datasets and indexing strategies that satisfied these parameters the most were selected for further processing. Where derivative data was collected in SAD experiments, attention was also paid to the anomalous signal and the resolution at which this could be detected.

The SHELX software suite (Sheldrick, 2010) was used with the HKL2MAP GUI (Pape and Schneider, 2004) to calculate phases and to output an initial electron density map. This map and the protein sequence were input into the model building program Buccaneer (Cowtan, 2006). A 3D model of the protein is output by Buccaneer in PDB format, which was then refined against the native data using Refmac (Murshudov *et al.,* 2011). Several cycles of refinement using Refmac and manual model refinement in Coot (Emsley *et al.,* 2010) were performed, until R_{factor} and R_{free} values reached convergence. The method of B-factor refinement was chosen dependent upon resolution and the number of unique reflections (Merritt, 2012).

The model was validated using the MolProbity (Chen *et al.,* 2010) program and POLYGON module (Urzhumtseva *et al.,* 2009) of the Phenix package. Further rounds of refinement were performed as necessary based on the results of the model validation.

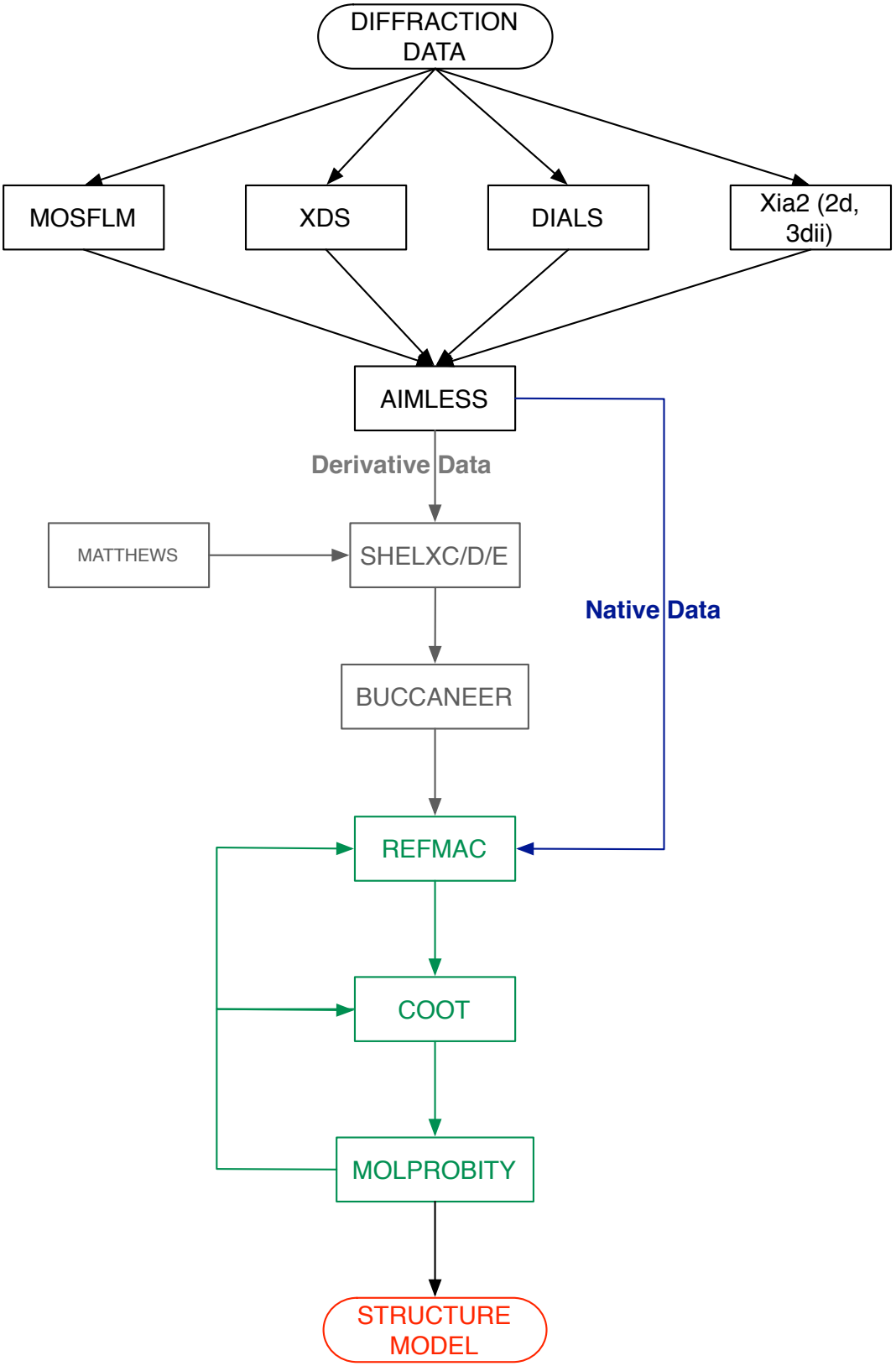


Figure 2.7.2 – **Single anomalous dispersion data processing workflow.** The data processing workflow that was followed for a SAD experiment. Indexing, integration and data reduction are outlined in grey. Steps involving only derivative data are outlined in grey and native data in blue. Refinement steps are outlined in green

2.8 Nuclear Magnetic Resonance

NMR experiments were conducted at the Astbury Centre for Structural Biology at the University of Leeds with support of Dr Arnout Kalverda and Dr Gary Thompson. 1D-NOESY spectra were collected on a DD2 600 MHz NMR spectrometer and 2D-HSQC spectra were collected on an Inova 750 MHz NMR spectrometer. To collect 1D-NOESY resonances, samples were prepared as previously described with the exception that the final sample buffer was 50 mM Na_2HPO_4 pH 6.0, 150 mm NaCl. The samples were concentrated to 300 μM in a volume of at least 300 μl. D_2O was added to the 300 μl sample to a final concentration of 10% v/v before aliquoting into an NMR sample tube. 1D nuclear overhauser spectroscopy (NOESY) experiments were conducted using a double pulsed field gradient spin echo (DPFGSE) pulse sequence for suppression of the water signal at 4.7 ^{1}H/ppm. The 1D-NOESY spectra were processed using NMRPipe and nmrDraw (Delaglio *et al.,* 1995) and visualised in the CCPN Analysis (Vranken *et al.,* 2005). All NMR spectra were recorded at 25° C.

Heteronuclear single quantum coherence (HSQC) experiments require 2 isotopes, ^{15}N labelled protein and the natural protons ^{1}H. ^{15}N incorporation is described in section 2.4.2 and protein was then purified and prepared as described for 1D experiments. Titrations involved incremental addition of the molar equivalent of unlabelled protein to the NMR tube which already contain ^{15}N labelled protein. After each addition of unlabelled protein spectra were recorded in the following labelled to unlabelled ratios: 1:0.25; 1:0.5; 1:0.75, 1:1; 1:2; 1:5. The spectra were processed using NMRPipe, Varian converter and nmrDraw (Delaglio *et al.,* 1995). 2D spectra were processed and visualised in CCPN Analysis (Vranken *et al.,* 2005).

Chapter 3

Biophysical characterisation of the SpoIIQ:SpoIIIAH complex in *Clostridium difficile*

3.1 Introduction

SpoIIQ and SpoIIIAH are essential sporulation gene products in *C. difficile* (Serrano *et al.,* 2015; Fimlaid *et al.,* 2015). Membrane proteins that are bound via N-terminal transmembrane domains, SpoIIQ is localised to the forespore membrane and SpoIIIAH to the mother cell membrane during sporulation, and they are concentrated to this interface during engulfment of the daughter cell by the mother cell (Figure 3.1.1) (Serrano *et al.,* 2015; Fimlaid *et al.,* 2015). Using fluorescently labelled fusion proteins, it has been shown that globular C-terminal domains of both SpoIIQ and SpoIIIAH are located on the outside of their respective membranes, in a region known as the intersporangial space (Figure 3.1.1) and associate across in a ratchet-like manner as the forespore is engulfed by the mother cell (Figure 3.1.1) (Serrano *et al.,* 2015).

The intersporangial domain of SpoIIQ is formed of an endopeptidase domain with homology to the LytM family of endopeptidases that belong to the M23 family of metalloproteases (Crawshaw *et al.,* 2014). These endopeptidases cleave peptide cross-bridges between peptidoglycan strands and have a pair of conserved motifs (HxxxD; HxH) that are required for the co-ordination of a Zn^{2+} metal ion essential for activity (Firczuk and

Bochtler, 2007). SpoIIIAH shares sequence homology with the Type III secretion proteins YscJ/FliF, which are pore forming proteins (Camp and Losick, 2008).

In *C. difficile,* SpoIIQ and SpoIIIAH are required for progression beyond forespore engulfment during sporulation (Fimlaid *et al.,* 2015; Serrano *et al.,* 2015). Mutants of either *spoIIQ* or *spoIIIAH* become stalled early during engulfment and do not proceed to mature spores. The *spoIIQ* and *spoIIIAH* mutants develop a partially engulfed forespore and bulged or collapsed membranes have been observed, indicating improper peptidoglycan processing during engulfment (Serrano *et al.,* 2015).

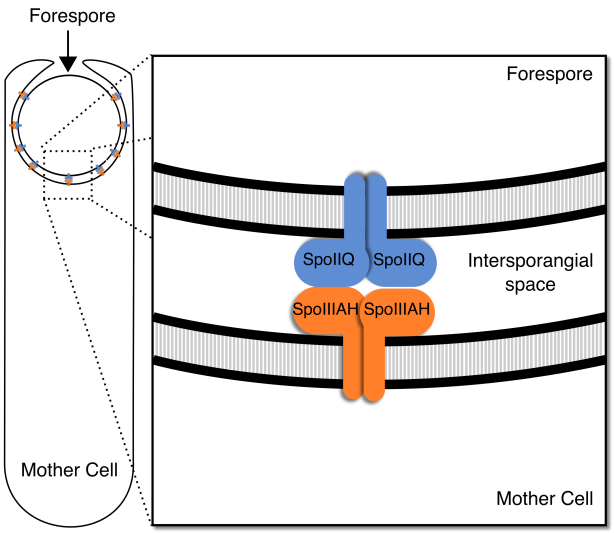


Figure 3.1.1 – Organisation of SpoIIQ:SpoIIIAH complex during engulfment. SpoIIQ and SpoIIIAH are expressed during engulfment of the forespore by the mother cell at early stages of sporulation. SpoIIQ (blue) is localised in the forespore membrane and SpoIIIAH (orange) is localised in the mother cell membrane. The C-terminal domains of both proteins are localised within the intersporangial space between the membranes. SpoIIQ and SpoIIIAH form an interaction via their intersporangial domains resulting in a ratchet-like association between the forespore and mother cell membranes.

The SpoIIQ:SpoIIIAH complex has been studied extensively in *Bacillus subtilis*. In *Bacilli*, only SpoIIQ is essential for the complete engulfment of the forespore with *spoIIQ* mutants stalled at engulfment (Londoño-Vallejo *et al.,* 1997). *B. subtilis* SpoIIQ also contains a degenerate LytM domain - the key metal co-ordinating residues that are required for peptidoglycan activity are not conserved (Crawshaw *et al.,* 2014). The intersporangial domains of *B. subtilis* SpoIIQ and SpoIIIAH form a complex, the crystal structure of which has been determined (Levdikov *et al.,* 2012; Meisner *et al.,* 2012). The SpoIIQ:SpoIIIAH

complex provides access between the cytosols of the *B. subtilis* forespore and mother cell suggesting that the complex forms a channel (Meisner *et al.,* 2008). The exact substrate of the channel is currently unknown but there are hypotheses that it could be a small signalling molecule or protein (Camp and Losick, 2009; Fimlaid *et al.,* 2015). The sequence identities between the *B. subtilis* and *C. difficile* proteins is 25% for SpoIIQ and 30% for SpoIIIAH, but the precise function of the SpoIIQ:SpoIIIAH complex in either *C. difficile* or *B. subtilis* is currently unknown.

The aim of the work presented in this chapter was to characterise the inter-sporangial domains of SpoIIQ and SpoIIIAH in *C. difficile* by answering a number of questions. Firstly, is the conserved LytM domain of *C. difficile* SpoIIQ able to co-ordinate Zn^{2+}? Do the inter-sporangial domains of *C. difficile* SpoIIQ and SpoIIIAH form a stable complex as observed in *B. subtilis*? Finally, what is the overall structure of the *C. difficile* SpoIIQ:SpoIIIAH complex and does it differ from that of *B. subtilis*?

3.2 Expression and purification of SpoIIQ and SpoIIIAH proteins

The open reading frames encoding the C-terminal domains of SpoIIQ and SpoIIIAH were amplified from coding sequences CD630_0125 (*spoIIQ*) and CD630_1199 (*spoIIIAH*) from genomic DNA of *C. difficile* strain 630. The constructs were designed to omit the first 30 residues of SpoIIQ and 28 residues of SpoIIIAH that encode the transmembrane spanning regions of each protein (Sonnhammer *et al.,* 1998; Möller *et al.,* 2001). The amplified gene fragments were ligated into the expression vector pET-M11 (Section 2.3) such that upstream of each insert there was an N-terminal His-6-tag that could be removed with tobacco etch virus (TEV) protease (Figure 3.2.1 and 3.2.2). These constructs of SpoIIQ and SpoIIIAH, referred to as sQ and sAH herein, were used to transform Rosetta (DE3) *E. coli* expression cells. Expression tests were conducted as outlined in Section 2.4 to determine the optimal expression conditions (20° C, 16 hrs), and these conditions were used throughout.

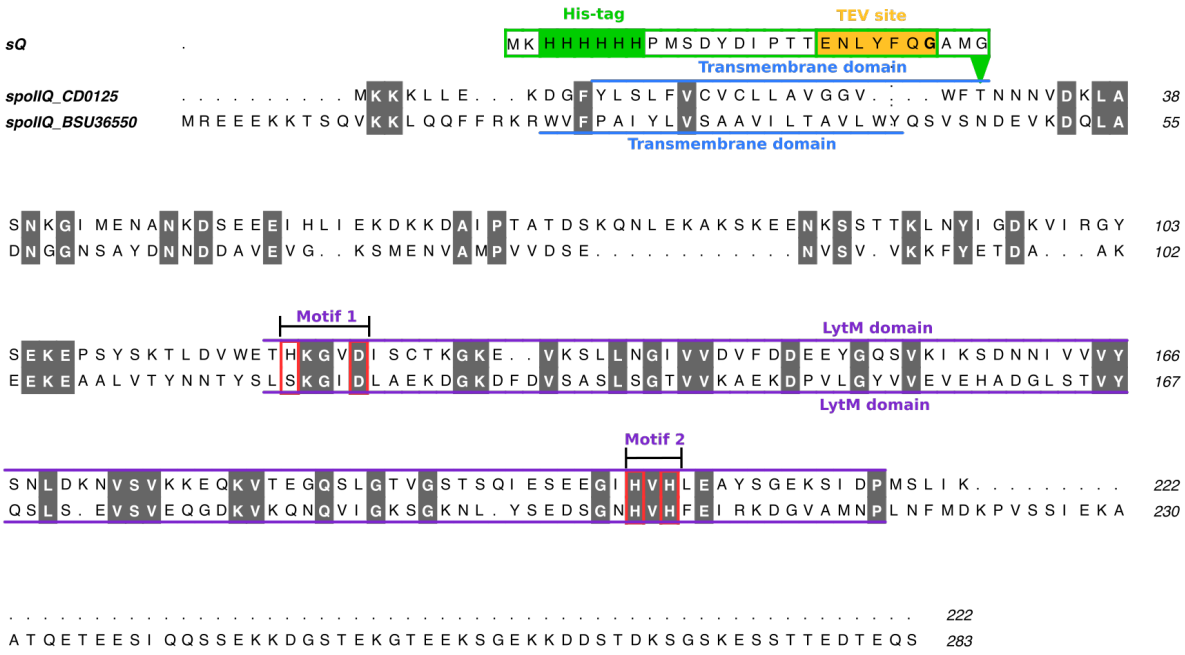


Figure 3.2.1 – sQ construct and alignment with full-length *C. difficile* and *B. subtilis* SpoIIQ. Alignment of *C. difficile* (CD0125) and *B. subtilis* (BSU36550) SpoIIQ. The boundaries and tags of the sQ construct are also shown. The N-terminal His-tag (green) and TEV protease recognition site are positioned above the *C. difficile* sequence, the green arrow indicates where the tags are appended to the protein coding sequences. Cleavage using TEV protease removed the His-tag and TEV protease site upstream of the glycine (bold). The sQ construct omits the first 30 residues of the open reading frame, which are predicted to form the transmembrane domain (blue lines). The C-terminal region of SpoIIQ contains a LytM domain (purple lines), a signature of which is two Zn^{2+} co-ordinating motifs, HXXXD (motif 1, red box) and HXH (motif 2, red box). Motif 1 in *B. subtilis* is described as degenerate due to a His to Ser mutation (SXXXD).

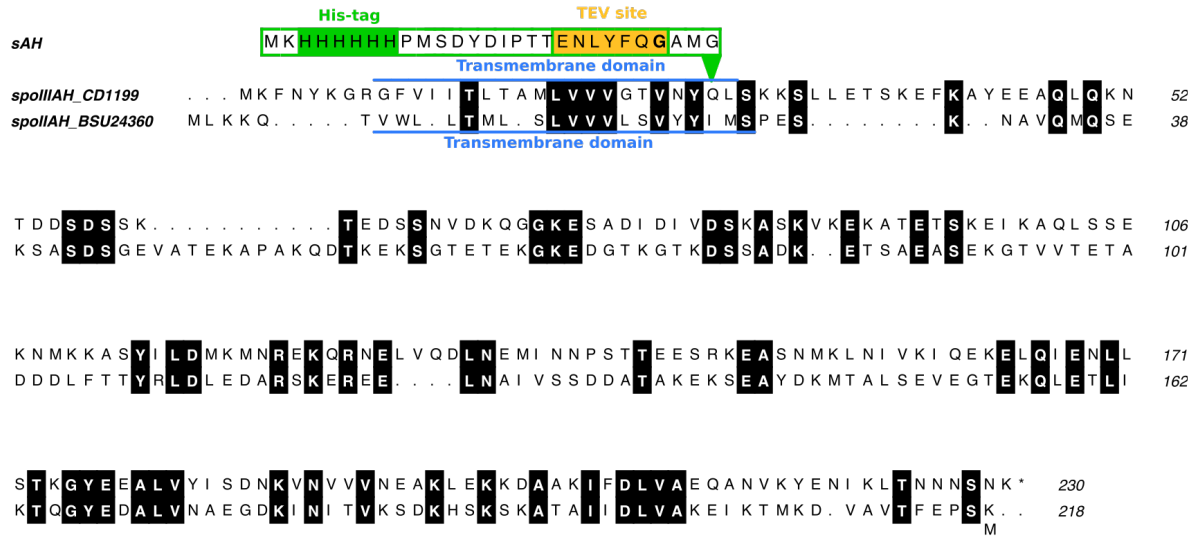


Figure 3.2.2 – sAH construct and alignment with full-length *C. difficile* and *B. subtilis* SpoIIIAH. *C. difficile* (CD1199) and *B. subtilis* (BSU24360) SpoIIIAH share 30% sequence identity. The boundaries and tags of the sAH construct are positioned above the *C. difficile* sequence. The N-terminal His-tag (green) and TEV protease recognition site (orange), the green arrow indicates where the tags are appended to the protein coding sequences. Cleavage using TEV protease removed the His-tag and TEV protease site upstream of the glycine (bold). The first 28 residues of the SpoIIIAH open reading frame are predicted to contain a hydrophobic transmembrane domain which has been excluded from the sAH construct.

3.2.1 Purification of sQ and sAH constructs

The soluble protein constructs sQ and sAH were purified as outlined in Section 2.5. Optimal step purification and buffer conditions were determined during the course of biochemical and biophysical characterisation of these proteins (Section 3.3), to ensure that sQ and sAH were folded and that such buffering conditions were appropriate for subsequent biochemical and biophysical analysis.

The yield of sAH after nickel affinity purification (Figure 3.2.3), His-tag removal and size exclusion chromatography was ~ 20 mg/L cell culture and ~ 50 mg/L for sQ. The SDS-PAGE in Figure 3.2.4B/D showed it was possible to purify stable sAH and sQ with few contaminants. The theoretical molecular masses of sQ and sAH were calculated using the ProtParam server (web.expasy.org/protparam) after the removal of the His-tag: sQ is 21.3 kDa and sAH is 22.8 kDa. Both proteins migrated the same distance as the 25 kDa molecular mass marker (Figure 3.2.4B/D), which is slightly greater than the theoretical molecular masses. Comparison of the elution volumes of these constructs from a Superdex 200 26/600 size exclusion column (Figure 3.2.4A/C) with the elution profile of calibration proteins suggested that both of these proteins eluted from the column in a dimeric state.

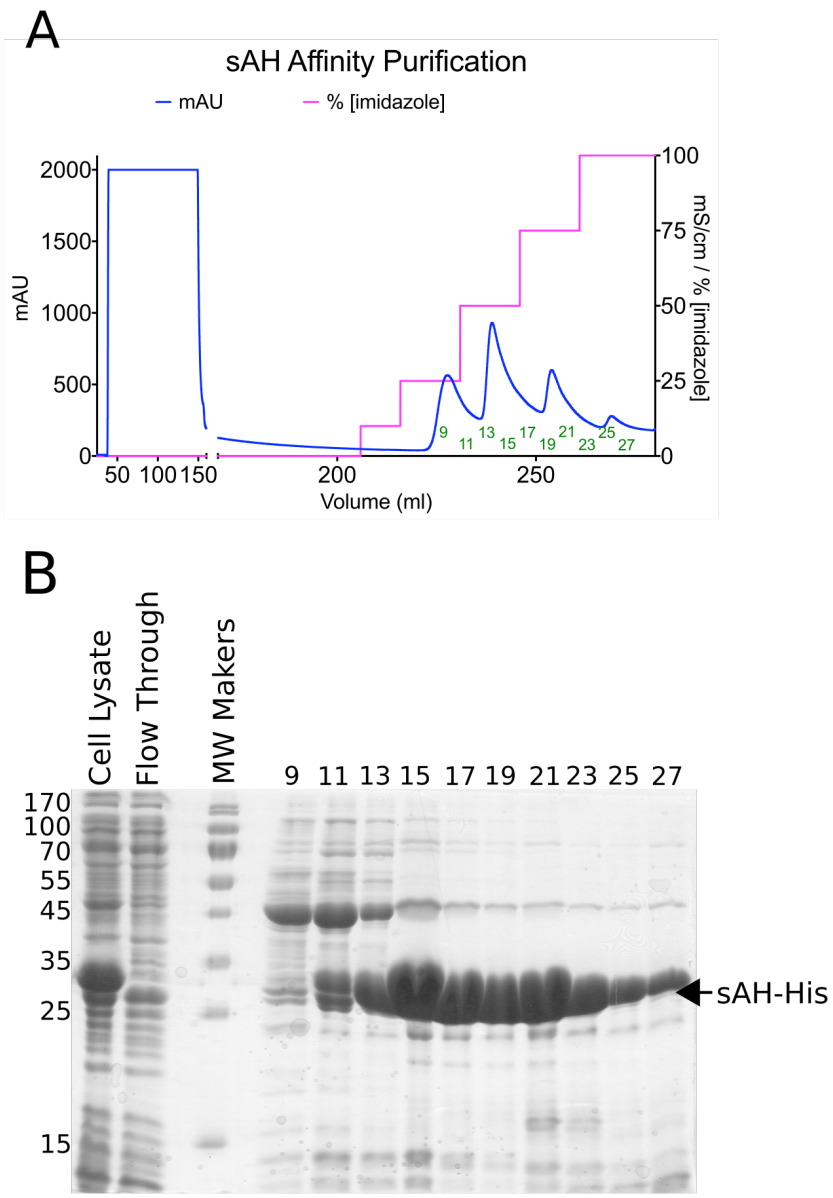


Figure 3.2.3 – **Nickel affinity purification of sAH. A:** Chromatogram of a nickel affinity purification of sAH showing absorbance at 280 nm (blue), concentration (% v/v) of imidazole in the elution buffer (pink) and fractions that were analysed by SDS-PAGE (green). Three elution peaks from the Hi-Trap NTA column are observed at imidazole concentrations of 125 mM, 250 mM and 375 mM. **B:** SDS-PAGE of the cell lysate, the flow through and peak fractions during elution. Of the three peaks observed in the chromatogram (A), fractions 9-12 contained a protein of ~ 45 kDa in the SDS-PAGE, whilst fractions 13-26 that contained eluant from the peaks at 240 and ~255 ml contained a protein that migrated to a mass of ~30 kDa. The theoretical mass of sAH with the His-tag and TEV linker was 26 kDa. Fractions 9 and 11 showed a doublet band at ~25 kDa, this was also observed in the cell lysate. The flow through contained a protein which migrated through the gel an equivalent distance to one of these doublets, indicating this is an impurity. DNase I (mass: 30 kDa) is added to the cell suspension before lysis and was identified as the contaminant.

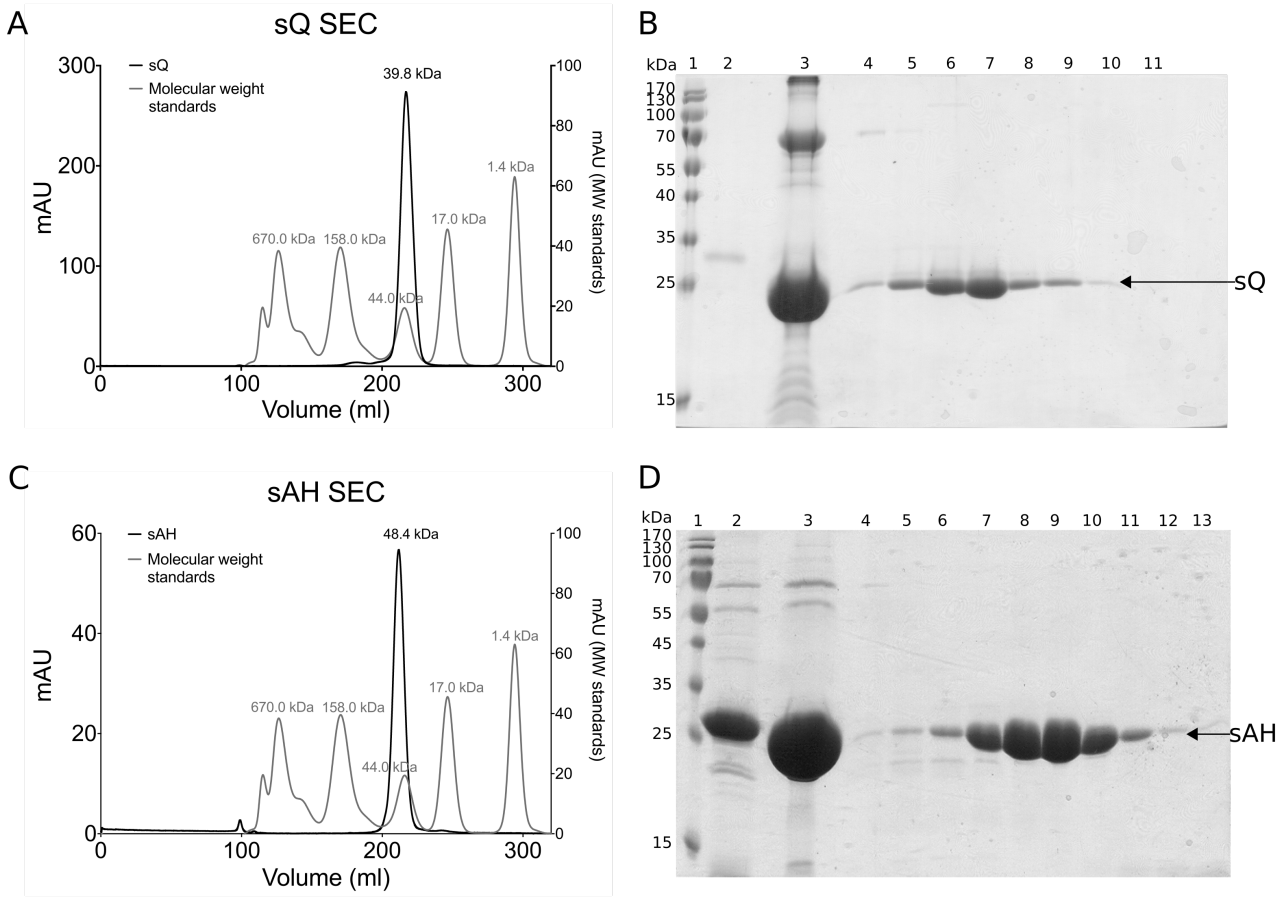


Figure 3.2.4 – **SEC purification of sQ and sAH.** Size exclusion chromatograms (SEC) of sQ (A) and sAH (C) purified using an S200 26/600 size-exclusion column. The UV elution profile of molecular mass standard proteins is shown in grey and each peak is labelled with the mass of these standards. The molecular mass standards were thyroglobulin (670.0 kDa), γ-globulin (158.0 kDa), ovalbumin (44.0 kDa), myoglobin (17.0 kDa) and Vitamin B_{12} (1.4 kDa). The elution volume of sQ was equivalent to a protein of ~40.0 kDa and for sAH was equivalent to a protein of 48.8 kDa. The loaded sample (lane 3), un-cleaved protein from His-trap purification (lane 2 B/D, lane 1 F) and the peak fractions (lanes 4-13) were analysed by SDS-PAGE for sQ (B) and sAH (D).

3.3 Characterisation of sQ and sAH

To determine if the over-expressed and purified sQ and sAH were folded, stable and suitable for further characterisation and crystallisation, circular dichroism (CD) and nuclear magnetic resonance (NMR) experiments were performed.

3.3.1 CD spectroscopy

The CD spectra of sQ and sAH were measured (Figure 3.3.1) and the secondary structure composition was calculated by CD deconvolution using the CDSSTR algorithm (Compton and Johnson, 1986) (Section 2.6.2). The secondary structure composition outputs by CDSSTR (www.dichroweb.org) for sQ and sAH are summarised in Table 3.3.1. The CD spectrum of sQ (Figure 3.3.1A) is representative of a protein that is either not well ordered, has a high proportion of β-strand structure or a combination of both. The CD spectrum of sAH presented in Figure 3.3.1B is typical of a folded protein with predominantly α-helical secondary structure.

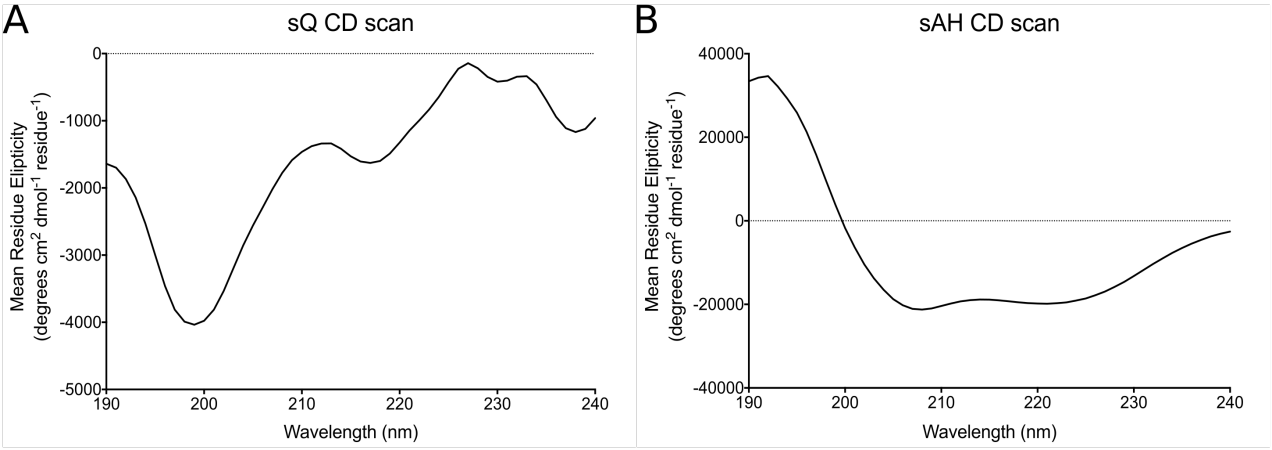


Figure 3.3.1 – **Circular dichroism spectra of sQ and sAH.** Secondary structure of sQ (A) and sAH (B) were analysed using CD and the data deconvoluted using the CDSSTR algorithm at the DICHROWEB server and reference set 4. The deconvolution results are presented in Table 3.3.1. These data are plotted above in units of mean residue ellipticity ([ϑ]MRE).

The peptide sequences of sQ and sAH were submitted to the PSIPRED secondary structure prediction server (bioinfo.cs.ucl.ac.uk/psipred) (Jones, 1999; Buchan *et al.,* 2013) and the results are summarised in Table 3.3.1. The secondary structure composition of SpoIIQ and SpoIIIAH observed in the crystal structures of the *B. subtilis* proteins

Protein	PDB code	Method	Helix (%)	Coil (%)	Strand (%)	Un-ordered (%)
sQ	-	PSIPRED	5	–	32	31
		CDSSTR	8	23	36	33
sAH	-	PSIPRED	50	–	9	63
		CDSSTR	59	7	12	22
B. subtilis SpoIIQ	3UZ0	-	0	4	43	53
	3TUF	-	4	8	39	49
	-	PSIPRED	17	–	25	45
B. subtilis SpoIIIAH	3UZ0	-	59	0	14	27
	3TUF	-	58	0	16	26
	-	PSIPRED	44	–	11	56

Table 3.3.1 – Secondary structure composition of sQ and sAH. Summary of secondary structure composition of sQ and sAH derived from CD spectra using the CDSSTR algorithm at the Dichroweb server and secondary structure predictions from the PSIPRED server. The secondary structure composition of *B. subtilis* SpoIIQ and SpoIIAH in the PDB models 3UZ0 (Meisner *et al.,* 2012) and 3TUF (Levdikov *et al.,* 2012) which share 30% and 25% sequence identity with the *C. difficile* proteins and the PSIPRED secondary structure prediction for the *B. subtilis* proteins is shown for reference.

are also presented. Deconvolution of the CD spectrum of sAH using CDSSTR determined that the secondary structure includes 59% helix, 7% coil, 12% strand and 22% unordered. The proportions of helix and strand were comparable with PSIPRED secondary structure prediction and with the secondary structure observed in the crystal structures of *B. subtilis* SpoIIIAH, that shares 25% sequence identity with the *C. difficile* protein. Comparison with these experimental measurements, predictions and the *B. subtilis* homologue indicate that sAH is folded, and likely observes a similar fold to that of the *B. subtilis* orthologue.

Processing of the sQ spectrum using CDSSTR returned a secondary structure composition that included 8% helix, 23% coil, 36% strand and 33% unordered residues. These proportions of secondary structure were similar to the PSIPRED predictions. Whilst the spectrum shown in Figure 3.3.1B of sQ is not typical of a folded protein, the deconvolution data, PSIPRED prediction and comparison with *B. subtilis* SpoIIQ indicate that sQ is folded. The β-strand content of sQ is high and the proportion of α-helix is low in both predicted and observed data. As α-helices result in stronger spectral features than β-strand structure in a CD scan measurement, especially at 208 nm and 220 nm, the low proportion of α-helical structure can result in the spectrum shown in Figure 3.3.1B (Kelly *et al.,* 2005). Interestingly, the PSIPRED secondary structure content prediction for the *B. subtilis* proteins were considerably different from those observed in the crystal structures.

To further characterise the two proteins, the thermal stability of sQ and sAH was probed using CD thermal melts at fixed wavelengths (Figure 3.3.2). The strongest spectral feature in the CD scan of sQ was measured at 200 nm (Figure 3.3.1A) and the low proportion of α-helical structure present in this sample indicated that neither 208 nm nor 220 nm were suitable for this experiment. Therefore the variable temperature scan of sQ was measured at 200 nm over a temperature range of 20-80° C. The melting temperature (T_m) of sQ was calculated to be 42° C. The sAH sample was measured at 220 nm due to the significant proportion of α-helical structure measured in the CD scan of sAH (Figure 3.3.1B). The T_m of sAH was calculated to be 46° C.

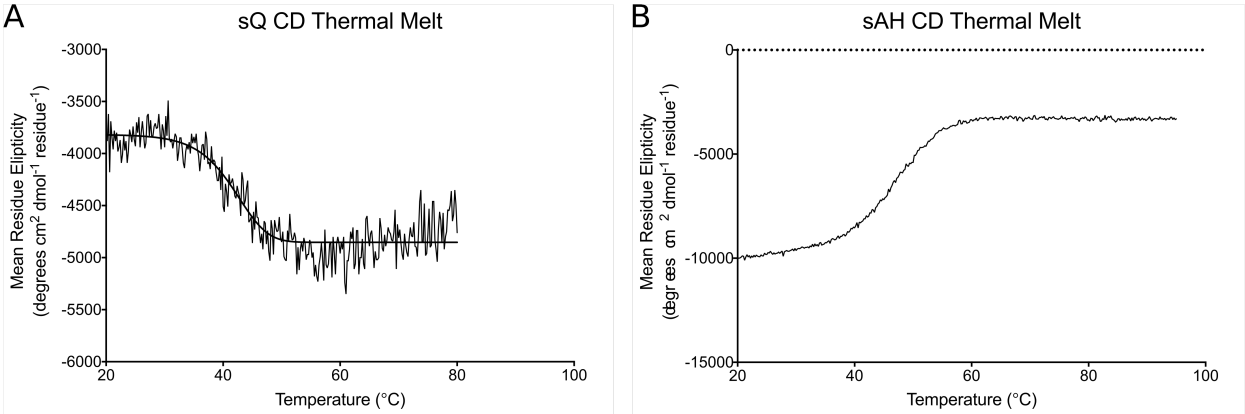


Figure 3.3.2 – **Thermal stability of sAH and sQ. A:** A CD spectrum of sQ was measured at 200 nm over a temperature gradient from 20° C to 80° C (+1° C min^{-1}) from which a T_m of 42° C was calculated by a non-linear regression fit. **B:** Circular dichroism spectra of sAH measured at 220 nm over a temperature range from 20° C to 90° C (+1° C min^{-1}) (A) and a T_m of 46° C was calculated.

3.3.2 Nuclear magnetic resonance

To assess further the stability and globularity of sQ and sAH in solution, NMR experiments were performed, first using 1D NMR and subsequently by 2D NMR. 1D ^{1}H NMR nuclear Overhauser spectra (NOESY) of sQ and sAH were collected using a 500 MHz INOVA NMR [Agilent] magnet at the Astbury Centre, Leeds University with support from Dr Arnout Kalverda and Dr Gary Thompson. The spectra showed (Appendix A, Figure A.0.1) that sQ and sAH contained ordered structure in the 6-9 ^{1}H/ppm region that correspond to the H-N bonds in both the peptide backbone and side chains. Downfield of (value above) 8

^{1}H/ppm, the peaks became broader and there were no peaks downfield of 8.5 ^{1}H/ppm. Folded, globular proteins result in peaks downfield of 8.5 ^{1}H/ppm, so the absence of such peaks suggested that sQ and sAH were not fully globular but did contain a proportion of ordered structure. Both sQ and sAH showed broad peaks in the α-H region of the 1D spectra (3.2-6 ^{1}H/ppm) indicating overlapping signals that could be better resolved in 2-dimension experiments.

The presence of some well resolved peaks in the 1D-NOESY spectra of both sQ and sAH showed that these proteins were appropriate for 2 dimension heteronuclear single quantum coherence (HSQC) NMR experiments. HSQC NMR required the presence of two magnetic isotopes ^{1}H and ^{15}N, the latter of which was incorporated into sQ and sAH during protein expression (Section 2.4.2).

The ^{15}N-HSQC spectrum of sQ (Figure 3.3.3A) contained 201 discrete N-H peaks out of a total 262 N-H bonds for 192 residues, 70 of which were part of amino acid side chains. The absence of peaks for all N-H bonds suggested that not all residues were observable, or that the peaks of these residues were broad or merged with those of other residues/chemical environments in the 7.6-8.5 ^{1}H/ppm region. Such instances are typical of unstructured regions of proteins (Felli and Pierattelli, 2012). There were 33 peaks downfield of 8.5 ^{1}H/ppm that could be described as N-H backbone bonds in ordered residues, indicating that sQ contained structured regions. The deconvoluted CD data (Table 3.3.1) indicated that a third of the residues were un-ordered and these may represent the broader peaks in the 7.6-8.5 ^{1}H/ppm region. Together, the CD and NMR data indicate that sQ was folded but contained regions that were disordered or flexible in solution.

The HSQC spectrum of sAH (Figure 3.3.3B) contained 260 peaks for a total of 275 N-H bonds in 201 residues, 74 of these bonds were located in side chains. The sAH HSQC spectra represented the majority of residues and missing peaks may be attributed to the peak broadness resulting in overlapping regions. A significant proportion of the backbone N-H peaks were upfield of 8.5 ^{1}H/ppm and only 15 peaks were observed downfield of 8.5 ^{1}H/ppm in the ordered backbone region. As observed in the sQ HSQC spectrum, the peaks observed in the spectrum of sAH appeared to represent an intrinsically disordered protein (Felli and Pierattelli, 2012). The CD spectrum deconvolution indicated that 22% of the residues were unstructured (Table 3.3.1). The CD and HSQC spectra were recorded in

similar buffers and pH, however, these data do not indicate the same level of order within the structure of sAH. CD is a measure of the secondary structure of a given protein, whilst NMR does not provide information on the secondary structure, only whether the protein is folded or not via the chemical environment of specific atoms. These observations seemed to indicate that sAH had significant elements of secondary structure but did not have the hallmarks of a globular folded protein in the HSQC-NMR spectra.

To understand whether the presence of the sQ and sAH influenced the structure of either protein, ^{1}H:^{15}N titration experiments were performed. Unlabelled sQ was added to ^{15}N labelled sAH (308 μM) in molar ratios of 1:0 (Figure 3.3.4A), 1:0.3, 1:0.6, 1:0.9 (Figure 3.3.4B), 1:2 and 1:5. After the addition of each unlabelled titre, ^{1}H:^{15}N HSQC spectra were recorded. A clear change in the ^{1}H:^{15}N spectra during the titration was a decrease in the number of observable peaks between ratios 1:0 and 1:0.9 (Figure 3.3.8B). Better resolved peaks, particularly in the H-N backbone region (7.6-8.5 ^{1}H/ppm), indicatied that there was more order in the residues that these peaks represented when in the presence of sQ. A small number of downfield (high ppm value) peak shifts were also observed in these spectra (Figure 3.3.5). As sQ was titrated into the experiment, these peaks may represent H-N bonds in the backbone of residues that become stabilised in sAH by an interaction with sQ. Without further NMR experiments, using protein labelled with both ^{13}C and ^{15}N to allow backbone assignment, it was not possible to attribute the changes in peak intensity and chemical shifts to specific residues. Further NMR experiments using triple labelled protein should be performed to assign the residues to be able to understand the influence of complex formation on the structure of sQ and sAH. Such experiments were considered to be beyond the scope of this project.

^{15}N-labelled sQ (110 μM) was titrated with unlabelled sAH in molar equivalents of 1:0, 1:0.2, 1:0.6, 1:1.2 and 1:2. The 1:0 and 1:2 spectra are shown in Figure 3.3.6. In contrast to the ^{15}N-sAH:sQ titration, the number of peaks observed remained relatively constant until a ratio of 1:2 was added where only 80% of the peaks were observed (Figure 3.3.8A). No peak shifts were observed in these titration spectra (Figure 3.3.7).

A significant number of resonance peaks disappeared in the ^{15}N-sAH titration as the concentration of sQ increased and to a lesser extent in the ^{15}N-sQ titration (Figure 3.3.8). There are three factors that could account for the change in peak numbers.

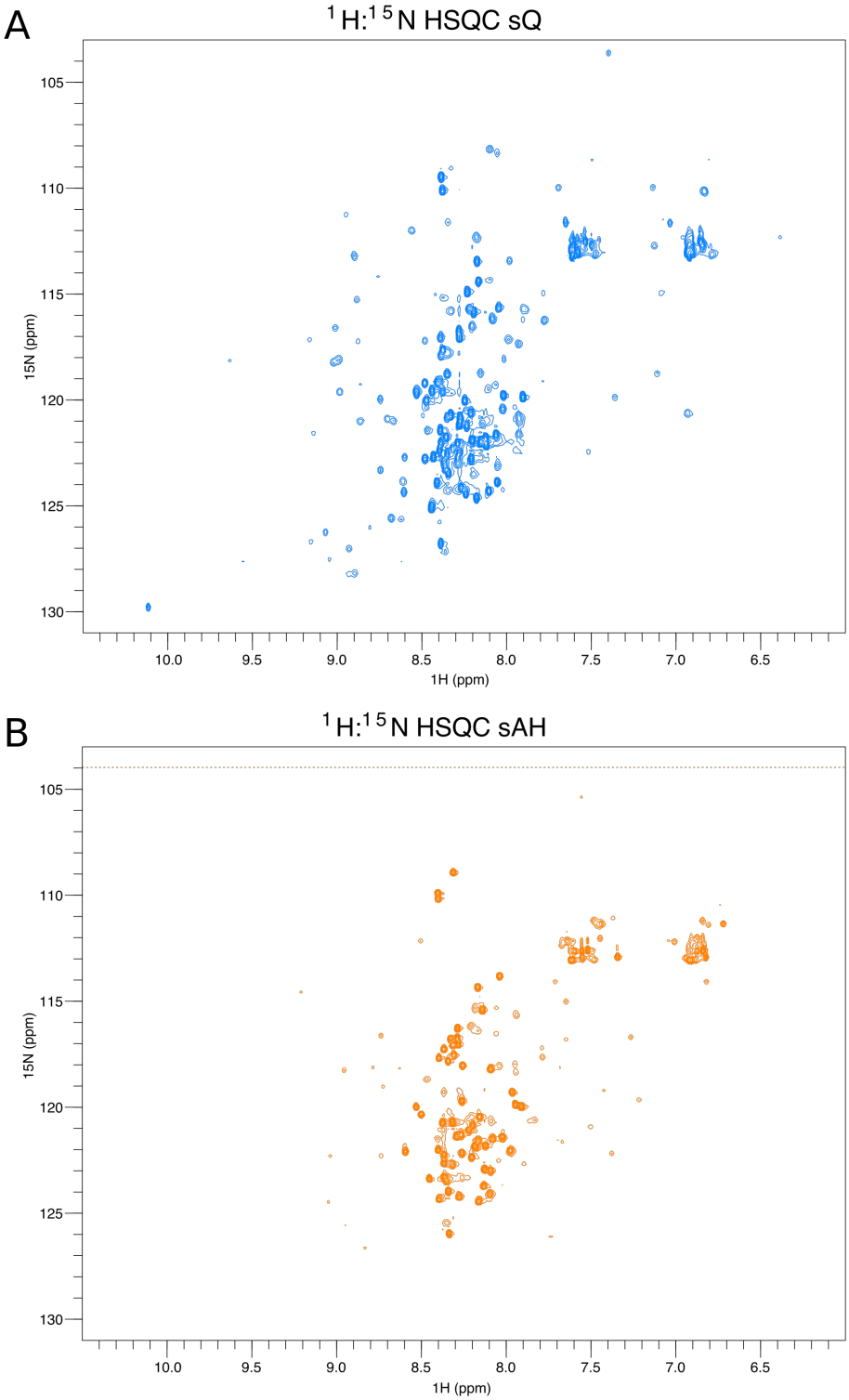


Figure 3.3.3 – ^{1}H:^{15}N-HSQC spectra of sQ and sAH. 2-D ^{1}H:^{15}NHSQC spectra of sQ (A) and sAH (B) were measured in a 750 MHz INOVA NMR magnet [Agilent] at the Astbury Centre, University of Leeds with the support of Dr Arnout Kalverda and Dr Gary Thompson. The data were processed using nmrPipe and presented using the CCPN suite at a baseline contour level of 7.5 sigma. The spectra of sQ (A) contained 201 discrete amide peaks and sAH (B) contained 260 peaks. Peaks downfield of 8.5 ^{1}H/ppm represent N-H bonds in the backbone of ordered residues. Peaks in the region 8.5-7.6 ^{1}H/ppm represent N-H bonds in the backbone of disordered residues. Peaks upfield of 7.6 ^{1}H/ppm represent side chain N-H bonds. The peak observed downfield of 10 ^{1}H/ppm specifically represents the single tryptophan residue (Trp117) of sQ (A).

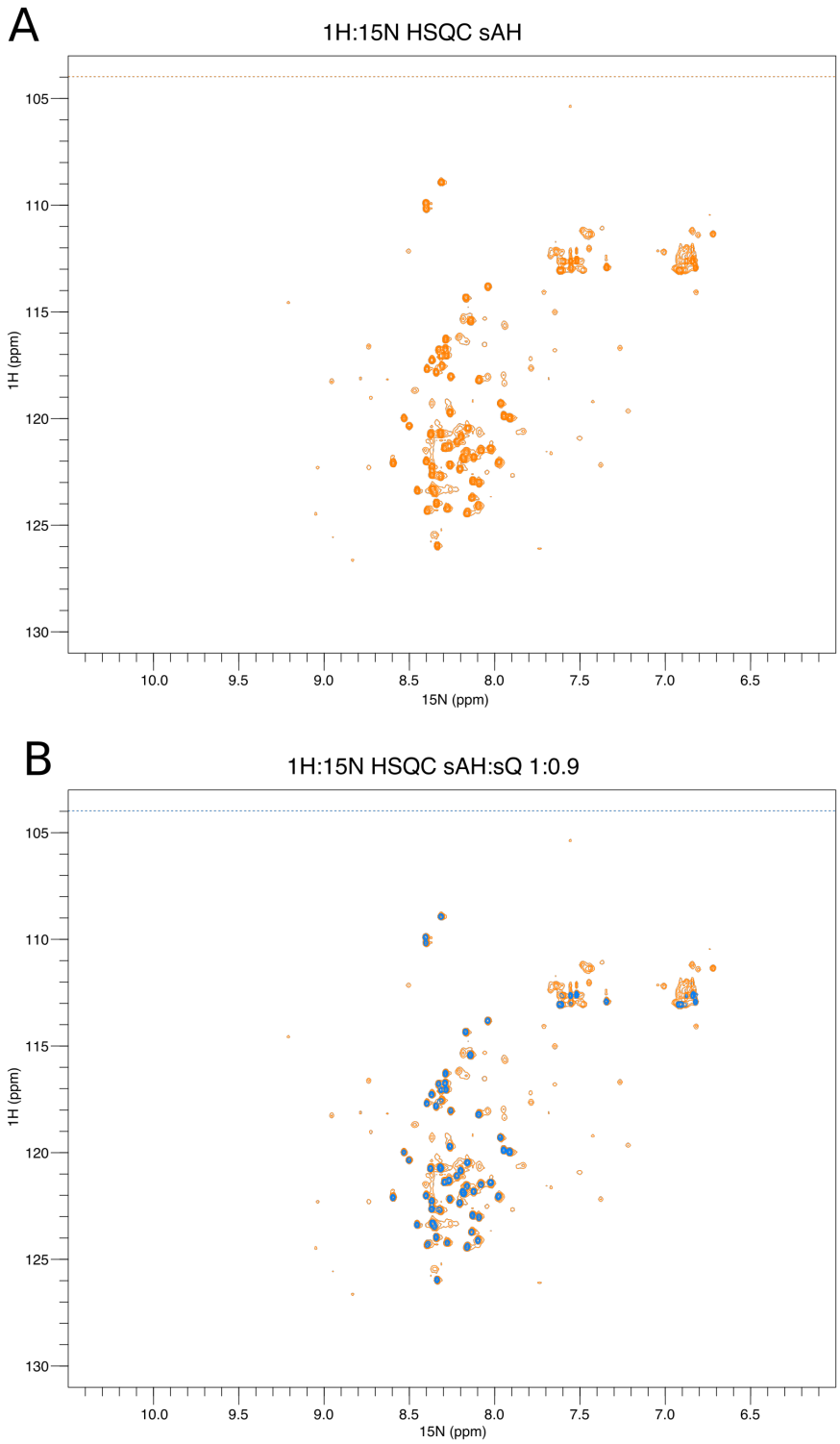


Figure 3.3.4 – ^{1}H:^{15}N HSQC titration of ^{15}N-sAH vs sQ. Molar equivalents of unlabelled sQ were titrated into 308 μM ^{15}N-sAH and ^{1}H:^{15}N HSQC spectra recorded. The number of peaks (260) in the sAH 1:0 (A, orange), was reduced upon addition of unlabelled sQ. 35% of the peaks observed in the 1:0 spectrum disappeared in the 1:0.9 (B, blue) spectra. The number of peaks and protein concentrations in each titre are summarised in Figure 3.3.8A/C. The peaks in the 7.6-8.5 ^{1}H/ppm region became less broad during the titration indicating changes in backbone structure.

First, peaks may no longer be detected because the resultant complex was of a mass that slowed the tumble time of the particle and as a result peaks were unobservable due to signal broadening effects. A 1:1 complex of sQ and sAH had a theoretical mass of 44 kDa which was close to the mass size limit for a 750 MHz INOVA magnet (~60 kDa) in which the experiment was conducted. Secondly, if the rate of the complex formation and dissociation is too fast it will not be observable in the timescales of the NMR experiment. Finally, the titration was conducted by the addition of unlabelled protein to the sample tube. Even though the total protein concentration was maintained, the increase in sample volume resulted in the dilution of the ^{15}N labelled protein (Figure 3.3.8). In the titration containing ^{15}N-sAH, the labelled protein was 53% of the starting concentration in the 1:1.2 experiment and in the ^{15}N-sQ titration at 1:1.2 was only 43% of the starting concentration (Figure 3.3.8). The intensity of peaks is proportional to the concentration of the isotope and dilution of the ^{15}N-labelled sample could have resulted in a decrease in the number of observable peaks at the contour level used (base level 7.5). However, the decrease in the number of observable N-H peaks does not appear to correlate to the concentration of the ^{15}N-labelled protein but rather with the molar stoichiometry to unlabelled protein (Figure 3.3.8). In the ^{15}N-sAH titration the number of peaks counted stabilised at molar equivalents of 1:0.9 and above, whilst in the ^{15}N-sQ titration a significant reduction in peaks was only observed at 1:2. To ensure that the concentration of ^{15}N-labelled protein was more strictly controlled, high concentrations (~1 mM) of unlabelled protein should be used in future titration experiments to minimise an increase in sample volume and subsequent dilution of the labelled protein.

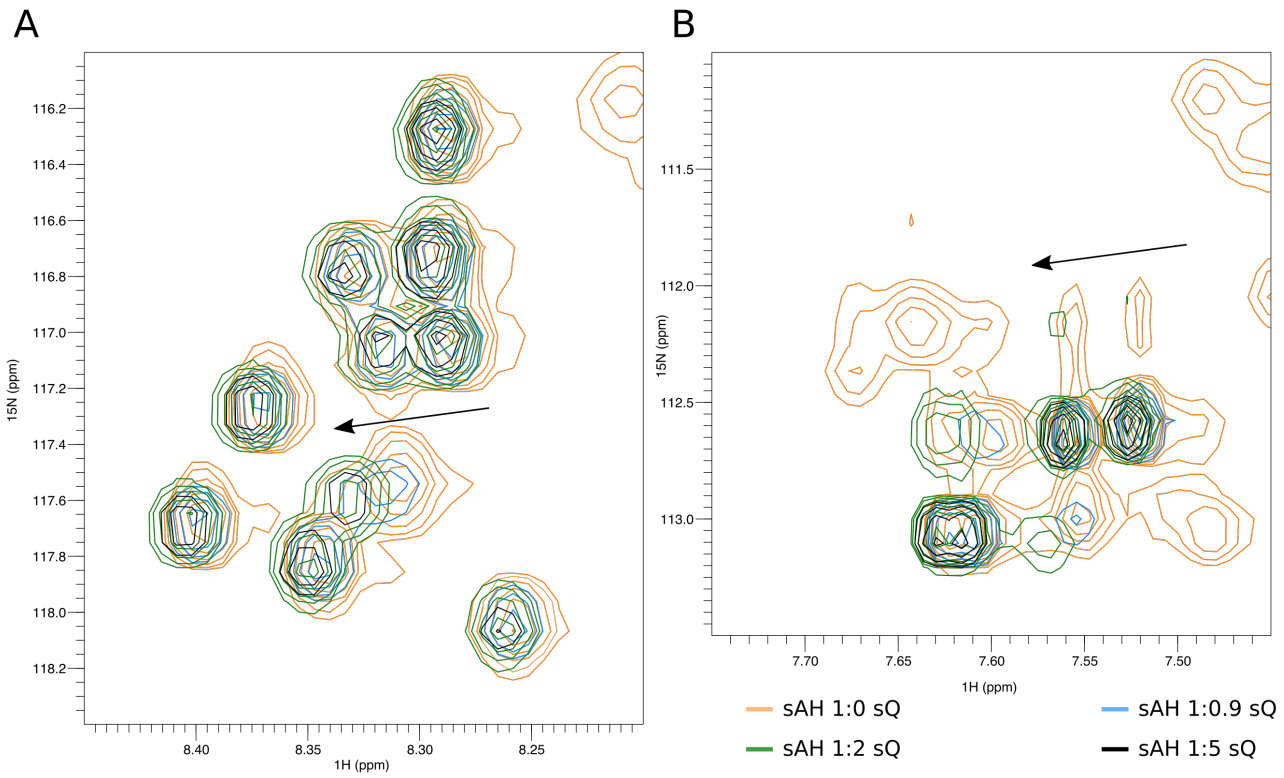


Figure 3.3.5 – ^{1}H:^{15}N HSQC peak shifts observed in ^{15}N-sAH vs sQ titration. Downfield chemical shifts observed in HSQC titration experiments with ^{15}N labelled sAH and unlabelled sQ. The overlaid spectra are 1:0 AH:Q (orange), 1:0.9 (blue), 1:2 (green) and 1:5 (black). **A:** A cluster of downfield shifts in the 8.2-0.45 ^{1}H/ppm, 116-118 ^{15}N/ppm region (indicated by the arrow). **B:** In addition to downfield shifts (shifting in direction of arrow), the 7.65-7.50 ^{1}H/ppm, 113-111 ^{15}N/ppm region included a number of peaks which were no longer visible after sQ was titrated.

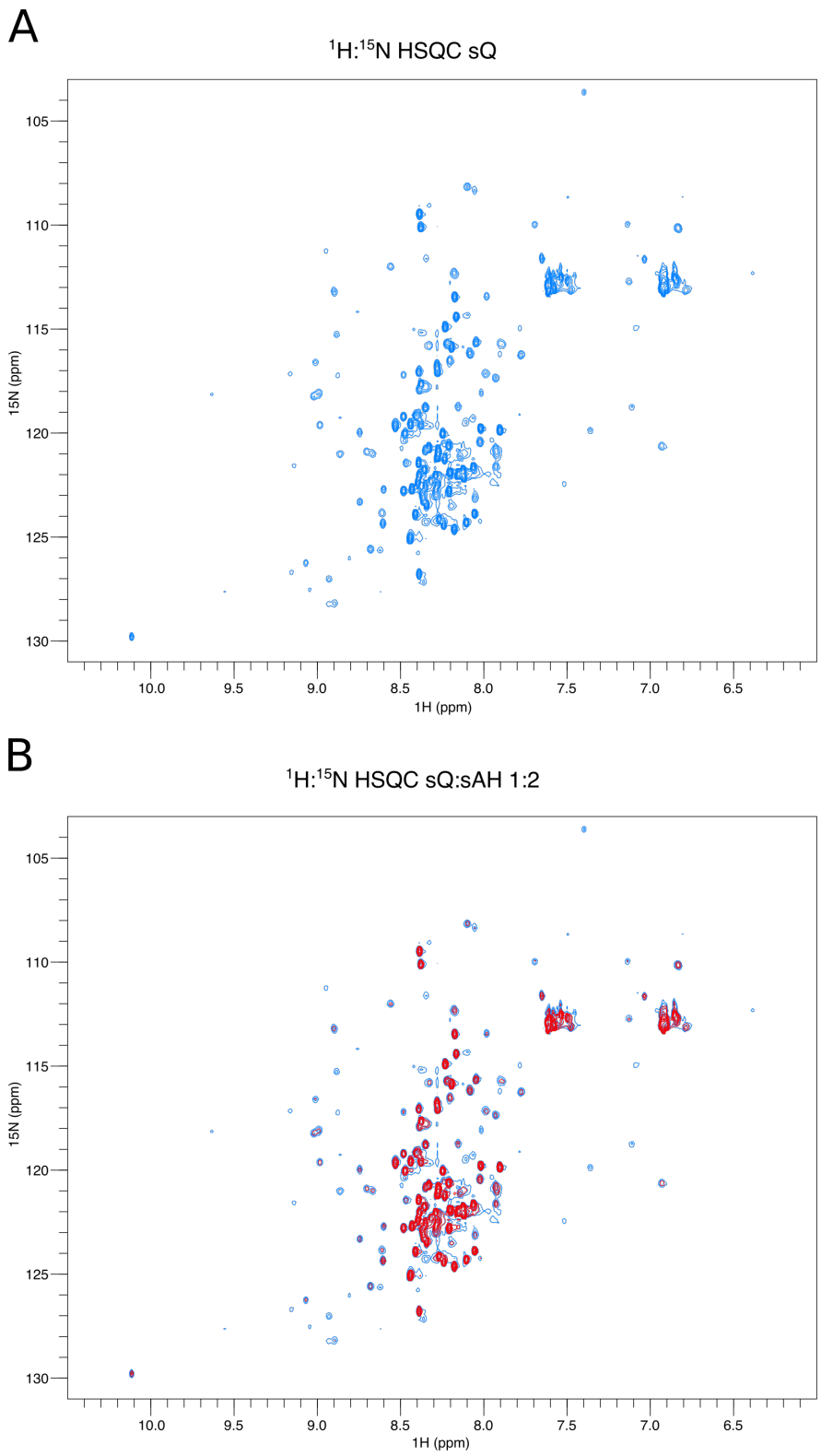


Figure 3.3.6 – ^{1}H:^{15}N HSQC titration of ^{15}N-sQ vs sAH. Unlabelled sAH was titrated into 110 μM ^{15}N-sQ and ^{1}H:^{15}N spectra were recorded (A, blue). Unlabelled sAH was added to the sample tube in equivalent molar ratios of 1:0.22, 1:0.62, 1:1.2 and 1:2 (B, red). There were 20% fewer peaks in the 1:2 spectrum compared with the 1:0 spectrum. The peak number changes are summarised in Figure 3.3.8B/D.

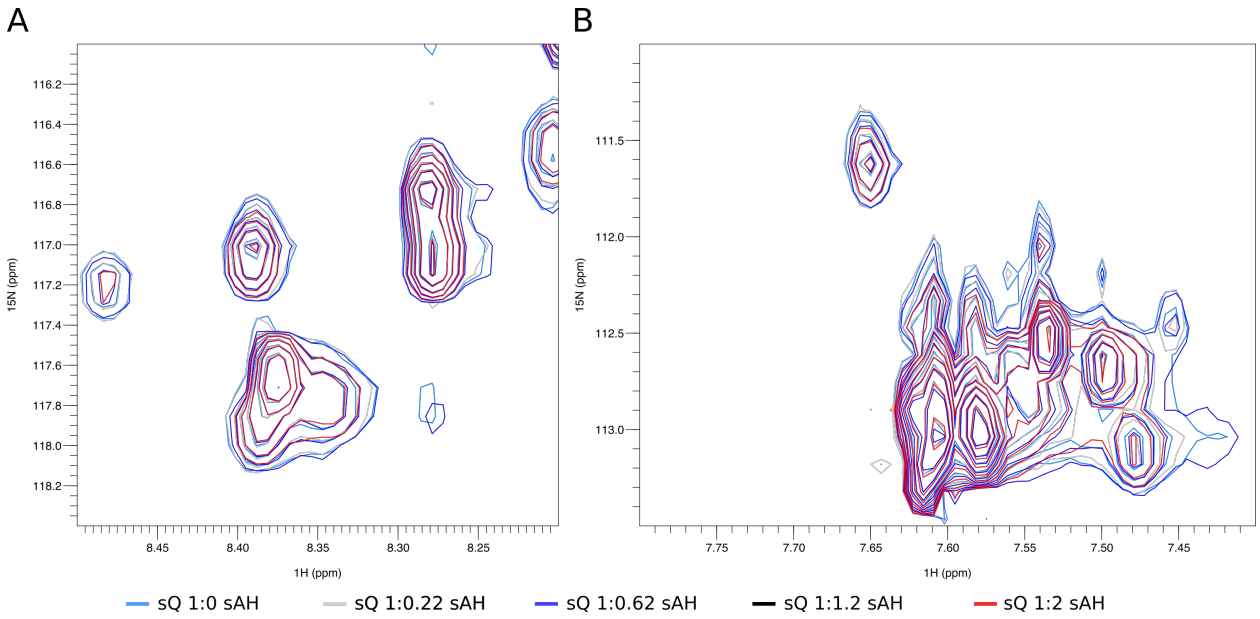


Figure 3.3.7 – Zoom of sQ ^{1}H:^{15}N HSQC titration with sAH. The 8.2-0.45 ^{1}H/ppm, 116-118 ^{15}N/ppm (A) and 7.65-7.50 ^{1}H/ppm, 113-111 ^{15}N/ppm of the 15N-sQ ^{1}H:^{15}N HSQC titration with sAH did not contain any peak shifts as observed in these regions for the 15N-sAH titration with sQ (Figure 3.3.5). The overlaid spectra molar ratios of sQ:sAH of 1:0 (light blue), 1:0.22 (grey), 1:0.62 (dark blue), 1:1.2 (black) and 1:2 (red).

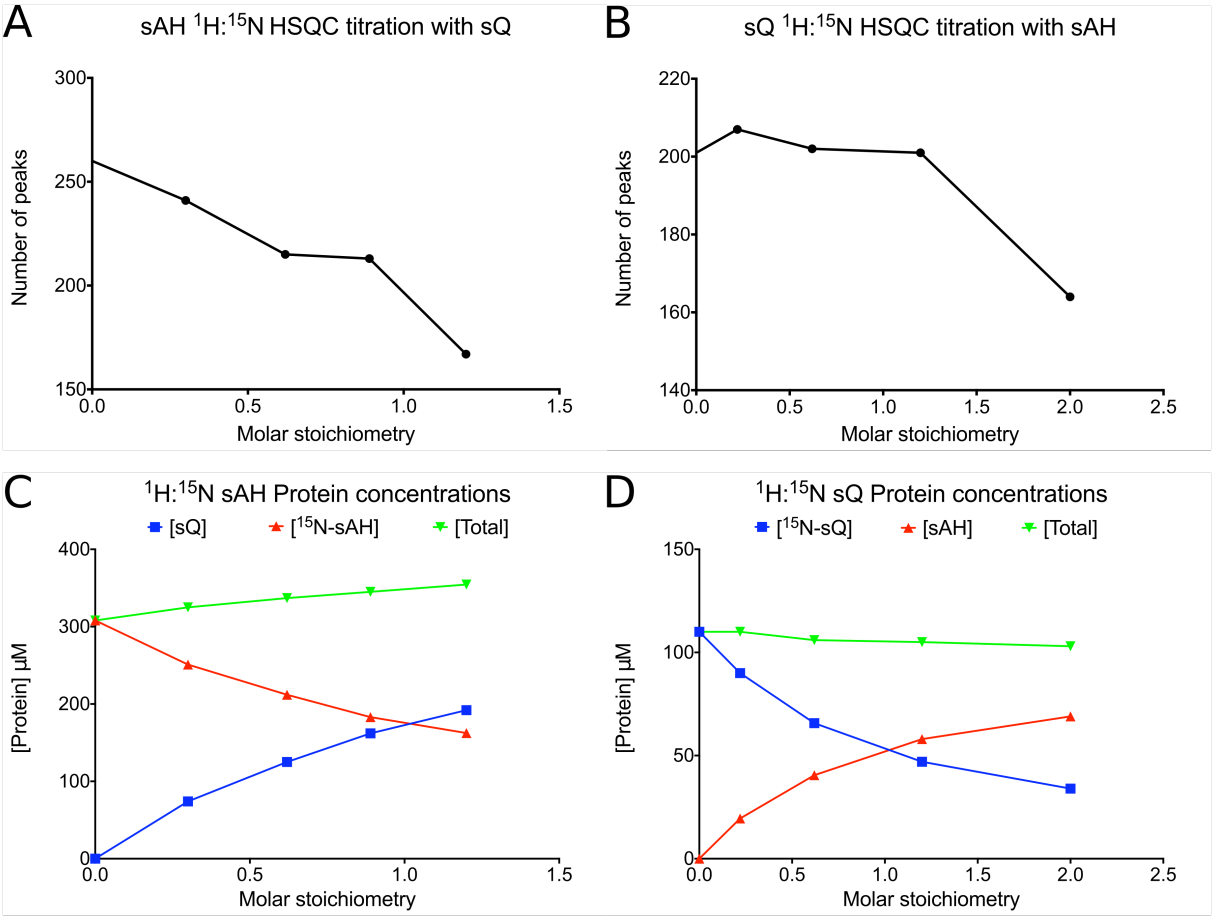


Figure 3.3.8 – Number of H-N peaks vs molar stoichiometry in sQ and sAH ^{1}H:^{15}N HSQC titrations. ^{15}N-sAH and unlabelled sQ (A) showed a downward trend in observed peaks across the titration. The number of peaks in the ^{15}N-sQ titration with sAH (B) were maintained to a stoichiometry of 1:0.9, at 1:2 there was a significant decrease in the number of observed H-N peaks. The concentration of sQ (blue), sAH (red) and the total protein (green) concentration in each HSQC experiment are shown for ^{15}N-sAH (C) and ^{15}N-sQ (D). The total protein concentration remained within 6% and 16% of the initial concentrations of ^{15}N-labelled sQ and sAH titrations, respectively. As unlabelled protein was titrated into the sample tube, the concentration of ^{15}N labelled protein was reduced (C/D).

3.4 Zn^{2+} is required for formation of a stable SpoIIQ:SpoIIIAH complex

3.4.1 sQ binds Zn^{2+}

To investigate whether the LytM conserved motifs in sQ were able to co-ordinate a metal ion, ICP-MS (Section 2.6.4) was used to detect and quantify the presence of bound metal ions. Samples of sQ were either incubated with a 5-fold excess of $ZnCl_2$, in 1 mM of the metal chelator ethylene-diamine-tretra-acetic acid (EDTA) or untreated (without the addition of $ZnCl_2$ during purification). The samples were re-purified using an S200 GL Increase column to separate protein from free ions and fractions were collected for ICP-MS analysis (Figure 3.4.1). In the presence of excess Zn^{2+}, 70% of sQ protein molecules contained the metal ion at a stoichiometry of 1:1 (Figure 3.4.1A). In the presence of EDTA, a strong metal chelator with a K_D in the subfemto molar range (10^{-16} M)(Nyborg and Peersen, 2004), no Zn^{2+} co-elution was observed in sQ, indicating that zinc binding to sQ is weaker than binding to EDTA since the chelator completely removes the metal from its protein binding site.

As a tool for understanding the importance of zinc binding on SpoIIQ, a point mutation was made using reverse PCR to mutate the His120 of the LytM Zn^{2+} binding motif (HXXXD) to Ser, mimicking the motif found in *B. subtilis* SpoIIQ (SXXXD). The point mutant construct, referred to as sQ^{H120S}, was expressed and purified in the same way as sQ (Appendix A, Figure A.0.2).

The point mutant sQ^{H120S} , treated in the same way as sQ, did not contain any Zn^{2+} in any conditions (Figure 3.4.1B) indicating that substitution of the histidine to serine in motif 1 prevented Zn^{2+} coordination by sQ. As sQ/sQ^{H120S} were purified using an NTA HiTrap, it is possible that Ni^{2+} could form a metal complex with sQ. Zn^{2+} sites are also able to coordinate Ni^{2+} but with greater stability than Zn^{2+} (Irving and Williams, 1953). Both metal ions can be coordinated by the N of His and O of Asp/Glu and can both coordinate with coordination numbers of 4, 5 and 6 with similar bond distances (Harding, 2006). Therefore, the concentration of Ni^{2+} was also measured during the ICP-MS experiments (Appendix A, Figure A.0.3) but no co-elution of sQ with Zn^{2+} was observed in the untreated

sample.

These data show that sQ is able to bind Zn^{2+}, in a manner dependent on His120, which could be important for the structure and stability of sQ. The influence of Zn^{2+} could be further studied using techniques such as CD and NMR, however these experiments were conducted in phosphate-containing buffers. Such buffers are not compatible with studies involving Zn^{2+} due to phosphate co-ordination of the metal resulting in the formation of insoluble $Zn_3(PO_4)_2$ crystals. Although MES buffer is compatible with Zn^{2+} binding studies (Table 2.5.1), it is not an appropriate buffer for CD experiments due to its strong absorption in the far-UV region (210-190 nm). It was not possible to find a suitable buffer that could buffer at pH 6.0 (the buffer pH at which sQ was shown to be folded and stable), did not interact with free Zn^{2+} and did not absorb light in the far-UV region. Buffers such as citrate, Bis-Tris or cacodylate did not fulfil these requirements. As a result, the influence of Zn^{2+} upon secondary structure and stability could not be studied using these methods. Alternatively, if the metal ion site was able to co-ordinate a different metal, such as Ni^{2+}, this could be used to alleviate the incompatibility of the metal ion with the buffer systems required for these experiments. However, when analysed by ICP-MS co-elution of protein with Ni^{2+}(Figure A.0.3) and Mn^{2+} was not observed, indicating that the LytM motif has a clear preference for Zn^{2+}.This observation showed that the LytM motifs in *C. difficile* SpoIIQ could co-ordinate a metal ion and that His120 was required for Zn^{2+} binding.

Based on the data from these ICP-MS experiments, all further experiments in this chapter were conducted with sQ where a 5-fold excess of Zn^{2+} was added before size-exclusion chromatography (Section 2.5.4) to ensure complete loading of the metal site. Care had to be taken when incubating Zn^{2+} with sQ, incubation with a 10-fold molar excess of Zn^{2+} with sQ resulted in precipitation of the protein and so a 5-fold molar excess of Zn^{2+} was used. Additionally it was essential to remove any free metal from the sample which may interfere with downstream experiments such as crystallisation and size-exclusion chromatography provided a mechanism to separate metal loaded sQ and free metal ions.

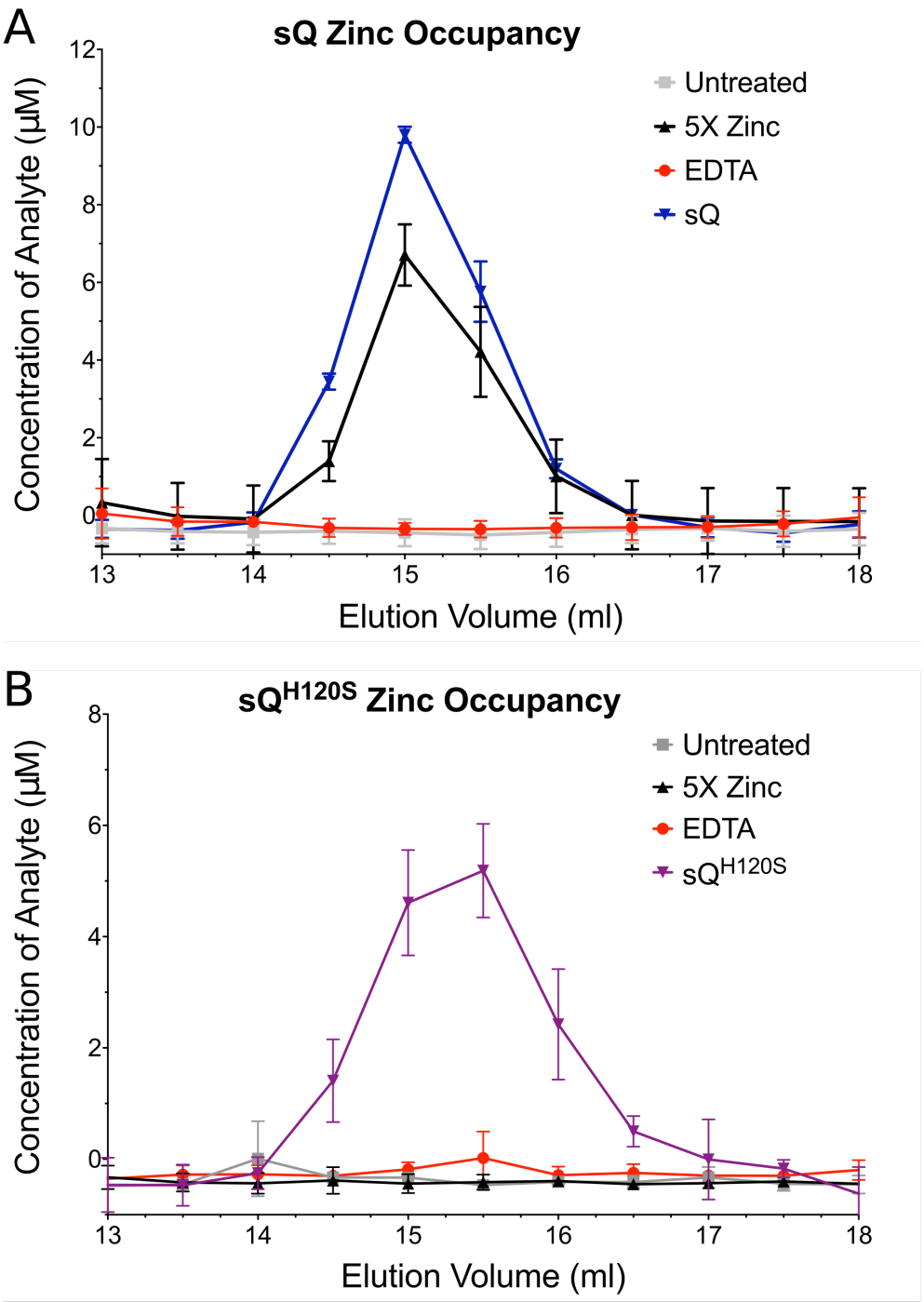


Figure 3.4.1 – Metal binding analysis of sQ and sQ^{H120S} by ICP-MS. A: 10 µM sQ was incubated with either 1 mM EDTA or a 5-fold excess of $ZnCl_2$ before injection on an S200 10/300 GL Increase column. A 10 µM sample of sQ was left untreated and also applied to the column. Fractions of 0.5 ml were collected during elution of each protein. The metal content was analysed using ICP-MS and the protein concentration determined by absorbance at 280 nm. Zinc concentration in the fractions of the samples are shown: 1 mM EDTA (red), untreated (grey) and 5-fold Zn^{2+} (black). The sQ protein concentration is also shown (blue). Only in the presence of excess $ZnCl_2$ was sQ observed to co-elute with Zn^{2+}. No Zn^{2+} was detected co-eluting with sQ in the untreated or EDTA containing samples. **B:** 10 µM of sQ^{H120S} was treated and analysed in an identical manner to sQ. sQ^{H120S} protein concentration is shown in purple. No Zn^{2+} is detected in the fractions containing sQ^{H120S} under any condition showing that His120 is required for Zn^{2+} metal binding in sQ.

3.4.2 Determination of complex formation by SEC-MALLS

The molecular masses for sQ, sQ^{H120S}, sAH and complexes were determined using SEC-MALLS. The theoretical masses of the proteins as calculated using ProtParam (Wilkins *et al.,* 1999) were: sAH, 22.8 kDa; sQ and sQ^{H120S}, 21.3 kDa. The SEC-MALLS experiments determined masses of 22.3 kDa, 20.9 kDa and 20.8 kDa for sAH, sQ and sQ^{H120S} respectively, which are comparable to the theoretical values. Therefore, individual proteins sQ (loaded with Zn^{2+}), sQ^{H120S} and sAH were determined to be in a monomeric state in solution (Figure 3.4.2A). The masses of sQ (loaded with Zn^{2+}) and sQ^{H120S} differ by 0.1 kDa in the SEC-MALLS experiment, demonstrating that Zn^{2+} loading did not alter the oligomeric state of sQ and the mutation of histidine-120 to serine did not have a significant global effect on the protein.

These results are contrary to the migration profile observed in an S200 26/600 column during purification that suggested the proteins were in a dimeric state when compared with molecular mass protein standards (Figure 3.2.4). The same running buffer was used in the purification by size exclusion chromatography and SEC-MALLS experiments. It is most likely that the shape of sAH and sQ differ from the generally globular proteins used to calibrate the size exclusion column, resulting in sAH and sQ displaying characteristics of proteins of greater mass than the mass for their actual oligomeric state. SEC-MALLS is a direct measurement of the light scattering properties of a molecule in solution while the determination of the mass of a molecule using its retention properties by a size exclusion column is a comparative method. As such, SEC-MALLS is a preferred method for determining the oligomeric state of proteins. The HSQC-NMR spectra presented in Section 3.3 also indicated that sAH and sQ did not have an overall globular shape but these spectra did not give detailed information on the oligomeric state of sAH or sQ.

When 119 μM sQ (loaded with Zn^{2+}) and 110 μM sAH were mixed, a stable sQ:sAH complex was formed (Figure 3.4.2B) as indicated by a significant peak containing both proteins as visualised by SDS-PAGE (Figure 3.4.2C). The major peak observed by SEC-MALLS when sQ and sAH were mixed on a 1:1 molar ratio had a calculated mass of 43.1 kDa (Figure 3.4.2B), which corresponded to the sum of the SEC-MALLS calculated masses of sQ and sAH (Figure 3.4.2A). A minor peak with a mass of 23.6 kDa was also observed in the sQ:sAH sample that indicated there was a small excess of one of the

proteins in the mixed sample, which analysis of the fractions by SDS-PAGE suggested was an excess of sQ. The relatively stable molar mass across the major peak indicated that the sQ:sAH complex was stable under the conditions of the SEC-MALLS.

In the presence of the metal chelator EDTA, sQ and sAH did not form a stable complex (orange, Figure 3.4.2B). No stable complex was observed also when sQ^{H120S} is mixed with sAH (magenta, Figure 3.4.2B). When analysing both of the Zn^{2+} deficient samples, single elution peaks were observed that had a tail profile, indicative of a dissociating complex. The molecular mass decreased across the peak, resulting in sloping molar masses (Figure 3.4.2B) that indicated that the molecules within this peak were not homogenous. The absolute mass across this peak was ~27.5 kDa, 25% greater than the masses of sQ and sAH monomers alone, indicating that there were some molecules that had greater mass than monomeric sQ or sAH, presumably a sQ:sAH complex. Since this measured mass was closer to the masses of the individual monomeric proteins, it suggested that the majority of the species measured by the light scattering and refractive index detectors in these samples were in a monomeric state. The presence of both sAH and sQ in the peak fractions, which were from lower elution volumes than in the individual samples, suggested that the proteins can interact in the absence of Zn^{2+}, albeit in a less stable manner. Only in the presence of Zn^{2+} was a stable sQ:sAH complex observed, indicating that the metal ion is important for complex formation and stability. The $Zn^{2+}$is perhaps affecting the fold stability of the sQ protein that is required for a stable interface with SpoIIIAH to be formed.

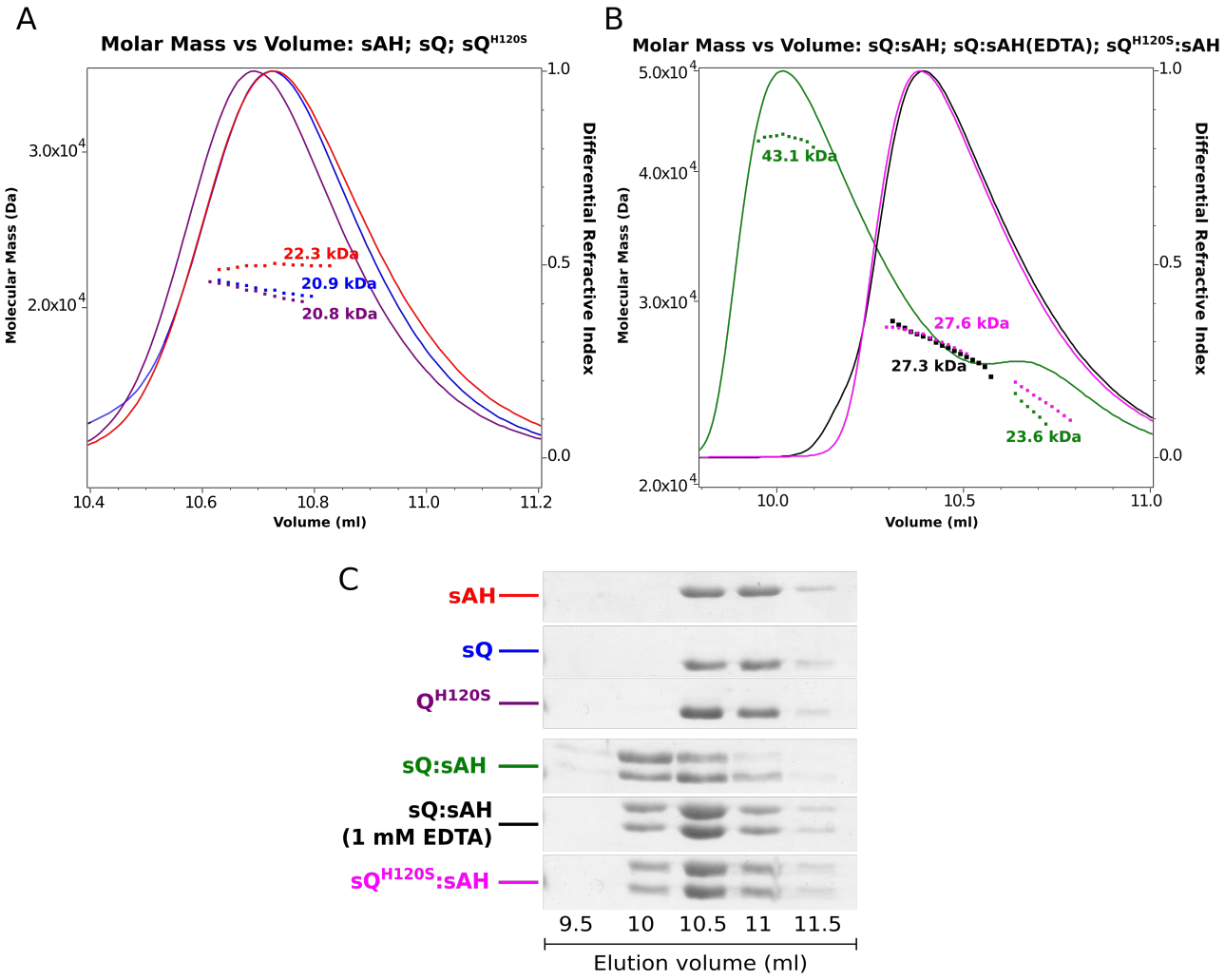


Figure 3.4.2 – SEC-MALLS analysis of sAH, sQ and in complex. A: Samples of the individual proteins (sQ: 119 μM; sAH: 111 μM) were determined to have the following masses: sAH (red) at 22.3 kDa ±0.15%; sQ (blue) at 20.9 kDa ±0.13%; sQ^{H120S} (purple) at 20.8 kDa ±0.14%. **B:** The proteins were mixed at a final concentration of ~115 μM as follows: sQ:sAH with Zn^{2+} (1:1, green), sQ:sAH (1:1) in the presence of 1 mM EDTA (black) and sQ^{H120S}:sAH (1:1) (magenta). In the presence of Zn^{2+} two species elute in sQ:sAH with masses of 43.1 kDa ±0.11% and 23.6 kDa ± 0.23%. The theoretical mass of the sQ:sAH complex was 44.1 kDa indicating that a stable heterodimeric complex had been formed. The smaller peak mass of 23.6 kDa indicated an excess of one of the proteins. Single elution peaks of 27.6 kDa ±0.10% and 27.3 kDa ±0.12% were observed in samples of sQ^{H120S}:sAH and sQ:sAH in the presence of EDTA which were indicative of a dissociating complex. **C:** SDS-PAGE analysis of 0.5 ml fractions collected from 9.5-11.5 ml was performed. In the sQ:sAH sample, the SDS-PAGE confirmed an excess of sQ protein. Co-elution of both sAH and sQ (EDTA) or sQ^{H120S} was observed in the SDS-PAGE in the 10 ml fraction indicating that a partially dissociating complex was formed. An Wyatt SEC-050S5 column was connected to an Äkta Pure [GE Healthcare] with Wyatt DAWN HELEOS 8 and Optilab T-rEX detectors connected downstream.

3.4.3 Determination of complex affinity by microscale thermophoresis

Microscale thermophoresis (MST) was used to measure the binding affinities between sQ and sAH (Section 2.6.6). Having determined using SEC-MALLS that Zn^{2+} was required for stable complex formation, MST experiments were performed with sQ that had been previously incubated with Zn^{2+}. Both sQ and sAH were labelled with the Cy5 fluorophore via an amine coupling reaction, and the partner protein titrated over a 20 point 2:1 serial dilution. Normalised fluorescence data were plotted and the K_D determined by fitting the data using the law of mass action (Section 2.6.6). In the MST experiment where sAH was titrated against labelled sQ, a K_D of 44.6±7.2 μM was fitted (Figure 3.4.3A). In the opposing experiment, where sQ was titrated against labelled sAH, a K_D of 77.8±22.7 μM was fitted (Figure 3.4.3B). These binding affinities are within 2-fold of each other and, therefore, are likely to represent the same interaction between sQ and sAH. The data at high titre concentrations in both Figure 3.4.3A and B are less consistent than at lower titres. The highest concentration of unlabelled protein in these titrations was 5.4 mM. At such concentrations, the viscosity of the sample is high and this may change the thermophoretic properties of the sample, resulting in less consistent diffusion of molecules through solution and thus less consistent changes in fluorescence. Changes in solution viscosity indicate that the hydration shell of the molecules in the sample has changed, an intrinsic variable in thermophoresis and as a result can be an artefact in MST data (Baaske *et al.,* 2010; Duhr and Braun, 2006b; Jerabek-Willemsen *et al.,* 2011; Wienken *et al.,* 2010).

SEC-MALLS showed that without Zn^{2+}, sQ and sAH formed an unstable complex. An attempt was made to quantify the effect of Zn^{2+} by MST using sQ^{H120S} instead of wildtype sQ. However, it was not possible to obtain reliable data from MST experiments that could be used to calculate a K_D(Appendix A, Figure A.0.4). The SEC-MALLS was performed using ~110 μM of each protein and so the K_D for the interaction between sAH and non-Zn^{2+} containing sQ must be higher than 110 μM. To quantify this interaction would require high concentrations of unlabelled protein that can lead to highly inconsistent results due to viscosity artefacts, as discussed.

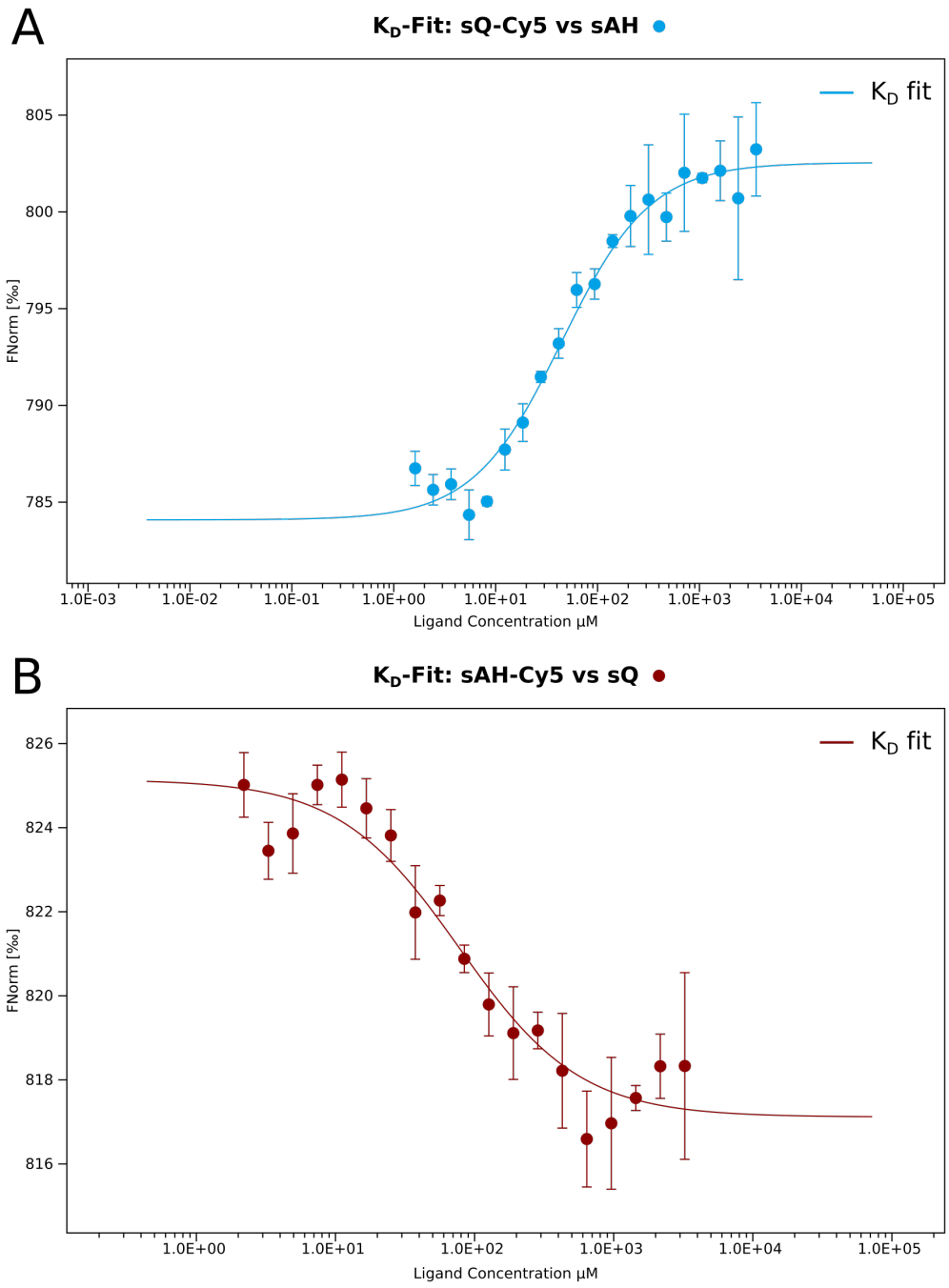


Figure 3.4.3 – sQ and sAH binding affinities by MST. MST binding curves with sQ labelled with Cy5 fluorophore (A) at a constant concentration of 0.01 µM and a 2:1 serial dilution of sAH at the highest concentration of 5.4 mM. The K_D of the fitting of this curve is 44.6±7.2 µM (A, blue): χ^2 = 4.8; standard error = 1.0. sAH labelled with Cy5 fluorophore (B) at a constant concentration of 0.17 µM with a 2:1 serial dilution of sQ at the highest concentration of 7.3 mM. A K_D of 77.8±22.7 µM was fitted (B, red): χ^2 = 1.1; standard error = 0.8.

3.5 Crystallisation of sQ and sAH

Nearly 11,000 individual crystallisation trial drops (57 X 96-well 2-drop plates) of sQ (loaded with Zn^{2+}) and sAH, individually and mixed together, were set up in either commercial sparse matrix crystallisation screens available in the laboratory (Section 2.7.1) or custom made optimisation screens (summarised in Table 3.5.1). Starting at a total protein concentration of 9 mg/ml, the protein concentration was increased over a number of crystallisation trials to a maximum of 100 mg/ml as the majority of the drops were clear and contained little precipitation. High protein concentrations such as 100 mg/ml resulted in phase separation or improper mixing with crystallisation solutions.

Protein sample	Total protein concentration	Crystallisation screen	Drop ratio	Total drop volume
sQ	19 mg/ml	Index, JCSG+, Morpheus, PACT, Structure	1:1, 2:1	200 nl
sQ	58.4 mg/ml	$AMSO_4$, Index, Morpheus, PACT, Structure	1:1, 2:1	100 nl
sQ	58.4 mg/ml	JCSG+, Structure	1:1, 2:1	300 nl
sQ + chymotrypsin (1:10000)	19 mg/ml	JCSG+, PACT	1:1, 2:1	100 nl
sQ^{H120S}	19 mg/ml	Index, JCSG+, Morpheus, PACT, Structure	1:1, 2:1	200 nl
sAH	84 mg/ml	Index, Structure	1:1, 2:1	100 nl
sAH	100 mg/ml	Morpheus, PACT	1:1, 2:1	100 nl
sAH-K(CH_3)	18 mg/ml	Cryo, Index, JCSG+, Morpheus, PACT, Structure	1:1, 2:1	100 nl
sAH + sQ (1:1)	25 mg/ml	$AMSO_4$	1:1, 2:1	100 nl
sAH + sQ (1:1)	48.5 mg/ml	$AMSO_4$, Index, Morpheus, PACT, Structure	1:1, 2:1	100 nl
sAH + sQ (1:1)	50 mg/ml	MIDAS	1:1, 2:1	100 nl
sAH + sQ (1:1)	84 mg/ml	Index, JCSG+, PACT, Structure	1:1, 2:1	100 nl
sAH-K(CH_3) + sQ-K(CH_3) (1:1)	9 mg/ml	Opt1_Meisner, Opt1_Levdikov	1:1, 2:1	100 nl
sAH-K(CH_3) + sQ-K(CH_3) (1:1)	20 mg/ml	$AMSO_4$, Cryo, Index, JCSG+, Morpheus, PACT, Structure, Opt1_Meisner, Opt1_Levdikov	1:1, 2:1	100 nl
BS-$Q_{25\text{-}218}$ + CD-AH (1:1)	49 mg/ml	$AMSO_4$, Index, JCSG+, Morpheus, PACT, Structure	1:1, 2:1	100 nl

Table 3.5.1 – Summary of sQ, sAH and sQ:sAH crystallisation experiments. A list of the crystallisation experiments performed at 20° C using sQ or sAH individually or mixed together at a 1:1 molar ratio. The total protein concentration in the sample (sum of each protein concentration) is given. Lysine methylated samples (sQ-K(CH_3)), sAH-K(CH_3)) and complexes are shown in Figure 3.5.1. Screens Opt1_Meisner and Opt1_Levdikov were developed based on the successful crystallisation conditions found by Levdikov *et al.,* and Meisner *et al.,* for the crystallisation of *B. subtilis* SpoIIQ:SpoIIIAH complex (Meisner *et al.,* 2012; Levdikov *et al.,* 2012) and are described in Appendix A, Tables A.0.1 and A.0.2. *B. subtilis* SpoIIQ (BS-$Q_{25\text{-}218}$) expression vector produced by Dr Jeffrey Meisner was kindly provided by Prof Adriano Henriques.

In the absence of crystallisation hits, so-called 'crystallisation rescue strategies' (Section 2.7.2) such as *in situ* proteolysis and lysine methylation were attempted (Figure 3.5.1) to reduce protein solubility (sAH: 16% Lys; sQ: 14% Lys). Successful methylation of sQ, sAH and the sQ:sAH complex was confirmed by a shift in elution volume from an S200 16/600 size exclusion column as shown in Figure 3.5.1, indicating that surface properties of the proteins had changed. The complex was formed before the lysine methylation reaction was performed in an effort to not disrupt the sQ:sAH interaction. After lysine methylation, sAH and the sQ:sAH complex remained stable and were concentrated to ~ 20 mg/ml for crystallisation trials. However, sQ-K(CH_3) precipitated during concentration and it was not possible to set up crystallisation trials.

The data presented in Section 3.4 indicates that the sQ:sAH complex does not have a strong affinity (K_D 40-80 μM) and in certain conditions is unstable, particularly if sQ does not co-ordinate Zn^{2+}. These data suggest that crystallisation of sQ and sAH would be challenging and this has proved to be the case. Proteins that form crystals are well structured with few regions of flexibility which enables the formation of a lattice arrangement of molecules and sQ, sAH or the complex do not fulfil this requirement as shown by HSQC-NMR (Figure 3.3.3).

In the absence of crystals of Clostridial sQ and sAH, formation of a chimeric complex using *B. subtilis* and *C. difficile* proteins was attempted. *B. subtilis* SpoIIQ$_{25\text{-}218}$ and SpoIIIAH$_{43\text{-}283}$ expression vectors were produced by Dr Jeffrey Meisner and were kindly provided by Prof Adriano Henriques. These constructs were expressed and purified following the published protocols by Meisner *et al.,* (Meisner and Moran, 2011; Meisner *et al.,* 2012). Figure 3.5.2 shows that *B. subtilis* SpoIIQ$_{25\text{-}218}$ (BS-Q) and *C. difficile* sAH (CD-AH) could form a stable complex. Crystallisation trials of this complex were set up in commercial sparse matrix screens at a total protein concentration of ~49 mg/ml but did not yield crystals.

Interestingly, no complex was detected upon incubation of *C. difficile* sQ and *B. subtilis* sAH, suggesting some intrinsic difference between the two SpoIIQ proteins.

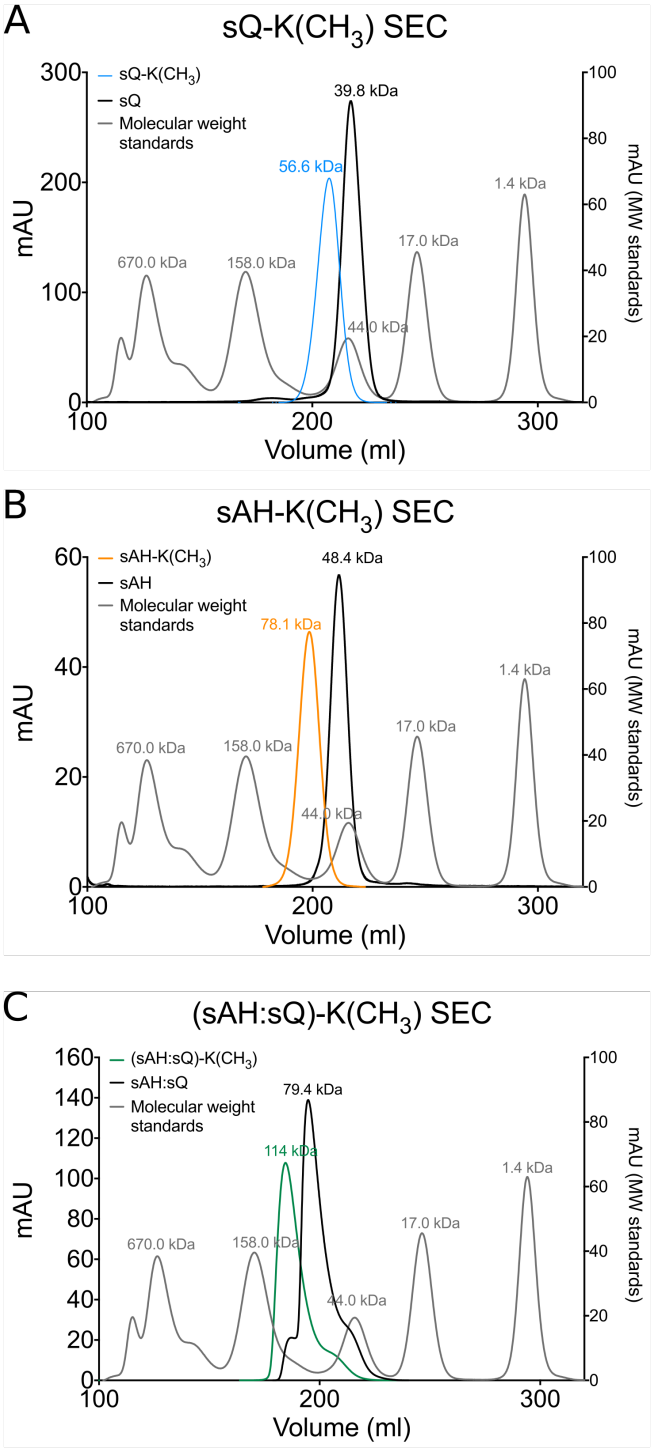


Figure 3.5.1 – SEC of lysine methylated sQ, sAH and sAH:sQ. Comparison of UV chromatograms of lysine methylated -K(CH_3) protein with non-methylated protein on SEC. Non-methylated sQ (A, black) eluted with a volume equivalent to 39.8 kDa, sQ-K(CH_3) (A, blue) eluted with an apparent molecular mass of 56.6 kDa. Non-methylated sAH (B, black) eluted at a volume equivalent to 48.4 kDa and sAH-K(CH_3) (A, orange) at 78.1 kDa. When sQ:sAH complex was lysine methylated (C, green) the complex eluted at a lower elution volume, equivalent to 114.0 kDa. Non-treated sQ:sAH complex eluted at a volume equivalent to 79.4 kDa (C, black). The UV trace of molecular mass standards are shown in grey. The molecular mass standards were thyroglobulin (670.0 kDa), γ-globulin (158.0 kDa), ovalbumin (44.0 kDa), myoglobin (17.0 kDa) and Vitamin B_{12} (1.4 kDa). The increase in equivalent elution mass indicates that sQ, sAH and sAH:sQ were methylated.

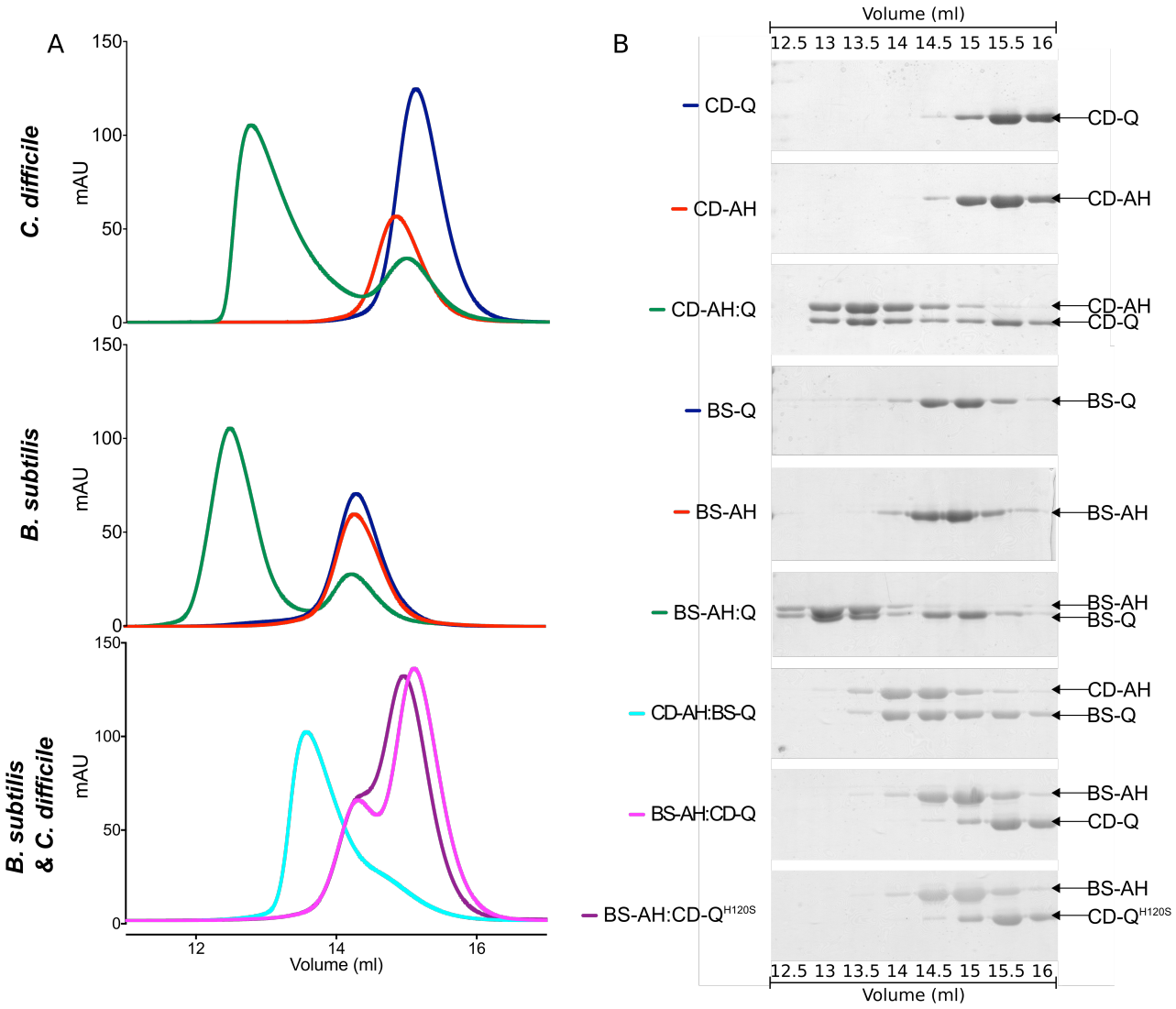


Figure 3.5.2 – SEC of chimeric *B. subtilis* and *C. difficile* complex. A: S200 Increase size-exclusion UV traces (A280) of *C. difficile* sQ (CD-Q, blue), sAH (CD-AH, red) and sQ:sAH (CD-AH:Q, green). *B. subtilis* SpoIIQ$_{25\text{-}218}$ (BS-Q, blue), SpoIIIAH$_{43\text{-}283}$ (BS-AH, red) and BS-Q:AH (green). The chimeric mixes of CD-AH:BS-Q, BS-AH:CD-Q and BS-AH:CD-Q^{H120S}. **B:** Fractions from elution volume 12.5-16 ml are compared from each size-exclusion performed in A. Only when CD-AH was mixed with BS-Q was a complex formed as observed by a single peak (at a lower elution volume than the individual proteins) and co-elution in the elution fractions. CD-AH:BS-Q sample was used to setup crystallisation trials at a total protein concentration of 49 mg/ml.

3.6 Discussion

The data presented in this chapter on the C-terminal, intersporangial domains of SpoIIQ and SpoIIIAH in *C. difficile* give new insight into this essential sporulation complex. Knowledge of the SpoIIQ:SpoIIIAH complex has, until recently, been based on observations in *B. subtilis*, the model Gram-positive spore forming bacterium. The aim of this work was to characterise SpoIIQ and SpoIIIAH from *C. difficile,* investigate the metal binding capabilities of the intact LytM domain of SpoIIQ, complex formation between SpoIIQ and SpoIIIAH, and determine the crystal structure of the complex. This work demonstrates that the SpoIIQ:SpoIIIAH complex has different characteristics in *C. difficile* from *B. subtilis*, highlighting an important role of SpoIIQ Zn^{2+} binding upon the stability of complex formation, a property not observed in *B. subtilis*.

3.6.1 Structural studies of SpoIIQ:SpoIIIAH in *C. difficile*

Crystallisation of SpoIIQ and SpoIIIAH from *C. difficile* proved unsuccessful. Whilst CD shows that both sQ and, to a greater extent, sAH contain secondary structure, the HSQC-NMR data presented in Section 3.3 show that both of these proteins contain unordered, flexible regions. Proteins that crystallise are well ordered, with few regions of flexibility and the CD and NMR data suggest that sQ and sAH do not fulfil these criteria and are therefore challenging targets for crystallisation. Both sQ and sAH could be concentrated to a very high concentration (max. 130 mg/ml) and not precipitate. Both proteins contain a high proportion of lysine residues (sQ 14%, sAH 16%) which may help to explain the high solubility of these proteins. Lysine methylation was performed (Figure 3.5.1), however, sQ became unstable and sAH and sAH:sQ samples did not yield crystals. *In situ* proteolysis was attempted to remove flexible, unordered regions of sQ protein to enable crystallisation of a more ordered product but this again did not yield crystals.

Co-crystallisation of sQ and sAH in complex also proved challenging with a K_D in the 40-80 μM range, suggesting that these molecules would not form a complex stable enough to enable nucleation and crystal growth to occur. Co-crystallisation of sQ and sAH was performed at concentrations in excess of 2 mM per protein. The proven dependency of Zn^{2+} occupation of sQ and the effects upon complex formation also provide another

variable that may affect crystallisation.

All of these factors reduce the possibility of crystallising sQ and sAH. A remaining strategy would be to change the sQ and sAH constructs by either removing residues at the termini or including more residues than had been truncated from sQ and sAH. Secondary structure prediction (PSIPRED) and disorder prediction (PONDR) could be used as a guide for this approach, removing random coil regions at the termini.

Both *B. subtilis* crystal structures were determined with different length truncations of the N-terminal regions of SpoIIQ and SpoIIAH although the full extent of the crystallised protein constructs were not observed in the calculated electron densities. The construct used for 3TUF contains residues 43-283 of SpoIIQ (residues 75-232 modelled) and residues 25-218 of SpoIIIAH (residues 103-207 modelled), where as 3UZ0 contains SpoIIQ residues 73-220 (residues 78-220 modelled) and SpoIIIAH residues 90-218 (residues 104-217 modelled). The *C. difficile* sQ and sAH constructs used here mimicked the expression constructs of the *B. subtilis* complex by Levdikov *et al.*, by consensus sequence. Since considerable regions of the N-terminus of the SpoIIQ were missing in the electron density of both 3TUF and 3UZ0, constructs that mimic the coverage observed in these crystals may prove more successful in crystallisation trials, such as $SpoIIQ_{61\text{-}212}$ and $SpoIIIAH_{111\text{-}231}$.

Co-expression of sQ and sAH using the pET-Duet system, enabling formation of the complex in *E. coli* expression cells, may also lead to a more stable complex that is more readily crystallised. To improve the stability of the sQ:sAH complex formation it may be possible to co-express the complex as a fusion protein with a linker peptide between the termini of sAH and sQ (Reddy-Chichili *et al.,* 2013). This approach reduces the space in which sAH and sQ can move by tethering the two proteins together and in addition ensures a 1:1 ratio of both proteins. Poly-glycine linkers of varying lengths could provide a flexible peptide chain enabling sAH and sQ to form their native interaction. However, such a linker may need to be relatively long as the greatest distance between the termini of the observed peptide chain in the crystal structure of *B. subtilis* SpoIIQ:SpoIIIAH is ~50 Å, and such a lengthy linker will itself likely prove disadvantageous towards crystallisation.

Structural studies could instead be carried out using other techniques such as electron microscopy (EM) of full-length membrane associated SpoIIQ and SpoIIIAH constructs, or

NMR. Full-length SpoIIQ and SpoIIIAH are predicted to form ring structures within their respective membranes of possibly at least 12 subunits. Such assemblies would be at least 360 kDa in mass and within the mass ranges for EM studies (~150 kDa). Additionally, detergent-free sample preparation methods (SMALPS) have proven highly useful in the determination of membrane associated structures using EM (Postis *et al.,* 2015). Backbone assignment of sQ and sAH using 3D NMR (^{1}H:^{15}N:^{13}C) would enable more detailed analysis of the structural changes that may take place in sQ when Zn^{2+} is present and when complex is formed with sAH. Since the NMR experiments in this thesis were conducted, the dissociation constant of the complex was determined, which would enable more refined titrations with unlabelled binding protein to be carried out. Refinement of the HSQC pulses during the experiments to control for fast on-off times between the interacting molecules and their tumble time in relation to mass may reveal greater detail in titrations. To determine a structural ensemble by NMR of sAH and sQ would be a challenging last resort to acquire structural information of sAH and sQ at the atomic level.

Whilst there is no evidence in the data presented in this chapter that sQ and sAH could form higher order structures in solution, *in vivo* the domains studied here would be located on a membrane surface. Their localisation upon a surface may aid their assembly into higher order structures. Both SpoIIQ and SpoIIIAH are predicted to oligomerise in the forespore and mother-cell membrane, respectively, forming pore-like ring structures within the membranes. The SpoIIQ and SpoIIIAH ring assemblies dock, forming a channel between the two cells. Molecular energy minimising modelling and homologue modelling using the *B. subtilis* crystal structures have provided models of assemblies of 12 (Figure 3.6.1)(Levdikov *et al.,* 2012) and 15 or 18 unit pore-rings (Meisner *et al.,* 2012). These models are based on the sequence similarity (22% sequence identity) of SpoIIIAH to the type III secretion protein EscJ of which the crystal structure has 6-fold symmetry arranged in a helix, a complete turn containing 24 molecules. If *C.* difficile SpoIIIAH forms such structures, with at least 12 subunits, the mass of such a assembly (>360 kDa) would be large enough to observe using EM. Negative stain EM would be a suitable technique to determine whether full-length SpoIIQ and SpoIIIAH are able to form ring-like structures on a surface/lipid environment. If so, further experiments such as single-particle analysis could be used to determine the diameter of such structures and the number of molecules

of SpoIIQ and SpoIIIAH required to form such rings.

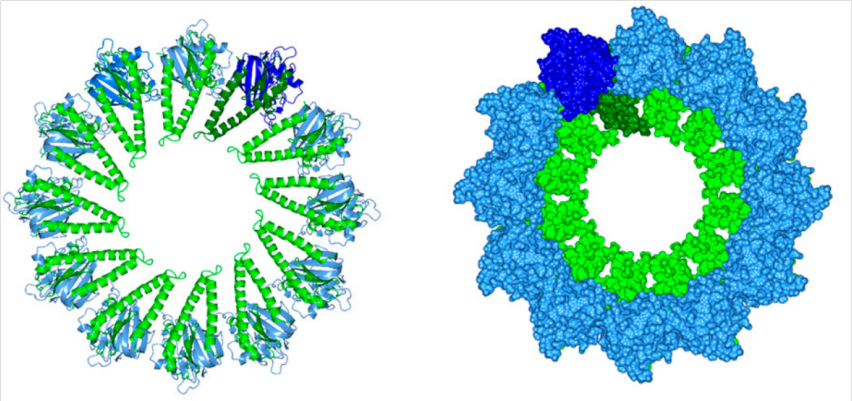


Figure 3.6.1 – SpoIIQ:SpoIIIAH channel assembly proposed by Levdikov et al. Based upon structural homology to the ring forming EscJ protein from E. coli, Levdikov et al., proposed an assembly of 12 SpoIIQ (blue), SpoIIIAH (green) dimers to form a channel assembly (Levdikov *et al.,* 2012).

3.6.2 The *C. difficile* SpoIIQ:SpoIIIAH interaction is dependent on Zn^{2+}.

The NMR, SEC-MALLS and MST experiments presented in this chapter show that the inter-sporangial domains of SpoIIQ and SpoIIIAH can form a 1:1 complex. The interaction is, however, only stable when sQ contains Zn^{2+}, suggesting that the metal ion co-ordinating site in the LytM domain of SpoIIQ has an important structural role. In the crystal structures of *B. subtilis* SpoIIQ:SpoIIIAH complex, β-strands 4 and 5 form an anti-parallel interaction in SpoIIQ with β-strands 1, 2 and 3 of SpoIIIAH (Figure 3.6.2) (Meisner *et al.,* 2012; Levdikov *et al.,* 2012). The degenerate metal co-ordination site motif 1 (SxxxD) is located between β-strands 5 and 6 (Figure 3.6.2) and is in close proximity to the interaction site with SpoIIIAH.

C. difficile SpoIIQ is predicted to have fewer β-strands (PSIPRED) than observed (PDB: 3TUF) in *B. subtilis* SpoIIQ (Figure 3.6.2A). Whereas *B. subtilis* SpoIIQ has an α-helix and two β-strands proximal to motif 1 that form the interaction interface with SpoIIIAH, *C. difficile* SpoIIQ is not predicted to have any such secondary structure elements in this region, formed of disordered backbone (Figure 3.6.2A, Table 3.3.1). It is possible that the docking of Zn^{2+} in the metal co-ordination site induces formation of secondary structure in

SpoIIQ or restricts the movement of the disordered region, which results in the formation of a competent interface with SpoIIIAH. Conformational changes could be analysed by use of HSQC-NMR of apo-sQ and a titration of $ZnCl_2$. Such an experiment would be most useful if the peptide backbone was assigned so that residues that undergo conformational changes could be identified.

The sQ:sAH complex has a K_D between 40-80 μM as observed by MST. This is not a strong protein:protein interaction, however, a tight binding interaction between SpoIIQ and SpoIIIAH may not be necessary *in vivo*. Both of these proteins are membrane bound and their respective membranes are in close proximity during engulfment of the forespore. This reduces the number of dimensions along which SpoIIQ and SpoIIIAH can diffuse to two. In *B. subtilis*, the intersporangial domains of SpoIIQ and SpoIIIAH were determined to have a K_D of 1 μM by ITC, 40-fold tighter than observed between *C. difficile* sQ and sAH (Levdikov *et al.,* 2012) which was measured using MST. To compare the binding affinities between the *C. difficile* and *B. subtilis* complexes, MST experiments using the *B. subtilis* proteins should be performed as a control. The *C. difficile* complex may be weaker due to the reliance of a metal ion being bound to SpoIIQ for stable complex formation, a variable that does not affect the *B. subtilis* complex. It was not possible to determine the affinity between the Zn^{2+} deficient mutant sQ^{H120S} and sAH.

In vivo studies support the *in vitro* data presented here that Zn^{2+} is required for stable complex formation. It has been shown by our collaborators in Professor Henriques' lab, using fluorescently labelled (via a SNAP tag) $SpoIIQ^{H120S}$ and SpoIIIAH, that in the absence of Zn^{2+} binding capability, a less stable complex was formed (Serrano *et al.,* 2015). However, the $SpoIIQ^{H120S}$ mutant could progress further in engulfment and did not result in membrane collapse or bulging as observed in *spoIIQ* mutants but the sporulation efficiency was lower than the wildtype (Serrano *et al.,* 2015).

In *B. subtilis* the LytM domain is not likely to have catalytic function due to the absence of metal co-ordinating residues and no metal ion is observed in the crystal structures. Conversely, *C. difficile* SpoIIQ, requires metal binding for stable complex formation with SpoIIIAH. The secondary structure prediction of this region in *C. difficile* lacks elements observed in the *B. subtilis* crystal structure models that are required for interaction with SpoIIIAH. Conservation of the secondary structure between that observed in the *B. subtilis*

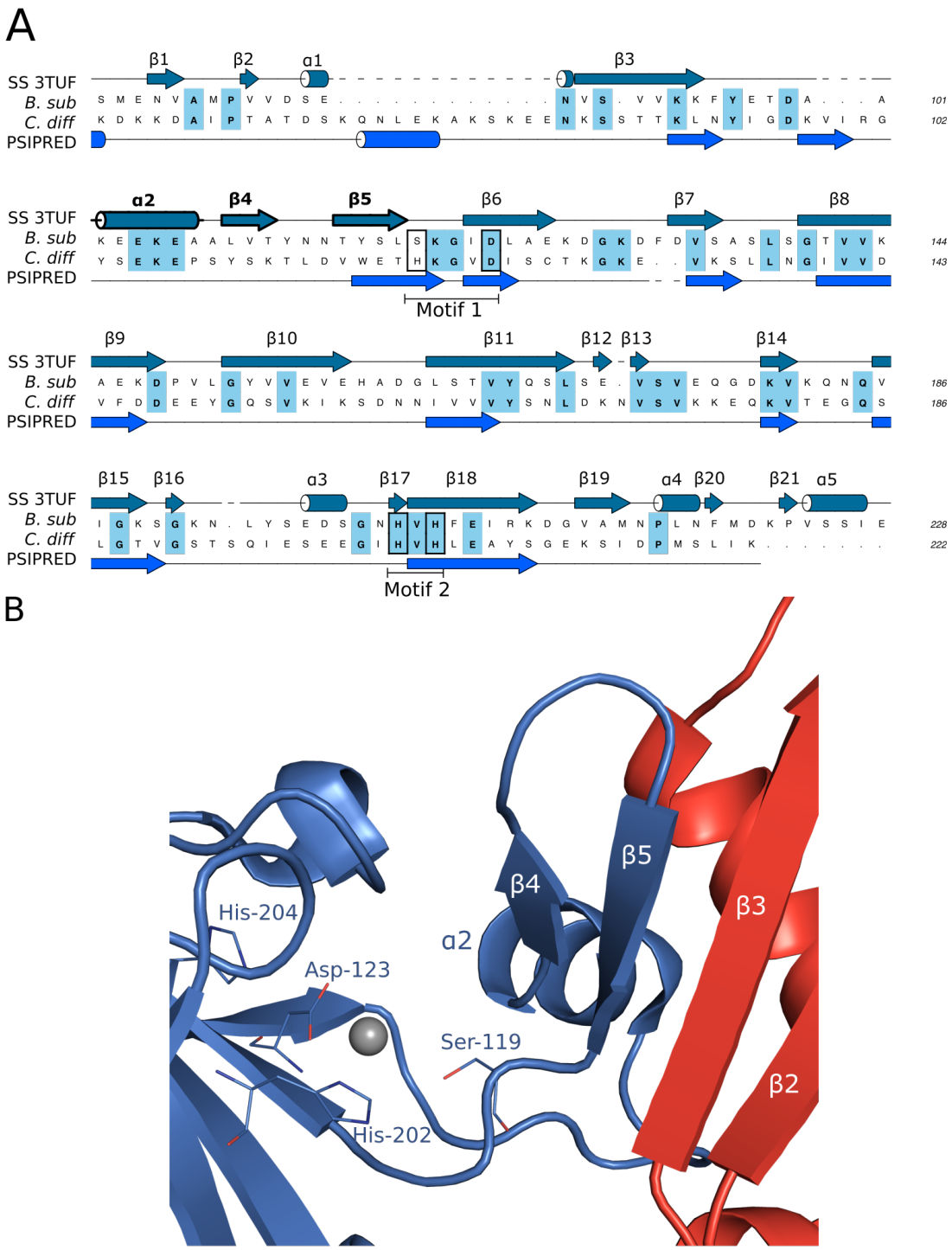


Figure 3.6.2 – Comparison of the LytM domain and SpoIIQ:SpoIIIAH interaction site A: Alignment of the LytM domain of *B. subtilis* SpoIIQ$_{43\text{-}283}$ and *C. difficile* SpoIIQ (30% sequence identity). The secondary structure in the crystal structures of *B. subtilis* SpoIIQ (teal) as analysed by DSSP (Kabsch and Sander, 1983) and PSIPRED secondary structure prediction of *C. difficile* SpoIIQ is also shown (blue). Secondary structure that forms the SpoIIIAH interaction is labelled and outlined in bold. The metal co-ordinating residues of motif 1 and motif 2 are highlighted in black boxes. **B**: Close up view of the SpoIIQ (blue), SpoIIIAH (red) β-strand interaction in *B. subtilis* (PDB: 3TUF) including the degenerate Zn^{2+} co-ordination residues (represented by sticks) in *B. subtilis* SpoIIQ with the Zn^{2+} superimposed the *S. aureus* LytM structure (PDB ID: 4ZYB).

crystal structures and the predicted secondary structure of *C. difficile* SpoIIIAH suggests that the interface for the interaction with SpoIIQ is structurally conserved (Figure 3.6.4). Such similarity between the *C. difficile* and *B. subtilis* SpoIIIAH is supported by the fact it was possible to form a chimeric complex of *B. subtilis* SpoIIQ with *C. difficile* SpoIIIAH (Figure 3.5.2). However, the inability to form a complex between *C. difficile* SpoIIQ and *B. subtilis* SpoIIIAH suggests that there are sufficient differences between the *C. difficile* SpoIIQ and *B. subtilis* SpoIIIAH interfaces that prevents this combination of chimeric complex formation (Figure 3.6.3). Therefore, it is likely that *C. difficile* SpoIIQ and SpoIIIAH interact in a similar manner to the *B. subtilis* homologues, albeit the *C. difficile* complex requires Zn^{2+} to be present within the LytM domain of SpoIIQ to allow the formation of the correct structure to interact with SpoIIIAH. Such a phenomena indicates that there is not complete structural conservation of the interacting regions between the *C. difficile* and *B. subtilis* complexes.

In the *B. subtilis* SpoIIQ structures, the region between β3 and β6 (where the Zn^{2+} site would be) forms the interaction with SpoIIIAH, and would be likely to block any possible substrate access by peptidoglycan to the Zn^{2+} site. Therefore the conserved motifs in the LytM domain of *C. difficile* are required to co-ordinate Zn^{2+} and form the necessary interface to interact with SpoIIIAH and in doing so block possible access to the bound Zn^{2+}. To confirm this hypothesis, the atomic structure of SpoIIQ and SpoIIIAH from *C. difficile* must be determined. It is, however, possible that conformational changes could occur to enable peptidoglycan binding SpoIIQ when it is not in complex with SpoIIIAH.

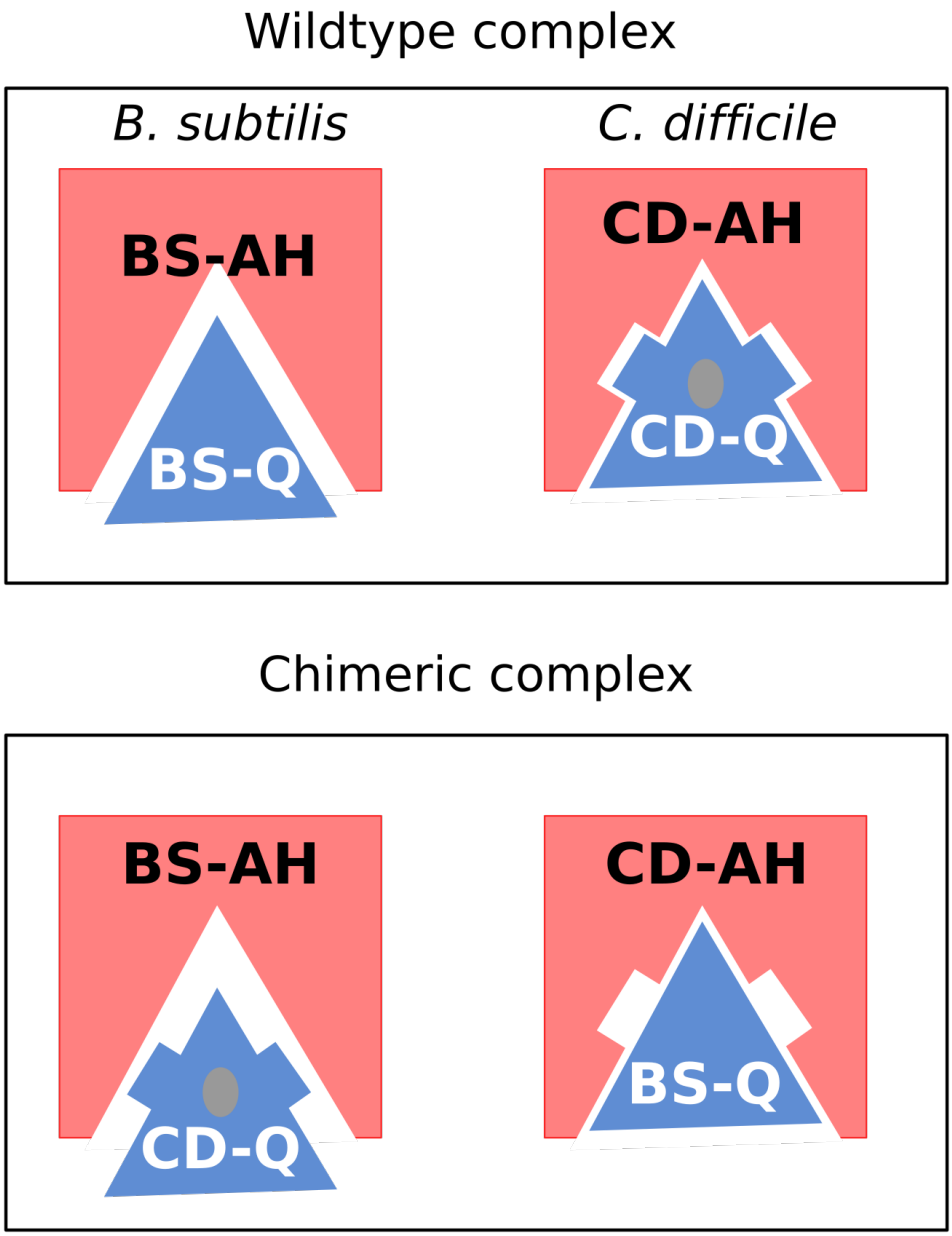


Figure 3.6.3 – Interface schemes for chimeric SpoIIQ:SpoIIIAH complexes. *B. subtilis* (BS) SpoIIQ (blue) can form a chimeric complex with *C. difficile* (CD) SpoIIIAH (red) but not vice versa suggesting that while the interfaces of *B.* subtilis and *C. difficile* SpoIIIAH are similar, the interaction interfaces of SpoIIQ differ in such a way that the *C. difficile* SpoIIQ (which binds $Zn^{2+,}$ grey circle) cannot form a complex with *B. subtilis* SpoIIIAH.

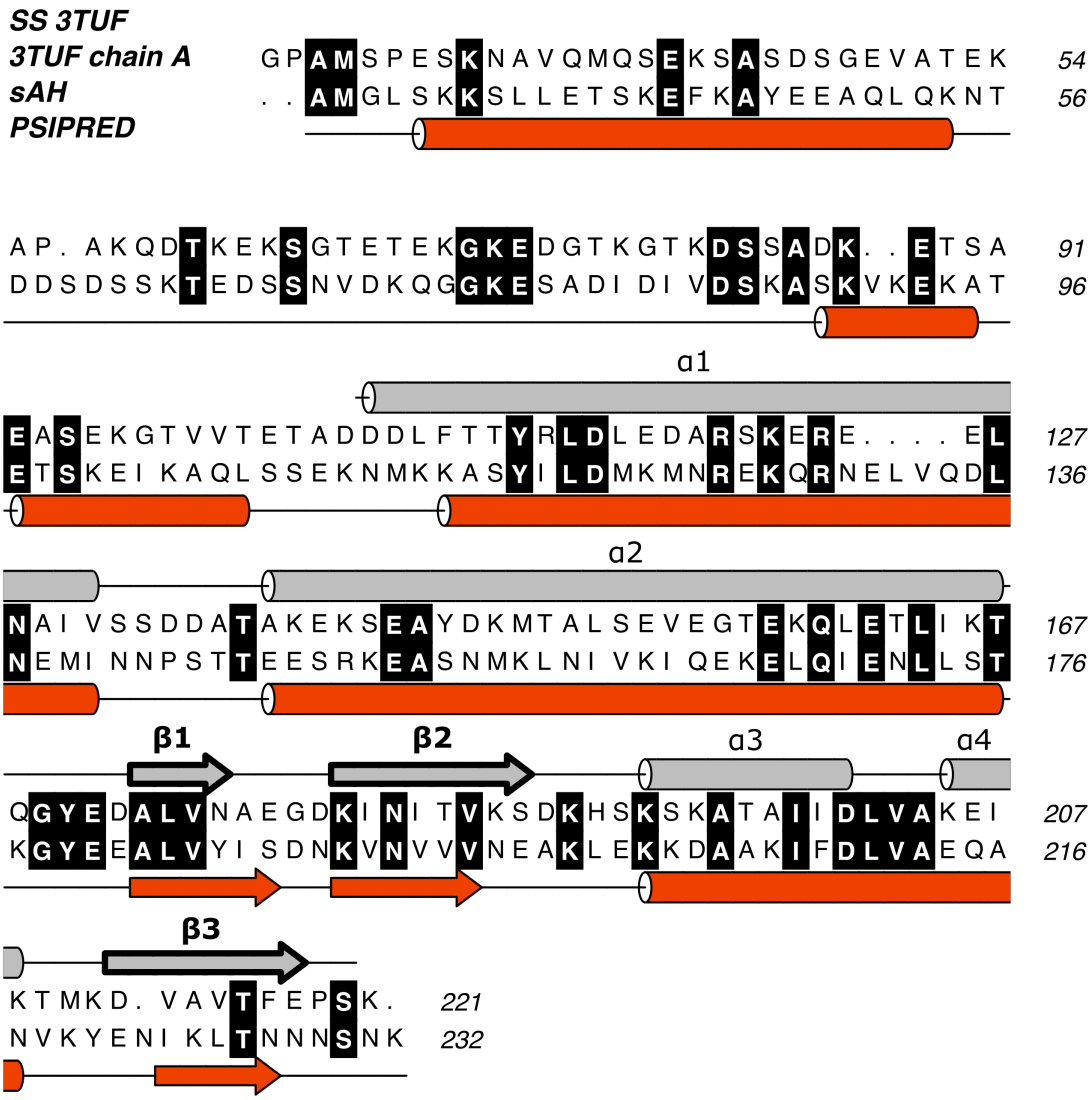


Figure 3.6.4 – Clustal Omega alignment and secondary structure of sAH and *B. subtilis* SpoIIIAH$_{25\text{-}218}$. Secondary structure from 3TUF (chain A) is shown in grey and PSIPRED secondary structure prediction for sAH is shown in red. The β-strands labelled and outlined in bold form the interaction with SpoIIQ in the *B. subtilis* crystal structure 3TUF, and are predicted to be conserved in *C. difficile* SpoIIIAH. *B. subtilis* and *C. difficile* SpoIIIAH have 25% sequence identity.

3.6.3 SpoIIQ has a complete metal binding LytM domain

The inter-sporangial domain of *C. difficile* SpoIIQ has 38% sequence identity (HHPred) with the M23 metallopeptidase family which includes Gly-Gly endopeptidases such as the LytM endopeptidase (Firczuk and Bochtler, 2007; Meisner and Moran, 2011). *B. subtilis* SpoIIQ shares 47% sequence identity (HHPred) with the M23 metallopeptidase family. A signature of LytM domains are two motifs, HxxxD and HxH, which coordinate a metal ion to generate a nucleophile required for the breakage of glycine-glycine peptide cross-bridges within peptidoglycan (Firczuk and Bochtler, 2007; Grabowska *et al.,* 2015). However, even though Gly is present in the peptidoglycan of *B. subtilis* and *C. difficile*, Gly-Gly bonds have not been observed in either species (Peltier *et al.,* 2011). The LytM motifs are conserved in *C. difficile* SpoIIQ and many other Clostridial SpoIIQ genes (Crawshaw *et al.,* 2014) are able to co-ordinate Zn^{2+} unlike *Bacilli* which contain a degenerate motif (SxxxD). Several crystal structures of LytM proteins exist in the PDB, including 4ZYB (orange, Figure 3.6.5), which contain a Zn^{2+}. In 4ZYB, the Zn^{2+} is also co-ordinated by a tetraglycine phosphinate transition state analogue (Grabowska *et al.,* 2015).

Comparison of the surface of SpoIIQ in 3TUF and the peptidoglycan analogue bound LytM in 4ZYB shows that it would not be possible for such a tetraglycine phosphinate ligand to bind *B. subtilis* SpoIIQ without major structural rearrangements (Figure 3.6.6). The potential peptidoglycan activity of sQ was not tested in this thesis due to the unavailability of *C. difficile* peptidoglycan or a suitable transition state analogue. Comparison with the *B. subtilis* SpoIIQ crystal structures and *S. aureus* LytM suggests that *C. difficile* SpoIIQ would not be able to bind a tetra-glycine fragment due to an additional α-helix and β-strand that forms part of the interaction with SpoIIIAH. Therefore, SpoIIQ is not likely to have a catalytic role in the processing of peptidoglycan when in complex with SpoIIIAH.

It is not clear whether the LytM fold of *C. difficile*, complete with a Zn^{2+} ion, would have endopeptidase activity. *In vivo*, SpoIIQ, in complex with SpoIIIAH, is localised to the forespore engulfment septum where peptidoglycan must be processed to enable complete and proper engulfment (Serrano *et al.,* 2015). In *spoIIQ* knockout mutants, membrane bulging and incomplete engulfment are observed in *C. difficile*, suggesting that SpoIIQ may have a direct role in peptidoglycan processing during engulfment (Serrano *et al.,* 2015). However, disruption of the metal binding site in spoIIQ^{H120S} mutants *in vivo* does

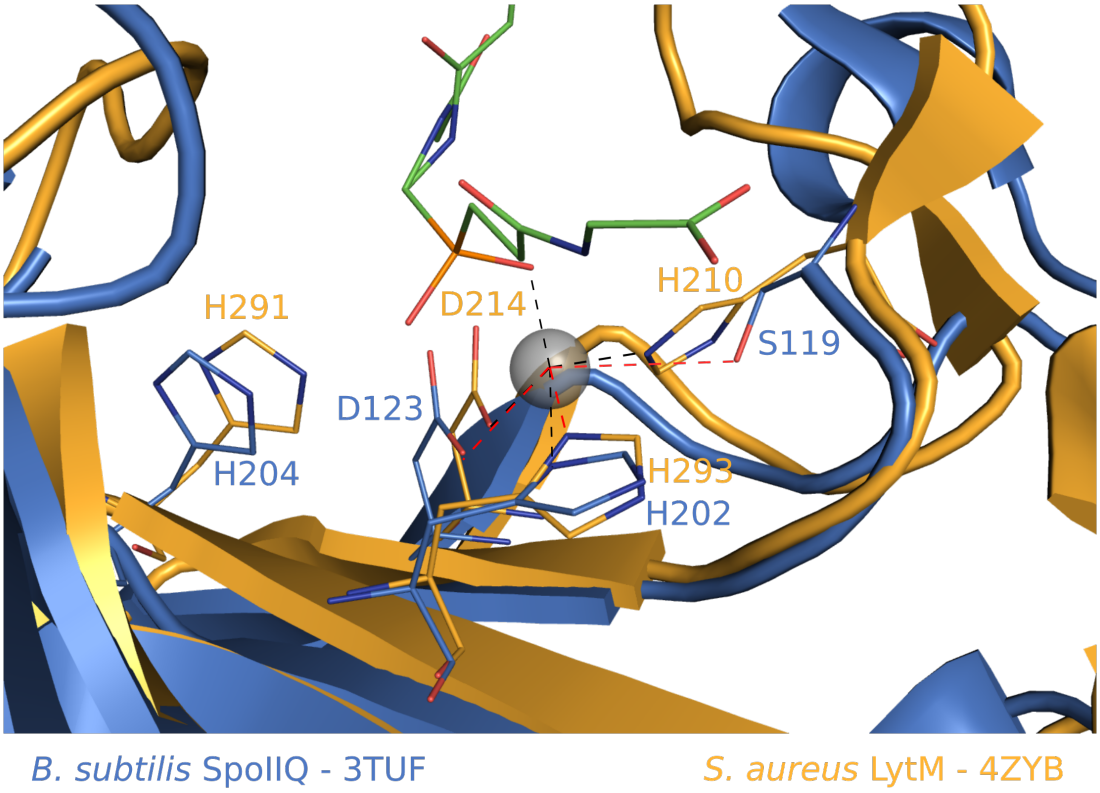


Figure 3.6.5 – Superimposition of *B. subtilis* SpoIIQ and *S. aureus* LytM. Secondary structure superimposition (Coot SSM (Emsley *et al.,* 2010)) of *B. subtilis* SpoIIQ (blue, PDB:3TUF) (Levdikov *et al.,* 2012) and *S. aureus* LytM (gold, PDB:4ZYB) crystal structures (core RMSD: 1.8 Å). The *S. aureus* LytM has 27% sequence identity with *B. subtilis* SpoIIQ and 24% identity with *C. difficile* SpoIIQ. The crystal structure of the LytM in 4ZYB has tetraglycine phosphinate bound (transition state analogue) and contains a Zn^{2+} ion co-ordinated by the tetraglycine fragment and the HxxxD and HxH metal co-ordinating motifs (displayed as sticks, black dotted lines), with each residue ~2 Å from the centre of the metal ion (co-ordination distances are summarised in Table 3.6.1). LytM motif 1 of 3TUF is SxxxD and the crystal structure does not contain metal ion at this site. Residues His204 and Asp123 are in approximately equivalent positions to His293 and Asp214 in 4ZYB but Ser119 is 3.9 Å away (red dotted lines) from the Zn^{2+} ion position, which is not close enough to co-ordinate the metal ion. *B. subtilis* SpoIIQ distances are represented in black and *S. aureus* LytM distances are shown in red dashed lines.

Residue (*S. aureus*/ *B. subtilis*)	*S. aureus* LytM	*B. subtilis* SpoIIQ
His210/Ser119	2.1 Å	3.8 Å
Asp123/Asp214	1.9 Å	2.3 Å
His293/His202	2.3 Å	1.9 Å

Table 3.6.1 – Zn^{2+} co-ordination distances in *S. aureus* LytM and superimposed on *B. subtilis* SpoIIQ. Co-ordination distances between the Zn^{2+} metal ion and co-ordinating residues His210, Asp123 and His293 in *S. aureus* LytM (Figure 3.6.5, gold) and the equivalent residues in *B. subtilis* SpoIIQ (Ser119, Asp214 and His202) after SSM superimposition (Figure 3.6.5, blue).

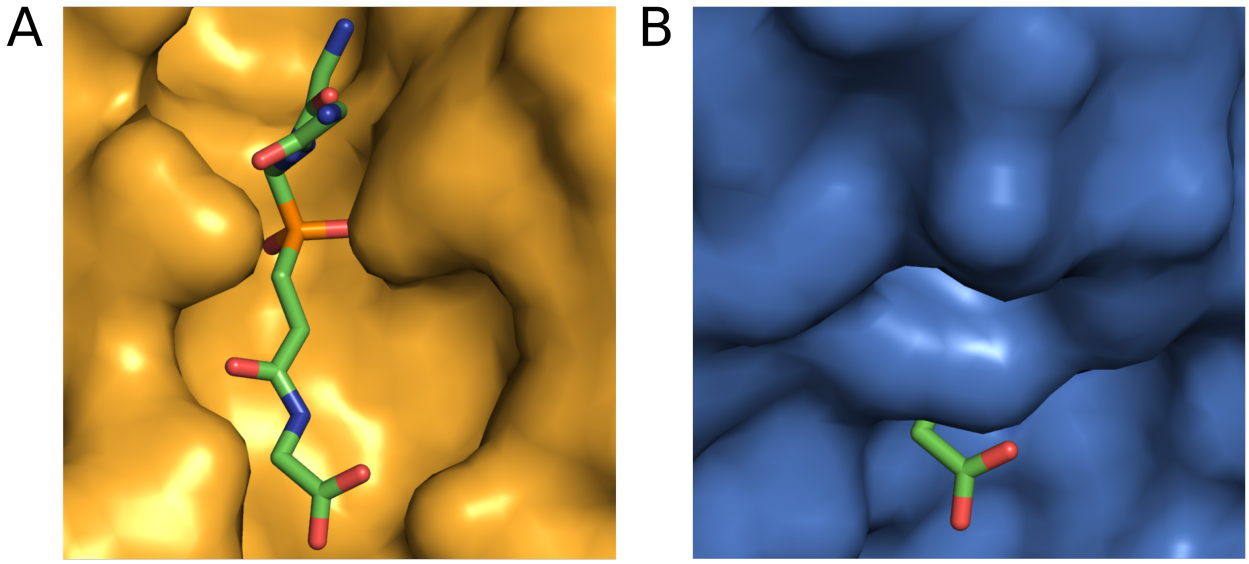


Figure 3.6.6 – Comparison of 4ZYB and 3TUF surfaces and peptidoglycan position. Surface representation of 4ZYB (A) and 3TUF (B), zoomed in on the peptidoglycan fragment in 4ZYB and the equivalent position in 3TUF. The peptidoglycan fragment in 4ZYB (stick representation) has been superimposed (via secondary structure superimposition of the protein models with a core RMSD of 1.8 Å) on to 3TUF from 4ZYB (B). Even though there is a small pocket close to the degenerate Zn^{2+} co-ordination site in 3TUF there is not sufficient space for a peptidoglycan fragment, as a loop containing 2 anti-parallel β-strands blocks access to the site, as opposed to the open groove in 4ZYB (A). This loop is involved in the interaction with SpoIIIAH.

not result in such membrane bulging (Serrano *et al.,* 2015). This suggests that SpoIIQ may not be directly involved in peptidoglycan cross-bridge scission but may play a role in the recruitment of other proteins that are involved in peptidoglycan processing. In both *C. difficile* and *B. subtilis* there is another complex containing peptidoglycan processing proteins, SpoIID and SpoIIP, that are localised to the engulfment septum (Aung *et al.,* 2007; Serrano *et al.,* 2015). SpoIID has sequence homology with the LytB family (HHPred, 62% ID) of peptidoglycan degrading proteins and SpoIIP has sequence homology to an N-acetylmuramoyl-L-alanine amidase (HHPred, 23%). In *B. subtilis, spoIID* mutants result in a membrane bulging phenotype, consistent with excess peptidoglycan that has not been processed (Rodrigues *et al.,* 2013). SpoIIQ, along with another protein SpoIIM, may play a role in their recruitment and localisation (Rodrigues *et al.,* 2013).

3.6.4 Conclusions and future work

The characterisation of the SpoIIQ:SpoIIIAH complex has added to the understanding of sporulation in *C. difficile*. Unlike in *B. subtilis*, the SpoIIQ:SpoIIIAH complex requires the LytM domain of SpoIIQ to co-ordinate a Zn^{2+} ion in order to form a stable complex with SpoIIIAH *in vitro*. In collaboration with Professor Adriano Henriques Group at ITQB, Lisbon, Portugal, it was established that both SpoIIQ and SpoIIIAH were required for complete engulfment of the forespore by the mother cell prior to spore maturation. Although, it has been shown that a $SpoIIQ^{H120S}$ mutant still formed a complex *in vivo* with SpoIIIAH this mutant could also not fully progress through engulfment of the forespore (Serrano *et al.,* 2015). As discussed in Section 3.6.3, it is unlikely that SpoIIQ is directly involved in peptidoglycan degradation. However, further experiments are required to fully characterise the peptidoglycan of *C. difficile* throughout its life cycle from vegetative cell to spore and to test possible substrates for sQ and determine whether the protein has enzymatic activity. The LytM domain of SpoIIQ may have had an arcane function during sporulation. However, as the bacteria have diverged, the function involving Zn^{2+} has become obsolete and in the *Bacilli* the residues required for co-ordination of the ion have evolved (Crawshaw *et al.,* 2014). For the moment, the exact role of the Zn^{2+} in the SpoIIQ:SpoIIIAH complex is undetermined.

Further *in vitro* work is required on the SpoIIQ:SpoIIIAH complex, studying the full-

length proteins to investigate channel formation, function and determine the structure and overall assembly of the complex. As the complex is understood to form a channel between the membranes of the forespore and the mother cell a number of questions are raised: what substrate may pass through the channel and how is the channel regulated? SpoIIIAH is expressed from the *spoIIIA* operon that contains 8 genes AA-AH, which includes ATPases and secretion proteins that resemble proteins from the Type-I, -II and -IV secretion systems, which may be involved in substrate transport across a channel (Crawshaw *et al.,* 2014).

In *B. subtilis* SpoIIQ can be recruited to the forespore interface with the mother cell by the SpoIID, SpoIIM and SpoIIP (DMP) complex. The relationship between SpoIIQ:SpoIIIAH and DMP complexes in *C. difficile* is yet to be studied fully. *In vivo*, the phenotype of double mutants of *spoIID-spoIIP* in *B. subtilis* resembles the phenotype observed in *spoIIQ* and *spoIIIAH* mutants in *C. difficile*, yet the *spoIIIAH* mutant in *B. subtilis* does not affect sporulation (Serrano *et al.,* 2015). In *C. difficile*, it appears that SpoIIIAH, in addition to SpoIIQ, is required for the recruitment of the DMP complex for peptidoglycan processing. *In vitro* characterisation of the soluble domains of the DMP complex should be carried out and potential interactions with the SpoIIQ:SpoIIIAH investigated.

The function of the SpoIIQ:SpoIIIAH complex in *C. difficile* is still unknown. However, the work presented in this thesis provides basic biophysical and biochemical information on the complex upon which further studies can be based. Understanding the role of the complex in protein recruitment at the mother cell to forespore interface, peptidoglycan processing during engulfment and the substrate of the resulting channel are important questions that must be addressed.

Understanding basic mechanisms of sporulation in an important pathogen may have useful implications. *C. difficile* is unable to persist in the aerobic environment in a vegetative state and it is essential that it can form spores. *C. difficile* spores are recognised as the infectious agent and enable *C. difficile* infection to spread between individuals. The majority of the knowledge on sporulation is based upon work on the *Bacillus* genus which includes some important pathogens such as *B. anthracis*. *Clostridia* represent a large proportion of the Gram-positive bacteria, including a number of important pathogens like *C. difficile*, *C. perfringens* and *C. botulinum*, and it is only recently that studies on the sporu-

lation mechanisms of these bacteria have come to light. While many of the genes involved in controlling the sporulation mechanism are conserved between *Clostridia* and *Bacilli*, it is increasingly clear that the pathways and mechanisms that they form are not strictly conserved and the understanding of these mechanisms is only now being developed (Fimlaid *et al.,* 2013; Saujet *et al.,* 2014; Fimlaid and Shen, 2015).

Chapter 4

Structure of major Type IV Pilins from *Clostridium difficile*

4.1 Introduction

Type IV pilins (TFP) are fibre-like appendages that extend from the cell into the external media and are involved in cell-to-surface adhesion and the formation of biofilms (Maier and Wong, 2015; Giltner *et al.,* 2012). These fibres can be extended and retracted in a manner that allows the cell to move across a surface (Varga *et al.,* 2006). The filaments are formed of protein units known as pilin proteins that are divided into two classes, major and minor pilins (Melville and Craig, 2013). Major pilins are the predominant pilin unit in the fibre, and in *Clostridium difficile* are known as *pilA* (Maldarelli *et al.,* 2014). There are two TFP loci in genomes of *C. difficile*, the main locus is the largest and most complete containing 11 TFP components whilst a secondary locus contains only a major pilin gene (*pilA2*) and three membrane associated components (Melville and Craig, 2013; Maldarelli *et al.,* 2014).

Pilin proteins can be identified by a highly conserved signal peptide sequence (Figure 4.1.1A) (Imam *et al.,* 2011), that is cleaved by the pre-pilin peptidase PilD before insertion into the pilin filament (Melville and Craig, 2013). The signal peptide is used to define the type of TFP, type-a pilins (TFPa) have a short signal sequence (5-7 residues) and the first residue of the mature pilin is a Phe. The Gram-positive TFP pilins that have been identified so far have TFPa (Melville and Craig, 2013). Type-b (TFPb) have a highly hydrophobic

sequence of ~30 residues and the first residue is a hydrophobic residue other than Phe.

There are 20 protein crystal structures of pilin proteins in the PDB, 50% of which are of pilins from *Pseudomonas aeruginosa*. The structures of pilins from *P. aeruginosa*, *Neisseria gonorrhoeae* and *N. meningitidis* exhibit a long N-terminal α-helix (α-1) that forms a polymerisation stalk domain and a globular C-terminal domain described as the headgroup, that consists of an α-β region linking α-1 with a β-sheet of at least three strands. (Figure 4.1.1B)(Craig and Li, 2008). A highly variable region is located at the C-terminus of the headgroup, known as the D-region, which is predicted to be the most solvent accessible element of the headgroup and can form interactions with host-cell surfaces (Craig and Li, 2008). The D-region is also delimited a disulphide bridge in the TFP pilin structures so far solved in Gram-negative bacteria (Craig and Li, 2008).

The aim of this project was to determine the crystal structures of the major pilins from the main locus, PilA1. To increase the likelihood of determining a crystal structure, PilA1 from two strains were pursued, the hyper-virulent R20291 and the better studied lab strain 630. These two PilA1 proteins are 91% identical and the sequence divergence is limited to residues 139-170 at the C-terminal ends of the proteins (Figure 4.2.1). The C-terminal region of the pilin units is solvent exposed and in the TFP of other pathogenic species, such as *N. gonorrhoea* and *N. meningitidis,* sequence variability has been observed in this region across strains as a means of host-immune evasion (Takahashi *et al.,* 2012). Sequence variability can result in structural variability and may be an important virulence mechanism in host cell adhesion during colonisation in an anaerobic environment such as the gut .

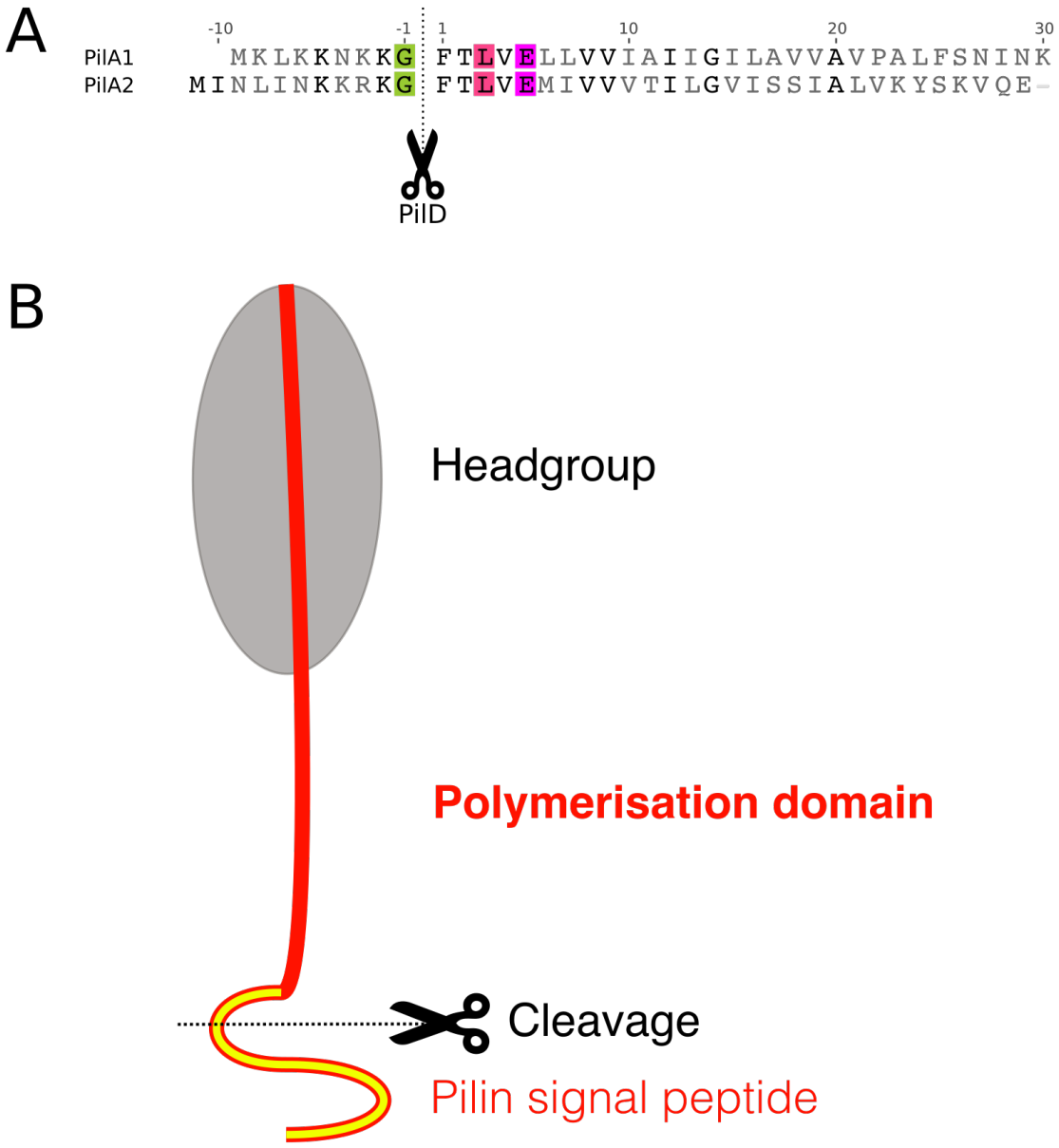


Figure 4.1.1 – **Pilin organisation and signal peptide. A:** The signal peptide of type-a Type IV pili (TFPa) includes a conserved Gly-1 which is cleaved at the carboxyl end by the pre-pilin peptidase PilD. Leu3 and Glu5 are also highly conserved in the TFPa signal peptide. The N-terminal residue of mature TFPa pilin proteins is a conserved Phe residue. **B:** General schematic of a type-a type IV pilin protein (TFPa) with N-terminal signal peptide (filled yellow), α-helical polymerisation domain stem (red) and C-terminal globular head-group (grey).

4.2 Purification of the Major Type IV pilin: PilA1

The 34 residues at the N-terminus of PilA1 are largely hydrophobic and submission of the full length peptide sequences of PilA1 from R20291 and 630 to transmembrane prediction servers such as TMHMM indicated that these residues formed a hydrophobic transmembrane domain (Möller *et al.,* 2001). It has been observed that long hydrophobic N-terminal helices proceeding a conserved signal peptide, which resemble transmembrane helices, formed the TFPa pilin polymerisation domain (Craig and Li, 2008). To ensure protein solubility, these N-terminal helices were not included in constructs of PilA1. The R20291 PilA1$\triangle$1-34 construct was produced by Edward Couchman (Fairweather group, Imperial College) by amplification of CD3355 from R20291 genomic DNA and inserted into pET-28a such that when expressed there was a non-cleavable N-terminal 6-His-tag (Figure 4.2.1). The 630 PilA1$\triangle$1-34 (CD3513) was amplified from a plasmid containing the full length gene (pRPF227, produced by Dr Robert Fagan, Fairweather group) and inserted into pET-28a in the same orientation as the PilA1$\triangle$1-34 R20291 construct (Figure 4.2.1).

The R20291 PilA1$\triangle$1-34 was expressed in *Escherichia coli* cells (Section 2.5.5) and purified by nickel affinity purification followed by size exclusion chromatography (SEC) (Figure 4.2.2A/B, Section 2.5). A total yield of ~8 mg/L cell culture was achieved and the protein was stable and contained few impurities. The 630 PilA1$\triangle$1-34 was purified in an identical manner to the R20291 protein and a total yield of 16 mg/L cell culture was achieved (Figure 4.2.2C/D). The R20291 PilA1$\triangle$1-34 protein eluted from the SEC column as a single peak at a volume equivalent to that of a 20.4 kDa protein and the 630 PilA1$\triangle$1-34 eluted in a single peak at a volume equivalent to a mass of 15.4 kDa. The R20291 and 630 constructs had a theoretical mass of 15.9 kDa and 15.5 kDa respectively. Comparison of the elution masses from SEC and the theoretical masses of these proteins indicated that both were purified in a monomeric state.

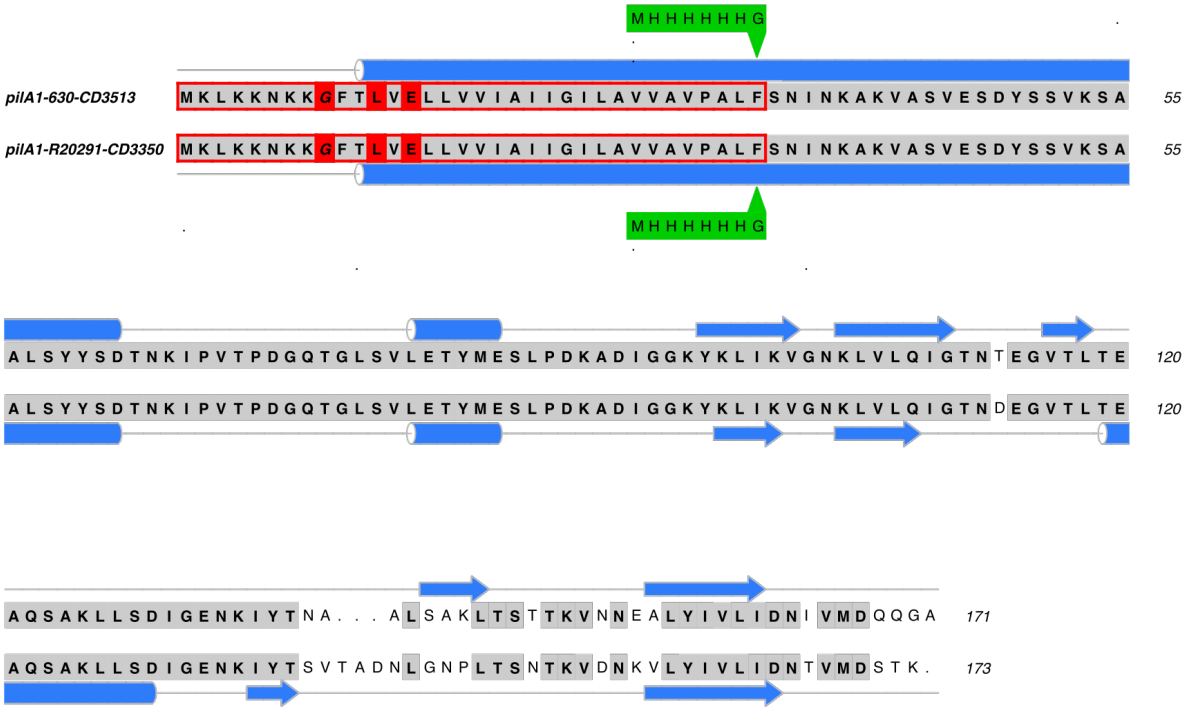


Figure 4.2.1 – PilA1 alignment and construct design. Clustal Omega alignment of PilA1 from strains 630 (CD3513) and R20291 (CD3350) which share 91% sequence identity (conserved residues are highlighted in grey). The polymerisation domains are outlined in red. The conserved recognition residues are highlighted in red boxes and the cleaved glycine is italicised. The position of the His-tag (green) of both PilA1Δ1-34 constructs is indicated by the green arrow. The peptide sequence upstream of this point, which is part of the polymiersation domain was not included in the PilA1Δ1-34 constructs. PSIPRED secondary structure predictions have been added above (630) and below (R20291) the respective sequence. The masses calculated using ProtParam (Wilkins *et al.,* 1999) of the 630 protein was 15.5 kDa and the R20291 protein was 15.9 kDa.

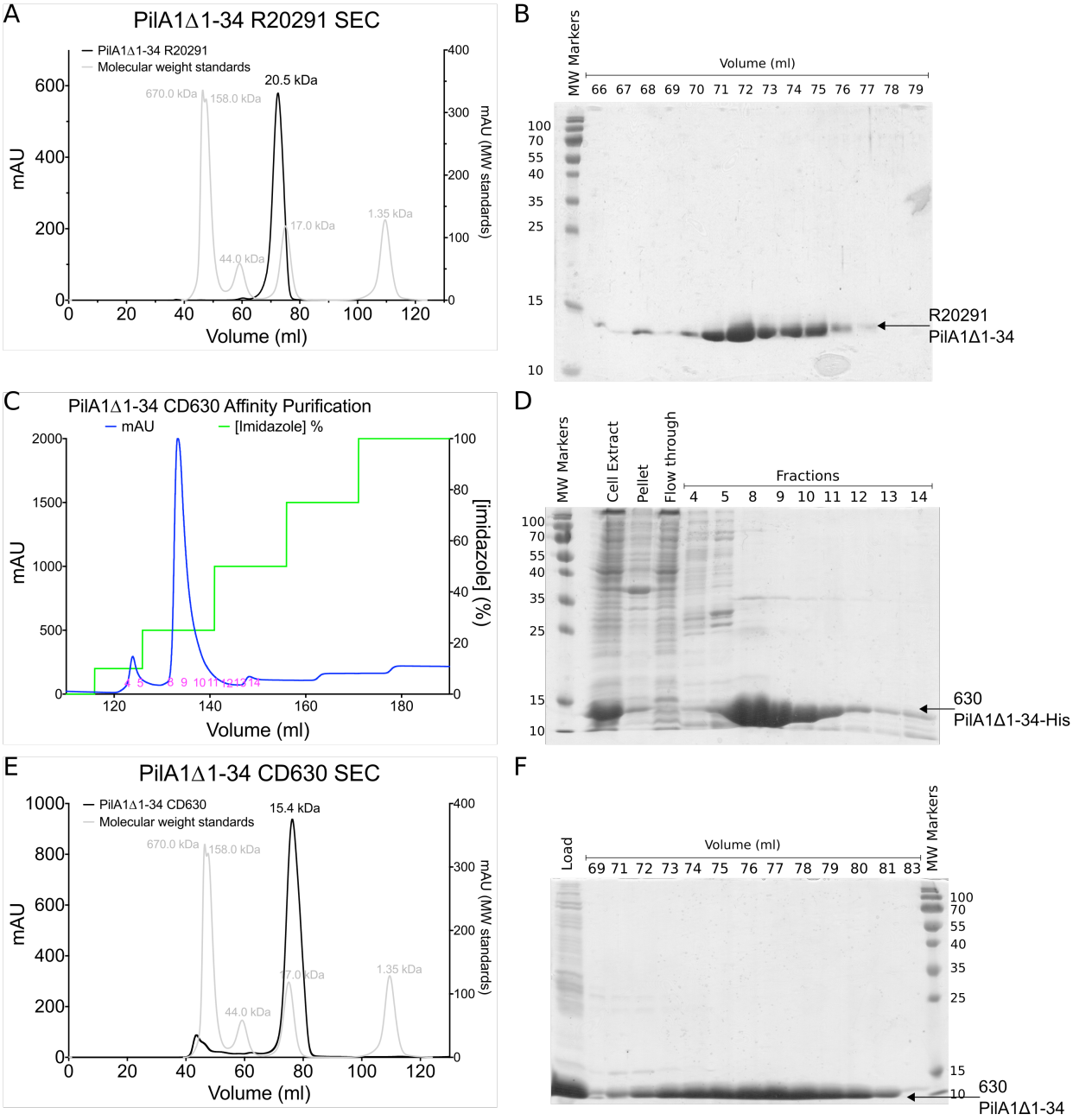


Figure 4.2.2 – **Purification of PilA1Δ1-34 constructs. A:** Size-exclusion chromatography (SEC) of R20291 PilA1Δ1-34. The R20291 protein eluted at a volume equivalent to 20 kDa when compared with elution profile of molecular mass (MW) markers (grey). The molecular mass standards were thyroglobulin (670 .0kDa), γ-globulin (158.0 kDa), ovalbumin (44.0 kDa), myoglobin (17.0 kDa) and Vitamin B_{12} (1.4 kDa). **B:** SDS-PAGE analysis showing peak fractions of R20291 PilA1Δ1-34. **C:** Chromatogram of the nickel affinity purification of 630 PilA1Δ1-34: the UV absorbance (blue), concentration of imidazole (% v/v, green)**.** Two elution peaks were observed at 60 mM and 125 mM imidazole. Fractions that were analysed by SDS-PAGE are labelled in pink. **D:** SDS-PAGE analysis of the cell extract, pellet, flow through and peak fractions from the nickel affinity purification of 630 PilA1Δ1-34. The majority of the 630 PilA1Δ1-34 protein eluted in 125 mM imidazole (fractions 8-12), with a small amount of protein eluting at 60 mM imidazole. **E:** SEC of 630 PilA1Δ1-34. A single elution peak at a volume equivalent to 15.5 kDa was observed when compared with elution profile of MW markers (grey). **F:** Peak fractions from the SEC of 630 PilA1Δ1-34 were analysed by SDS-PAGE, and showed that these proteins were stable and pure.

4.3 Crystal structures of major Type IV pilins from *C. difficile*

4.3.1 Structure determination of R20291 PilA1△1-34

4.3.1.1 Crystallisation

Crystallisation trials of R20291 PilA1△1-34 were set up as described in Section 2.7.1 at a concentration of 16 mg/ml. Crystals of R20291 PilA1△1-34 of approximately 150 x 150 μm formed within four days of dispensing in the Structure [Hampton] crystallisation condition H12, which contained 1.6 M sodium citrate at pH 6.5 (Figure 4.3.1A). Crystals formed in both 1:1 and 2:1 protein:crystallisation solution ratios, although larger crystals formed in the 2:1 drop ratio. The crystals could be reproduced reliably using protein that had been flash frozen in liquid nitrogen and stored at -80° C, however, crystals dissolved within 8 days of tray set up. 0.2 M Sodium citrate is a proven protein crystal cryo-protectant (Bujacz *et al.,* 2010) and so crystals were harvested and directly flash cooled in liquid nitrogen. Crystals were sent to the Diamond Light Source synchrotron, Harwell for diffraction experiments (Section 4.3.1.2).

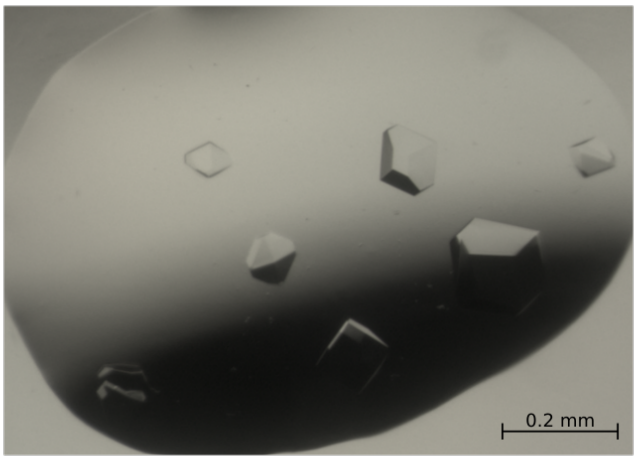


Figure 4.3.1 – **Crystals of PilA1△1-34 from R20291. A:** Crystals of pilA1△1-34 from R20291 crystallised in a 100 nl :100 nl drop with crystallisation solution Structure H12 [Hampton] (1.6 M sodium citrate, pH 6.5).

4.3.1.2 Data collection and processing from native crystals

Datasets of four independent native crystals were collected on the fixed-wavelength beamline I04-1 at Diamond Light Source by Dr Arnaud Baslé as outlined in Section 2.7.4. Test images of each crystal at 0° and 90° were recorded and the programme iMosflm (Battye *et al.,* 2011) was used to index the data and determine the spacegroup and unit cell parameters of the crystal. The strategy function of iMosflm was utilised to determine the optimum starting position of the crystal for a full native dataset collection. During the diffraction experiment the crystal was rotated 200° with an exposure time of 0.5s, each image recorded an oscillation of 0.1°. Such a strategy enabled the collection of a highly redundant native dataset with high completeness in order to simplify the structure solution pipeline.

Each of the four datasets were indexed and integrated with three widely used programs: XDS (Kabsch, 2010); iMosflm (Battye *et al.,* 2011); DIALS (Gildea *et al.,* 2014). The three programs use different approaches to indexing and integration, Mosflm treats the data in a two-dimensional fashion indexing each image individually (Battye *et al.,* 2011). XDS and DIALS treat the images in wedges, which enables a three-dimensional approach (Kabsch, 2010; Gildea *et al.,* 2014). The readout capabilities of modern area detectors such as the Pilatus detectors mean that each image contains reflections from very small oscillations (0.1°) of the crystal and as a result the same reflection is likely to be recorded over several images. The 3D approach enables single reflections recorded over many images to be treated as a single reflection. The reflections output by these programs were then reduced using Aimless in the CCP4 suite (Evans and Murshudov, 2013; Evans, 2006; Evans, 2011). For comparison, the dataset parameters, indexed using the three programs described and scaled in Aimless, are presented in Appendix B, Table B.0.1. Since each of the four crystals diffracted in the range of 2.2-1.6 Å, no crystallisation optimisation was performed.

No images were removed from the 2000 collected of each crystal during the indexing and integrating stage. The quality and resolution limits of the datasets were judged on four factors: R_{merge}; R_{meas}; I/σI; CC1/2 (Correlation coefficient) of the mean intensity. The internal quality of the data in each dataset is described by the R_{merge} and R_{meas}(Evans and Murshudov, 2013; Evans *et al.,* 2011). The signal-to-noise ratio (I/σI) of the recorded intensities is an important factor and in the outer resolution shell this can be as low as 1.5

for high resolution data that has been recorded using a Pilatus detector [Dectris] (Evans *et al.,* 2011). Another assessment of the quality of the intensities is the mean intensity CC1/2, which divides the data into two bins and a correlation co-efficient is calculated for the two bins. A CC1/2 close to 1.0 indicates that the intensities in the two bins agree and are therefore consistent throughout all images.

Since processing of all datasets resulted in similarly low R values, I/σI >1.5 and mean intensity CC1/2 of at least 0.996, the data that had been successfully processed to the highest resolution was chosen. Dataset 4, indexed and integrated using Mosflm, to a resolution of 1.65 Å and with 100% completeness was used as the 'native' data for all downstream processing (Table 4.3.1).

As part of the Aimless workflow, Pointless (Evans, 2006; Evans, 2011) was used to confirm that the correct point group had been chosen; the software indicated that all four datasets were $P4_12_12$ or $P4_32_12$. The spacegroup $P4_12_12$ is non-centrosymmetric meaning the centre of inversion is at the origin of the coordinate system and when inverted results in the change of the spacegroup to $P4_32_12$ (Shmueli, 2008); this is known as enantiomorphic ambiguity and the correct spacegroup will only be determined during structure solution. Resolution of the unit cell content ambiguity was attempted during molecular replacement (Section 4.3.1.4) but only resolved during experimental phasing (Section 4.3.1.5).

Data Collection statistics	
Resolution (Å)	51.21 - 1.65
Unit cell dimensions	
a=b, c (Å)	102.42, 104.21
α=β=γ (°)	90
Spacegroup	$P4_12_12$ or $P4_32_12$
R_{merge}*	0.136 (1.887)
Total Reflections	904124
Unique Reflections	67115 (3391)
I/σI	10.7 (1.5)
Mean intensity CC1/2	0.998 (0.323)
Completeness (%)	100.0 (100.0)

Table 4.3.1 – Table of dataset parameters for native R20291 PilA1Δ1-34 crystal 4. The reflections and intensities were then scaled and reduced using Aimless (Evans and Murshudov, 2013; Evans, 2006; Evans, 2011), the summary of the dataset parameters is presented above. These data were collected at a wavelength of 0.92 Å. Values in parentheses represent the highest resolution shell. *$R_{merge} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where e th observation of reflection and is the massed average intensity for all observations of reflection .

4.3.1.3 Unit cell content of native R20291 PilA1△1-34 crystals.

The unit cell contents were calculated using the Matthews program in CCP4 (Matthews, 1968; Kantardjieff and Rupp, 2003). The highest probability calculated was for the presence of four protein molecules (P(1.65): 0.78) in the asymmetric unit (ASU) (Table 4.3.2). The Matthews coefficient suggested that there were most likely either three copies (Matthews coef: 2.87 $Å^3$/Da) or four copies (Matthews coef: 2.15 $Å^3$/Da) in the ASU. Protein crystals are composed of between ~25-75% solvent (Matthews, 1968).

Nmol/asym	Matthews coef. ($Å^3$/Da)	solvent (%)	P(1.65)	P(tot)
1	8.60	85.71	0.00	0.00
2	4.30	71.41	0.00	0.02
3	2.87	57.12	0.20	0.35
4	2.15	42.82	0.78	0.62
5	1.72	28.53	0.02	0.01
6	1.43	14.23	0.00	0.00

Table 4.3.2 – Unit cell contents of a native PilA1△1-34 R20291 crystal. Table of unit cell contents properties by number of molecules in the asymmetric unit.

4.3.1.4 Molecular replacement

The peptide sequence of R20291 PilA1 was used to search the PDBe (www.ebi.ac.uk/pdbe) for potential molecular replacement search models. The search returned two PDB entries: 2HI2, a 2.3 Å resolution X-ray crystal structure of PilE1 from *Neisseria gonorrhoeae* (Craig *et al.,* 2006); and 2PIL, an X-ray crystal structure of the same PilE1 from *N. gonorrhoeae* at a resolution of 2.6 Å (Forest *et al.,* 1999). R20291 PilA1 has 39% sequence identity with PilE1 from *N. gonorrhoeae*. Both of these structures were used as molecular replacement search models using the programs Phaser (McCoy *et al.,* 2007) and MolRep (Vagin and Teplyakov, 2010). Molecular replacement calculation in which four or three copies in the ASU were sought was performed. The results of these molecular replacement processes using either search model did not produce interpretable electron density maps. The phasing statistics for the solutions produced by Phaser did not have translation function Z-score (TFZ) values greater than 5 which indicated that the phase problem had not

been solved (McCoy *et al.,* 2007). The log likelihood gain (LLG) values of the Phaser solutions did not change significantly as the solution progressed, again indicating the solution was not correct (McCoy *et al.,* 2007). The solutions from MolRep did not have a solution score greater than 0.3 indicating that the solution was not likely to be correct (Vagin and Teplyakov, 2010). In addition, the R-values of all presented solutions were greater than 50%. Since it was possible to readily and reliably reproduce R20291 PilA1$\triangle$1-34 crystals, experimental phasing methods were pursued instead.

4.3.1.5 Experimental phasing

Single-wavelength anomalous dispersion (SAD) experiments were carried out using heavy atom derivatives. Crystals were reproduced and were soaked in solutions containing ~ 10 mM K_2PtCl_4, $AuCl_3$ and $Pb(CH_3CO_2)_2$. After a 10 minute incubation in the derivative solution, the crystals were back soaked in the crystallisation solution from the drop reservoir to remove unbound and excess heavy atom derivates before flash cooling in liquid nitrogen (Garman and Murray, 2003).

Diffraction experiments were carried out at the I04 beamline at Diamond Light Source using a Pilatus 6M detector [Dectris]. To check for the presence of a heavy atom, fluorescence scans were performed at energy ranges that covered the L-III absorption edge of each heavy atom (Pt, Au and Pb) (Figure 4.3.2). The raw fluorescence data were processed in Chooch (Evans and Pettifer, 2001) at the beamline to calculate the peak and inflection points. The incident radiation wavelength was then set to that corresponding to the peak energy for the heavy atom present (Figure 4.3.2).

Two test images (90° apart) of each crystal were collected for indexing and optimum data collection strategy (subsection 4.3.1.2), as described previously. High redundancy is key for a successful SAD experiment as anomalous scatterers result in only small differences in intensities that are comparable to or lower than the noise in a dataset, therefore excellent collection statistics are required to detect these differences (Dauter *et al.,* 2002). To achieve a high redundancy, a total crystal oscillation of 720° was performed (Dauter *et al.,* 2002). To minimise radiation damage, the beam transmission was attenuated to 40% (2.66 x 10^{12} photons/second at the Pb peak energy), for all derivative datasets, to ensure that meaningful data were still being collected towards the end of the experiment as crys-

tals that contain heavy atoms suffer from greater radiation damage than native crystals, especially at the atom absorption edges (Owen and Sherrell, 2016).

The datasets were indexed and integrated using XDS in which the Friedel's law parameter was set to 'false' so that I+ and I- were handled separately during data processing (Kabsch, 2010), since the law breaks down when anomalous scatters are present (Dauter *et al.,* 2002). These data were then scaled and reduced using Aimless, and particular attention was paid to the anomalous signal, anomalous completeness and anomalous multiplicity, in addition to the usual indicators of dataset quality previously described (Table 4.3.3). Soaking the crystals with the heavy atom derivatives did not change significantly the unit cell parameters or spacegroup, indicating crystals of native and derivative were isomorphous. Reflections were observed to a range of 3-1.5 Å, similar to the native crystals (Table B.0.1).

Derivative and native data (Section 4.3.1.2) were prepared for use in the SHELX (Sheldrick, 2010) and hkl2map (Pape and Schneider, 2004) suite using Aimless and each derivative dataset was processed using SHELXC. Attention was paid to the anomalous difference signal (d"/sig) of each derivative dataset in given resolution bins. The difference signal describes the strength of the anomalous signal (differences in intensities between and $-h, -k, -l$) which is used to derive crystallographic phases (Dauter *et al.,* 2002). The difference signal *vs* resolution for each of the derivative datasets is plotted in Figure 4.3.3. The dataset with the highest resolution bin with a difference signal >0.8 was chosen for further processing, this was the Pb dataset in which the anomalous difference signal remained >0.8 to a resolution of 2.7 Å but overall lower than the others (Figure 4.3.3D). The remaining datasets had a difference signal >0.8 to a resolution ~5 Å and as such these datasets were much less likely to yield an interpretable electron density map into which the atomic structure of PilA1 could be built.

SHELXD (Schneider and Sheldrick, 2002) was used to find the Pb substructure with the initial search of 8 Pb sites in the ASU. The correct number of heavy atom sites would be resolved later using SHELXE. SHEXLD was instructed to perform 1000 tries to find the heavy atoms. The CCall/CCweak statistics (Figure 4.3.4A) showed two distinct populations of solutions, suggesting that a correct heavy atom substructure had been determined (Sheldrick, 2007). The best heavy atom substructure solution was output in a PDB file.

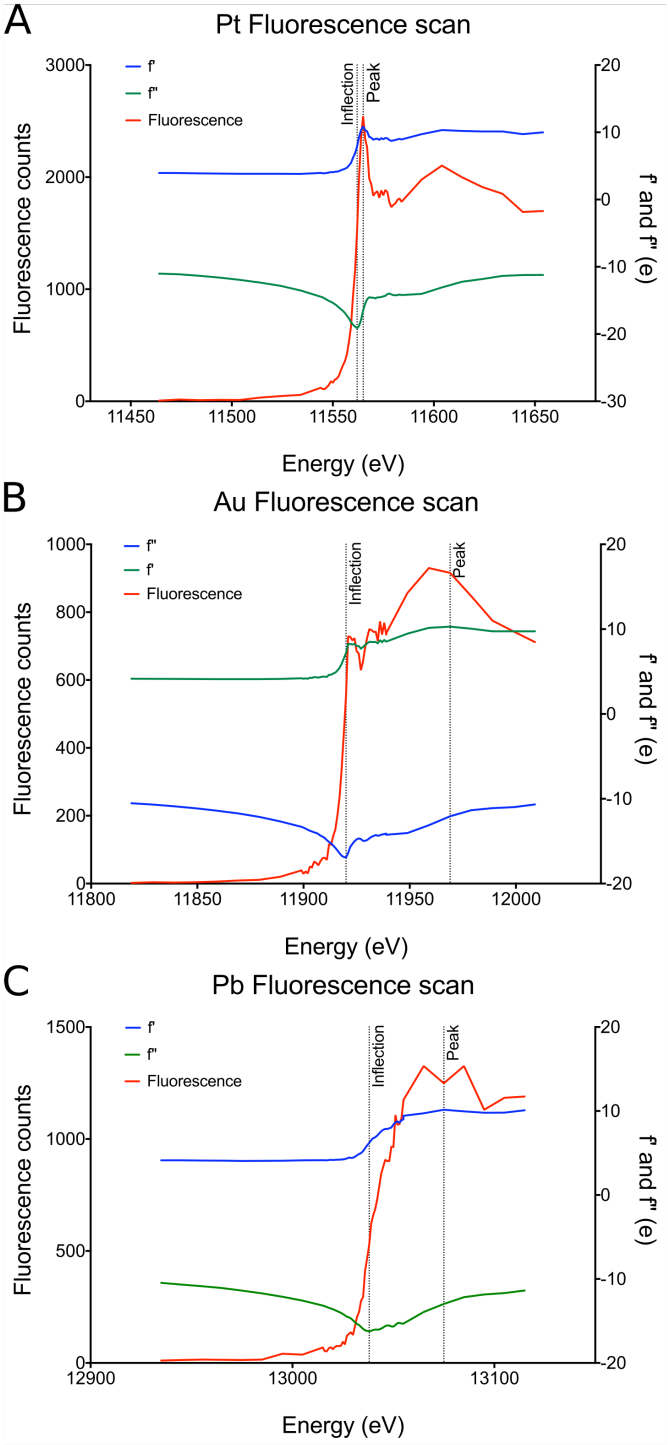


Figure 4.3.2 – Fluorescence scans of derivative soaked PilA1△1-34 R20291 crystals. Fluorescent scans were performed on the derivative crystals to determine the presence of heavy atom and the energy of the absorption edge (L-III) for each derivative atom tested (Pt, Au and Pb). Fluorescent counts were processed using Chooch (Evans and Pettifer, 2001) to calculate the f' and f" of each scan in addition to the peak and inflection points. The Pt (A) scan had a peak at 11562 eV (1.0723 Å) and inflection point at 11565 eV (1.0721 Å), Au (B) peak at 11969 eV (1.0359 Å) and inflection at 11920 eV and Pb (C) peak at 13075 eV (0.9483 Å) and inflection at 13038 eV.

	Pt-1	Pt-2	Au	Pb
Resolution (Å)	46.13 - 3.00	46.26 - 2.50	46.61 - 1.54	46.6 - 2.00
Wavelength (Å)	1.07206	1.07206	1.03588	0.94825
Unit cell dimensions				
a=b, c (Å)	101.74, 103.50	102.04, 103.79	103.16, 104.49	102.37, 104.67
α=β=γ (°)	90.00	90.00	90.00	90.00
Spacegroup	$P4_12_12$ or $P4_32_12$	$P4_12_12$ or $P4_32_12$	$P4_12_12$ or $P4_32_12$	$P4_12_12$ or $P4_32_12$
R_{merge}*	0.156 (1.773)	0.207 (4.342)	0.202 (1.543)	0.100 (1.095)
Total Reflections	241978	983182	5193752	1958556
Unique Reflections	11401	19604	81820	38380
Mean I/σI	14.8 (1.7)	16.5 (1.4)	16.6 (0.7)	38.2 (5.1)
Mean intensity CC1/2	0.999 (0.707)	0.999 (0.659)	1.0 (-0.027)	1.0 (0.9)
Completeness (%)	100.0 (100.0)	100.0 (100.0)	97.9 (61.2)	99.9 (98.4)
Multiplicity	21.2	50.2	63.5	51.0
Anomalous signal (Aimless, CC_{anom} >0.15) (Å)†	5.63	4.8	5.3	2.8
Anomalous Completeness (%)	100.0 (100.0)	100.0 (100.0)	97.1 (47.6)	99.8 (97.2)
Anomalous Multiplicity	11.6 (9.8)	26.9 (26.6)	32.7 (3.6)	26.9 (25.1)
Anomalous signal (SHELXC) <d"/sig> (Å)‡	5	5	5	2.7

Table 4.3.3 – Table of crystal parameters of derivative R20291 PilA1Δ1-34 crystals. Crystal parameters of derivative soaked R20291 PilA1Δ1-34 crystals. Values in parentheses represent the highest resolution shell. *$R_{merge} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is the ith observation of relfection hkl and $\langle I(hkl) \rangle$ is the massed average intensity for all observations i of reflection hkl. †CC_{anom} = resolution at which CC_{anom} is greater than 0.15**.** ‡Resolution at which <d"/sig > < 0.8.

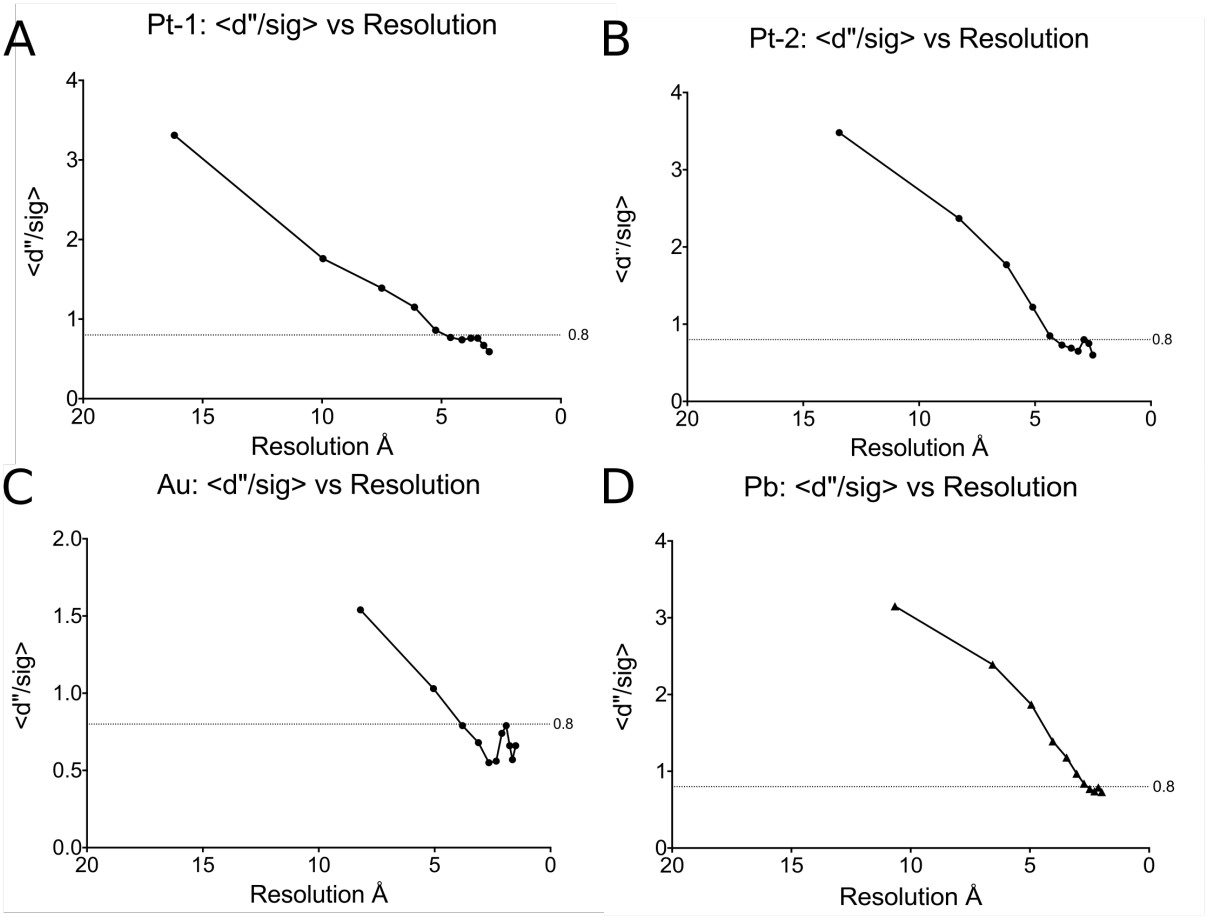


Figure 4.3.3 – SHELXC - Difference signal vs resolution in heavy atom derivative datasets. The difference signal in datasets Pt-1 (A), Pt-2 (B) and Au (C) remained >0.8 to a resolution of ~5 Å. The dataset containing Pb-heavy atom (D), had a difference signal >0.8 to a highest resolution of 2.7 Å.

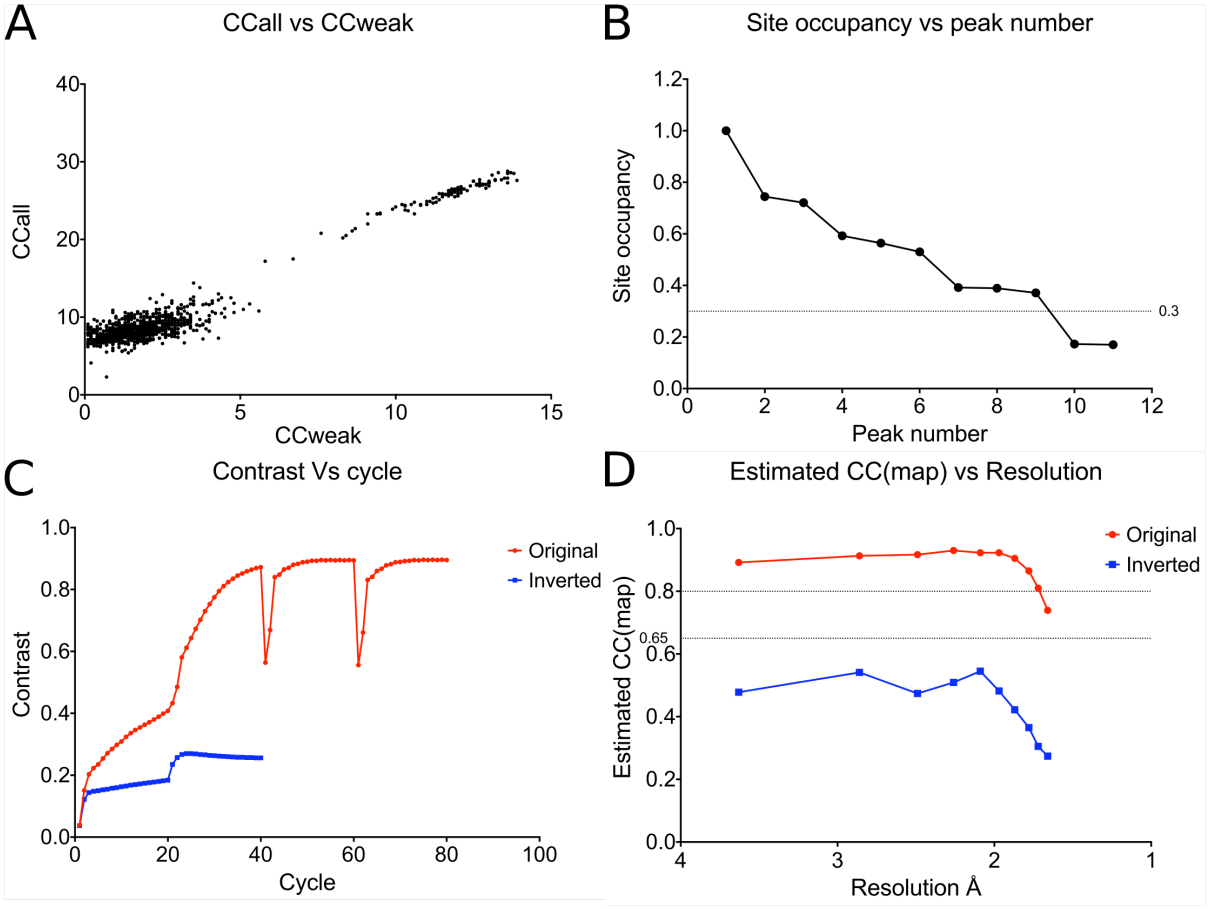


Figure 4.3.4 – Pb dataset statistics from SHELXD/E for three molecules in the ASU. A: CCall vs CCweak output from SHELXD, two distinct populations of solutions are seen, indicating that a correct solution has been found. **B:** Heavy atom site occupancy vs peak number. 9 heavy atom sites were found by SHELXE as indicated by the drop in occupancy below 0.3 at peak numbers greater than 9. **C:** Contrast vs cycle during auto-tracing cycles in SHELXE of which there were 80 for the original enantiomorph (red) and 40 for the inverted enantiomorph (blue). **D:** The estimated map CC vs resolution. A map (CC) above 0.8 indicates an interpretable map although a value above 0.65 is indicative of a good solution (Sheldrick, 2010). The inverted enantiomorph did not reach the 0.65 threshold while the original enantiomorph is above 0.8 to a highest resolution of 1.7 Å. The graphs in panels C and D indicate that the original enantiomorph is the correct solution. The final mean figure of merit (FOM) was 0.71 and pseudo-free CC was 77.8%.

The substructure solution from SHELXD was used in SHELXE for phase calculation for both possible enantiomorphs. The number of heavy atom sites found by SHELXD was 9 (Figure 4.3.4B), as determined by the occupancy threshold. As 9 is divisible by 3, it was most probable that there were three copies of PilA1 in the ASU and not four as suggested by the Matthews coefficient calculation (Table 4.3.2). Therefore, SHELXE calculations were performed using a fractional solvent content of 0.57, corresponding to three copies of PilA1 in the ASU.

SHELXE was instructed to invert the heavy atom substructure, this enabled the correct anomalous hand to be determined. From the statistics output by SHELXE, it was clear that the original hand was the correct one (Pape and Schneider, 2004). The contrast per cycle (Figure 4.3.4B) increased for the original hand during the first 40 cycles and remained greater than 0.8, while the inverted hand did not increase above a contrast of 0.3 and SHELXE stopped after 20 cycles. A contrast of 1.0 is indicative of a correct hand so the observed CCmap greater than 0.8 to a resolution of 1.7 Å confirms that the original hand is correct (Figure 4.3.4D). The inverted solution has a highest CCmap of 0.5 even at lower resolutions. A final mean FOM of 0.71 and pseudo-free CC of 77% was calculated suggesting that the final phases were correct (Sheldrick, 2010), and when the electron density map was visualised in Coot, clear contrast between protein features and solvent could be seen, confirming that the structure had been solved in spacegroup $P4_12_12$. When the calculations were repeated in $P4_32_12$, no interpretable electron density map could be obtained.

4.3.1.6 *Ab initio* structure solution of R20291 PilA1△1-34

Methods for determining a crystal structure without prior phases such as experimentally derived phases or a search model for replacement have been developed. Known as *ab initio* phasing, a fragment based search method using secondary structure elements has been developed in the program Arcimboldo (Rodríguez *et al.,* 2009; Millan *et al.,* 2015). Arcimboldo uses a combination of helix fragment placement utilising Phaser (McCoy *et al.,* 2007) followed by density modification and main chain tracing using SHELXE (Sheldrick, 2010). Such a method is the equivalent of determining the substructure as in experimental phasing methods, using secondary structure fragments rather than heavy atoms, and then

extending and improving the phases to produce a reliable map. Model and map quality are judged on the CC output by SHELXE, and Arcimboldo will conduct a number of rounds of these cycles until the CC value is greater than 25%, indicating a reliable solution has been found (Millan *et al.,* 2015). The output from Arcimboldo is a set of phases for the input intensities and traced coordinates (poly-alanine fragments), that can be used as a basis for model building in a program such as Buccaneer (Cowtan, 2006).

The requirements to use Arcimboldo are data to a resolution greater than 2 Å, information about helix content (maximum helix length), number of molecules in the ASU and the molecular mass of the target protein. As native data for R02921 PilA1$\triangle$1-34 are of a higher resolution that 2 Å and are predicted to contain a helix ~30 residues long, phasing using Arcimboldo was tested (Figure 4.3.5). A final CC of 47.4% was calculated and 395 residues were traced. The phases were input into the model building program Buccaneer that built 411 residues out of 441. An initial round of refinement using Refmac5, followed by manual model inspection and adjustment in Coot, resulted in an R_{free} of ~25 %. This is comparable with the model that had been built using experimentally derived phases (Section 4.3.1.8), and after comprehensive refinement and validation had the potential to provide a final model of equal quality.

The ability to solve the phase problem without experimentally derived phases or a search model for molecular replacement methods is a great advantage. However, the R20291 PilA1$\triangle$1-34 dataset was of a high resolution and quality (low R_{merge}, high I/σI and completeness), which is a strict requirement for successful *ab initio* phasing (Rodríguez *et al.,* 2009).

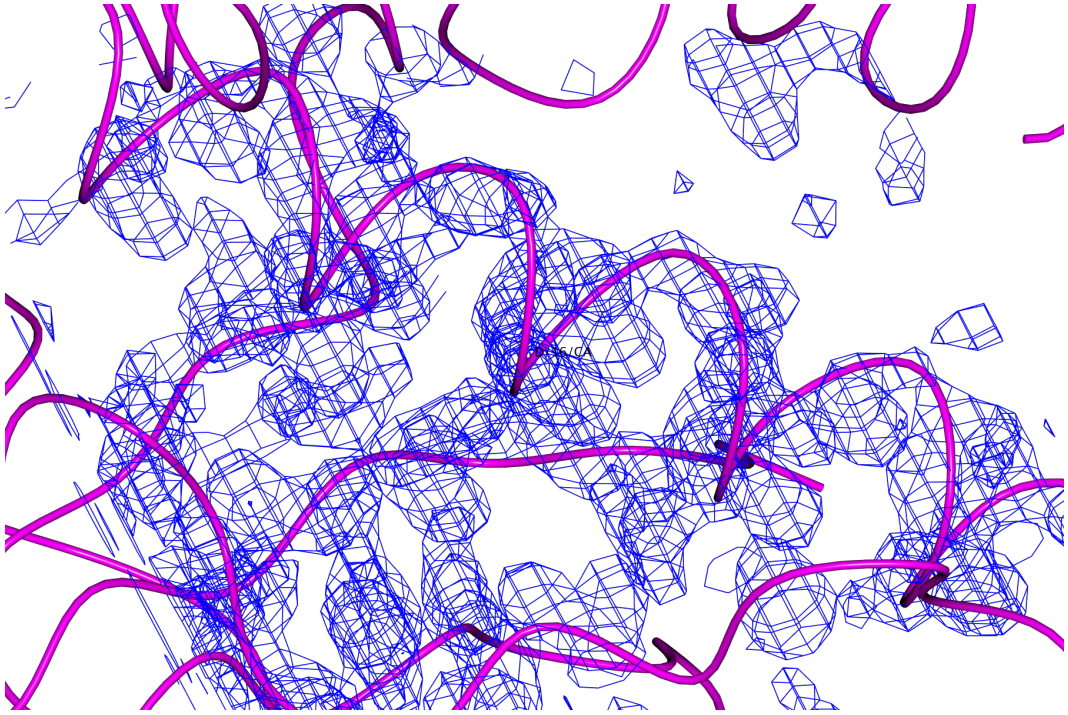


Figure 4.3.5 – R02921 PilA1△1-34 best coordinates and phases displayed in CCP4MG. Zoomed on the longest helix in R02921 PilA1△1-34 (α1), the resulting electron density map (map level: 1.3 e/A^3) from Arcimboldo is of good quality and side chains are clearly visible. The traced coordinates are represented in C-alpha/backbone (McNicholas *et al.,* 2011).

4.3.1.7 Model building

The phases output by SHELXE after Pb-SAD phasing and the sequence of the R20291 PilA1△1-34 construct were input into the automated model building program, Buccaneer, and after five cycles of building and refinement using Refmac5, 402 residues (out of 441) were built in 3 fragments (Cowtan, 2006). R_{work} and R_{free} after the final cycle were 20.8% and 20.5%, respectively.

4.3.1.8 Model refinement and validation

Cycles of automated refinement in Refmac5 (Murshudov *et al.,* 2011), using the native data, followed by manual model inspection and adjustment using Coot were performed (Emsley *et al.,* 2010). Waters were arranged around the 3 protein molecules in the ASU using the 'arrange waters' tool in Coot (Emsley *et al.,* 2010). The positions of the water molecules were checked in Coot based upon distances that were less than 2.3 Å or greater than 3.5 Å and with a B-factor value lower than 80 Å^2. Waters that did not

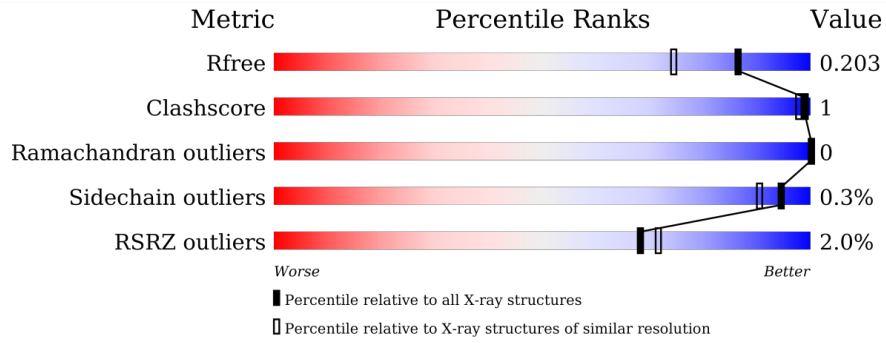
fulfil these criteria were removed or refined. To validate the model, geometry, clashes, rotamers, Ramachandran outliers, bond lengths and angles were analysed and checked using MolProbity (Chen *et al.,* 2010). MolProbity also added riding hydrogen atoms to the model and flipped any asparagine, glutamine or histidine molecules that clashed as a result, increasing hydrogen bonding potential. The resulting coordinate file was used for further refinement in Coot and Refmac5. The refinement cycles were continued until all the parameters analysed by MolProbity were within acceptable ranges. Anisotropic B factors were selected since the data had a resolution of 1.6 Å and the number of unique reflections in the dataset was greater than the number of atoms in the ASU multiplied by six (Merritt, 2012). There are six parameters that describe a thermal ellipsoid of an atom that displays anisotropy at high resolution, therefore at least six reflections per atom are required to describe each atom (Merritt, 2012). After several rounds of manual inspection and building in Coot and automated refinement in Refmac5, the R_{work} and R_{free} values were reduced to less than 25%. However, the bond angles and bond length root mean squared deviations (RMSD) were outside acceptable ranges of 1.5° - 1.3° and 0.01 Å - 0.02 Å, respectively (Rossmann and Arnold, 2001). To ensure the bond angles and lengths were within reasonable, expected values, the refinement weighting term was adjusted in Refmac5 over a number of refinement rounds to a final value of 0.2. To ensure that the refinement geometry was not 'over' massed, attention was paid to the R_{work} and R_{free} values, which did not change as a result of defining a massing term in Refmac5.

Once all the peaks in the electron density map had been accounted for and the parameters in MolProbity were satisfied, refinement was stopped (Figure 4.3.6A). The R20291 PilA1$\triangle$1-34 model was submitted to the wwPDB validation server (validate-rcsb-1.wwpdb.org) and a validation report was generated. The report included information comparing the R20291 PilA1$\triangle$1-34 model statistics with other structures in the PDB, which indicated that the quality of the model was within the top 30% of structures at 1.65 Å (Figure 4.3.6B). The programme Polygon (Urzhumtseva *et al.,* 2009) from the Phenix suite was also used to assess the refinement parameters of the R20291 PilA1$\triangle$1-34 model with other structures in the PDB of the same atomic resolution (Figure 4.3.6C). The polygon indicated that the RMSD bond lengths, angles, R_{free} and mean B factor were in the average range. The R_{work} was slightly higher than average and the clash score was lower than aver-

A

All-Atom Contacts	Clashscore, all atoms:	0.16		100th percentile* (N=829, 1.65Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	1	0.29%	Goal: <0.3%
	Favored rotamers	344	98.29%	Goal: >98%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	394	98.50%	Goal: >98%
	MolProbity score^	0.56		100th percentile* (N=9033, 1.65Å ± 0.25Å)
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad bonds:	0 / 3048	0.00%	Goal: 0%
	Bad angles:	1 / 4153	0.02%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 12	0.00%	Expected: ≤1 per chain, or ≤5%

B



C

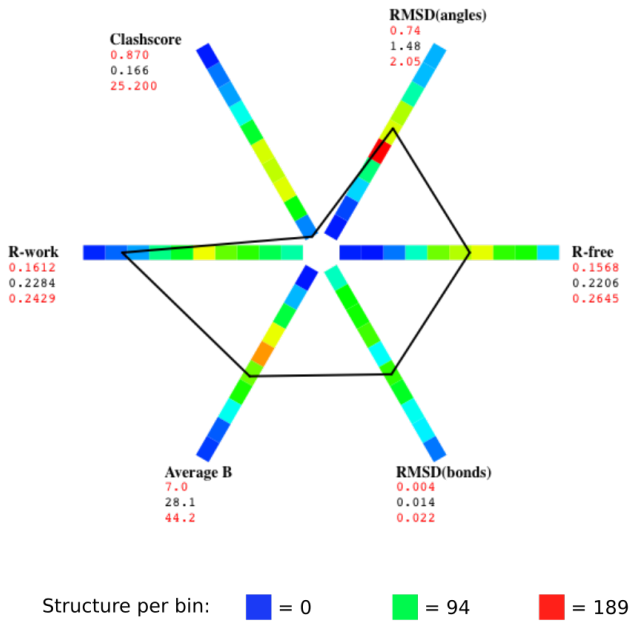


Figure 4.3.6 – R20291 PilA1△1-34 model validation. MolProbity validation statistics (A) and Polygon (C) refinement comparison. Model geometry was analysed using MolProbity (A) after each round of refinement and used to identify problematic residues. The R20291 PilA1△1-34 model statistics were compared with other structures in the PDB using the wwPDB validation tool (B) and the Phenix program Polygon (C). The wwPDB validation sliders (B) rank model statistics from worse (red) to better (blue). All of the parameters shown by the sliders (B) are in the better range compared to other structures of similar resolution - in this case 1.65 Å. Polygon compares each parameter using bins. These bins are coloured upon the number of structures in each bin at similar resolution to the model provided from the PDB (see key). Under each parameter the highest (top) and lowest (bottom) bins are defined and the value from the model is shown in red. Black lines describe where the model statistics for each parameter are located within the bins and should lie in the bins with the highest number of structures or better. The RMSD bond lengths, RMSD angles, R_{free} and average B factor are average for a structure of this resolution. The clash score is better while the R_{work} is higher than average but within the range observed for structures of 1.65 Å.

age. Together the wwPDB validation sliders and Polygon results showed that the R20291 PilA1△1-34 model was of a good quality. The final refinement statistics are shown in Table 4.3.4.

Refinement statistics for native R20291 PilA1△1-34	
R_{work} (%)	14.60
R_{free} (%)	19.07
No. of non-H atoms	
No. of protein atoms	3010
No. of solvent atoms	440
RMSD	
Bond angle (°)	1.49
Bond length (Å)	0.01
Average B factor (Å^2)	28.12
Ramachandran plot, residues in	
Most favoured regions (%)	98.25
Allowed (%)	1.75
Outliers (%)	0

Table 4.3.4 – Table of R20291 PilA1△1-34 model refinement statistics. Refinement statistics output by Refmac5 and MolProbity. $R = \frac{\Sigma_{hkl}|F\frac{obs}{hkl} - F\frac{calc}{hkl}|}{\Sigma_{hkl}F\frac{obs}{hkl}}$.

4.3.2 Structure determination of 630 PilA1△1-34

4.3.2.1 Crystallisation

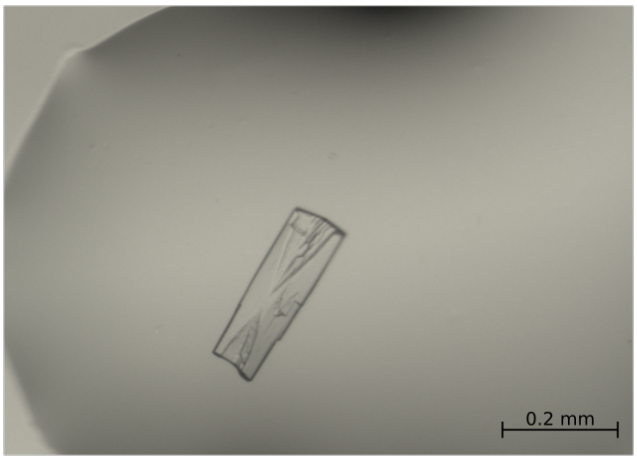


Figure 4.3.7 – 630 PilA1△1-34 crystal. A single crystal of PilA1△1-34 from 630 obtained in a 100 nl:100 nl drop with crystallisation solution Structure D1 [Hampton] (0.2 M sodium acetate, 0.1 M Tris pH 8.5 and 30% PEG 400). The crystal formed within 10 days of the protein and crystallisation screen being dispensed and appeared to have defects that suggested there were hollow regions.

Crystallisation trials of 630 PilA1△1-34 were set up at 18 mg/ml and 9 mg/ml in the commercial screens Structure, Index, Morpheus and JCSG+. A single crystal of PilA1△1-34 from 630 was obtained (Figure 4.3.7B) in 0.2 M sodium acetate, 0.1 M Tris pH 8.5 and 30% PEG 400 (Structure, D1) in a 1:1 protein:crystallisation solution drop ratio. The crystallisation solution did not contain any cryo-protecting compounds, so crystallisation solution from the reservoir was used to dilute PEG 400 to a final concentration of 25% v/v and the crystal was harvested from the crystallisation drop, soaked for 10 seconds in the cryo-solution and flash cooled in liquid nitrogen (Garman and Schneider, 1997). The crystal was tested on a MicroMax home source X-ray generator [Rigaku] equipped with an Raxis IV++ detector [Rigaku]. Diffraction spots were observed to a resolution of ~2 Å (Figure 4.3.8), however it was not possible to index these data using Mosflm or XDS. The single crystal obtained appeared to have defects that suggested that there were hollow regions which could have resulted in poor quality diffraction, particularly when a wide beam, as produced by the in house X-ray generator, was used. The diffraction image

shown in Figure 4.3.8 was of poor quality and diffraction spots were smeared which most likely inhibited indexing of these images. However, the images confirmed that this crystal is formed of protein and not salt, due to the number of diffraction spots and the close proximity of these reflections indicating a unit cell compatible with protein molecules. The crystal was stored in liquid nitrogen and sent to Diamond Light Source synchrotron.

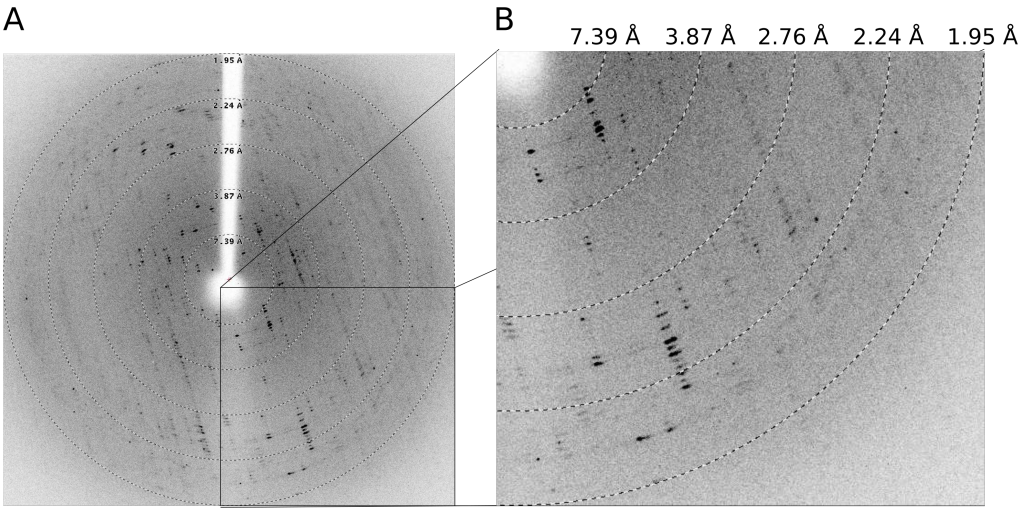


Figure 4.3.8 – Home source diffraction image of 630 PilA1△1-34. A: Diffraction images of the single crystal of 630 PilA1△1-34 (Figure 4.3.7) taken on a Rigaku 007 using a Raxis IV++ detector. **B:** shows a zoomed in corner of the diffraction image contains diffraction spots to a resolution of ~2 Å. These images were presented using the ADXV package (http://www.scripps.edu/tainer/arvai/adxv.html).

4.3.2.2 Synchrotron data collection

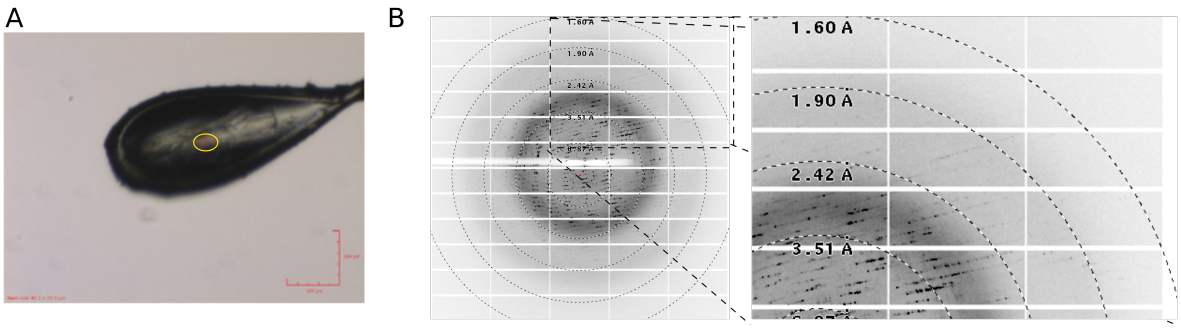


Figure 4.3.9 – 630 PilA1△1-34 synchrotron crystal diffraction images. A: Test images were taken at a central position along the crystal that appeared to contain fewer defects using a 43 x 30 μm beam (yellow circle). **B:** Smearing of diffraction spots was observed during diffraction, as observed in Figure 4.3.8 but to a lesser extent.

Diffraction experiments were performed by Dr Arnaud Baslé at the I04 beamline at Diamond Light Source. Test diffraction images were taken at a central position along the crystal (Figure 4.3.9A). Even though the crystal produced smeared diffraction spots, they were still indexable to a resolution of ~1.7 Å. The test images were indexed using Mosflm (Battye *et al.,* 2011) to determine the unit cell parameters and the spacegroup of the crystal as $P2_12_12_1$. The strategy function in Mosflm was used to calculate the optimum starting position for data collection in this spacegroup. A highly redundant dataset was collected over 200° with an oscillation of 0.1° per image.

The dataset was indexed using DIALS (Gildea *et al.,* 2014) before scaling and reduction using Aimless in CCP4i (Evans and Murshudov, 2013; Evans, 2006); the data collection statistics are summarised in Table 4.3.5.

	630 PilA1△1-34
Resolution (Å)	242.09 - 1.61
Wavelength (Å)	0.97951
Unit cell dimensions	
a, b, c (Å)	28.89, 54.67, 242.09
α=β=γ (°)	90.00
Spacegroup	$P2_12_12_1$
R_{merge}*	0.119 (1.421)
Total Reflections	345516
Unique Reflections	51272
I/σI	8.60 (1.6)
Mean intensity CC1/2	0.995 (0.607)
Completeness (%)	100.0 (100.0)

Table 4.3.5 – Table of crystal parameters of native 630 PilA1△1-34 crystals. The dataset collected of PilA1△1-34 630 was indexed and integrated using DIALS. Values in parentheses represent the highest resolution shell. *$R_{merge} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is the ith observation of reflection hkl and $\langle I(hkl) \rangle$ is the massed average intensity for all observations i of reflection hkl.

4.3.2.3 Molecular replacement and model building

The Matthews program (Matthews, 1968; Kantardjieff and Rupp, 2003) in the CCP4i suite was used to determine the unit cell contents (Table 4.3.6), indicating that the highest probability was three molecules in the ASU, with a 40% solvent content and Matthews coefficient of 2.05 $Å^3$/Da.

Nmol/asym	Matthews coef. ($Å^3$/Da)	Solvent (%)	P(1.65)	P(tot)
1	6.15	80.03	0.00	0.00
2	3.08	60.05	0.10	0.33
3	2.05	40.08	0.90	0.66
4	1.54	20.11	0.00	0.00
5	1.23	0.13	0.00	0.00

Table 4.3.6 – Unit cell contents of native 630 PilA1△1-34 crystals. Table of the unit cell content properties by number of molecules in the asymmetric unit.

Molecular replacement was carried out in the program MolRep (Vagin and Teplyakov, 2010) using the complete R20291 PilA1△1-34 model, cwith the 3 chains in the ASU but without water molecules. MolRep was instructed to search for three chains in the ASU. The phasing statistics from the MolRep solution were a final CC of 0.56 and contrast of 5.59 which indicated that the the phases were correct. A model was output by MolRep that included 408 residues out of a possible 438 for three molecules of the 630 PilA1△1-34 construct.

4.3.2.4 Model refinement and validation

The model produced by MolRep was refined in Refmac5 against the native data of 630 PilA1△1-34. The initial round of refinement calculated an R_{work} of 33.2% and R_{free} of 39.2%. The model was inspected and compared with the electron density map in Coot and manual refinement across all residues was carried out to improve the fit of the model to the electron density. The sequence was verified manually. Water molecules were added in Coot using the 'arrange waters' tool and checked as previously described (Emsley *et al.,* 2010). Anisotropic B factors were selected since the data contained a number of unique

reflections 6-fold greater than the number of atoms in the ASU (Merritt, 2012). Selection of anisotropic B factors decreased the R_{work} and R_{free} values.

Further rounds of refinement were performed using Refmac5 and Coot. The models output by Refmac5 were input into the MolProbity program to analyse model geometry and identify residues that did not fulfil normal geometric ranges (Chen *et al.,* 2010; Rossmann and Arnold, 2001). MolProbity also identified Asn/Gln/His residues that required flipping to remove clashes, these were flipped and new coordinates output which were used for further rounds of refinement.

Poor regions of density were present in all three chains between residues 70-85 in which it was difficult to accurately model residue side chains. Moreover, backbone density was missing in all chains for residue 71 (Asp), while in chain B no density was observed between residue 71 and 74. Other problematic residues included the final 4 residues at the C-terminal of chain C. To address the missing density, these residues were either replaced with alanine or their occupancies optimised during automated refinement with Refmac5 (Murshudov *et al.,* 2011). Changes in the R_{work} and R_{free} were negligible but a small improvement in geometry was observed when these sides were mutated to alanine and therefore the model containing alanine in substitution for these residues was used for further refinement and analysis. After several rounds of refinement using these settings it was not possible to improve the model statistics any further. The final refinement statistics are shown in Table 4.3.7.

The 630 PilA1$\triangle$1-34 model was submitted to the wwPDB validation server and a validation report was produced. The refinement parameters were compared with all structures and structures of a similar resolution in the PDB (Figure 4.3.10B), which indicated that there were problems with the 630 PilA1$\triangle$1-34 model. Most striking is the poor R_{free} in comparison to other structures. Analysis using Polygon also shows that the R_{free} was in the upper range for this resolution and in a R_{free} bin higher than the most common bin (Figure 4.3.10C). Polygon did indicate that RMSD bond lengths and angles and average B values were in average bins and the clash score was below average.

Overall these validation data show that the 630 PilA1$\triangle$1-34 model is not the best quality, however, further refinement and manual model adjustment does not improve the model. The model reflects the observed diffraction data and includes regions of poor or

Refinement statistics 630 PilA1△1-34	
R_{work} (%)	20.3
R_{free} (%)	26.8
No. of non-H atoms	
No. of protein atoms	2964
No. of solvent atoms	141
RMSD	
Bond angle (°)	1.5
Bond length (Å)	0.01
Average B factor (Å^2)	22.9
Ramachandran plot, residues in	
Most favoured regions (%)	97.7
Allowed (%)	2.0
Outliers (%)	0.25

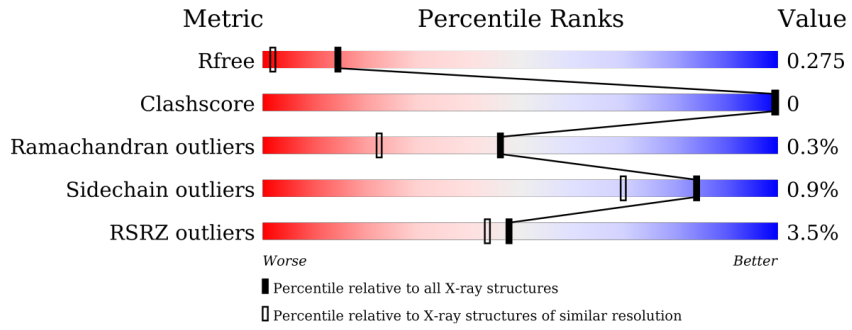
Table 4.3.7 – Table of 630 PilA1△1-34 model refinement statistics. Refinement statistics output by Refmac5 and MolProbity. $R = \frac{\Sigma_{hkl}|F\frac{obs}{hkl} - F\frac{calc}{hkl}|}{\Sigma_{hkl}F\frac{obs}{hkl}}$.

missing density. This may be the result of poor quality data, which correlates to the imperfect diffraction patterns for this unique crystal, which could not be reproduced. Even though the completeness and noise quality measurements of the collected data were within an acceptable range (Table 4.3.5), the R_{merge} of the native 630 PilA1△1-34 data was relatively high in the highest resolution shell (1.421) and inspection of the diffraction images revealed diffractions spots that were smeared and not always well defined. The data were collected using 100% beam transmission that may have resulted in radiation damage to the crystal during the experiment and therefore the data at the end of the experiment is not as high quality as the data at the start of the experiment.

A

All-Atom Contacts	Clashscore, all atoms:	0.33		100th percentile* (N=711, 1.61Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	2	0.60%	Goal: <0.3%
	Favored rotamers	317	95.77%	Goal: >98%
	Ramachandran outliers	1	0.25%	Goal: <0.05%
	Ramachandran favored	387	97.97%	Goal: >98%
	MolProbity score^	0.63		100th percentile* (N=7200, 1.61Å ± 0.25Å)
	Cβ deviations >0.25Å	2	0.53%	Goal: 0
	Bad bonds:	0 / 2992	0.00%	Goal: 0%
	Bad angles:	0 / 4062	0.00%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 9	0.00%	Expected: ≤1 per chain, or ≤5%

B



C

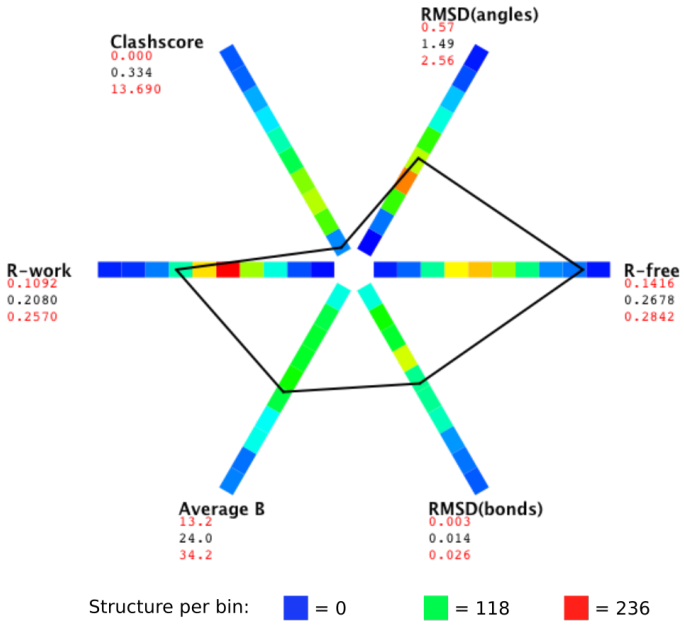


Figure 4.3.10 – 630 PilA1△1-34 model validation. Model geometry was analysed using MolProbity (A) after each round of refinement to identify problematic residues. MolProbity indicated that there were some problems with the model geometry including poor rotamers, Ramachandran outliers and Cβ deviations (>0.25 Å). The 630 PilA1△1-34 model statistics were compared with other structures in the PDB using the wwPDB validation tool (B) and the Phenix program Polygon (C). The wwPDB validation sliders (B) rank model statistics from worse (red) to better (blue). Clashscore and side chain outliers are in the better region however, the R_{free} is worse (high) for a structure based on data of this resolution (1.61 Å).

4.3.3 Analysis of the PilA1△1-34 crystal structures

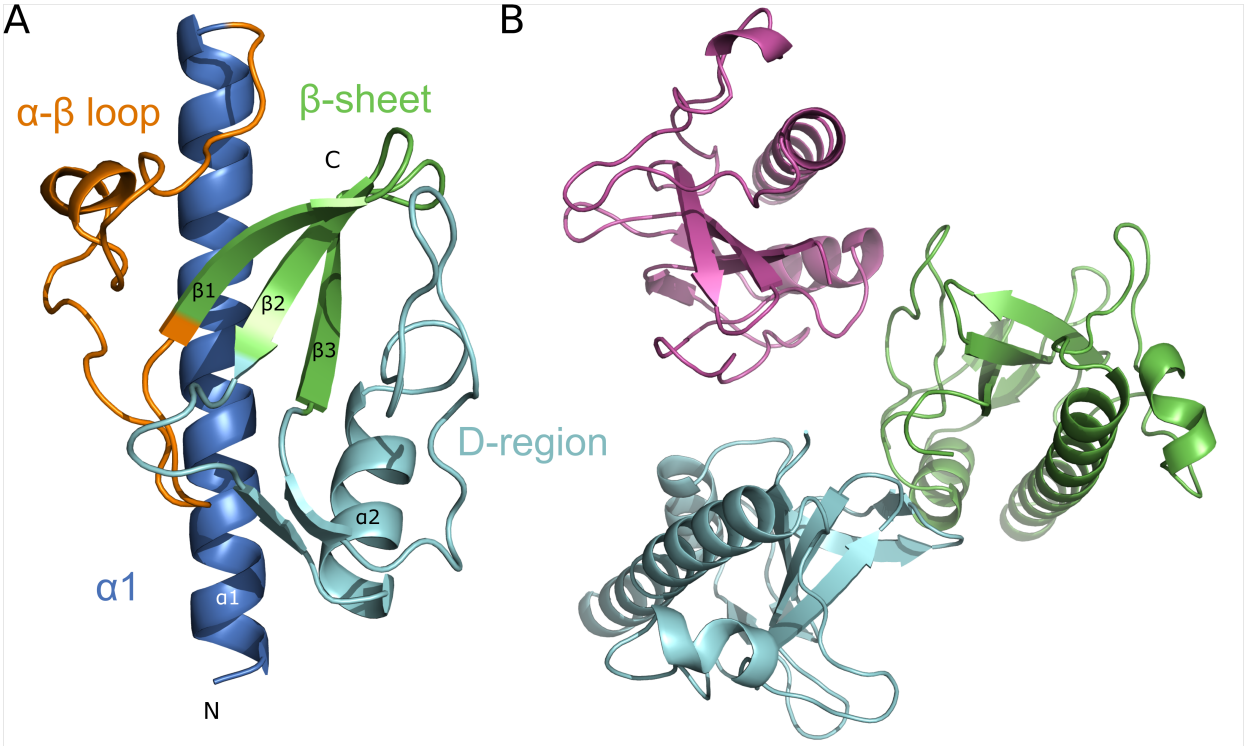


Figure 4.3.11 – R20291 PilA1△1-34 crystal structure model. A: Chain A of R20291 PilA1△1-34 represented in cartoon viewed from the predicted solvent exposed side. The R20291 PilA1△1-34 contains structural regions associated with TFPa pilin protein, including a long α-1 helix (dark blue) followed by a α-β loop, that connects the long α-helix and anti-parallel β-sheet (green). The variable D-region (cyan) contains a shorter α-helix, 10 residues long but is largely loop. **B:** Three molecules were observed in the ASU of R20291 PilA1△1-34 shown viewed from the C-terminal end of the α-1 helix.

The structure of R20291 PilA1△1-34 (Figure 4.3.11A) is typical of the TFPa pilins, containing a long N-terminal α-helix (α1), which is linked by an α-β loop to an anti-parallel β-sheet (β1-β2-β3) which encapsulates the D-region containing a shorter α-helix (α2) and the rest of this sub-domain is largely formed of loop. As there are no Cys residues in the PilA1 ORF, R20291 PilA1△1-34 lacks the disulphide bridge that delimits the D-region of Gram-negative pilins.

The crystal structures of PilA1△1-34 from the *C. difficile* strains R20291 and 630 are very similar (core RMSD: 0.7 Å). As such the 630 PilA1△1-34 structure contains the same structural elements as the R20291 structure. This is not a surprising result as the *pilA1* open reading frames from these two strains are 91% identical, although the differences

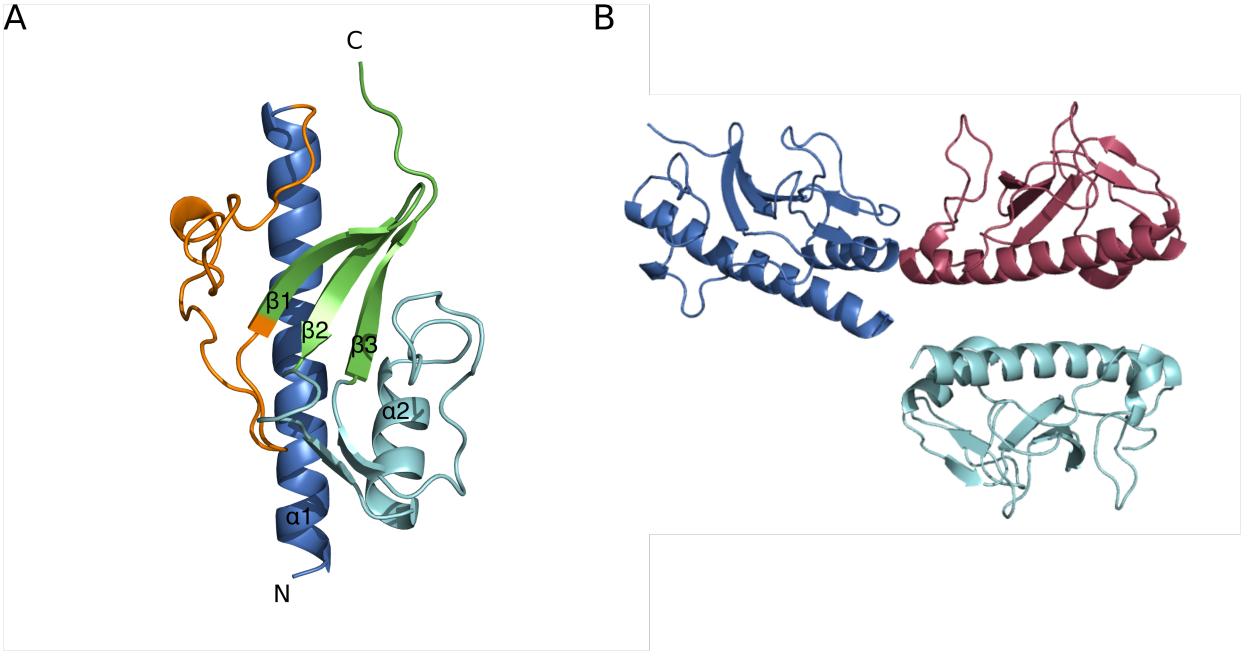


Figure 4.3.12 – 630 PilA1Δ1-34 crystal structure model. A: Viewed from the predicted solvent exposed side, 630 PilA1Δ1-34 is also representative of a typical of TFPa pilin protein. The long α-1 helix (dark blue) that is linked to an anti-parallel β-sheet (green) by an α-β loop. The variable D-region (cyan) contains the most sequence and structural divergence from the R20291 PilA1 (Figure 4.3.13).**B:** Three molecules of 630 PilA1Δ1-34 were present in the ASU and differ in C-termini residue conformation. Interestingly, the arrangement of the three moleulces in the ASU is different to that seen in R20291 crystals.

are restricted to the C-terminal region (Figure 4.3.13). Noticeable differences in these structures are observed between residues 139-145 (between β4 and β5), where there is an extended loop region in the R20291 structure and at the C-termini of the structures, which diverge (Figure 4.3.13). The extended loop in the R20291 structure is the result of a 3 residue insertion in the R20291 peptide sequence when aligned with the 630 peptide sequence (4.3.13).

Both the R20291 and 630 structures contained three molecules in the ASU. As part of the structure analysis, each of the coordinate files were submitted to the protein interfaces, surfaces and assembles (PISA) server (Krissinel and Henrick, 2007). PISA scores interfaces between molecules using a complex formation significance score (CSS), 0.0 indicating no biological significance and a value of 1.0 for a significant complex formation (Krissinel and Henrick, 2007). The CSS score can be used to assess whether molecular interfaces are biologically significant or a product of crystallisation. A summary of the interfaces detected by PISA are presented in Appendix B, in Table B.0.2 and Table B.0.3. However, a salt bridge between Asp62 (chain A) and Lys96 (chain C) in R20291 PilA1 drew attention with an apparent interface surface area totalling 7% of the solvent accessible surface area of both proteins (Figure 4.3.14, Table B.0.2, ID 1). A number of hydrogen bonding interactions were also predicted to contribute to this interface. Although the calculated solvation energy of the interaction had a P-value in chain C interface of 0.6, suggesting the interface was a product of crystal packing, the solvation energy P-value for chain A was 0.3 and suggested the interface on chain A could be biologically relevant. To test this hypothesis, three point mutants of R20291 PilA1Δ1-34 were produced, D62A, K96A and a double mutant D62A-K96A with the aim of attempting to crystallise them. However, it was not possible to solubly express and purify these mutants, suggesting these residues were important for protein stability.

Of the interfaces detected for the 630 PilA1Δ1-34 structure by PISA, none had a CSS score greater than 0 and none of the interfaces observed in the R20291 structure were seen in 630 PilA1Δ1-34. However, one interface was formed of two salt bridges Asp62-C:Lys39-B and Glu80-C:Lys150-B, forming a complex of 630 PilA1Δ1-34, with molecules aligned in the same orientation (Figure 4.3.15A/B) and with the polymerisation helix (α-1) on the same face (Table B.0.3, ID 4). The interface on chain C had a solvation energy P-

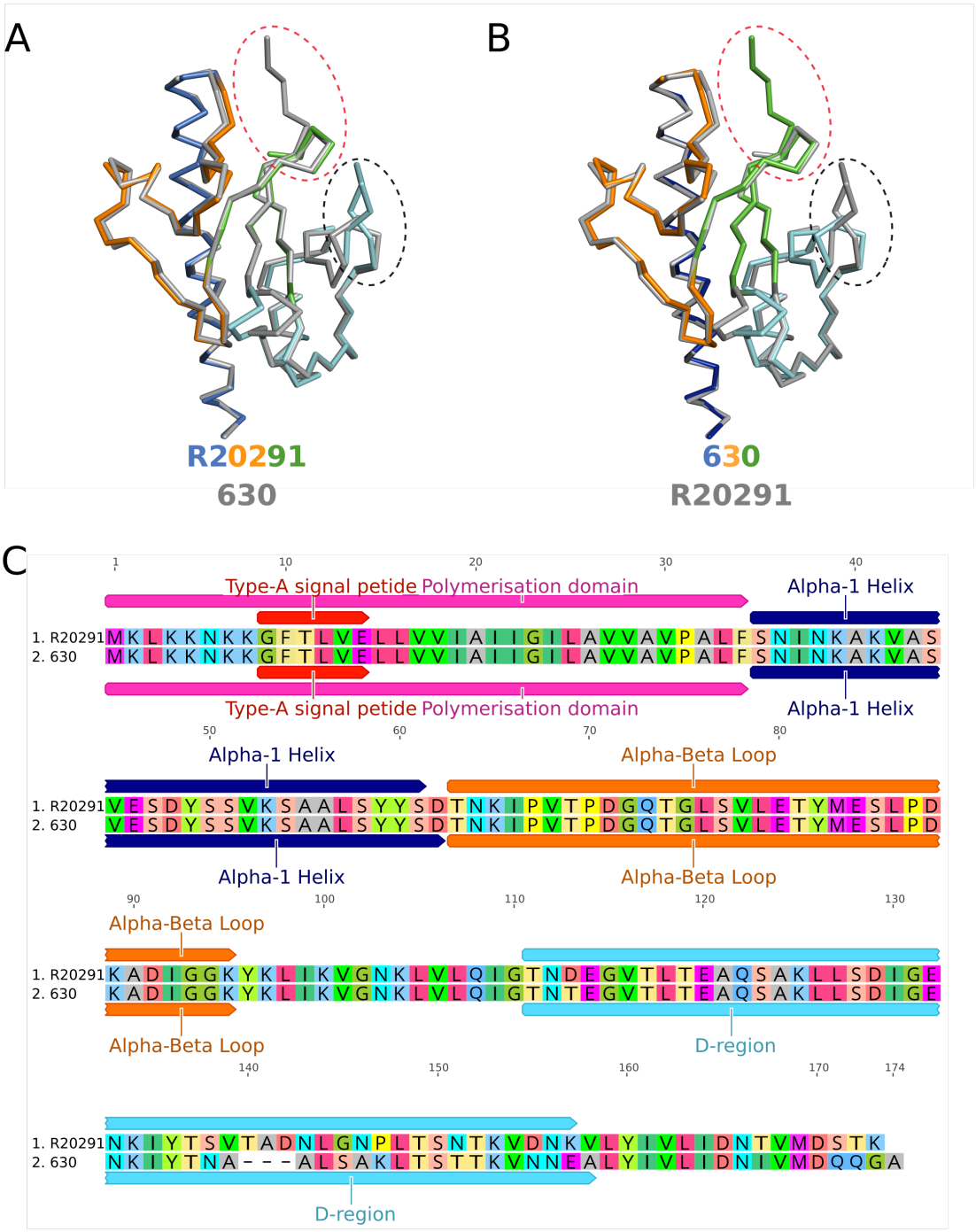


Figure 4.3.13 – Comparison of R20291 and 630 PilA1Δ1-34 backbone structures. The C-alpha backbone structures of R20291 and 630 were superimposed using SSM on chain A of each structure (Coot). The core RMSD was 0.7 Å. Structures are coloured by region: α-1 dark blue; α-β loop orange; β-sheet (green); D-region (cyan). For clarity R20291 is coloured and 630 is grey in panel A and 630 is coloured with R20291 in grey in panel B. The D-regions of the R20291 and 630 PilA1 are most divergent (C). While the α-helix (α2) within the D-region of both R20291 and 630 is structurally conserved, there are differences in the loop structure downstream of α2. In addition to these differences (A/B black dashed circle), five residues at the C-terminus of 630 PilA1Δ1-34 extend away from the β-sheet region, whilst in the R20291 structures this terminal region loops back towards the β-sheet (red dashed circle).

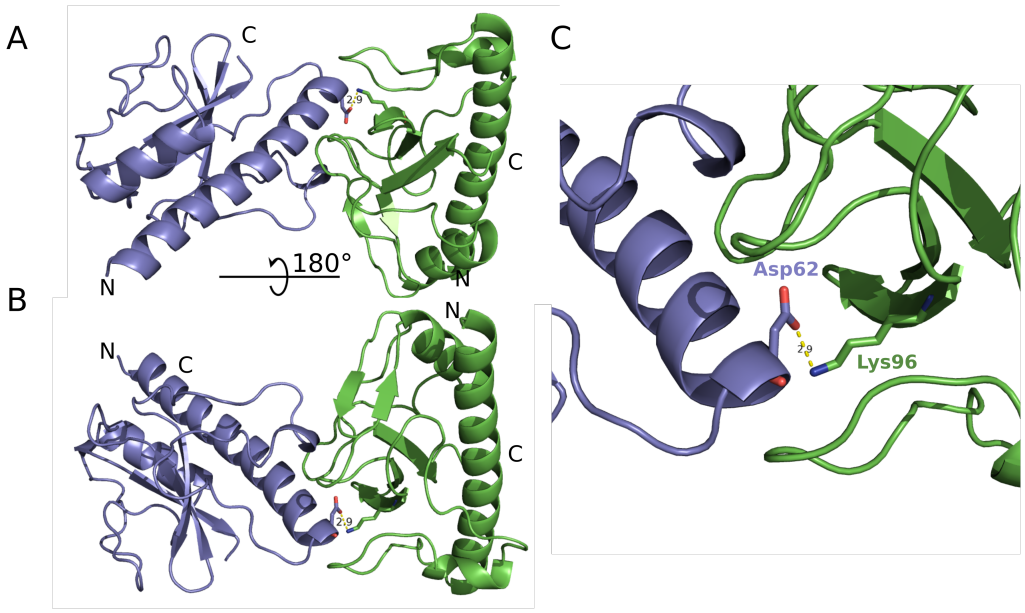


Figure 4.3.14 – R20291 PilA1△1-34 interface. Interface analysis using PISA indicated a possible interface between R20291 PilA1△1-34 molecules via a Asp62:Lys96 salt bridge (A/B are view rotated by 180°). The Asp62:Lys96 residues were 2.9 Å apart (C). The resulting orientation of the two R20291 PilA1△1-34 molecules is unlikely to be compatible with filament formation.

value of 0.6 and on chain B of 0.4, the overall P-value for the interface was 0.5 suggesting that the interface could be a result of crystal packing. It was not possible to mutate these residues within the scope of this work to test whether these residues were important, however, they should be the focus of future experiments. Interestingly, Asp62 in both R20291 and 630 structures were predicted to be involved in salt-bridges with different interfaces, which suggests that these salt-bridges could be significant. However, there was also no experimental evidence that either R20291 or 630 PilA1△1-34 formed multimers; during size-exclusion chromatography only monomer species were observed (Figure 4.2.2A, Figure 4.2.2E).

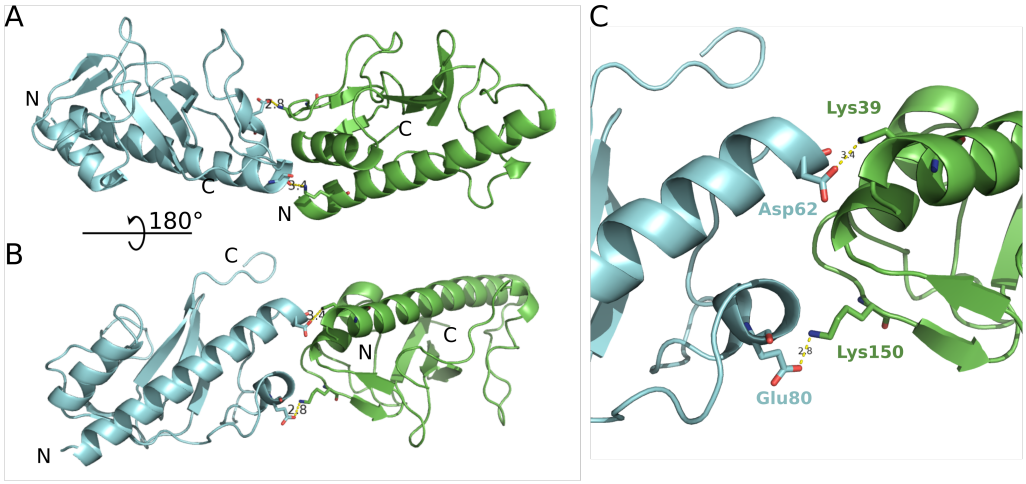


Figure 4.3.15 – 630 PilA1△1-34 interface. Interface analysis with PISA defined an interface between 630 PilA1△1-34 molecules that involved two salt bridges (Asp62:Lys39 and Glu80:Lys150), which arranged both 630 PilA1△1-34 molecules in the same direction (A/B). The Asp62:Lys39 were 3.4 Å apart and the Glu80:Lys150 residues were 2.8 Å apart. The acidic residues were positioned at the C-terminal end of the α-1 helix (Asp62) and within the α-β loop (Glu80), while the Lys residues in the other chain were positioned at the N-terminal end of the α-1 helix (Lys39) and within the D-region (Lys150).

4.4 Discussion

4.4.1 Comparison of Major Type IV pili structures

During the course of this project, the structures of PilA1 from three strains of *C. difficile* were published (Piepenbrink *et al.,* 2015), including R20291 (PDB: 4TSM, Figure 4.4.1A/B), NAP08 (PDB: 4OGM) and CD160 (PDB: 4PE2) at resolutions between 1.7 Å and 2.3 Å. These structures were of PilA1 that had been truncated to residue 26 and were fused to maltose binding protein (MBP) at the N-terminus and were His-tagged at the C-terminus. The MBP fusion was not removed prior to crystallisation and it appears in the structures (Figure 4.4.1A). The sequence coverage in the Piepenbrink structures is the same as observed in the structures presented in this chapter (residue 35-176).

The structures determined by Piepenbrink *et al.,* were aligned with the R20291 and 630 structures determined in this chapter using the secondary structure matching superimposition (SSM) tool in Coot to calculate the core RMSD between them. All of these structures are highly similar (within 1.0 Å) with the exception of the CD160. Protein sequence conservation of PilA1 proteins across *C. difficile* strains varies; R20291, 630 and

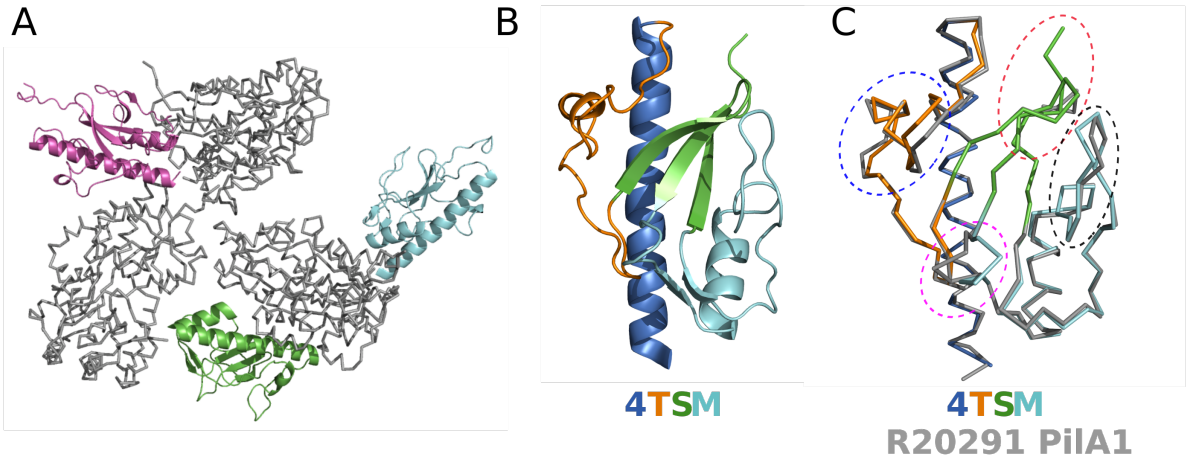


Figure 4.4.1 – R20291 PilA1 structure determined by Piepenbrink et al. 2015. A: Three copies of MBP-PilA1 (PDB: 4TSM) were present in the ASU of the R20291 structure determined by Piepenbrink *et al.,* The MBP tags (grey ribbon) formed a trimer with PilA1 at the outer edges (represented in cartoon). **B:** Chain A of MBP-PilA1 in cartoon representation showing the α1 helix (dark blue), α-β loop (orange), a 3-stranded β-sheet (green) and the D-region (cyan). **C:** The PilA1 structure by Piepbrink et al. 2015, (coloured) SSM superimposed onto the R20291 PilA1Δ1-34 (grey) and displayed in ribbon format for clarity. The structures have a core RMSD of 0.8 Å. Small structural differences are observed in the α-β loop and D-region which are highlighted by dashed circles.

NAP08 are between 82-90% sequence identical (Figure 4.4.2C). PilA1 from the strain CD160 has the lowest level of conservation of the *C. difficile* structures discussed here (<70% sequence identity). The sequence variability in the *pilA1* genes from these strains is limited to the α-β loop and D-region (Figure 4.4.2A/B). The D-region of TFPa pilin proteins has been modelled on the surface of pilin filaments (Craig and Li, 2008; Piepenbrink *et al.,* 2014; Piepenbrink *et al.,* 2015) and is widely recognised across many TFP pilins as forming a pilin-surface interface (Miller *et al.,* 2014). Therefore sequence variability is an important property for cell-surface and host cell adhesion, particularly for pathogens such as *C. difficile* and well studied examples including *N. meningitidis* (Szeto *et al.,* 2011; Miller *et al.,* 2014). Meanwhile, the N-terminal region of the TFP contains a highly conserved signal peptide sequence and polymerisation domain formed of an extensive α-helix (α1), and as expected there is little structural or sequence divergence in this domain across the *C. difficile* strains (Figure 4.4.2A/B). In addition to the conserved α1 helix, the β-sheet forming regions (residues 95-133 and 160-170) are also well conserved (Figure 4.4.2B).

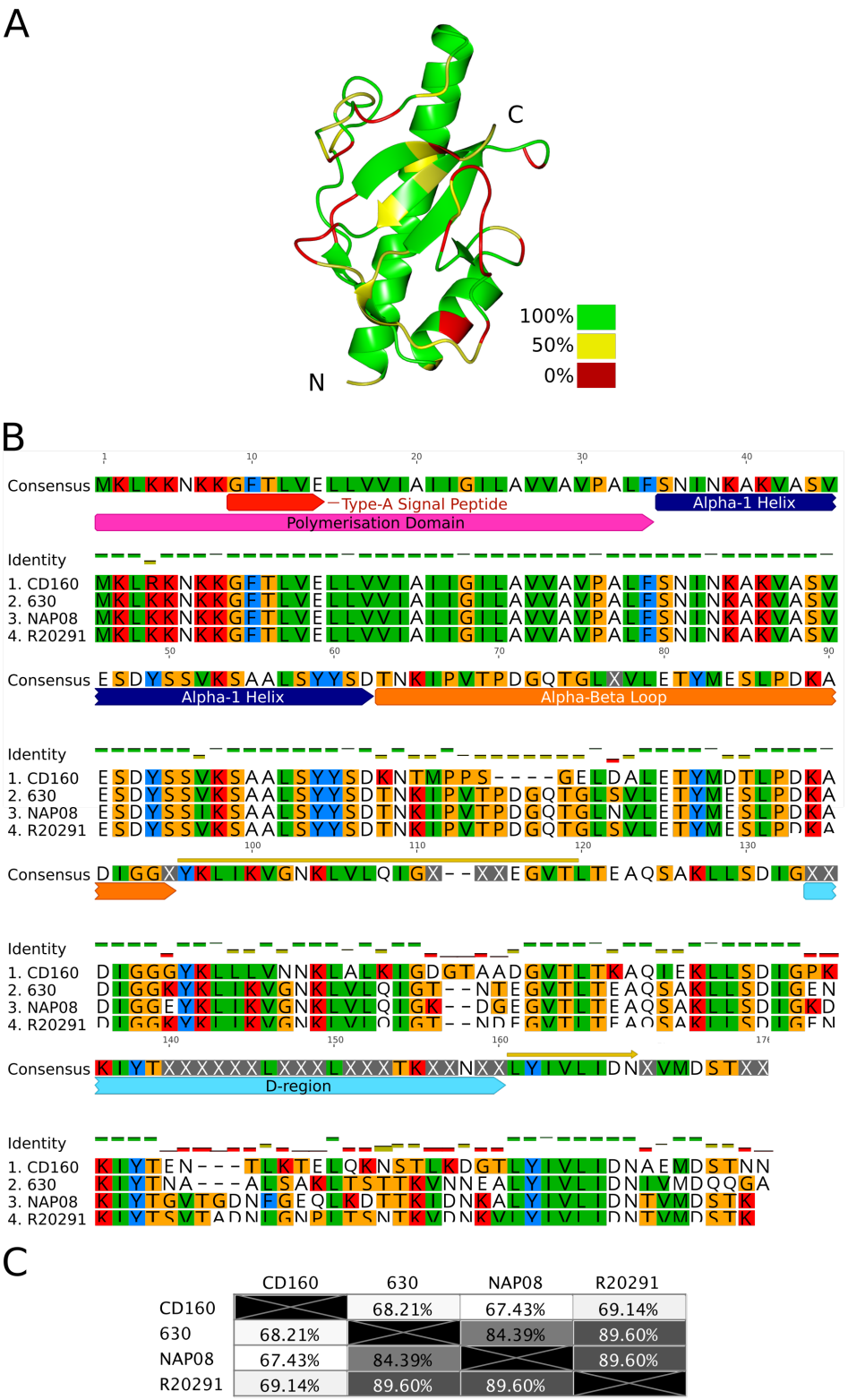


Figure 4.4.2 – PilA1 sequence conservation across *C. difficile* strains. **A:** R20291 PilA1Δ1-34 structure coloured by sequence conservation across R20291, 630, NAP08 and CD160. Green: 100% conservation; yellow: 50%; red: 0% conservation. **B:** The open reading frames of *pilA1* from R20291, 630, NAP08 and CD160 were aligned with ClustalO and the structural domains annotated in Geneious. The sequences are most divergent in the α-β loop (orange) and D-region (cyan). **C:** Identity matrix of *pilA1* from the R20291, 630, NAP08 and CD160 strains. CD160 *pilA1* is the most divergent of these *C. difficile* strains.

4.4.1.1 Structural similarities of Gram-positive TFP, Gram-negative TFP and pseudopilins

The first structure determined of a Gram-positive TFP pilin protein was of PilJ from *C. difficile* at a resolution of 1.98Å (PDB:4IXJ) (Piepenbrink *et al.,* 2014). PilJ is a minor pilin that has been shown to interact with PilA1 in *C. difficile* along the pilin fibre in a ratio of 1:2000 PilJ:PilA1 (Piepenbrink *et al.,* 2014; Piepenbrink *et al.,* 2015). The *pilJ* gene is not located within the main or secondary TFP loci in *C. difficile* but is a remote satellite gene (Piepenbrink *et al.,* 2014). The N-terminus of PilJ contained common TFP features including the conserved signal peptide and a hydrophobic α-helix for pilin polymerisation, which were truncated from the crystallised protein (Figure 4.4.3A). PilJ has two well defined domains that each contain the commonly observed TFP structural features including a prominent α1 helix, α-β loop and β-sheet, while only the C-terminal domain contains a D-region (Figure 4.4.3A). Superposition of the C-terminal region of PilJ returned a core RMSD of 2.5 Å (Figure 4.4.3B). These structures show that pilin proteins in *C. difficile* are formed of conserved structural units (Figure 4.4.3B).

A search of the PDB using the peptide sequence of the R20291 (CD3355) and 630 (CD3513) PilA1 proteins retrieved the 2.3 Å resolution structure of a full-length TFP pilin protein known as PilE1 (PDB:2HI2) from *Neisseria gonorrhoeae* (Craig *et al.,* 2006) that shared 39% sequence identity with the *C. difficile* proteins (Figure 4.4.4A). The PilE1 structure represents mature TFP, including the complete hydrophobic N-terminal α-helix responsible for pilin polymerisation that extends away from the globular element of the protein. Superimposition of PilE1 and R20291 PilA1$\triangle$1-34 showed the conservation of structural domains in the headgroup of PilE1 (Figure 4.4.4B). Unlike the Gram-positive TFP structures presented in this chapter, PilE1 contains a disulphide bond that delimits the D-region.

The models of the R20291 and 630 structures were submitted to the PDBe fold (Krissinel and Henrick, 2004) and the DALI (Holm and Rosenström, 2010) servers to identify similar structures in the PDB. PDBe fold search using the R20291 model retrieved the 1.6 Å structure of a truncated pseudopilin known as PulG (PDB:1T92) from a Gram-negative bacterium, *Klebsiella pneumoniae* (Köhler *et al.,* 2004). Pseudopilins are part of the Type II secretion system (TIISS) which are similar in architecture to TFP and export pilin-like

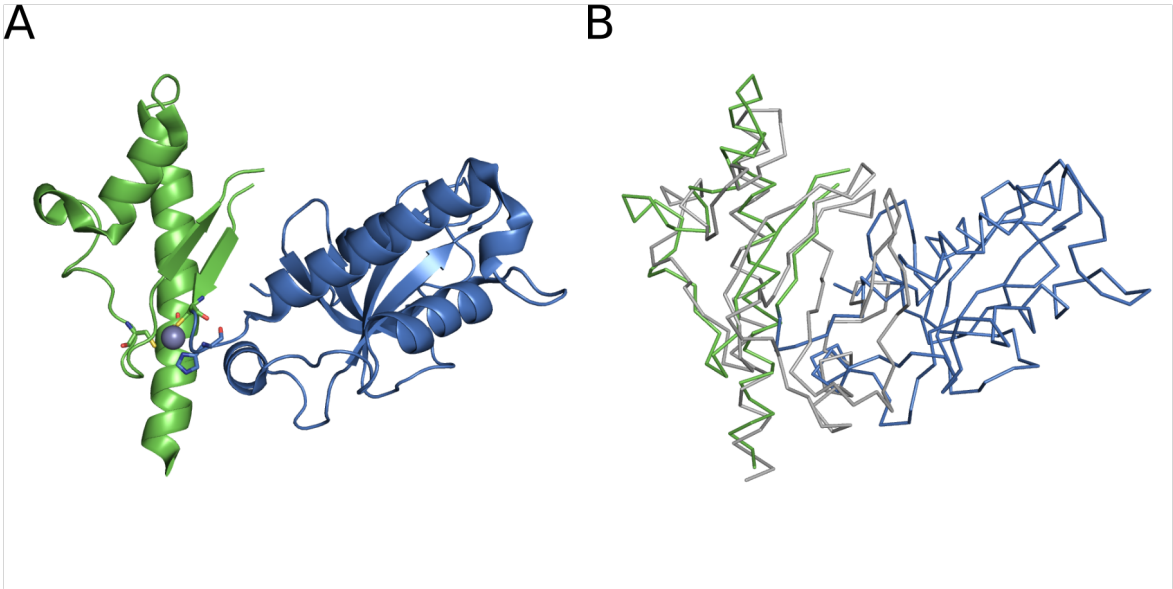


Figure 4.4.3 – Structure of *C. difficile* PilJ. A: Cartoon representation of PilJ structure, the N-terminal domain is shown in green and the C-terminal domain is shown in blue. There is a loop missing between the β-strands that connect the two domains, as electron density did not support model building in this loop region (Piepenbrink *et al.,* 2014). A structural Zn^{2+} was found to be coordinated by three Cys residues (36, 81, 111, coloured green) in the N-terminal domain, plus a single His (114, blue) residue from the C-terminal domain (shown in stick representation) and was predicted to stabilise the interaction between the two domains of PilJ. **B:** Superimposition of R20291 PilA1Δ1-34 structure and the N-terminal domain of the PilJ structure (core RMSD: 2.5 Å).

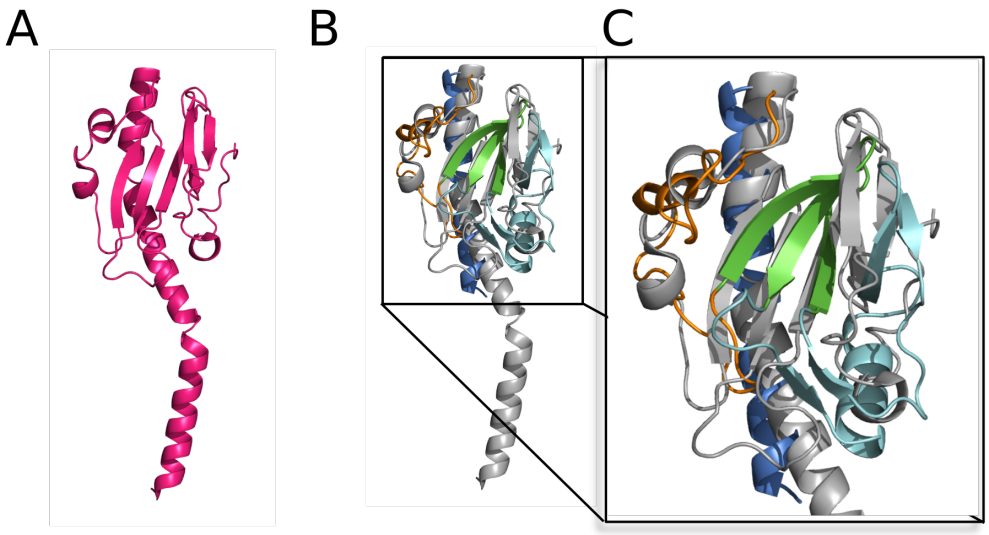


Figure 4.4.4 – Cartoon representation of PilE1 from *Neisseria gonorrhoeae* and superimposition with R20291 PilA1Δ1-34. A: Cartoon representation of the full-length PilE1 from *Neisseria gonorrhoeae* with intact N-terminal α-helix forming an oligomerisation domain with other pilin proteins in the pilin fibre (Craig *et al.,* 2006). **B:** Superposition of PilE1 (grey), excluding the residues 1-25 in the superposition calculation, with the R20291 PilA1Δ1-34 (coloured by structural domain as in Figure 4.3.11) (core RMSD: 4.4 Å).

units, however, they do not readily form filaments as observed in TFP (Korotkov *et al.,* 2012).

Superimposition using SSM of PulG with the R20291 PilA1Δ1-34 structure calculated a core RMSD of 1.6 Å. As observed in the TFP pilin structures determined in this work and discussed here, the pseudopilin PulG contains the characteristic structural regions observed (Figure 4.4.5B). Two molecules were observed in the ASU of the PulG structure, where the C-termini were domain swapped and form an anti-parallel β-sheet with the three β-strands of the partner molecule (Figure 4.4.5A). The swapped C-termini coordinate a Zn^{2+} between them. The dimer form of PulG was only observed in crystal form and the authors of the PulG structure concluded that the dimerisation was a crystallisation artefact (Köhler *et al.,* 2004; Lewis *et al.,* 2000).

Search results using R20291, 630 model and the DALI server returned a number of TFP pilins and pseudopilins from TIISSs. The structures included the major pilins, TcpA from *Vibrio cholerae* (PDBs: 1OQV;3HRV) (Craig *et al.,* 2003), CofA from *Escherichia coli* (PDB:3S0T) (Lim *et al.,* 2010) and PilS from *Salmonella typhi* (PDB:3FHU). The TIISS structures included EspG from *V. cholerae* (PDB:4LW9) and GspG from *E. coli*

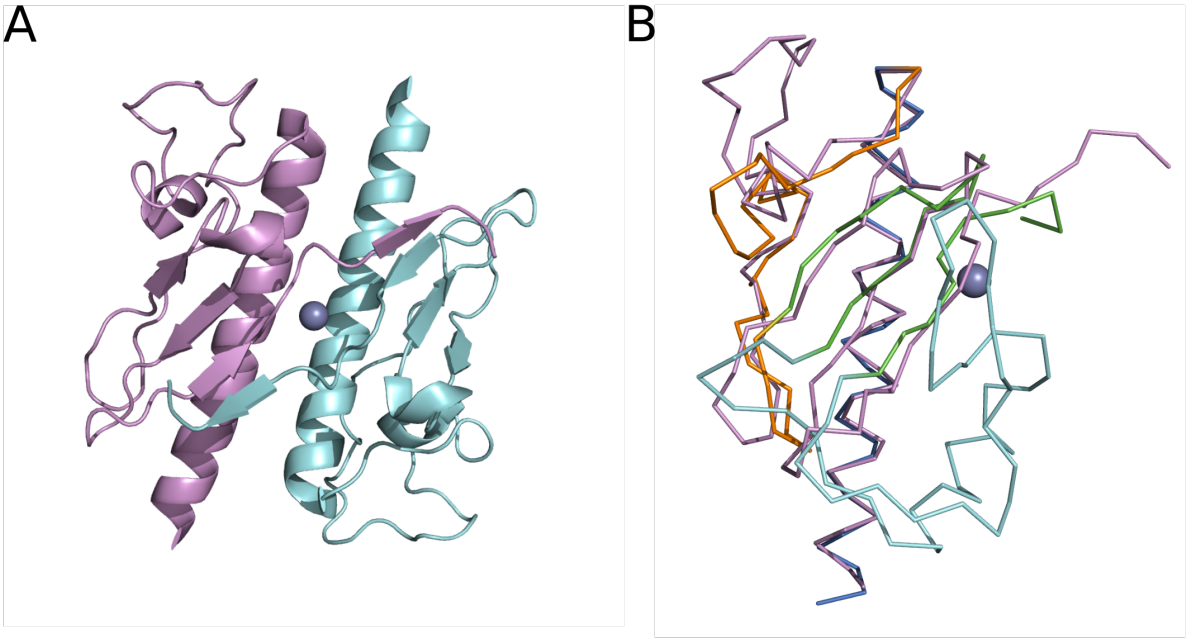


Figure 4.4.5 – Cartoon representation of PulG and SSM superimposition with R20291 PilA1△1-34. A: Cartoon representation of a dimer of PulG from *Klebsiella pneumoniae*. The C-termini of each molecule swap and form an anti-parallel β-sheet with the other molecule in what was likely to be a crystallisation artefact. **B:** SSM superimposition of PulG monomer and R20291 PilA1△1-34, with a core RMSD of 1.6 Å these proteins share 23% sequence identity.

(PDB:3G20) (Korotkov *et al.,* 2009).

All of the structures discussed here have the commonly observed folds of TFP including an N-terminal α-helix (α1), connected to a β-sheet of at least 3 anti-parallel β-strands by an α-β loop, the C-termini are formed of a structurally variable D-region. There are clear similarities in the TFP proteins of Gram-positive proteins with the pseudopilins of the Type II secretion system in Gram-negative bacteria. These structures also contain a disulphide bond between the β-sheet containing region and the variable D-region that is observed in Gram-negative TFP (Giltner *et al.,* 2012). Non-TFP pilins have been observed in Gram-positive bacteria such as *Streptococcus pyogenes* that have pilins that are covalently attached to the cell membrane and are non-retractable (Proft and Baker, 2009). Structures of such Gram-positive pilins have revealed Ig-like structural domains (Proft and Baker, 2009). Additionally, the pilin units in *S. pyogenes* interact via covalent iso-peptide links that enable the formation of a strong helix (Kang *et al.,* 2007). The determination of TFP pilin structures from a Gram-positive bacterium has revealed structural conservation of the TFP secretion system and has shown that these structures in Gram-positive

organisms are very different to the non-TFP Gram-positive pilins.

4.4.2 Major pilin filament formation

Models of filament formation have been proposed for the Gram-negative TFP of *N. meningitidis, N. gonorrhoeae* and *V. choleraee* (Craig *et al.,* 2003; Craig *et al.,* 2006; Craig and Li, 2008; Hartung *et al.,* 2011). The models have been built using crystallographic and NMR structures of the pilin proteins and electron-microscopy data of the overall filament structure to fit the pilin units (Craig and Li, 2008). Such models place the extended, hydrophobic α-helix within the core of the filament, interacting with other pilin units via hydrophobic interactions. The soluble headgroups form the solvent accessible surface of the filament and are responsible for interactions with surfaces (Figure 4.4.6).

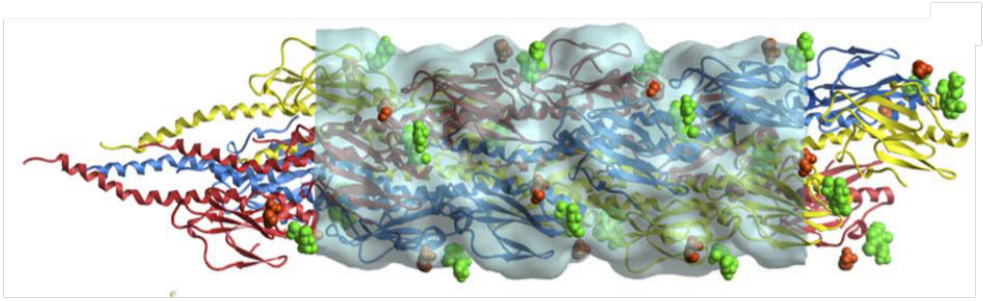


Figure 4.4.6 – ***Neisseria* TFP filament model.** Adapted from Craig et al. 2006. *N. gonnorrhoeae* TFP filament model built by fitting the full-length crystal structure of PilE1 into the cryo-EM density of the filament (Craig *et al.,* 2006). This model also shows filament surface bound molecules (green spheres: disaccharide Gal-DADDGlc; red spheres:phosphoethanolamine).

The structural similarity of the PilA1 structures from *C. difficile* to PilE1 and Tcp of *N. gonorrhoeae* and *V. cholerae*, respectively enabled Piepenbrink et al. to propose a model for PilA1 filament formation in *C. difficile* (Piepenbrink *et al.,* 2015). In this model, they observed a possible salt bridge between PilA1 units between Lys30 and Glu75 that was in a similar position to a salt bridge observed in the TcpA filament model (Arg26:Glu83), proposing that these two residues are important in pilin interface formation within the filament (Figure 4.4.7A). Further, they determined that these residues were well conserved across *C. difficile* strains and observed Lys and Glu residues within PilJ in structurally similar positions (Figure 4.4.7B). Piepenbrink et al. produced charge reversing point mutants (K30E

and E75K) of *pilA1* in *C. difficile* which abrogated the formation of TFP in these mutants. Although a double mutant was also unable to produce TFP, they conclude this was due to an imperfect recreation of the interaction (Piepenbrink *et al.,* 2015).

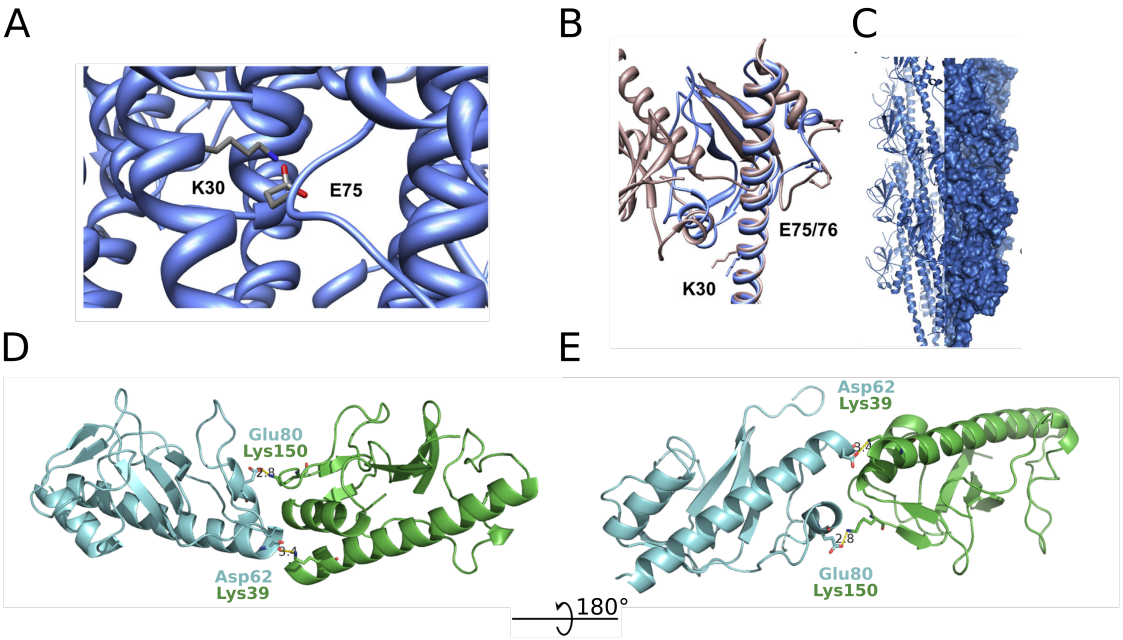


Figure 4.4.7 – Piepenbrink et al. PilA1 filament model and 630 PilA1△1-34 interface. Adapted from Piepenbrink et al. Possible PilA1 filament formation was modelled based upon data of the *V. cholerae* TFP model (C). From the model they proposed that K30 and E75 formed a salt-bridge between PilA1 monomers within the filament (A) and this was structurally conserved in PilJ (B). Point mutants in *C. difficile* PilA1 that altered the charge at these residue positions could not form TFP. Interface analysis of the 630 PilA1△1-34 structure highlighted a possible interface between molecules in which they were arranged in the same orientation (D/E).

Analysis of the PilA1△1-34 structures from R20291 and 630 determined in this project using PISA did not reveal a salt-bridge in the positions proposed by Piepenbrink et al., 2015. The structures we have determined have the advantage that they only interface with other PilA1 molecules, unlike the Piepenbrink structures that also contain contacts with MBP and as a result they are unable to observe inter-PilA1 contacts on all interfaces of the PilA1 molecules. Moreover, analysis of the 630 structure determined an interface that contained 2 salt bridges (Asp62:Lys39, Glu80:Lys150) and allowed the molecules to be arranged in such a way that could be compatible with filament formation (Figure 4.4.7D/E). Further work is required both *in vitro* and *in vivo* to determine whether these salt-bridges are significant in TFP filament formation.

4.4.3 Comparison of protein structure determination methods

The methods used in this project and by Piepenbrink et al. in the determination of the structures of PilA1 from a number of *C. difficile* strains have been quite different. Piepenbrink et al. reported that an MBP fusion to PilA1 was required for solubilisation. They excluded the first 34 residues of the open reading frames of all their PilA1 constructs, as was the case in the constructs used in our project and His-tagged the construct at the C-terminus. The His-tag of the PilA1$\triangle$1-34 constructs used here were positioned at the N-terminus and were non-cleavable, however, we did not experience any issues regarding solubility of these constructs.

During structure determination Piepenbrink et al. were able to use the structure of maltose binding protein to determine phases using Phaser and to build the PilA1 structures (Piepenbrink *et al.,* 2015). Despite collecting high resolution data of R20291 PilA1$\triangle$1-34, it was not possible to calculate phases and an interpretable electron density maps using crystal structures of PilE1 as a search model (39% sequence identity). Since R20291 PilA1$\triangle$1-34 crystals were readily reproducible, heavy atom soaking and experimental phasing was pursued and it was possible to calculate phases from a Pb-containing derivative crystal and determine the structure of R20291 PilA1$\triangle$1-34 to a resolution of 1.65 Å with good figures of merit (FOM: 0.71 ;CC: 77%). The refinement of the R20291 model resulted in an R_{work} of 14.6% and R_{free} of 19.1%, which were slightly better (1%) than those achieved by Piepenbrink et al. At the same time, phases were calculated using Arcimboldo, a relatively recent *ab initio* phasing method that combines substructure determination using alpha helical fragments using Phaser and density modification using SHELXE (Millan *et al.,* 2015). The final mapCC using Arcimboldo was 47.2%. Arcimboldo traced 395 residues during *ab initio* phasing whilst Buccaneer built 411 using experimental derived phases. The native dataset of R20291 PilA1$\triangle$1-34 was of a high resolution (1.65 Å), high completeness, low R_{merge} and high I/σI and as a result was an ideal candidate for *ab initio* phasing using Arcimboldo (Rodríguez *et al.,* 2009). Even though the phases provided by Arcimboldo were not used to build a model which was refined, Arcimboldo provides a useful alternative to experimental phasing or molecular replacement that can be biased by the search model if crystal data is of suitable quality.

Chapter 5

Structural studies of minor pilin proteins from Type IV pili in *C. difficile*

5.1 Introduction

Type IV pili (TFP) are fibre-like appendages that extend from the cell and provide the cell with twitching abilities (Varga *et al.,* 2006). The pili are predominantly formed by the major pilin PilA1 (Chapter 4), however, they are also decorated with pilin proteins known as minor pilins. *C. difficile* has three minor pilin genes in the major TFP locus: *pilV*, *pilU* and *pilK* (Melville and Craig, 2013). Pilin units are formed of an N-terminal hydrophobic stem that enables polymerisation and a C-terminal globular head-group that is on the exterior of the pilin fibre. Mature PilU and PilV proteins are 166 residues and 178 residues in length respectively, which is a similar size to the mature major pilin, PilA1 (164 residues) (Maldarelli *et al.,* 2014).

PilK differs from the other minor pilin proteins in that it is the only pilin protein in the major locus that does not have a conserved pilin leader sequence. Glu5 is substituted for Leu (Figure 5.1.1), and unlike the other pilins, the N-terminal residue of mature PilK is Ala. At 500 residues, mature PilK protein is also much larger than the other pilins discussed here. A single large minor pilin has been identified in the TFPa loci of a number of Gram-positive and pseudopilin loci of Gram-negative species, and has been grouped in the GspK family of proteins that have been shown to cap pseudopili in TIISS (Type II secretion system) (Korotkov and Hol, 2008).

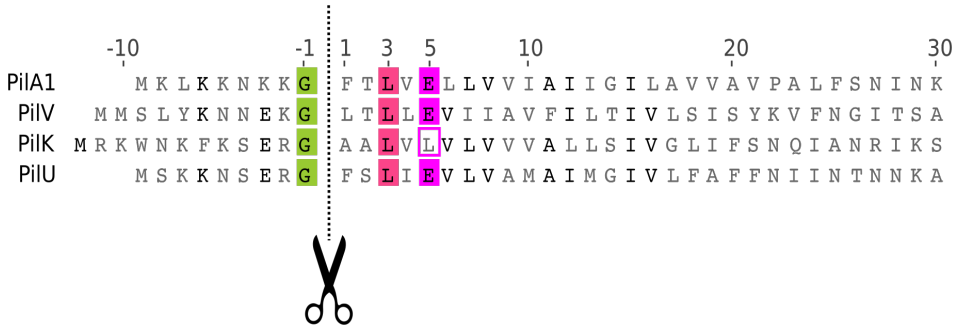


Figure 5.1.1 – Organisation of TFP proteins and signal peptide. The type-a N-terminal signal peptide sequences of PilA1, PilV, PilK and PilU aligned using ClustalO. Residues are numbered by their position in the mature pilin protein and bold residues are at least partially conserved. The conserved Gly-1 at which cleavage by the pre-pilin peptidase PilD occurs is highlighted in green. The first residue of mature PilA1 and PilU is Phe1 and is conserved amongst TFPa proteins. Leu3 is conserved across all 4 pilin proteins. Glu5 is conserved in all but PilK.

The specific roles of minor pilins within the TFP assembly are unclear, particularly in Gram-positive bacteria. The functions of a number of TFP minor pilins in Gram-negative bacteria have been identified, including DNA binding (Cehovin *et al.,* 2013), cell-adhesion (Helaine *et al.,* 2005) and filament control (Szeto *et al.,* 2011). Particularly interesting is the role of PilK, which has a non-conserved signal peptide, and is three-fold larger than the other minor pilins. A hypothesis for the role of PilK is as an initiator pilin unit which is localised at the tip of the pilus (Melville and Craig, 2013).

The aim of the work in this chapter was the structural characterisation of the minor pilins PilU, PilV and PilK. Some preliminary work into pilus assembly was also carried out to understand whether PilU, PilV or PilK interact with the major pilin, PilA1.

5.2 Expression and purification of minor pilins

All of the constructs used in this chapter were produced by Edward Couchman in Professor Neil Fairweather's group at Imperial College, London, as part of an ongoing collaboration. The pilin signal peptide sequence, identified using PilFind (Imam *et al.,* 2011), and the N-terminal hydrophobic residues that form the polymerisation domain, were not included in these constructs. The truncated open reading frames were inserted into the expression vector pET-28a such that the expressed proteins were tagged at the C-terminal with an uncleavable 6-His-tag, linked to the protein by a Leu-Glu linker.

The proteins were expressed in Rosetta *Escherichia coli* DE3 cells, as described in Section 2.4. The expressed proteins were purified by nickel affinity purification followed by size exclusion chromatography (SEC) using a calibrated Superdex 75 16/600 column.

5.2.1 PilV

The PilV$\triangle$1-35 construct (pECC74) contained residues 36-189 of the open reading frame from the *pilV* gene CD3508 (Figure 5.2.1A). Three peaks eluted from the SEC at peak elution volumes of ~46, ~60 and ~71 ml and SDS-PAGE analysis showed that a single protein that had migrated between the 15 and 25 kDa markers was present in all fractions analysed (Figure 5.2.2B). The theoretical mass of PilV$\triangle$1-35 is 19.1 kDa (ProtParam). These results indicated that PilV$\triangle$1-35 was able to form multimeric species as well as a monomer. Elution of PilV$\triangle$1-35 at volumes equivalent to 21 kDa and 52 kDa suggest that the protein was present in monomer and dimer or trimer species, respectively. The remaining peak eluted at the void volume of the S75 SEC column (46.5 ml), suggesting that the sample contained protein aggregates or higher oligomeric states. There was no indication that the protein was in a dynamic equilibrium between monomer and dimer or trimer species, as would have been evidenced by a broader peak with a shoulder, suggesting that the monomers and multimers are stable entities, given the timescale of the analyses. The final yield of monomer PilV$\triangle$1-35 was 4.5 mg/L cell culture and a dimer yield of 2.3 mg/L cell culture.

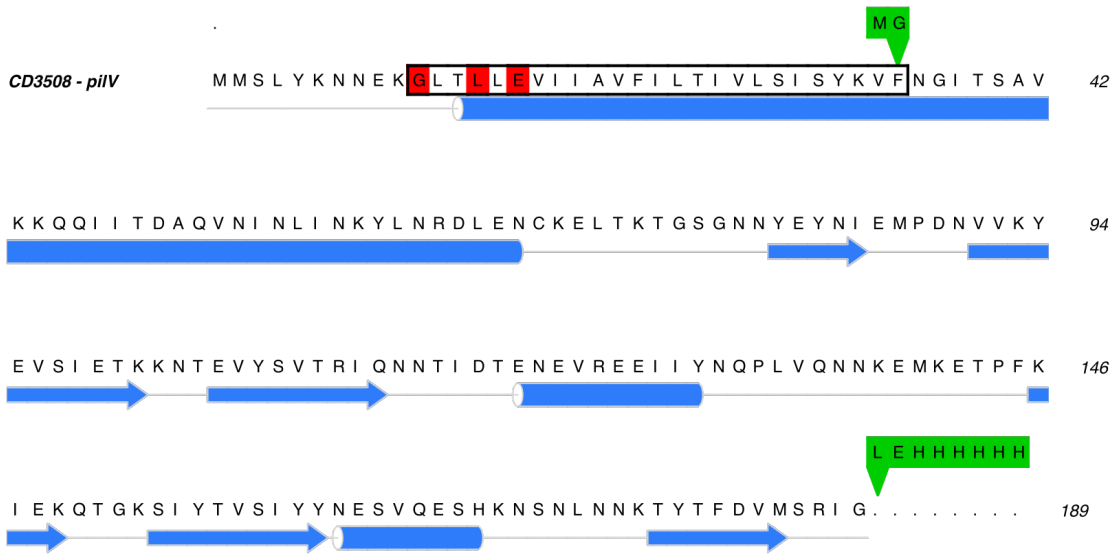


Figure 5.2.1 – PilV△1-35 construct design. Open reading frame of *pilV* (CD3508), the predicted polymerisation domain is highlighted in black, the conserved Gly, Leu and Glu residues of the pilin leader sequence are coloured red. PSIPRED secondary structure prediction is annotated below the peptide sequence. The limits of the PilV△1-35 construct are described by the sequences highlighted in green, including a C-terminal 6-His-tag. The PilV△1-35 construct has a calculated mass of 19.1 kDa (ProtParam)

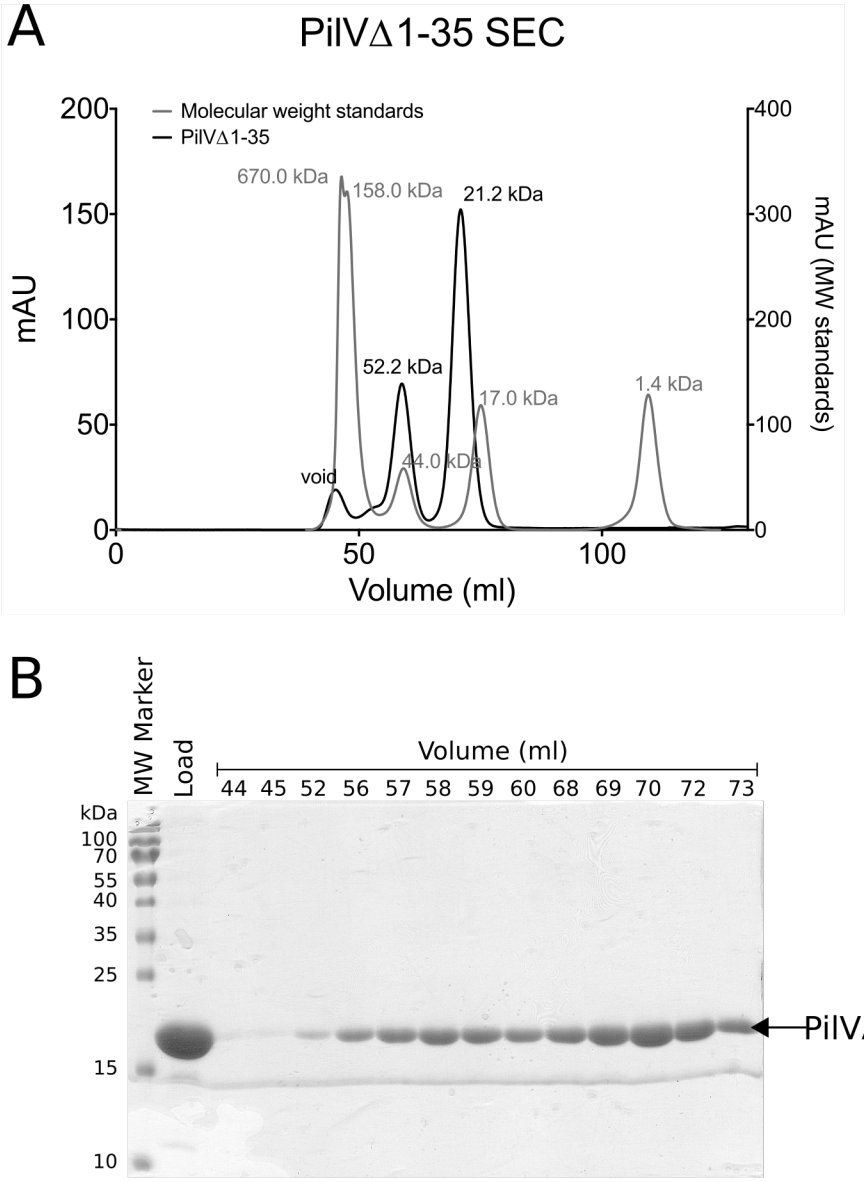


Figure 5.2.2 – **PilVΔ1-35 purification. A:** The UV trace from size-exclusion chromatography (SEC) of PilVΔ1-35 (black) after initial nickel affinity purification. A Superdex 75 16/600 column was used and the elution profile of molecular mass marker proteins are shown in grey. PilVΔ1-35 elutes from the S75 column in two peaks at volumes equivalent to masses of 21.2 kDa and 52.2 kDa indicating the presence of monomer and dimer forms of PilVΔ1-35. A peak is also observed at the void volume. **B:** SDS-PAGE of the injected sample and peak elution fractions from SEC. Fractions from the void (44-45 ml), 52.2 kDa peak (56-60 ml) and 21.2 kDa peak (68-73 ml) all contain a protein that has migrated to a distance equivalent to ~20 kDa.

5.2.2 PilU

The PilU construct excluded 33 residues at the N-terminus and is referred to as PilU△1-33 (encoded on pECC73) (Figure 5.2.3A). A single elution peak was observed the S75 column used, at an elution volume equivalent to 19.1 kDa. The theoretical mass of PilU△1-33 is 17.4 kDa (ProtParam), indicating that PilU△1-33 was in a monomeric form in solution. The final yield of pure PilU△1-33 was ~15 mg/L cell culture.

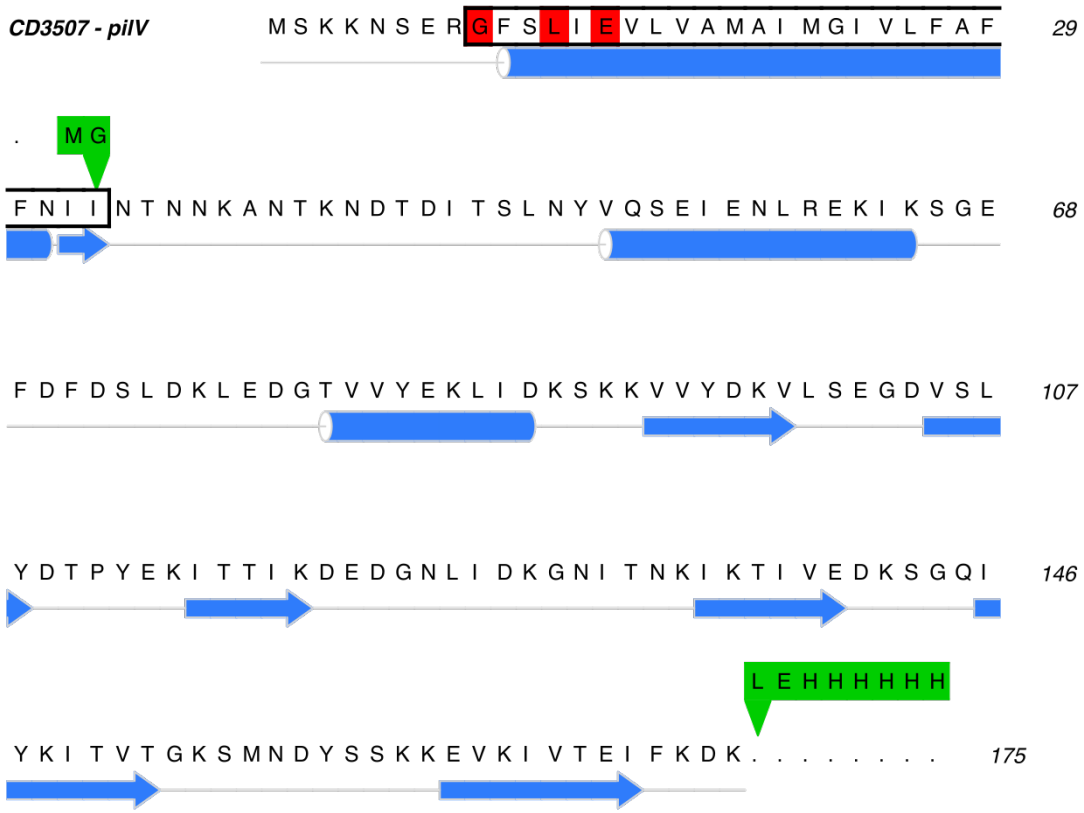


Figure 5.2.3 – :**PilU△1-33 construct design.** Open reading frame of *pilU* (CD3507), the predicted polymerisation domain is highlighted in black, the conserved Gly, Leu and Glu residues of the pilin leader sequence are coloured red. PSIPRED secondary structure prediction is annotated below the peptide sequence. The limits of the PilU△1-33 construct (pECC73) are described by the sequences highlighted in green, including a C-terminal 6-His-tag. The PilU△1-33 construct has a calculated mass of 17.4 kDa (ProtParam).

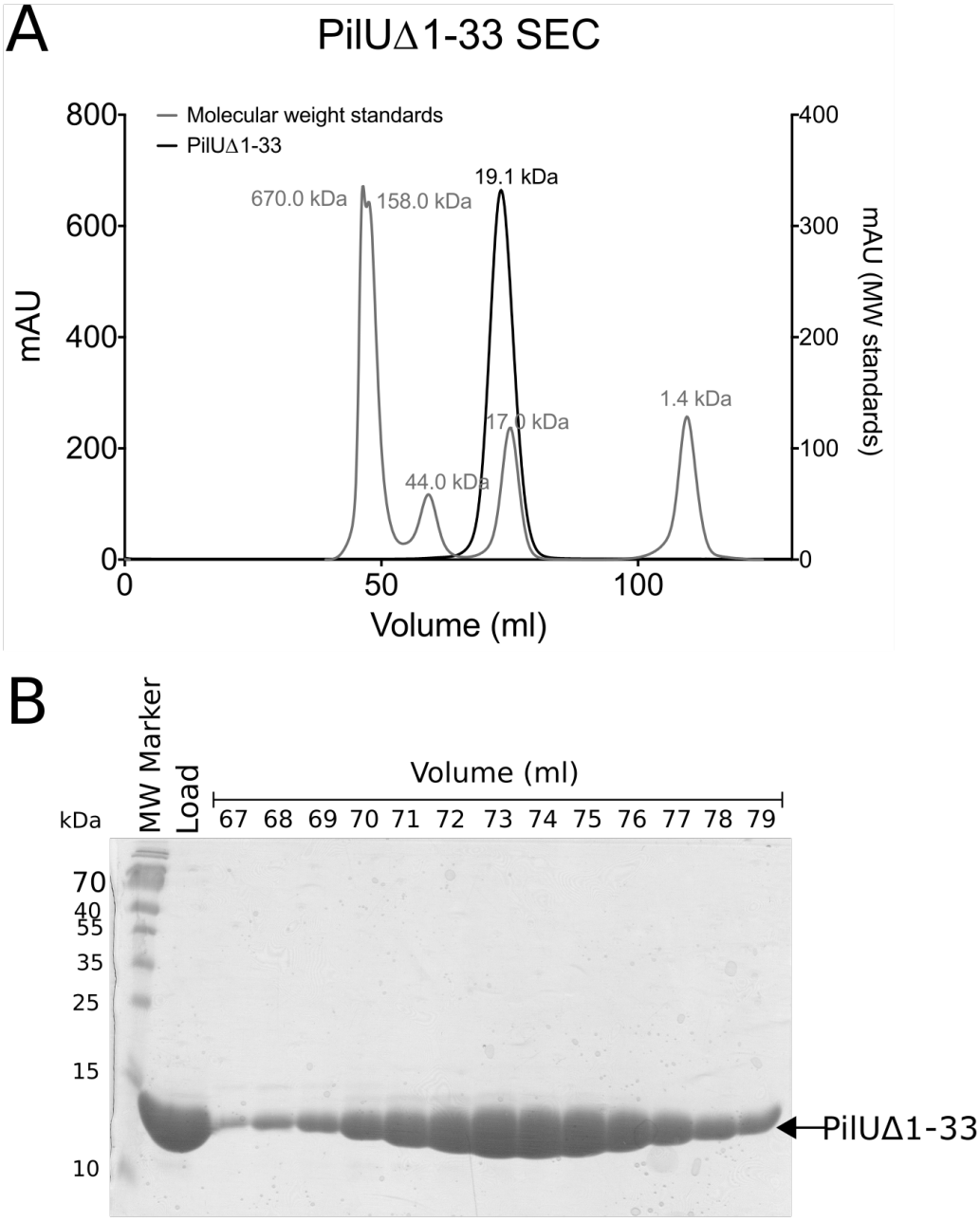


Figure 5.2.4 – **PilUΔ1-33 purification. A:** The UV trace from size-exclusion chromatography (SEC) of PilUΔ1-33 (black) after nickel affinity purification. A Superdex 75 16/600 column was used and the elution profile of molecular mass marker proteins are shown in grey. PilUΔ1-33 elutes from the S75 column in a single peak at a volume equivalent to 19.1 kDa indicating a monomeric species. **B:** SDS-PAGE of the injected sample and peak elution fractions from SEC. Peak elution fractions contain a single protein that has migrated to a distance equivalent to ~20 kDa.

5.2.3 PilK

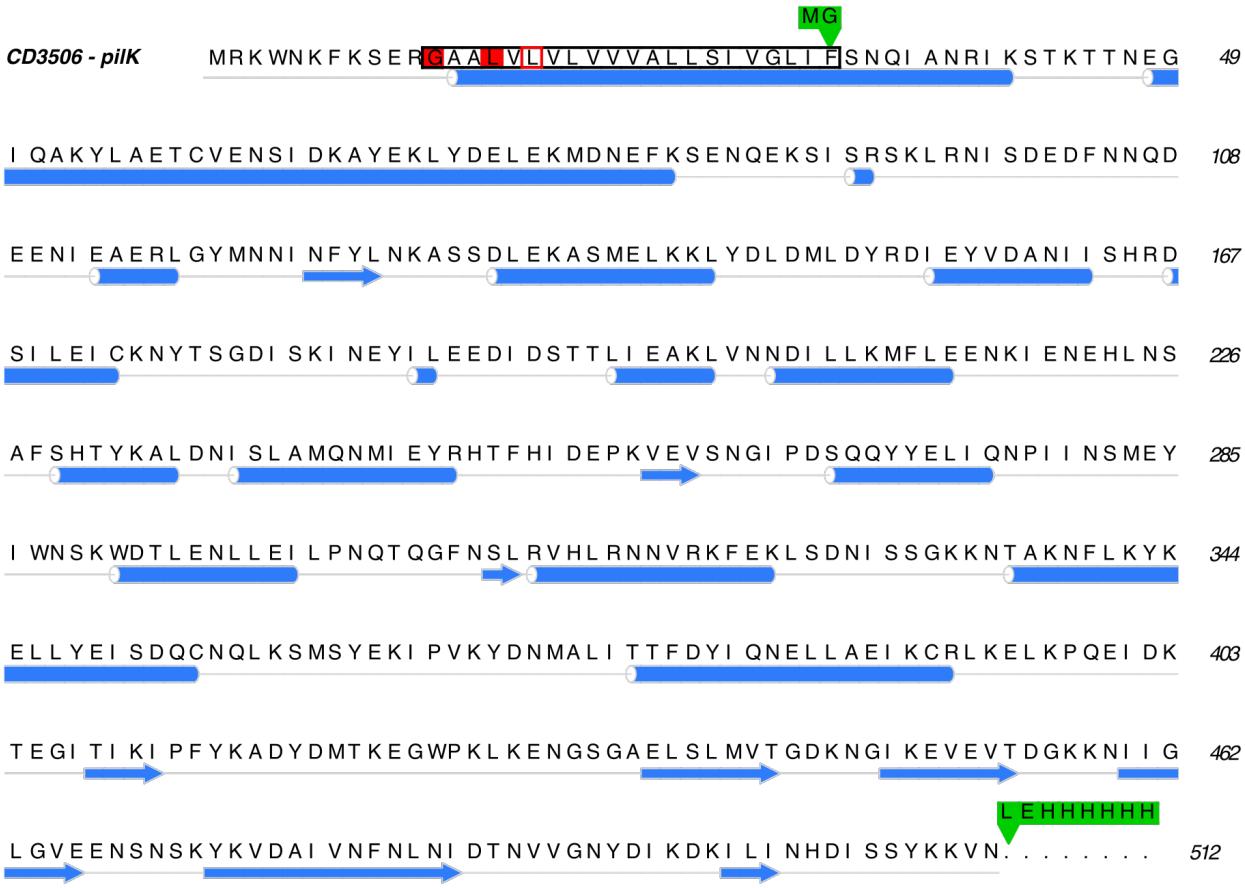


Figure 5.2.5 – PilKΔ1-32 construct design. The open reading frame of *pilK* from the gene CD3506 from strain 630 of *C. difficile*. The conserved pilin recognition residues are shaded red, the non-conserved position of this sequence is outlined in red. The truncated hydrophobic polymerisation domain is highlighted in black. Additional residues from the pET-28a vector are aligned above the sequence and shaded in green, the arrow indicating where these insert into the open reading frame of PilK to produce the PilKΔ1-32 truncation (pECC75) with a C-terminal 6-His-tag. The theoretical mass of PilKΔ1-32 is 56.9 kDa. The PilK peptide sequence has been annotated with PSIPRED secondary structure prediction (blue).

The truncated PilK construct, PilKΔ1-32 (encoded on pECC75) is illustrated in Figure 5.2.5. Two elution peaks were observed in the SEC (Figure 5.2.6C) and SDS-PAGE analysis indicated the presence of PilKΔ1-32 in each peak (Figure 5.2.6D).Two significant bands were observed at migration distances equivalent to 55 kDa and 25 kDa. The band at 25 kDa was exclusive to fractions from the higher molecular mass peak. Comparison with molecular mass protein standards suggested that the protein's molecular mass in these SEC peaks was 70.0 kDa and 120.9 kDa; the theoretical mass of PilKΔ1-32 is 56.9 kDa. Such an elution profile indicated the likely presence of monomeric and dimeric

forms of PilK△1-32. The fractions from the two peaks, LMW (C, yellow shaded region) and HMW (C, blue shaded region) were pooled separately and injected onto the SEC as HMW and LMW (see Figure 5.2.7). Degradation of PilK△1-32 was also observed in the SDS-PAGE analysis; more degradation or impurities were observed in the fractions from the higher molecular mass peak than in the monomer peak. Intriguingly, a band at an apparent molecular mass of 25 kDa is observed in the SDS-PAGE analysis of the higher molecular mass peak, but not in the 70 kDa peak. To understand whether this apparent degradation and oligomeric state were stable, the fractions from each peak were pooled and reanalysed separately (Figure 5.2.3E/G). The pooled fractions of the higher molecular mass peak resulted in a peak at an elution volume equivalent to 129 kDa and a similar pattern of bands were observed in the SDS-PAGE. Meanwhile the low molecular mass peak fractions resulted in an elution peak equivalent to 75 kDa, again indicating a monomeric species (Figure 5.2.7C). A shoulder was also observed on this peak at volumes of a higher molecular mass, similar to the profile observed in Figure 5.2.3A. The SDS-PAGE showed a decreased level of degradation and no clear bands ~ 25 kDa. The shoulder contains protein of the same apparent molecular mass on SDS-PAGE as the monomeric peak, suggesting that PilK△1-32 can form dimers.

To understand the content of the most significant bands observed in the SDS-PAGE analyses, LC/MS/MS was performed on the ~55 kDa and ~25 kDa bands (Figure 5.2.7). The use of LC/MS/MS enabled the isolation of specific peptides and subsequent identification of the proteins present in each band. The band at ~55 kDa contained residues 52-510, representative of 93 % of the peptide sequence of PilK△1-32 (Figure 5.2.8B). The band at ~25 kDa covered residues 52-211, representing the N-terminal region (32%) of the peptide sequence of PilK△1-32 (Figure 5.2.8C).

The SEC and SDS-PAGE in Figure 5.2.6C/D suggested PilK△1-32 was present in monomer and dimer species. Intriguingly, the SEC, SDS-PAGE and LC/MS/MS of the HMW sample (Figure 5.2.8A) suggest that PilK△1-32 can form a stable interaction with its degradation product, specifically the region between residues 52-211.

The final yields of the PilK△1-32 purifications were 8 mg/L cell culture of PilK△1-32 monomer (LMW) and 9 mg/L cell culture PilK△1-32 HMW sample.

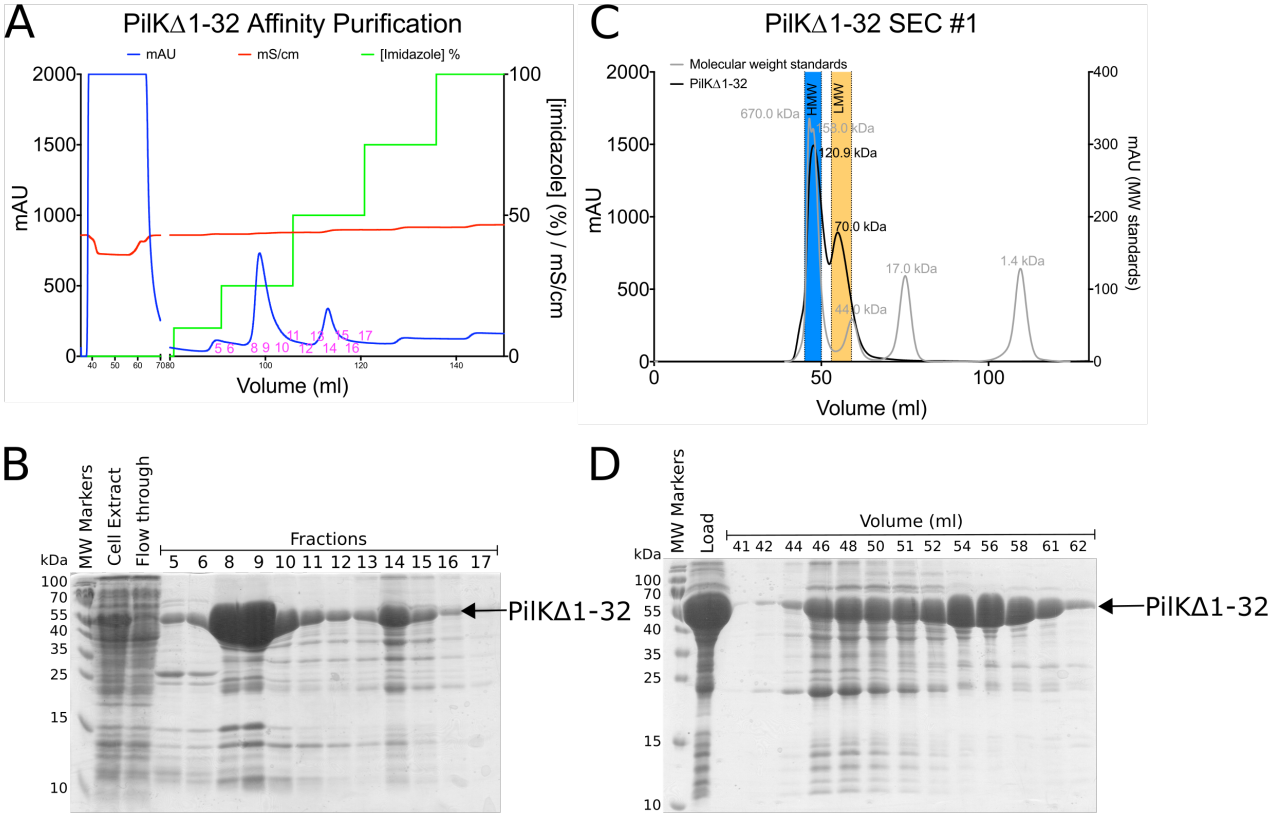


Figure 5.2.6 – PilKΔ1-32 purification. After lysis and clarification, PilKΔ1-32 was purified using affinity purification by Ni-NTA chromatography (A), the cell extract, flow-through and peak elution fractions were analysed by SDS-PAGE (B). The majority of PilKΔ1-32 eluted from the Ni-NTA column in 125 mM imidazole and a smaller quantity eluted in 250 mM imidazole. SDS-PAGE showed the peak elution fractions from the affinity purification were impure. Further purification by size-exclusion chromatography (SEC) (C) using a Superdex 75 column [GE Healthcare] resulted in two elution peaks equivalent to 70.0 kDa (LMW: low molecular mass) and 120.9 kDa (HMW: high molecular mass). The theoretical mass of PilKΔ1-32 is 56.9 kDa, indicating the presence of monomer and dimer species in these peaks. The molecular mass standards (grey) were thyroglobulin (670.0 kDa), γ-globulin (158.0 kDa), ovalbumin (44.0 kDa), myoglobin (17.0 kDa) and Vitamin B_{12} (1.4 kDa). SDS-PAGE of the elution peaks from the S75 indicated impurities and degradation. Two significant bands were observed at migration distances equivalent to 55 kDa and 25 kDa.

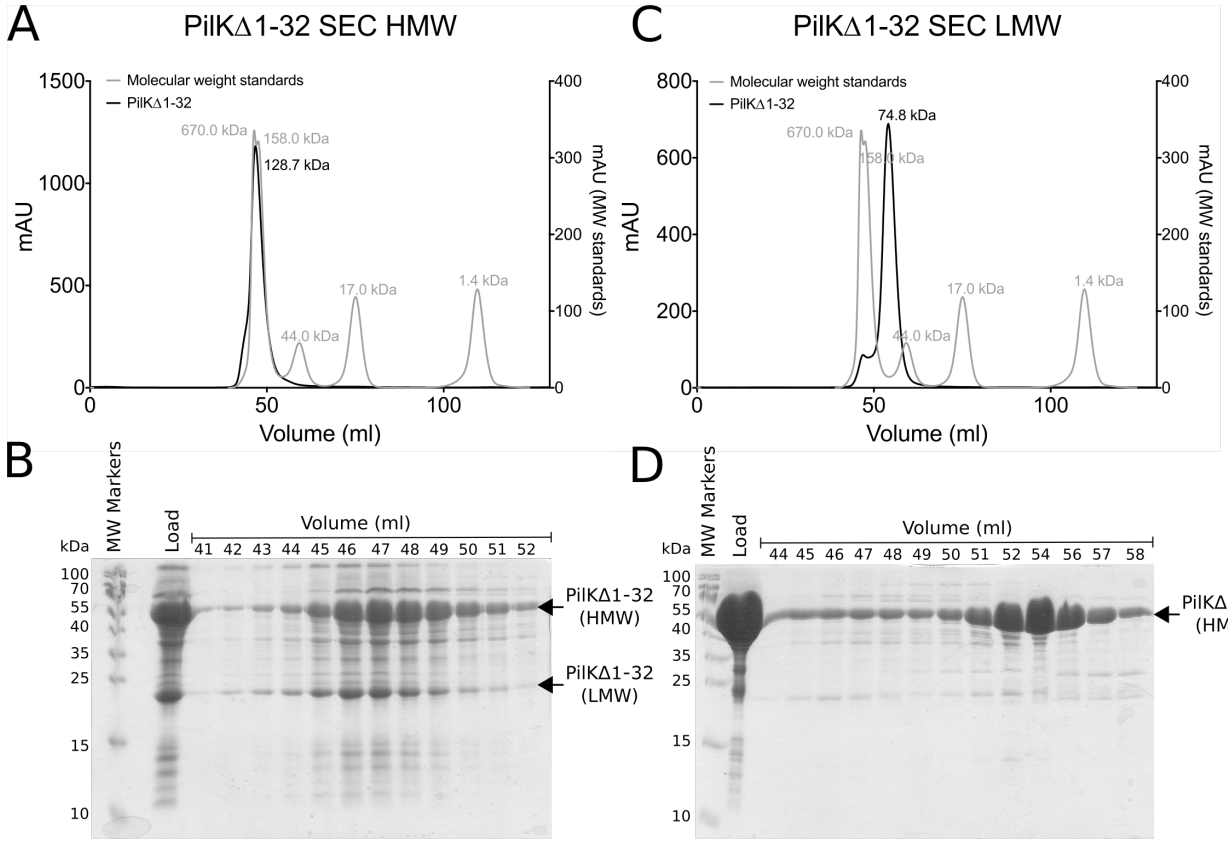
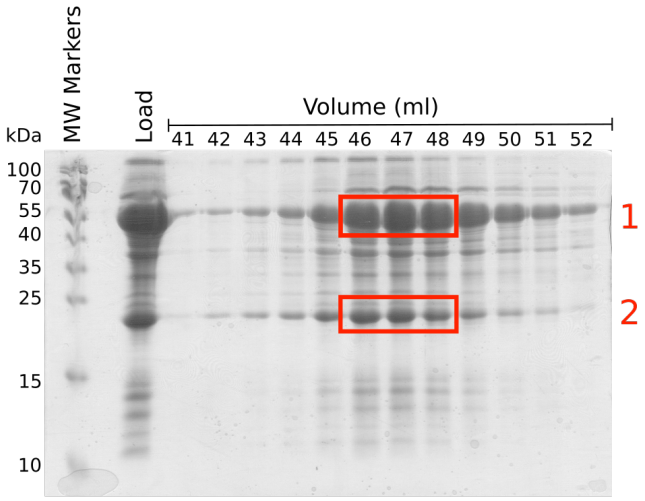


Figure 5.2.7 – **Analysis of PilKΔ1-32 MW species.** The two significant peaks observed in the SEC of PilKΔ1-32 (Figure 5.2.6C) were individually pooled and re-run on the SEC. The samples were labelled HMW (blue shaded region, Figure 5.2.6C) and LMW (yellow shaded region, Figure 5.2.6C). The HMW sample resulted in a single peak of 128.7 kDa and bands were observed at 55 kDa and 25 kDa in the SDS-PAGE of the peak elution fractions (B). The LMW sample (C) resulted in a single peak with a shoulder at lower volume edge. The equivalent mass of the peak was 74.8 kDa and the shoulder was ~122 kDa. SDS-PAGE analysis (D) showed the presence of a protein that had migrated at a distance equivalent to 55 kDa in the fractions of both the peak and shoulder. The molecular mass standards (grey) were thyroglobulin (670.0 kDa), γ-globulin (158.0 kDa), ovalbumin (44.0 kDa), myoglobin (17.0 kDa) and Vitamin B_{12} (1.4 kDa).

A



B

```
  1 MRKWNKFKSE RGAALVLVLV VVALLSIVGL IFSNQIANRI KSTKTTNEGI
 51 QAKYLAETCV ENSIDKAYEK LYDELEKMDN EFKSENQEKS ISRSKLRNIS
101 DEDFNNQDEE NIEAERLGYM NNINFYLNKA SSDLEKASME LKKLYDLDML
151 DYRDIEYVDA NIISHRDSIL EICKNYTSGD ISKINEYILE EDIDSTTLIE
201 AKLVNNDILL KMFLEENKIE NEHLNSAFSH TYKALDNISL AMQNMIEYRH
251 TFHIDEPKVE VSNGIPDSQQ YYELIQNPII NSMEYIWNSK WDTLENLLEI
301 LPNQTQGFNS LRVHLRNNVR KFEKLSDNIS SGKKNTAKNF LKYKELLYEI
351 SDQCNQLKSM SYEKIPVKYD NMALITTFDY IQNELLAEIK CRLKELKPQE
401 IDKTEGITIK IPFYKADYDM TKEGWPKLKE NGSGAELSLM VTGDKNGIKE
451 VEVTDGKKNI IGLGVEENSN SKYKVDAIVN FNLNIDTNVV GNYDIKDKIL
501 INHDISSYKK VN
```

C

```
  1 MRKWNKFKSE RGAALVLVLV VVALLSIVGL IFSNQIANRI KSTKTTNEGI
 51 QAKYLAETCV ENSIDKAYEK LYDELEKMDN EFKSENQEKS ISRSKLRNIS
101 DEDFNNQDEE NIEAERLGYM NNINFYLNKA SSDLEKASME LKKLYDLDML
151 DYRDIEYVDA NIISHRDSIL EICKNYTSGD ISKINEYILE EDIDSTTLIE
201 AKLVNNDILL KMFLEENKIE NEHLNSAFSH TYKALDNISL AMQNMIEYRH
251 TFHIDEPKVE VSNGIPDSQQ YYELIQNPII NSMEYIWNSK WDTLENLLEI
301 LPNQTQGFNS LRVHLRNNVR KFEKLSDNIS SGKKNTAKNF LKYKELLYEI
351 SDQCNQLKSM SYEKIPVKYD NMALITTFDY IQNELLAEIK CRLKELKPQE
401 IDKTEGITIK IPFYKADYDM TKEGWPKLKE NGSGAELSLM VTGDKNGIKE
451 VEVTDGKKNI IGLGVEENSN SKYKVDAIVN FNLNIDTNVV GNYDIKDKIL
501 INHDISSYKK VN
```

Figure 5.2.8 – Sequence coverage of PilK fragments determined by LC/MS/MS. The two significant bands (A) that migrated a distance equivalent to 55 kDa (1) and 25 kDa (2) were excised and analysed by LC/MS/MS. The 55 kDa band (1) was shown to represent residues 52-510, the majority of PilKΔ1-32 (B, sequence in red). The 25 kDa band (2) was determined to represent residues 52-211 (C, sequence in red). The LC/MS/MS was performed and analysed by Dr Joe Gray, Pinnacle Laboratory, Newcastle University.

5.3 Biophysical Characterisation of pilin proteins

To ensure that the recombinant proteins that were purified and subsequently crystallised were correctly folded, circular dichroism (CD) scans, variable temperature CD spectroscopy and differential scanning fluorimetry (DSF) experiments were carried out.

5.3.1 PilV

The CD spectrum of PilV$\triangle$1-35 is presented in Figure5.3.1A, showing that PilV$\triangle$1-35 is folded. It was not possible to measure a CD spectrum below a wavelength of 195 nm that did not cause high tension overload, despite adjusting the protein concentration. A wavelength range of 260-190 nm is the minimum required for CD deconvolution using the Dichroweb tools (Whitmore and Wallace, 2004). Therefore, quantification of the secondary structure content in PilV using CD deconvolution was not possible.

The thermal stability of PilV$\triangle$1-35 protein was also probed using CD spectroscopy at 220 nm and a melting temperature (T_m) of 64.5° C was calculated from the CD spectrum in Figure 5.3.1B. The T_m value was verified using differential scanning fluorimetry (DSF), which calculated a melting temperature of 59° C (Figure 5.3.1C) - a value within 5° C of the CD determined value.

Despite it not being possible to calculate the proportions of secondary structure, the thermal stability assays show that PilV$\triangle$1-35 is likely to contain secondary structure. If PilV$\triangle$1-35 did not, it is unlikely that protein would have a T_m of ~60° C since intrinsicaly disordered proteins cannot unfold.

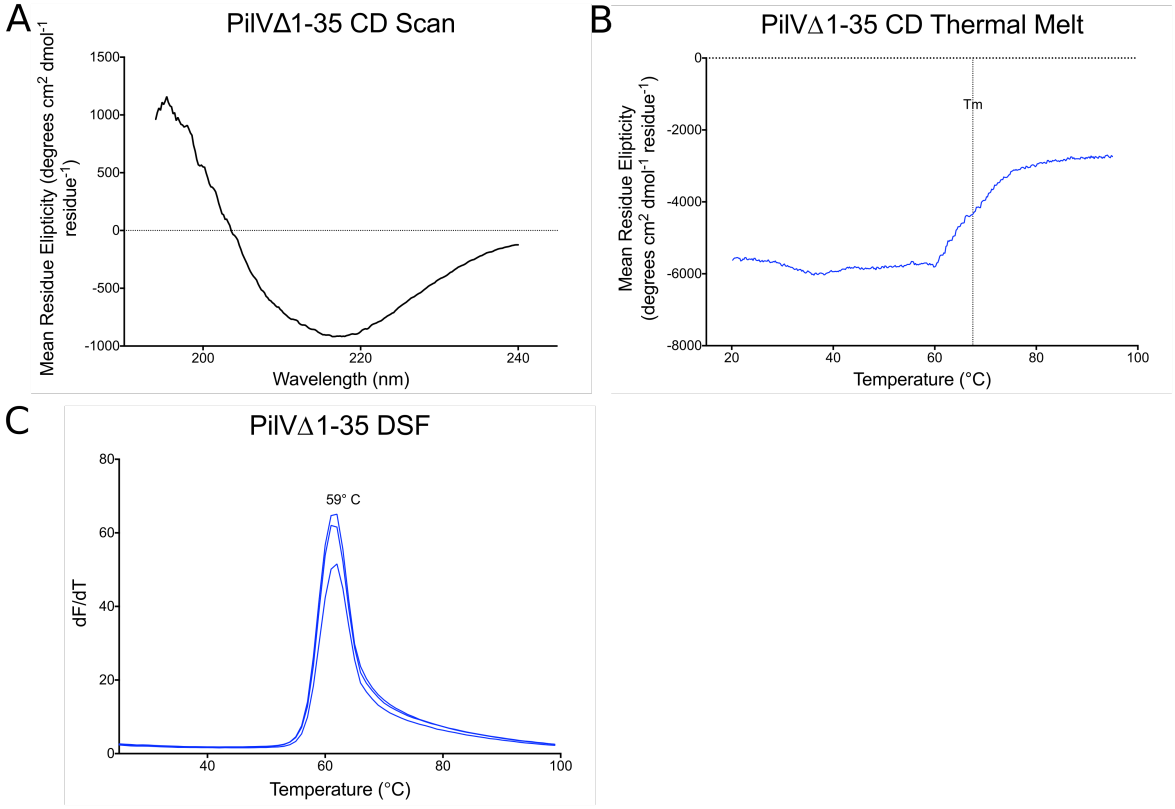


Figure 5.3.1 – CD spectroscopy and DSF of PilVΔ1-35. A: CD spectrum of PilVΔ1-35 measured between 240-195 nm at 20° C. The wavelength range of the scan was limited by the sample quality and as such it was not possible to deconvolute the data. **B:** CD was measured at 220 nm over a temperature gradient of 20-95° C (+1° C min^{-1}) and a T_m of 64.5° C (dotted line) was calculated from these data. **C:** Triplicate DSF were performed of PilVΔ1-35 during a temperature gradient of 25-95° C (+1° C min^{-1}). A T_m of 59° C was calculated.

5.3.2 PilU

A CD spectrum of PilU△1-33 was measured (Figure 5.3.2A), however it was not possible to measure a greater wavelength range than 240-195 nm. The limited wavelength range meant it was not possible to deconvolute the data using the Dichroweb server (Whitmore and Wallace, 2004). The spectrum contained minima at 220 and 208 nm indicating the presence of alpha-helical structure. The thermal stability of PilU△1-33 was measured using CD spectroscopy at 208 nm since the peak at this wavelength was better defined than at other wavelengths (Figure 5.3.2B). A T_m value of 65.3° C was calculated from these data. It was not possible to measure the T_m of PilU△1-33 using DSF because there was no change in fluorescence during the assay.

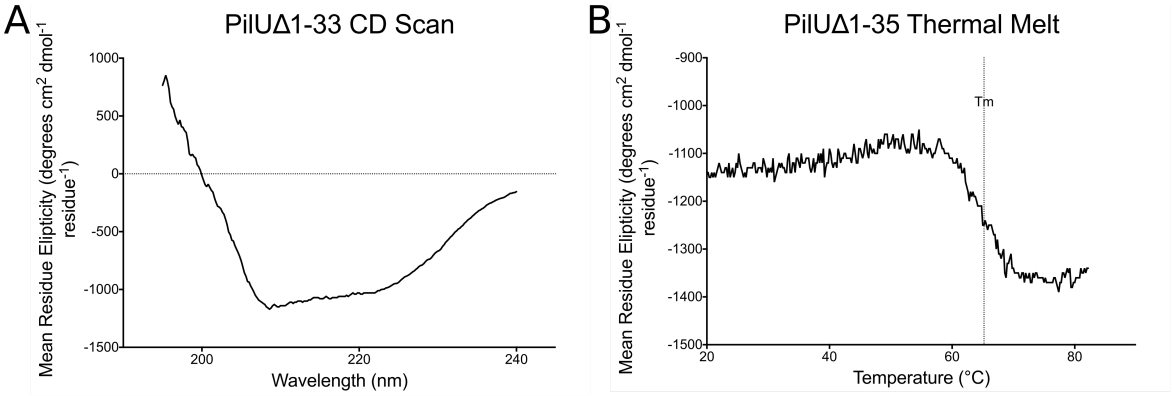


Figure 5.3.2 – CD spectroscopy of PilU△1-33. A: CD scan of PilU△1-33 between 240-195 nm at 20° C, with features at 208 and 220 nm indicating the presence of α-helical secondary structure. The limited range of the spectrum prevented deconvolution. **B:** CD was measured at 208 nm over a temperature range of 20-95° C (+1° C min^{-1}) and a T_m of 65.3 (dotted line) was calculated from these data.

The CD spectroscopy of PilU△1-33 did not conclusively show that the protein was well folded despite the indication of the presence of secondary structure features. Moreover, it was possible to observe a change in the ellipticity at this wavelength during a temperature gradient indicating that the protein is at least partially folded.

5.3.3 PilK

Deconvolution of the PilK△1-32 CD spectrum (Figure 5.3.3A) using the CDSSTR algorithm on the Dichroweb server (Whitmore and Wallace, 2004), revealed that PilK△1-32

contained 52% helix, 10% strand and 15% coiled residues. The proportions of secondary structure calculated from the CD spectrum was similar to those predicted by the PSIPRED secondary structure prediction server (Table 5.3.1) (Buchan *et al.,* 2013). The CD spectrum and resulting deconvolution indicate that PilK△1-32 is a highly folded protein. CD spectroscopy at 220 nm over a temperature range of 4-95° C revealed a clear unfolding event at 70° C (Figure 5.3.3B). Interestingly, a weaker event may have occurred at 36° C but was not significant enough to be described as an unfolding event. This minor even may represent destabilisation of the interface between two domains within PilK△1-32 or that a domain within PilK△1-32 is partially unfolded and only required a relatively small amount of energy to result in the complete unfolding of a subdomain. While DSF of PilK△1-32, reaffirmed the the significant unfolding event observed at 70° C using CD, it also revealed a similar weak event at 36° C (Figure 5.3.3C). The minor event at 36° C, a relatively low temperature, indicated that care should be taken in the purification and storage of PilK△1-32 protein and was a factor to consider in crystallisation experiments.

	Helix (%)	**Strand (%)**	**Coil (%)**	**Disordered (%)**
CDSSTR	52	10	15	22
PSIPRED	48	18	–	19

Table 5.3.1 – PSIPRED secondary structure prediction and deconvolution of CD spectrum of PilK△1-32. Deconvolution of the PilK△1-32 CD scan spectrum (Figure 5.3.3A) using the CDSSTR algorithm calculated the secondary structure content described above. The average helix length was calculated to be 11 residues and average strand length ~5 residues. The deconvoluted CD is similar to the predicted secondary structure content of PilK△1-32 calculated using PSIPRED, indicating that PilK△1-32 is folded.

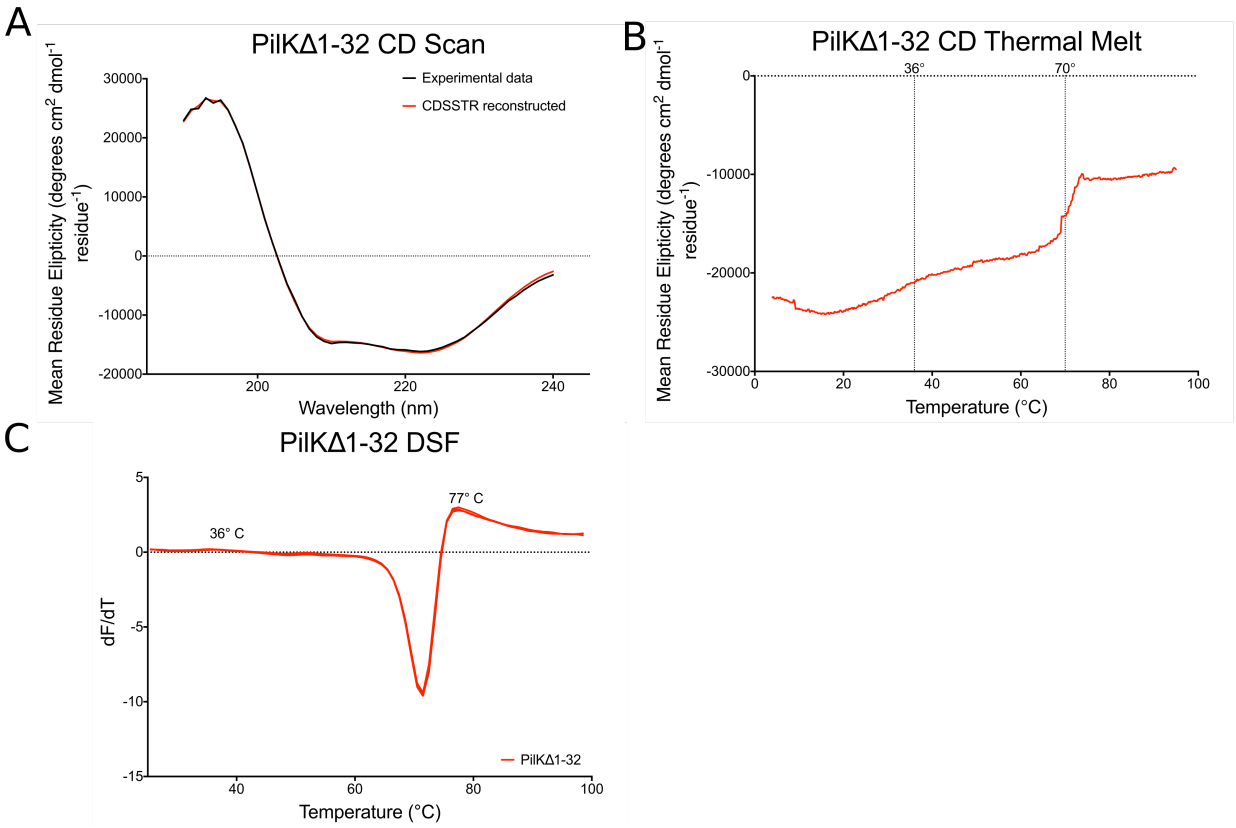


Figure 5.3.3 – CD spectroscopy and DSF of PilK△1-32. A: The CD spectrum of PilK△1-32 was measured over a range of 260-190 nm at 20° C. The experimental data (black) and reconstructed data (red) from CDSSTR strongly agree giving a high level of confidence in the deconvolution results. **B:** CD was measured at 220 nm over a temperature range of 4°-95° C (+1° C min^{-1}). An unfolding event was observed at 70° C (dotted lines), resulting in a much more significant change in ellipticity at 220 nm. A much smaller event may have also occurred at 36° C. **C:** A significant unfolding event was observed in triplicate DSF data at 77±0.2° C, although a small event was also observed at 36±0.5° C.

5.4 Crystallisation of minor TFP

5.4.1 PilV and PilU

Sparse matrix crystallisation trials of PilV△1-35 and PilU△1-33 were conducted using commercial crystallisation screens (Table 5.4.1). Many of these trials resulted in clear drops and increasing the protein concentration had no effect. PilV△1-35 and PilU△1-33 contain 15 Lys (9%) and 23 Lys (15%) residues respectively, making them good candidates for lysine methylation in an effort to reduce the solubility of the proteins and aid crystallisation (Figure 5.4.1). Even though lysine methylation changed the elution properties of the proteins in SEC, suggesting that the proteins had been methylated (Figure 5.4.1), crystals of these samples were not obtained.

Protein Sample	Total protein concentration (mg/ml)	Crystallisation screen	Drop ratio	Drop volume
PilV△1-35	9	Index, Structure, Morpheus, JCSG+	1:1, 2:1	100 nl
PilV△1-35	19	Index, Structure, Morpheus, JCSG+	1:1, 2:1	100 nl
PilV△1-35-K(CH_3)	12	Index, Structure, Morpheus, JCSG+, PACT	1:1, 2:1	100 nl
PilU△1-33	10	Index, Structure, Morpheus, JCSG+	1:1, 2:1	100 nl
PilU△1-33	40	Index, Structure, Morpheus, JCSG+	1:1, 2:1	100 nl
PilU△1-33-K(CH_3)	8	Index, Structure, Morpheus, JCSG+, PACT	1:1, 2:1	100 nl

Table 5.4.1 – Table of PilV and PilU crystallisation trials. Total protein concentration, crystallisation screens and drop ratio and volumes attempted in the pursuit of PilV△1-35 and PilU△1-33 crystals.

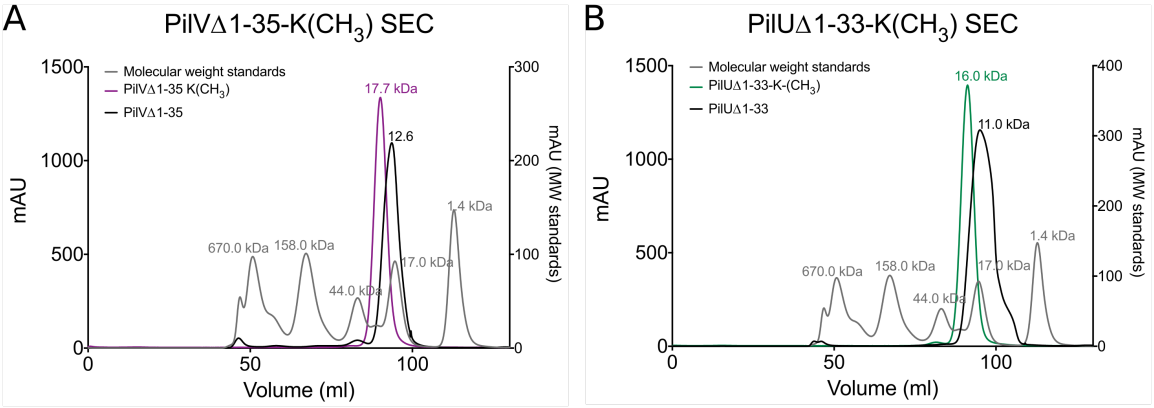


Figure 5.4.1 – Lysine methylation of PilVΔ1-35 and PilUΔ1-33. Lysine methylation was performed on purified PilVΔ1-35 (A, purple) and PilUΔ1-33 (B, green) proteins. After the reaction the proteins were purified by SEC using an S200 16/600 column. The peak elution volumes of the lysine methylated samples (coloured curves) were compared to non-methylated samples (black curves).

5.4.2 PilK

Sparse matrix crystallisation trials of both PilKΔ1-32 samples of monomer (Figure 5.2.7C/D) and HMW (Figure 5.2.7A/B) were set up at 16 mg/ml and 10.5 mg/ml respectively. The crystallisation screens probed were Index, Structure, Morpheus and JCSG+. Within 12 days of dispensing drops of the PilKΔ1-32 monomer sample, micro-crystals of approximately 10-20 μm across (Figure 5.4.2) had formed in 0.2 M $MgCl_2$, 0.1 M Na/HEPES pH 7.5, 30% PEG 400. The crystals were identified as protein crystals using a UV-scope [Molecular Dimensions] (Figure 5.4.2B/C/D). The crystals were harvested using micro-mesh crystal mounts with 10 μm apertures (Figure 5.4.3A) and directly flash cooled in liquid nitrogen as the crystallisation condition contained a cryo-protecting compound (30% PEG 400). The crystals were then tested at the micro-focus beamline at Diamond Light Source, I24.

An intrinsic problem in using micro-mesh crystal mounts is the removal of mother liquor from the very small apertures. The liquid surrounding crystals contributes to the background scatter during a diffraction experiment and data maybe lost if the diffraction spots from the crystal are weak. New methods for the mounting of microcrystals were trialled at Diamond Light Source, as work carried out during the internship included in this PhD, are discussed in Appendix D and published in, Acta Crystallographica, section D

(Warren *et al.,* 2015). By wrapping the mounted crystal samples in X-ray invisible multi-layer graphene, the mother liquor was drawn away from the crystals and the diffraction quality of the crystals was improved (Warren *et al.,* 2015).

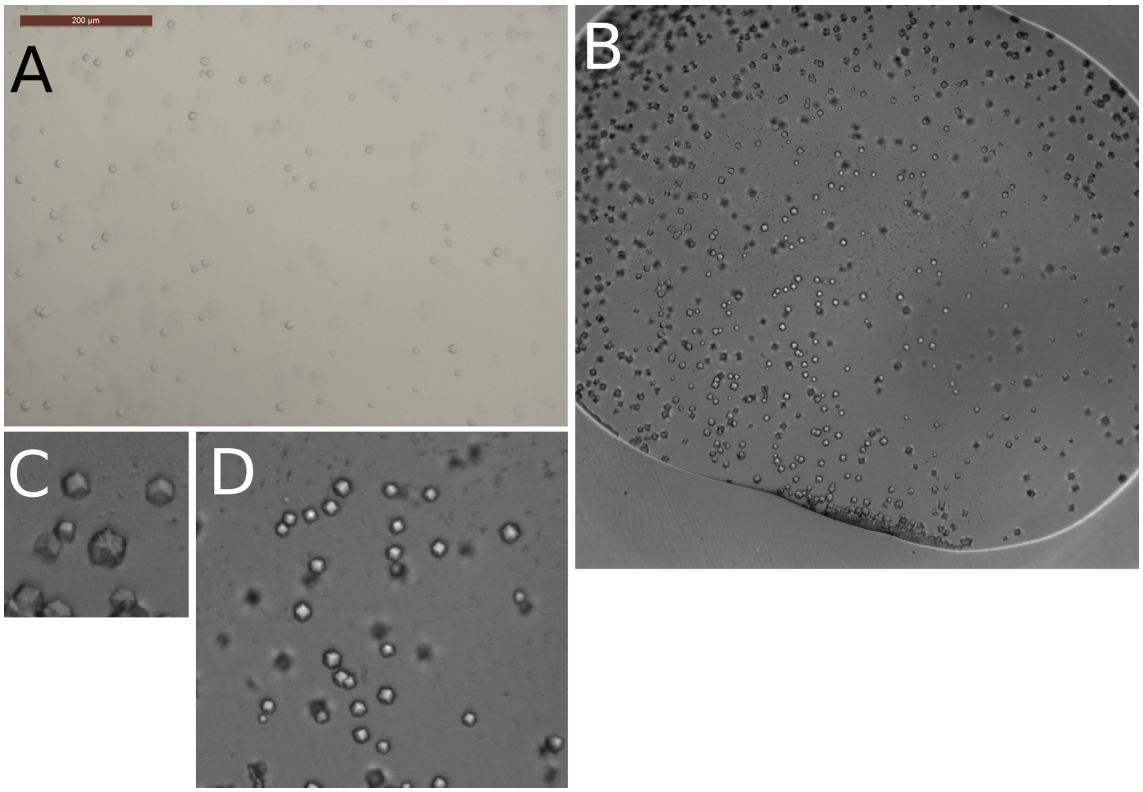
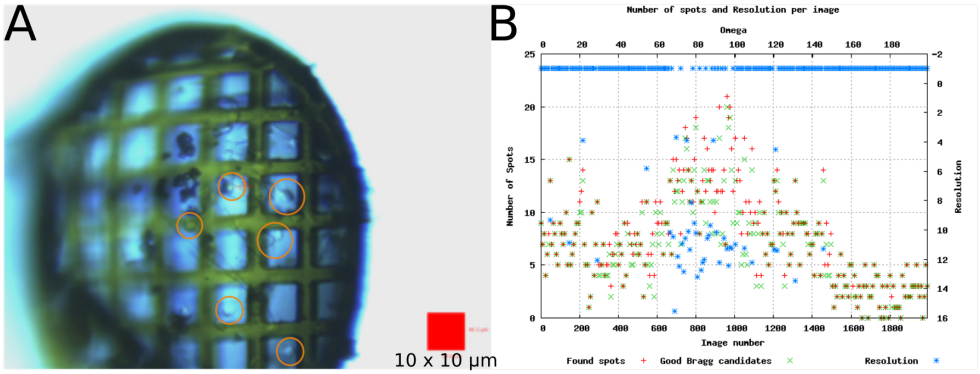
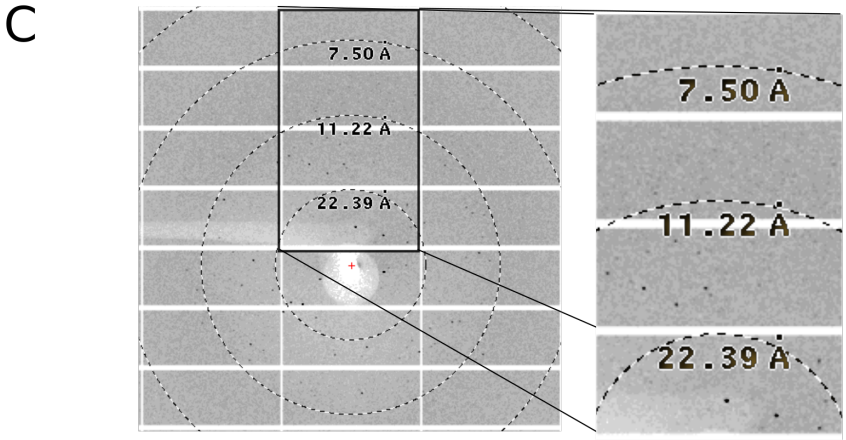


Figure 5.4.2 – PilK△1-32 micro-crystals. PilK△1-32 micro-crystals of approximately 10-20 µm across formed in 1:1 and 2:1 drops of PilK△1-32 crystals at 16 mg/ml in 0.2 M $MgCl_2$, 0.1 M Na/HEPES pH 7.5, 30% PEG 400 two weeks after dispensing. Crystals of such a size were difficult to identify by light microscopy (A). Crystals were identified as protein crystals using a UV-scope [Molecular Dimensions] (B). Zoomed in (C/D), fluorescent crystals are clearly observed.

Using grid scans to identify the location of the crystal on the mesh-mount, test diffraction images were taken of the crystals using a 10 x 10 µm beam at 100% transmission. The crystals diffracted to a resolution of ~7.5 Å (Figure 5.4.3C). A dataset of 2000 images over a total oscillation of 200° was collected from a crystal (Figure 5.4.3A/B) to determine the unit cell dimensions and spacegroup of the crystal. Automatic data processing pipelines on the I24 beamline at Diamond Light Source, the Xia2 pipeline using DIALS (Gildea *et al.,* 2014) and 3dii (XDS) (Winter *et al.,* 2013) suggest that the spacegroup of the crystal is $I4_1$ 3 2, with a large unit cell (a=b=c=~187 Å, α=β,=γ=90°) (Table 5.4.2). DIALS indicated that there were useful reflections to a resolution of 6.8 Å, however, the R_{merge} across the dataset at this resolution was high (>0.5) indicating the data were of





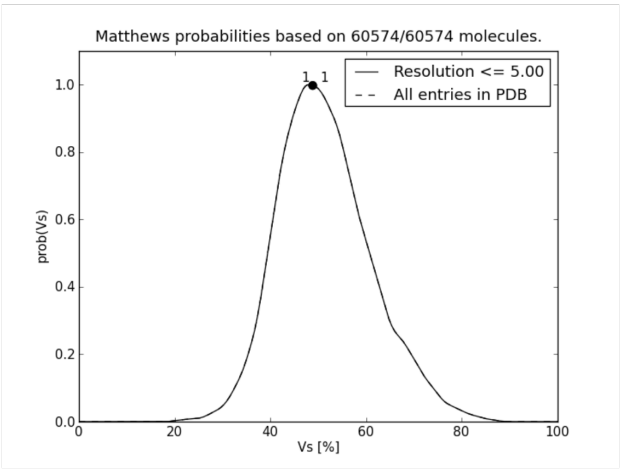


Figure 5.4.3 – **PilKΔ1-32 crystal diffraction. A:** PilKΔ1-32 crystals supported by a 10 x 10 µm aperture mesh crystal mount (crystals are highlighted in orange circles),. The beam size was 10 x 10 µm (red box). **B:** 2000 diffraction images were recorded over a 200° oscillation of the crystal mount. The number of spots found (red) and resolution (blue) per image are generated automatically in this DISTL plot (Zhang *et al.,* 2006). The number of spots that fulfil Bragg's law per image are also given (green). **C:** A test image taken to confirm alignment of the crystal within the X-ray beam and determine diffraction quality. The image represents a 0.5° oscillation. Reflections are observed to beyond a 10 Å resolution. **D:** The unit cell parameters were used to calculate the Matthews coefficient and solvent content (48.8%) of the PilKΔ1-32 crystal (Kantardjieff and Rupp, 2003). The graph shows the probability of a given solvent volume (Vs) against the percentage solvent volume (Vs[%]). The probability of 1.0 for one molecule in the ASU was calculated.

poor quality. The solvent content of the crystal and number of molecules in the asymmetric unit were calculated using the Matthews coefficient (Matthews, 1968; Kantardjieff and Rupp, 2003), indicating that there was one molecule in the asymmetric unit that has a solvent content of 48.8 %. The low resolution of these data did not permit any further processing or analysis. Using graphene based crystal mounting methods as trialled during the internship at Diamond Light Source might also have resulted in improvement of data quality. Optimisation of the crystals to produce larger crystals with greater diffraction power was therefore pursued.

	DIALS	**Xia2 3dii**
Wavelength (Å)	0.96858	
High resolution limit (Å)	6.81	7.66
Low resolution limit (Å)	66.24	66.31
Completeness (%)	99.9 (100.0)	99.9 (100.0)
I/σI	7.5 (0.9)	10.2 (2.1)
R_{merge}*	0.491 (6.102)	0.387 (2.382)
Unit cell dimensions:		
a=b=c (Å)	187.341	187.550
α=β=γ (°)	90	90
Spacegroup	$I4_1$ 3 2	$I4_1$ 3 2

Table 5.4.2 – Automatic processing statistics for PilKΔ1-32 crystals. Automated data processing pipelines at Diamond Light Source DIALS and Xia2-3dii successfully indexed and integrated the PilKΔ1-32 crystal data. The indexing and integration statistics are shown above. The pipeline processed the data to ~7 Å. Statistics in the highest resolution shell are in parentheses. *$R_{merge} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is the ith observation of reflection hkl and $\langle I(hkl) \rangle$ is the massed average intensity for all observations i of reflection hkl.

Many attempts were made to reproduce and improve upon the initial crystallisation hit of the PilKΔ1-32 crystals (Figure 5.4.2) and are summarised in Table 5.4.3. These trials included varying the crystallisation temperature, using additives and micro-seeding, as well as fine screening matrices around the original crystallisation solution. Crystallisation temperature was a factored variable due to the apparent two-step unfolding event observed by CD spectroscopy and DSF (Figure 5.3.3B/C)

Since there were no structures of homologous proteins in the PDB and PilKΔ1-32

Protein Sample	[Total protein] (mg/ml)	Crystallisation screen	Additive	Drop ratio	Drop volume	Temperature (°)	Tray Type
PilK△1-32	16	Index, Structure, Morpheus, JCSG+, PilK-Opt#1, PilK-Opt#2, PilK-Opt#3, Van-den-Berg Screen: #1,#2,#3	–	1:1, 2:1, 3:1	100 nl, 300 nl, 2 µl, 3 µl	20, 4	MRC, IQ, Crystal quick, 24-well (hanging drop)
PilK△1-32	10	PilK-Opt#2	–	1:1, 2:1	200 nl	20, 4	MRC
PilK△1-32	14	PilK-Opt#2	–	1:1, 2:1	200 nl	20, 4	MRC
PilK△1-32	19	PilK-Opt#2	–	1:1, 2:1	200 nl	20, 4	MRC
PilK△1-32	16	Structure, PilK-Opt#2	Hampton Additive screen	1:1, 2:1	100 nl	20, 4	MRC
PilK△1-32	16	Structure, PilK-Opt#2	Micro-seeding with PilK△1-32 micro-crystals	1:1, 2:1	100 nl	20, 4	MRC, IQ
PilK△1-32	8	Structure, PilK-Opt#2	Microseeding with PilK△1-32 micro-crystals	1:1, 2:1	100 nl	20, 4	MRC, IQ
PilK△1-32	16	Index, Structure, JCSG+, PilK-Opt#2	PilA1△1-34 (16 mg/ml)	1:1, 2:1	100 nl	20	MRC
PilK△1-32 (SeMet)	8	Structure, PilK-Opt#2	–	1:1, 2:1	200 nl	20, 4	MRC, IQ
PilK△1-32 (SeMet)	8	Structure, PilK-Opt#2	Microseeding with native PilK△1-32 micro-crystals	1:1, 2:1	200 nl	20, 4	MRC
PilK△1-32 (SeMet)	14	Index, Structure, Morpheus, JCSG+, PilK-Opt#2, PACT	–	1:1, 2:1, 3:1	100 nl	20, 4	MRC, IQ
PilK△1-32 (SeMet)	16	Index, Structure, Morpheus, JCSG+, PilK-Opt#2, PACT	–	1:1, 2:1, 3:1	100 nl	20, 4	MRC, IQ

Table 5.4.3 – Table of PilK△1-32 crystal optimisation trials. Total protein concentration, crystallisation screens and drop ratio and volumes attempted in the pursuit of better diffracting PilK△1-32 crystals. MRC plates have conical wells, while IQ and Crystal quick have flat, square drop wells. The crystallisation screens PilK-Opt#1, -Opt#2 and -Opt#3 are described in Appendix C, Table C.0.1, Table C.0.2 and Table C.0.3.

contained 13 methionines, a seleno-methionine preparation of PilK△1-32 protein was prepared for structure solution by SAD (Figure 5.4.4). In addition, incorporation of seleno-methionine into PilK△1-32 could have a positive effect on the crystallisation of the protein and, possibly, could have produced better diffracting crystals. The final yield of SeMet PilK△1-32 protein was 4.1 mg/L cell culture which was used in sparse matrix crystallisation screens and optimisation trials based upon the initial native hit at 16, 14 and 8 mg/ml (See Table 5.4.3). Incorporation of SeMet into PilK△1-32 appeared to improve the stability of the protein when compared with the native PilK△1-32 preparation (Figure 5.2.6 and Figure 5.2.7), since the significant N-terminal degradation product of ~25 kDa incorporating residues 52-211 was not observed in this sample.

Despite the extensive crystallisation optimisation trials listed in Table 5.4.3, which included ~10,000 crystallisation drops, it was not possible to obtain larger crystals of PilK△1-32 or SeMet PilK△1-32.

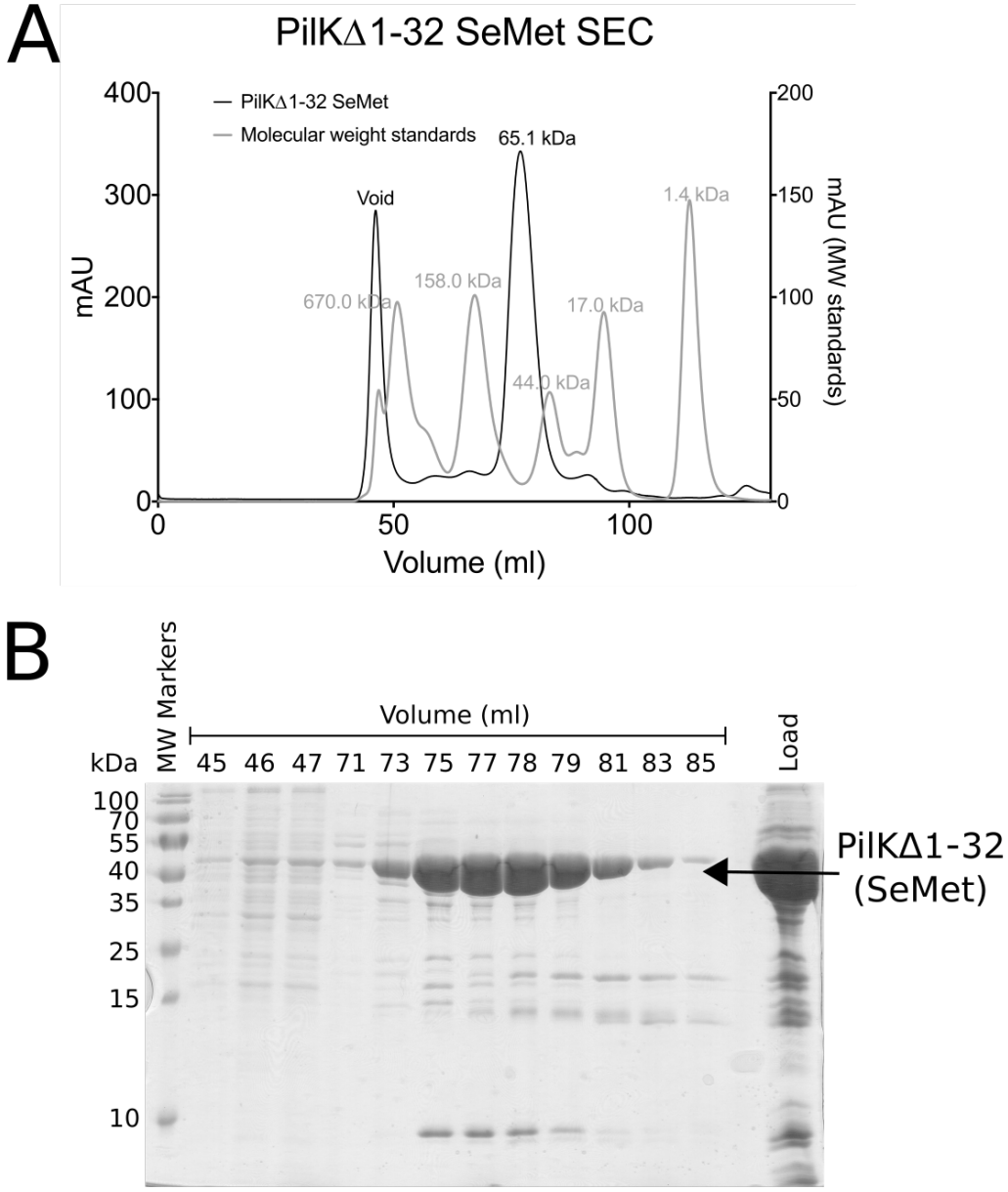


Figure 5.4.4 – Purification of SeMet PilK△1-32 protein. A: Chromatogram of SeMet PilK△1-32 purified by SEC after nickel affinity purification. SEC was performed using an S200 SEC column [GE Healthcare]. Two significant peaks eluted from the column at the void volume and a volume equivalent to 65.1 kDa. The molecular mass standards (grey) were thyroglobulin (670.0 kDa), γ-globulin (158.0 kDa), ovalbumin (44.0 kDa), myoglobin (17.0 kDa) and Vitamin B_{12} (1.4 kDa). **B:** SDS-PAGE analysis of the peak elution fractions from the SEC. Fewer impurities and degradation products were observed in the SeMet PilK△1-32 SEC purification compared with the native purification (Figure 5.2.6), specifically the degradation product observed at 25 kDa by SDS-PAGE of native PilK△1-32 was not as significant in the SeMet PilK△1-32 purification.

5.5 Inter-pilin interactions

To determine whether the head-groups of the minor pilins of PilV, PilU or PilK interacted with the head-group of the major pilin PilA1 (Chapter 4), preliminary interaction studies were performed. Surface plasmon resonance (SPR) was performed to identify interactions using a PilA1Δ1-34 bound surface that was challenged with PilVΔ1-35, PilUΔ1-33 or PilKΔ1-32. The PilA1Δ1-34 surface as also challenged with PilA1Δ1-34 protein to probe self-interaction. The PilVΔ1-35, PilUΔ1-33 and PilA1Δ1-34 proteins bound non-specifically to both the reference chip surface and the PilA1Δ1-34 bound surface. It was not possible to optimise sample conditions to minimise non-specific binding within the scope of this project. However, a specific interaction between PilKΔ1-32 and PilA1Δ1-34 was observed (Figure 5.5.1).

The K_D of the PilKΔ1-32:PilA1Δ1-34 interaction was calculated to be 13 μM. The SPR indicates that there is a weak affinity between the PilA1 and PilK head-groups and suggests that PilK could be located within the pilin fibre of the major pilin PilA1. The constructs used here of all the pilin proteins lack the hydrophobic polymerisation domains, which is expected to be required for the stable polymerisation of the pilin proteins into a fibre-like assembly. Therefore, it is interesting that PilA1 and PilK are able to form an interaction in the absence of the polymerisation domain.

Further experiments are required to confirm the data in Figure 5.5.1, including orthogonal techniques such as isothermal titration calorimetry (ITC) and microscale thermophoresis (MST).

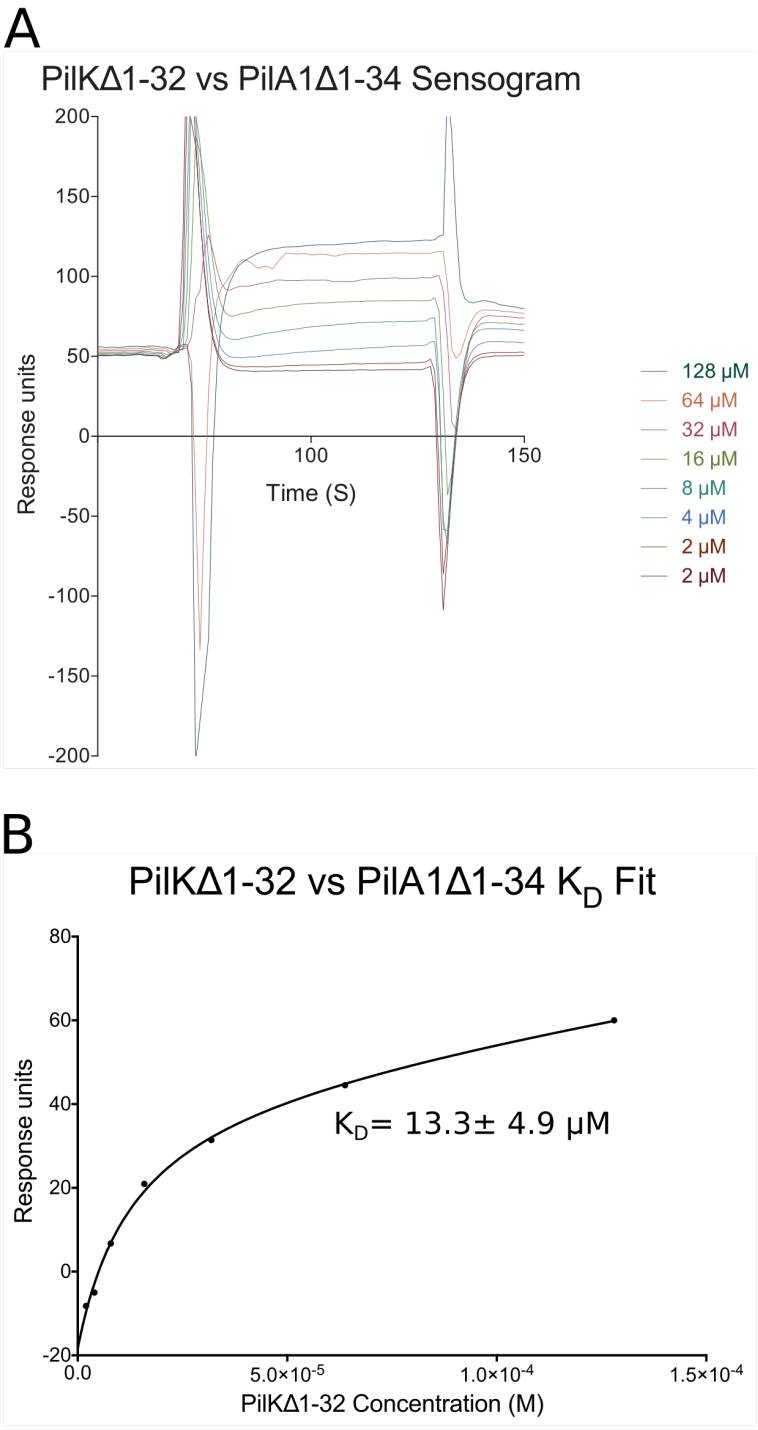


Figure 5.5.1 – SPR binding analysis of PilK△1-32 and PilA1△1-34. PilA1△1-34 was covalently bound to a CM5 sensor chip and PilK△1-32 was flowed across the PilA1 surface through a concentration range of 2-225 µM. The saturation K_D fit (B) of the change in response units vs concentration determined a K_D of 13.3±4.9 µM. The sensograms for each PilK△1-32 concentration are shown in panel A.

5.6 Discussion

5.6.1 Structural studies of minor pilins from TFP

During the course of this project it was only possible to obtain crystals of PilK△1-32 and not of the other minor pilins studied, PilV△1-35 and PilU△1-33. The micro-crystals of PilK△1-32 had limited diffraction, most likely due to their very small size and as a result it was not possible to obtain data that proved suitable for structure determination. Optimisation of the crystallisation conditions that produced PilK△1-32 crystals failed to reproduce diffracting crystals and it was not possible to obtain larger crystals. Characterisation of PilK△1-32 using CD indicated that the protein was well folded in solution. Even though the thermal stability assays suggested that PilK△1-32 unfolded at 70-77°C, a small event was also observed at the relatively low temperature of 36°C and it is possible this result influenced the crystallisation of the protein. In addition, some degradation of PilK△1-32 protein was observed during purification, although it was possible to purify apparently stable protein. The protein sample that did not contain significant degradation and eluted as an apparent monomer from SEC yielded crystals. The behaviour of PilK△1-32 protein observed in the experiments presented in this chapter suggest why an initial crystallisation hit was identified that was not subsequently optimisable to obtain larger, better diffracting crystals or even reproduce the micro-crystals of PilK△1-32. It is possible that the micro-crystals were a product of degraded PilK△1-32 or formed of a complex between degraded and full-length PilK△1-32 protein. Improving the diffraction quality and size of small crystals is highly challenging and in many cases micro-crystals cannot be improved upon (McPherson and Cudney, 2014).

In addition to optimising the purification and crystallisation conditions for PilK△1-32, alternative construct designs could have a greater crystallisation success. Such changes in construct design could include changing the position of the 6-His-tag from the C-terminus to the N-terminus of the construct or use removable tags. His-tags are intrinsically disordered and can hinder crystallisation success. Constructs designed around the degradation pattern observed during purification of PilK△1-32 (Figures 5.2.7 and 5.2.8) where the degradation product was determined to include residues 52-211 of PilK△1-32 may have greater stability and be more amenable to crystallisation. To understand whether the

degradation products are significant, two constructs PilK$\triangle$1-211 and $PilK_{211\text{-}512}$ would be appropriate. The micro-crystals obtained of PilK$\triangle$1-32 may have been the product of degradation at the protein termini, and to test this hypothesis, constructs that differ in length by a small number of residues could identify crystallisable protein. Constructs with triplet residue truncations at the N- or C-termini could aid crystallisation if terminal degradation enabled the micro-crystals to form. Analysis of the full-length PilK peptide sequence using the disorder prediction tool PONDR (Xue *et al.,* 2010), reveals a region of 32 residues (84-115) that are predicted to be disordered (Figure C.0.1). PilK constructs that exclude this region such as PilK$\triangle$1-115 could have greater overall order and be more amenable to crystallisation.

Crystallisation of the PilV$\triangle$1-35 and PilU$\triangle$1-33 proteins was unsuccessful in the scope of this project. Despite increasing protein concentrations and methylation of exposed lysines, both PilV$\triangle$1-35 and PilU$\triangle$1-33 proved to be highly soluble and limited precipitation was observed in crystallisation trials. Subsequent biophysical analysis of the PilV$\triangle$1-35 and PilU$\triangle$1-33 proteins using CD did not conclusively show that the proteins were well folded in solution and it was not possible to deconvolute the data to calculate secondary structure proportions. Further work is required to optimise the constructs and buffering conditions of both PilV and PilU. The PONDR disorder prediction tool (Xue *et al.,* 2010) predicts that each of the proteins have a single region of disorder: PilV residues 109-132 (Figure C.0.3); PilU residues 51-56 (Figure C.0.2). Constructs that do not include these regions, for example PilU$\triangle$1-56 or PilV35-109 may provide a better opportunity to obtain crystals. Since the C-termini of the ORFs of both PilU and PilV are predicted to be ordered, N-terminal His-tags maybe more appropriate, especially as the N-terminal regions of these proteins have been truncated.

The identification of an interaction between PilA1 and PilK suggests that co-crystallisation of the protein constructs presented in Chapter 4 and this Chapter could yield crystals. As PilA1 is the ubiquitous pilin unit, co-crystallisation trials with PilU, PilV or PilK proteins should be attempted in addition to PilK with PilU or PilV.

In the absence of crystals of the minor pilins studied here and thus structure models, modelling was attempted using the Phyre2 (Kelley *et al.,* 2015) and Swiss-Model tools (Arnold *et al.,* 2006; Biasini *et al.,* 2014; Bordoli *et al.,* 2009). Submission of the complete

ORF peptide sequences of PilV, PilU and PilK to these model building tools did not result in substantial models, however, the modelling tools identified potentially similar structures. The PilV and PilU models resemble the long N-terminal α-helix (Figure 5.6.1) that forms the polymerisation domains as observed in the PilA1 structures from *C. difficile* (Chapter 4) and TFP pilin proteins such as PilE1 from *Neisseria gonorrhoeae* (Craig *et al.,* 2006) and PAK pilin from *Pseudomonas aeruginosa* (Craig *et al.,* 2003). The structure templates used by the modelling programs are predominately pilin proteins in the PDB, PilU being modelled upon PilA (PDB: 2M7G) from *Geobacter sulfurreducens* (Reardon and Mueller, 2013) in Swiss-Model while Phyre2 used PAK (PDB: 1OQW) from *P. aeruginosa* (Craig *et al.,* 2003) as the major templates for PilU and PilV. Swiss-Model also used 1OQW as a major template in the production of the PilV model. The coverage and confidence of the PilV and PilU models and the templates that were chosen by the modelling tools suggest that these proteins have similar structure to major Type IVa pilin proteins. However, it seems both tools proposed models restricted to the long α-helix, characteristic of TFP pilins.

The modelsof PilK covered only a small proportion of the sequence and have low confidence scores meaning they are unlikely to be representative of the actual PilK structure. Unlike the PilV and PilU models, the templates used by the modelling tools are not related to TFP proteins and mostly include DNA binding or viral proteins. Structure modelling does not give any greater insight into the structure or function of PilK, this is likely due to the absence of apparent homologues of this protein.

It is interesting to note that while the templates for the PilV and PilU model include major pilins in the PDB, neither Swiss-model or Phyre2 utilised the *C. difficile* PilA1 or PilJ structures available in the PDB (Piepenbrink *et al.,* 2014; Piepenbrink *et al.,* 2015). PilA1 shares 16% identity and 9% identity with PilV and PilU, respectively, and are most highly conserved at the N-terminus (residues 1-30). PilJ shares 15% and 20% sequence identity with PilV and PilU respectively. It is likely that the structure of the polymerisation domain of the pilin proteins is conserved and sequence alignment and structure modelling provides evidence to suggest this. However, differences between the minor pilin headgroups and those observed in the PilA1 and PilJ proteins cannot be elucidated without crystal structures of the minor pilins studied in this chapter or from a close homologue

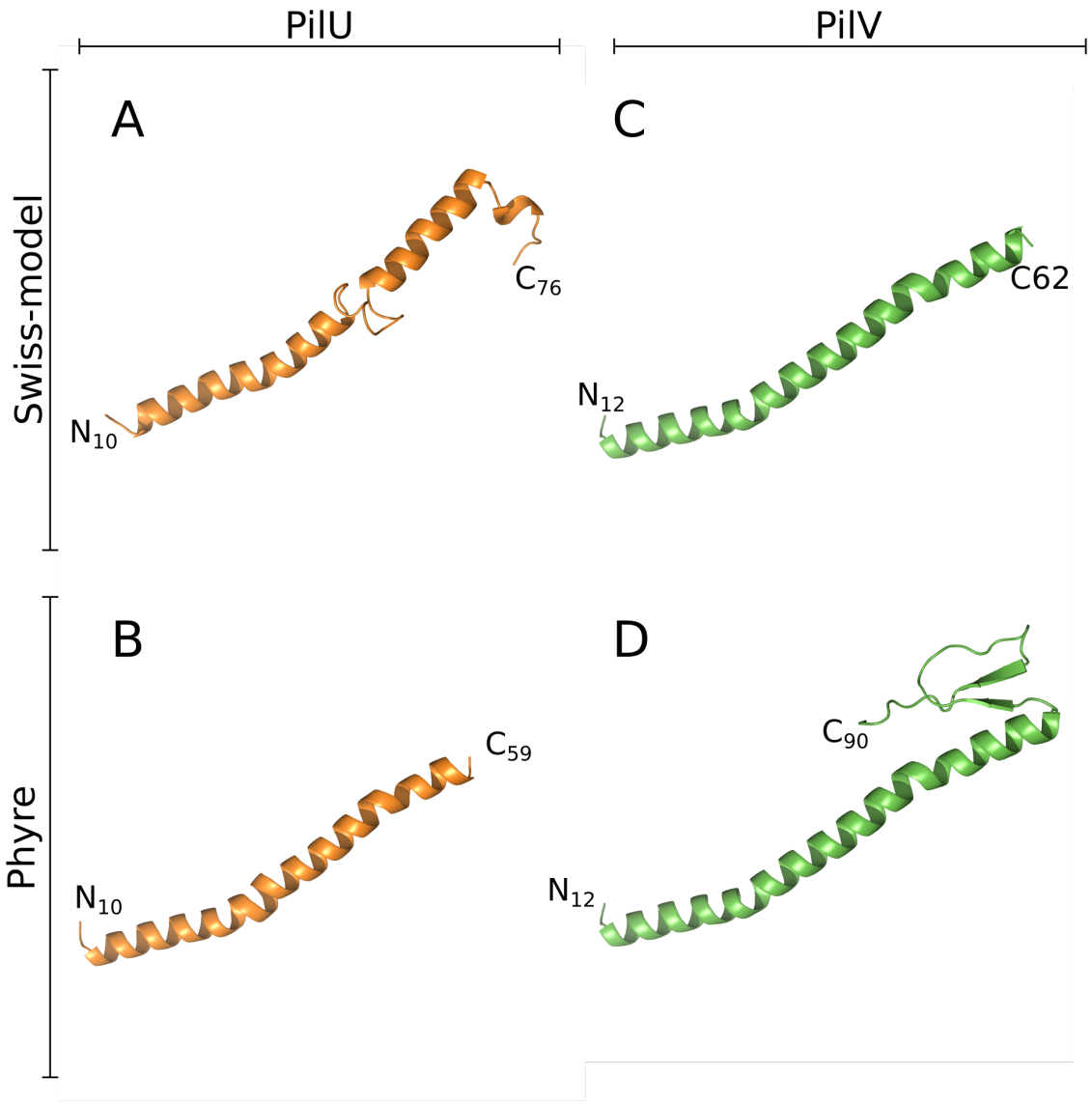


Figure 5.6.1 – Swiss model and Phyre2 models for PilV and PilU. A: PilU model (GMQE: 0.18) from Swiss-model covering residue 10-76. GMQE is a global model quality estimation which reflects the accuracy of the model against the alignment and template and is scored between 0 and 1. **B:** PilU model (confidence: 94%; coverage: 28%) from Phyre2 covering residues 10-59. **C:** PilV model from Swiss-model (GMQE: 0.12) covering residues 12-62. **D:** PilV model (confidence: 83%; coverage: 42%) from Phyre2 covering residues 12-90.

such as *Clostridium perfringens*.

5.6.2 TFP pilin assembly

Interactions between the head-groups of the minor pilins with the major pilin PilA1 were investigated in this project and it was determined that the PilA1 and PilK head-groups formed an interaction with a K_D of 13 μM. Previously Piepenbrink *et al.* had determined an interaction between the PilA1 head-group and another pilin PilJ with a K_D of 70 μM (Piepenbrink *et al.,* 2014). PilJ was also demonstrated to be localised in the pilin fibre with PilA1 (Piepenbrink *et al.,* 2015). The majority of the interactions between the pilin proteins within the fibre is expected to occur via the hydrophobic polymerisation domain which forms a hydrophobic core inside the fibre. The PilA1 and PilJ structures determined by Piepenbrink *et al.* have been modelled in fibre assemblies (Piepenbrink *et al.,* 2015) based on the cryo-EM structure of PAK pilin (Craig *et al.,* 2003). In the modelled fibre, the truncated polymerisation domains of PilA1 and PilJ were modelled (Figure 5.6.2). The protein molecules were organised in such a way that the hydrophobic polymerisation domains formed the core of the fibre, the N-termini of the proteins pointing towards the base of the fibre (Figure 5.6.2A). Where PilJ was added, the larger protein provided a more bulky head-group (Figure 5.6.2B).

The identification of an interaction between PilA1 and PilK is interesting due to the unknown role of the PilK minor pilin within the fibre. It has been hypothesised that PilK is similar to the GspK family of proteins from the TTIIS system that forms a cap on the end of the pseudopilus (Korotkov and Hol, 2008). If indeed PilK plays this role, it is likely that the protein interacts with the head-group of PilA1, since the head-groups are positioned at the top of the filament (red block, Figure 5.6.3). Even though PilK is predicted to have a polymerisation domain that will anchor the protein into the core of the filament, the head-groups of several PilA1 units are likely to be exposed at the top of the filament. As PilK is much larger than the other minor pilins, it may be large enough that it interacts with more than one PilA1 unit to completely cap the end of the filament. Further investigation of the PilA1 interaction with PilK could be conducted with more limited PilK constructs (PilKΔ1-211 and $PilK_{211\text{-}512}$ for example) than have been used here, to enable identification of the interaction domains. Truncations of the PilA1 head-group designed in a similar manner would enable identification of the domain interacting with PilK.

While PilA1 is the major pilin unit, PilJ and PilV have been identified on the surface

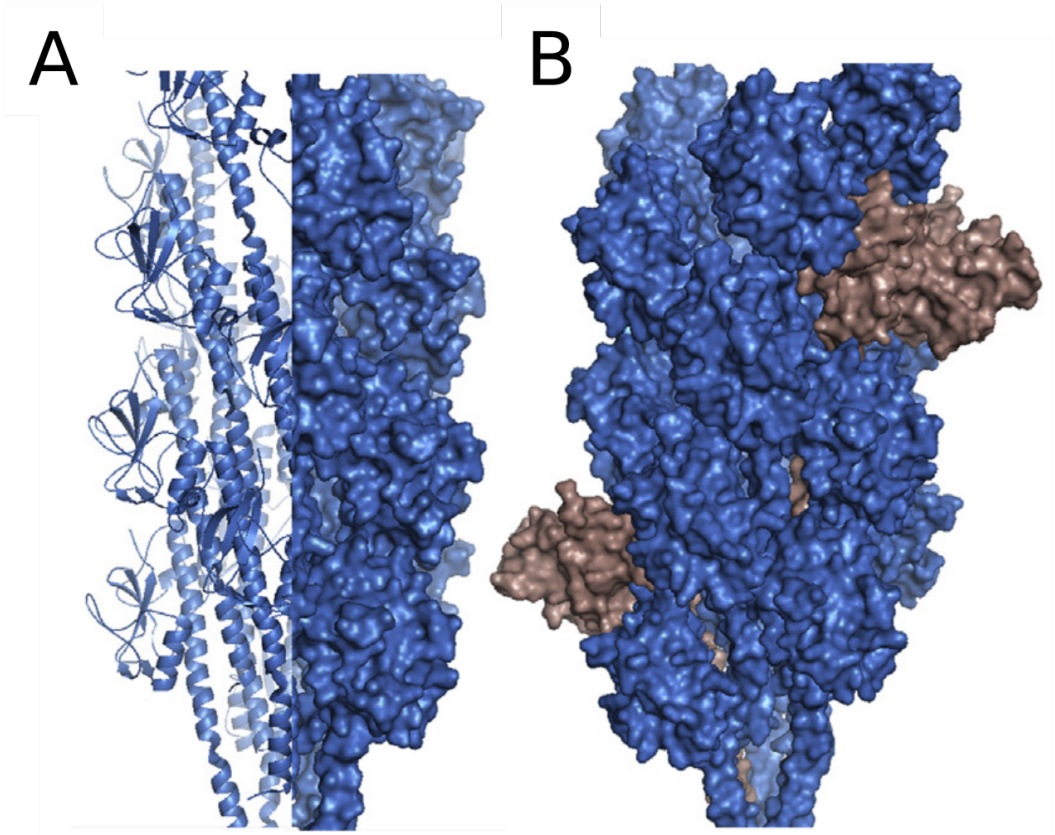


Figure 5.6.2 – Model of *C. difficile* fibre assembly using PilA1 and PilJ. Adapted from Piepenbrink et al., 2015. PilA1 is structurally similar to the *Vibrio cholerae* major pilin TcpA of which an electron microscopy model of the filament has been produced (Craig *et al.,* 2003; Lim *et al.,* 2010). Based upon the model of the *V. cholerae* TFP filament, Piepenbrink et al. were able to fit their PilA1 structure within the electron density to produce a model of a filament formed exclusively of PilA1 (A) and with PilJ present (B).

of *C. difficile*, it is not unreasonable that PilK may also interact with these proteins and potential interactions between the head-groups of PilJ, PilV and PilU with PilK should also be probed.

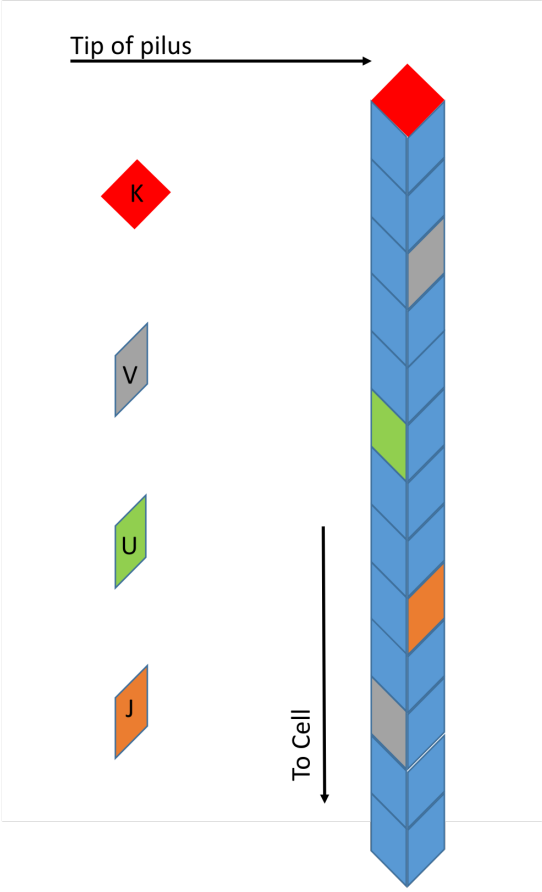


Figure 5.6.3 – **Predicted organisation of TFP filaments.** The TFP filament is predominantly formed of PilA1 units (blue) with the minor pilins, PilV (grey), PilU (green) and PilJ (orange) interspersed along the pilin filament. PilK (red) is predicted to sit at the end of the filament as a filament cap.

5.6.3 The role of minor pilins

As PilA1 forms the bulk mass of the TFP pilin fibre in *C. difficile* (Piepenbrink *et al.,* 2015), the exact role of the minor pilins in *C. difficile* is yet to be elucidated. Minor pilins feature in many TFP loci of both Gram-positive and Gram-negative bacteria, and in the latter the function of some minor pilins has been determined. Minor pilins have been studied extensively in the Gram-negative pathogen, *Neisseria meningitidis* and their functions include the

ability to bind DNA (Cehovin *et al.,* 2013), host-cell adhesion (Helaine *et al.,* 2005; Helaine *et al.,* 2007) and extension/retraction of the pilin filament (Szeto *et al.,* 2011). *N. meningitidis* has Type IVa pili genes and therefore the same signal peptide as observed in *C. difficile* (Craig and Li, 2008). It is recognised that the minor pilins are accessory pilin units in the fibre and can provide TFP with secondary functions (Melville and Craig, 2013).

Modelling of PilV and PilU (Subsection 5.6.1), and sequence conservation at the N-terminus suggests that these proteins follow the structural paradox of TFP pilin proteins of an N-terminal hydrophobic stem that is involved in forming the core of the pilin subunit. The PilV and PilU proteins are similar in size to the major PilA1, suggesting conformity in the pilin units that form the fibre. PilV expressed from the locus known as CD3507 (previously described as PilA) has been observed on the surface of *C. difficile* strain 630 cells in infected hamsters, indicating the presence of PilV pilin units within TFP fibres (Goulding *et al.,* 2009).

The function of the largest minor pilin unit, PilK, is the most curious. In many species that express Type IVa pilins, a single larger minor pilin gene is located in the major loci, amongst minor pilin genes (Melville and Craig, 2013). Unlike the rest of the pilin proteins in these loci, they do not conform to the TFPa signal peptide and have a hydrophobic residue at the position where glutamate is usually conserved. Additionally, in *Clostridia*, the N-terminal residue of mature TFP is not phenylalanine but an alanine residue. This phenomena is also observed in TIISS and loci encoding pseudopilins in Gram-negative bacteria, these large minor pilins have been named the GspK family and are also observed in TFPb systems (Melville and Craig, 2013). The structure of a GspK pseudopilin from the *E. coli* TIISS was determined in complex with two other pseudopilins (GspI and GspJ) that form a pseudopilus (Figure 5.6.4A-C) (Korotkov and Hol, 2008). All of the proteins in the GspK complex have the α-β fold that is observed in many pseudopilins and TFP proteins such as PilA1 and PilE (Figure 5.6.4) (Forest, 2008). In addition, GspK also has a large α-helix flanked by two β-strands as is observed in the major pilins (Forest, 2008). The complex formed a quasi-helical structure that could form the tip of the pseudopilin. The sequence identity between *E. coli* GspK and *C. difficile* PilK is 17%. However, the similarity of the larger size, consistent differences in signal peptide sequence and position in their respective loci across TFP and TIISS/pseudopilin loci suggests that the role of PilK

could be to cap the pilin fibre.

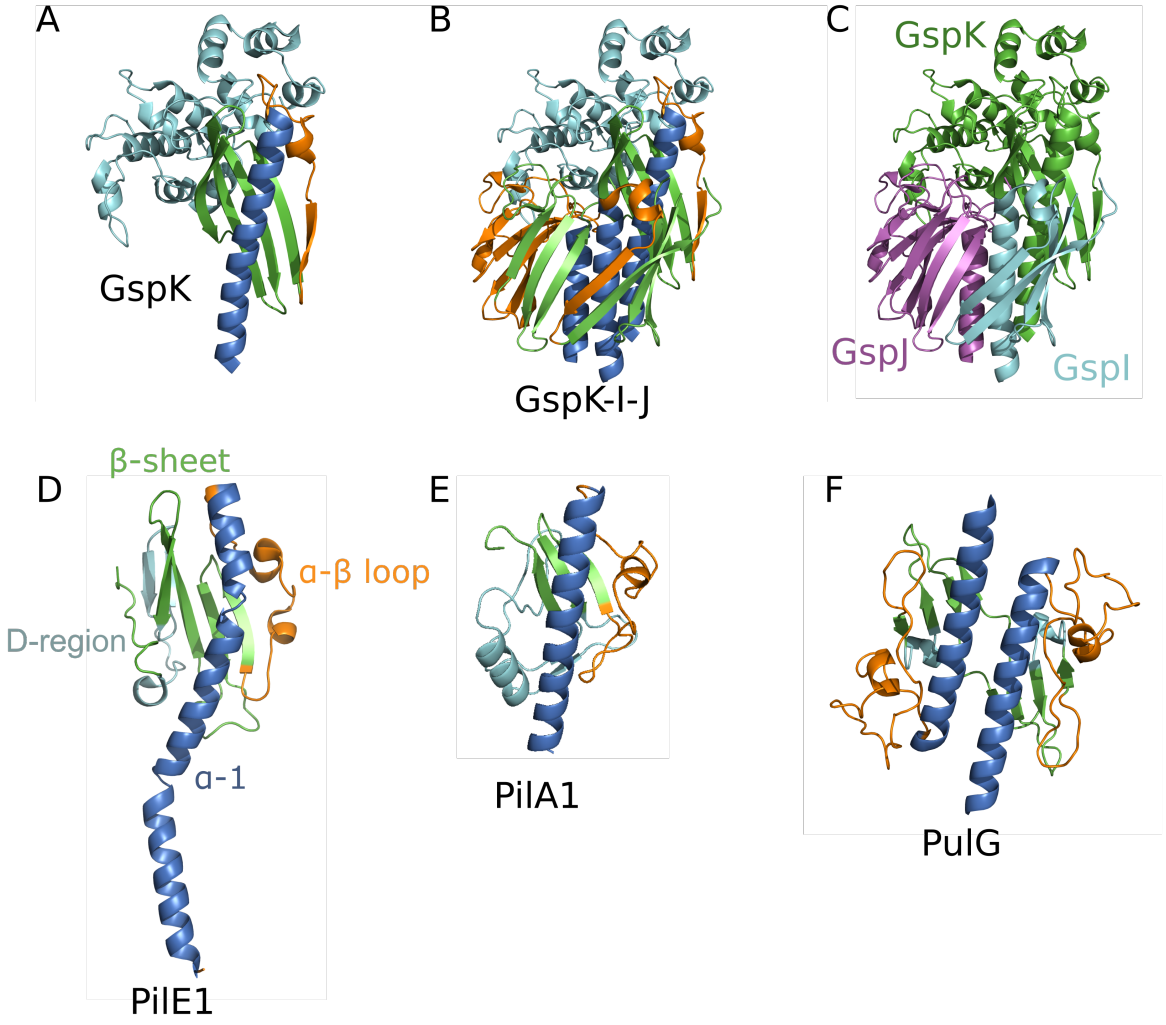


Figure 5.6.4 – Comparison of GspK, GspK-I-J complex, PilE1 and R20291 PilA1 structures. The crystal structure of *E. coli* GspK (A), a TIISS pseudopilin. The α-helix 1 is coloured dark blue, the α-β region is coloured orange and conserved β-strands are coloured green. The D-region or variable region is coloured in cyan. The structure of GspK was determined in complex (B) with GspI (C, blue) and GspJ (C, purple). For comparison the structures of PilE1 (D), R20291 PilA1 (E) and the pseudopilin PulG (F) are also shown (Craig *et al.,* 2006; Köhler *et al.,* 2004).

The structure of GspK suggests that the larger pilin could have a larger D-region that could enable the tip of the fibre to have greater variability. In an attempt to find any internal repeats within PilK, both the PilK and GspK peptide sequences were submitted to Radar, an internal repeat search tool (Heger and Holm, 2000; Goujon *et al.,* 2010). Radar revealed three possible internal repeats within PilK between 61 and 88 residues in length. Interestingly, despite the internal repeats not being highly conserved (16-23% identity),

the class of side chain was more conserved (Figure 5.6.5). GspK was found to contain 2 internal repeats of an average length of 113 residues and shared 25% sequence identity (Appendix C, Table C.0.4).

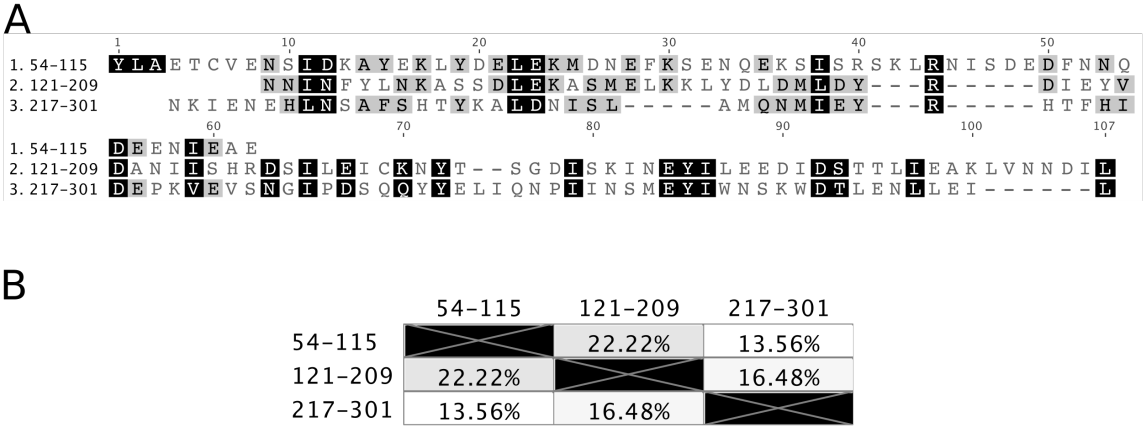


Figure 5.6.5 – **Alignment of PilK internal repeats.** Radar detected internal repeats between residues 54-115, 121-209 and 217-301 within PilK. The repeats contain conserved amino acid types at positions highlighted in black (A). The identity distances between these repeats are between 16 and 23% (B).

The TFP of Gram-positive must pass through the thick peptidoglycan layer. In Gram-negative bacteria the pilin fibre passes through the peptidoglycan layer in the periplasm and through the outer membrane via a series of channels formed of proteins known as PilQ (Martin *et al.,* 1993; Chang *et al.,* 2016). No proteins have so far been identified in the TFP expressing Gram-positive bacteria that may form a secretion channel through the peptidoglycan layer. If positioned at the tip of the pilin filament, PilK maybe required for forcing the filament through the peptidoglycan layer.

To test whether PilK is indeed at the tip of the pilin fibres in *C. difficile* and the other minor pilins in the fibre*, in vivo* experiments are needed. Labelling of the pilin proteins with fluorescent tags and conducting single molecule fluorescent experiments to verify incorporation of pilin units within the fibre and to quantify the ratio of different pilins could be carried out. Electron microscopy could be used to identify immunogold antibody labelled pilin units using specific antibodies raised against individual pilin units, this technique has been successfully utilised to quantify the incorporation of PilJ into PilA1 filaments (Piepenbrink *et al.,* 2015). Methods to control the expression of TFP fibres in *C. difficile* using cGMP have been developed could be advantageous for such localisation experiments

(Bordeleau *et al.,* 2015).

Chapter 6

Discussion

A major hospital associated pathogen, *C. difficile* still represents a significant challenge in the healthcare setting. Increasing resistance to antibiotics that are commonly used to treat *C. difficile* infection (CDI) is a problem and novel therapeutics are now being sought. In order to achieve this, basic understanding of the *C. difficile* life cycle must be acquired. Sporulation and colonisation are two important pathogenicity associated traits that are still poorly understood. The *C. difficile* spore is a highly resistant dormant cell type that is the infective agent of CDI and enables the spread of infection. Colonisation is the stage of the cycle during which *C. difficile* populates the host gut and produces enterotoxins that cause the clinical symptoms of infection which can in some cases result in fatality.

6.1 The *C. difficile* SpoIIQ:SpoIIIAH sporulation complex

The SpoIIQ:SpoIIIAH complex is required for the engulfment of the forespore by the mother cell during sporulation. SpoIIQ, located in the forespore, and SpoIIIAH, in the mother cell, form a bridge between the two cells and are predicted to provide a communication mechanism that enables coordination of sporulation gene expression.

6.1.1 Aims and outcomes

The aims of the work on the SpoIIQ:SpoIIIAH complex presented in this thesis were the *in vitro* biophysical and structural characterisation of the individual proteins and their com-

plex from *C. difficile*. SpoIIQ contains a conserved LytM domain that includes signature metal coordinating motifs. It was shown that these conserved motifs did indeed coordinate a Zn^{2+}, as observed in active LytM proteins. Surprisingly, it was determined that regardless of whether Zn^{2+} is required for catalytic activity of SpoIIQ, presence of the metal ion was essential for the formation of a stable 1:1 complex between SpoIIQ and SpoIIIAH *in vitro*. This result complemented our collaborator's work that showed that a *spoIIQ* mutant incapable of binding Zn^{2+}, formed fewer and less stable complexes with SpoIIIAH, than those formed by wildtype SpoIIQ (Serrano *et al.,* 2015). This was a different behaviour that had not been observed in the SpoIIQ:SpoIIIAH complex from *B. subtilis* that has been extensively studied but does not have metal binding capabilities.

As the basic understanding of the sporulation pathway has increased over the past 4 years, it has become clear that although many of the sporulation expressed proteins and their regulators are conserved across the *Bacilli* and *Clostridia*, the exact mechanisms are not. The *Clostridia* represent an older genera of the Firmicutes than *Bacilli* and it has been hypothesised that *Bacilli* have evolved a more complex sporulation pathway. It is possible that the conserved LytM domain of *C. difficile* SpoIIQ represents a now arcane function but the Zn^{2+} binding capability is conserved in this bacteria as it is required for stable complex formation.

6.1.2 Future outlook: SpoIIQ:SpoIIIAH

The role of the Zn^{2+} bound to SpoIIQ in the *C. difficile* sporulation complex is still not completely understood. Further studies on the potential endopeptidase activity of SpoIIQ are required to determine whether the metal has a dual role or is exclusively a structural requirement. Understanding of the exact SpoIIQ:SpoIIIAH complex interface is lacking in relation to why the Zn^{2+} is required for complex stabilisation. It is hypothesised that the metal ion stabilises the formation of secondary structure in SpoIIQ at the interface required for SpoIIIAH binding, which otherwise forms a predicted disordered loop. Indeed, *C. difficile* SpoIIIAH was shown to be compatible with *B. subtilis* SpoIIQ but not vice versa, indicating that the SpoIIQ binding interface of *C. difficile* SpoIIIAH is similar to that observed in *B. subtilis*. Further structural studies are required to understand the exact structural mechanism of Zn^{2+} binding in SpoIIQ, using techniques such as NMR or X-ray

crystallography should crystals of the complex become obtainable using new constructs.

The SpoIIQ and SpoIIIAH proteins are membrane anchored proteins with N-terminal transmembrane domains, whilst the work presented in this thesis was conducted using SpoIIQ and SpoIIIAH constructs that lacked these regions. Membrane anchoring may change the biophysical properties of these proteins and it is therefore important to study them in this context. Thus far there is no data from either *C. difficile* or *B. subtilis* regarding the predicted self-polymerisation of the SpoIIQ or SpoIIIAH proteins. Assembly models have so far been based on the sequence similarity of SpoIIIAH with the Type III secretion protein EscJ.

Optimisation of membrane purification of full-length SpoIIQ and SpoIIIAH should be carried out. This would enable experiments such as SEC-MALLS to determine the oligomeric state of these proteins, binding studies of SpoIIQ vs SpoIIIAH, in which the presence of the N-terminal region could have an as yet unknown effect, and electroporation experiments to ascertain whether full-length SpoIIQ and SpoIIIAH can form pores and to determine the size of such assemblies. *In vitro* formation of the SpoIIQ:SpoIIIAH assembly may also provide a mechanism to test the transport of potential substrates such as proteins, small molecule metabolites or DNA, as their exact nature is currently unknown.

Although membrane proteins are substantially more challenging crystallisation targets than soluble proteins, full-length SpoIIQ and SpoIIIAH may prove more amenable. Failing crystallisable targets, if full-length SpoIIQ and SpoIIIAH can form assemblies of at least several molecules, they would be suitable for structural studies using cryo-electron microscopy which would provide valuable information regarding the SpoIIQ:SpoIIIAH complex assembly.

Finally, *in vivo* studies in *C. difficile* have suggested that the SpoIIQ:SpoIIIAH complex does not act alone during engulfment of the forespore. Interaction with another sporulation complex, the SpoIIDMP machinery, which is predicted to have endopeptidase activity, appears to be recruited by the SpoIIQ:SpoIIIAH complex to the engulfment septum (Serrano *et al.,* 2015). Such association of the SpoIIQ:SpoIIIAH and SpoIIDMP complexes could be important for the proper peptidoglycan and membrane processing during engulfment. Also, SpoIIIAH is the final gene in the 8 gene *spoIIIA* operon, which includes membrane proteins and a predicted ATPase. It has been predicted that some of the *spoIIIA* expressed

genes may be required for the regulation of the SpoIIQ:SpoIIIAH complex from the mother cell side. *In vitro* characterisation of both the SpoIIDMP proteins and products of the *spoIIIA* operon should be carried out and possible interactions of these proteins with the SpoIIQ:SpoIIIAH should be investigated. Indeed, interaction of one or more of the *spoIIIA* products could stabilise the SpoIIQ:SpoIIIAH complex and provide a more amenable crystallisation target than SpoIIQ and SpoIIIAH alone.

6.2 Type IV pilins in *C. difficile*

Type IV pili (TFP) have only relatively recently been identified in Gram-positive bacteria, including *C. difficile*. So far characterisation of the expression, function and structure of TFP in *C. difficile* is at an early stage. However, this early work and their apparent similarity to pili that have been extensively characterised, suggests that TFP may play an important role in *C. difficile* colonisation, biofilm formation and virulence.

6.2.1 Aims and outcomes

The work on TFP in this thesis aimed to determine the crystal structures of the major and minor pilins and determining potential interactions between the major and minor TFP proteins beyond the polymerisation domain.

The structure of the major pilin, PilA1, was determined from two *C. difficile* strains, the 630 lab strain and the hyper-virulent R20291 to 1.6 Å and 1.7 Å, respectively. The PilA1 structures were determined to be highly similar and the most significant structural variations were observed in the sequence variable D-region which is predicted to be exposed on the surface of the pilin filament. The structures also shared the common features of type-a pilin proteins as had been determined from Gram-negative species including a predominant α-1 helix, α-β loop, 3 stranded anti-parallel β-sheet, and the variable D-region. However, PilA1 lacks any Cys residues and therefore the D-region was not delimited by a disulphide bridge. PilA1 is unusual in this aspect as this is a common feature of not only TFP but many other pilin proteins.

During the course of this work, the crystal structure of the PilA1 from R20291 and two other strains was published by Piepenbrink et al. and although the structure determination

methods varied considerably, the structure of the R20291 PilA1 monomer was essentially the same (Piepenbrink *et al.,* 2015). However, Piepenbrink et al. determined the structure of MBP-PilA1 fusions, whilst our structures contained no fused tags, allowing us to observe PilA1 trimers which provide possible PilA1 interfaces. In particular, the 630 structure formed an interface that could be compatible with PilA1 assembly in the pilin filament.

Pilins have been shown to interact via their α-1 helix, a hydrophobic helix that forms a polymerisation domain but it is also recognised that the head-groups of the pilin proteins will be in contact within the filament. Therefore, interactions between the head-groups of the major pilin, PilA1, and the minor pilins, PilV, PilU and PilK were probed. An interaction was identified between PilA1 and PilK with an K_d of less than 20 μM. While PilV and PilU are ~17 kDa in mass and contain the conserved signal peptide, PilK is almost 3-fold greater in mass and contains a Glu-Leu substitution within the signal peptide. The function of PilK as a minor pilin is currently unknown, however, it has been hypothesised that PilK may form a cap and in this role it is likely that PilK could interact with the headgroup of PilA1.

6.2.2 Future outlook: Type IV pili

Determination of the structures of the minor pilins will provide insights into their function and possibly into their assembly within the *C. difficile* pilin filament. In particular, optimisation of PilK crystals and structure determination would provide significant information as there are no homologues or domain information available. An interesting feature of the PilK peptide sequence is that it contains three internal repeats, which could represent repeating structural domains. A proposed homologue of PilK, the pseudopilin GspK, which contains repeating structural domains, is also proposed to form the cap of the pseudopilus (Forest, 2008; Korotkov and Hol, 2008; Melville and Craig, 2013). No structures of filament cap proteins from any TFP expressing species currently exist and therefore PilK represents an important and novel structure target. Work is now ongoing to fluorescently label PilK in *C. difficile* in an attempt to determine the localisation pattern of PilK and confirm the hypothesis that it caps the filament.

The PilA1 structures share significant structural homology with the PilE1 and TcpA structures from *Neisseria gonorrhoeae* and *Vibrio cholerae*. The structures of the TFP

filaments from these species have been determined and the crystal structures modelled within them. Considering the similarities between PilA1 and these structures, it is likely that *C. difficile* forms pilus filaments which are also structurally similar. The filaments of the Gram-negative bacteria were constitutively expressed and isolated for cryo-EM experiments. This is currently a challenging aim in *C. difficile* despite the identification of pathways, such as c-di-GMP, that control TFP expression. In the mean time, *in vitro* characterisation of possible interactions between the head-groups of the major and minor pilins will give information on filament formation.

Potential biological interfaces were identified in the crystal structures of the PilA1 proteins presented here. While there was no experimental evidence that PilA1 could form oligomeric states greater than monomers, point mutations that remove or reverse the charge polarity of possible salt-bridge forming residues should be produced and purification attempted to assess protein stability and probe formation of high order states.

6.3 Final remarks

The work presented in this thesis provides new insight into the basic mechanisms of sporulation and colonisation of *C. difficile*. These stages in the life cycle of *C. difficile* represent important pathogenicity factors. Understanding of these areas has so far been limited and represents potential new routes towards the control and treatment of *C. difficile* infection. This work extends our knowledge of these processes and can be the basis for future development of such strategies.

Appendix A

SpoIIQ:SpoIIIAH complex in *Clostridium difficile*

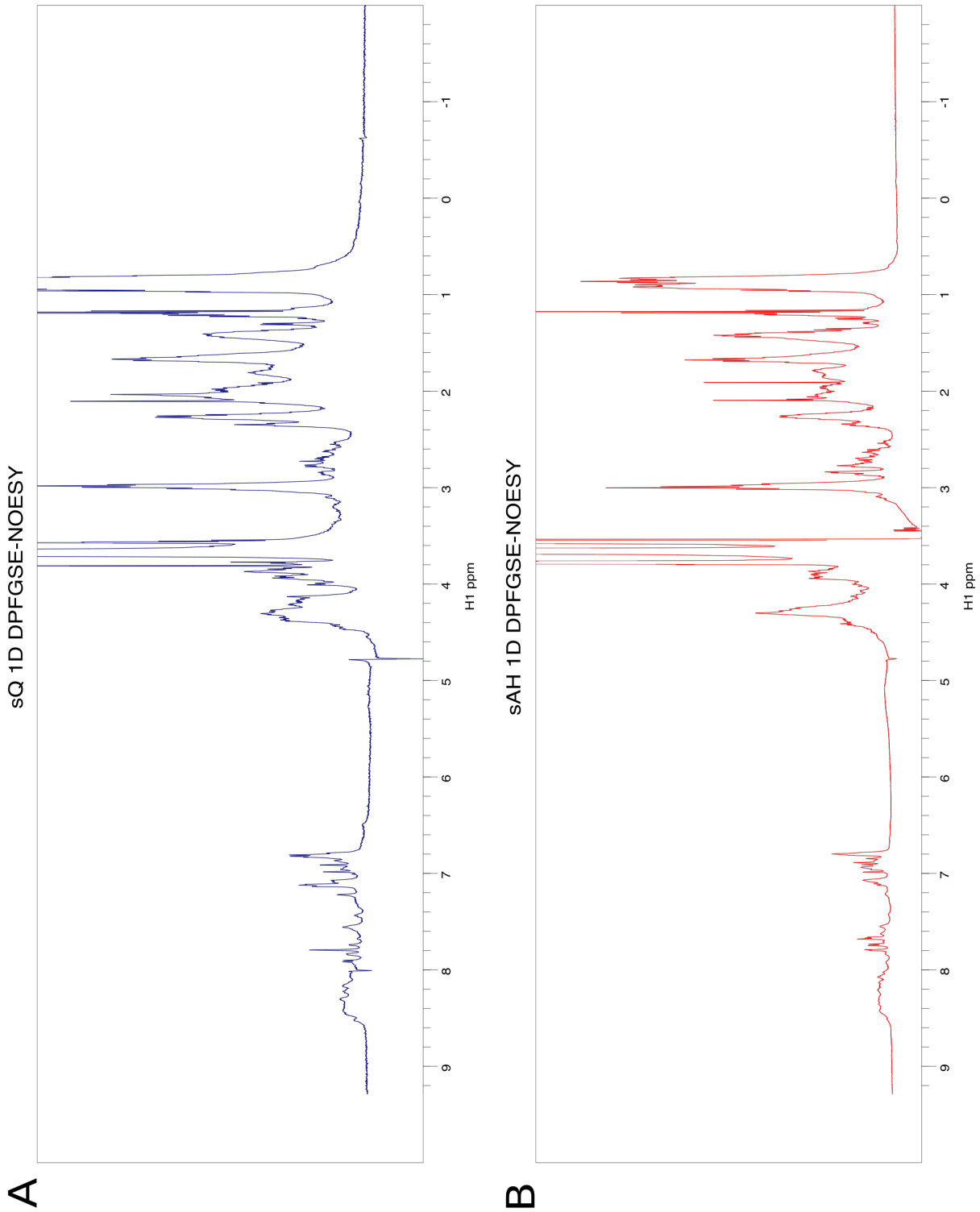


Figure A.0.1 – 1-D NOESY NMR spectra of sQ and sAH. 1-D ^{1}H NMR NOESY spectra of sQ (A) and sAH (B) were collected using a 500 MHz INOVA NMR [Agilent] magnet at the Astbury Centre, University of Leeds. For water suppression, double pulsed field gradient spin echo (DPFGSE) pulse sequence was used. The regions at 6.5-9.2 ^{1}H/ppm, 3.2-6 ^{1}H/ppm and -1-1.2 ^{1}H/ppm represent the backbone N-H, α-H and methyl H-protons respectively. These spectra show that sQ and sAH contain structured elements but are not representative of highly folded globular proteins. The negative peak at 4.7 ^{1}H/ppm represents the suppressed water signal.

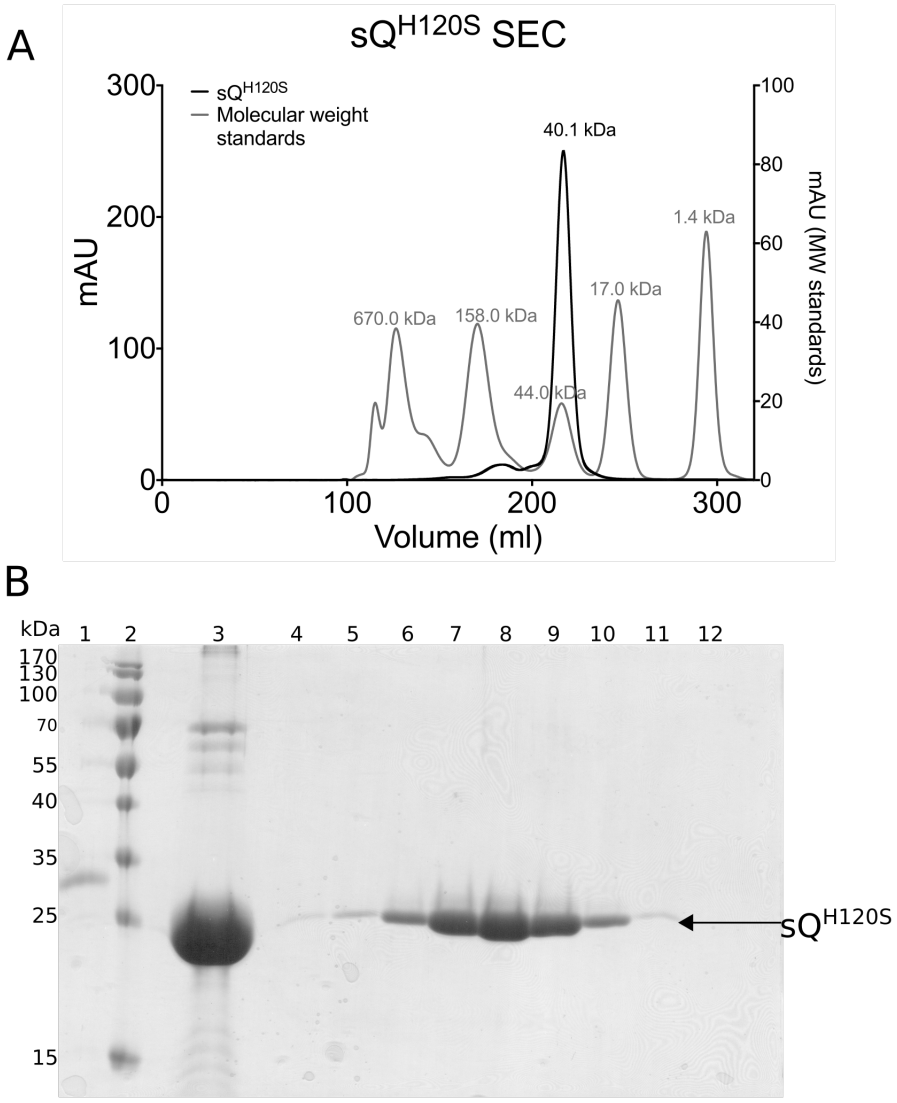


Figure A.0.2 – SEC purification of sQ^{H120S}. **A:** Size-exclusion chromatogram of sQ^{H120S} which eluted from the SEC at a mass equivalent to 40.1 kDa of the wildtype sQ (Figure 3.2.4A). The UV elution profile of molecular weight standard proteins is shown in grey and each peak is labelled with the mass of these standards. The molecular weight standards were thyroglobulin (670.0 kDa), γ-globulin (158.0 kDa), ovalbumin (44.0 kDa), myoglobin (17.0 kDa) and Vitamin B_{12} (1.4 kDa). **B:** SDS-PAGE of uncleaved Q^{H120S} (1), the sample loaded to SEC (3) and peak elution fractions (4-12) showing the protein was stable and pure.

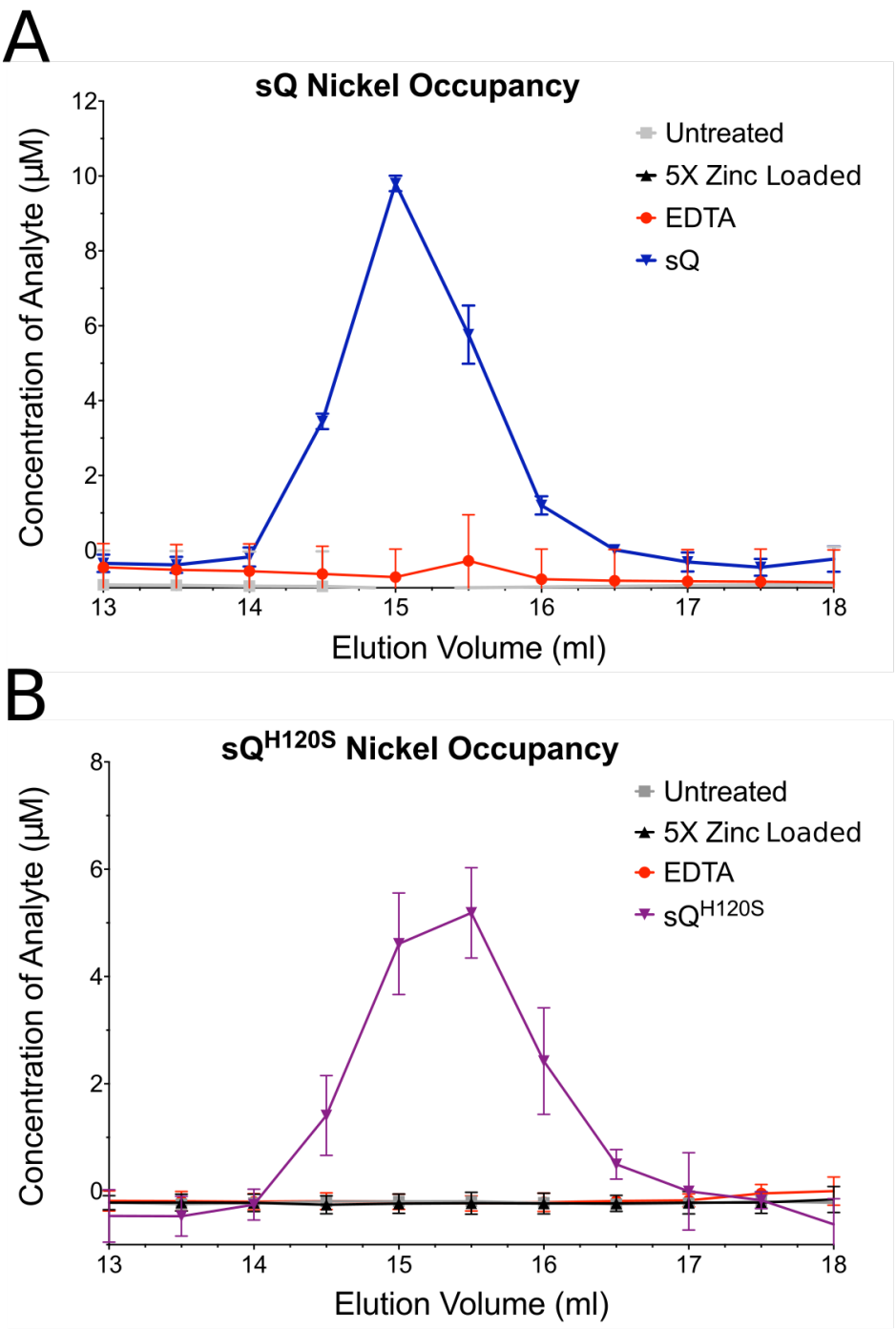


Figure A.0.3 – Nickel binding analysis of sQ and sQ^{H120S} by ICP-MS. A: 10 μM sQ was incubated with either 1 mM EDTA or a 5-fold excess of $ZnCl_2$ before injection on to an S200 10/300 GL Increase column. A 10 μM sample of sQ was left untreated and also applied to the column. Fractions of 0.5 ml were collected during elution. The metal content was analysed using ICP-MS and the protein concentration determined by absorbance at 280 nm. Ni^{2+}concentration in the fractions of of the samples are shown: 1 mM EDTA (red), untreated (grey) and 5-fold Zn^{2+} (black). The sQ protein concentration is also shown (blue). Ni^{2+} is not observed to elute with sQ. **B:** 10 μM of sQ^{H120S} was treated and analysed in an identical manner to sQ. sQ^{H120S} protein concentration is shown in purple. Ni^{2+} is not observed to elute with sQ^{H120S}.

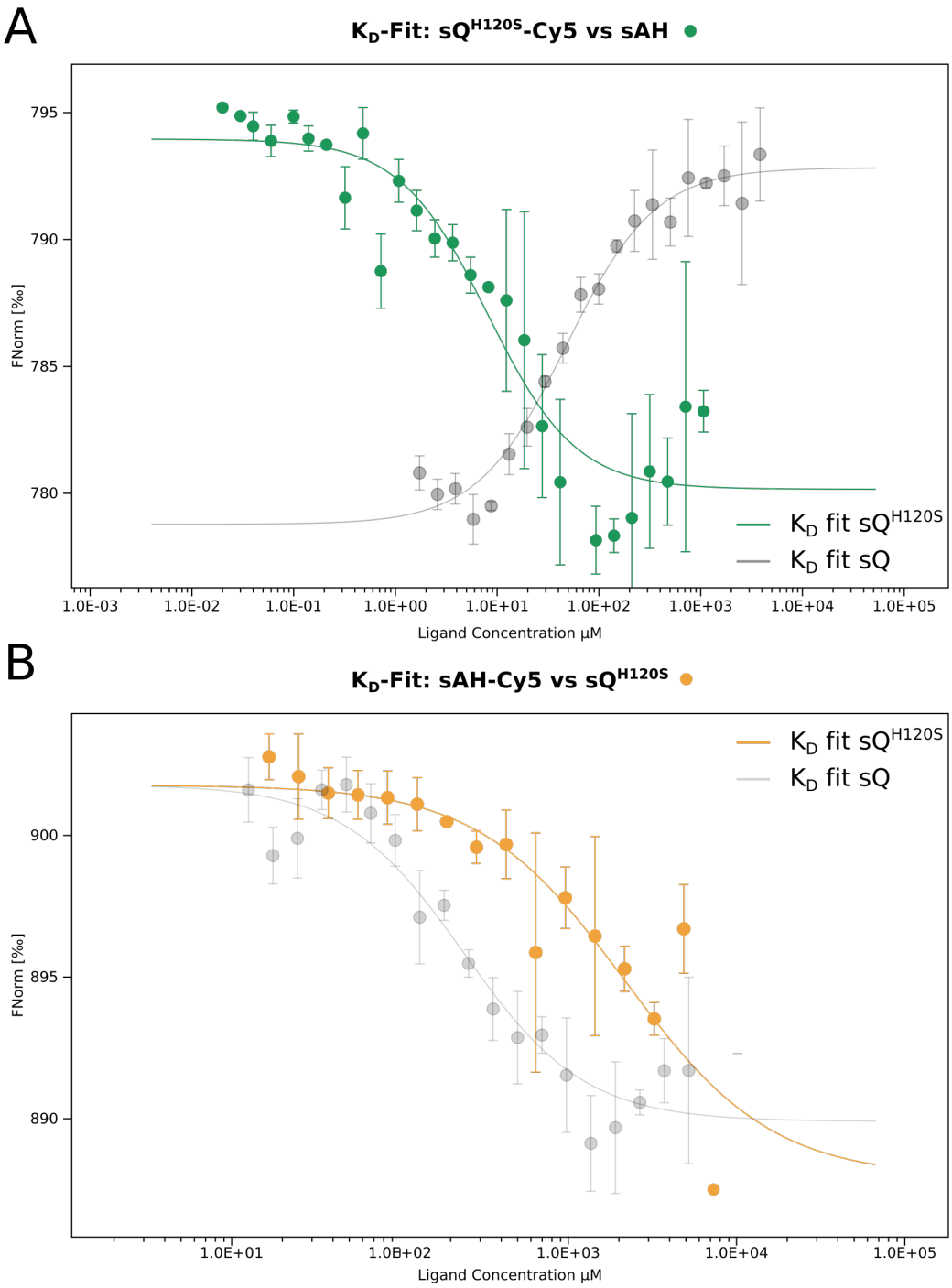


Figure A.0.4 – sQ^{H120S} vs sAH MST experiments. A: MST binding curves with sQ labelled with Cy5 fluorophore at a constant concentration of 0.01 µM and a 2:1 serial dilution of sAH at the highest concentration of 4.7 mM. The MST data of wildtype sQ are shown in grey. Due to the highly inconsistent data it was not possible to determine a reliable K_D value. **B:** sAH labelled with Cy5 fluorophore at a constant concentration of 0.17 µM with a 2:1 serial dilution of sQ at the highest concentration of 6.8 mM. The MST data using wildtype sQ are shown in grey. A K_D value could not be fitted due to the unreliable nature of the data.

	1	2	3	4	5	6
A	0.05 M Na Acetate 30 % PEG 400	0.05 M Na Acetate 35 % PEG 400	0.05 M Na Acetate 40 % PEG 400	0.05 M Na Acetate 42 % PEG 400	0.05 M Na Acetate 46 % PEG 400	0.05 M Na Acetate 50 % PEG 400
B	0.10 M Na Acetate 30 % PEG 400	0.10 M Na Acetate 35 % PEG 400	0.10 M Na Acetate 40 % PEG 400	0.10 M Na Acetate 42 % PEG 400	0.10 M Na Acetate 46 % PEG 400	0.10 M Na Acetate 50 % PEG 400
C	0.15 M Na Acetate 30 % PEG 400	0.15 M Na Acetate 35 % PEG 400	0.15 M Na Acetate 40 % PEG 400	0.15 M Na Acetate 42 % PEG 400	0.15 M Na Acetate 46 % PEG 400	0.15 M Na Acetate 50 % PEG 400
D	0.20 M Na Acetate 30 % PEG 400	0.20 M Na Acetate 35 % PEG 400	0.20 M Na Acetate 40 % PEG 400	0.20 M Na Acetate 42 % PEG 400	0.20 M Na Acetate 46 % PEG 400	0.20 M Na Acetate 50 % PEG 400

Table A.0.1 – Opt1_Levdikov: a crystallisation screen developed around the crystallisation conditions for *B. subtilis* SpoIIQ:SpoIIIAH by Levdikov *et al.,* (Levdikov *et al.,* 2012) in the determination of 3TUF. This 24 condition block was replicated 4 times (in 96-well format) at pH 4.0, 4.5, 5.0 and 5.5.

	1	2	3	4	5	6
A	2 M $(NH_4)SO_2$ 0.004 M Mg Acetate 0.05 M MES	2 M $(NH_4)SO_2$ 0.006 M Mg Acetate 0.05 M MES	2 M $(NH_4)SO_2$ 0.008 M Mg Acetate 0.05 M MES	2 M $(NH_4)SO_2$ 0.010 M Mg Acetate 0.05 M MES	2 M $(NH_4)SO_2$ 0.012 M Mg Acetate 0.05 M MES	2 M $(NH_4)SO_2$ 0.014 M Mg Acetate 0.05 M MES
B	2.5 M $(NH_4)SO_2$ 0.004 M Mg Acetate 0.05 M MES	2.5 M $(NH_4)SO_2$ 0.006 M Mg Acetate 0.05 M MES	2.5 M $(NH_4)SO_2$ 0.008 M Mg Acetate 0.05 M MES	2.5 M $(NH_4)SO_2$ 0.010 M Mg Acetate 0.05 M MES	2.5 M $(NH_4)SO_2$ 0.012 M Mg Acetate 0.05 M MES	2.5 M $(NH_4)SO_2$ 0.014 M Mg Acetate 0.05 M MES
C	2.8 M $(NH_4)SO_2$ 0.004 M Mg Acetate 0.05 M MES	2.8 M $(NH_4)SO_2$ 0.006 M Mg Acetate 0.05 M MES	2.8 M $(NH_4)SO_2$ 0.008 M Mg Acetate 0.05 M MES	2.8 M $(NH_4)SO_2$ 0.010 M Mg Acetate 0.05 M MES	2.8 M $(NH_4)SO_2$ 0.012 M Mg Acetate 0.05 M MES	2.8 M $(NH_4)SO_2$ 0.014 M Mg Acetate 0.05 M MES
D	3 M $(NH_4)SO_2$ 0.004 M Mg Acetate 0.05 M MES	3 M $(NH_4)SO_2$ 0.006 M Mg Acetate 0.05 M MES	3 M $(NH_4)SO_2$ 0.008 M Mg Acetate 0.05 M MES	3 M $(NH_4)SO_2$ 0.010 M Mg Acetate 0.05 M MES	3 M $(NH_4)SO_2$ 0.012 M Mg Acetate 0.05 M MES	3 M $(NH_4)SO_2$ 0.014 M Mg Acetate 0.05 M MES

Table A.0.2 – Opt1_Meisner: a crystallisation screen developed around the crystallisation conditions for *B. subtilis* SpoIIQ:SpoIIIAH by Meisner *et al.,* (Meisner *et al.,* 2012) in the determination of 3TUF. This 24 condition block was replicated 4 times (in 96-well format) at pH 5.6, 6.0, 6.2 and 6.5.

Appendix B

Crystal structures of PilA1 from *Clostridium difficile*

	Dataset 1			Dataset 2			Dataset 3			Dataset 4		
Software	XDS	Mosflm	DIALS	XDS	Mosflm	DIALS	XDS	Mosflm	DIALS	XDS	Mosflm	DIALS
Resolution (Å)	46.7 - 1.85	52.31 - 1.65	60.17 - 1.8	46.79 - 2.23	72.89 - 2.2	72.70 - 2.3	46.72 - 1.87	59.66 - 1.78	73.63 - 1.75	46.64 - 1.76	51.21 - 1.65	73.32 - 1.67
Unit cell dimensions												
a=b, c, (Å)	103.60, 104.74	103.62, 104.62	103.86, 104.97	102.61, 105.15	101.99, 104.21	102.82, 105.58	103.19, 104.81	102.82, 104.39	103.33, 104.94	102.78, 104.67	102.42, 104.21	102.80, 104.61
α=β=γ (°)		90.00			90.00			90.00			90.00	
Spacegroup		$P4_12_12$			$P4_12_12$			$P4_12_12$			$P4_12_12$	
R_{merge}*	0.120 (1.327)	0.177 (1.646)	0.151 (0.965)	0.074 (0.999)	0.141 (1.245)	0.134 (0.75)	0.077 (1.062)	0.108 (1.149)	0.100 (1.136)	0.103 (1.144)	0.136 (1.887)	0.124 (2.297)
Total Reflections	354141	908077	770621	197896	233046	302039	333420	688483	808333	320494	904124	941365
Unique Reflections	44796 (2978)	68857 (3371)	53740 (3111)	27964 (2524)	26237 (2340)	25772 (2483)	47291 (3025)	54192 (3046)	57815 (3138)	45939 (2919)	67115 (3391)	65512 (3288)
I/σI	8.6 (1.4)	7.8 (1.5)	9.1 (1.5)	14.8 (1.8)	31.5 (1.9)	9.7 (1.9)	13.5 (1.7)	14.0 (2.0)	13.3 (1.8)	11.2 (1.6)	10.7 (1.5)	11.0 (1.6)
Mean intensity CC1/2	0.996 (0.682)	0.996 (0.410)	0.997 (0.814)	0.999 (0.662)	0.997 (0.657)	0.998 (0.930)	0.999 (0.627)	0.999 (0.522)	0.998 (0.650)	0.998 (0.678)	0.998 (0.323)	0.998 (0.605)
Completeness (%)	99.8 (100.0)	99.9 (100.0)	100.0 (100.0)	99.9 (99.9)	99.6 (99.3)	99.9 (100.0)	99.9 (99.9)	100.0 (100.0)	100.0 (100.0)	99.6 (99.9)	100.0 (100.0)	100.0 (100.0)

Table B.0.1 – Table of dataset parameters for native R20291 PilA1$\triangle$1-34 crystals. Each of the four datasets were integrated and indexed using XDS (Kabsch, 2010), iMosflm (Battye *et al.,* 2011) and DIALS (Gildea *et al.,* 2014). The reflections and intensities were then scaled and reduced using Aimless (Evans and Murshudov, 2013; Evans, 2006; Evans, 2011), the summary of the dataset parameters is presented above. These data were collected at a wavelength of 0.92 Å. Values in parentheses represent the highest resolution shell. *$R_{merge} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is the ith observation of reflection hkl and $\langle I(hkl) \rangle$ is the massed average intensity for all observations i of reflection hkl.

N#	Structure 1				Structure 2					Interface						
	Range	N_{at}	N_{res}	Surface Å^2	Range	Symmetry op.	N_{at}	N_{res}	Surface Å^2	Interface area Å^2	△G kcal/mol	△G P-value	N_{HB}	N_{SB}	N_{DS}	CSS
1	C	59	15	7078	A	y,x,-z	62	20	7251	584.9	-0.3	0.508	14	1	0	0.0
2	C	61	21	7078	A	-y+1/2,x-1/2,z+1/4	58	17	7251	509.7	-0.6	0.468	12	1	0	0.0
3	B	61	15	7038	B	-y+1/2,x-1/2,z+1/4	59	21	7038	508.1	-0.1	0.476	12	1	0	0.0
4	C	46	16	7078	C	y,x,-z	48	16	7078	374.0	-1.3	0.415	8	0	0	0.0
5	B	37	15	7038	A	y,x,z	35	12	7251	362.8	0.0	0.513	7	0	0	0.0
6	C	32	11	7078	A	y,x,z	30	11	7251	274.9	1.2	0.657	5	0	0	0.0
7	B	36	14	7038	A	-y+1/2,x-1/2,z+1/4	23	10	7251	228.5	-2.7	0.235	1	0	0	0.0
8	C	7	3	7078	B	y,x,z	9	3	7038	77.7	0.0	0.541	1	0	0	0.0

Table B.0.2 – PISA results for R20291 PilA1△1-34. Interfaces between the protein chains A, B and C of R20291 PilA1△1-34 were analysed, two chains at a time. The chain, number of interfacing atoms (N_{at}), number of interfacing residues (N_{res}) and total solvent accessible area (Å^2) for the two interacting chains are described. The symmetry operation that must be applied to the second structure to obtain the respective interface is given. The surface area (Å^2) of the interface, the solvation free energy upon complex formation (△G kcal/mol) and the specificity of the formation (△G P-value) as well as the number and type of bonding: hydrogen-bonding (N_{HB}); salt-bridges (N_{SB}); disulphide bonds (N_{DS}) are indicated. Complex formation significance score (CSS), describes the biological significance of the interface, a value of 1.0 is highly significant and 0.0 is not biologically significant.

N#	Structure 1				Structure 2					Interface						
	Range	N_{at}	N_{res}	Surface Å^2	Range	Symmetry op.	N_{at}	N_{res}	Surface Å^2	Interface area Å^2	△G kcal/mol	△G P-value	N_{HB}	N_{SB}	N_{DS}	CSS
1	A	53	13	7001	A	-x-1,y-1/2,-z-1/2	44	13	7001	463.1	-3.6	0.237	3	1	0	0.0
2	C	51	15	7011	B	x-1,y,z	43	12	6586	444.1	-3.9	0.261	6	0	0	0.0
3	C	41	13	7011	C	x-1,y,z	45	16	7011	376.2	-3.6	0.311	5	0	0	0.0
4	C	34	11	7011	B	x-1/2,-y+1/2,-z	27	9	6586	311.3	-1.0	0.514	3	2	0	0.0
5	A	30	9	7001	C	-x-1,y-1/2,-z-1/2	25	7	7011	298.1	-1.7	0.406	3	1	0	0.0
6	B	32	13	6586	B	x-1,y,z	29	11	6586	269.7	0.6	0.698	5	1	0	0.0
7	A	31	12	7001	A	-x,y-1/2,-z-1/2	25	8	7001	259.0	2.9	0.799	6	1	0	0.0
8	B	27	9	6586	C	x,y-1,z	21	6	7011	214.2	1.1	0.654	1	0	0	0.0
9	B	27	9	6586	A	y,x,z	23	8	7001	255.6	-2.8	0.253	2	0	0	0.0
10	A	25	10	7001	A	x-1,y,z	16	5	7001	169.5	0.1	0.557	2	0	0	0.0
11	C	4	2	7011	B	x-3/2,-y+1/2,-z	11	5	6586	70.8	0.1	0.592	0	0	0	0.0
12	C	6	4	7011	B	y,x,z	9	6	6586	48.7	-0.2	0.505	0	0	0	0.0

Table B.0.3 – PISA results for 630 PilA1△1-34. The chain, number of interfacing atoms (N_{at}), number of interfacing residues (N_{res}) and total solvent accessible area (Å^2) for the two interacting chains are described. The symmetry operation that must be applied to the second structure to obtain the respective interface is given. The surface area (Å^2) of the interface, the solvation free energy upon complex formation (△G kcal/mol) and the specificity of the formation (△G P-value). The number and type of bonding: hydrogen-bonding (N_{HB}); salt-bridges (N_{SB}); disulphide bonds (N_{DS}) are indicated. Complex formation significance score (CSS), describes the biological significance of the interface, a value of 1.0 is highly significant and 0.0 is not significant.

Appendix C

Structural studies of Type IV minor pilins from *C. difficile*

	1	2	3	4	5	6
A	0.1 M HEPES pH 7.5 0 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES pH 7.5 0 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES pH 7.5 0 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES pH 7.5 0 M $MgCl_2$ 32 % PEG 400	0.1 M HEPES pH 7.5 0 M $MgCl_2$ 34 % PEG 400	0.1 M HEPES pH 7.5 0 M $MgCl_2$ 36 % PEG 400
B	0.1 M HEPES pH 7.5 0.075 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES pH 7.5 0.075 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES pH 7.5 0.075 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES pH 7.5 0.075 M $MgCl_2$ 32 % PEG 400	0.1 M HEPES pH 7.5 0.075 M $MgCl_2$ 34 % PEG 400	0.1 M HEPES pH 7.5 0.075 M $MgCl_2$ 36 % PEG 400
C	0.1 M HEPES pH 7.5 0.150 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES pH 7.5 0.150 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES pH 7.5 0.150 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES pH 7.5 0.150 M $MgCl_2$ 32 % PEG 400	0.1 M HEPES pH 7.5 0.150 M $MgCl_2$ 34 % PEG 400	0.1 M HEPES pH 7.5 0.150 M $MgCl_2$ 36 % PEG 400
D	0.1 M HEPES pH 7.5 0.225 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES pH 7.5 0.225 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES pH 7.5 0.225 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES pH 7.5 0.225 M $MgCl_2$ 32 % PEG 400	0.1 M HEPES pH 7.5 0.225 M $MgCl_2$ 34 % PEG 400	0.1 M HEPES pH 7.5 0.225 M $MgCl_2$ 36 % PEG 400

Table C.0.1 – PilK△1-32 Optimisation Screen #1. The HEPES buffer was prepared at pH 7.0, pH 7.2, pH 7.5 and pH 7.7. One of these pH values would be selected for a 24-well screen or all four used in a 96-well screen.

	1	2	3	4	5	6
A	0.1 M HEPES 0.075 M $MgCl_2$ 24 % PEG 400	0.1 M HEPES 0.075 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES 0.075 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES 0.075 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES 0.075 M $MgCl_2$ 32 % PEG 400	0.1 M HEPES 0.075 M $MgCl_2$ 34 % PEG 400
B	0.1 M HEPES 0.125 M $MgCl_2$ 24 % PEG 400	0.1 M HEPES 0.125 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES 0.125 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES 0.125 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES 0.125 M $MgCl_2$ 32 % PEG 400	0.1 M HEPES 0.125 M $MgCl_2$ 34 % PEG 400
C	0.1 M HEPES 0.150 M $MgCl_2$ 24 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 32 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 34 % PEG 400
D	0.1 M HEPES 0.2 M $MgCl_2$ 24% PEG 400	0.1 M HEPES 0.2 M $MgCl_2$ 26% PEG 400	0.1 M HEPES 0.2 M $MgCl_2$ 28% PEG 400	0.1 M HEPES 0.2 M $MgCl_2$ 30% PEG 400	0.1 M HEPES 0.2 M $MgCl_2$ 32% PEG 400	0.1 M HEPES 0.2 M $MgCl_2$ 34% PEG 400

Table C.0.2 – PilK△1-32 Optimisation Screen #2. The HEPES buffer was prepared at pH 7.0, pH 7.2, pH 7.5 and pH 7.7. One of these pH values would be selected for a 24-well screen or all four used in a 96-well screen.

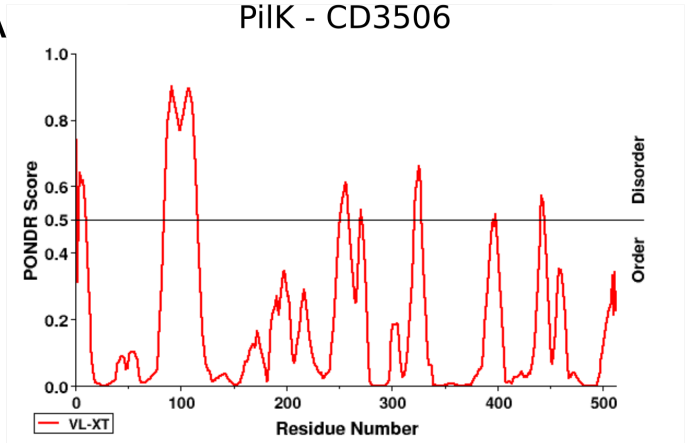
	1	2	3	4	5	6	7	8	9	10	11	12
A	0.1 M HEPES 0.150 M $MgCl_2$ 18 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 20 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 21 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 22 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 23 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 24 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 25 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 27 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES 0.150 M $MgCl_2$ 0 % PEG 400
B	0.1 M HEPES 0.175 M $MgCl_2$ 18 % PEG 400	0.1 M HEPES 0.175 M $MgCl_2$ 20 % PEG 400	0.1 M HEPES 0.175 M $MgCl_2$ 21 % PEG 400	0.1 M HEPES 0.175 M $MgCl_2$ 22 % PEG 400	0.1 M HEPES 0.175 M $MgCl_2$ 23 % PEG 400	0.1 M HEPES 0.175 M $MgCl_2$ 24 % PEG 400	0.1 M HEPES 0.175 M $MgCl_2$ 25 % PEG 400	0.1 M HEPES 0.175 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES 0.175 M $MgCl_2$ 27 % PEG 400	0.1 M HEPES 0.175 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES 0.175 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES 0.175 M $MgCl_2$ 0 % PEG 400
C	0.1 M HEPES 0.200 M $MgCl_2$ 18 % PEG 400	0.1 M HEPES 0.200 M $MgCl_2$ 20 % PEG 400	0.1 M HEPES 0.200 M $MgCl_2$ 21 % PEG 400	0.1 M HEPES 0.200 M $MgCl_2$ 22 % PEG 400	0.1 M HEPES 0.200 M $MgCl_2$ 23 % PEG 400	0.1 M HEPES 0.200 M $MgCl_2$ 24 % PEG 400	0.1 M HEPES 0.200 M $MgCl_2$ 25 % PEG 400	0.1 M HEPES 0.200 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES 0.200 M $MgCl_2$ 27 % PEG 400	0.1 M HEPES 0.200 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES 0.200 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES 0.200 M $MgCl_2$ 0 % PEG 400
D	0.1 M HEPES 0.250 M $MgCl_2$ 18 % PEG 400	0.1 M HEPES 0.250 M $MgCl_2$ 20 % PEG 400	0.1 M HEPES 0.250 M $MgCl_2$ 21 % PEG 400	0.1 M HEPES 0.250 M $MgCl_2$ 22 % PEG 400	0.1 M HEPES 0.250 M $MgCl_2$ 23 % PEG 400	0.1 M HEPES 0.250 M $MgCl_2$ 24 % PEG 400	0.1 M HEPES 0.250 M $MgCl_2$ 25 % PEG 400	0.1 M HEPES 0.250 M $MgCl_2$ 26 % PEG 400	0.1 M HEPES 0.250 M $MgCl_2$ 27 % PEG 400	0.1 M HEPES 0.250 M $MgCl_2$ 28 % PEG 400	0.1 M HEPES 0.250 M $MgCl_2$ 30 % PEG 400	0.1 M HEPES 0.250 M $MgCl_2$ 0 % PEG 400

Table C.0.3 – PilK△1-32 Optimisation Screen #3. The HEPES buffer was prepared at pH 7.2.

Protein	Number of repeats	Total score	Coverage	Length of repeat
PilK	3	299	54-115	61
			121-209	88
			217-301	84
GspK	2	300	39-157	118
			162-270	108

Table C.0.4 – **Radar results for *C. difficile* PilK and *E. coli* GspK.** Radar was used to search for internal repeats in the PilK and GspK peptide sequences. Three possible internal repeats were found in PilK of an average length of 78 residues and two internal repeats were detected in GspK of an average number of 113 residues.(Heger and Holm, 2000; Goujon *et al.,* 2010)

A

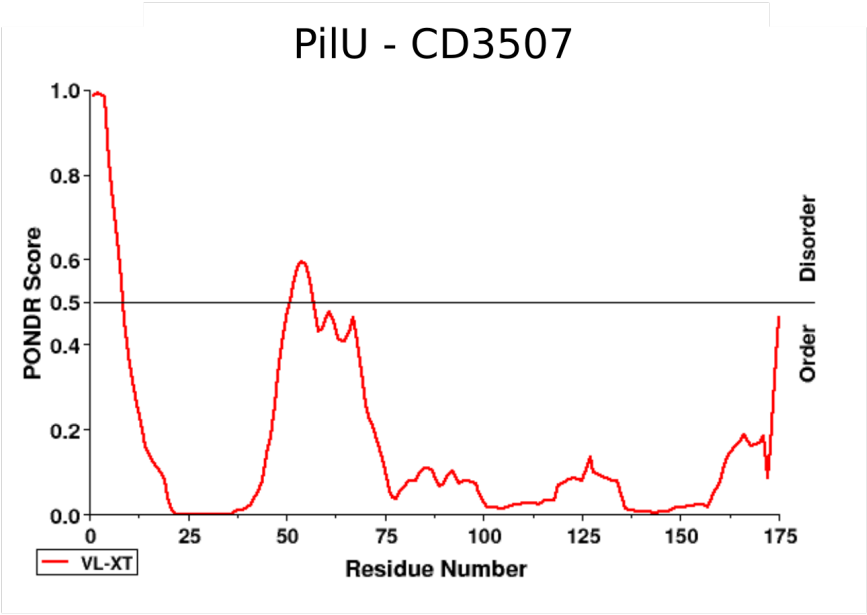


B

```
==============================VLXT NNP STATISTICS==============================
Predicted residues: 512                 Number Disordered Regions: 8
Number residues disordered: 64          Longest Disordered Region: 32
Overall percent disordered: 12.50       Average Prediction Score: 0.1861
Predicted disorder segment [3]-[9]      Average Strength= 0.5975
Predicted disorder segment [84]-[115]   Average Strength= 0.7943
Predicted disorder segment [251]-[258]  Average Strength= 0.5657
Predicted disorder segment [270]-[271]  Average Strength= 0.5238
Predicted disorder segment [321]-[328]  Average Strength= 0.5967
Predicted disorder segment [396]-[396]  Average Strength= 0.5012
Predicted disorder segment [398]-[398]  Average Strength= 0.5144
Predicted disorder segment [441]-[444]  Average Strength= 0.5457

==============================PREDICTOR OUTPUT================================
          "D" = Disordered                          " " = Ordered
================================================================================

1       MRKWNKFKSE RGAALVLVLV VVALLSIVGL IFSNQIANRI KSTKTTNEGI
VLXT    D DDDDDDD

51      QAKYLAETCV ENSIDKAYEK LYDELEKMDN EFKSENQEKS ISRSKLRNIS
VLXT                                    DDDDDDD DDDDDDDDDD

101     DEDFNNQDEE NIEAERLGYM NNINFYLNKA SSDLEKASME LKKLYDLDML
VLXT    DDDDDDDDDD DDDDD

151     DYRDIEYVDA NIISHRDSIL EICKNYTSGD ISKINEYILE EDIDSTTLIE
VLXT

201     AKLVNNDILL KMFLEENKIE NEHLNSAFSH TYKALDNISL AMQNMIEYRH
VLXT

251     TFHIDEPKVE VSNGIPDSQQ YYELIQNPII NSMEYIWNSK WDTLENLLEI
VLXT    DDDDDDDD           D D

301     LPNQTQGFNS LRVHLRNNVR KFEKLSDNIS SGKKNTAKNF LKYKELLYEI
VLXT                        DDDDDDDD

351     SDQCNQLKSM SYEKIPVKYD NMALITTFDY IQNELLAEIK CRLKELKPQE
VLXT                                              D D

401     IDKTEGITIK IPFYKADYDM TKEGWPKLKE NGSGAELSLM VTGDKNGIKE
VLXT                                        DDDD

451     VEVTDGKKNI IGLGVEENSN SKYKVDAIVN FNLNIDTNVV GNYDIKDKIL
VLXT

501     INHDISSYKK VN
VLXT
```

Figure C.0.1 – **PONDR disorder prediction of PilK-CD3506. A:** Graphical output of PONDR score vs residue number. Residues scored below 0.5 are ordered and above 0.5 are disordered (Xue *et al.,* 2010). **B:** Summary of the disordered regions which are also annotated (red boxes) on the peptide sequence. Disordered residues are marked below the sequence with a D. The largest region of predicted disorder in the PilK peptide sequence includes 32 residues at 84-115.

A

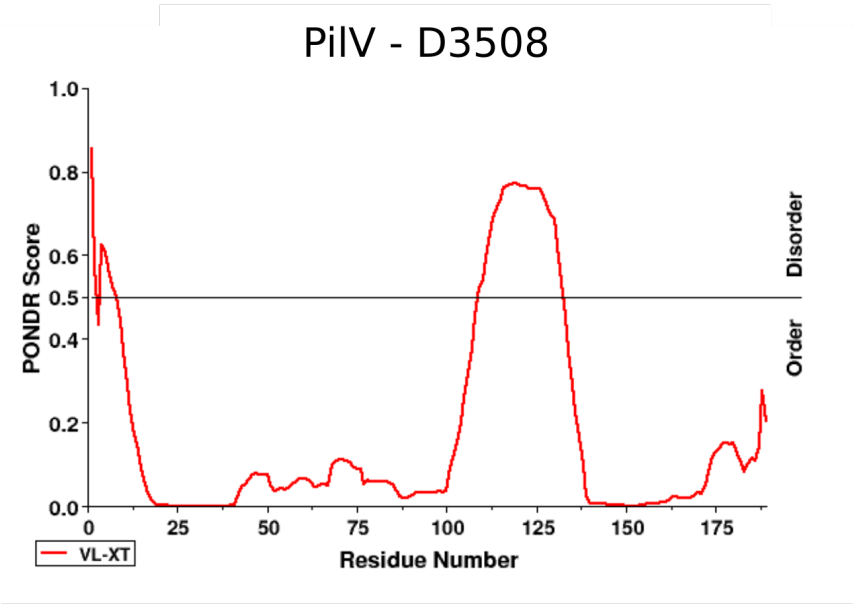


B

```
================================VLXT NNP STATISTICS================================
Predicted residues: 175                 Number Disordered Regions: 2
Number residues disordered: 14          Longest Disordered Region: 8
Overall percent disordered: 8.00        Average Prediction Score: 0.1612
Predicted disorder segment [1]-[8]      Average Strength= 0.8390
Predicted disorder segment [51]-[56]    Average Strength= 0.5635

================================PREDICTOR OUTPUT================================
           "D" = Disordered                          " " = Ordered
================================================================================

1       MSKKNSERGF SLIEVLVAMA IMGIVLFAFF NIINTNNKAN TKNDTDITSL
VLXT    DDDDDDDD

51      NYVQSEIENL REKIKSGEFD FDSLDKLEDG TVVYEKLIDK SKKVVYDKVL
VLXT    DDDDDD

101     SEGDVSLYDT PYEKITTIKD EDGNLIDKGN ITNKIKTIVE DKSGQIYKIT
VLXT

151     VTGKSMNDYS SKKEVKIVTE IFKDK
VLXT
```

Figure C.0.2 – **PONDR disorder prediction of PilU-CD3507. A:** Graphical output of PONDR score vs residue number. Residues scored below 0.5 are ordered and above 0.5 are disordered (Xue *et al.,* 2010). **B:** Summary of the disordered regions which are also annotated (red boxes) on the peptide sequence. Disordered residues are marked below the sequence with a D. The largest region of predicted disorder in the PilU peptide sequence is the N-terminal 8 residues, the only other region includes 5 residues from 51-56.

A



B

```
================================VLXT NNP STATISTICS================================
Predicted residues: 189                 Number Disordered Regions: 3
Number residues disordered: 30          Longest Disordered Region: 24
Overall percent disordered: 15.87       Average Prediction Score: 0.1747
Predicted disorder segment [1]-[2]      Average Strength= 0.7136
Predicted disorder segment [4]-[7]      Average Strength= 0.5755
Predicted disorder segment [109]-[132]  Average Strength= 0.7017

================================PREDICTOR OUTPUT================================
           "D" = Disordered                          " " = Ordered
================================================================================

1       MMSLYKNNEK GLTLLEVIIA VFILTIVLSI SYKVFNGITS AVKKQQIITD
VLXT    DD DDDD

51      AQVNINLINK YLNRDLENCK ELTKTGSGNN YEYNIEMPDN VVKYEVSIET
VLXT

101     KKNTEVYSVT RIQNNTIDTE NEVREEIIYN QPLVQNNKEM KETPFKIEKQ
VLXT             DD DDDDDDDDDD DDDDDDDDDD DD

151     TGKSIYTVSI YYNESVQESH KNSNLNNKTY TFDVMSRIG
VLXT
```

Figure C.0.3 – **PONDR disorder prediction of PilV-CD3508.A:** Graphical output of PONDR score vs residue number. Residues scored below 0.5 are ordered and above 0.5 are disordered (Xue *et al.,* 2010). **B:** Summary of the disordered regions which are also annotated (red boxes) on the peptide sequence. Disordered residues are marked below the sequence with a D. A 24 residue region of disorder is predicted between residues 109-132.

Appendix D

Professional Internship for Postgraduate students

For 11 weeks (Sept-Nov 2014) the professional internship for postgraduate students (PIPs) was completed at Diamond Light Source under the supervision of VMXm principal beamline scientist Dr Gwyndaf Evans. Working with the VMXm group I developed a technique to protect protein crystals from harsh environments such as *in vacuo* and room temperature conditions using multi-layer graphene. The work involved understanding the multi-layer graphene material using resources at Diamond such as interferometry, designing new tools to wrap the crystals using 3D printing and testing the quality of standard crystals (lysozyme, thaumatin and glucose isomerase) wrapped in graphene on the beamline I04 during *in vacuo* and room temperature data collection. By wrapping protein crystals in multi-layer graphene datasets with improved diffraction quality and lifetime in these extreme conditions were achieved over those that were not protected in graphene. This technique is now in trial on the new long-wavelength beamline I23. These results have also formed the basis of a paper published in Acta Crystallographica section D (Warren *et al.,* 2015). The PIPs was a valuable experience and enabled me to gain experience of methods development at a cutting edge facility and develop skills that I was able to apply to my PhD studies.

Publications

- **Crawshaw AD**, Serrano M, Stanley WA, Henriques AO and Salgado PS. A mother cell-to-forespore channel: current understanding and future challenges. FEMS Microbiology Letters. 2014; 358(2):129-136
- Warren AJ, **Crawshaw AD**, Trincao J, Aller P, Alcock S, Nistea I, Salgado PS and Evans G In vacuo X-ray data collection from graphene-wrapped protein crystals. Acta Crystallographica Section D. 2015;71(10):2079-2088
- Serrano M, **Crawshaw AD**, Dembek M, Monteiro JM, Pereira FC, de Pinho MG, Fairweather NF, Salgado PS and Henriques AO. The SpoIIQ-SpoIIIAH complex of Clostridium difficile controls forespore engulfment and late stages of gene expression and spore morphogenesis. Molecular Microbiology. 2015;100(1):204-28

Bibliography

Abgottspon, D., Rölli, G., Hosch, L., Steinhuber, A., Jiang, X., Schwardt, O., Cutting, B., Smiesko, M., Jenal, U., Ernst, B. and Trampuz, A. (2010). 'Development of an aggregation assay to screen FimH antagonists'. In: *Journal of Microbiological Methods* 82.3, pp. 249–255.

Abt, M. C., McKenney, P. T. and Pamer, E. G. (2016). '*Clostridium difficile* colitis: pathogenesis and host defence.' In: *Nature Reviews Microbiology* 14.10, pp. 609–20.

Akerlund, T., Persson, I., Unemo, M., Noren, T., Svenungsson, B., Wullt, M. and Burman, L. G. (2008). 'Increased sporulation rate of epidemic *Clostridium difficile* type 027/NAP1'. In: *Journal of Clinical Microbiology* 46.4, pp. 1530–1533.

Ammann, A. A. (2007). 'Inductively coupled plasma mass spectrometry (ICP MS): a versatile tool.' In: *Journal of mass spectrometry : JMS* 42.4, pp. 419–427.

Arnold, K., Bordoli, L., Kopp, J. and Schwede, T. (2006). 'The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling'. In: *Bioinformatics* 22.2, pp. 195–201.

Aung, S., Shum, J., Abanes-De Mello, A., Broder, D. H., Fredlund-Gutierrez, J., Chiba, S. and Pogliano, K. (2007). 'Dual localization pathways for the engulfment proteins during *Bacillus subtilis* sporulation.' In: *Molecular microbiology* 65.6, pp. 1534–1546.

Ayers, M., Sampaleanu, L. M., Tammam, S., Koo, J., Harvey, H., Howell, P. L. and Burrows, L. L. (2009). 'PilM/N/O/P proteins form an inner membrane complex that affects the stability of the Pseudomonas aeruginosa type IV pilus secretin.' In: *Journal of Molecular Biology* 394.1, pp. 128–142.

Baaske, P., Wienken, C. J., Reineck, P., Duhr, S. and Braun, D. (2010). 'Optical thermophoresis for quantifying the buffer dependence of aptamer binding.' In: *Angewandte Chemie (International ed. in English)* 49.12, pp. 2238–2241.

Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. and Leslie, A. G. W. (2011). 'iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM.' In: *Acta crystallographica Section D, Biological crystallography* 67.Pt 4, pp. 271–281.

Biais, N., Higashi, D. L., Brujic, J., So, M. and Sheetz, M. P. (2010). 'Force-dependent polymorphism in type IV pili reveals hidden epitopes'. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.25, pp. 11358–11363.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L. and Schwede, T. (2014). 'SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information.' In: *Nucleic acids research* 42, W252–8.

Blaylock, B. (2004). 'Zipper-like interaction between proteins in adjacent daughter cells mediates protein localization'. In: *Genes & Development* 18.23, pp. 2916–2928.

Bordeleau, E. and Burrus, V. (2015). 'Cyclic-di-GMP signaling in the Gram-positive pathogen *Clostridium difficile*.' In: *Current genetics* 61.4, pp. 497–502.

Bordeleau, E., Purcell, E. B., Lafontaine, D. A., Fortier, L.-C., Tamayo, R. and Burrus, V. (2015). 'Cyclic di-GMP riboswitch-regulated type IV pili contribute to aggregation of *Clostridium difficile*.' In: *Journal of bacteriology* 197.5, pp. 819–832.

Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J. and Schwede, T. (2009). 'Protein structure homology modeling using SWISS-MODEL workspace.' In: *Nature Protocols* 4.1, pp. 1–13.

Brissac, T., Mikaty, G., Duménil, G., Coureuil, M. and Nassif, X. (2012). 'The meningococcal minor pilin PilX is responsible for type IV pilus conformational changes associated with signaling to endothelial cells.' In: *Infection and Immunity* 80.9, pp. 3297–3306.

Brown, L., Wolf, J. M., Prados-Rosales, R. and Casadevall, A. (2015). 'Through the wall: extracellular vesicles in Gram-positive bacteria, mycobacteria and fungi.' In: *Nature Reviews Microbiology* 13.10, pp. 620–630.

Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K. and Jones, D. T. (2013). 'Scalable web services for the PSIPRED Protein Analysis Workbench.' In: *Nucleic acids research* 41.Web Server issue, W349–57.

Buffie, C. G. (2013). 'Microbiota-mediated colonization resistance against intestinal pathogens.' In: *Nature reviews. Immunology* 13.11, pp. 790–801.

Bujacz, G., Wrzesniewska, B. and Bujacz, A. (2010). 'Cryoprotection properties of salts of organic acids: a case study for a tetragonal crystal of HEW lysozyme'. In: *Acta crystallographica Section D, Biological crystallography* 66.7, pp. 789–796.

Burns, D. A., Heeg, D., Cartman, S. T. and Minton, N. P. (2011). 'Reconsidering the sporulation characteristics of hypervirulent *Clostridium difficile* BI/NAP1/027.' In: *PLoS ONE* 6.9, e24894.

Camp, A. H. and Losick, R. (2009). 'A feeding tube model for activation of a cell-specific transcription factor during sporulation in *Bacillus subtilis*'. In: *Genes & Development* 23.8, pp. 1014–1024.

Camp, A. H. and Losick, R. (2008). 'A novel pathway of intercellular signalling in *Bacillus subtilis* involves a protein with similarity to a component of type III secretion channels'. In: *Molecular microbiology* 69.2, pp. 402–417.

Camp, A. H., Wang, A. F. and Losick, R. (2011). 'A small protein required for the switch from sigmaF to sigmaG during sporulation in *Bacillus subtilis*.' In: *Journal of bacteriology* 193.1, pp. 116–124.

Carter, G. P., Rood, J. I. and Lyras, D. (2012). 'The role of toxin A and toxin B in the virulence of *Clostridium difficile*.' In: *Trends in Microbiology* 20.1, pp. 21–29.

Cartman, S. T., Heap, J. T., Kuehne, S. A., Cockayne, A. and Minton, N. P. (2010). 'The emergence of hypervirulence in *Clostridium difficile*'. In: *International Journal of Medical Microbiology* 300.6, pp. 387–395.

Cehovin, A., Simpson, P. J., McDowell, M. A., Brown, D. R., Noschese, R., Pallett, M., Brady, J., Baldwin, G. S., Lea, S. M., Matthews, S. J. and Pelicic, V. (2013). 'Specific DNA recognition mediated by a type IV pilin.' In: *Proceedings of the National Academy of Sciences* 110.8, pp. 3065–3070.

Chang, Y. W., Rettberg, L. A., Treuner-Lange, A., Iwasa, J., Sogaard-Andersen, L. and Jensen, G. J. (2016). 'Architecture of the type IVa pilus machine'. In: *Science* 351.6278, pp. 2001–2001.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. and Richardson, D. C. (2010). 'MolProbity: all-atom structure validation for macromolecular crystallography.' In: *Acta crystallographica Section D, Biological crystallography* 66.Pt 1, pp. 12–21.

Chiba, S., Coleman, K. and Pogliano, K. (2007). 'Impact of membrane fusion and proteolysis on SpoIIQ dynamics and interaction with SpoIIIAH.' In: *The Journal of biological chemistry* 282.4, pp. 2576–2586.

Choudhury, D., Thompson, A., Stojanoff, V., Langermann, S., Pinkner, J., Hultgren, S. J. and Knight, S. D. (1999). 'X-ray structure of the FimC-FimH chaperone-adhesin complex from uropathogenic *Escherichia coli*.' In: *Science* 285.5430, pp. 1061–1066.

Compton, L. A. and Johnson, W. C. (1986). 'Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication.' In: *Analytical biochemistry* 155.1, pp. 155–167.

Cowtan, K. (2006). 'The Buccaneer software for automated model building. 1. Tracing protein chains'. In: *Acta crystallographica Section D, Biological crystallography* 62.9, pp. 1002–1011.

Craig, L. and Li, J. (2008). 'Type IV pili: paradoxes in form and function.' In: *Current Opinion in Structural Biology* 18.2, pp. 267–277.

Craig, L., Pique, M. E. and Tainer, J. A. (2004). 'Type IV pilus structure and bacterial pathogenicity'. In: *Nature Reviews Microbiology* 2.5, pp. 363–378.

Craig, L., Taylor, R. K., Pique, M. E., Adair, B. D., Arvai, A. S., Singh, M., Lloyd, S. J., Shin, D. S., Getzoff, E. D., Yeager, M., Forest, K. T. and Tainer, J. A. (2003). 'Type IV pilin structure and assembly: X-ray and EM analyses of *Vibrio cholerae* toxin-coregulated pilus and *Pseudomonas aeruginosa* PAK pilin.' In: *Molecular cell* 11.5, pp. 1139–1150.

Craig, L., Volkmann, N., Arvai, A. S., Pique, M. E., Yeager, M., Egelman, E. H. and Tainer, J. A. (2006). 'Type IV pilus structure by cryo-electron microscopy and crystallography: implications for pilus assembly and functions.' In: *Molecular cell* 23.5, pp. 651–662.

Crawshaw, A. D., Serrano, M., Stanley, W. A., Henriques, A. O. and Salgado, P. S. (2014). 'A mother cell-to-forespore channel: current understanding and future challenges.' In: *FEMS microbiology letters* 358.2, pp. 129–136.

Crooks, G. E., Hon, G., Chandonia, J.-M. and Brenner, S. E. (2004). 'WebLogo: a sequence logo generator.' In: *Genome research* 14.6, pp. 1188–1190.

Dapa, T. and Unnikrishnan, M. (2013). 'Biofilm formation by *Clostridium difficile*.' In: *Gut microbes* 4.5, pp. 397–402.

Dauter, Z., Dauter, M. and Dodson, E. (2002). 'Jolly SAD.' In: *Acta crystallographica Section D, Biological crystallography* 58.Pt 3, pp. 494–506.

Deakin, L. J., Clare, S., Fagan, R. P., Dawson, L. F., Pickard, D. J., West, M. R., Wren, B. W., Fairweather, N. F., Dougan, G. and Lawley, T. D. (2012). 'The *Clostridium difficile* spo0A Gene Is a Persistence and Transmission Factor'. In: *Infection and Immunity* 80.8, pp. 2704–2711.

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. and Bax, A. (1995). 'NMRPipe: a multidimensional spectral processing system based on UNIX pipes.' In: *Journal of biomolecular NMR* 6.3, pp. 277–293.

Dembek, M., Barquist, L., Boinett, C. J., Cain, A. K., Mayho, M., Lawley, T. D., Fairweather, N. F. and Fagan, R. P. (2015). 'High-throughput analysis of gene essentiality and sporulation in *Clostridium difficile*.' In: *mBio* 6.2, e02383.

Duhr, S. and Braun, D. (2006a). 'Optothermal molecule trapping by opposing fluid flow with thermophoretic drift.' In: *Physical review letters* 97.3, p. 038103.

Duhr, S. and Braun, D. (2006b). 'Why molecules move along a temperature gradient.' In: *Proceedings of the National Academy of Sciences of the United States of America* 103.52, pp. 19678–19682.

Dümmler, A., Lawrence, A.-M. and Marco, A. de (2005). 'Simplified screening for the detection of soluble fusion constructs expressed in *Escherichia coli* using a modular set of vectors.' In: *Microbial cell factories* 4, p. 34.

Eddy, S. R. (1998). 'Profile hidden Markov models.' In: *Bioinformatics* 14.9, pp. 755–763.

Eichenberger, P., Fujita, M., Jensen, S. T., Conlon, E. M., Rudner, D. Z., Wang, S. T., Ferguson, C., Haga, K., Sato, T., Liu, J. S. and Losick, R. (2004). 'The program of gene transcription for a single differentiating cell type during sporulation in *Bacillus subtilis*.' In: *PLoS Biology* 2.10, e328.

Eichenberger, P., Jensen, S. T., Conlon, E. M., Ooij, C. van, Silvaggi, J., González-Pastor, J. E., Fujita, M., Ben-Yehuda, S., Stragier, P., Liu, J. S. and Losick, R. (2003). 'The sigmaE regulon and the identification of additional sporulation genes in *Bacillus subtilis*.' In: *Journal of Molecular Biology* 327.5, pp. 945–972.

Eijk, E. van, Anvar, S. Y., Browne, H. P., Leung, W. Y., Frank, J., Schmitz, A. M., Roberts, A. P. and Smits, W. K. (2015). 'Complete genome sequence of the *Clostridium difficile*

laboratory strain 630∆erm reveals differences from strain 630, including translocation of the mobile element CTn5.' In: *BMC Genomics* 16.1, p. 31.

Emsley, P., Lohkamp, B., Scott, W. G. and Cowtan, K. (2010). 'Features and development of Coot.' In: *Acta crystallographica Section D, Biological crystallography* 66.Pt 4, pp. 486–501.

Errington, J. (2003). 'Regulation of endospore formation in *Bacillus subtilis*.' In: *Nature Reviews Microbiology* 1.2, pp. 117–126.

Evans, G., Axford, D. and Owen, R. L. (2011). 'The design of macromolecular crystallography diffraction experiments.' In: *Acta crystallographica Section D, Biological crystallography* 67.Pt 4, pp. 261–270.

Evans, G. and Pettifer, R. F. (2001). 'CHOOCH: a program for deriving anomalous-scattering factors from X-ray fluorescence spectra'. In: *Journal of applied crystallography* 34.1, pp. 82–86.

Evans, P. (2006). 'Scaling and assessment of data quality.' In: *Acta crystallographica Section D, Biological crystallography* 62.Pt 1, pp. 72–82.

Evans, P. R. (2011). 'An introduction to data reduction: space-group determination, scaling and intensity statistics.' In: *Acta crystallographica Section D, Biological crystallography* 67.Pt 4, pp. 282–292.

Evans, P. R. and Murshudov, G. N. (2013). 'How good are my data and what is the resolution?' In: *Acta crystallographica Section D, Biological crystallography* 69.Pt 7, pp. 1204–1214.

Fagan, R. P. and Fairweather, N. F. (2014). 'Biogenesis and functions of bacterial S-layers'. In: *Nature Reviews Microbiology* 12.3, pp. 211–222.

Fawcett, P., Eichenberger, P., Losick, R. and Youngman, P. (2000). 'The transcriptional profile of early to middle sporulation in *Bacillus subtilis*.' In: *Proceedings of the National Academy of Sciences of the United States of America* 97.14, pp. 8063–8068.

Felli, I. C. and Pierattelli, R. (2012). 'Recent progress in NMR spectroscopy: Toward the study of intrinsically disordered proteins of increasing size and complexity'. In: *IUBMB Life* 64.6, pp. 473–481.

Fellmeth, G., Yarlagadda, S. and Iyer, S. (2010). 'Epidemiology of community-onset *Clostridium difficile* infection in a community in the South of England.' In: *Journal of infection and public health* 3.3, pp. 118–123.

Fimlaid, K. A., Bond, J. P., Schutz, K. C., Putnam, E. E., Leung, J. M., Lawley, T. D. and Shen, A. (2013). 'Global analysis of the sporulation pathway of *Clostridium difficile*.' In: *PLoS genetics* 9.8, e1003660.

Fimlaid, K. A., Jensen, O., Donnelly, M. L., Siegrist, M. S. and Shen, A. (2015). 'Regulation of *Clostridium difficile* Spore Formation by the SpoIIQ and SpoIIIA Proteins'. In: *PLoS genetics* 11.10, e1005562–35.

Fimlaid, K. A. and Shen, A. (2015). 'Diverse mechanisms regulate sporulation sigma factor activity in the Firmicutes.' In: *Current Opinion in Microbiology* 24, pp. 88–95.

Firczuk, M. and Bochtler, M. (2007). 'Folds and activities of peptidoglycan amidases.' In: *FEMS Microbiology Reviews* 31.6, pp. 676–691.

Firczuk, M., Mucha, A. and Bochtler, M. (2005). 'Crystal structures of active LytM.' In: *Journal of Molecular Biology* 354.3, pp. 578–590.

Flanagan, K. A., Comber, J. D., Mearls, E., Fenton, C., Wang Erickson, A. F. and Camp, A. H. (2016). 'A membrane-embedded amino acid couples the SpoIIQ channel protein to anti-sigma factor transcriptional repression during *Bacillus subtilis* sporulation.' In: *Journal of bacteriology* 198.9, pp. 1451–63.

Forest, K. T., Dunham, S. A., Koomey, M. and Tainer, J. A. (1999). 'Crystallographic structure reveals phosphorylated pilin from *Neisseria*: phosphoserine sites modify type IV pilus surface chemistry and fibre morphology.' In: *Molecular microbiology* 31.3, pp. 743–752.

Forest, K. T. (2008). 'The type II secretion arrowhead: the structure of GspI-GspJ-GspK.' In: *Nature structural & molecular biology* 15.5, pp. 428–430.

Forest, K. T. (2013). 'Enterotoxigenic *Escherichia coli* CS1 pilus: not one structure but several.' In: *Journal of bacteriology* 195.7, pp. 1357–1359.

Fredlund, J., Broder, D., Fleming, T., Claussin, C. and Pogliano, K. (2013). 'The SpoIIQ landmark protein has different requirements for septal localization and immobilization'. In: *Molecular microbiology* 89.6, pp. 1053–68.

Frye, S. A., Assalkhou, R., Collins, R. F., Ford, R. C., Petersson, C., Derrick, J. P. and Tønjum, T. (2006). 'Topology of the outer-membrane secretin PilQ from *Neisseria meningitidis*.' In: *Microbiology* 152.Pt 12, pp. 3751–3764.

Galperin, M. Y., Mekhedov, S. L., Puigbo, P., Smirnov, S., Wolf, Y. I. and Rigden, D. J. (2012). 'Genomic determinants of sporulation in *Bacilli* and *Clostridia*: towards the minimal set of sporulation-specific genes'. In: *Environmental Microbiology* 14.11, pp. 2870–2890.

Garman, E. F. and Schneider, T. R. (1997). 'Macromolecular cryocrystallography'. In: *Journal of applied crystallography* 30.3, pp. 211–237.

Garman, E. and Murray, J. W. (2003). 'Heavy-atom derivatization'. In: *Acta crystallographica Section D, Biological crystallography* 59.11, pp. 1903–1913.

Gerding, D. N., Muto, C. A. and Owens, R. C. (2008). 'Measures to control and prevent *Clostridium difficile* infection.' In: *Clinical infectious diseases* 46 Suppl 1, S43–9.

Gildea, R. J., Waterman, D. G., Parkhurst, J. M., Axford, D., Sutton, G., Stuart, D. I., Sauter, N. K., Evans, G. and Winter, G. (2014). 'New methods for indexing multi-lattice diffraction data.' In: *Acta crystallographica Section D, Biological crystallography* 70.Pt 10, pp. 2652–2666.

Giltner, C. L., Nguyen, Y. and Burrows, L. L. (2012). 'Type IV pilin proteins: versatile molecular modules.' In: *Microbiology and molecular biology reviews* 76.4, pp. 740–772.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J. and Lopez, R. (2010). 'A new bioinformatics analysis tools framework at EMBL-EBI.' In: *Nucleic acids research* 38.Web Server issue, W695–9.

Goulding, D., Thompson, H., Emerson, J., Fairweather, N. F., Dougan, G. and Douce, G. R. (2009). 'Distinctive profiles of infection and pathology in hamsters infected with *Clostridium difficile* strains 630 and B1.' In: *Infection and Immunity* 77.12, pp. 5478–5485.

Grabowska, M., Jagielska, E., Czapinska, H., Bochtler, M. and Sabala, I. (2015). 'High resolution structure of an M23 peptidase with a substrate analogue.' In: *Scientific Reports* 5, p. 14833.

Haraldsen, J. D. and Sonenshein, A. L. (2003). 'Efficient sporulation in *Clostridium difficile* requires disruption of the sigmaK gene.' In: *Molecular microbiology* 48.3, pp. 811–821.

Harding, M. M. (2006). 'Small revisions to predicted distances around metal sites in proteins'. In: *Acta crystallographica Section D, Biological crystallography* 62.6, pp. 678–682.

Hartung, S., Arvai, A. S., Wood, T., Kolappan, S., Shin, D. S., Craig, L. and Tainer, J. A. (2011). 'Ultrahigh resolution and full-length pilin structures with insights for filament assembly, pathogenic functions, and vaccine potential.' In: *Journal of Biological Chemistry* 286.51, pp. 44254–44265.

Heger, A. and Holm, L. (2000). 'Rapid automatic detection and alignment of repeats in protein sequences.' In: *Proteins* 41.2, pp. 224–237.

Helaine, S., Carbonnelle, E., Prouvensier, L., Beretti, J.-L., Nassif, X. and Pelicic, V. (2005). 'PilX, a pilus-associated protein essential for bacterial aggregation, is a key to pilus-facilitated attachment of Neisseria meningitidis to human cells.' In: *Molecular microbiology* 55.1, pp. 65–77.

Helaine, S., Dyer, D. H., Nassif, X., Pelicic, V. and Forest, K. T. (2007). '3D structure/function analysis of PilX reveals how minor pilins can modulate the virulence properties of type IV pili.' In: *Proceedings of the National Academy of Sciences of the United States of America* 104.40, pp. 15888–15893.

Higgins, D. and Dworkin, J. (2011). 'Recent progress in *Bacillus subtilis* sporulation'. In: *FEMS Microbiology Reviews* 36.1, pp. 131–148.

Hilbert, D. W. and Piggot, P. J. (2004). 'Compartmentalization of gene expression during *Bacillus subtilis* spore formation.' In: *Microbiology and molecular biology reviews* 68.2, pp. 234–262.

Holm, L. and Rosenström, P. (2010). 'Dali server: conservation mapping in 3D.' In: *Nucleic acids research* 38.Web Server issue, W545–9.

Hoon, M. J. L. de, Eichenberger, P. and Vitkup, D. (2010). 'Hierarchical evolution of the bacterial sporulation network.' In: *Current biology : CB* 20.17, R735–45.

Huang, H., Weintraub, A., Fang, H. and Nord, C. E. (2009). 'Antimicrobial resistance in *Clostridium difficile*.' In: *International journal of antimicrobial agents* 34.6, pp. 516–522.

Iber, D., Clarkson, J., Yudkin, M. D. and Campbell, I. D. (2006). 'The mechanism of cell differentiation in *Bacillus subtilis*.' In: *Nature* 441.7091, pp. 371–374.

Illing, N. and Errington, J. (1991). 'The spoIIIA operon of *Bacillus subtilis* defines a new temporal class of mother-cell-specific sporulation genes under the control of the sigma E form of RNA polymerase.' In: *Molecular microbiology* 5.8, pp. 1927–1940.

Imam, S., Chen, Z., Roos, D. S. and Pohlschröder, M. (2011). 'Identification of surprisingly diverse Type IV Pili, across a broad range of Gram-Positive bacteria'. In: *PLoS ONE* 6.12, e28919.

Irving, H. and Williams, R. J. P. (1953). 'The stability of transition-metal complexes'. In: *Journal of the Chemical Society* 0, pp. 3192–19.

Jerabek-Willemsen, M., Wienken, C. J., Braun, D., Baaske, P. and Duhr, S. (2011). 'Molecular interaction studies using microscale thermophoresis.' In: *Assay and drug development technologies* 9.4, pp. 342–353.

Jin, F., Conrad, J. C., Gibiansky, M. L. and Wong, G. C. L. (2011). 'Bacteria use type-IV pili to slingshot on surfaces.' In: *Proceedings of the National Academy of Sciences* 108.31, pp. 12617–12622.

Johnson, T. L., Fong, J. C., Rule, C., Rogers, A., Yildiz, F. H. and Sandkvist, M. (2014). 'The Type II secretion system delivers matrix proteins for biofilm formation by *Vibrio cholerae*.' In: *Journal of bacteriology* 196.24, pp. 4245–4252.

Jones, D. T. (1999). 'Protein secondary structure prediction based on position-specific scoring matrices.' In: *Journal of Molecular Biology* 292.2, pp. 195–202.

Kabsch, W. and Sander, C. (1983). 'Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features'. In: *Biopolymers* 22.12, pp. 2577–637.

Kabsch, W. (2010). 'XDS.' In: *Acta crystallographica Section D, Biological crystallography* 66.Pt 2, pp. 125–132.

Kachrimanidou, M. and Malisiovas, N. (2011). '*Clostridium difficile* infection: A Comprehensive Review'. In: *Critical Reviews in Microbiology* 37.3, pp. 178–187.

Kang, H. J., Coulibaly, F., Clow, F., Proft, T. and Baker, E. N. (2007). 'Stabilizing isopeptide bonds revealed in gram-positive bacterial pilus structure.' In: *Science* 318.5856, pp. 1625–1628.

Kantardjieff, K. A. and Rupp, B. (2003). 'Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals.' In: *Protein science* 12.9, pp. 1865–1871.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. and Drummond, A. (2012). 'Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data.' In: *Bioinformatics* 28.12, pp. 1647–1649.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. and Sternberg, M. J. E. (2015). 'The Phyre2 web portal for protein modeling, prediction and analysis.' In: *Nature Protocols* 10.6, pp. 845–858.

Kelly, S. M., Jess, T. J. and Price, N. C. (2005). 'How to study proteins by circular dichroism'. In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1751.2, pp. 119–139.

Khanna, S. and Gupta, A. (2014). 'Community-acquired *Clostridium difficile* infection: an increasing public health threat'. In: *Infection and Drug Resistance* Volume 7, pp. 63–10.

Khanna, S., Pardi, D. S., Kelly, C. R., Kraft, C. S., Dhere, T., Henn, M. R., Lombardo, M.-J., Vulic, M., Ohsumi, T., Winkler, J., Pindar, C., McGovern, B. H., Pomerantz, R. J., Aunins, J. G., Cook, D. N. and Hohmann, E. L. (2016). 'A novel microbiome therapeutic increases gut microbial diversity and prevents recurrent *Clostridium difficile* infection.' In: *The Journal of infectious diseases* 214.2, pp. 173–181.

Köhler, R., Schäfer, K., Müller, S., Vignon, G., Diederichs, K., Philippsen, A., Ringler, P., Pugsley, A. P., Engel, A. and Welte, W. (2004). 'Structure and assembly of the pseudopilin PulG.' In: *Molecular microbiology* 54.3, pp. 647–664.

Kohlstaedt, M., Hocht, I. von der, Hilbers, F., Thielmann, Y. and Michel, H. (2015). 'Development of a thermofluor assay for stability determination of membrane proteins using the Na+/H+ antiporter NhaA and cytochrome c oxidase'. In: *Acta crystallographica Section D, Biological crystallography* 71, pp. 1112–1122.

Korotkov, K. V., Gray, M. D., Kreger, A., Turley, S., Sandkvist, M. and Hol, W. G. J. (2009). 'Calcium is essential for the major pseudopilin in the type 2 secretion system.' In: *Journal of Biological Chemistry* 284.38, pp. 25466–25470.

Korotkov, K. V. and Hol, W. G. J. (2008). 'Structure of the GspK-GspI-GspJ complex from the enterotoxigenic *Escherichia coli* type 2 secretion system.' In: *Nature structural & molecular biology* 15.5, pp. 462–468.

Korotkov, K. V., Sandkvist, M. and Hol, W. G. J. (2012). 'The type II secretion system: biogenesis, molecular architecture and mechanism.' In: *Nature Reviews Microbiology* 10.5, pp. 336–351.

Krissinel, E. and Henrick, K. (2004). 'Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions.' In: *Acta crystallographica Section D, Biological crystallography* 60.Pt 12 Pt 1, pp. 2256–2268.

Krissinel, E. and Henrick, K. (2007). 'Inference of macromolecular assemblies from crystalline state.' In: *Journal of Molecular Biology* 372.3, pp. 774–797.

Kurka, H., Ehrenreich, A., Ludwig, W., Monot, M., Rupnik, M., Barbut, F., Indra, A., Dupuy, B. and Liebl, W. (2014). 'Sequence similarity of *Clostridium difficile* strains by analysis of conserved genes and genome content is reflected by their ribotype affiliation.' In: *PLoS ONE* 9.1, e86535.

Levdikov, V. M., Blagova, E. V., McFeat, A., Fogg, M. J., Wilson, K. S. and Wilkinson, A. J. (2012). 'Structure of components of an intercellular channel complex in sporulating *Bacillus subtilis*.' In: *Proceedings of the National Academy of Sciences* 109.14, pp. 5441–5445.

Lewis, R. J., Muchová, K., Brannigan, J. A., Barák, I., Leonard, G. and Wilkinson, A. J. (2000). 'Domain swapping in the sporulation response regulator Spo0A.' In: *Journal of Molecular Biology* 297.3, pp. 757–770.

Lim, M. S., Ng, D., Zong, Z., Arvai, A. S., Taylor, R. K., Tainer, J. A. and Craig, L. (2010). '*Vibrio cholerae* El Tor TcpA crystal structure and mechanism for pilus-mediated microcolony formation.' In: *Molecular microbiology* 77.3, pp. 755–770.

Lobley, A., Whitmore, L. and Wallace, B. A. (2002). 'DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra.' In: *Bioinformatics* 18.1, pp. 211–212.

Londoño-Vallejo, J. A., Fréhel, C. and Stragier, P. (1997). 'SpoIIQ, a forespore-expressed gene required for engulfment in *Bacillus subtilis*.' In: *Molecular microbiology* 24.1, pp. 29–39.

Maier, B. and Wong, G. C. L. (2015). 'How Bacteria Use Type IV Pili Machinery on Surfaces'. In: *Trends in Microbiology*, pp. 1–14.

Maldarelli, G. A., De Masi, L., Rosenvinge, E. C. von, Carter, M. and Donnenberg, M. S. (2014). 'Identification, immunogenicity, and cross-reactivity of type IV pilin and pilin-like proteins from *Clostridium difficile*.' In: *Pathogens and disease* 71.3, pp. 302–314.

Maldarelli, G. A., Piepenbrink, K. H., Scott, A. J., Freiberg, J. A., Song, Y., Achermann, Y., Ernst, R. K., Shirtliff, M. E., Sundberg, E. J., Donnenberg, M. S. and Rosenvinge, E. C. von (2016). 'Type IV pili promote early biofilm formation by *Clostridium difficile*.' In: *Pathogens and disease* 74.6, ftw061.

Manetti, A. G. O., Zingaretti, C., Falugi, F., Capo, S., Bombaci, M., Bagnoli, F., Gambellini, G., Bensi, G., Mora, M., Edwards, A. M., Musser, J. M., Graviss, E. A., Telford, J. L., Grandi, G. and Margarit, I. (2007). '*Streptococcus pyogenes* pili promote pharyngeal cell adhesion and biofilm formation.' In: *Molecular microbiology* 64.4, pp. 968–983.

Martin, P. R., Hobbs, M., Free, P. D., Jeske, Y. and Mattick, J. S. (1993). 'Characterization of pilQ, a new gene required for the biogenesis of type 4 fimbriae in *Pseudomonas aeruginosa*.' In: *Molecular microbiology* 9.4, pp. 857–868.

Matthews, B. W. (1968). 'Solvent content of protein crystals.' In: *Journal of Molecular Biology* 33.2, pp. 491–497.

Mattick, J. S. (2002). 'Type IV pili and twitching motility'. In: *Annual Review of Microbiology* 56.1, pp. 289–314.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. and Read, R. J. (2007). 'Phaser crystallographic software.' In: *Journal of applied crystallography* 40.Pt 4, pp. 658–674.

McKenney, P. T. and Eichenberger, P. (2012). 'Dynamics of spore coat morphogenesis in *Bacillus subtilis*.' In: *Molecular microbiology* 83.2, pp. 245–260.

McNicholas, S., Potterton, E., Wilson, K. S. and Noble, M. E. M. (2011). 'Presenting your structures: the CCP4mg molecular-graphics software.' In: *Acta crystallographica Section D, Biological crystallography* 67.Pt 4, pp. 386–394.

McPherson, A. and Cudney, B. (2014). 'Optimization of crystallization conditions for biological macromolecules'. In: *Acta Crystallographica Section F* 70.11, pp. 1445–1467.

Meisner, J., Maehigashi, T., André, I., Dunham, C. M. and Moran, C. P. (2012). 'Structure of the basal components of a bacterial transporter.' In: *Proceedings of the National Academy of Sciences* 109.14, pp. 5446–5451.

Meisner, J. and Moran, C. P. (2011). 'A LytM domain dictates the localization of proteins to the mother cell-forespore interface during bacterial endospore formation'. In: *Journal of bacteriology* 193.3, pp. 591–598.

Meisner, J., Wang, X., Serrano, M., Henriques, A. O. and Moran, C. P. (2008). 'A channel connecting the mother cell and forespore during bacterial endospore formation.' In: *Proceedings of the National Academy of Sciences* 105.39, pp. 15100–15105.

Melville, S. and Craig, L. (2013). 'Type IV pili in Gram-positive bacteria.' In: *Microbiology and molecular biology reviews* 77.3, pp. 323–341.

Merrigan, M., Venugopal, A., Mallozzi, M., Roxas, B., Viswanathan, V. K., Johnson, S., Gerding, D. N. and Vedantam, G. (2010). 'Human hypervirulent *Clostridium difficile* strains exhibit increased sporulation as well as robust toxin production'. In: *Journal of bacteriology* 192.19, pp. 4904–4911.

Merritt, E. A. (2012). 'To B or not to B: a question of resolution?' In: *Acta crystallographica Section D, Biological crystallography* 68.4, pp. 468–477.

Millan, C., Sammito, M. and Uson, I. (2015). 'Macromolecular ab initio phasing enforcing secondary and tertiary structure'. In: *IUCrJ* M2, pp. 95–105.

Miller, F., Phan, G., Brissac, T., Bouchiat, C., Lioux, G., Nassif, X. and Coureuil, M. (2014). 'The hypervariable region of meningococcal major pilin PilE controls the host cell response via antigenic variation.' In: *mBio* 5.1, e01024–13.

Möller, S., Croning, M. D. and Apweiler, R. (2001). 'Evaluation of methods for the prediction of membrane spanning regions.' In: *Bioinformatics* 17.7, pp. 646–653.

Monot, M., Boursaux-Eude, C., Thibonnier, M., Vallenet, D., Moszer, I., Medigue, C., Martin-Verstraete, I. and Dupuy, B. (2011). 'Reannotation of the genome sequence of *Clostridium difficile* strain 630.' In: *Journal of medical microbiology* 60.Pt 8, pp. 1193–1199.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. and Vagin, A. A. (2011). 'REFMAC5 for the refinement of macro-

molecular crystal structures.' In: *Acta crystallographica Section D, Biological crystallography* 67.Pt 4, pp. 355–367.

Myers, G. S. A., Rasko, D. A., Cheung, J. K., Ravel, J., Seshadri, R., DeBoy, R. T., Ren, Q., Varga, J., Awad, M. M., Brinkac, L. M., Daugherty, S. C., Haft, D. H., Dodson, R. J., Madupu, R., Nelson, W. C., Rosovitz, M. J., Sullivan, S. A., Khouri, H., Dimitrov, G. I., Watkins, K. L., Mulligan, S., Benton, J., Radune, D., Fisher, D. J., Atkins, H. S., Hiscox, T., Jost, B. H., Billington, S. J., Songer, J. G., McClane, B. A., Titball, R. W., Rood, J. I., Melville, S. B. and Paulsen, I. T. (2006). 'Skewed genomic variability in strains of the toxigenic bacterial pathogen, *Clostridium perfringens*.' In: *Genome research* 16.8, pp. 1031–1040.

Nyborg, J. K. and Peersen, O. B. (2004). 'That zincing feeling: the effects of EDTA on the behaviour of zinc-binding transcriptional regulators.' In: *The Biochemical journal* 381.Pt 3, e3–4.

Ojkic, N., López-Garrido, J., Pogliano, K. and Endres, R. G. (2014). 'Bistable forespore engulfment in *Bacillus subtilis* by a zipper mechanism in absence of the cell wall'. In: *PLoS Computational Biology* 10.10, e1003912–13.

Ottow, J. C. (1975). 'Ecology, physiology, and genetics of fimbriae and pili.' In: *Annual Review of Microbiology* 29.1, pp. 79–108.

Owen, R. L. and Sherrell, D. A. (2016). 'Radiation damage and derivatization in macromolecular crystallography: a structure factor's perspective.' In: *Acta crystallographica. Section D, Structural biology* 72.Pt 3, pp. 388–394.

Pansegrau, W. and Bagnoli, F. (2016). 'Pilus Assembly in Gram-Positive Bacteria.' In: *Current topics in microbiology and immunology*, pp. 1–31.

Pape, T. and Schneider, T. R. (2004). 'HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs'. In: *J. Appl. Cryst* 37, pp. 843–844.

Paredes-Sabja, D., Shen, A. and Sorg, J. A. (2014). '*Clostridium difficile* spore biology: sporulation, germination, and spore structural proteins.' In: *Trends in Microbiology* 22.7, pp. 406–416.

Peltier, J., Courtin, P., El Meouche, I., Lemée, L., Chapot-Chartier, M.-P. and Pons, J.-L. (2011). '*Clostridium difficile* has an original peptidoglycan structure with a high level of

N-acetylglucosamine deacetylation and mainly 3-3 cross-links.' In: *Journal of Biological Chemistry* 286.33, pp. 29053–29062.

Pereira, F. C., Saujet, L., Tomé, A. R., Serrano, M., Monot, M., Couture-Tosi, E., Martin-Verstraete, I., Dupuy, B. and Henriques, A. O. (2013). 'The spore differentiation pathway in the enteric pathogen *Clostridium difficile*.' In: *PLoS genetics* 9.10, e1003782.

Piepenbrink, K. H., Maldarelli, G. A., Peña, C. F. M. de la, Dingle, T. C., Mulvey, G. L., Lee, A., Rosenvinge, E. von, Armstrong, G. D., Donnenberg, M. S. and Sundberg, E. J. (2015). 'Structural and evolutionary analyses show unique stabilization strategies in the Type IV pili of *Clostridium difficile*.' In: *Structure/Folding and Design* 23.2, pp. 1–13.

Piepenbrink, K. H., Maldarelli, G. A., Peña, C. F. M. de la, Mulvey, G. L., Snyder, G. A., De Masi, L., Rosenvinge, E. C. von, Günther, S., Armstrong, G. D., Donnenberg, M. S. and Sundberg, E. J. (2014). 'Structure of *Clostridium difficile* PilJ exhibits unprecedented divergence from known type IV pilins.' In: *Journal of Biological Chemistry* 289.7, pp. 4334–4345.

Pirie, N. W. (1949). 'Structure and activities of the bacterial surface'. In: *Nature* 163.4154, pp. 897–898.

Postis, V., Rawson, S., Mitchell, J. K., Lee, S. C., Parslow, R. A., Dafforn, T. R., Baldwin, S. A. and Muench, S. P. (2015). 'The use of SMALPs as a novel membrane protein scaffold for structure study by negative stain electron microscopy.' In: *Biochimica et biophysica acta* 1848.2, pp. 496–501.

Proft, T. and Baker, E. N. (2009). 'Pili in Gram-negative and Gram-positive bacteria — structure, assembly and their role in disease'. In: *Cellular and Molecular Life Sciences* 66.4, pp. 613–635.

Purcell, E. B., McKee, R. W., Bordeleau, E., Burrus, V. and Tamayo, R. (2015). 'Regulation of Type IV Pili Contributes to Surface Behaviors of Historical and Epidemic Strains of *Clostridium difficile*.' In: *Journal of bacteriology*, JB.00816–15.

Rakotoarivonina, H., Jubelin, G., Hebraud, M., Gaillard-Martinie, B., Forano, E. and Mosoni, P. (2002). 'Adhesion to cellulose of the Gram-positive bacterium *Ruminococcus albus* involves type IV pili.' In: *Microbiology* 148.Pt 6, pp. 1871–1880.

Reardon, P. N. and Mueller, K. T. (2013). 'Structure of the type IVa major pilin from the electrically conductive bacterial nanowires of *Geobacter sulfurreducens*.' In: *Journal of Biological Chemistry* 288.41, pp. 29260–29266.

Reddy-Chichili, V. P., Kumar, V. and Sivaraman, J. (2013). 'Linkers in the structural biology of protein-protein interactions'. In: *Protein Science* 22.2, pp. 153–167.

Rodrigues, C. D. A., Marquis, K. A., Meisner, J. and Rudner, D. Z. (2013). 'Peptidoglycan hydrolysis is required for assembly and activity of the transenvelope secretion complex during sporulation in Bacillus subtilis'. In: *Molecular microbiology* 89.6, pp. 1039–52.

Rodríguez, D. D., Grosse, C., Himmel, S., González, C., Ilarduya, I. M. de, Becker, S., Sheldrick, G. M. and Usón, I. (2009). 'Crystallographic ab initio protein structure solution below atomic resolution'. In: *Nature Methods* 6.9, pp. 651–653.

Rossmann, M. G. and Arnold, E. (2001). *International tables for crystallography, volume F*. Crystallography of biological macromolecules. Wiley.

Rotem, O., Nesper, J., Borovok, I., Gorovits, R., Kolot, M., Pasternak, Z., Shin, I., Glatter, T., Pietrokovski, S., Jenal, U. and Jurkevitch, E. (2015). 'An extended cyclic di-GMP network in the predatory bacterium *Bdellovibrio bacteriovorus*'. In: *Journal of bacteriology* 198.1, pp. 127–137.

Rubio, A. and Pogliano, K. (2004). 'Septal localization of forespore membrane proteins during engulfment in *Bacillus subtilis*.' In: *The EMBO Journal* 23.7, pp. 1636–1646.

Saujet, L., Pereira, F. C., Henriques, A. O. and Martin-Verstraete, I. (2014). 'The regulatory network controlling spore formation in *Clostridium difficile*.' In: *FEMS microbiology letters* 358.1, pp. 1–10.

Saujet, L., Pereira, F. C., Serrano, M., Soutourina, O., Monot, M., Shelyakin, P. V., Gelfand, M. S., Dupuy, B., Henriques, A. O. and Martin-Verstraete, I. (2013). 'Genome-wide analysis of cell type-specific gene transcription during spore formation in *Clostridium difficile*.' In: *PLoS genetics* 9.10, e1003756.

Schirmer, T. and Jenal, U. (2009). 'Structural and mechanistic determinants of c-di-GMP signalling.' In: *Nature Reviews Microbiology* 7.10, pp. 724–735.

Schneider, T. R. and Sheldrick, G. M. (2002). 'Substructure solution with SHELXD'. In: *Acta crystallographica Section D, Biological crystallography* 58, pp. 1772–1779.

Sebaihia, M., Wren, B. W., Mullany, P., Fairweather, N. F., Minton, N., Stabler, R., Thomson, N. R., Roberts, A. P., Cerdeño-Tárraga, A. M., Wang, H., Holden, M. T., Wright, A., Churcher, C., Quail, M. A., Baker, S., Bason, N., Brooks, K., Chillingworth, T., Cronin, A., Davis, P., Dowd, L., Fraser, A., Feltwell, T., Hance, Z., Holroyd, S., Jagels, K., Moule, S., Mungall, K., Price, C., Rabbinowitsch, E., Sharp, S., Simmonds, M., Stevens, K., Unwin, L., Whithead, S., Dupuy, B., Dougan, G., Barrell, B. and Parkhill, J. (2006). 'The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome'. In: *Nature Genetics* 38.7, pp. 779–786.

Serrano, M., Crawshaw, A. D., Dembek, M., Monteiro, J. M., Pereira, F. C., Pinho, M. G. de, Fairweather, N. F., Salgado, P. S. and Henriques, A. O. (2015). 'The SpoIIQ-SpoIIIAH complex of *Clostridium difficile* controls forespore engulfment and late stages of gene expression and spore morphogenesis.' In: *Molecular microbiology* 100.1, pp. 204–28.

Setlow, P. (2006). 'Spores of *Bacillus subtilis*: their resistance to and killing by radiation, heat and chemicals.' In: *Journal of applied microbiology* 101.3, pp. 514–525.

Sheldrick, G. M. (2007). 'A short history of SHELX'. In: *Acta Crystallographica Section A* 64, pp. 112–122.

Sheldrick, G. M. (2010). 'Experimental phasing with SHELXC/D/E: combining chain tracing with density modification.' In: *Acta crystallographica Section D, Biological crystallography* 66.Pt 4, pp. 479–485.

Shimizu, T., Ohtani, K., Hirakawa, H., Ohshima, K., Yamashita, A., Shiba, T., Ogasawara, N., Hattori, M., Kuhara, S. and Hayashi, H. (2002). 'Complete genome sequence of *Clostridium perfringens*, an anaerobic flesh-eater.' In: *Proceedings of the National Academy of Sciences of the United States of America* 99.2, pp. 996–1001.

Shmueli, U. (2008). *International tables for crystallography, reciprocal space*. Springer Science & Business Media.

Skotnicka, D., Petters, T., Heering, J., Hoppert, M., Kaever, V. and Søgaard-Andersen, L. (2015). 'c-di-GMP regulates type IV pili-dependent-motility in *Myxococcus xanthus*.' In: *Journal of bacteriology*, JB.00281–15.

Smits, W. K., Lyras, D., Lacy, D. B., Wilcox, M. H. and Kuijper, E. J. (2016). '*Clostridium difficile* infection.' In: *Nature reviews. Disease primers* 2, p. 16020.

Sonnhammer, E. L., Heijne, G. von and Krogh, A. (1998). 'A hidden Markov model for predicting transmembrane helices in protein sequences.' In: *Proceedings: International Conference on Intelligent Systems for Molecular Biology* 6, pp. 175–182.

Sorg, J. A. and Sonenshein, A. L. (2008). 'Bile salts and glycine as cogerminants for *Clostridium difficile* spores'. In: *Journal of bacteriology* 190.7, pp. 2505–2512.

Stabler, R. A., He, M., Dawson, L., Martin, M., Valiente, E., Corton, C., Lawley, T. D., Sebaihia, M., Quail, M. A., Rose, G., Gerding, D. N., Gibert, M., Popoff, M. R., Parkhill, J., Dougan, G. and Wren, B. W. (2009). 'Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium.' In: *Genome Biology* 10.9, R102.

Steil, L., Serrano, M., Henriques, A. O. and Völker, U. (2005). 'Genome-wide analysis of temporally regulated and compartment-specific gene expression in sporulating cells of *Bacillus subtilis*.' In: *Microbiology* 151.Pt 2, pp. 399–420.

Stephenson, K. and Lewis, R. J. (2005). 'Molecular insights into the initiation of sporulation in Gram-positive bacteria: new technologies for an old phenomenon.' In: *FEMS Microbiology Reviews* 29.2, pp. 281–301.

Sun, Y. L., Sharp, M. D. and Pogliano, K. (2000). 'A dispensable role for forespore-specific gene expression in engulfment of the forespore during sporulation of *Bacillus subtilis*.' In: *Journal of bacteriology* 182.10, pp. 2919–2927.

Szeto, T. H., Dessen, A. and Pelicic, V. (2011). 'Structure/function analysis of *Neisseria meningitidis* PilW, a conserved protein that plays multiple roles in type IV pilus biology.' In: *Infection and Immunity* 79.8, pp. 3028–3035.

Takahashi, H., Yanagisawa, T., Kim, K. S., Yokoyama, S. and Ohnishi, M. (2012). 'Meningococcal PilV potentiates *Neisseria meningitidis* type IV pilus-mediated internalization into human endothelial and epithelial cells.' In: *Infection and Immunity* 80.12, pp. 4154–4166.

Taylor, G. L. (2010). 'Introduction to phasing'. In: *Acta crystallographica Section D, Biological crystallography* 66, pp. 325–338.

Ton-That, H. and Schneewind, O. (2003). 'Assembly of pili on the surface of *Corynebacterium diphtheriae*.' In: *Molecular microbiology* 50.4, pp. 1429–1438.

Underwood, S., Guan, S., Vijayasubhash, V., Baines, S. D., Graham, L., Lewis, R. J., Wilcox, M. H. and Stephenson, K. (2009). 'Characterization of the Sporulation Initiation Pathway of *Clostridium difficile* and Its Role in Toxin Production'. In: *Journal of bacteriology* 191.23, pp. 7296–7305.

Urzhumtseva, L., Afonine, P. V., Adams, P. D. and Urzhumtsev, A. (2009). 'Crystallographic model quality at a glance.' In: *Acta crystallographica Section D, Biological crystallography* 65.Pt 3, pp. 297–300.

Vagin, A. and Teplyakov, A. (2010). 'Molecular replacement with MOLREP.' In: *Acta crystallographica Section D, Biological crystallography* 66.Pt 1, pp. 22–25.

Varga, J. J., Nguyen, V., O'Brien, D. K., Rodgers, K., Walker, R. A. and Melville, S. B. (2006). 'Type IV pili-dependent gliding motility in the Gram-positive pathogen *Clostridium perfringens* and other Clostridia.' In: *Molecular microbiology* 62.3, pp. 680–694.

Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L., Markley, J. L., Ionides, J. and Laue, E. D. (2005). 'The CCPN data model for NMR spectroscopy: development of a software pipeline.' In: *Proteins* 59.4, pp. 687–696.

Wang, S. T., Setlow, B., Conlon, E. M., Lyon, J. L., Imamura, D., Sato, T., Setlow, P., Losick, R. and Eichenberger, P. (2006). 'The forespore line of gene expression in *Bacillus subtilis*.' In: *Journal of Molecular Biology* 358.1, pp. 16–37.

Warren, A. J., Crawshaw, A. D., Trincao, J., Aller, P., Alcock, S., Nistea, I., Salgado, P. S. and Evans, G. (2015). 'In vacuo X-ray data collection from graphene-wrapped protein crystals'. In: *Acta crystallographica Section D, Biological crystallography* 71.10, pp. 2079–2088.

Whitchurch, C. B., Tolker-Nielsen, T., Ragas, P. C. and Mattick, J. S. (2002). 'Extracellular DNA required for bacterial biofilm formation.' In: *Science* 295.5559, pp. 1487–1487.

Whitmore, L. and Wallace, B. A. (2004). 'DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data.' In: *Nucleic acids research* 32.Web Server issue, W668–73.

Wienken, C. J., Baaske, P., Rothbauer, U., Braun, D. and Duhr, S. (2010). 'Protein-binding assays in biological liquids using microscale thermophoresis.' In: *Nature communications* 1.7, pp. 100–7.

Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C., Williams, K. L., Appel, R. D. and Hochstrasser, D. F. (1999). 'Protein identification and analysis tools in the ExPASy server.' In: *Methods in molecular biology* 112, pp. 531–552.

Winter, G., Lobley, C. M. C. and Prince, S. M. (2013). 'Decision making in xia2.' In: *Acta crystallographica Section D, Biological crystallography* 69.Pt 7, pp. 1260–1273.

Wu, H. and Fives-Taylor, P. M. (2001). 'Molecular strategies for fimbrial expression and assembly.' In: *Critical reviews in oral biology and medicine* 12.2, pp. 101–115.

Wüst, J., Sullivan, N. M., Hardegger, U. and Wilkins, T. D. (1982). 'Investigation of an outbreak of antibiotic-associated colitis by various typing methods.' In: *Journal of Clinical Microbiology* 16.6, pp. 1096–1101.

Xue, B., Dunbrack, R. L., Williams, R. W., Dunker, A. K. and Uversky, V. N. (2010). 'PONDR-FIT: a meta-predictor of intrinsically disordered amino acids.' In: *Biochimica et biophysica acta* 1804.4, pp. 996–1010.

Yutin, N. and Galperin, M. Y. (2013). 'A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia.' In: *Environmental Microbiology* 15.10, pp. 2631–2641.

Zhang, Z., Sauter, N. K., Bedem, H. van den, Snell, G. and Deacon, A. M. (2006). 'Automated diffraction image analysis and spot searching for high-throughput crystal screening'. In: *Journal of applied crystallography* 39.1, pp. 112–119.