UNIVERSITY OF CAMBRIDGE

A simulation approach to the study of bacterial secretion proteins

Alexandra East

Wolfson College, University of Cambridge

September 2016

This dissertation is submitted for the degree of Doctor of Philosophy.

~ For my family ~

Human beings: little bags of thinking water held up
briefly by fragile accumulations of calcium.
– Pyramids, Terry Pratchett

Disclaimer

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. No substantial part of this work has been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. In accordance with the regulations of the Physics and Chemistry Degree Committee, this thesis does not exceed the 60,000 word limit.

Alexandra East

Abstract

Knowledge of the structure and dynamics of cellular protein complexes is essential for understanding their functionally relevant interactions. In Gram-negative bacteria, the complex machinery associated with the type II secretion system (T2SS) polymerises inner membrane pseudopilin proteins into thin filaments, to export substrates such as toxins, hydrolases and cytochromes. Here, computational simulations were used to study proteins from the *Klebsiella oxytoca* T2SS, focusing on the substrate pullulanase PulA, the major pseudopilin PulG, and the putative chaperone PulM.

Chapter 3 contains an *in silico* study of both post-translationally acylated PulA (lipoPulA) and non-acylated PulA ($PulA_{NA}$) in association with a lipid bilayer, representing an approximation of the biological state prior to secretion; this study examined PulA dynamics and the possible role of the acyl tail in protein-membrane interactions before secretion. Novel insights into the interactions of a key residue necessary for Type 2 secretion were gained *via* simulations performed on a $PulA_{NA}$ D2S variant, extending prior *in vitro* results. In Chapter 4, PulA was simulated in conditions closer to the physiological environment, using counter-ions to investigate the possible effect of the high periplasmic calcium concentration on protein conformation and lipid interactions prior to secretion. In Chapter 5, variants of the major pseudopilin PulG containing one transmembrane helix were simulated, demonstrating N-terminal interactions made possible by wild-type methylation of residue Phe1. Simulations of several monomeric PulG variants provided insight into the roles of the essential residue Glu5 and Phe1 methylation, previously identified by experimental work to be important. Simulations of the PulG dimer demonstrated the dynamic nature of the membrane-embedded dimer interface, and showed how computational analysis can predict *in vivo* contacts. Finally, Chapter 6 extended the T2SS studies to coarse-grained methods, sampling possible conformations and predicting the PulG-PulM interface within the membrane, prior to PulG presentation to the remaining secretion apparatus.

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Peter Bond for his support and academic expertise throughout this Ph.D., especially considering the physical distance between us after his move to Singapore. Many thanks to Dr. Mark Williamson for his advice, proof-reading, and excellent teaching on the methods I have used throughout this thesis. Thanks also to Prof. Bobby Glen for taking care of the group in Pete's absence, and to Prof. Syma Khalid from Southampton University for allowing me to visit her group and learn coarse-grain methods from them; particular thanks go to Jenny and Firdaus for their assistance with Chapter 6. My unending gratitude goes to my collaborator Dr. Olivera Francetic of the Pasteur Institut in Paris for both her expertise and emotional support throughout this research and the resulting publications. I could not have asked for a better collaborator and I am sorry that our partnership is drawing to a close. I also thank Javier for his help and advice.

I would like to thank the other members of the Bond group for their friendship and assistance as we have muddled through together: Tere for her unfailing cheerfulness, sense of humour and technical expertise; Maite for the times we shared in the lab, her help with the software, and her listening ear when things got tough either inside or outside the lab; Florian for his positive attitude and his critical eye that helped me to honestly examine both my work and myself; Namita for always sharing a joke and helping me through the writing phase *via* Skype! I cannot fail to thank the other members of the UCC (now the CMI) for their friendship, particularly Nele, Siti, Ain, Oscar, Avid, Shardul, Aakash, and Fatima (for all our unexpectedly deep conversations!). My gratitude and affection goes to Matt and Tim, for their good humour that helped me see the funny side of everything, offering excellent advice when I needed it, and accepting me as one of the lads- their support has been invaluable. Heartfelt thanks go to Katie for her ongoing support, friendship and PMA; I'm delighted to leave Cambridge with a new friend for life. Many thanks also to Toby for his encouragement and advice throughout this degree. I would also like to thank Maitreyi for sharing my trials and tribulations over the last several years; I'm excited to live in the same city as her yet again for the next stage of our lives! Love

and thanks go to my sister and One, Meera, who has been a listening ear and my partner in crime through every stage of our lives over the last 15 years.

I initially only researched opportunities at Cambridge in order to be closer to Gonzalo, and I cannot adequately express my happiness in our relationship and subsequent marriage. He has been the best friend and husband I could have asked for: sharing in my small victories, believing in me when I didn't believe in myself any more, and providing a shoulder to cry on but also giving me tough love when I needed it. I am excited to start the next part of our life together, hopefully as Dr. and Dr. Garcia! Most importantly, I thank my parents and brother for their help and unfailing love throughout my academic career and especially my Ph.D. They have done everything they can to make my life and degrees easier, and have always supported me, encouraged me, and believed in me. If I can demonstrate even half of their love, professionalism and kindness in my life going forward, I will be very fortunate indeed. Without them, I would not be the woman I am today, and I am eternally grateful.

Publications

Efficient Characterization of Protein Cavities within Molecular Simulation Trajectories: *trj_cavity*.

Teresa Paramo, Alexandra East, Diana Garzón, Martin B. Ulmschneider, Peter J. Bond

J. Chem. Theory Comput., 2014, 10 (5), pp 2151–2164

Structural Basis of Pullulanase Membrane Binding and Secretion Revealed by X-Ray Crystallography, Molecular Dynamics and Biochemical Analysis.

Alexandra East, Ariel E. Mechaly, Gerard H.M. Huysmans, Cédric Bernarde, Diana Tello-Manigne, Nathalie Nadeau, Anthony P. Pugsley, Alejandro Buschiazzo, Pedro M. Alzari, Peter J. Bond, Olivera Francetic

Structure, 2016, 24 (1), pp 92 - 104

Polar N-terminal residues conserved in type 2 secretion pseudopilins determine subunit targeting and membrane extraction steps during fibre assembly.

Javier Santos-Moreno, Alexandra East, Ingrid Guilvout, Nathalie Nadeau, Peter J. Bond, Guy Tran Van Nhieu and Olivera Francetic

Accepted by *Journal of Molecular Biology*, April 2017

Contents

List of Abbreviations

ABC	ATP-binding cassette
AP	Assembly platform
CDs	Cyclodextrins
CG	Coarse-grained
CT	Cholera Toxin
E. coli	*Escherichia coli*
EHEC	Enterohaemorrhagic *Escherichia coli*
ETEC	Enterotoxigenic *Escherichia coli*
IM	Inner membrane
K. oxytoca	*Klebsiella oxytoca*
K. pneumoniae	*Klebsiella pneumoniae*
LPS	Lipopolysaccharide
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
OM	Outer membrane
PDB	Protein Data Bank
PE	Phosphatidylethanolamine
PG	Phosphatidylglycerol
POPE	Palmitoyl-oleoyl phosphatidylethanolamine
PulA	Pullulanase
$PulA^{Kox}$	PulA from *Klebsiella oxytoca*
$PulA^{Kpn}$	PulA from *Klebsiella pneumoniae*
$PulA_{NA}$	Non-acylated pullulanase

PulG	Major pseudopilin from *Klebsiella oxytoca*
PulM	Inner membrane complex protein from *Klebsiella oxytoca*
T2SS	Type 2 Secretion System
T4P	Type 4 pilin
TM	Transmembrane
TMS	Transmembrane segment
T*n*SS	Type *n* Secretion System
WT	Wild-type

Chapter 1 – Introduction

1.1 Biological Background

Proteins can be considered the most fundamental molecules for life on Earth. Formed from combinations of a class of organic molecules called amino acids, they perform a colossal array of functions by interacting with each other and other moieties, including lipids, carbohydrates and ions. Proteins can perform structural, enzymatic, transport, recognition, and signalling roles, to name just a few. Specific ligand binding is a particularly useful capability, enabling specialised cellular functions including directed immune responses[1], signal transduction[2], and substrate transport[3]. This characteristic also facilitates drug and small molecule design, to potentially manipulate these functions.

All organisms on Earth require proteins, including eukaryotes, archaea and bacteria. Bacteria outnumber humans by a staggering amount; for every human on the planet, there are an estimated 10 million trillion microbes on the ocean floor alone[4]. Bacteria are ubiquitous single-celled prokaryotes, with at least one membrane enclosing the cell contents (cytoplasm) and acting as a barrier between the internal and external environments. Due to their relative ease (usually) of cultivation and relevance to human industry and health, bacteria have been the focus of copious protein studies. To date, the Protein Data Bank (PDB, www.rcsb.org) contains almost 44,000 bacterial protein structures[5]. Bacterial proteins include both structural molecules and those produced by bacteria, either for use by the cell (eg. for adhesion) or for secretion into the environment, including exocellular polymeric substances[6], quorum sensing autoinducers[7], and virulence factor proteins[8].

Bacteria have colonised every known habitat on Earth, and many populate mucosal surfaces by using hairlike appendages. These protrusions, mostly known interchangeably in the literature as pili or fimbriae, are referred to here solely as pili. Pili are oligomeric proteinaceous non-flagellar filaments projecting from bacterial cells. They are usually several micrometres in length, may number in the hundreds on a single cell, are usually encoded by plasmids and can play a role in bacterial

conjugation. Various systems, such as complexes involved in secretion, contain such filaments or analogous pseudopili. Fully assembled pilial systems are located in the outer membrane (OM) of Gram-negative diderm bacteria, which will be detailed subsequently, and their secretion mechanisms remain under investigation. However, extensive advances have been made in our understanding since Antony van Leeuwenhoek discovered bacterial cells in the 1660s[9].

Comprehensive knowledge of bacterial secretion would assist in preventing or treating pathogenic behaviours, and potentially improve several areas of human health and agriculture. Experimental studies continue to provide invaluable data, particularly analyses of protein variants both *in vitro* and *in vivo*, but since the development of the computer in the middle of the 20th century, *in silico* investigations and analysis have revolutionised scientific research. Modern day computational simulations act as a "molecular microscope" to replicate the behaviour of a chosen system on an atomic level and observe internal interactions. They provide dynamical, structural and thermodynamic data that is unavailable from static crystal structures and cannot be extrapolated from the results of bench experiments; *in silico* data can be used to supplement and interpret experimental data.

A truly atomic view of OM secretion systems is desirable, and the work presented here combines molecular dynamics (MD) simulation approaches with existing experimental results to provide details on the role of specific interactions mediated by proteins present in a bacterial secretion system that breaches the bacterial OM, the final barrier between the cell and the environment. With protein crystal structures being solved at an ever-accelerating rate, there is a wealth of systems to choose between as the broad architecture of the field has been constructed and yet endless details remain unexplored. This is an exciting time for microbiology and computational chemistry to overlap!

1.2 Bacterial Membranes and Secretion Systems

Bacteria are classified as either Gram-negative or Gram-positive[10], depending on the characteristics they exhibit upon staining with crystal violet. Gram-positive bacteria

retain the dye due to a thick peptidoglycan layer surrounding their sole cell membrane. Gram-negative bacteria contain two membranes: a cytoplasmic inner membrane (IM) composed of a mixed lipid bilayer[11], and an outer membrane (OM) with a similar structure but different composition. The OM inner leaflet primarily contains phosphatidylethanolamine (PE), with phosphatidylglycerol (PG) and cardiolipin also present, and the outer leaflet includes large rigid lipopolysaccharide (LPS) molecules[11]. These membranes are separated by periplasmic space, containing a thin peptidoglycan layer (Figure 1.1) and a high concentration of calcium ions. The OM is destroyed by alcohol used in the Gram staining process and the remaining thin peptidoglycan layer does not retain the dye. The Gram-negative IM is impermeable (unlike the OM), allowing cellular processes to be coupled to the energetic gradient across this membrane[11].

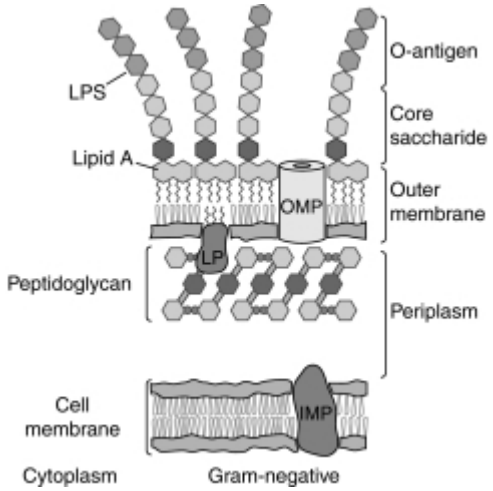


Figure 1.1 – Gram negative bacterial membrane structure
Adapted from Silhavy *et al.,* 2010[12]. Gram-negative bacteria have a cytoplasmic inner membrane composed of a mixed lipid bilayer and an outer membrane with a similar structure. These membranes are separated by aqueous periplasmic space, containing a thin peptidoglycan layer. Depiction of a Gram-negative cell envelope: LPS, lipopolysaccharide; OMP, outer membrane protein; LP, lipoprotein; IMP, inner membrane protein.

Among their many life processes, bacteria often produce and secrete substrates, many of which are toxic, such as cholera toxin (CT) produced by *Vibrio cholerae* and botulinum toxin from *Clostridium botulinum*. Nine systems that export and secrete proteins from the cytoplasm into the environment have been described in the literature to date. Type I Secretion Systems (T1SS) contain an ATP-Binding Cassette

(ABC) transporter, a membrane fusion protein and an OM protein, which transport substrates out of the cell in a single-step process[13]. Type II Secretion Systems (T2SS) allow transport across the OM for folded proteins that enter the periplasm either *via* the Sec (unfolded precursors) or Tat (proteins folded in the cytoplasm) transport pathways[14]; the T2SS will subsequently be described in more detail. Type III Secretion Systems (T3SS) – closed complexes spanning both membranes – perform a one-step process in which semi-unfolded effectors are believed to infect eukaryotic cells through a needle-like oligomer[15]. The Type IV Secretion System (T4SS), reviewed elsewhere[16], comprises a diverse group of systems that mostly rely on cell-to-cell contact to transfer substrates. It contains several subfamilies, including conjugation systems, effector translocators, and release/uptake systems. There are two groups of Type V systems (T5SS): Two Partner secretion[17] and the Autotransporter pathway[18], both requiring transporters from the Omp85 super-family. The former allows secretion of long β-helical proteins across the OM, in a process requiring both the secreted "cargoes" and their binding "transporters". In the latter pathway, proteins mediate their own translocation as their β-domains form OM pores through which the passenger domain is then secreted. The recently identified Type VI Secretion System (T6SS) secretes proteins in a one-step process using a bacteriophage-like needle to deliver the substrate into target cells[19]. The Type 7 secretion system (T7SS) is a specialised secretion pathway necessary for mycobacterial virulence, such as that of *Mycobacterium tuberculosis* (attenuated strains, with the T7SS locus removed, are used in vaccinations). This path resembles the T2SS, but does not include a pseudopilus structure, and the translocation mechanism remains unknown[20]. The Type VIII system (T8SS) – also known as the extracellular nucleation-precipitation (ENP) pathway – exports so-called curli subunits, which subsequently oligomerise to form functional amyloid fibres involved in host immunity interactions and biofilm formation[21]. Details of the most recently discovered system, Type IX, remain sparse. However, it appears to contain at least 12 essential proteins, with a large periplasmic collar complex found around the OM pore, and translocation may be linked with substrate attachment to the outer cellular surface[22]. Interestingly, evidence has emerged of this system cooperating with a T2SS in *Burkholderia cenocepacia*, in a novel mechanism for inflammasome activation inside infected macrophages[23]. Bacteria can also release membrane vesicles, used for

long-distance protein delivery/secretion[24].

Many Gram-negative bacteria are either confirmed or believed to use the T2SS (or "secreton") to release anywhere between one and over twenty different specific substrates with important roles in pathogenesis, such as toxins[25], hydrolases[26], cytochromes[27] and adhesins. The T2SS is a specialised protein ensemble that secretes folded exoproteins from the periplasm into the extracellular medium in a two-step process, through the secretin protein spanning the bacterial OM[26]. T2SSs contain approximately 12-15 different types of proteins essential for function, which are usually encoded in a single operon[28]. Most of these proteins are present as oligomers that form 4 sub-assemblies: the pseudopilus, the OM complex, the IM complex, and the secretion ATPase (Figure 1.2). The pathway was first discovered in *Klebsiella oxytoca,* where it secretes a single enzyme of the amylase family, pullulanase, and was hence designated the Pul (pullulanase) system; those interested in further details are invited to read recent comprehensive reviews on the T2SS[26,28–30], as only an outline of the system can be provided here.

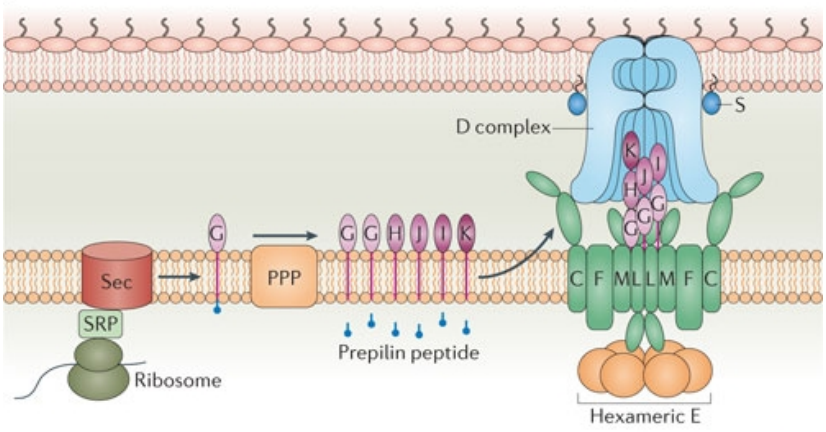


Figure 1.2 – T2SS outline
Adapted from Korotkov *et al.*, 2012[29]. Schematic of the Type II Secretion System (T2SS) sub-assemblies, and pilus biogenesis. For T2SS proteins, only the generic capital letter is shown (the Pul prefix is omitted). The OM complex is shown in blue, pseudopilins in pink, the IM platform in green and the hexameric secretion ATPase as orange spheres. Pseudopilin processing and pseudopilus assembly (with the tip of the pseudopilus added to the inner- and outer-membrane complexes) are shown, although the stage at which pseudopilins are incorporated has not been confirmed. SRP, signal recognition particle; PPP, prepilin peptidase.

Flexible periplasmic filaments, known as pseudopili and formed from oligomers of the major pseudopilin PulG, are capped by a complex of four other pilins (PulH/I/J/K)[31]. Cytoplasmic secretion ATPase PulE is believed to hydrolyse ATP, providing the energy required for the T2SS to translocate proteins across the OM. This hexameric protein is a member of the Type II/IV secretion ATPase family[32–34]. The OM complex consists of the secretin PulD and a lipoprotein designated PulS. PulD belongs to a family of multimeric channels that includes secretins from the Type III and IV secretion systems, and is brought to the periplasm *via* the Sec pathway mentioned previously[29]. PulS is a lipoprotein and a pilotin, a member of a particular chaperone protein subclass that targets the secretin monomer to the bacterial OM[35]. PulS uses the lipoprotein OM localisation (Lol) sorting pathway (discussed in more detail in section 1.3) to reach the OM[36]. Interaction with PulS targets the secretin to the OM[37], necessary for correct and efficient insertion of PulD into the bilayer. The IM complex forms a platform containing at least four core proteins – PulM/L/F/C[38]. Both pilus assembly and secretion are absolutely dependent on PulM, PulL and PulC. The IM complex is thought to transform conformational changes of the ATPase into pseudopilus motion leading to secretion, although the mechanism remains unknown and the proton motive force has been shown to be involved in substrate translocation[39,40]. Studies show that the T2SS shares multiple features with Type 4 pili (T4P), archaeal flagellae, and the natural competence-DNA transformation system[41,42]. T4P are the best-known and most studied members of the pilus-assembly nano-machine family[43]. Notably, experimental work on T4P can be extrapolated to the T2SS as T4P are indistinguishable from T2SS pseudopili in their dimensions, flexibility and bundling properties[44].

There are extensive ramifications for the improvement of agriculture and medicine if effective T2SS antagonists can be discovered or designed, as many human pathogens possess one or more T2SSs. Such pathogens include *Klebsiella sp*[45,46]*, Vibrio cholerae*[47], enterohaemorrhagic and enterotoxigenic *E. coli* (EHEC and ETEC)[48–50], *Pseudomonas aeruginosa*[51,52], and *Legionella pneumophila*[53]. Fish pathogen *Aeromonas hydrophila*[54] and plant pathogens such as *Dickeya dadantii* also contain a T2SS, demonstrating the prevalence of this system and the far-reaching implications of its pathogenic capabilities.

1.3 Pullulanase PulA and the Lol Avoidance Signal

Currently, the only known substrate secreted by the *K. oxytoca* T2SS is pullulanase PulA, a 116-kDa lipoprotein of glycoside hydrolase family 13[55]. Lipoproteins fulfil diverse functions in bacteria[56,57], including adhesion, membrane biogenesis and nutrient uptake. PulA secretion allows bacteria to depolymerise complex carbohydrates in the environment into metabolites, assisting colonisation of the host. $PulA^{Kox}$ performs random hydrolytic cleavage to degrade pullulan, a chain of α-1,6-linked maltotriose units (three glucose entities linked *via* α-1,4 glycosidic bonds). Secreted PulA remains surface-associated through its N-terminal acyl chains, which are embedded in the bacterial OM, unlike most other proteins secreted by the T2SS[58–60]. The $PulA^{Kox}$ crystal structure (Figure 1.3) has been determined by the team of Pedro Alzari in collaboration with Dr. Olivera Francetic and colleagues[61], and details are found in Chapter 3.

Following transport into the periplasm *via* the Sec machinery, PulA folds into the correct conformation in preparation for secretion, and after the Sec signal peptide is cleaved, the terminal Cys residue is triacylated[62]. The protein remains localised to the IM, according to the +2 sorting rule[63]. This states that presence of Asp at position 2 of an OM protein causes retention at the IM and prevents the protein entering the specific Lol pathway, which targets most lipoproteins to the OM. Asp2 is thereby acting as a Lol Avoidance Signal (LAS). The Lol pathway comprises cytoplasmic ATPase LolD, cytoplasmic membrane proteins LolC and LolE, periplasmic chaperone LolA and OM receptor LolB. It has been the subject of two thorough recent reviews[57,64] and is mentioned here briefly to provide context. Mature lipoproteins destined to the OM initially interact with LolE, are transferred to LolC, and are subsequently moved to LolA following ATP hydrolysis by LolD[65]. N-acylation is essential for the Lol-dependent release of OM lipoproteins[66] and LolCDE has been shown to recognise only triacylated Cys1, the sole common structure among lipoproteins[64]. Studies suggest that LolA contains a large hydrophobic cavity likely to bind acyl chains and its hydrophobic nature is necessary for LolA-dependent release of lipoproteins from the IM[57]. Interestingly, the IM localisation signal has been shown *via* studies of protein variants to specifically block the release

step catalysed by LolCDE[67–69], however the mechanism by which lipoproteins interact with the LolCDE complex and are released from the IM remains unknown.

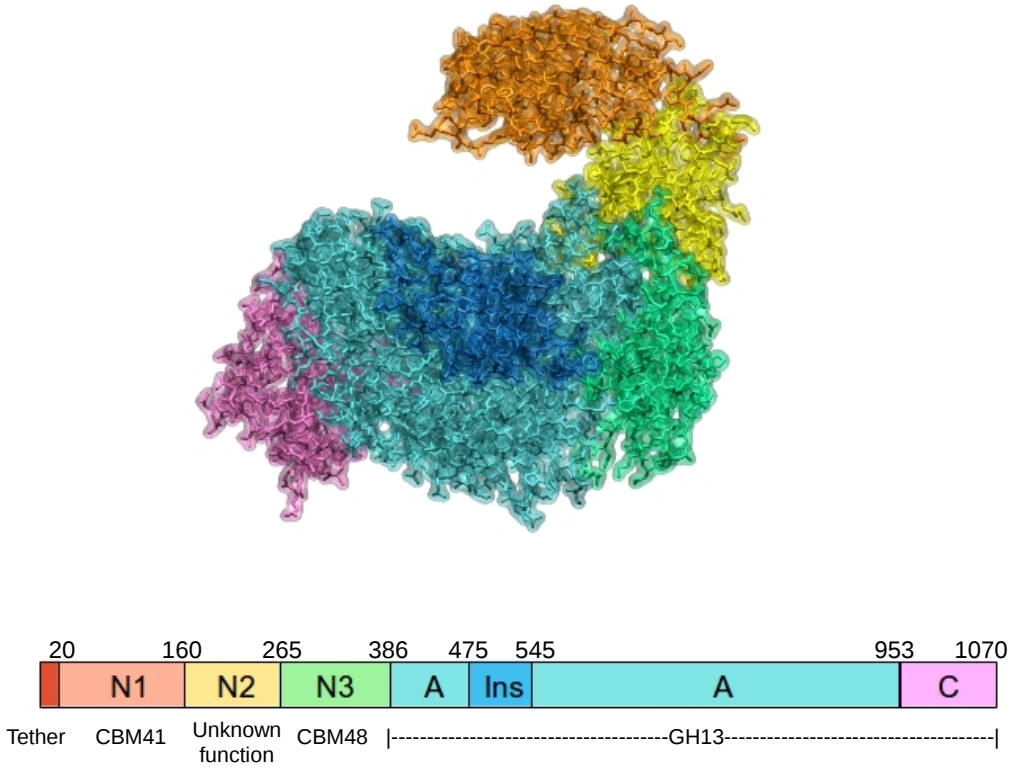


Figure 1.3 – Pullulanase PulA crystal structure
K. oxytoca PulA structure (PDB ID: 2YOC) and domain organisation. Surface representation of the 3D structure of a PulA protomer (subunit A) coloured by structural domains, with the domain architecture (according to the CAZy definition[70]) of the 1070 residue protein shown below it. PulA domains are shown in different colours: the N-terminal 19-residue tether (invisible in the crystal structure) is in red, N1 is in orange, N2 in yellow, N3 in green, A in cyan and C in purple, with the Inserted (Ins) domain in dark blue. See Chapter 4 for details on the catalytic site.

Significantly, substituting another residue for Asp2 in an IM-specific protein causes it to localise to the OM. (In the *Enterobacteriaceae* family, this rule appears to be fully conserved, although some IM lipoproteins with residues other than Asp at position 2 have been discovered in other branches of Gram-negative bacteria[71].) Even when Asp2 is present, the Lol avoidance function is abolished when the negative charge is chemically dissipated (and conversely the oxidation of Cys1 to cysteic acid generates an avoidance signal). The distances between the Cα and negative charges of these two residues (Asp2 and cysteic acid at position 1) are calculated to be similar, therefore proximity of the negative charge to Cα at the 2 position has been

concluded to be critical[72]. It was later discovered that the residue at position 3 is also important, with His3 or Lys3 allowing only partial lipoprotein retention at the IM, even with Asp2 present[73]. Native IM lipoproteins in *E. coli* contain either Asp, Glu or Gln at position 3; PulA from *K. oxytoca* contains Asn3. Unusually, Asn2 also functions as a retention signal when the residue at position 3 is Asp, as in *E. coli* protein AcrE[74,75]. Systematic substitution of residues 2 and 3 indicated that a negative charge or amide group should be present at position 3 for retention of lipoproteins with Asp2 at the IM[76].

Investigations into the effects of membrane phospholipids on lipoprotein sorting in *E. coli* have been clouded by the putative formation of non-physiological structures and the presence of ions in experiments. Reconstituting varied proteoliposomes did show that the positive charge of zwitterionic PE is essential for Lol avoidance in *E. coli*, likely due to electrostatic and steric complementarity between Asp2 and PE[72]. In the absence of PE, *E. coli* can grow and correctly sort lipoproteins only in the presence of a high concentration of magnesium, where unphysiological non-bilayer lipid structures are believed to form[77]. Further research is required to fully elucidate the role of lipids in Lol avoidance.

No MD simulations of bacterial lipoproteins and specifically of the LAS are present in the literature to date, although an overview of simulations relating to the T2SS is found in section 1.7. With the flexible tether regions of secreted lipoproteins often poorly characterised in X-ray crystal structures due to poor electron density, questions remain regarding the interactions of this protein region with the membrane, and how this enables Lol avoidance. Chapter 3 contains *in silico* investigations performed into PulA interactions with a lipid bilayer, with triacylated Cys1, a substituted Met1, and a D2S variant. Chapter 4 details simulation studies of PulA in an environment containing a high calcium concentration, extending the previous study by mimicking the physiologically relevant periplasm and exploring the possible effect of this environment on protein dynamics and protein-lipid interactions.

1.4 Role of Pseudopilin PulG and Possible Secretion Mechanisms

The key T2SS pseudopilin protein, PulG, consists of an extended amino-terminal helix with a globular head domain (Figure 1.4D), the fold of which is fairly conserved between bacterial species[78]. It is designated the major pseudopilin as it is the most abundant and the only pilin capable of forming long homopolymers[79]. The Signal Recognition Pathway recognises the N-terminus and targets it to the Sec translocon, which inserts PulG into the IM[80]. The N-terminal segment is then cleaved by prepilin peptidase PulO[51,81], which also methylates the N-terminus of the conserved phenylalanine (Phe1) residue[82]. The hydrophobic N-terminal helix is proposed to be required for pilin export, retention in the membrane, and the interactions promoting oligomer assembly[83]. PulG subunits are thought to oligomerise in the periplasm, capped by a tetramer of PulH/I/J/K[31]. PulK is believed to be the pseudopilus tip, as its larger globular domain extends above the smaller globular domains of PulI and PulJ[84], and it may prevent the pseudopilus extending beyond the outer membrane[44,85]. PulG/I/J/K are all required for protein secretion[86].

Sequence similarity between the 30 N-terminal residues of PulG and the T4Ps[87] suggests that the major pseudopilin may also assemble into pilus-like structures. Studies demonstrate that, following PulG overproduction, the T2SS produces extended, flexible filaments (pseudopili) composed of major pilin subunits, which are not present under physiological conditions[44,79]. Assembly of T2SS pili in *K. oxytoca* and *P. aeruginosa* correlates with functional secretion, implying the presence of short periplasmic pseudopili that promote secretion under native conditions[29]. These pseudopili play a crucial but poorly understood role in secretion, and studies into the relationship between their structure and function are ongoing. To develop this understanding, it is essential to determine their structure at high resolution and the significance of inter-protomer contacts at an atomic level. Truncated $PulG_{25\text{-}134}$(His)$_6$, has been crystallised[83], and the 1.6 Å resolution structure solved. Along with results from scanning transmission electron microscopy (STEM) analysis of PulG pili, this structure has subsequently been used to build a near-atomic pseudopilus model[83]. The resulting oligomer model has a rough surface with profound grooves separating helix strands and a very narrow hydrophobic central

cavity. The modelled pseudopilus has a helical pitch of 43.8 Å and 4.25 units per turn with an outer diameter of 65 Å, similar to the filaments studied by STEM containing an average of 4 turns and 17 PulG subunits in each helical repeat, with an outer diameter of 82 Å. Each monomer interacts with three protomers above (*P+1*, *+3* and *+4*) and below (*P-1*, *-3* and *-4*). The PulG monomer and dimer used in the work described here have been sourced from this structure (Figure 1.4). Recent electron microscopy has demonstrated continuously variable twist angles in the pili, supporting the idea of a spool-like mechanism for pilus assembly[88].

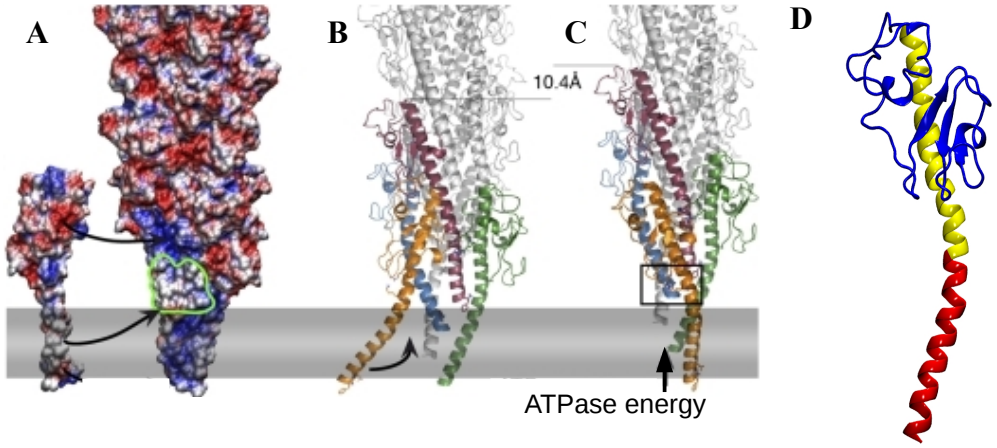


Figure 1.4 – Structure of major pseudopilin PulG
Adapted from Campos *et al.*, 2010[89]. **(A)** *K. oxytoca* PulG structure, following modelling using a 1.6 Å resolution crystal structure and results from scanning transmission electron microscopy. The electrostatic envelope of the protein monomer (left) and pilus (right) is shown inserted into the membrane (grey). The modelled pseudopilus has a helical pitch of 43.8 Å and 4.25 units per turn with an outer diameter of 65 Å. The arrows indicate the Asp48-Arg87 and Glu44-Arg88 salt bridges that tether the incoming monomer, and the hydrophobic patch is outlined in green. **(B)** The pilus and incoming subunit are shown in ribbon view, with each monomer differentiated by colour. **(C)** Incorporation of an incoming monomer *P* is associated with the membrane extraction of *P+1*, driven by the PulE ATPase, adding 10.4 Å of length to the oligomer. **(D)** Structure of PulG; the helix is shown in red and yellow (25 hydrophobic N-terminal residues shown in red), with the globular domain shown in blue.

In most T4Ps, the only charged residue in the first 25 amino acids is the conserved Glu5. This residue has been shown to be essential for assembly of pili from the *Pseudomonas* PAK strain, and for pilin methylation, which may be required at an earlier stage[44]. However it does not appear to interact with the N-terminal of the *P+1* protomer in the pilus to promote assembly[88], as previously suggested[90]. Possibly Glu5

interacts with the positive charge on the methylated N-terminus, as the side-chain is positioned to potentially form a salt bridge with the amine from methylated Phe1 in the PAK pilin structure. However, existence of this bridge under physiological conditions has not been confirmed. Also, a Glu5Asp substitution conserves the negative charge, allowing partial pilus assembly, but does not complement secretion, indicating that the role of this residue amounts to more than conserving charge. The periplasmic part of the PulG N-terminal helix has seven negatively and six positively charged residues, more than are present in T4P, and these may be engaged in inter- or intra-promoter contacts. Two salt bridges that form between Asp48-Arg87 and Glu44-Arg88 have been shown to be essential in PulG assembly and PulA secretion[89]. PulG binds a crucial calcium cation at residues Asp117 and Asp125, which stabilises a globular domain loop of the protein. Charge inversions (Asp124Arg and Asp117Arg substitutions) decrease PulG monomer stability and lead to piliation/secretion defects in these variants, consistent with the proposed role of Ca^{2+} in stabilising the loop[89]. The Ca^{2+}-binding site appears to be present across the pseudopilin family, despite the differences in the coordinating Asp residues[91]. In Chapter 5 I present MD simulations of PulG monomer variants and a homodimer, performed to complement the experimental work to date and provide an atomistic understanding of the interactions between methylated Phe1 and Glu5, and surrounding solvent, lipids or the adjacent PulG monomer. Once more, these simulations provided novel insights, and formed an original contribution to the field as no previous simulation-based data were published.

At least two models have been suggested to explain the currently unknown T2SS secretion mechanism – the “piston” and the “Archimedes' screw” models[92]. In the piston model, assembly/disassembly of major pseudopilins into oligomers extends and retracts the fibres repeatedly to push the secreted substrate protein through the secretin channel, like a piston, with the minor pseudopilins providing binding specificity for the substrate. Such specificity has been observed in the T2SS in *P. aeruginosa*, supporting this idea[93]. Retraction may occur *via* pseudopilin degradation and collapse, an idea supported by evidence that PulK over-production abolishes piliation but does not affect secretion, suggesting a role in controlling pilus length[44]. It has therefore been proposed that upon contact with PulD, PulK acquires the

capability to interact with PulG and orchestrate pseudopilus collapse – however currently this is merely a hypothesis and there is little other evidence for pseudopilus retraction in the T2SS. In the Archimedes' screw model, the substrate binds to the pseudopilus base and gradually leaves the cell as subsequent pseudopilins bind consecutively to the growing pseudopilus, in an action coupled to its rotation. Less biochemical evidence supports this model than the piston model. However, studies do imply that during T4P assembly, the fibres rotate relative to the assembly ATPase upon addition of each new subunit[94,95]. Similarly, there is evidence for rotation during/upon assembly of helical fibres from archaella[96], and conservation between these systems and the T2SS renders it plausible that the pseudopili have similar dynamics. Clearly, more research is needed to identify the correct model. MD simulations of a PulG homodimer, reported in Chapter 5, provided an initial atomic appraisal of the dynamic interaction interface between monomers and suggested directions for future experimental work.

1.5 Inner Membrane Assembly Platform Component PulM

As seen in Figure 1.2, the T2SS IM platform contains multiple copies of four different proteins; PulC, PulF, PulL and PulM. All of these components are believed to use the Sec machinery for membrane insertion[29]; PulC, PulL and PulM each span the membrane once, whereas PulF contains three TM helices. PulC/L/M demonstrably protect each other from proteolysis *via* protein-protein interactions[90,97–100], although their mechanism of interaction or assembly remains unclear. Putatively, the OM complex assists assembly of the IM platform, as fluorescently-labelled PulC and PulM only form foci in the *V. cholerae* envelope when the T2SS secretin is present[100].

PulM proved of particular interest in this thesis, following on logically from the studies of *K. oxytoca* PulG; after transport into the IM *via* the Sec transport pathway, research indicates that PulG may be chaperoned by PulM to the IM assembly platform (AP) complex. PulM is believed to interact with the pseudopilin in the IM, move it to the secretion machinery, and facilitate oligomerisation. Biochemical studies have assisted in gaining this understanding. Bacterial protein-fragment

complementation assays[101] have shown that the meningococcal major pilin PilE can interact individually with each of PilG (the PulF homologue), PilO (PulM homologue) and PilN (PulL homologue) – the IM platform components. Notably, this PilE-PilO interaction is apparently mediated by the globular periplasmic domains of both proteins and the 39 N-terminal residues of PilE[102]. A truncated version of the *Thermus thermophilus* major pilin PilA lacking the N-terminal hydrophobic α-helix interacted *in vitro* with both PilM/N (homologous to the cytoplasmic, and TM/periplasmic domains of PulL, respectively) and PilM/N/O complexes[103]. These studies point to the involvement of both the N-terminal and globular domains in pilin-PulM interactions.

As well as interacting with PulG homologues, the PulM periplasmic domain was considered a potential interaction partner for the periplasmic domain of PulL; yeast two-hybrid systems have been used to test for this interaction[38]. Following domain fusion to either LexA or its transcriptional activator motif B42, the proteins were co-expressed and the subsequent activation of *lacZ* demonstrated that PulL and PulM can indeed interact *via* their periplasmic domains. Repetition using truncated PulM showed that the final 79 amino acids of this protein interact with PulL and also that PulM can homodimerise *via* these residues[38]. GspM is demonstrably necessary for GspL (PulL homologue) stability[98]; in *P. aeruginosa* two stabilising domains were identified at the start and end of the periplasmic domain, and a third – localised next to the TMS – required the presence of GspC. GspL dissociated faster from a GspL/M homodimer than a GspC/L/M trimer in *X. campestri*, demonstrating the influence of GspC (PulC)[104]. Antibodies against GspM have co-immunoprecipitated GspL/C/F from cell extracts, confirming the existence of this quarternery complex[86].

An extensive study of *pilM/N/O/P* variants (PilM/N are homologous to the cytoplasmic, and TM/periplasmic domains of PulL, respectively, PilO is a PulM homologue) using stability and complementation analyses again suggested a complex of all four is required for optimal OM secretin function[105]. In *P. aeruginosa*, full-length pilin PilA co-eluted with a putative trans-envelope complex containing PilN/O/P and the PilQ secretin (homologous to PulD)[106]. Another study in *Pseudomonas* provided co-expression and biochemical evidence for direct

interactions between the periplasmic domains of PilO and PilN, which are crucial for T4P function[107].

Structural studies have assisted this understanding by uncovering the crystal structures of several secretion system components, to link these to protein function. For example, the homodimeric crystal structure of the periplasmic domain of *V. cholerae* PulM (EpsM) has been established and interestingly contains two αββ repeats that form a circular permutation of the ferredoxin fold[104]. Notably, this structure contains a peptide-like unidentified electron density in the cleft between the oligomers, suggesting the location of a partner binding site. GspM, a PulM homologue, is a bitopic protein[98], consisting of a short cytoplasmic domain, a TMS and a periplasmic domain that is involved in protein homodimerisation[38,97,108]. The crystal structure of *K. oxytoca* PulM remains unknown.

However, biochemical analyses have demonstrated that PulM is necessary for both pilus formation and PulA secretion in *K. oxytoca* (Santos-Moreno *et al.*, accepted). Notably, PulM and PulG have been shown to interact directly and independently of the other T2SS components, and BAC2H has shown that the highly conserved PulG E5 residue is the key determinant of this interaction (Santos-Moreno *et al.*, accepted). Biochemical and quantitative immunofluorescence studies have demonstrated that *ΔPulM* variants are severely defective in both pseudopilus assembly initiation and PulA secretion under physiological conditions (Santos-Moreno *et al.*, accepted). Formaldehyde cross-linking has shown that PulG and PulM interact in the context of the complete T2SS machinery (Santos-Moreno *et al.*, accepted).

No MD simulation data exploring the role and interactions of *K. oxytoca* PulM with PulG is found in the literature to date, as no structure has been available. With biochemical studies failing to yield data at atomistic resolution, questions remain regarding the interactions of PulM with the IM complex and the T2 pseudopilins. Extending the analyses of PulG presented in Chapter 5, this thesis contains an *in silico* study into the PulG-PulM interface using coarse-grained simulation methods to obtain dynamic data. Chapter 6 details the analysis of CG simulations including a PulM monomer, a PulG-PulM heterodimer and a heterotrimer.

1.6 Importance of the Work

There are multiple facets of the chosen systems that make their study valuable. *Klebsiella spp.* and *E. coli* are ubiquitous pathogens with relevance to human disease, industry and the environment. Among the eight species in the *Klebsiella* genus, *K. pneumoniae* is the most pathogenic to humans, followed by *K. oxytoca,* and these are responsible for the most human *Klebsiella* infections[109]. They are opportunistic pathogens, found in the environment and on mammalian mucosal surfaces, and are the cause of many nosocomial infections, which can be fatal[110]. Both *K. pneumoniae* and *K. oxytoca* cause community-acquired meningitis and brain abscesses, predominantly in Asia[109]. *K. oxytoca* is implicated in neonatal bacteraemia and septicaemia, especially in neonatal intensive care units (where it is among the top 4 infectious pathogens) and among premature infants[149]. It is the second most frequent causative agent of Gram-negative neonatal bacteraemia, and is also responsible for many cases of antibiotic-associated hemorrhagic colitis[111]. *K. oxytoca* can also cause bovine mastitis, an infection of the mammary glands that can prove fatal[112]. The prevalence and severity of *Klebsiella* infections demonstrate the need for deeper understanding of the Type 2 secretion system, which may allow more effective treatment and prevention protocols to be developed.

The role of the T2SS in *Klebsiella* pathogenesis has not been extensively studied, while in many species including ETEC, *Dickeya* or *Vibrio* it is essential for virulence and the secretion of virulence factors. T2SS proteins may be required for proteolysis[152], lipolysis[153] and phospholipolysis[154] assisting colonisation or tissue destruction, and enable an infection to be established on mucosal surfaces such as in the gastrointestinal or respiratory tracts. The T2SS is prevalent among bacteria including several extracellular pathogens, such as those noted at the end of section 1.2. Studies have shown that inactivating several T2SS genes in these species prevents translocation of multiple proteins, such as various plant cell wall-degrading enzymes in *Erwinia* species and *X. campestris,* and exotoxin A, elastase, phospholipase C, alkaline phosphatase, and lipases LipA and LipC of *P. aeruginosa*[113–115]. In *V. cholerae*, the T2SS secretes cholera toxin (CT), hemagglutinin-protease, and chitinase[47,116,117]. Cholera remains an important disease

in several parts of the world, causing potentially fatal diarrhoea. CT is the primary virulence factor, and probably the most extensively studied T2SS substrate. *P. aeruginosa*, an opportunistic human pathogen that often causes infections in cancer, burn and cystic fibrosis patients, also uses the T2SS. It can likewise grow in soil and water, infecting animals and plants. *Burkholderia cepacia* is also present in soil, water and plants, is associated with infections in cystic fibrosis patients too, and was recently discovered to contain a T2SS for protease and lipase secretion. T2SS genes have also been found in *Legionella pneumophila*[118–120], which causes Legionnaires' disease (a severe form of pneumonia).

On a non-pathogenic note, pullulanase is an industrially valuable enzyme, used alongside saccharifying amylases (such as gluco-amylase or β-amylase) for the production of sugar syrups, as it improves yield, reduces reaction time, and allows both an increase in substrate concentration and a reduction in the use of gluco-amylase[121,122]. The enzymatic conversion of starch into glucose, maltose, and fructose is an important and growing area for the food industry. Pullulanase can be used in the production of high-maltose corn syrup (used in food processing, with new interest from the pharmaceutical sector)[123] and high-glucose syrup (a carbon source in fermentation, and used for making crystalline glucose and high-fructose syrup[124,125]). Pullulanase has likewise been used to prepare high-amylose starches, which can be processed into nutritionally beneficial "resistant starch"[126]. Pullulanase is also helpful in carbohydrate structural studies[127]. Pullulanases are used in branched cyclodextrins (CD) production, and catalyse condensation of malto-oligosaccharides and CDs[128]. CDs have a variety of applications in complexing materials in foods, plastics, pharmaceuticals, and agricultural products as antioxidants and emulsifiers[129]. Knowledge of pullulanase interactions with the lipid membrane may allow us to secrete it more effectively, for commercial gain.

Bacteria with T2SSs are also important environmental agents, recycling a huge mass of biopolymers and renewing life resources; cellulose and chitin are the most abundant biopolymers on earth. The bacteria also allow reduction of metal oxides and render iron and manganese, for example, bio-available[27]. Understanding their mode of secretion may have implications for how we negotiate the ongoing

challenges of improving our environment and revoking damage to our ecosystems.

1.7 Using Molecular Dynamics (MD) Approaches

To understand the mechanisms of protein complexes such as the T2SS, and their interactions with localised cell membrane or ions, knowledge of the structure and dynamics of the system is essential. Although protein structures can frequently be determined by X-ray crystallography or nuclear magnetic resonance (NMR) methods, these techniques yield a static, time-averaged structure, obtained under *in vitro* conditions, the dynamics of which (beyond local vibrations) remain unknown. Few lipid-bound protein structures are available, due to factors such as poorly resolved electron densities or the use of non-native detergents prior to crystallisation[130]. Forster Resonance Energy Transfer (FRET) techniques can be used to examine interactions between tagged biomolecules, but do not yet yield atomistic detail[131]. Therefore MD simulations can be used to observe atomic interactions within the system. Computational techniques are now routinely used to complement experimental data and have provided multiple insights into, for example, the dynamics of biomolecular systems. Predictive results are also used regularly used to direct future experimental work, which can subsequently corroborate *in silico* results.

MD calculates the time-dependent behaviour of molecules and provides atomistic representation of the conformational fluctuations of a biological system in a model of a physiologically relevant environment. A crystal structure is usually used to provide the required initial configuration of atomic coordinates, as mentioned previously, with modelling programmes or homology modelling used to model atoms missing from the structure or to create variants. Simulations in which all system atoms (protein, lipid and solvent) are represented in atomistic or near-atomistic detail enable extraction of detailed information about structural and dynamic system properties. Coarse-grained (CG) methods, in which single particles represent small groups of atoms, and allow fewer degrees of freedom, enable simulation of longer time-scales and larger components, at the expense of accuracy. The MD approach is a demonstrably powerful technique to explore biological processes such as secretion, but arguably its full potential has yet to be realised in the field.

Studies of secretion systems using MD approaches, which have emerged in the last 15 years as computational resources become faster, more accessible and less expensive, cover two main areas: effector mechanisms and pilus formation. Simulations have provided insights into the mechanisms of effectors (both apo and in complex with ligands), an effector docking site present on a chaperone protein, and key determinants of effector substrate specificity. For example, the newly discovered crystal structure of phospholipase effector Tle1, from an unexplored super-family of T6SS effectors in *P. aeruginosa*, comprises two domains, one of which is a putative membrane-anchoring domain. MD simulations demonstrated that this domain remains stable over 50 ns when embedded in a lipid bilayer, supporting a cell-toxicity assay showing that the domain is crucial for antibacterial activity[132]. In a study of SpvC, a T3SS effector from *Salmonella serovars*, 10 ns MD simulations of protein variants with peptide substrates confirmed the stability of the binding site, thereby supporting the use of a truncated active-site model for quantum mechanics calculations[133]. This method permits a compromise between accuracy and efficiency, for effective use of resources. Short MD simulations of ExoT, another T3SS effector, with its homodimeric chaperone SpcS identified Asn65, Phe67 and Trp88 as the interfacial chaperone residues strongly affecting effector-chaperone interactions, complementing site-directed mutagenesis and demonstrating the hitherto unknown effects of substitutions on distal regions[134]. Similarly, simulations of cycle-inhibiting factor homologue in *Burkholderia pseudomallei* (CHBP) demonstrated that the electrostatic interactions mediated by residue Glu31 are the key determinants of substrate specificity and preference for NEDD8 (ubiquitin-like neural precursor cell expressed developmentally down-regulated 8) over ubiquitin[135]. MD approaches have also been used to perform secreton structure studies. Simulations of secretion system components have provided evidence of the structural elasticity of YscD, an IM component from the *Y. enterocolitica* T3SS injectisome, over 70 ns[136]. FlhB membrane protein, a component of the flagellar T3SS of *Salmonella typhimurium*, has been examined using MD techniques[137]. The flexibility of the tether between the cytoplasmic and transmembrane (TM) domains has been shown to depend upon the presence of a short loop in the globular cytoplasmic domain, which subsequently has implications for decreased functional activity.

MD research has also provided mechanistic insight into pilus formation and mechanism. Steered MD to observe force-induced conformational changes in an 18 subunit oligomer has demonstrated that the strength of the T4P is due to hydrophobic contacts between buried α-1 domains of different pilins interacting in the filament core[138]. As previously mentioned in section 1.4, Cisneros *et al.* have used MD to confirm that the anticipated helical register of the pseudopilus is observed in simulations of the membrane-embedded tip complex, that the PulJ homologue undergoes conformational changes when bound into the tip complex, and that the PulK homologue does not fully insert into the membrane, causing deformation. The computational and experimental results concur to suggest that PulI/J/K self-assemble into a pseudopilus-like complex in the bacterial IM. Evidence for an electrostatic export mechanism for MxiH, the needle monomer of the T3SS in *Shigella flexneri*, to move through its needle has been obtained *via* steered MD, using implicit solvent[139]. The results suggested that the aromatic groove (with Trp10 playing a key role) of the substrate is important for export, and that there is an energetic barrier for MxiH to enter the needle, which may be overcome by force provided from the ATPase. To date, the literature focuses on T3SS and T4P, with *Yersinia* proving the most prevalent model organism, and no simulations except those of the Pul pilus relate to T2SSs, thereby necessitating work such as that presented in this thesis to expand the field.

1.8 Thesis Aims

To provide novel insight in this field, an extensive series of both atomistic and CG MD simulations on multiple systems from the *K. oxytoca* Pul secreton, over hundreds of nanoseconds, has been performed. Detailed *in silico* examination of PulA, PulG and PulM – essential elements of the T2SS – has been performed, supplemented by comprehensive experimental data previously available. PE-containing lipids are the most prevalent in several bacterial species[140], including those that produce pili[141], and so POPE was chosen as the model for the single-lipid systems in this thesis.

Chapter 3 contains analyses of simulations of lipoPulA and non-acylated PulA ($PulA_{NA}$) with a POPE lipid bilayer representing an approximation of the biological state in which the substrate is found prior to secretion, which attempt to elucidate the role of the acyl tail in protein-membrane interactions before secretion. Likewise novel insights into the role of the Lol avoidance pathway are provided, in the light of simulations performed on the D2S $PulA_{NA}$ variant. Chapter 4 extends the PulA study and contains $PulA_{NA}$ simulations containing calcium chloride in the presence and absence of the lipid bilayer, produced to investigate the possible effect of the periplasmic environment on the protein. Chapter 5 contains atomistic simulations of PulG, the pilus monomer from the Pul system, and a series of variants, which provide insight on the crucial roles of essential residue Glu5 and the methylation of Phe1, previously identified through experimental work. Simulations of PulG dimers demonstrate the formation of salt bridge contacts when the protein is embedded in the membrane, and provide a perfect example of how MD can predict contacts that are likely to happen *in vivo*, as the membrane provides a native-like environment and constrains the number of possible conformations that the proteins might otherwise adopt if they were in solution. Chapter 6 contains analysis of CG systems containing combinations of PulG and PulM, identifying for the first time the likely protein interface as a direction for future experimental mutational studies.

Chapter 2 – Methods

To understand the mechanisms of nano-machines composed of multiple proteins in complex, it is essential to study the structure and dynamics of each system. However, studying such complexes in real time and/or *in vivo* presents a formidable challenge, particularly at atomic resolution. Although appropriate biophysical techniques continue to be developed and improved, difficulties remain with respect both to isolating required proteins and their subsequent characterisation.

Protein structures are usually obtained *via* X-ray crystallography or nuclear magnetic resonance (NMR) studies. Even under the assumption that the desired protein can be purified and crystallised, these methods yield only a static structure averaged over time, without providing information on the associated dynamics (beyond thermal motion around equilibrium, for example in the form of B factors)[142]. These challenges make protein complexes ideal candidates for theoretical studies; for example, MD simulations can be used to observe atomic interactions within these systems, providing structural, dynamic and thermodynamic data. Importantly, simulations also allow molecules to be modelled in their native environment, enabling the study of proteins in membranes rather than solely in the crystal lattice state, if appropriate. To date, such simulations have been used primarily to complement existing experimental data[143,144] in the study of numerous dynamic processes[143,145–149], such as protein folding, conformational changes and ion transport. In this thesis, MD methods are used to investigate the steady-state properties of bacterial components in physiologically relevant environments, directing and supplementing experimental work on protein interactions in secretion systems.

2.1 Statistical Mechanics

The macroscopic properties of molecular systems can be described by simple thermodynamic laws relating to the energy and temperature of the system. Statistical mechanics describes microscopic system properties by using the mechanical properties to explain system thermodynamics when the system state is uncertain, relying on a probabilistic approach. Each system state, for example every possible

conformation of a protein, is represented by a unique point in phase space – the *6N* dimensional space that represents all possible system states, as each particle has 3 position and 3 momentum coordinates. Phase space thus contains all possible combinations of momentum and position variables, and MD simulations explore part of this space. Statistical mechanics is required to relate such microscopic data to observable/macroscopic properties, such as temperature, pressure or energy.

Each MD simulation is computed within a particular ensemble (further explained below) whereby three macroscopic properties are fixed and the system is held in equilibrium. The remaining properties of the system are expected to fluctuate around a steady value. Assuming that an indefinitely evolving system will eventually pass through all accessible energy levels, the ergodic hypothesis (equation 2.1) states that the ensemble and time averages of a system property X should be the same. Here N_e is the number of the accessible energy levels and X_i is the value of X at energy level i.

$$X_{macro} = \langle X \rangle_{ensemble} = \frac{1}{N_e}\left(\sum_{i=1}^{N_e} X_i\right) \qquad \text{Equation 2.1}$$

As long as there is sufficient sampling of phase space, simulation data can be related to the ensemble averages of macroscopic system properties.

The probability P_i of any particular state i being sampled, according to the underlying theory, is proportional to the Boltzmann factor (equation 2.2) for that state:

$$P_i \propto e^{\left(\frac{-U_i}{k_B T}\right)} \qquad \text{Equation 2.2}$$

where U_i is the free energy of state i, k_B is the Boltzmann constant and T is the temperature. These factors can be summed to give the system partition function Q, which contains complete information on the overall system:

$$Q = \sum_{i=1}^{N_s} \rho_i P_i \qquad \text{Equation 2.3}$$

where ρ_i represents the degeneracy of state i. P_i is then given by:

$$P_i = \frac{\rho_i e^{\left(\frac{-U_i}{k_B T}\right)}}{Q} \qquad \text{Equation 2.4}$$

If Q is known, then the system's average internal energy can be found *via* differentiation:

$$U_{system} = -kT \frac{dQ}{d\left(\frac{1}{kT}\right)}, \qquad \text{Equation 2.5}$$

and all other thermodynamic properties can be derived *via* similar operations on Q. It is only possible to sample all the configurations necessary to approximate the ergodic hypothesis when there are no kinetic barriers between system states that prevent some of these being explored within the simulation time scale. Limited time-scale simulations may otherwise not sample high energy regions (or even many low-energy regions separated by multiple barriers) sufficiently, leading to errors in calculating the absolute values of macroscopic properties. However, simulation results are usually compared with experimental work relating to the difference between two states, so a relative and not absolute value of macroscopic properties will suffice and supports the use of MD.

2.2 Thermodynamic ensembles

As mentioned, MD simulations produce a series of points in phase space as a function of time in a particular thermodynamic ensemble. Traditionally, MD was performed in the microcanonical NVE ensemble (Number of Particles, Volume, Energy remain constant) because the total energy of the system is conserved by the application of Newton's laws. However, most physical experiments are not carried

out in the NVE ensemble. It is also possible to study the NVT (constant Number of Particles, Volume, Temperature) and NPT (constant Number of Particles, Pressure, Temperature) ensembles, by connecting the system to a thermo- or barostat respectively. Thermodynamic ensembles will be described here, with further detail on molecular mechanics provided subsequently.

A system's temperature is related to its kinetic energy by equation 2.6, which also relates to equations 2.7 and 2.8, where N is the number of particles in the system, M_i is the mass of atom i, v_i is the velocity of atom i, k_B is the Boltzmann constant, and N_f is the number of degrees of freedom:

$$K = \frac{1}{2}\sum_{i=1}^{N} M_i v_i \qquad \text{Equation 2.6}$$

$$K = \frac{1}{2} k_B T N_f \qquad \text{Equation 2.7}$$

$$T = \frac{2K}{N_f k_B} = \frac{1}{N_f k_B}\sum_{i=1}^{N} M_i v_i \qquad \text{Equation 2.8}$$

Scaling the velocities to satisfy a temperature constant is the most simple way to control the system temperature, termed a thermostat. It is possible to weakly couple the system to an external heat bath, fixed at the desired temperature, which acts to supply/remove heat to/from the system as required – known as the Berendsen thermostat[150]. The velocities are scaled at each step of the dynamics calculations, and the rate of change is proportional to the temperature difference between the system and the thermal bath:

$$\frac{dT(t)}{dt} = \frac{1}{\tau} T_{bath} - T(t) \qquad \text{Equation 2.9}$$

where τ is a coupling factor, determining how tightly the bath and system are coupled. The greater the value of τ, the weaker the coupling. This method provides exponential decay towards the desired T, and allows the system to fluctuate around the required temperature.

It is also possible to keep the system pressure constant, by scaling the volume of the simulation box appropriately, termed a barostat. Volume fluctuation in the isothermal-isobaric ensemble is related to the isothermal compressibility, κ:

$$\kappa = \frac{-1}{V}\left(\frac{\delta V}{\delta P}\right)_{\tau} \qquad \text{Equation 2.10}$$

The resulting scaling factor can be applied to the system's volume equally in the *x*, *y* and *z* dimensions (isotropically), the same in two and independent in the third (semi-isotropically), or be calculated independently for each (anisotropically). Analogous to temperature coupling, the pressure can be coupled to a pressure bath, and the resulting rate of change is given by equation 2.11, where $P(t)$ is the pressure at time t, τ_P is the coupling constant, and $P_{(bath)}$ is the bath pressure:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_p}\left(P_{bath} - P(t)\right) \qquad \text{Equation 2.11}$$

The volume is scaled by a factor, λ, defined in equation 2.12, which is equivalent to scaling the atom coordinates by $\lambda^{\frac{1}{3}}$, and the new atomic positions are obtained using equation 2.13:

$$\lambda = 1 - \kappa\frac{\delta t}{\tau_p}\left(P_{bath} - P(t)\right) \qquad \text{Equation 2.12}$$

$$r_i = \lambda^{\frac{1}{3}} r_i \qquad \text{Equation 2.13}$$

To improve sampling of the *NPT* ensemble, another method should be used, such as the Andersen barostat[151], which adds an extra degree of freedom, corresponding to the system volume. This has kinetic energy of $\frac{1}{2}m(dV/dt)^2$ where m equals the mass, and acts on the system like a piston, varying the pressure. The extended system coordinates relate to the real coordinates as follows:

$$r_i = V^{\frac{-1}{3}} r_i \qquad \text{Equation 2.14}$$

This method can be extended to allow the box to change shape, as in the Parrinello-Rahman barostat[152]. Three vectors (a, b and c) describe the box, and the box volume is given by the triple product of these vectors:

$$V = a \cdot (b \times c) = det(H), where\, H = (a, b, c) \qquad \text{Equation 2.15}$$

Most experiments are performed at constant pressure, reflecting *in vivo* conditions, so the NPT ensemble is most suitable for comparison with experimental data and has been used in the simulations described in this thesis. This is also a useful ensemble for membrane simulations, since it allows the system size – and hence the lipid area – to adjust spontaneously. In membrane systems the *xy* and *z* directions may be coupled separately, to ensure correct lipid behaviours.

2.3 Molecular mechanics

Molecular mechanics generally uses a framework of classical mechanics calculations to model molecular systems by generating ensembles of thermally relevant states. The potential energy of all system components in a particular ensemble is calculated using a "force field" (FF), a classical potential function containing both bonded and non-bonded terms. Classical FFs only use nuclear positions and omit the electronic aspect (unlike quantum mechanical methods), which greatly reduces the number of calculations required, making this a viable method for simulating large systems, as long as pair-wise approaches are used and the many-body effects are omitted. Classical representation is possible due to the validity of the Born-Oppenheimer approximation, which assumes that electrons respond instantaneously to any nuclear motion and always occupy the ground-state of that nuclear configuration, resulting in the separation of electronic and nuclear motions in an atom. Therefore the electronic aspect of the atom is not considered, enabling representation of each atom as a point particle that follows classical Newtonian dynamics. It is important to note that, due to this, molecular mechanics cannot be used to simulate electronically excited states, electron transfer processes or chemical reactions, and omits basic polarisability effects[153] (unless using a polarisable FF, such as AMOEBA).

In an all-atom molecular mechanics approach, each atom in the system is represented by a van der Waals sphere with a charge. Bonds between atoms are typically treated with harmonic potentials, with the equilibrium distance equal to the experimental or calculated length specific to the atom types within that bond (the reference length). United atom approaches consider terminal methyl or intermediate methylene units as a single particle[154] so do not represent non-polar hydrogens explicitly, and coarse-grained representations group even larger clusters of atoms together[155]. The increasing abstraction here facilitates longer simulation time, at the expense of atomic resolution.

2.4 Force fields

A molecular mechanics FF describes the potential energy of a molecule as a function of its atomic position. In turn, the potential energy derivative can be used to calculate forces on atoms and hence generate dynamics. FFs contain certain key components; energetic penalties associated with bond lengths or angles deviating from equilibrium reference values, a function representing the energy changes as bonds rotate, and terms describing non-bonded interactions. Atomic positions are represented by a pair-additive sum of terms that approximate the bonded and non-bonded interactions within the system:

$$V = V_{bonded} + V_{non\text{-}bonded} \qquad \text{Equation 2.16}$$

$$V_{bonded} = V_{angle} + V_{bond} + V_{dihedral} \qquad \text{Equation 2.17}$$

$$V_{non\text{-}bonded} = V_{electro} + V_{vdW} \qquad \text{Equation 2.18}$$

The bonded energies model stretching bond lengths using harmonic springs (V_{bond}), describe bond angle deformation energies (V_{angle}), and use sinusoidal functions to characterise the energies associated with dihedral angles ($V_{dihedral}$). The non-bonded interactions include both Coulombic electrostatic charges ($V_{electro}$) and the Lennard-Jones (LJ) 6-12 potential that represents van der Waals forces (V_{vdW}). The non-bonded interactions account for the largest number of interactions in a calculation. These terms will be discussed in more detail below.

Recently, attempts have been made to incorporate polarisability effects into FFs. An

external electric field (**E**) can cause changes in atomic charge distribution, inducing a dipole moment proportional to **E**. Polarisability is isotropic for isolated atoms, but the polarisability of molecules is often anisotropic. Attempts have been made to integrate polarisability into FFs by modelling polarisation effects at the atomic level[156], or representing the polarisation centre as a group of closely spaced charges[157]. However, polarisation calculations remain computationally expensive and can be problematic; for example, current models do not account for such induced charge distributions on two atoms influencing each other. Therefore polarisability effects are not incorporated into most FFs, including the ones used here.

FFs have specific applicability domains; one may be applicable for lipids but not for proteins, hence selecting the most appropriate FF is essential for ensuring system accuracy and reliability. For example, CHARMM lipid parameters and protein parameters are based on similar principles to ensure compatibility but are developed independently to ensure correct structural/thermodynamic properties, such as proteins adopting the correct folds, lipids assembling correctly and reproducing electron density profiles correctly. Numerous highly accurate FFs are available for use with biomolecular systems, including CHARMM[158,159], GROMOS[154,160], AMBER[161–163] and OPLS[164]. FFs are designed to be transferable, i.e. the parameters generated for a particular atom or group of atoms should be applicable to those atoms in other molecules. To this end, parameters are developed either from *ab initio* theoretical calculations or experimental data such as NMR or IR spectra, and are tested computationally. Such spectra are used for structural parameters, however other experimental data are often used for non-bonded parameters, such as free energies of solvation or partitioning. Parameter derivation methods and forms vary, so parameters cannot be transferred between FFs. FFs are empirical and there is no single correct FF – simply different compromises between accuracy and computational efficiency. As efficiency improves due to computational hardware developments, it is becoming possible to develop more accurate functional forms of FFs.

2.4.1 Bonded interactions

Bonded terms characterise the interactions of atoms linked by covalent bonds. Although bond stretching occurs, bond lengths rarely deviate significantly from their reference values in molecular mechanics calculations, so a simple expression such as the Hooke's law formula is sufficient in an FF (equation 2.19). Here, the energy varies with the square of the displacement from the reference bond length (l_0) , and the functional form is a suitable approximation to the bottom of the potential energy well curve, at distances corresponding to bonding in ground state molecules. Here v is the bonded potential, k is the bond's force constant, l is the current bond length and l_0 is the reference bond length:

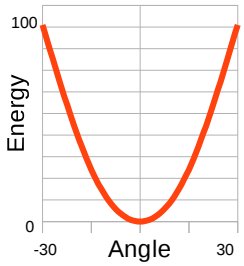
$$v(l)=\frac{k}{2}(l-l_0)^2 \qquad \text{Equation 2.19}$$

The deviation of bond angles (θ) from their reference values can also be described using a harmonic potential (equation 2.20), where again k is the spring constant:

$$\upsilon(\theta)=\frac{k}{2}(\theta-\theta_0)^2 \qquad \text{Equation 2.20}$$

Many major changes in molecular conformation are due to bonds rotating; to simulate this correctly, it is crucial to correctly represent the energy profiles of such changes in FFs. It is possible to use either hard or soft potentials; with the former, the angle needs to change significantly in order to have a large effect on the energy, whereas in the latter, the energy does not change significantly regardless of the change in dihedral angles, for example (Figure 2.1). The energy barrier is lower using soft potentials, which allows more extensive exploration of phase space.

Hard potential:



Soft potential:

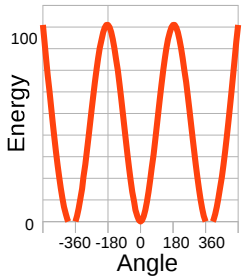


Figure 2.1 – Hard and soft potentials applied to dihedral bonds
Graphical representation of the relationship between rotational bond energy and angle using either a hard (left) or soft (right) potential. With the former, the angle needs to change significantly in order to have a large effect on the energy, whereas with the latter, the energy oscillates around a value and does not change significantly regardless of the change in dihedral angles.

Torsional potentials are almost always expressed as a sum of cosine functions (equation 2.21), where ω is the torsion angle and n is the multiplicity (the number of minima in the function as the bond rotates through 360°):

$$\upsilon(\omega)=\sum_{n=0}^{N} C_n \cos(\omega)^n \qquad \text{Equation 2.21}$$

2.4.2 Non-bonded interactions

Non-bonded interactions in a simulation occur through space, do not depend on a specific bonding relationship between atoms, and are usually modelled as an inverse function of distance. These interactions are usually considered in two terms - electrostatics and van der Waals.

Interaction energy between two atoms is zero at infinite distance and decreases as proximity increases, reaching a minimum value and subsequently increasing rapidly as proximity continues to increase. The force between the atoms equals minus the first derivative of the potential energy (V) with respect to distance (r).

Attractive forces are dispersive and long range, whereas repulsive forces act at short distances. Dispersive force occurs when an instantaneous dipole – arising during electron cloud fluctuations – induces a dipole in neighbouring atoms, thereby producing an attractive inductive effect. The Drude oscillator models a point charge

attached to an opposite point charge by a spring and provides a reasonable representation of this phenomenon, despite its simplicity[165]. The repulsive forces are in turn explained by the Pauli principle, which forbids any two electrons in a system from having the same set of quantum numbers and therefore from occupying the same spatial region.

An FF must contain a function to model the interatomic potential curve accurately, as calculating van der Waals interactions using quantum mechanics is not trivial and has proven impractical. The LJ 6-12 function (equation 2.22,) is the best known for this purpose and takes the following form for interactions between two atoms, where ε is the depth of the potential well and σ is the finite distance at which the potential is 0:

$$\upsilon(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \qquad \text{Equation 2.22}$$

An FF must also contain a function to model the interactions between molecules due to permanent dipole moments; the charges are approximated to point charges and the Coulomb potential is then applied. This is an effective pair potential that acts along the line connecting two charges to describe the interaction between them, and is given by equation 2.23, where r is the distance between the charges, q_1 and q_2 are the electric charge in coulombs carried by each particle respectively, and ε_0 is the electrical permittivity of space. The details of how this is achieved are explained in section 2.7.4.

$$V_{Coulomb} = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r} \qquad \text{Equation 2.23}$$

Notably, the interactions between two particles can be affected by the presence of a third, fourth or more particles. Nevertheless, it is often preferable to avoid calculating three-body effects as these greatly increase computational requirements. However, a significant proportion of these effects can be assimilated into a pairwise model if appropriately parameterised. Molecular modelling most commonly uses effective pairwise potentials, which are parameterised to include many-body effects in the pairwise energy. Hydrogen bonding is also incorporated into FFs and modelled in

various ways; some FFs incorporate an explicit hydrogen-bonding term, using a 10-12 LJ potential (equation 2.24) to replace the LJ 6-12 term between hydrogen-bonding atoms, and others reproduce hydrogen bonding solely using electrostatics and van der Waals interactions.

$$\upsilon(r)=\frac{A}{r^{12}}-\frac{C}{r^{10}} \qquad \text{Equation 2.24}$$

where $A=4\varepsilon\sigma^{12}$ and $C=4\varepsilon\sigma^{10}$. Often, FF performance is particularly affected by torsional and non-bonded terms, so optimising those parameters is usually more beneficial than focusing on other parameters that affect the results to a lesser extent (e.g. angle bending and bond stretching).

2.5 United atom and coarse-grained approaches

Decreasing the number of atomic interaction sites by subsuming atoms into their bonded neighbours reduces the computational resources needed and increases modelled time-scales by 2-3 orders of magnitude, relative to atomistic simulation time-scales. Due to the reduced degrees of freedom, the time step can be increased to 30-40 fs. Usually hydrogen atoms are subsumed, and this leads to segments such as methyl groups being modelled as a single "united atom"[154]. Usually only non-polar hydrogens (those bonded to carbon atoms) are subsumed, as hydrogens bonded to nitrogen or oxygen can participate in hydrogen bonding, which is modelled significantly better by an all-atom approach.

A number of coarse-grained (CG) approaches are possible[155], which vary by the number of atoms represented by a single particle. For example, residue-based modelling substitutes groups of ~ 10 atoms with a single bead, and represents amino acids by two CG beads - one for the backbone and one for the side-chain. The beads, like the spheres in atomistic simulations, are assumed to be point-like masses obeying Newtonian physics. The beads are linked by harmonic springs, and harmonic angular potentials assist in maintaining the overall molecular shape. Long-range interactions are modelled using LJ 6-12 and Coulombic potentials. Reference bond angles and lengths for the CG model are usually obtained by averaging the corresponding angles and distances over representative all-atom structures.

Alternatively, shape-based CG systems model the large-scale motions of macromolecular assemblies. Biomolecules assume numerous shapes, which often cannot be predicted from their sequences, and can vary from compact globular domains to elongated linkers. Shape-based CG methods allow all types of domain to be modelled with equal accuracy; the CG beads are placed using a topology-conserving algorithm which copies the molecule shape. Interactions in all types of CG simulations are parameterised using all-atom simulations and experimental data, and solvent is modelled implicitly, using the Langevin equation which allows solvent viscosity to be considered.

The MARTINI FF[166] was used to complete the work presented here. MARTINI parameterisation is based on the reproduction of thermodynamic data, specifically on the partitioning free energies between polar and apolar environments. These partitioning coefficients are crucial for the correct representation of processes such as peptide-membrane binding. On average, four heavy atoms are substituted by a single bead, which is categorised as one of four main interaction types: charged, apolar, polar or non-polar. These are separated into subtypes that allow for better representation of the chemical attributes of the group, and consider both its degree of polarity and its hydrogen-bonding capability (donor, acceptor, both, or none). The MARTINI extension for proteins uses the same method and is based upon the partitioning free energy of amino acids side-chains between water and oil phases.

The functional form of the MARTINI FF is very similar to the ones used in all-atom simulations, with the MARTINI functional for bonded terms described by weak harmonic potentials (see equation 2.25).

$$V_{MARTINI} = \sum_{bonds} \frac{1}{2} K_b (b - b_0)^2 + \sum_{angles} \frac{1}{2} K_\varphi (\cos\theta - \cos\theta_0)^2 + \sum_{dihedrals} k_\varphi (1 + \cos(n\varphi - \delta)) + \sum_{impropers} k_\omega (\omega - \omega_0)^2 \qquad \text{Equation 2.25}$$

As mentioned above, the CG beads are linked by harmonic springs. The equilibrium bond distance b_0, the force constant K_b, the equilibrium angle θ_0 and the angle force constant K_θ are fixed, with corrections for double bonds. A dihedral term is added for proteins, and an improper dihedral term is added to prevent angular

distortions out-of-plane. As mentioned, an LJ 6-12 potential function is used for the non-bonded interactions. Full charges on ionised groups are represented by a Coulombic term and an additional relative dielectric constant for explicit screening.

CG simulations present specific challenges. For example, using reduced representations affects the interpretation of time-scales; fewer particles of increased size, simulated using softer potentials, result in a smoother energy landscape. Due to fewer degrees of freedom, friction from atomic motion is missing and dynamics across the Potential Energy Surface are more easily sampled. As a result, the dynamics in CG simulations are faster than those in atomistic simulations, and the time axis must be scaled to compensate for this effect; a scaling factor of four matches experimental data well[166]. Likewise, ring particles contain a large density of CG bonds, and as a result both the minimum distance and strength of interactions are scaled down appropriately[165]. Also, it may be useful to conserve the protein conformation relative to the initial structure, in which case harmonic bonds can be used to implement an elastic network model mimicking the structure and large-scale dynamics of a particular state[167–170].

2.6 Lipid simulations

The interactions between lipids and proteins are key to our biological understanding; approximately 20-30 % of most genomes code for membrane proteins[171]. Simulating lipid membranes is desirable yet complex, partly due to the variety of membrane structures and compositions that exist. For example, mammalian endoplasmic reticulum contains up to 5 mol% cholesterol, whereas plasma membrane contains 25-40 mol% cholesterol[172]. Likewise, lipid rafts have recently been of particular interest in the field but these dynamic small membrane regions provide many challenges as they are densely packed with membrane proteins, cholesterol and sphingomyelin[173]. It is often difficult to observe crystallised native lipids bound to proteins due to poorly resolved lipid electron densities[174]. Also, detergents and precipitant molecules used prior to crystallisation can displace lipids from protein surfaces[130].

As in the case of protein simulations, lipid bilayer atom coordinates are required for

membrane simulations. Various pre-made and equilibrated membrane models are available online, and the CHARMM membrane builder server can be used to create mixed lipid systems (http://www.charmm-gui.org/?doc-1⁄4input/membrane). The bilayer of choice must be sufficiently large to avoid interactions between any proteins included in the system, and their periodic images (see section 2.8). CHARMM lipid parameters are used extensively, having been developed consistently with parameter sets for proteins and other biomolecules[159,175]. CHARMM36, used for the studies in this thesis, is an improved version of CHARMM27, which had previously produced inaccurate surface area values and incorrect gel-phase membranes[176]. CG approaches are popular for lipid simulations; the Marrink model is often used and groups four heavy atoms into each CG particle[166,177].

Notably, it is important to ensure that any proteins are positioned appropriately in or on the membrane during atomistic simulations, to subsequently obtain biologically relevant results. For example, hydrophilic and hydrophobic areas along the TM domain should be matched up to the changing character of the bilayer, and amphipathic aromatic residues should be placed at the interface between lipid and solvent. However, CG approaches allow the spontaneous assembly of lipids around proteins, and specific positioning is not crucial.

2.7 Energy minimisation

It is advisable to perform an energy minimisation of the system - identifying the nearest low energy local minimum - before performing an MD simulation. Notably, this process does not involve calculating dynamics. Minimisation relieves any unfavourable interactions in the initial system conformation that may lead to a subsequent MD simulation crashing, such as those caused by solvent addition during system preparation, overlapping atoms, homology modelling, or the use of a different FF to optimise experimentally deduced X-ray or NMR structures.

The minima on the potential energy surface of a system represent stable states. At a minimum point, the first derivative of the function, V , (which depends on the degrees

of freedom of the system) is 0 with respect to each derivative, and the second derivatives are positive. Minima cannot usually be identified analytically using standard calculus methods, as the energy varies with the coordinates in a complex manner. However, numerical methods can be used to change the coordinates progressively and produce conformations with increasingly lower energies, until a minimum is attained.

No minimisation algorithm can yet locate the global energy minimum from an arbitrary starting point. Energy minimisation produces a new system configuration, rather than a time-dependent trajectory. Derivative minimisation methods are the most popular, although non-derivative methods exist. The first derivative of the potential energy indicates the location of the minimum, and the magnitude of the value demonstrates the steepness of the surrounding gradient. The second derivative indicates the function curvature, which can be used to calculate when the surface will change direction (pass through a stationary point). Nevertheless, algorithms have been developed that locate the minimum closest to the starting point (by only taking steps to decrease the system energy), and some even move up the energy gradient, thereby possibly locating points with even lower energy than the closest minimum.

To locate a local minimum, the steepest descents (SD) method can be employed. SD moves in the direction parallel to net force, determined by identifying the largest interatomic forces. Once the direction is chosen, it is necessary to define how far to move along the gradient, and either of two approaches can be employed here. The arbitrary step approach takes a step of arbitrary length along the gradient vector, then calculates the energy, and iterates to repeatedly reduce the energy, adjusting step size until the minimum is located. The line search method identifies three points along a line such that the energy of the middle point is lower than the other two, demonstrating the existence of a local minimum. This is iterated, with each direction being orthogonal to the previous one, until the minimum is located. SD is a useful method for removing the highest-energy features in an initial conformation, and the method is robust even when the local minimum is far from the starting point. Notably, small steps are used when moving down a narrow energy “valley” and SD can over-correct, introducing errors.

Second derivative methods exist, such as the Newton-Raphson method, but these are computationally expensive and require a large amount of memory, as they involve inverting the inverse Hessian matrix of second derivatives. This method is therefore only suitable for systems containing fewer than 100 atoms. In general, the choice of minimisation algorithm depends on computational and storage requirements, method robustness, the availability of analytical derivatives, and the speed at which the calculations can be performed.

2.8 Production algorithm

Following minimisation of the system, production simulations are performed to determine the time-dependent behaviour of the atoms and elucidate how the system changes between conformations. The force on any particle can be calculated by differentiating the energy function (equation 2.16). The acceleration can be calculated from the force using Newton's second law (equation 2.26) where F = force, m = mass and a = acceleration. Integration of the equations of motion over time subsequently yields a trajectory describing the variation of particle coordinates, velocities and acceleration during the simulation.

$$F = ma \qquad \text{Equation 2.26}$$

During simulation set-up, periodic boundary conditions (PBC) are applied which allow the simulation to be run with relatively few particles, yet these particles experience forces as if they were in bulk solvent, thus avoiding edge effects. This is achieved by repeating the box of particles in all directions to give a periodic array, such that if a particle leaves the box during the simulation, it is replaced by an image particle entering from the opposite side of the box. In this way, the system is fully maintained. The system then undergoes equilibration, which allows it to transition from the minimised starting conformation to equilibrium. Equilibration continues until properties such as temperature, pressure and energy have stabilised, and is an important stage of system preparation, especially for membrane systems where the initial conformation of the membrane/protein complex is unknown and the lipids must be allowed to adjust around the protein, and solvent and ions to relax at the

membrane surface.

2.8.1 Calculation of forces

Historically the first MD simulations used simple potentials, such as the hard-sphere potential. Since then, more realistic potentials have been developed in which the force between two atoms varies continuously depending on their proximity. This continuous nature requires the equations of motion to be integrated over a series of very short time steps (usually 1-10 femtoseconds), which is a function of the fastest motion in the system. At each step, the force on each atom is calculated and the current positions and velocities are used to calculate new positions and velocities a short time ahead. Particles are then moved to the new coordinates, a new set of forces is calculated, and this process is repeated to produce a time-dependent trajectory that describes how the system dynamics change. MD simulations are usually run for tens to hundreds of nanoseconds, although increasingly powerful computers are extending this time-scale to microseconds[178].

As mentioned above, the motions of all the particles in the system are coupled due to the influence they exert upon one another, and the resulting many-body problem must be solved numerically using a finite difference method. Using this method, the integration is segmented into small steps, separated by a fixed time (δt). The total force on each atom at time t is calculated as the vector sum of all the interactions it experiences, and the acceleration can be calculated from this force, assuming the force is constant during the time step. The acceleration is combined with the velocities at time t to produce the positions and velocities at time $t+\delta t$. This process is then iterated, and the forces on the particles in their new positions are calculated, leading to the most recent positions and velocities at time $t+2\delta t$. Such algorithms assume that the particle positions, velocities and accelerations can be approximated as Taylor series expansions, where r is the position, v is the velocity, and a is the acceleration:

$$r(t+\delta t)=r(t)+\delta t\,v(t)+\frac{1}{2}\delta t^2 a(t)+\frac{1}{6}\delta t^3 b(t)+\frac{1}{24}\delta t^4 c(t)... \quad \text{Equation 2.27}$$

$$v(t+\delta t)=v(t)+\delta t a(t)+\frac{1}{2}\delta t^2 b(t)+\frac{1}{6}\delta t^3 c(t)... \quad \text{Equation 2.28}$$

$$a(t+\delta t)=a(t)+\delta t b(t)+\frac{1}{2}\delta t^2 c(t)... \quad \text{Equation 2.29}$$

$$b(t+\delta t)=b(t)+\delta t c(t)... \quad \text{Equation 2.30}$$

For these calculations, the Verlet algorithm[179] is the most popular, which uses the positions and accelerations at time t, and positions (r) from $(t-\delta t)$ to calculate $r(t+\delta t)$. These quantities are related to the velocities as follows:

$$r(t+\delta t)=r(t)+\delta t v(t)+\frac{1}{2}\delta t^2 a(t)+... \quad \text{Equation 2.31}$$

$$r(t-\delta t)=r(t)-\delta t v(t)+\frac{1}{2}\delta t^2 a(t)-... \quad \text{Equation 2.32}$$

And combining these expansions cancels the third-order terms to give:

$$r(t+\delta t)=2r(t)-r(t-\delta t)+\delta t^2 a(t) \quad \text{Equation 2.33}$$

The Verlet algorithm is simple to implement, with few memory storage requirements. However, there are drawbacks to this method: it is necessary to use other means to obtain $(r(t-\delta t))$ initially; the positions at $(t+\delta t)$ may be calculated imprecisely as they are obtained by adding a small term $(\delta t^2 a(t))$ to a much larger term $(2r(t)-r(t-\delta t))$; the lack of an explicit velocity term makes it difficult to obtain the velocities, and $v(t)$ cannot be calculated until $r(t+\delta t)$ is known.

A variation on this method, the leap-frog algorithm, explicitly includes the velocity and does not require calculation of the difference between large numbers. It calculates the velocities $\left(v\left(t+\frac{1}{2}t\right)\right)$ from $\left(v\left(t-\frac{1}{2}t\right)\right)$ and $(a(t))$ (equation 2.34). Atom positions at $\left(t+\frac{1}{2}t\right)$ are then calculated using the newly obtained velocities and the positions at (t) (equation 2.34).

$$v\left(t+\frac{1}{2}\delta t\right)=v\left(t-\frac{1}{2}\delta t\right)+\delta t a(t) \quad \text{Equation 2.34}$$

$$r(t+\delta t)=r(t)+\delta t\, v\left(t+\frac{1}{2}\delta t\right) \qquad \text{Equation 2.35}$$

Velocities at time (t) can be obtained as follows:

$$v(t)=\frac{1}{2}\left[v\left(t+\frac{1}{2}\delta t\right)+v\left(t-\frac{1}{2}\delta t\right)\right] \qquad \text{Equation 2.36}$$

This method is so called because the velocities "leap-frog" over the positions to give their values at $\left(t+\frac{1}{2}\delta t\right)$ and the positions then leap-frog over the velocities to give their new values at $(t+\delta t)$. An advantage over the Verlet approach is that the velocities are explicitly calculated, hence can be operated on by a thermostat. However the clear disadvantage of this algorithm is that r and v are not synchronised, so it is not possible to determine the potential energy of the system at the same time as the positions are defined.

When running a simulation, it is important to decide on an appropriate time step (δt) to use, in order to optimise computational efficiency, simulate the "correct" trajectory (with no instabilities due to high energy overlaps between atoms, and reproducing all pertinent motions), and cover the phase space. The time step should be approximately 1/10th the time of the shortest period of system motion, i.e. C-H bond vibrations, namely 1 fs. This is a very small time scale, 1 fs to a second is as 1 second to ~ 30 million years! Therefore many iterative steps need to be taken in an MD simulation to model anything on biologically relevant time scales. However, for CG simulations the time step can be increased; united-atom FFs can use 5-7 fs, and MARTINI-type models can use 20-50 fs.

2.8.2 Constraints

As stated, the time step of a simulation is dictated by the highest frequency motion, but these are often of less interest than the lower frequency motions that correspond to large conformational system changes. It is possible to extend the time step by applying system constraints to prevent high vibrational frequency bonds deviating

from their equilibrium lengths. Such constraints limit bond distances (or angle motions) without affecting other internal degrees of freedom, and allow standard integration of the equations of motion. The use of constraints also narrows the range of frequencies in the system, decreasing the inaccuracies caused by computing a broad range of forces.

Constraint algorithms enforce a list of constraints, which system coordinates must fulfil. These are holonomic and depend only on the coordinates and time, not velocity, keeping the motions of the particle on the surface of a sphere, and take the form:

$$\sigma_k = (r_i - r_j)^2 - d_{ij}^2 = 0 \qquad \text{Equation 2.37}$$

r_i represents the position of atom i, r_j the position of atom j, and d_{ij} is the constraint distance. In such a system, the equations of motion involve two types of force: the forces arising from the bonded/non-bonded interactions already described at length, and the forces due to the constraints, which are as follows:

$$Fc_{ki} = \lambda_k \frac{\delta \sigma_k}{\delta r_i} = \lambda_k (r_i - r_j) \qquad \text{Equation 2.38}$$

There are currently two algorithms in GROMACS to calculate the Lagrangian multiplier(s) λ_k, which satisfy(ies) all the constraints simultaneously. Known as SHAKE[180] and LINCS[181], the former iteratively calculates each constraint until all are satisfied to within an arbitrary tolerance, and the latter approximates the solution by using a series expansion to calculate the multiplier. The SHAKE algorithm is more useful as it can be used to constrain coupled angles, but it is more computationally demanding than LINCS, due to the iterative calculations required. LINCS provides an answer to the same accuracy 3-4 times faster than SHAKE, and therefore has been employed for the simulations here.

2.8.3 Non-bonded cutoff distance considerations

The calculated velocities depend upon interactions between atoms, as discussed in

section 2.2, and in MD simulations cutoffs can define which interactions are taken into account, to assist computational efficiency. For example, the LJ potential decreases rapidly with distance: at 2.5 σ , the LJ potential has just 1% of its value at σ , reflecting the r^{-6} distance dependence of the dispersion force. Both non-bonded cutoffs and the minimum image convention (MIC) can help to model this. Using non-bonded cutoffs, the interactions between all pairs of atoms that are further apart than the cutoff value are set to zero. Notably, when PBC are employed, the cutoff should be small enough to prevent the particle seeing its own image or any particle twice. The cutoff must be much greater when long-range electrostatics are used - at least 10 Å is recommended. This is because the Coulombic interaction is larger, dying off after ~ 30 Å, therefore Particle Mesh Ewald is used (see below). However, evidence suggests that using any cutoffs leads to errors. Using the MIC, each atom "sees" at most one image of every other atom in the system and only the closest image is accounted for when calculating the potential energy.

A non-bonded neighbour list can also increase efficiency and speed up calculations. The list stores all atoms within, and slightly outside, the cutoff distance. It is updated regularly throughout the simulation, and is used to identify the nearest neighbours to any given atom between updates. Update frequency must be chosen to provide efficiency without compromising accuracy; usually the list is updated between every 10 and 20 time steps.

2.8.4 Particle Mesh Ewald

To calculate the coulombic interaction energy of an infinite system, the electrostatic interactions between all atom pairs and their periodic images must be calculated, which can be performed using the Ewald summation method[182]. However, as this method scales at $O\left(N^{\frac{3}{2}}\right)$, MD simulations use the more efficient Particle Mesh Ewald method[183,184] instead. The charge distribution is discretised using a grid, and 3D fast Fourier transforms are used to calculate the Coulombic potential, which can be interpolated to give the potential on each individual particle. This approach increases the speed of calculation of electrostatics to order $O\left(Nlog\left(N\right)\right)$, and

therefore is widely used in MD simulations.

2.9 Details of presented atomistic simulations

All atomistic simulations were performed using the GROMACS 4.6 simulation package[185], with the exception of the PulA simulations, which used GROMACS 4.5.5 and PulM simulations, which used GROMACS 5.0.7. The CHARMM36 FF[175] was used to treat the lipid molecules, and CHARMM22/CMAP parameters[186] to treat the PulA (PDB ID: 2YOC) and PulG (PDB ID: 1T92) proteins. Membrane embedded proteins were inserted into an equilibrated and hydrated membrane using the *g_membed* tool provided in the GROMACS package[187]. *g_membed* reduces the protein in size around its centre of mass in the *xy*- and *z*-planes, and any lipid or solvent molecules that overlap with the resized protein are subsequently removed. The protein is subsequently iteratively resized in the *xy*-plane around its centre of mass, using an incrementally increasing resize factor, and overlapping molecules are removed until the protein reaches its original size (i.e. the resize factor = 1) in this plane. The same process is then repeated in the *z* direction, until the protein reaches its original size and is fully embedded in the membrane.

Each system lacking lipid was placed in a box with at least 2 nm between the protein and the adjacent periodic boundary. Systems containing a membrane bilayer were placed in a box with at least 2 nm between each protein and its periodic image. The systems were then explicitly solvated *via* superimposition of a pre-equilibrated box of TIP3P water molecules[188], followed by removal of any waters inserted into the lipid bilayer. The systems were electrically neutralised by random replacement of waters by counter-ions to a ~ 0.1 M final concentration, imitating physiological salt conditions. Counter-ion identities varied; details are provided in each chapter. Prior to and following solvation, energy minimisation was performed using an SD algorithm, to minimise any steric overlap between system components. This was followed by an equilibration simulation, allowing the lipid and solvent components to relax around the restrained protein. All the protein and lipid non-hydrogen atoms were harmonically restrained, with the constraints gradually reduced in 3 distinct 0.5 ns phases. Finally, production MD was executed in at least triplicate on the

unrestrained system for varying simulation lengths (detailed in each chapter), within the NPT ensemble.

All simulations were completed using the leap-frog algorithm with a 2 fs time step, and trajectory data was collected every 10 ps. LINCS[181] was used to constrain bond lengths and the neighbour list was updated every 10 steps using a 1.4 nm cutoff. A cutoff of 1.2 nm was used for LJ (excluding 1-4 scaled) interactions, with a smooth switch off between 1 and 1.2 nm. Long-range electrostatic interactions were corrected using the Particle Mesh Ewald method[184], using a real-space cutoff of 1.2 nm. The velocity-rescale thermostat[189] was used to maintain absolute temperature throughout the simulations as listed in each chapter respectively, to ensure the system temperature was above the membrane phase transition temperature. Pressure was set to 1 bar semi-isotropically (Parrinello-Rahman semi-isotropic barostat, 5 ps coupling constant), under PBC.

The methods and simulation details for all presented CG simulations are found in Chapter 6, section 6.2.

2.10 Analysis

Numerous thermodynamic properties can be calculated from MD simulations; comparing experimental and calculated values for such properties enables quantification of the simulation and energy model accuracy. MD also allows prediction of the thermodynamic system properties for which experimental data do not exist or cannot be obtained.

All simulation analysis presented here was performed using tools available in the GROMACS software package[190] and/or locally written code. VMD 1.9.1[191] was used for visualisation and for the production of all molecular graphics. Graphs were constructed using Grace (http://plasma-gate.weizmann.ac.il/Grace/).

Conformational change over the time course can be quantified by calculating root-mean-square deviation (RMSD) between particle coordinates, which demonstrates

the degree of structural drift of atomic groups of interest during simulation. The measurement is made with respect to a reference structure (usually the initial simulation conformation, or a crystal structure) on which the simulation structures are superimposed. Given structures at t_1 and t_2, the structural RMSD is defined as:

$$RMSD(t_1, t_2) = sqrt \frac{1}{n} \sum_{i=1}^{n} \left((t_{1ix} - t_{2ix})^2 + (t_{1iy} - t_{2iy})^2 + (t_{1iz} - t_{2iz})^2 \right) \quad \text{Equation 2.39}$$

where t_{1ix} is the x position of atom i at time t_1. However, it is also possible to calculate RMSD compared to a structure at $(t_2 = t_1 - \tau)$, rather than the initial structure. This provides some insight into the structural mobility as a function of τ.

RMSD gives an average taken over the particles, providing time-resolved values. Calculating root-mean-square fluctuation (RMSF) instead provides an average taken over time, giving a value for each particle, i. The RMSF for a subset of atoms over time t is defined as:

$$RMSF_t = N^{-1} S^{-1} \sum_{i=1}^{N} \sum_{j=1}^{S} \left(r_{ij} - (r_i) \right)^2 \quad \text{Equation 2.40}$$

where N is the number of atoms in the subset, S is the number of frames in the sampling window, r_{ij} is the positional vector of atom i in frame j, and (r_i) is the average of the latter over all the frames in the time period. From this, simulated atomic B factors can be calculated and compared with experimental values. Various factors, such as the system state (crystalline or in solution) and the time-scales for calculation, cause the experimental and MD-derived B factors to differ although usually ideally these values should be equal.

Chapter 3 – Protein-lipid interactions of PulA prior to secretion

Disclaimer
A portion of the work described in this chapter was included in a manuscript published in 2016, under the title "Structural determinants of pullulanase membrane association and secretion revealed by X-ray crystallography, molecular dynamics and biochemical assays". The non-computational experimental work found in this chapter was carried out by other authors on the manuscript, led by Dr. Olivera Francetic, Institut Pasteur, Paris. Parts of the text therefore were written in collaboration, however all the figures are my own work unless otherwise stated in the legend.

3.1 Introduction

Pullulanase PulA is the only known substrate of the T2SS in *K. oxytoca* ($PulA^{Kox}$) and is a model protein for secretion mechanism research, as discussed in Chapter 1. Pedro Alzari, Olivera Francetic and their colleagues crystallised a periplasmic nonacylated PulA variant ($PulA_{NA}$) (containing a C-terminal thrombin cleavage site followed by a hexahistidine tag) and resolved the structure to 2.9 Å (PDB ID: 2YOC; Figure 3.1A)[61]. To generate $PulA_{NA}$, the native signal peptide of lipoPulA was replaced with that of PelB (which directs the protein to the *E. coli* periplasmic membrane) and Cys1 was substituted by Met, in line with previously published work[192]. $PulA^{Kox}$ is organised into 5 domains, as shown in Figure 3.1A: N1 (residues 20-160), N2 (residues 161-265), N3 (residues 266-386), A (residues 387-953) and C (residues 954-1070). Residues 1-19 comprise the peptide tether, invisible in the crystal structure and likely to be disordered. Domains N1, N2 and N3 constitute the N-terminal section, with N1 and N3 known to bind carbohydrates[193]. The role of the immunoglobulin-like N2 domain is unknown. Domains A and C form the catalytic core. Domain A comprises a TIM-barrel (eight α-helices and eight parallel β-strands), which is well conserved among glycosyl hydrolases, as is the N3 domain. Domain C has a β-sandwich fold structure. Sequence analysis identified an extra "inserted" domain (Ins) in PulA, relative to related amylases, neopullulanases and pullulanases[194–197]. This domain, also of unknown function, is located between residues 475 and 545 in domain A (see Figure 3.1B) and is only present in pullulanases from Gram-negative bacteria including *Klebsiella, Aeromonas* and *Vibrio,* and not those from Gram-positive species, implying it may represent a secretion determinant[61]. The Ins subdomain is rich in helical and loop secondary

structure, and is presumably stabilised by the disulphide bond formed between Cys491 and Cys506 (see Figure 3.1C) and its two Ca^{2+} ions.

PulA from *K. oxytoca* shares 90% sequence identity with PulA from *Klebsiella pneumoniae*, which was crystallised by Bunzo Mikami and colleagues[193]. PulA from *K. pneumoniae* was crystallised from the secreted protein, whereas $PulA^{Kox}$ was purified from the periplasmic fraction of bacterial cells. Cα alignment of $PulA^{Kox}$ and $PulA^{Kpn}$ generates a Q-score of 0.84 (Q-score = 1 for identical structures); the structural similarity between the two suggests that PulA does not undergo changes in architecture upon secretion. The $PulA^{Kpn}$ crystal contains a protein of dimensions 102 x 65 x 71 Å, with a cavity (diameter 25 Å) surrounded by the N1, N2 and A domains, similar to the $PulA^{Kox}$ structure. Domains N2, N3, A and C are tightly bound by van der Waals interactions and hydrogen bonds, whereas the N1 domain lacks hydrogen bonds to any other domains, indicating its flexible nature. $PulA^{Kpn}$ was crystallised in complex with each of glucose, maltose, isomaltose, maltotriose and maltotetraose. Comparison of the apo- and maltotetraose-bound pullulanase static structures demonstrated an induced-fit motion of residue Trp708 and subsequently Glu706 as a result, leading to the catalytic residue and binding site residue adopting a conformation appropriate for catalysis[193]. However, more detailed understanding of the catalytic mechanism of the enzyme remains elusive.

Both $PulA^{Kox}$ and $PulA^{Kpn}$ were found to crystallise in dimers; however, it appears likely that PulA is secreted as a monomer, and carbohydrate binding sites have been shown to function as catalytic sites during enzymatic action, with no indication dimerisation is required[193]. Anchoring of the lipoprotein in the outer surface of the bacterial cell is likely to constrain PulA, potentially preventing dimerisation *in vivo*. Secretion of $PulA^{Kox}$ requires the T2SS composed of 15 components, 12 of which are essential for transport across the OM. $PulA^{Kox}$ was shown *via* chemical cross-linking analysis and bacterial two-hybrid assay (BACTH) to interact with PulG (pilin protein) and PulM (an IM platform protein of unknown function) (Dr. Olivera Francetic, personal correspondence). This indicates that these T2SS components come into close contact with pullulanase during secretion and play a role in the process.

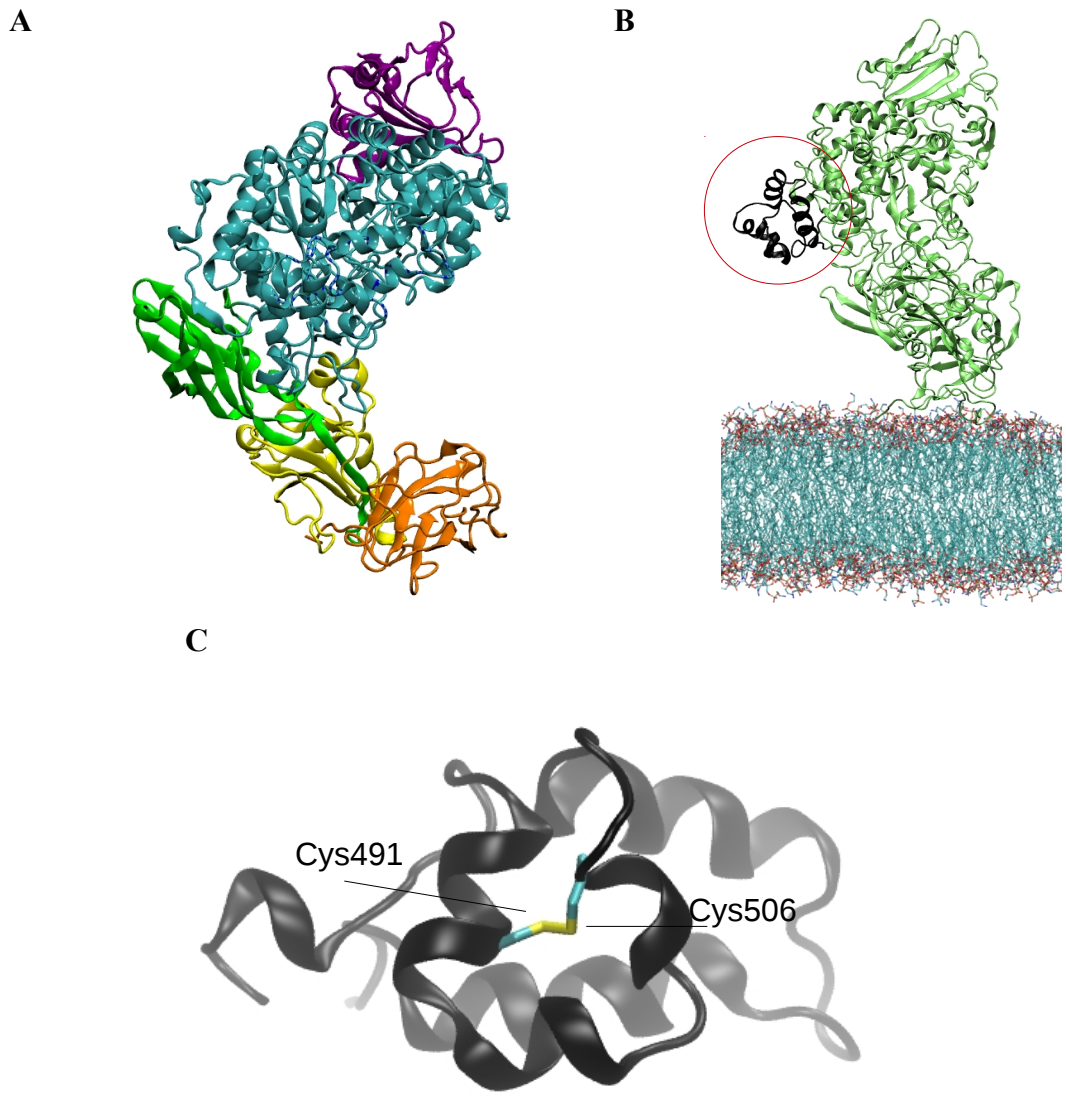


Figure 3.1 – PulA$_{NA}$ structure
(A) Modelled PulA$_{NA}$ structure, coloured according to domain (N1 – orange, N2 – yellow, N3 – green, A – cyan, C – purple, Ins – blue). Disordered tether not shown **(B)** A snapshot of an MD simulation system with PulA$_{NA}$ (green) placed on a POPE bilayer (coloured by atom; C – cyan, O – red, N – dark blue, P – gold) and the Ins subdomain shown in black/circled in red. **(C)** A snapshot of the Ins domain showing the three short helices forming a triangular structure, with the sulphur atoms forming a disulphide bridge between Cys491 and Cys506 shown in yellow. Two other α-helices are also present in the Ins domain; the bound calcium ions are omitted.

Initially, pullulanase is synthesised as a precursor with a Sec signal peptide (SP) containing a conserved lipobox sequence (consensus sequence: L^{-3}-A/S^{-2}-G/S^{-1}-C^{+1}; PulA sequence: L^{-3}-A^{-2}-G^{-1}-C^{+1}). The conserved Cys1 undergoes attachment of an *sn*-1,2-diacylglyceryl moiety from phosphatidylglycerol (PG) by membrane-embedded enzyme Lgt (phosphatidylglycerol:prolipoprotein diacylglyceryl transferase)[198].

Prolipoprotein signal peptidase LspA then cleaves the SP, releasing the N-terminal α-amino group of diacylglyceryl-cysteine for further N-acylation by apolipoprotein N-acyltransferase Lnt[198]. As a result of this maturation process (outlined in Figure 3.2), the lipid anchor is believed to remain embedded in the bacterial IM prior to secretion. Likewise *N*-acyl-*S*-diacylglyceryl pullulanase (lipoPulA) is not released directly into the extracellular environment following secretion, and remains attached to the bacterium.

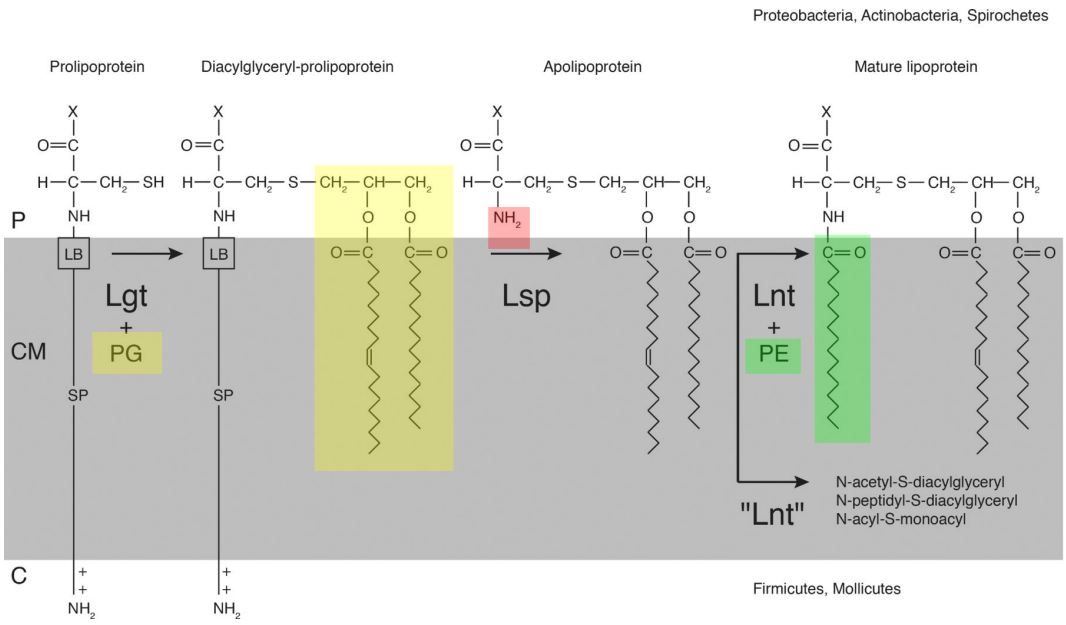


Figure 3.2 – Pathway of formation of lipoPulA from precursor
Adapted from Buddelmeijer, 2015[199]. Following transport from the cytoplasm (C) to the periplasm (P), prolipoproteins are inserted into the cytoplasmic membrane (CM) *via* recognition of the signal peptide (SP), which includes the lipobox sequence (LB). Lgt catalyses the transfer of a diacylglyceryl group (highlighted in yellow) from PG. Lsp cleaves the SP, freeing the α-amino group of diacylglycerol-Cys (pink). Lnt acylates this group using PE as a donor (green). Different modifications have been discovered in firmicutes and mollicutes.

As outlined in section 1.3, prior to secretion mature lipoproteins can either remain in the periplasmic IM leaflet or reach the OM *via* the Lol machinery (shown schematically in Figure 3.3). In the latter case, upon extraction from the IM by the LolCDE complex, lipoproteins bind to periplasmic chaperone LolA, which escorts them to the OM receptor LolB. In *E. coli*, and probably most other enterobacteria, all lipoproteins are sorted to the OM, except those with an Asp residue at position +2 (adjacent to the fatty-acylated N-terminal cysteine)[63]. Studies have identified

exceptions to this "+2 rule"[71] and shown that +3 position residues influence the strength of the IM retention[72,76]. Comprehensive *in vitro* analyses of these signals led to the current model of Lol avoidance by the strongest IM retention signal Asp2-Asp3, relying on charge and distance-specific interactions with the amine groups of phosphatidyl-ethanolamine[72]. However, the interactions of the +2 and +3 residues with the membrane bilayer have not been further characterised. Wild-type PulA^{Kox} contains Asp2 and Asn3.

In diderm bacteria, some of the lipoproteins that avoid the Lol system also eventually reach the cell surface, either through well-characterised protein secretion systems like the type II[88] or type V[200], or through mechanisms that are currently unclear as in *Borrelia burgdorferi*[201]. When produced in *E. coli* in the absence of the cognate T2SS, the Asp2 sorting signal leads to lipoPulA anchoring in the IM[61]. However, substituting Asp2 for Ser targets PulA to the OM and impairs its secretion[202], suggesting that IM localisation is required for the correct presentation of PulA to the T2SS. The molecular reason for this remains unknown, and MD simulations were considered an excellent strategy to attempt to explain the different interactions of the sorting signal with the lipid membrane.

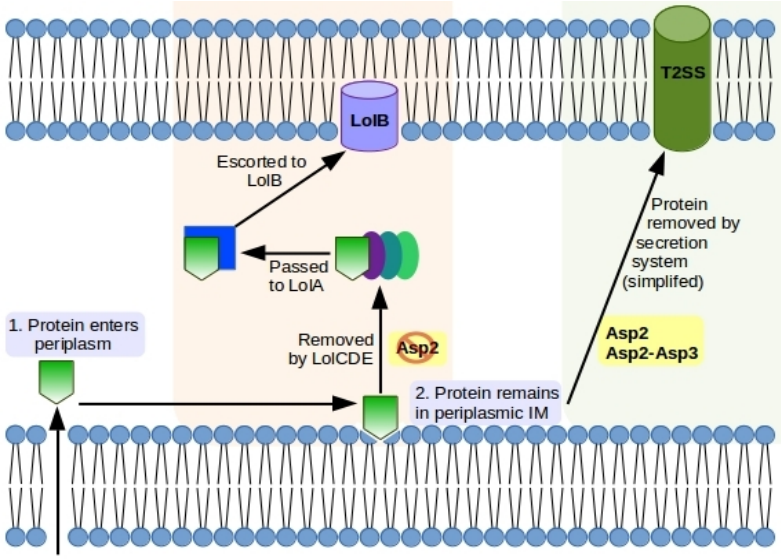


Figure 3.3 – The Lol removal pathway and effect of the Lol Avoidance Signal
Upon entering the periplasm, lipoproteins are either removed from the IM by the Lol pathway (in orange) or by a secretion system (green pathway). The Lol Avoidance Signal (LAS), Asp2, retains lipoproteins in the IM. If a protein lacks Asp at position 2, it enters the Lol pathway. Asp2 and Asp2-Asp3 form the strongest retention signals. PulA contains Asp2 and Asn3, and exits the cell *via* the T2SS.

PulA is secreted in the folded form and determinants on the protein surface may play a role in its specific recognition and transport across the OM[29,88]. The 430 N-terminal residues of PulA (consisting of the linker, domains N1-N3, and the first 43 residues of domain A) promoted secretion of reporter proteins[192], while previous deletion and gene fusion analyses specifically implicated two distinct regions of PulA primary sequence, A (residues 11-78, i.e. partial linker and first 58 residues of domain N1) and B (residues 735-814, found in the A domain), in T2SS-mediated secretion[203]. Likewise, peptide insertions in region C (residues 234-284, within the N2 domain) abolished secretion of $PulA_{NA}$. Discovery of the Ins domain suggested another region that also has functional relevance for secretion, as this segment was subsequently identified in pullulanases from other Gram negative genera with functional T2SSs[61]. Deletion of the Ins domain had virtually no effect on $PulA_{NA}$ secretion under over-expression conditions, yet abolished secretion at physiological production levels without affecting PulA stability[61]. However, these secretion determinants have not been characterised further.

Pullulanase is a model protein for the study of lipoprotein sorting and type 2 secretion, and several questions remained unresolved concerning both processes. In particular, what is the molecular nature of its secretion signal and does the lipid anchor play a role in the process? Does removal of the lipid anchor lead to pullulanase release from the membrane? Which regions of PulA are exposed and interact with the T2SS components prior to secretion? What is the atomic view of the LAS and its interactions with the bacterial membrane? Can structural determinants of secretion be identified, and what is the role of the newly identified Ins domain? In order to characterise the interactions of PulA and the LAS with the IM, I have modelled each of lipoPulA, $PulA_{NA}$ and a $PulA_{NA}$-D2S variant on a palmitoyl-oleoyl phosphatidylethanolamine (POPE) bilayer. Three 100 ns replicates of each have been carried out to characterise the dynamics and protein-lipid interactions of the systems.

3.2 Methods

The full-length model of lipoPulA was built by incorporating the 19 residues of the flexible N-terminal tether that was disordered, and hence missing, in the

crystallographic density. The tether was built and subsequently refined using the DOPE module in the Modeller suite[204]. Three fatty-acyl chains derived from palmitate to post-translationally modify the Cys1 residue were built as an N-α-palmitoyl-S-[2,3-bis(palmitoyloxy)-(2RS)-propyl]-L-cysteine moiety (CYP) (Figure 3.4) using PyMol (The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.). The final protein structure was composed of 1070 amino acids, including CYP, and all ionisable groups were assigned to their most probable charged states at neutral pH. LipoPulA was then manually positioned above the surface of a pre-equilibrated POPE membrane, and the CYP acyl chains were partially inserted into the bilayer. The system was placed in a triclinic box with dimensions ~ 100 x 100 x 180 Å. The system was explicitly solvated with the TIP3P water model, *via* superimposition of a pre-equilibrated box of waters, and electrically neutralised by replacing random water molecules with sodium chloride counter-ions to ~ 0.1 M concentration. Overlapping solvent and two POPE lipids were removed, resulting in a system containing ~ 34,000 water molecules and 318 lipids. Set up of the first lipoPulA system and parameterisation of CYP were performed by Dr. Peter Bond.

Figure 3.4 – Structure of the acyl anchor
Representation of the palmitoylated Cys (CYP) residue 1 of lipoPulA. The segment is coloured by atom (C – cyan, O – red, S – yellow, N – blue, H – transparent white), and there are 3 extended tail group chains visible, two in the foreground and one curving behind, which anchor into the lipid bilayer.

For the $PulA_{NA}$ systems, PyMOL was used to replace CYP with a methionine (Met) residue to produce the non-acylated structure, and to replace Asp2 with a serine (Ser)

residue to form the D2S variant. $PulA_{NA}$ or $PulA_{NA}$-D2S proteins were placed above the 318 lipid bilayer described previously, with the placing of each system in a box, solvation and neutralisation steps performed as above. Prior to and following solvation of each system, successive energy minimisations were performed using an SD algorithm, to minimise steric overlap between the components of each system. Subsequent equilibration simulations allowed the solvent to relax around each restrained biomolecular complex. Harmonic restraints were applied to all the protein non-hydrogen atoms and gradually removed in three distinct 0.5 ns phases. Finally, production MD was executed in triplicate on the unrestrained systems for 100 ns, within the NPT ensemble.

All PulA simulations were performed using the GROMACS 4.5.5 simulation package[185], with the CHARMM36 all-atom FF[175] used to treat the lipid molecules and CHARMM22/CMAP parameters[175] to treat the protein. Simulations were completed using the leap-frog algorithm with a 2 fs time-step, and trajectory data were collected every 10 ps. The LINCS algorithm[181] was used to constrain bond lengths and the neighbour list was updated every 10 steps. The long-range electrostatic interactions were corrected using the PME method[184]; the real-space electrostatic and van der Waals interactions were cut off at 12 and 10 Å, respectively. Simulations were performed at an absolute temperature of 310 K using the V-rescale thermostat. The pressure was set to 1 bar (Parrinello-Rahman semi-isotropic barostat, 5 ps coupling constant), under PBC.

Table 3.1 contains the nomenclature of the structures and simulations analysed in this chapter, for the clarity of the reader.

Table 3.1 – Systems analysed in this chapter

Name	System	POPE	Counter-ions	Simulations
LipoPulA	POPE + crystal structure of PulA (2YOC) with modelled tether (residues 1-19) and acyl anchor (CYP)	YES	Na^+Cl^-	3 x 100 ns
$PulA_{NA}$	POPE + crystal structure of PulA (2YOC) with modelled tether (residues 2-19) and with CYP replaced by Met	YES	Na^+Cl^-	3 x 100 ns
$PulA_{NA}$-D2S	POPE + crystal structure of PulA (2YOC) with modelled tether (residues 2-19), CYP replaced by Met, Asp2 replaced by Ser	YES	Na^+Cl^-	3 x 100 ns

Revised FFs have become available more recently, however the FF combination used in this study was chosen to maintain consistency with previously published work on the Pul secretory system using these two FFs. The CHARMM36 parameters for proteins and for lipids were not developed concomitantly and are not formally equivalent. Except for adjustments to the CHARMM36 lipid parameter set allowing simulations to run in the tensionless ensemble to yield the correct head group surface area for various phospholipid bilayers, CHARMM36 lipids are fully consistent with the additive CHARMM22/CMAP protein parameters[175]. However, the CHARMM36 protein parameters were developed mostly to correct an equilibrium imbalance in the sampling of sheet versus helical conformations, which tends to only be relevant for shorter peptides or on folding time scales. Calculating the total amount of α-helix and β-sheet secondary structure in each system demonstrated no significant loss of structure in any simulation (see Table 3.2) and validated use of CHARMM22/CMAP. Compared to the crystal structure, there was an average increase of < 1.8 residues of α-helicity across the systems, indicating that they are all well described by the FF and not unduly biased by any α/β imbalance.

Table 3.2 – Amount of Secondary Structure per Simulation

Name	Replicate	No. of α-helical residues*	Change in α-helix**	No. of β-sheet residues*	Change in β-sheet**
LipoPulA	**I**	220	+ 1.6	215	+ 6.8
	II	235	+ 16.8	216	+ 8.0
	III	214	- 3.8	218	+ 10.0
$PulA_{NA}$	**I**	221	- 1.4	213	+ 4.8
	II	221	- 1.2	213	+ 5.4
	III	215	- 7.5	218	+ 10.0
$PulA_{NA}$-D2S	**I**	227	- 4.4	220	+ 9.5
	II	221	+ 9.6	214	+ 4.0
	III	224	+ 6.5	215	+ 5.0

*** Refers to average number of residues engaging in the respective secondary structure over the final 20 ns of the simulation.**

**** "Change" in secondary structure refers to difference between the number of residues holding a given structure (A) during the first frame, and (B) averaged over the final 20 ns.**

3.3 Results

3.3.1. Molecular dynamics simulations of lipoPulA

Initially I sought to gain insight into the intermediate state of lipoPulA, anchored in the periplasmic leaflet of the IM and important for the correct presentation of the secretion signal(s) on the protein surface to the T2SS. To this end, I performed MD simulations of lipoPulA on a POPE model membrane. As described in section 3.2 above, the protein was initially placed near the membrane with the fatty acyl tail (shown as spheres) partially inserted into the bilayer (Figure 3.5A). Following equilibration and triplicate 100 ns production simulation, the fatty acids inserted fully into the bilayer, pulling the N-terminal tether onto the membrane surface (Figure 3.5B). The CYP lipid tails and glycerol groups became more deeply buried than the equivalent groups of the membrane lipids, and the CYP amide group was observed at a position similar to that of the POPE ethanolamine groups.

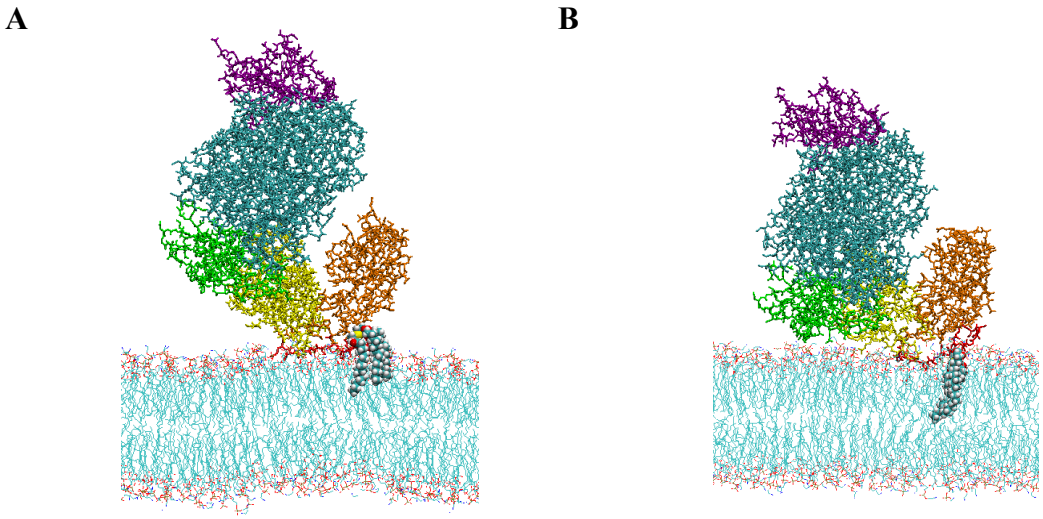


Figure 3.5 – MD simulations of lipoPulA anchored in the POPE model membrane
The lipoPulA conformation at the **(A)** start and **(B)** end of a representative 100 ns simulation. The N-terminal 19-residue tether is shown in red, and the fatty acyl tail as spheres. PulA domains are shown in different colours: N1 is in orange, N2 in yellow, N3 in green, A in cyan and C in purple. The POPE membrane atoms are shown in light blue (carbon), red (oxygen) and dark blue (nitrogen).

The two monomers observed in the PulA crystal structure demonstrated a 47° rotation between the N1 domains, notably replicated during one simulation as domain N1 pivoted around the rest of the protein – clearly demonstrated *via* superposition of all lipoPulA domains except N1 and the tether (Figure 3.6A). Crystal subunit A was used as the initial structure; by the end of the simulation the N1 domain moved to occupy a very similar position to crystal subunit B (Figure 3.6B). The position of the rest of the protein adjusted relative to the N1 domain in a manner consistent with the X-ray structure, so that the N2/N3 domains also approached the membrane surface. This is illustrated in Figure 3.6C by changes in the location of each domain centre-of-mass along the *z*-axis (the normal to the membrane plane) relative to the membrane centre-of-mass, during each replicate. After 60 ns of simulation, the N2 and N3 domains moved closer to the membrane surface than N1, with the other domains (A/C/Ins) following.

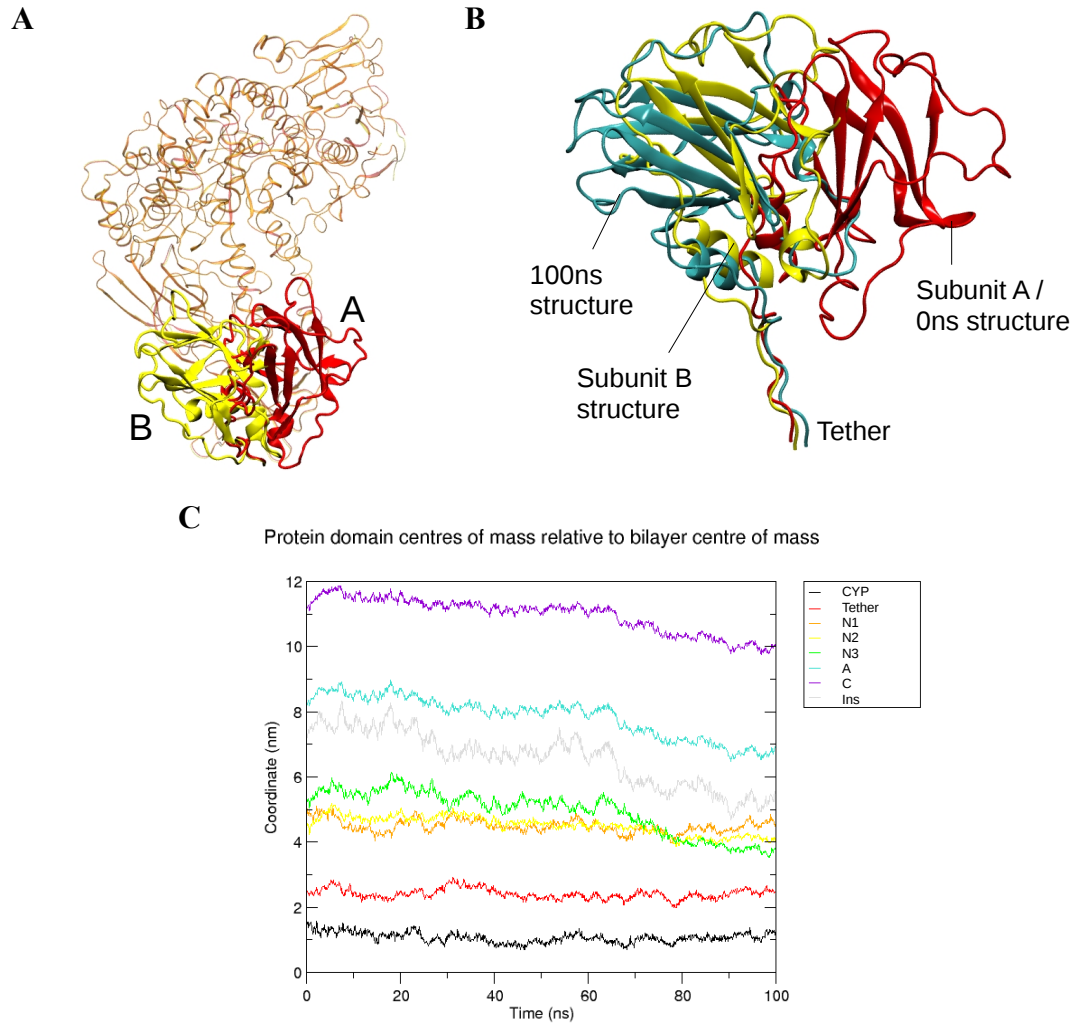


Figure 3.6 – LipoPulA conformational changes
(A) Superposition of the crystal monomers (subunit A – red, subunit B – yellow) highlights the observed conformational differences. Dimer asymmetry arises from the significant difference between the relative orientations of the N1 domains (shown in ribbon form) and rigid cores (including the N2, N3, A and C domains) of each monomer. **(B)** All domains except N1 and the tether were superimposed. Domain N1 (shown) pivots around the rest of the protein (not shown) during the simulations. Subunit A (red) was used as the initial structure and the N1 domain moved to occupy a similar position to subunit B (yellow) by the end of the simulation (cyan). **(C)** The distance between the centre of mass of different PulA domains and the membrane during a representative simulation. The lines, coloured according to domain as indicated on the right, represent movements of individual lipoPulA domains relative to the membrane centre of mass.

The number of hydrogen bonds and atomic contacts between the protein domains and POPE lipids remained fairly constant and high for the N-terminal tether and N1 domain throughout. The number of contacts for the N2 and N3 domains showed a

large increase over the final ~ 20 ns in one simulation in particular, as domain N3 came into contact with the membrane and relaxed onto its surface (Figure 3.7). The N3 domain did not interact with the lipids before 70 ns in any simulation, at which point hydrogen bonds were formed between the protein and POPE as their proximity increased.

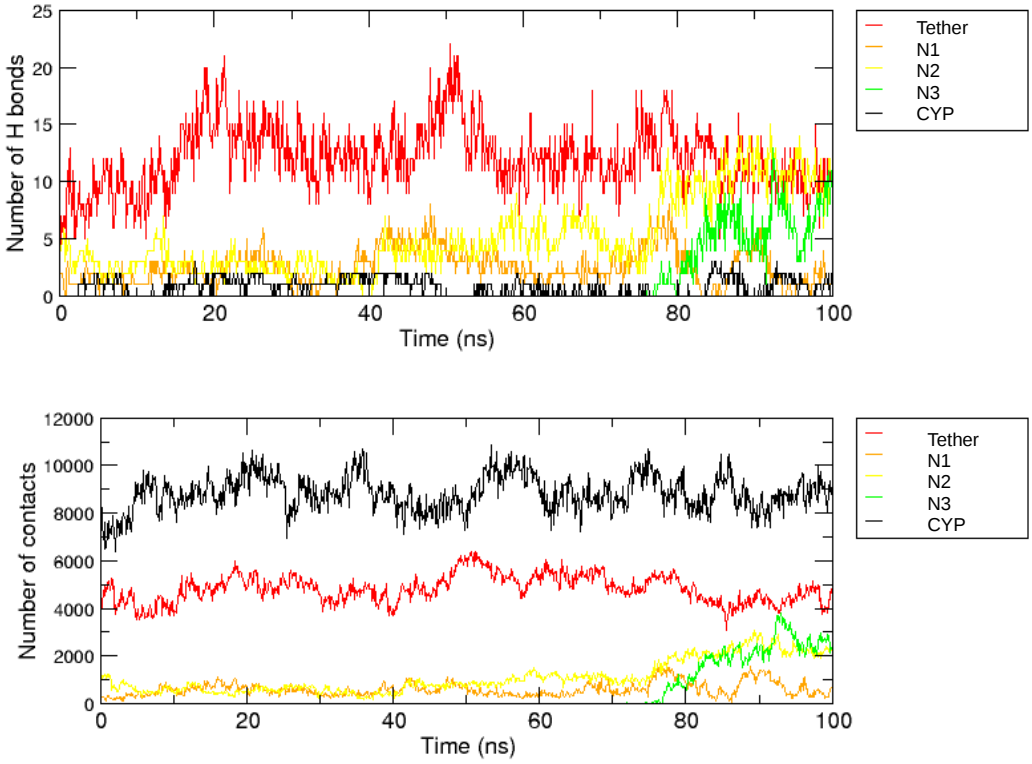


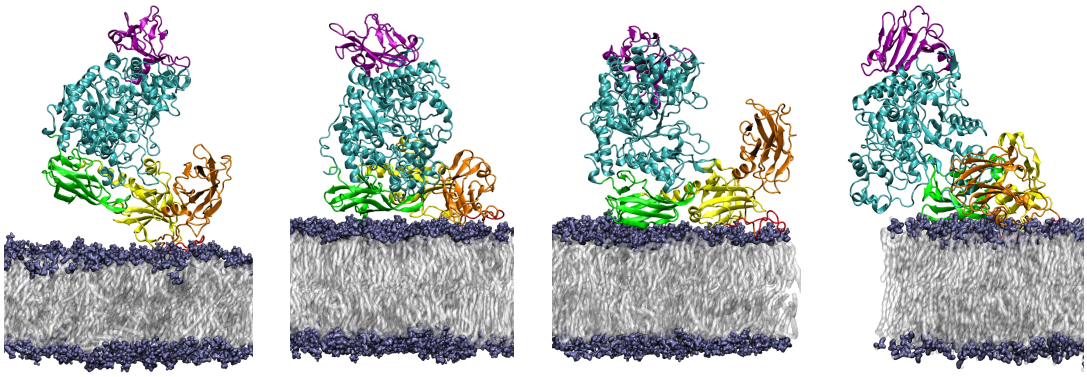
Figure 3.7 – Hydrogen bonding and atomic contacts of lipoPulA with the membrane
Main chain hydrogen bonds (top panel) and atomic contacts closer than 6 Å (bottom panel) of the tether and domains N1, N2 and N3 with the POPE membrane during a representative lipoPulA simulation. The graph colour codes are indicated in the legends.

3.3.2. Molecular dynamics simulations of $PulA_{NA}$

The requirement of IM binding for lipoPulA secretion and the observation that several lipoPulA domains interact with the POPE membrane in MD simulations suggested that $PulA_{NA}$ could also associate with the membrane. The ability of purified Pul_{NA} to bind *E. coli* and other liposomes, unaffected by Ca^{2+} in the buffer solution or liposome phospholipid variation, was also confirmed experimentally[61]. I

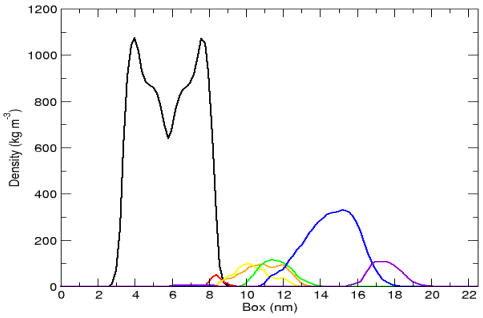
therefore performed three independent 100 ns simulations of $PulA_{NA}$ near the POPE bilayer to test this hypothesis, using the crystal structure of full-length PulA with an N-terminal Met replacing CYP. Overall the results of these simulations revealed that $PulA_{NA}$ adopted highly similar orientations to lipoPulA, relative to the membrane, and comparable protein-membrane interactions occurred as a result. Initially, the protein was placed near the membrane in the same orientation as the lipoPulA systems (Figure 3.8A), but with no part of the protein embedded in the lipid, due to the absence of the acyl anchor. Surprisingly, $PulA_{NA}$ did not diffuse away and instead relaxed onto the membrane surface by the end of the simulations (Figure 3.8B). The density changes of the system components during the simulation as a function of the *z*-axis (Figure 3.8C, left panel), and the movements of the centre of mass of each domain (Figure 3.8C, right panel), as well as the increasing number of hydrogen bonds formed between membrane and protein (Figure 3.8D), demonstrated the large extent to which $PulA_{NA}$ approached the membrane. The $PulA_{NA}$ membrane interactions occurred primarily *via* the tether but also *via* the N2 and N3 domains, as observed for lipoPulA. However, additional residues in the tether and N2 domain interacted with the membrane compared to lipoPulA. These differences might be due to the constraints imposed on these regions by the membrane-bound N-terminus of lipoPulA.

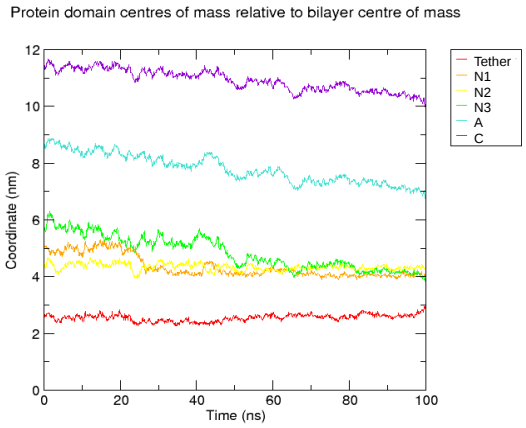
A **B** **I** **II** **III**



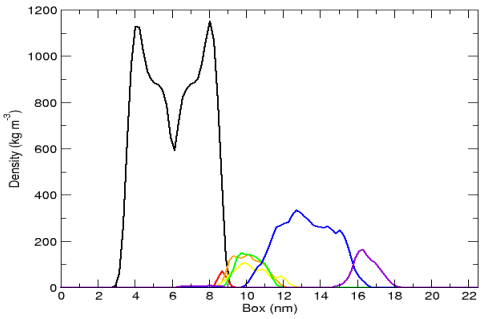
C

Initial frame – 0 ns





Final frame – 100 ns



D

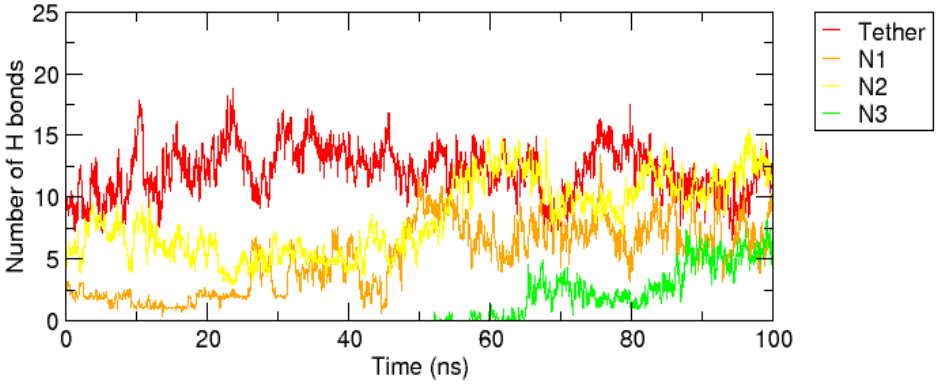


Figure 3.8 – Behaviour of $PulA_{NA}$ protein through simulations
(A) The initial structure, with $PulA_{NA}$ placed near the membrane. The protein is coloured by domain, as shown previously, with the lipid head groups are shown in lilac and the lipid tails in white. **(B)** The final conformations following three 100 ns production runs. The protein forms an increased number of contacts with the lipid bilayer, and adopts a similar conformation to lipoPulA (refer to Figure 3.6B). **(C)** Left: The changes in density of the system components between the initial (top) and final (bottom) frames of the simulations, shown as a function of the *z*-axis. The protein is observed approaching the membrane (shifts to the left across the graph). Right: The centres of mass of the protein domains change between the initial and final frames of the simulations relative to POPE, also demonstrating the $PulA_{NA}$ movements. **(D)** Graph of the hydrogen bonding between POPE and each of the tether, N1, N2 and N3 domains of $PulA_{NA}$ in a representative simulation.

3.3.3 Conformational dynamics of PulA

To study more precisely the conformational changes that occur during membrane binding, I quantified the conformational dynamics of the two PulA forms by calculating the RMSD of the Cα atoms in PulA with respect to the crystal structure of subunit A, which represented the initial form. As anticipated, the unstructured lipoPulA tether (residues 2-19) repeatedly exhibited the greatest structural variation among all the individual domains throughout the simulations (Figure 3.9A, red line). The structural drift of all other parts of the lipoprotein plateaued after ~ 20 ns, showing that the simulations are relatively stable. Interestingly, over some parts of each simulation, the RMSD of the ensemble formed by the N1, N2, and N3 domains was significantly higher than any of the individual N1/N2/N3 RMSDs, and this was particularly evident in replicate I, where the RMSD reached almost 1 nm. These values indicated the motion between these three domains mentioned previously, with N1 swivelling away from N2 and N3.

In the $PulA_{NA}$ simulations, the tether and N1 domain exhibited the greatest structural drift, and the internal motion between N1, N2 and N3 was again evident in replicate I. After 25 ns of simulation the RMSD increased abruptly to over 0.7 nm, quantifying the observed movement of the domains to mimic the conformational variation observed in the crystal. Visual analysis of this particular trajectory showed the marked motion of N1 relative to the rest of the protein. This is illustrated in Figure 3.10, where the protein was fitted to the rigid core formed by domains N2, N3, A and

C, and the N1 domain pivoted around the rest of the protein, in a manner consistent with the conformational differences captured in the crystallographic dimer. It appears that this $PulA_{NA}$ simulation, along with the lipoPulA simulation showing the same movement (see Figure 3.6B), has visualised an event whereby the protein moves as predicted by the crystal structures, with a motion that might therefore have functional significance.

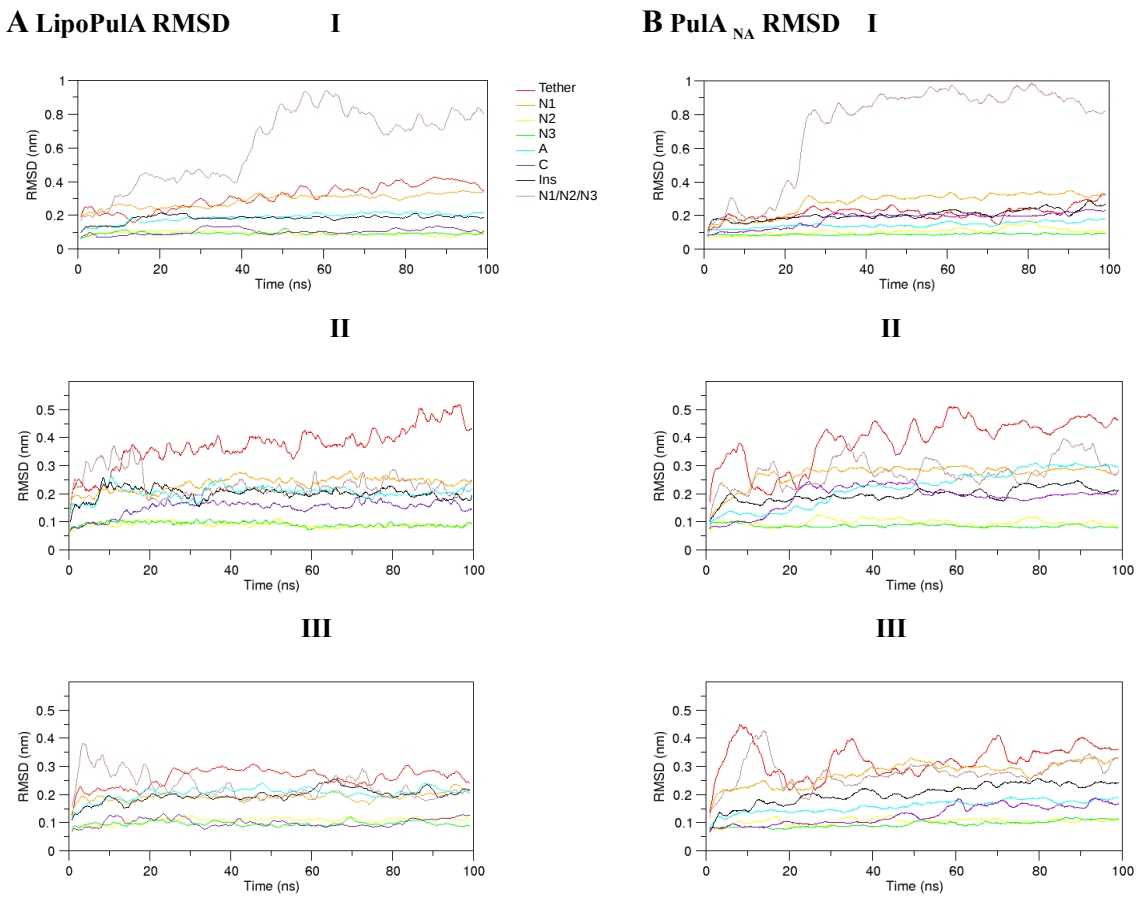


Figure 3.9 – PulA protein RMSD throughout lipoPulA and $PulA_{NA}$ simulations
Graphs showing the root mean square deviation of each **(A)** lipoPulA and **(B)** $PulA_{NA}$ domain (the colours are listed in the legend) from the initial crystal structure over the course of each 100 ns simulation. Note different scales for replicate I of each system, due to large domain movements.

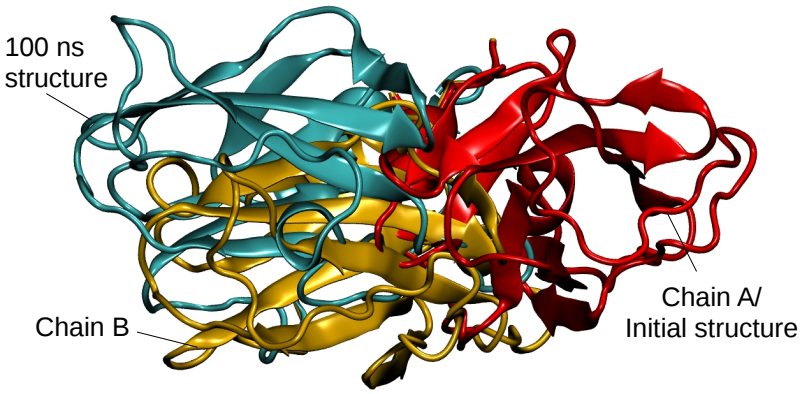


Figure 3.10 – Motion of $PulA_{NA}$ N1 domain
All domains except N1 and the tether were superimposed. Domain N1 (shown) pivots around the rest of the protein (not shown) during the simulations. Subunit A (red) was used as the initial structure and the N1 domain moved to occupy a similar position to subunit B (yellow) by the end of the simulation (cyan).

3.3.4 – Interactions of the Lol Avoidance Signal with the Inner Membrane

Deletion analysis was used to test the roles of the N-terminal tether and domain N1 in PulA secretion[61]. Deleting the distal part of the tether (residues 12-20) had little effect on PulA secretion or stability. However, variant $PulA_{\Delta 4\text{-}18}$ showed a partial secretion defect, which was accentuated under near-physiological conditions. Deleting the entire tether and the adjacent unstructured region of domain N1 in $PulA_{\Delta 2\text{-}34}$ resulted in complete PulA degradation under both conditions. These results suggested that the N-terminal IM-binding region, unstructured in solution, is essential for $PulA_{NA}$ stability.

Consequently, following the overview of PulA dynamics and membrane interactions, analysis focused on the interactions of the LAS with water and POPE. I examined whether the presence of CYP increases Asp2/Asn3 anchoring to the bilayer, and the interactions of the LAS in the absence of CYP. I also performed 3 x 100 ns simulations of a non-acylated PulA D2S variant ($PulA_{NA}$-D2S), placed on POPE in a manner analogous to $PulA_{NA}$. $PulA_{NA}$-D2S also remained close to POPE and did not diffuse away from the membrane. This was expected since $PulA_{NA}$ did not diffuse away and also the LolCDE complex, which removes proteins lacking the LAS from the bilayer, was not included in the simulations. Even still, this is consistent with the

fact that other residues in the tether also contribute to membrane binding. The variant demonstrated average overall protein RMSD (measured between the first and final frames) statistically similar to the other systems (lipoPulA: 4.3 ±3.1 Å; $PulA_{NA}$: 5.0 ±2.5 Å; $PulA_{NA}$-D2S: 4.0 ±0.4 Å). In one simulation $PulA_{NA}$-D2S exhibited similar dynamics to lipoPulA and $PulA_{NA}$, with the protein approaching the bilayer (Figure 3.11A) and the number of hydrogen bonds between the tether, N1 and N2 domains and POPE increasing over the course of each trajectory. However, in two simulations the protein remained positioned above the POPE surface, with minimal contacts (Figure 3.11B), indicating that the substitution of a single residue may indeed have an effect on the overall protein conformation.

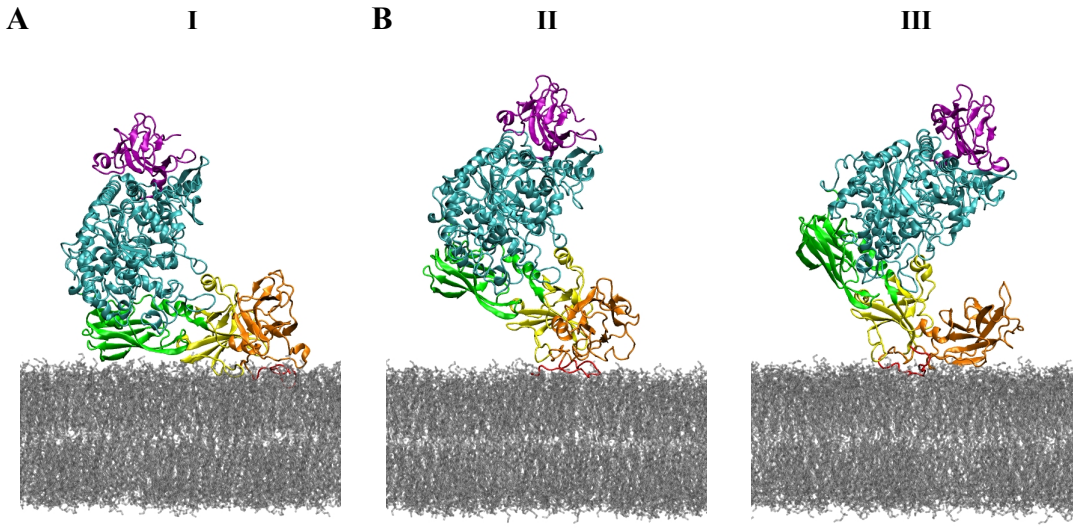


Figure 3.11 – Snapshots of protein conformations during $PulA_{NA}$-D2S simulations
(A) The final conformation of $PulA_{NA}$-D2S replicate I, following a 100 ns production run. The protein reclines on the bilayer surface and adopts a similar conformation to lipoPulA and $PulA_{NA}$ (refer to Figures 3.6B and 3.8B respectively). **(B)** The final conformation of $PulA_{NA}$-D2S replicates II and III. The protein remained above the bilayer and no contacts were found between lipid and the N3 domain, with fewer between the N1/N2 and POPE than in the other systems.

The MD simulations presented here provide detailed molecular insight into a lipoprotein tether region and LAS, typically disordered and invisible in lipoprotein crystal structures. Comparison of tether contacts with the IM across the systems showed how they evolve between the beginning and the end of MD simulations (Figure 3.12). Calculation of the number of contacts within 3 Å between POPE and

each of PulA residues 2, 3 and 4 showed fluctuations among all the simulations, indicating that substituting Asp2 to Ser does not lead to very marked differences in atomic LAS-POPE interactions. This suggests that the D2S substitution may have a different effect in lipoPulA, and the PulA membrane binding energy might be lowered sufficiently to allow protein extraction by LolCDE. These simulations would be a control for such a future exploration and they show that the tether promotes initial membrane binding, with the protein subsequently adopting a membrane-proximal position.

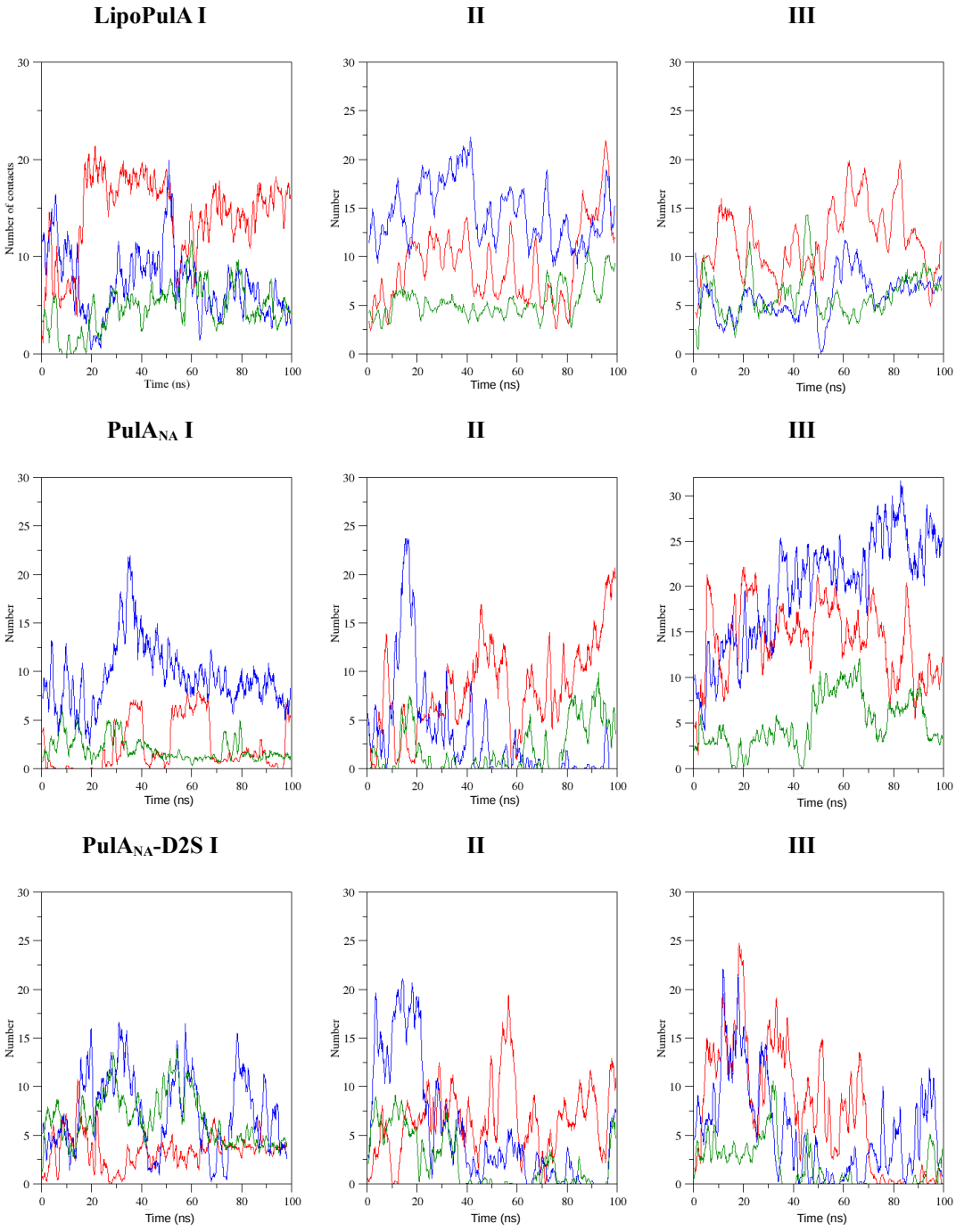


Figure 3.12 – Number of contacts between PulA residues Asp/Ser2, Asn3, Gly4 and POPE
Plots showing the number of contacts closer than 3 Å between residues 2 (red), 3 (blue), 4 (green) and POPE. Fluctuations are observed among all the simulations, indicating that substituting Asp2 to Ser does not lead to very marked differences in atomic LAS-POPE interactions and suggesting that the substitution may have a different effect in lipoPulA, such as affecting the energetics of protein removal from POPE.

Examining the number of LAS contacts within 4 Å of POPE, differentiated by

residue main chain and side-chain, produced a more nuanced analysis. CYP formed a consistent and large number of contacts across all three lipoPulA replicates, whilst Met1 contacts with POPE fluctuated more among the $PulA_{NA}$-D2S replicates than the $PulA_{NA}$. In two replicates the $PulA_{NA}$-D2S Met 1 side-chain consistently engaged in over 50 close contacts with POPE, unlike the $PulA_{NA}$ Met1 which averaged ~ 30 such contacts. Notably, both the main- and side-chain of Asp2 in the lipoPulA system formed contacts with the lipid at all times in all replicates. This was in stark contrast to the $PulA_{NA}$-D2S system, in which S2 in all three replicates was more than 4 Å from the bilayer during one or more simulation frames, and the $PulA_{NA}$ system, in which the number of contacts between D2 and POPE varied widely from 0-90 contacts. Asn3 contact with POPE was not affected significantly by the residue 2 substitution or the absence of CYP.

In a representative frame at the end of one lipoPulA replicate, residues Asp2, Asn3 and Gly4 (comprising the IM retention signal), and the CYP acyl chains fostered polar contacts with amine and phosphate groups of seven different POPE molecules. The backbone nitrogen of CYP formed a hydrogen bond to a POPE molecule. D2 interacted with the amine groups of three POPE molecules *via* two side-chain, and the main chain, oxygen atoms, and formed hydrogen bonds with phosphate groups of a different POPE molecule. The Asn3 side-chain and the Gly4 backbone oxygen contacted one of these POPE molecules. Oxygen atoms of the three fatty-acyl chains made polar contacts with two additional POPE molecules, reinforcing the network and bringing the number of POPE molecules in the complex to seven.

The details of the $PulA_{NA}$ membrane contacts closely resembled those of lipoPulA, with the notable exception that fewer POPE molecules (maximum 5) interacted with residues 1-4; 11 POPE molecules contacted the IM retention signal in the lipoPulA systems. Similarly, in the final structures of the $PulA_{NA}$-D2S trajectories there were between 2 and 5 POPE molecules present within 5 Å of protein residues 1-4. This is probably due to the lack of embedded acyl tails pulling the tether into more extensive contacts with the membrane and the significantly smaller size of residue 1 in systems lacking CYP may also contribute, as there is a smaller surface area available for contacts.

Comparison of the solvation of the first four residues across the systems (Table 3.3) showed that residues 2 and 3 were generally the most solvated, and Gly4 the least, as the glycine side-chain cannot form hydrogen bonds. Ser2 was less solvated than Asp2, which may be expected as charged Asp is significantly more hydrophilic than polar neutral Ser at physiological pH. However, there was no significant difference in hydrogen bonding once standard deviations were taken into account (Figure 3.13), suggesting that the role of solvation is minor in this set of interactions.

Table 3.3 – Number of hydrogen bonds between the LAS and solvent

		Residue 1 (CYP/Met)	Residue 2 (Asp/Ser)	Asn3	Gly4
LipoPulA	**I**	3.0 ± 1.4	5.2 ± 1.5	2.9 ± 1.2	0.5 ± 0.6
	II	1.5 ± 0.9	5.1 ± 1.3	2.1 ± 1.1	1.1 ± 0.5
	III	1.5 ± 0.8	3.3 ± 1.0	4.2 ± 1.1	1.2 ± 0.8
PulA$_{NA}$	**I**	1.6 ± 0.7	6.1 ± 1.3	3.9 ± 1.2	1.5 ± 0.7
	II	2.3 ± 0.9	3.8 ± 1.0	4.8 ± 1.2	0.5 ± 0.7
	III	2.8 ± 0.9	5.8 ± 1.6	1.6 ± 0.8	0.5 ± 0.5
PulA$_{NA}$-D2S	**I**	0.5 ± 0.6	1.9 ± 1.1	3.9 ± 1.2	0.8 ± 0.4
	II	2.0 ± 1.0	2.4 ± 1.0	3.6 ± 1.2	1.0 ± 0.6
	III	3.2 ± 1.1	2.9 ± 0.9	4.2 ± 1.4	1.3 ± 0.9

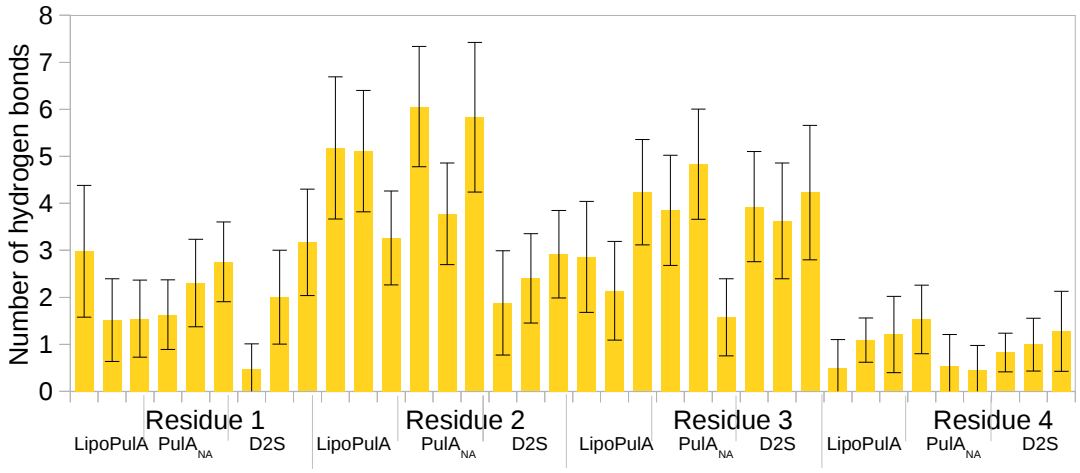


Figure 3.13 – Solvation of the Lol Avoidance Signal
Comparison of the solvation of the first four residues across the systems averaged over the full simulation time showed that residues 2 and 3 were generally the most solvated, and Gly4 formed the fewest hydrogen bonds to water. Ser2 was less solvated than Asp2, however, there was no significant difference in hydrogen bonding once standard deviations were taken into account. Error bars show standard deviation over the full simulation time.

3.3.5 – Dynamics of the Ins domain

Deletion of the Ins domain had virtually no effect on Pul_{NA} secretion under overexpression conditions, however, secretion of $Pul_{NA-\Delta INS}$ was abolished at physiological production levels, while its stability was not affected[61]. This evidence that the Ins domain may be a secretion determinant led me to compute the root mean square fluctuation (RMSF) for the protein, to observe whether the domain displayed unusual behaviour with potential functional relevance. I calculated the Cα RMSF of each residue for the final 10 ns of each trajectory, and converted this to a B factor value, providing an indication of the mean motion of each Cα atom around its average position (Figure 3.14). In agreement with the B factors derived from the X-ray data, the greatest fluctuations in the protein backbones of all three systems localised primarily around domain N1, followed by the Ins domain, which underwent relatively large fluctuations (B factors of up to 295 $Å^2$). These results indicate that the Ins domain, located on the side of the protein (Figure 3.1B) is surprisingly dynamic and may function as a secretion determinant *via* interactions with, putatively, PulG or PulM – or other proteins involved in the T2SS. Even with the presence of a disulphide bond, the domain is flexible and experiences motion. However, these simulations did not include two known bound calcium cations; divalent cations are known to rigidify proteins and therefore the next step would be to extend this study to include the ions.

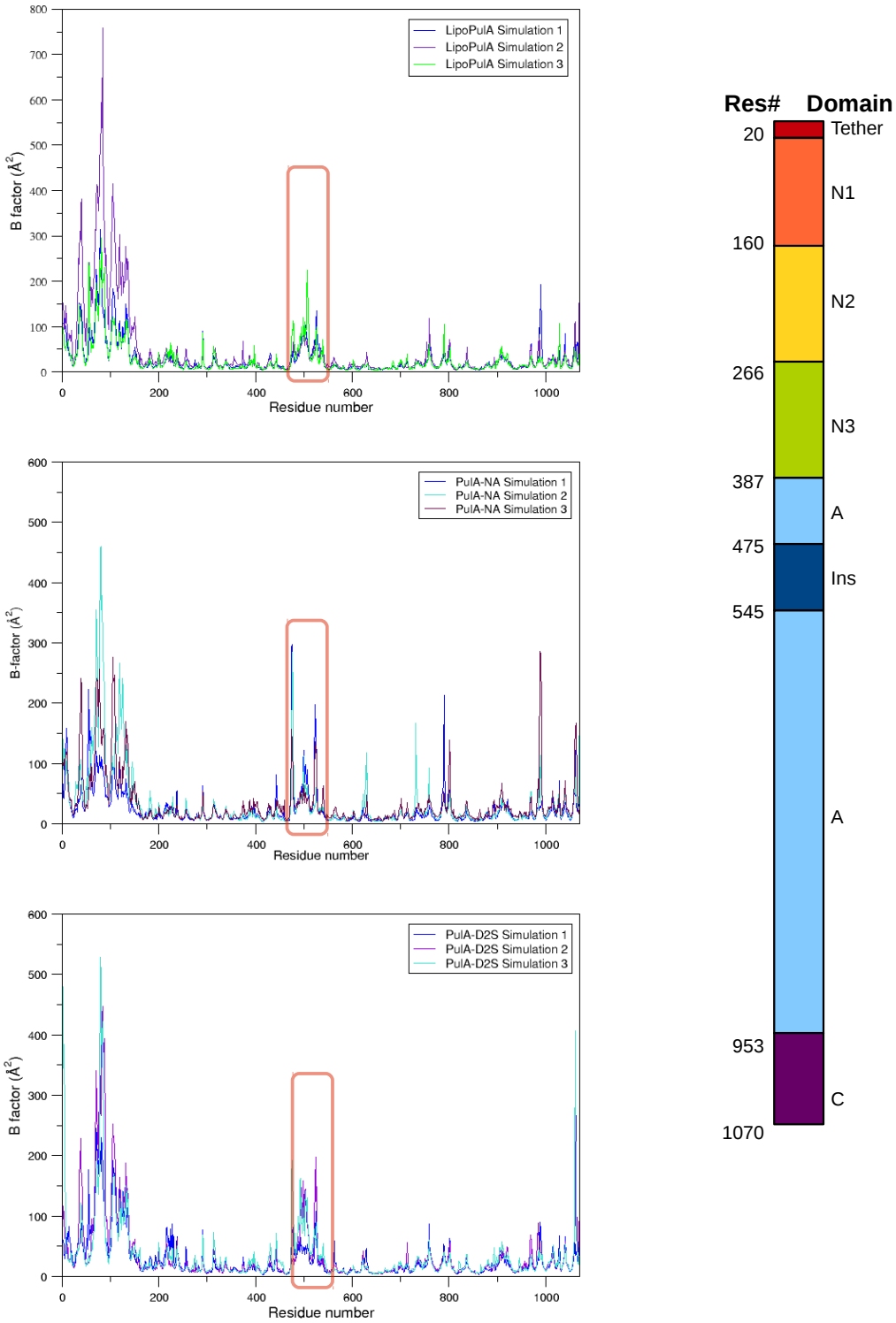


Figure 3.14 – RMSF of protein Cα atoms during final 10 ns of lipoPulA, Pul_{NA} and $PulA_{NA}$-D2S simulations

Comparison of the mean motion of each Cα atom during the final 10 ns of each simulation demonstrates that the N1 and Ins domains (highlighted in red) consistently demonstrate the most fluctuation. In the lipoPulA simulations, the fluctuations peak between residues 470-510 and 525-540. In both the Pul_{NA} and $PulA_{NA}$-D2S simulations the peaks occur between residues 470-480, 490-510 and 520-530.

3.4 Discussion

There were three primary questions that this MD simulation research, in conjunction with experimental collaborators, aimed to answer. Firstly, the regions of PulA that interact with the IM were examined. Deletion and gene fusion analyses had implicated the 430 N-terminal residues of PulA, as well as regions A, B and C (residues 11-78 in the N1 domain, 735-814 in the A domain and 234-284 in the N2/N3 domains, respectively) in T2SS-mediated secretion. MD simulations allowed visualisation and rationalisation of the protein-lipid interactions. In all the non-substituted simulations containing a lipid bilayer in an environment with sodium chloride, regardless of the presence of a Lol retention signal or acyl anchor, PulA approaches the membrane, and segments of the N1, N2 and N3 domains contact the lipid. Notably, rotation of the N1 domain was observed in both systems, with the observed conformations fully reproducing the difference between the two asymmetric units of the crystal. Drawing on analyses by Mikami and colleagues regarding the location of sugar binding sites and the pullulanase catalytic sites, this motion may have functional relevance, possibly moving the protein into position for repeated catalytic actions (as the pullulan polymer is degraded). The increasing hydrogen bonding and atomic contacts between domains N1/2/3 and POPE suggest a mechanism to protect the protein from proteolysis in the periplasm, in preparation for presentation of the substrate to the T2SS. The MD results corroborate the biochemical data showing that the 430 N-terminal amino acids (the tether, and domains N1-3 and the first 43 amino acids of domain A) are necessary for secretion and promote low levels of secretion when overproduced, and provide evidence of the behaviour of PulA in the periplasm, suggesting that the acyl anchor is not the sole area of interest. In addition, the MD simulations provided a molecular view of the tether region and guided mutational analysis, which showed that this segment plays a key role in $PulA_{NA}$ stability *in vivo*. MD simulations revealed the likely extended structure of the PulA tether, which established van der Waals interactions with the membrane surface *via* proline residues in addition to the polar side-chain contacts with the phospholipids head-groups.

Secondly, a molecular view of the Lol Avoidance Signal was sought. The Lol

Avoidance model based on *in vitro* analyses of lipoprotein release from the IM proposes a network of salt bridge and electrostatic interactions between the Asp2 and the amine groups of POPE lipids. The network theory explains the tight binding of model lipoproteins to the membrane, and the inability of Lol machinery to extract them and catalyse their targeting to the OM. MD provided an insight into this network, which proved to be more extensive than predicted, and involved phosphate in addition to amide groups of POPE as well as protein residues 3 and 4. This computational approach allows prediction and rationalisation of the behaviour of different LAS signals and to explain their relative strength, taking into account the geometry and flexibility of various tether sequences and particular phospholipid species present in the bilayer.

Thirdly, the motion of the Ins domain was examined. RMSF analysis determined that residues 470-510 (comprising unstructured loop residues, except residues 488-500 which form an α-helix) and 520-530 (an unstructured loop between two helices) consistently demonstrated large movements. These residues may act as a secretion determinant *in vivo* by interacting with other components of the T2SS, but as they are uninhibited by such contacts in these MD simulations the residues moved freely in solution.

Future experimental and computational work would aim to fully ascertain the role of the Ins domain, for example by including the bound calcium cations. Atomistic or CG simulations of PulA paired with components such as PulG or PulM, where crystal structures are available, may direct mutational experiments by identifying possible key residues necessary for substrate specificity. Further analysis is necessary to examine the proximity of known sugar binding PulA residues and demonstrate the functional relevance of this data.

Chapter 4 – Effect of calcium ions on PulA simulations

Disclaimer
A portion of the work described in this chapter was included in a manuscript published in 2014, under the title "Efficient Characterization of Protein Cavities within Molecular Simulation Trajectories: *trj_cavity*." Parts of section 4.3.3 were therefore written in collaboration, however the remaining text and all the figures are my own work unless otherwise stated in the legend.

4.1 Introduction

Metal ions are crucial to life and many, such as iron, zinc and magnesium, perform key functions at cellular and sub-cellular levels. Calcium is an abundant alkaline earth metal and is essential for living organisms. The human body contains more calcium than any other mineral, and calcium is likewise vital for single-celled organisms. Changes in calcium concentration, [Ca^{2+}], are associated with the regulation of numerous eukaryotic cellular processes, including transport, motility, gene expression, cell differentiation, pathogenesis and the cell cycle[205–210]. Calcium is also important for prokaryotes; Ca^{2+} ions are involved in maintaining bacterial cell structure, molecular transport, motility and differentiation[211]. Research increasingly suggests that Ca^{2+} coordination is an important feature of periplasmic proteins in Gram-negative bacteria. This coordination is necessary for chemotaxis and cytotoxicity, and performs structural roles, among other functions. The PDB currently lists more than 3,800 bacterial calcium-binding proteins, which contain binding sites that fall into three major categories: EF-hand, C2 and annexin (ANX)[212]. These sites are formed by polar and acidic amino acids and are defined by the pattern formed by the carbonyl and carboxyl oxygens from each type of residue, respectively. The sites can be further classified by loop geometry: EF-hand sites contain a helix–loop–helix motif; C2 domain sites contain multiple β-strands; ANX sites contain a non-adjacent pair of helix–loop–helix motifs[212]. Calcium ions can also play important roles on the outside surface of cells. For example, the outer leaflet of the bacterial OM contains LPS which is cross-linked by divalent cations; these are usually Mg^{2+}, however Ca^{2+} can also play a role. The OM is stabilised by divalent cations, which screen and bridge the repulsive forces that occur in the membrane due to an accumulation of negative charges from LPS[213]. PulA is the only T2SS substrate

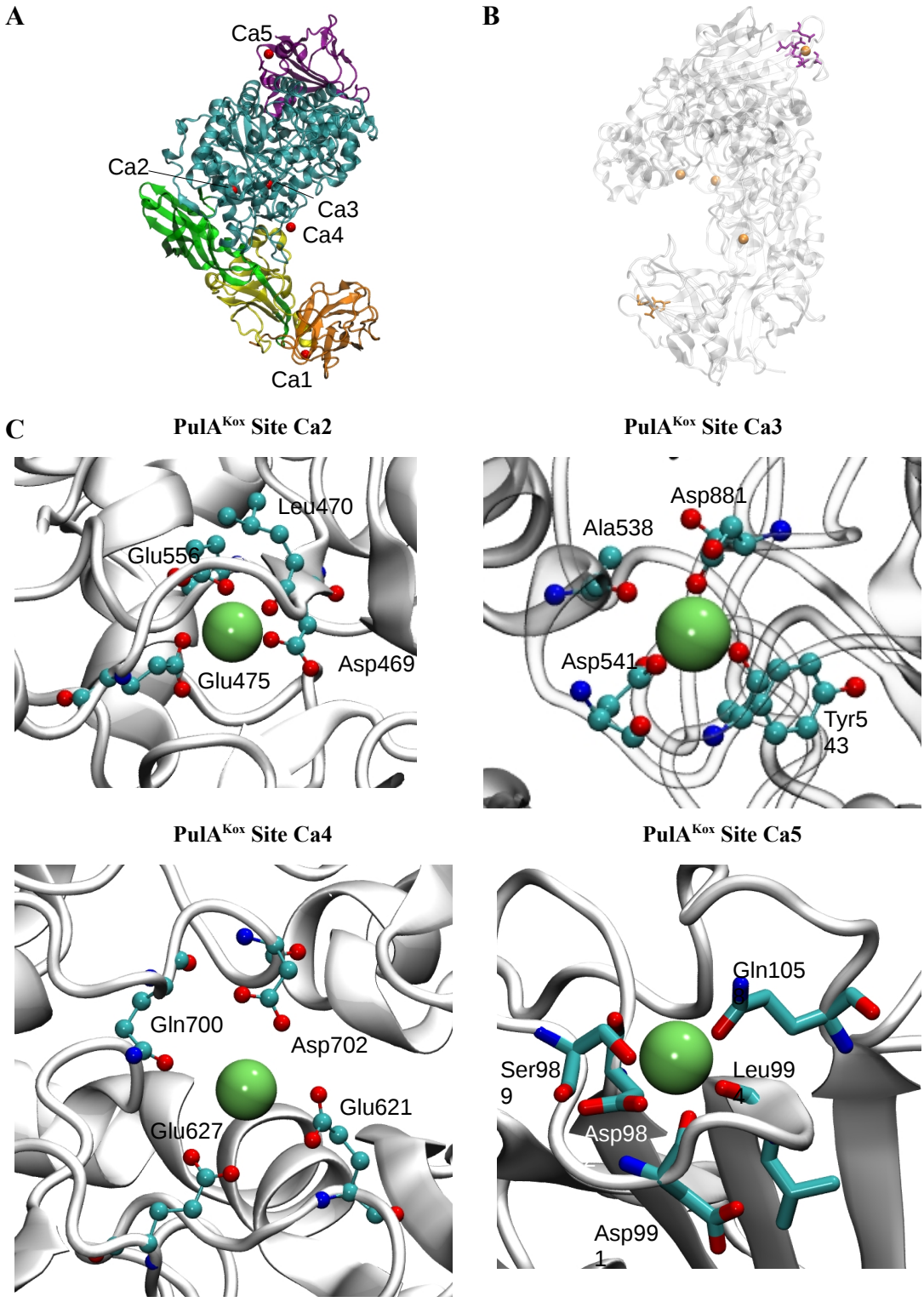
known to remain anchored to the cellular surface following secretion, and as such is likely to encounter calcium or other divalent ions after leaving the cell.

Internal bacterial [Ca^{2+}] was first measured in 1987 by Gangola and Rosen, who found the resting [Ca^{2+}] in energy-replete *E. coli* cells to be 90 ± 10 nM[214]. Subsequent studies constitutively expressed the gene for aequorin (a Ca^{2+}-activated photo-protein used to investigate the system) to continuously measure changes in cytosolic free [Ca^{2+}] across a range of bacterial species[215,216]. Using this technique, Jones *et al*. found that the overall internal [Ca^{2+}] was higher than previously thought, ranging between 170-300 nM[217]. Jones also specifically measured the [Ca^{2+}] in *E. coli* periplasm by targeting aequorin to this compartment, and showed that Ca^{2+} can accumulate in the periplasm at a concentration three- to six-times that of the external medium[218]. This suggests the OM and periplasm may regulate the availability of Ca^{2+} to transporters in the IM, and Ca^{2+} may act as an intracellular regulator in prokaryotes[211]. To date, calcium in bacteria has not been investigated systematically and many valuable studies have not been developed further. This may be due to experimental difficulties, for example toxic reagents, a lack of reliable measurement methods, and the challenges of loading indicator dyes into cells; MD simulations therefore have great potential as an additional tool in this research area.

Interestingly, five calcium binding sites (CaX) were identified in the $PulA^{Kpn}$ structure[193] (Figure 4.1A). Ca1 is located in the N1 domain; three are found in the A domain (Ca2 by the active site; Ca3 in loop 3; Ca4 on the interface with the N2 domain); and Ca5 is in the C domain, on the interface with the A domain. Three of these sites (Ca2/3/4) are identical in $PulA^{Kox}$ and one (Ca5) is coordinated by different residue types. Unlike the Ca1 site, these four are all occupied in the $PulA^{Kox}$ structure; the $PulA^{Kox}$ coordinating residues are shown in Figure 4.1B. Ca2 is coordinated by Asp469, Leu470, Glu475 and Glu556. Ca3 is coordinated by Ala538, Asp541, Tyr543 and Asp881. Ca4 is coordinated by Glu621, Glu627, Gln700 and Asp702. Ca5 is coordinated by Asp982, Ser989, Asp991, Leu994 and Gln1058. Mikami *et al*. commented that the ions they observed may be magnesium rather than calcium, due to lack of clarity with the electron densities, and the absence of calcium in the $PulA^{Kox}$ Ca1 site likewise suggests it may not be a true binding site. The

significance of these sites remains unclear. Mikami only describes the coordination sites and does not comment on any possible functional significance of the calcium cations. Discovery of these sites, along with research by Korotkov and colleagues which showed, by substituting coordinating residues, that calcium is essential for secretion by the major pseudopilin GspG in *E. coli*[91], led me to consider that Ca^{2+} – either bound or environmental – may play a functional role for PulA in the T2SS.

Next page: Figure 4.1 – Calcium binding sites of both $PulA^{Kox}$ and $PulA^{Kpn}$
(A) The $PulA^{Kpn}$ crystal structure, with bound maltotetraose (PDB ID: 2FHF) is shown coloured by domain. The structure contains five occupied Ca^{2+} binding sites, labelled Ca1-5 with the ions represented as red spheres. **(B)** The $PulA^{Kox}$ crystal structure also contains five Ca sites. The Ca1 site is unoccupied and the surrounding residues are shown in orange. Sites Ca2/3/4 are identical to those in $PulA^{Kpn}$ and site Ca5 has different coordination, with the coordinating residues in purple. Calcium ions are shown as orange spheres. **(C)** Residues coordinating each $PulA^{Kox}$ Ca site are labelled and shown as sticks, with the calcium ions represented as green spheres.



PulA contains a central cavity surrounded by the N1, N2 and A domains. This cavity is located between the pullulanase catalytic active site in the A domain and a carbohydrate-binding site in the N1 domain. The catalytic domain, which preferentially hydrolyses α-1,6-pullulan, includes four conserved regions common to

α-amylase family enzymes, and the consensus YNWGYDP sequence from Type I pullulanases. Mutational studies of pullulanase from *K. aerogenes* had shown that substitution of His607, Asp677 and His682 leads to complete loss of enzymatic activity, and substitution of the first two residues prevents binding of cyclodextrins[219], which are competitive inhibitors of pullulanase[219–223]. A subsequent crystallographic study by Mikami *et al*. shed further light on the enzymatic mechanism. PulA^{Kpn} residues 706-710, forming a loop on the domain A periphery, change conformation depending on which substrate analogues are bound, with the main chain torsion angle of each residue changing during the deformation. Comparison of apo- and maltotetrose-bound pullulanase structures demonstrated an induced-fit motion of residue Trp708 and subsequently Glu706 as a result, leading to the catalytic residue and the binding site residue adopting a conformation appropriate for pullulan hydrolysis; the predicted catalytic mechanism is shown in Figure 4.2A. Among the residues involved in substrate binding, Tyr559, Asp602, His607, Arg675, Asp677, Glu706, His 833 and Asp834 (highlighted in Figure 4.2B) are conserved among α-amylase family enzymes. Of these, Glu706 and Asp677 side chains correspond to the catalytic acid/base and nucleophile reported in this enzyme family[219,224–227]. The carbohydrate-binding site in the N1 domain was observed to bind maltose moieties of maltotriose and maltotetraose on the surface of a β-sheet opposite Ca1. The N1 domain is classified as a Carbohydrate Binding Module (CBM) from family 41[228], and the geometry of its maltose binding residues (Tyr78, Lys133, Asp138) is similar to the geometry of CBM20 (another starch binding module found in gluco-amylase, cyclodextrin-glucanotransferase, and β-amylase[229]). Crystallographic data also showed that N1 domain residues Trp80 and Trp95 engage in stacking interactions with glucose sugar rings on the bound substrate.

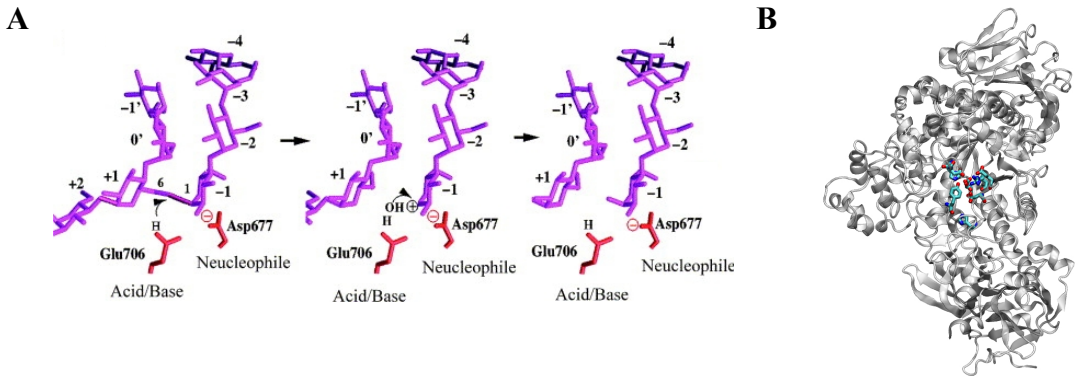


Figure 4.2 – Pullulanase catalytic mechanism and site
(A) Adapted from Mikami *et al.*, 2006. The predicted mechanism of pullulanase catalysis; prior to cleavage, the positions of glucose residues at sub-sites −1 and +1 are adjusted as the torsion angle between C4–O5–C6–O6 of Glc+1 rotates, and the glycosidic torsion angles between Glc−1 and Glc−2 of the maltotetraose complex shift to make an α-1,6 linkage between Glc−1 and Glc +1. This linkage is then hydrolysed by Glu706 and Asp677. **(B)** The conserved residues involved in substrate binding – Tyr559, Asp602, His607, Arg675, Asp677, Glu706, His 833 and Asp834 – are highlighted in colour against a white structure of $PulA^{Kpn}$; the active site is located inside the protein cavity.

For a broader overview of the functions of Ca^{2+} in proteins, due to space constraints the reader is directed elsewhere to thorough reviews of the roles of Ca^{2+} in binding proteins and altering conformations, by which calcium frequently provides regulatory control of biochemical processes, and directly enhances protein stability and enzyme activity[206,230]. Key examples include troponin C (TnC) and calmodulin (CaM), two small acidic proteins that undergo conformational changes when [Ca^{2+}] in the cell increases and modulate the activity of secondary proteins as a result[ref]. Found in vertebrate striated muscle, TnC binds calcium, transmits the resulting conformational changes through the tripartite troponin complex and allows the tropomyosin dimer to induce conformational change in actin. This in turn generates contractile force and allows muscle contraction. CaM is a ubiquitous eukaryotic Ca^{2+} receptor, which in its Ca^{2+}-saturated state activates numerous intracellular enzymes such as phosphodiesterase, phosphorylase kinase and myosin light chain kinase. Notably, a review of the existing literature, intended to provide direction for this investigation into the role of calcium in PulA structure or function, showed that no biomolecular MD simulations using calcium chloride as counter-ions appear to have been published to date.

Interestingly, however, the effect of non-standard counter-ion concentrations has been explored by Dr. Peter Bond and colleagues, who simulated two systems containing OpcA (an OM adhesion protein) and either ~ 0.1 M or ~ 1 M NaCl[231]. At the higher ionic concentration, the observed movement of loop L2 was greatly decreased, possibly due to increased ionic interactions with charged residues in the loop, and the protein was stabilised. Membrane stability also increased as the lipid was rigidified. This study suggests that increased protein stability may also be expected when using calcium chloride counter-ions to mimic the periplasmic environment, as cations with a greater charge will be present and can both interact with charged amino acids and stabilise the membrane.

With mounting evidence that PulA could be affected (and likely stabilised) by high periplasmic [Ca^{2+}], with a possible effect on function, MD simulations provide an interesting approach to further understanding this phenomenon. MD simulations usually include solely sodium and chloride counter-ions; however, this study used calcium and chloride counter-ions, to elucidate any specific and possibly functional effects of an environment containing Ca^{2+}. This work proved an exciting opportunity to experiment beyond standard protocols, provide a novel perspective and potentially produce increasingly physiologically relevant data with respect to PulA, possibly with wider implications in the research field. Full-length $PulA_{NA}$ was modelled adsorbed to a POPE bilayer in a calcium chloride solution, and also simulated in sodium chloride with bound crystal calcium ions. These simulations provided insights into the putative effect of calcium on periplasmic PulA dynamics.

4.2 Methods

In this smaller-scale study, two systems were simulated and analysed: $PulA_{IONS}$, a replicate of $PulA_{NA}$ with the crystallographic Ca^{2+} ions retained, and $PulA_{CALCIUM}$, containing solvated PulA, POPE, and calcium and chloride counter-ions (the crystallographic Ca^{2+} ions were not retained). In each, PulA lacked the wild-type acyl anchor but contained the modelled N-terminal tether. The $PulA_{IONS}$ system was created by adding the crystallographic ions from the PDB ID 2YOC structure into the $PulA_{NA}$ system from Chapter 3. The $PulA_{CALCIUM}$ system was created by adding Ca^{2+}

and Cl^- (rather than Na^+ and Cl^-) ions to the solvated full-length model situated on the POPE bilayer, used as a starting point for the $PulA_{NA}$ simulations in Chapter 3. Similarly to the systems described in Chapter 3, the $PulA_{CALCIUM}$ system was neutralised by replacing random water molecules with Ca^{2+} and Cl^- to a concentration of ~ 0.1 M. Overlapping solvent and two POPE lipids were removed, resulting in a system containing ~ 48,500 water molecules and 318 lipids. The system occupied a triclinic box with dimensions ~ 95 x 95 x 195 Å.

Following ionisation of $PulA_{CALCIUM}$, successive energy minimisations were performed using a steepest-descent algorithm, to minimise steric overlap between the components of each system. Subsequent equilibration simulations allowed the solvent to relax around each restrained biomolecular complex. Harmonic restraints were applied to all the protein non-hydrogen atoms and gradually removed in three distinct 0.5 ns phases. Finally, production MD was executed in triplicate on the unrestrained systems for 100 ns, within the NPT ensemble.

All the simulations in this study were performed using the GROMACS 4.5.5 simulation package[185], with the CHARMM36 all-atom FF[175] used to treat the lipid molecules in the $PulA_{CA}$ system, and CHARMM22/CMAP parameters[175] used to treat the protein in all three systems. As in Chapter 3, simulations were completed using the leap-frog algorithm with a 2 fs time-step, and trajectory data was collected every 10 ps. The LINCS algorithm[181] was used to constrain bond lengths and the neighbour list was updated every 10 steps. The long-range electrostatic interactions were corrected using the Particle Mesh Ewald method[184]; the real-space electrostatic and van der Waals interactions were cut off at 12 and 10 Å, respectively. Simulations were performed at an absolute temperature of 310 K using the V-rescale thermostat. The pressure was set to 1 bar (Parrinello-Rahman semi-isotropic barostat, 5 ps coupling constant), under PBC.

Table 4.1 contains the nomenclature of the structures and simulations analysed in this chapter, for the clarity of the reader.

Table 4.1 – Systems analysed in this chapter

Name	System	POPE	Counter-ions	Simulations
$PulA_{CALCIUM}$	POPE + crystal structure of PulA (2YOC) with modelled tether (res 1-19)	YES	$Ca^{2+}Cl^{-}$	3 x 100 ns
$PulA_{IONS}$	POPE + crystal structure of PulA (2YOC) with modelled tether (res 1-19) + crystallised calcium ions	YES	$Na^{+}Cl^{-}$	3 x 100 ns

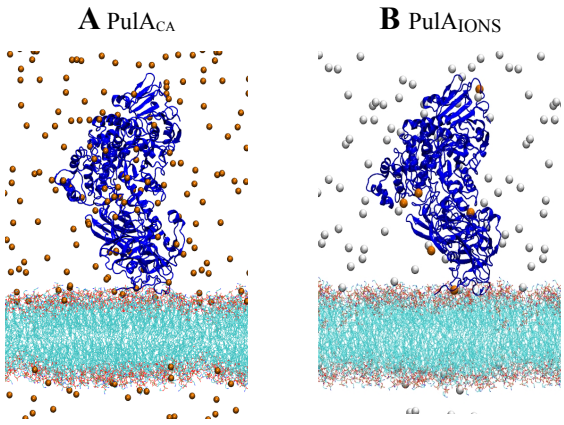


Figure 4.3 – Representations of systems analysed in this chapter **(A)** $PulA_{NA}$ is placed on a POPE bilayer and the system contains calcium ions, shown as orange spheres. **(B)** $PulA_{NA}$ and the crystallographic Ca^{2+} ions (orange) are placed on a POPE bilayer; the system contains sodium ions (white spheres).

4.3 Results

4.3.1 – Effect of calcium on protein dynamics

Analysis addressed whether the presence of either crystallographic calcium ions or calcium chloride solution affects PulA dynamics and lipid interactions. Visual analysis immediately demonstrated a significant change in protein motion; across all three $PulA_{CA}$ simulations, the protein remained upright and did not collapse onto the membrane, as in the simulations lacking any calcium cations in Chapter 3 (Figure 4.4). However, PulA in the $PulA_{IONS}$ simulations interacted extensively with the lipid bilayer, suggesting a high concentration of calcium ions in the environment is required to constrain the protein conformation.

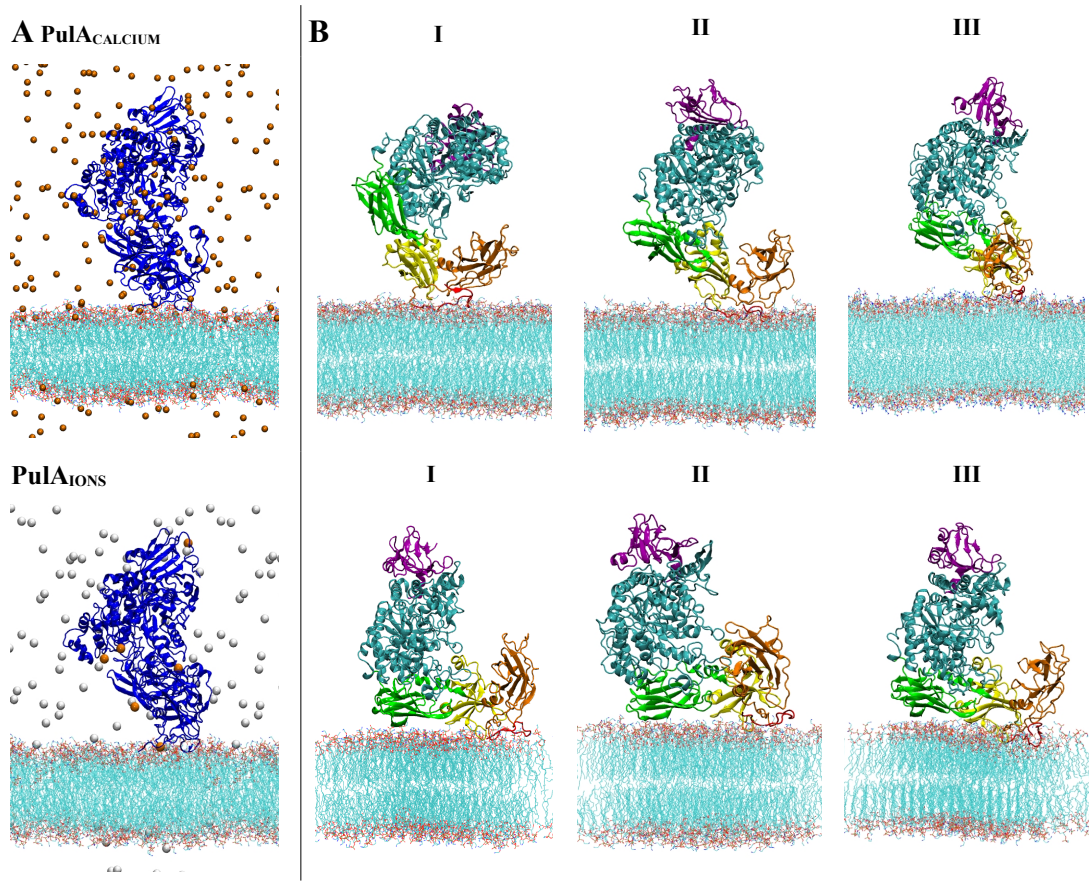


Figure 4.4 – Snapshots of protein conformations from PulA$_{CALCIUM}$ and PulA$_{IONS}$ simulations
(A) Initial structures of PulA$_{CALCIUM}$ and PulA$_{IONS}$ are shown, with PulA placed near the membrane. The protein is shown in blue, sodium ions as white spheres, calcium ions as orange spheres, and the lipid is coloured by atom (oxygen – red, carbon – cyan, nitrogen – blue, phosphorous – tan). **(B)** The final conformations following three 100 ns production runs. When solvated by calcium chloride solution, the protein (coloured by domain) remains upright and does not form an increased number of contacts with the membrane. However, in the presence of crystallographic calcium ions only, the protein increases interactions with the membrane as in the PulA$_{NA}$ simulations described in Chapter 3.

Calculating the RMSD of the protein domains of PulA$_{CALCIUM}$ and PulA$_{IONS}$ demonstrated that all three systems exhibited similar trends (shown in Figure 4.5). In the PulA$_{CALCIUM}$ simulations, the tether and N1 (and in one replicate, the C) domains exhibited the largest RMSD values, usually above ~ 0.4 nm. This result was unexpected given the visual observation suggesting that protein rigidity increased in the presence of calcium cations. The tether and N1 domains of the PulA$_{IONS}$ replicates exhibited higher RMSD values of ~ 0.4-0.6 nm, in comparison with these domains in the PulA$_{NA}$ simulations which had RMSD values that did not tend to increase above

0.4 nm. The remaining domains exhibited similar RMSD values of ~ 0.1-0.3 nm in both systems, suggesting that the presence of bound calcium cations is likely to affect localised regions and not the PulA conformation as a whole.

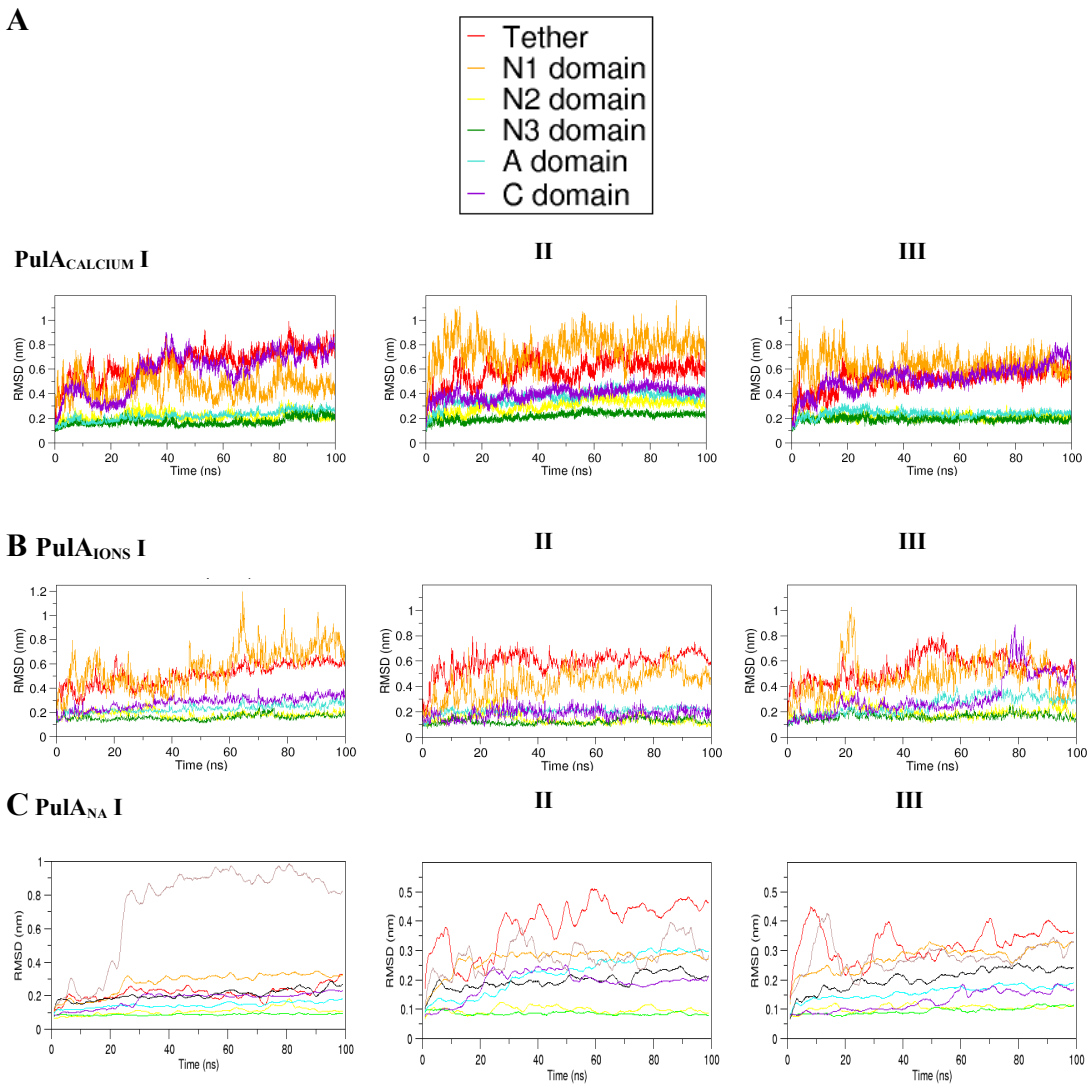


Figure 4.5 – RMSD of protein domains from PulA$_{CALCIUM}$ and PulA$_{IONS}$ systems
The RMSD of each protein domain – tether, N1, N2, N3, A, C – from **(A)** PulA$_{CALCIUM}$, **(B)** PulA$_{IONS}$ and **(C)** PulA$_{NA}$ (also contains RMSDs of the Ins domain, coloured in black, and the N1/N2/N3 domains combined in grey). The datasets are differentiated by colour, according to the legend.

The RMSF of PulA showed that during the final 10 ns of the simulations, the RMSFs of the Cα atoms of PulA$_{CALCIUM}$ and PulA$_{IONS}$ (Figure 4.6A/B) were remarkably similar to that of the PulA$_{NA}$ system (Figure 4.6C). This indicated that the protein did not undergoing any large movements during these time steps. Once more, the N1

(residues 20-160) and Ins (residues 475-545) domains consistently demonstrated the most fluctuation, and the presence of calcium did not have a significant effect of the fluctuation of the protein backbone.

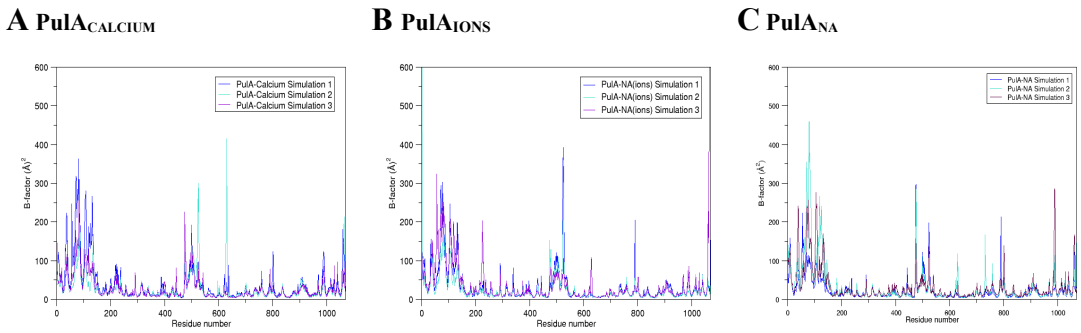


Figure 4.6 – RMSF of protein Cα atoms during final 10 ns of PulA$_{CALCIUM}$, PulA$_{IONS}$ and PulA$_{NA}$ simulations
Comparison of the mean motion of each Cα atom during the final 10 ns of each simulation demonstrates that the N1 (residues 20-160) and Ins (residues 474-545) domains consistently fluctuate the most. The presence of calcium does not have a significant effect of the fluctuation of the protein backbone.

Analogously to previous analysis, the final structures (except the tether and N1 domains) from each simulation was superimposed onto the crystallographic chain A and chain B structures. This aimed to ascertain the effect of the presence of calcium ions, both bound and in solution, on the 47° N1 domain rotation observed both in the crystal dimer and in the PulA simulations in Chapter 3. The results are visualised in Figure 4.7; in the PulA$_{IONS}$ system, the N1 domain remained extremely close in configuration to the starting structure and no rotational movement was observed. In two replicates, the domain even rotated slightly in the opposite direction to the chain B conformation. In the PulA$_{CALCIUM}$ simulations, the N1 domain exhibited a larger degree of motion, however the chain B conformation was still not mimicked fully. Relative to the PulA simulations lacking calcium (see Figures 3.6B and 3.10), it is evident that the presence of calcium ions increases the rigidity of the otherwise flexible N1 domain.

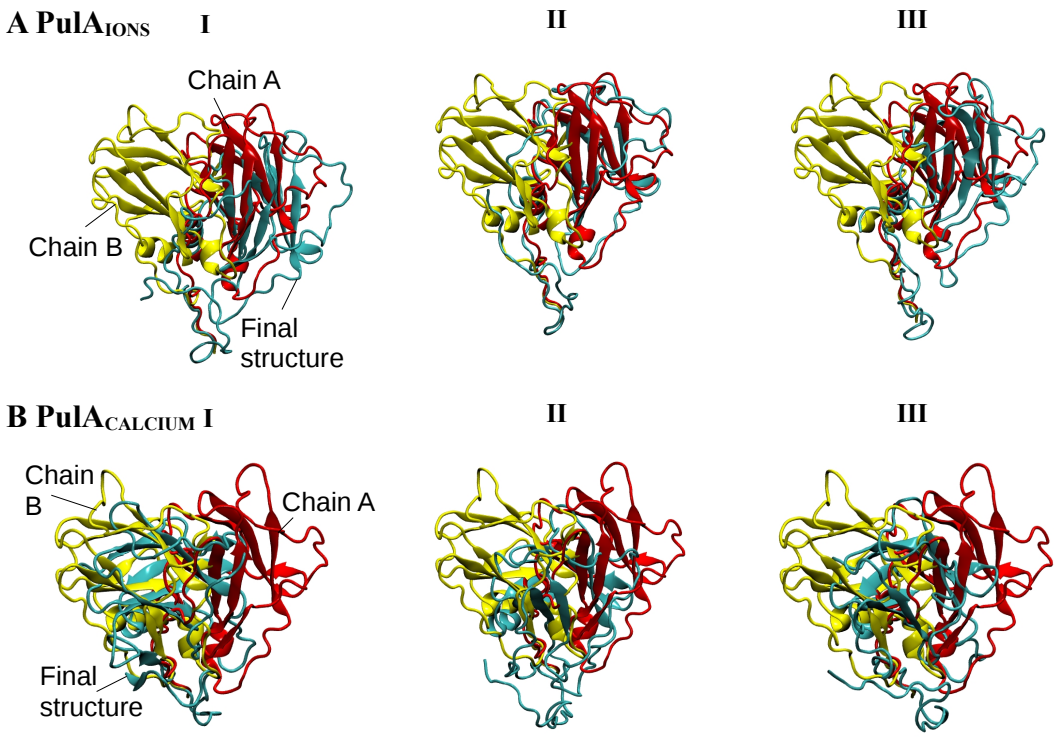


Figure 4.7 – N1 domain movements
All domains except N1 and the tether were superimposed. In the crystal structure, two N1 domain conformations are seen, chain A (red) and chain B (yellow), relative to the remaining protein (not shown). In the **(A)** PulA$_{IONS}$ system, the N1 domain remained very close to the subunit conformation after 100ns (cyan), whereas in the **(B)** PulA$_{CALCIUM}$ system, the final N1 domain position was repeatedly more similar to chain B.

4.3.2 – Effect of calcium on protein-lipid interactions

Assessing the number of hydrogen bonds between protein residues and POPE over time showed that the tether, N1 and N2 domains consistently formed a similar number of bonds to the lipid bilayer as in the simulations lacking any calcium ions (Figure 4.8). However, in the presence of calcium counter-ions the N3 domain did not form any hydrogen bonds to lipid, in line with the upright protein conformation observed visually. N3 domain in PulA with bound calcium ions formed up to 5 hydrogen bonds to lipid in the final 25 ns of two replicates only.

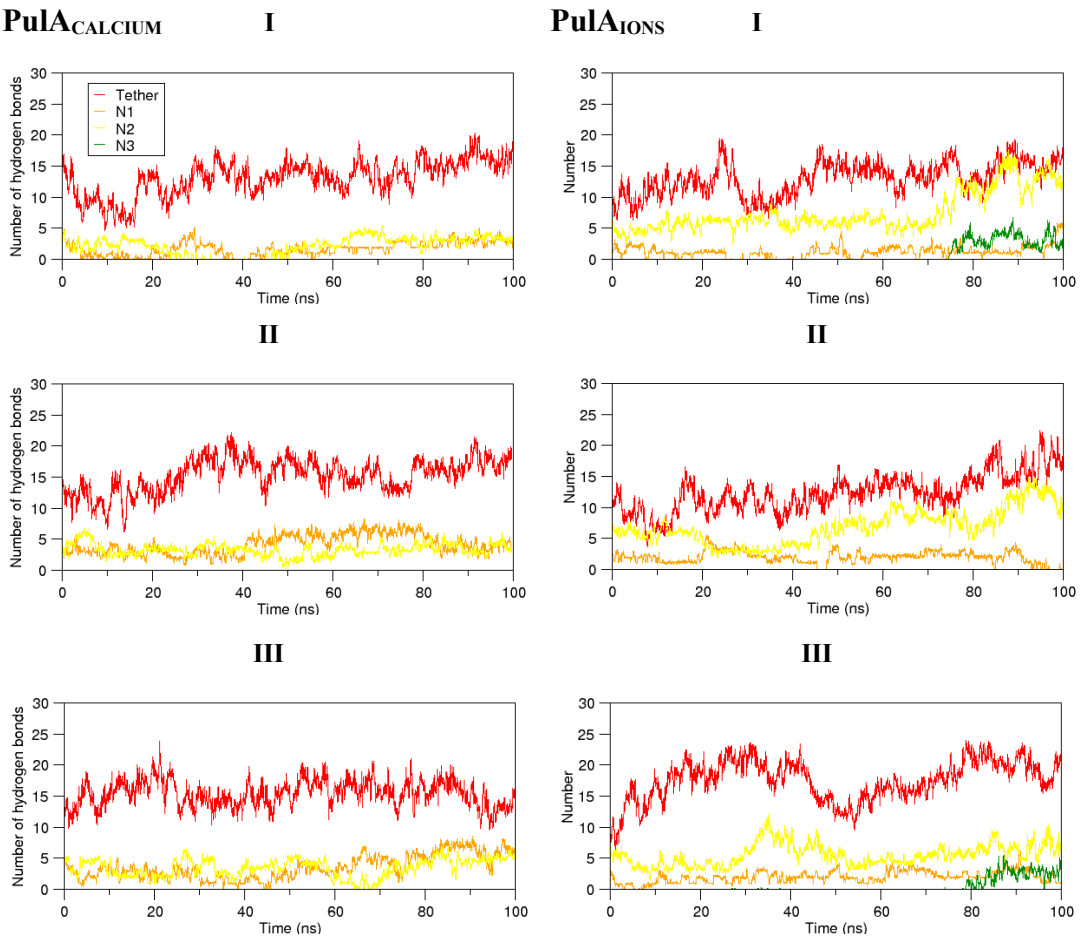
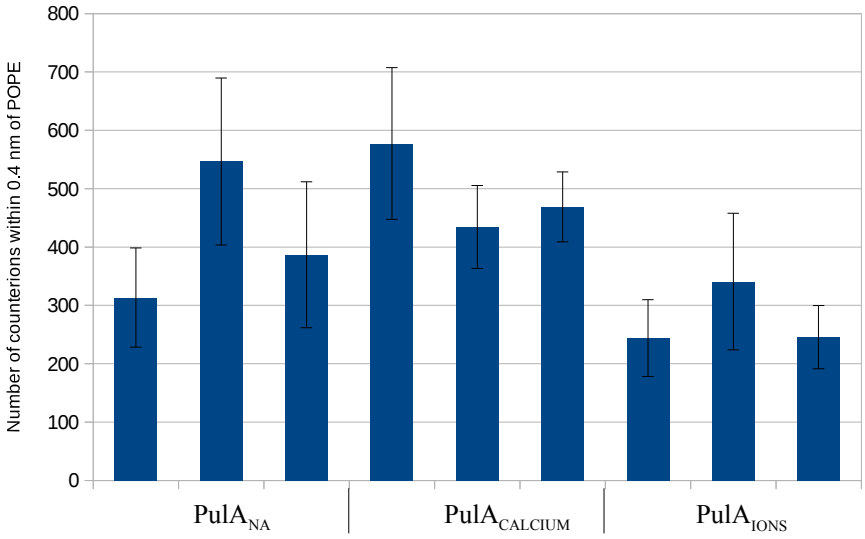


Figure 4.8 – Hydrogen bonding to POPE
Hydrogen bonds between protein residues within each of the tether (red line), N1 (orange), N2 (yellow) and N3 (green) domains, and POPE; the graph colour codes are also indicated in the legend. Tether/N1/N2 domains form a similar number of bonds to the lipid bilayer as in the simulations lacking any calcium ions. However, in the PulA$_{IONS}$ system the N3 domain does not form any hydrogen bonds to lipid, and in the PulA$_{CALCIUM}$ system, few hydrogen bonds are formed by N3.

To confirm that the lack of conformational change was not due to the calcium ions chelating the membrane and preventing protein-lipid interactions, the number of Ca^{2+} ions within 4 Å of POPE at any given time frame between 20 and 100 ns (following system equilibration) was calculated, and compared to the number of sodium ions within 4 Å of POPE in the PulA$_{NA}$ system (Table 4.2). This demonstrated no significant difference in the number of ions close to the lipid bilayer, and suggested that the lack of protein-lipid interactions is due to protein rigidity preventing PulA being "pulled" onto the membrane as contacts form, rather than cations on the surface of the lipid preventing the protein adsorbing fully.

Table 4.2 – Number of counter cations within 4 Å of POPE

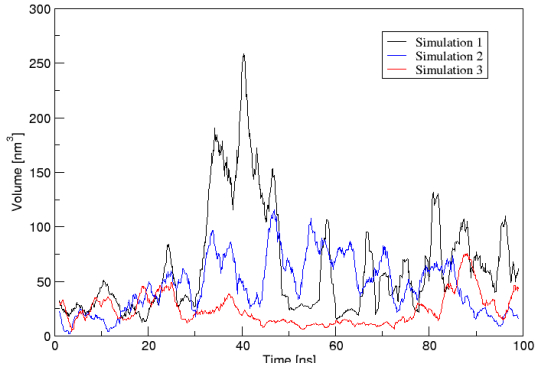
	I	STD	II	STD	III	STD
$PulA_{NA}$	313	85	547	143	387	125
$PulA_{CALCIUM}$	577	130	435	71	469	60
$PulA_{IONS}$	244	66	341	117	245	54



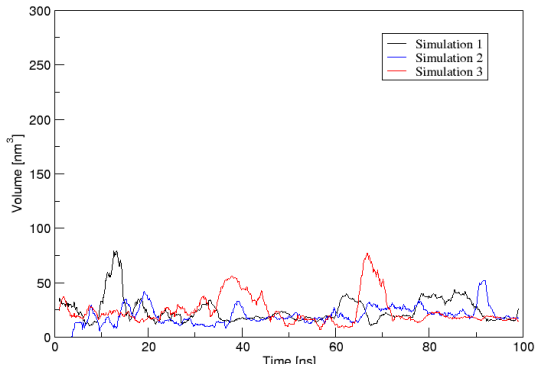
4.3.3 – Functional effect of calcium

The functional effect of calcium on PulA was analysed using the *trj_cavity* programme, which uses a novel grid-based algorithm to perform an efficient neighbour search. Any cuboid segment, allocated by the programme, not surrounded on all six sides by segments containing protein are considered to form part of the cavity. The results of this analysis are shown in Figure 4.9, and demonstrated that the presence of a high concentration of calcium cations in the environment decrease the volume of the catalytic cavity up to ~ 5 fold. The cavities in the $PulA_{NA}$ replicates varied as the protein conformation changed throughout the simulations, and also varied between replicates (with a particularly large cavity observed in replicate 1), with the cavity volume increasing to more than 50 nm^3 in all three replicates. The cavities were much smaller in the $PulA_{IONS}$ simulations, and smaller still (no larger than 45 nm^3) in the $PulA_{CALCIUM}$ replicates. These results suggest possible functional significance of the high periplasmic calcium concentration, as the protein remains rigid in the periplasm prior to secretion, presumably unable to perform catalytic function.

PulA$_{NA}$



PulA$_{IONS}$



PulA$_{CALCIUM}$

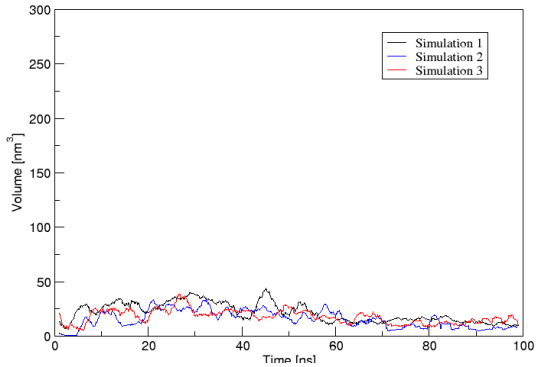


Figure 4.9 – Volume of the largest PulA cavity over time
Trj_cavity analysis allowed quantification of the PulA catalytic cavity over the course of each system replicate. The cavities in the PulA$_{NA}$ replicates varied as the protein conformation changed throughout the simulations, and also varied between replicates (with a particularly large cavity observed in simulation 1), with the cavity volume increasing to more than 50 nm^3 in all three replicates. The cavities were much smaller in the PulA$_{IONS}$ simulations, and smaller still (no larger than 45 nm^3) in the PulA$_{CALCIUM}$ replicates.

4.4 Discussion

The MD simulations presented here support the notion that PulA could be affected (and likely stabilised) by high periplasmic [Ca^{2+}], with a possible effect on function. However, certain data points appeared conflicting. On one hand, visual analysis and calculating protein-lipid hydrogen bonding showed that the presence of calcium chloride ions in the solvent increased PulA rigidity, and decreased the interactions of the protein with POPE. On the other hand, the B factor values demonstrated that the presence of calcium did not have a significant effect of the fluctuation of the protein backbone, and the RMSD values did not vary significantly from those of the $PulA_{NA}$ simulations. Calculating the volume of the cavity between domains N1 and A showed that the cavities in the $PulA_{NA}$ replicates varied as the protein conformation changed throughout the simulations, however the cavities were much smaller in the $PulA_{IONS}$ and $PulA_{CALCIUM}$ simulations, suggesting a functional implication of the periplasmic environment. Future analysis would include mapping the catalytic sites to the cavity, and even simulating the sugars binding.

Chapter 5 – MD Simulation Studies of Pseudopilin PulG

Disclaimer
A portion of the work described in this chapter was included in a manuscript accepted for publication in 2017, under the title "Polar N-terminal residues conserved in type 2 secretion pseudopilins determine subunit targeting and membrane extraction steps during fibre assembly". Parts of the Results section were therefore written in collaboration, however the remainder of the text and all of the figures are my own work unless otherwise stated in the legend.

5.1 Introduction

PulG, introduced in section 1.2 as the major pseudopilin of the *K. oxytoca* T2SS, comprises 133 amino acids that form an extended 54-residue amino-terminal helix with a globular head domain (Figure 5.1A/B), the fold of which is fairly conserved between bacterial species[78]. Details are found in section 1.4 and are briefly recapped here. PulG, the most abundant pseudopilin[79], is inserted into the IM, cleaved by prepilin peptidase[51,81] and methylated at the conserved amino-terminal phenylalanine residue[82] to form N-terminal-methylated Phe (designated "MPH" here) at the terminus. Methylation is believed to occur almost always, however in over-expressing conditions a fraction of the protein remains unmethylated. PulG is required for secretion, along with PulI, PulJ and PulK, which form a tetramer with PulH on the tip of the pseudopilus[31,84]. Following STEM analysis and X-ray crystallography of truncated $PulG_{25\text{-}134(His)6}$ to obtain a 1.6 Å resolution crystal structure, a pseudo-atomic model of the pseudopilus structure has been built by Manuel Campos and colleagues (Figure 5.1C)[89]. This modelled pseudopilus has a rough surface with deep grooves separating the helix strands and a narrow hydrophobic central cavity. The structure has 4.25 PulG monomers per turn, and each monomer (*P*) interacts with three monomers above (*P+1*, *P+3*, *P+4*) and below it (*P-1*, *P-3*, *P-4*). Truncated $PulG_{25\text{-}134(His)6}$ does not form pili or perform secretion, indicating that the TM amino terminus may play a key role in this process. Biochemical and functional validation of the model has shown that electrostatic interactions between *P* and *P+1* monomers play a key role in assembly and secretion. To develop existing knowledge of the role of PulG in secretion, it is crucial to study the system dynamics relating to assembly, oligomeric contacts, and protein-

lipid interactions at an atomic level.

Dr. Olivera Francetic and colleagues have been studying the secretion mechanism involving PulG and pseudopilus formation, and have gleaned a number of insights. In particular, electron microscopy (EM) and a structural modelling method based on self-organising maps (whereby structures are organised according to structural similarity and plotted on a 2D map) uncovered a possible structural transition path between three energy basins towards a low-energy pilus conformation, driven by a continuous increase in the helical twist[88]. The *P–P+1* contacts were shown to play a crucial role in the oligomer docking step necessary for pseudopilus assembly and PulA secretion[88].

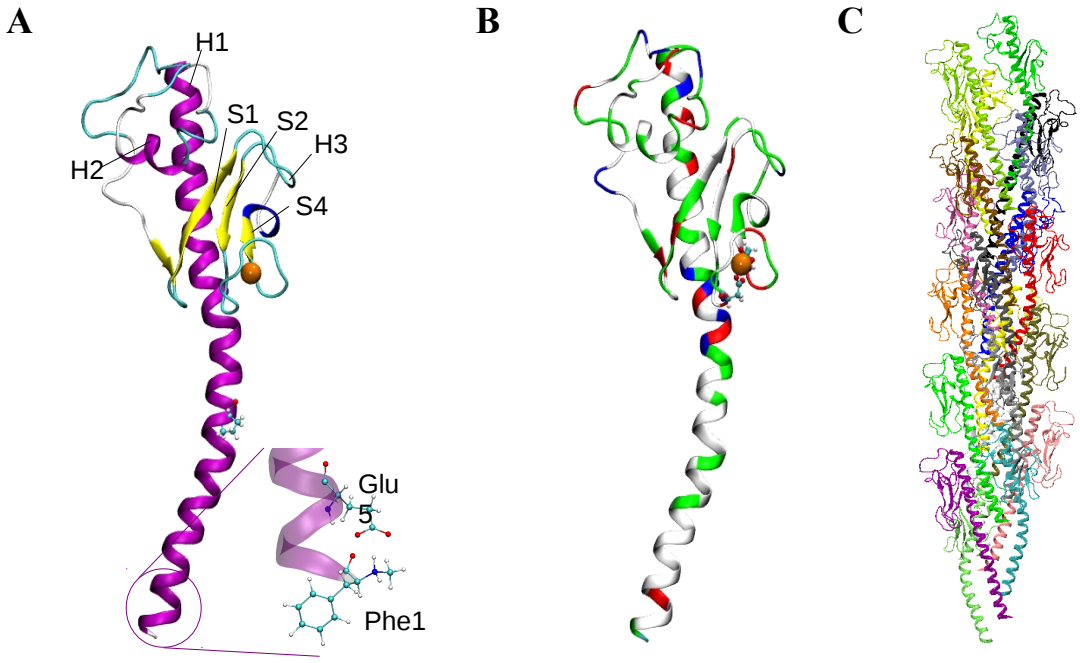


Figure 5.1 – PulG monomer and assembled pseudopilus models
(A) Model of a *K. oxytoca* PulG monomer, created using STEM and crystallographic data, followed by homology modelling. The protein is coloured according to secondary structure (α-helix – purple, 3-10 helix – blue, β-sheet – yellow, turn – cyan, coil – white). Pro22 is represented in stick format, as are methylated Phe1 and Glu5 (inset); the structural motifs are labelled. **(B)** PulG monomer coloured according to residue type (non-polar – white, basic – blue, acidic – red, polar – green). The calcium ion is shown (orange) coordinated by Asp117 and Asp125 (in stick format). **(C)** The modelled oligomeric pseudopilus is shown in ribbon view, with each PulG monomer differentiated by colour. The structure has a helical pitch of 43.8 Å and 4.25 units per turn, with an outer diameter of 65 Å.
Contrary to predictions from the current model for T4P assembly, in the T2SS Glu5

of *P* was shown not to interact with the N-terminal amine of *P+1* except in the initial, unstable and high-energy conformations[88]. The subsequent presence of *P–P+3* and *P–P+4* contacts in the more stable conformations demonstrated their role in downstream stabilisation steps. Overall the study suggested a spool-like one-start mechanism for assembly of pili from the T2SS-T4P super-family, providing new evidence for rotational assembly. To build on these results, this chapter contains analysis of PulG homodimer MD simulations generated to examine the *P* and *P+1* dimer interface in motion and also in a membrane environment, a novel undertaking omitted from the study by Nivaskumar *et al*[88].

The super-family is thus categorised because the T2SS bears a strong resemblance to the T4P assembly machineries, and the T2SS can produce T4P-like fibres on the bacterial cell surface when PulG is overproduced [232]. In most T4Ps, the only charged residue found among the 25 N-terminal amino acids is the conserved Glu5. In the T2SS pseudopilus model, (*P*)E5 interacts with K28 and K35 of the *P-3* monomer. Substitution of Glu5 by Ala ($PulG_{E5A}$) has been shown to prevent both piliation and secretion[88], and prevents interaction of PulG with PulM – the T2SS protein currently thought to act as a targeting factor. In biochemical studies, $PulG_{WT}$ was turned over with a half-life of 100 minutes but the assembly-defective $PulG_{E5A}$ variant was highly stable, presumably accumulating in the IM[88]. The E5 residue therefore appears to protect the protein from periplasmic proteases; $PulG_{E5A}$ may arrest at a step prior to membrane extraction and remain in a compartment protected from degrading enzymes. Alternatively, the protein may simply be anchored more strongly in the membrane, preventing extraction by proteases: perhaps the N-terminus of the variant interacts with the membrane lipids, whereas in $PulG_{WT}$ E5 may instead promote intra-molecular interactions with the positively charged N-terminal MPH. Analysis of PulG using mass spectrometry has confirmed that both $PulG_{WT}$ and $PulG_{E5A}$ are methylated, including in pili containing both variants. This suggests that $PulG_{E5A}$, which was previously shown to be incompetent for assembly as it cannot interact with the assembly factor – putatively PulM – may form a heterodimer with $PulG_{WT}$, allowing it to enter the oligomer assembly site. PulM is a T2SS protein, attributed to the IM platform. Its role continues to be defined, but it is currently proposed to function as the targeting factor, bringing the major pilin to the T2SS machinery IM

site where oligomerisation occurs (Santos-Moreno *et al.*, accepted). Unexpectedly, unprocessed PulG interacts more strongly than processed (i.e. cleaved and methylated) PulG with PulM (Santos-Moreno *et al.*, accepted). PulG has been shown to crosslink to both PulM and PulL (another T2SS IM protein) in complex, suggesting these are involved in the formation of pili, but more research will be necessary to clarify the details of the process.

The PulG crystal structure contains a bound calcium cation at residues Asp117 and Asp125, which is thought to stabilise the β2-3 loop of the protein. According to the pseudopilus model, the PulG monomers are hypothesised to interact within the membrane in a staggered fashion, tilting *via* interactions of the Ca^{2+} binding loop with phospholipids and potentially assisted by a kink induced by Pro22 – the only proline residue located in the helix. Notably, a P22A substitution has been shown to adversely affect the efficiency of oligomer assembly[89]. Cryo-EM of PulG pili shows that the zone around Pro22, including a number of downstream Gly residues, melts and is not α-helical (Edward Engelman, personal communication). MD simulations, as mentioned previously, are an excellent method to obtain information at atomic resolution, and this study aimed to examine the dynamics of the PulG dimer interface, shed light on the experimental observations regarding the role of Pro22, and clarify the extent of the calcium-binding loop interactions with phospholipids.

Four key residues have been consistently predicted to form salt bridges in PulG dimers, (*P*)Glu44–(*P*+1)Arg88 and (*P*)Asp48–(*P*+*1*)Arg87, and single charge-inversion variants decrease or abolish assembly[88]. The latter of these two salt bridges is believed to play a key role in the early docking stage of pseudopilus assembly. The double alanine variant $PulG_{E44A/D48A}$ is rapidly degraded, for unknown reasons. Dimerisation may be required for protection against proteolysis, and it may be that only a single substitution is responsible for the degradation effect, although this has yet to be determined experimentally; certainly the quadruple alanine variant abolished the dimerisation interaction. The current model predicts that $PulG_{E44A/D48R}$ would not be incorporated into pili while $PulG_{E44A/R87E}$ would. Simulating a PulG homodimer enabled me to examine the dimer interface in motion and identify key residues.

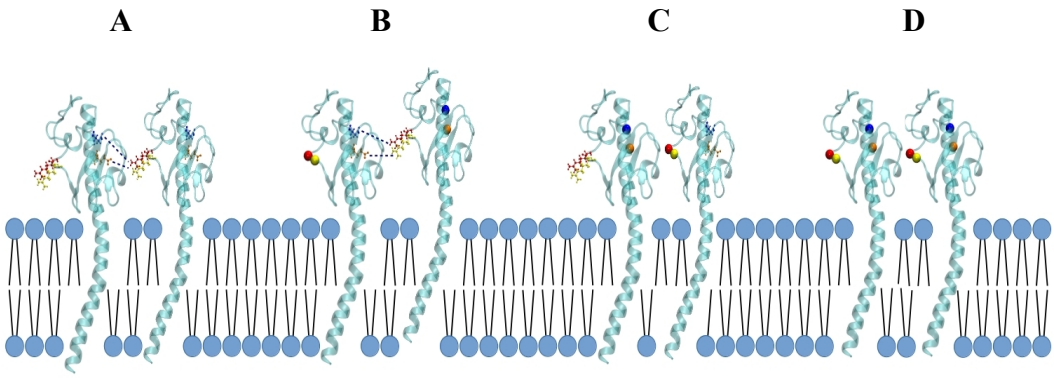


Figure 5.2 – Key residues form salt bridges during PulG dimerisation
Four key residues – (*P*)Glu44 (orange), (*P*)Asp48 (blue), (*P+1*)Arg87 (yellow), (*P+1*)Arg88 (red) – have been predicted to form salt bridges. Four sets of alanine mutants, shown here schematically with the substituted Ala residues shown as spheres, were tested using bacterial two-hybrid (BACTH) analysis, and demonstrated which residues were necessary for docking of the incoming protomer *P-1*. The BACTH system showed that **(A)** WT protein dimerised effectively, substituting two residues adjacent in the salt-bridge interaction **(B)** reduced or **(C)** abolished interactions, and **(D)** quadruple Ala substitution also abolished interactions.

It has been repeatedly emphasised throughout this thesis that MD simulations can complement experimental work and provide atomistic insights to results from the experimental bench. For this chapter, I performed computational substitutions of PulG, yielding a wild-type (WT) structure with methylated Phe1 (MPH), and E5A variant structures with/without the post-translational methyl modification. These simulations were expected to clarify whether methylation alters protein-membrane interactions or disrupts the lipid bilayer, providing evidence regarding whether methylation assists the monomer to be removed from the membrane more easily for pseudopilus formation. I also aimed to shed light on whether Glu5 interacts with the positive charge on MPH, given the possible steric hindrance caused by the bulky methyl group. Likewise, experimental data found that the unprocessed/unmethylated PulG was less stable than the processed and methylated protein – it was hoped that MD simulations may explain the molecular basis for this. The study included observation of the role of Pro22 in protein-membrane interactions and examination of any change in secondary structure around this region of the protein.

5.2 Methods

The full-length model of the *K. oxytoca* PulG monomer was derived from the X-ray crystal structure of the *K. pneumoniae* PulG periplasmic domain (PDB ID: 1T92[83]) solved by Rolf Koehler and colleagues. The missing 20 carboxy-terminal residues, involved in the domain swap in the crystallographic dimer, were modelled on the basis of close homology to GspG from EHEC and the TMS was modelled from PilA of *P. aeruginosa* (48 % identity, 84 % similarity in the TMS)[90]. All residue substitutions in this chapter were performed using PyMOL (see Figure 5.3). The final monomeric PulG structure was composed of 133 amino acids and included one calcium ion, and all ionisable groups were assigned to their most probable charged states at neutral pH. Dimeric PulG was obtained by selecting the bottom two monomers from the published multimeric model formulated by Manuel Campos and colleagues[233]; the dimer is shown in Figure 5.3D.

All PulG variants were then embedded in a pre-equilibrated POPE lipid bilayer, using the *g_membed* tool provided in the GROMACS package[187]. PulG position along the *z* axis was guided by the presence of non-polar and hydrophobic residues in the lower section of the α-helix (see Figure 5.1B), and interfacial aromatic residues Phe1 and Trp94. See section 2.8 for further details on *g_membed*. The systems were placed in triclinic boxes with dimensions ~ 90 x 90 x 160 Å (monomeric systems) or ~ 80 x 90 x 130 Å (dimeric systems). The system was explicitly solvated with the TIP3P water model, *via* superimposition of a pre-equilibrated box of waters, and electrically neutralised by replacing random water molecules with sodium chloride counter-ions to 0.1 M concentration. The resulting systems contained ~ 36,500 water molecules and 314 lipids (~ 26,000 water molecules and 247 lipids in the dimer system).

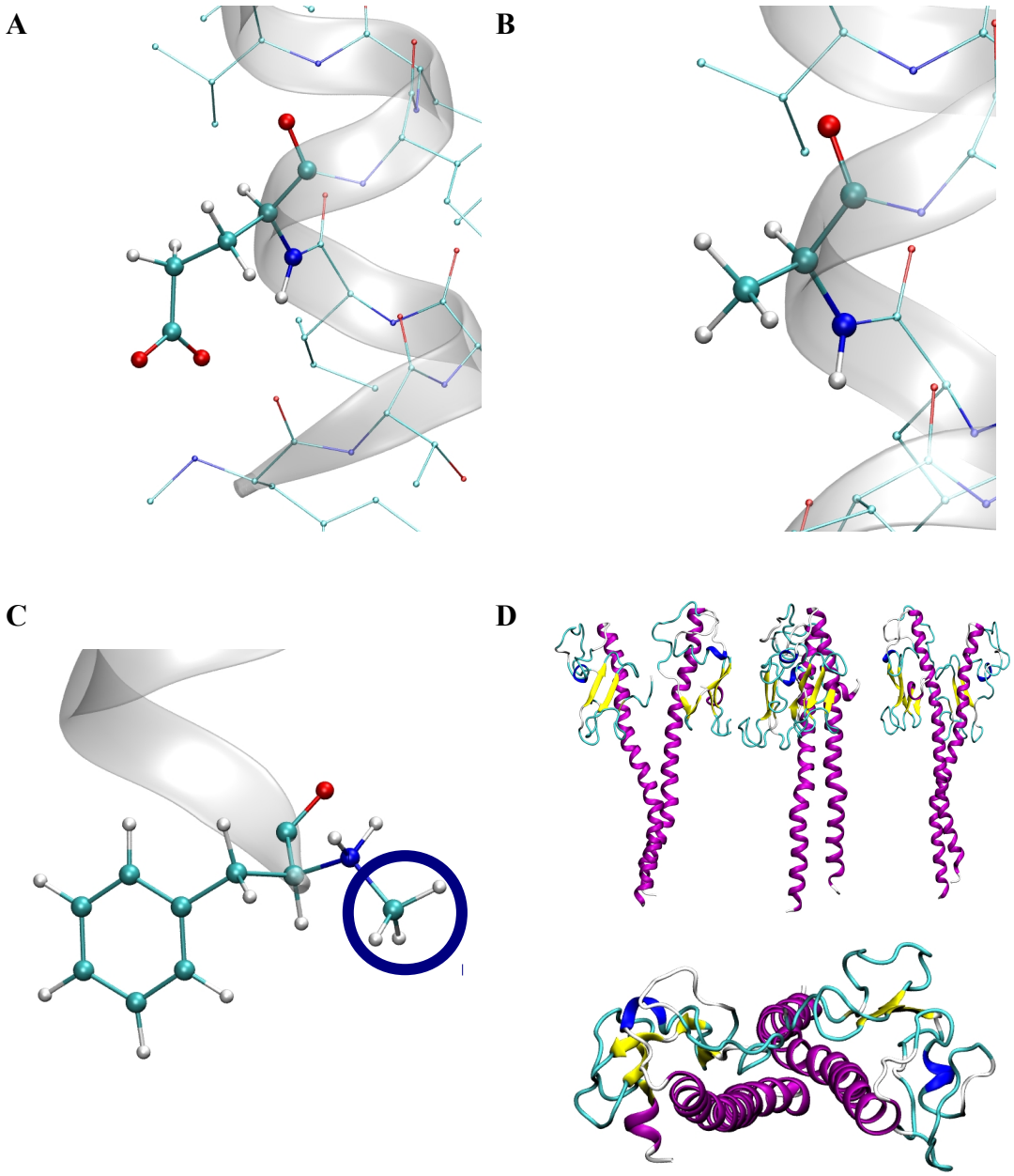


Figure 5.3 – PulG structures prior to simulation
(A) Illustration of the WT Glu5 of PulG, and **(B)** the substituted Ala5 residue, both represented in stick format and coloured by atom (carbon – cyan, nitrogen – dark blue, oxygen – red, hydrogen – white). **(C)** Illustration of the N-terminal methylated Met1 residue, MPH. The newly added methyl group is circled in blue. **(D)** The PulG dimer is shown in a ribbon format and coloured according to secondary structure (α-helix – purple, 3-10 helix – blue, π-helix – red, β-sheet – yellow, turn – cyan, coil – white), shown from multiple angles.

All of the variants were embedded in a lipid bilayer, placed in a box, solvated and neutralised as outlined above. Prior to and following solvation of each system, successive energy minimisations were performed using an SD algorithm, to minimise

steric overlap between the components of each system. Subsequent equilibration simulations allowed the solvent to relax around each restrained biomolecular complex; harmonic restraints were applied to all the protein non-hydrogen atoms, and gradually removed in three distinct 0.5 ns phases. The Ca^{2+} ion was unrestrained during the equilibration steps yet remained in the original bound position. Finally, production MD was executed in triplicate on the unrestrained systems, within the NPT ensemble.

All PulG simulations were performed using the GROMACS 4.5.5 simulation package[185], with the CHARMM36 all-atom FF[175] used to treat the lipid molecules and CHARMM22/CMAP parameters[175] used to treat the protein. The parameters for MPH were formulated by Dr. Peter Bond, based on existing parameterised fragments (see Appendix 1). Simulations were completed using the leap-frog algorithm with a 2 fs time-step, and trajectory data were collected every 10 ps. The LINCS algorithm[181] was used to constrain bond lengths and the neighbour list was updated every 10 steps. Long-range electrostatic interactions were corrected using the Particle Mesh Ewald method[184] and real-space electrostatic and van der Waals interactions were cut off at 12 and 10 Å, respectively. Simulations were performed at an absolute temperature of 310 K using the V-rescale thermostat. The pressure was set to 1 bar (Parrinello-Rahman semi-isotropic barostat, 5 ps coupling constant), under PBC.

Table 5.1 contains the nomenclature of the structures and simulations analysed in this chapter, for the clarity of the reader.

Table 5.1 – Systems analysed in this chapter

Name	System	Simulations
$PulG_{WT}$	POPE + PulG model monomer + N-terminal methylated Phe1 (MPH)	3 x 200 ns
$PulG_{MPH\text{-}E5A}$	POPE + PulG model monomer + MPH + E5 substituted by A5	3 x 350 ns
$PulG_{E5A}$	POPE + PulG model monomer + unmethylated Phe1 + E5 substituted by A5	3 x 350 ns
Dimer	POPE + unmethylated PulG model homodimer	3 x 100 ns

5.3 Results

To extend knowledge of T2SS secretion related to *K. oxytoca* PulG, the analyses presented here intended to answer several questions. The simulations aimed to clarify whether the N-terminus of the PulG E5A variant interacts with the membrane lipids but in the WT promotes intra-molecular interactions with the N-terminal MPH, stabilising the helical terminus. I examined whether the region around Pro22 was dynamic ("melted"), and the interactions of the calcium-binding loop with the bilayer. Following experimental demonstrations of the role of *P–P+1* contacts, I have used MD methods to study the dimer interface in motion.

5.3.1. Comparison of the dynamics of all PulG systems

There were no vast differences in Cα RMSD values or trends among the simulations of monomer variants (Figure 5.4). Interestingly, during the $PulG_{WT}$ simulations the RMSD of the Cα atoms in the helical segment (residues 2-54) was greater than that of the globular domain completely throughout two replicas, and frequently in the third. The RMSD of the helix oscillated between ~ 0.2-0.3 nm in one replicate (except a peak up to ~ 0.5 nm during the first 10 ns, as the system stabilised) and over a larger range, ~ 0.4-0.6 nm, in another. In the third replica, the helix RMSD remained ~ 0.5 nm. However, there were no abrupt changes in any $PulG_{WT}$ simulations, and comparison between RMSD datasets showed that this helix RMSD was due to the gradual bending motion of the protein within the bilayer and not caused by unwinding of the helical terminus. This supports the notion that the protein portion within the membrane may be able to contact the TMS of a chaperone such as PulM, and flexibility would promote the necessary contacts for PulG removal. In contrast, the RMSDs of the globular domain (residues 55-133) remained stable at ~ 0.2 nm in two simulations and ~ 0.35 nm in the third (during which the RMSD increased gradually over the first 100 ns from ~ 0.2 nm).

Substituting E5 was not anticipated to affect globular RMSD; one $PulG_{MPH\text{-}E5A}$ replicate mimicked the $PulG_{WT}$ results, with stable RMSD of ~ 0.2 nm. In the other replicates the RMSD remained stable around 0.2 nm for ~ 150 ns, although large

oscillations up to 0.5-0.6 nm were subsequently observed. Visual analysis confirmed that these fluctuations were due to a loop containing residues 70-80 moving away from the protein and flexing freely in solution. There was no great difference between the $PulG_{MPH-E5A}$ and $PulG_{WT}$ helix deviations; one $PulG_{MPH-E5A}$ replicate averaged ~ 0.4 nm throughout, with no abrupt fluctuations; another experienced an increase to ~ 0.5 nm after approximately 30 ns of simulation (following system stabilisation) and then oscillated with an overall downward trend to ~ 0.2 nm after 275 ns, followed by a second gradual increase in RMSD; the third replicate oscillated around 0.3 nm, experiencing one peak up to 0.5 nm around 170 ns, which visual analysis showed was due to the stretch of residues between Met25 and Lys35, located above the membrane and flexing as it is not supported by POPE molecules.

Two unmethylated $PulG_{E5A}$ replicates experienced the same extensive loop movements as the aforementioned $PulG_{MPH-E5A}$ replicate: the RMSDs remained stable at ~ 0.25 nm and ~ 0.2 nm for 100 ns and 290 ns of simulation, respectively, and then increased to ~ 0.5 nm after which the RMSD values of both plateaued. The third globular domain RMSD remained stable at ~ 0.2 nm throughout. The helical RMSD values of the $PulG_{E5A}$ replicates varied. In one simulation, the RMSD increased quickly to 0.5 nm and plateaued until 250 ns, when it increased gradually to 0.98 nm and then gradually decreased to 0.7 nm in the final 30 ns. In another, the RMSD averaged 0.35 nm except for an increase to ~ 0.5 nm between 180-280 ns. In the third replicate, more oscillations were observed overall, but the RMSD again remained in the range of 0.1-0.6 nm.

In the dimer simulations, there were no significant differences in RMSD between the two protomers, or between the domains within each (Figure 5.4B). In all three replicates, the RMSD of each domain remained between 0.1-0.4 nm, except for the globular domain of one protomer in one replicate which remained stable at higher values, between 0.4-0.6 nm. Overall, the α-helices appeared to be stabilised by their existence in a dimer compared to a monomer, regardless of the N-terminal variations. The RMSDs of the globular domains did not change noticeably depending on whether the simulation contained a monomer or dimer, which may be expected as the dimer simulations showed that 81 % of the contact interface between the protomers

was due to amino acids from the PulG α-helix (averaged across the final frames of the replicates), with no more than 5 globular domain residues closer than 3 Å to the other protomer.

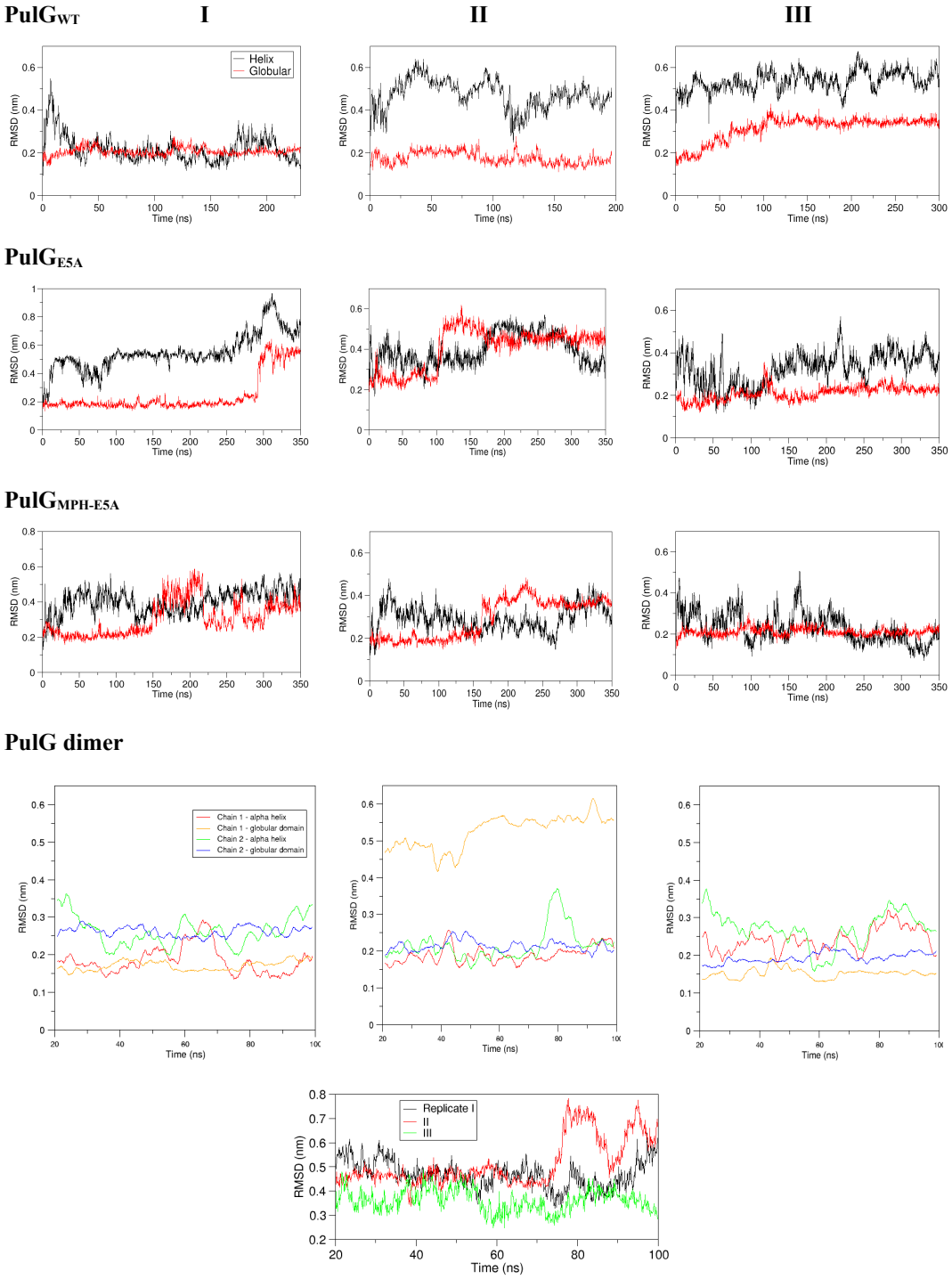


Figure 5.4 – RMSD of PulG protein throughout all simulations
Graphs showing the RMSD of the helical and globular domains of each PulG simulation relative to the initial structure over the course of each replicate in each system. The colours are shown in the respective legends.

The RMSF values of the Cα atoms of each residue were calculated for the final 50 ns of each monomer simulation and the final 20 ns of each dimer simulation (Figure 5.5). The RMSF values of two $PulG_{E5A}$ replicates displayed a peak between residues 70-80 (representative dataset shown in Figure 5.5A), corresponding to the segment of loops and coils at the top of the globular domain (Figure 5.5B) also responsible for the previously noted increases in globular RMSD. This peak was only observed in one of the $PulG_{MPH\text{-}E5A}$ simulations, and was observed to a much smaller extent in one $PulG_{WT}$ replicate. Dimerisation stabilised PulG, as demonstrated by the lower B factor values seen in Figure 5.5B, and only one PulG *P* monomer demonstrated an increased fluctuation of the loop between residues 70-80, among the dimer simulations. Notably, the RMSF and RMSD analyses did not indicate that the unmethylated protein was significantly less stable than the processed, as previously shown by experimental data, although B factor values demonstrated that the unmethylated E5A variant structure tended to fluctuate slightly more than the methylated variant.

Visual analysis confirmed that the peaks tended to occur when the protein loop did not settle onto the lipid bilayer, and instead moved freely in solution. In the $PulG_{E5A\text{-}MPH}$ and $PulG_{E5A}$ systems, there was a clear correlation between the bottom of the globular domain interacting with POPE, and a peak in B factor values. However, in the WT system this correlation was not observed, suggesting the loop movement in solution is, to an extent, random. The loop remained far removed from POPE in all the dimer simulations; the B factor values of the loop remained above 100 $Å^2$.

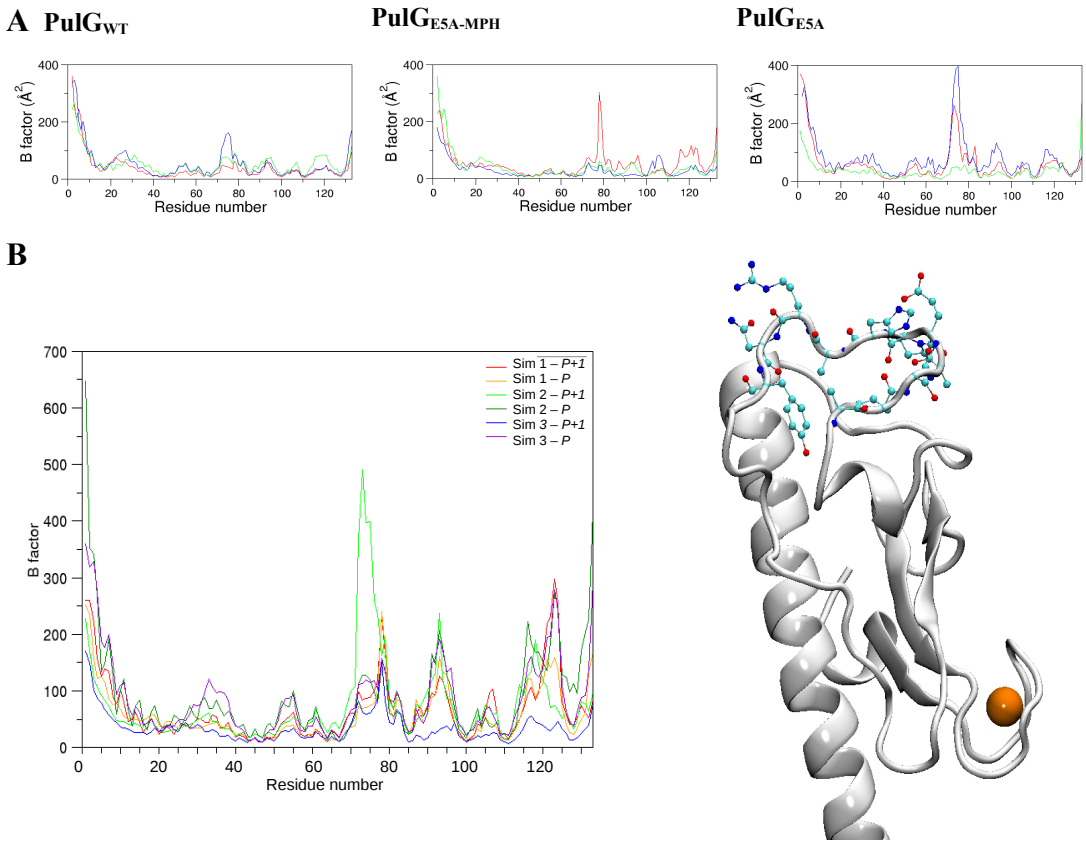


Figure 5.5 – RMSF of PulG
(A) Graphs of B factor values of PulG monomer simulation, in several cases displaying a peak between residues 70-80. **(B)** Left: B factor values for all three PulG dimer simulations; PulG proved more stable in a dimeric than monomeric conformation. Only one PulG *P* monomer (green line) demonstrated a notably large fluctuation of the loop between residues 70-80, among the dimer simulations. Right: Top of the PulG structure (shown in white) with residues 70-80 highlighted by displaying the residues in colour (C – cyan, O – red, N – blue) and the calcium ion shown as an orange sphere.

Key analysis of all the simulations included visual analysis; initial and final snapshots are shown in Figure 5.6. In all the monomer simulations, the protein appeared to become more deeply buried, and in all (except one PulG$_{E5A}$ replicate) the protein bent towards the lipid membrane surface to varying degrees. The globular domain interactions with membrane fluctuated; however the solvent-accessible surface area (SASA) of residues 60-125, the bulk of the globular domain, decreased on average between ~ 0.2 -3.1 nm over the course of the simulations. The standard deviations were not statistically significant, and variations in protein conformation and SASA suggest the globular domain does not maintain strong, consistent contacts with POPE. In the anomalous PulG$_{E5A}$ simulation, the bottom of the globular domain

nevertheless interacted more extensively with the lipid by the end of the simulation. In the dimeric system, the membrane was disrupted around the protein following equilibration, as seen in the initial snapshot. Following 100 ns of simulation, lipid head groups clustered around the dimer more closely yet the globular PulG domains did not interact as intensely as in the monomeric simulations, if at all. A kink around Pro22 that was not seen in the initial structure became visible in a monomer of one dimeric structure. Kinking at Pro22 was also observed in several monomer simulations, in particular all of the $PulG_{WT}$ replicates, warranting further investigation (detailed below in section 5.3.3).

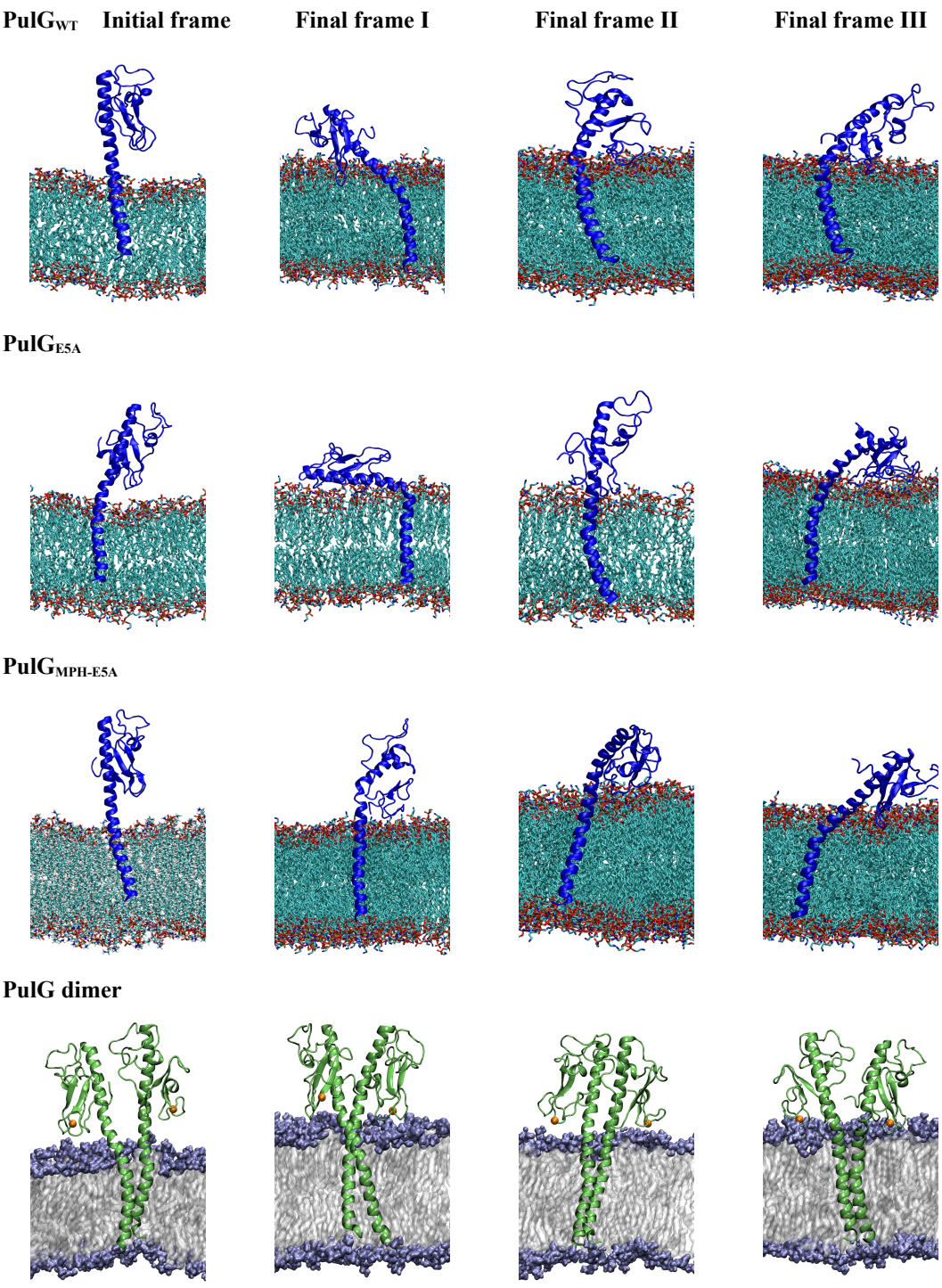


Figure 5.6 – Visualisation of initial and final structures from PulG simulations
Initial and final snapshots of all the systems presented in this chapter are shown below. The membrane representation allows the PulG α-helix to be observed clearly, and the lipid head groups and tails are differentiated by colour. PulG became more deeply buried in all monomer simulations, and bent towards the lipid. Lipid head groups clustered around the dimer more closely following the simulations yet the globular PulG domains did not interact as intensely with POPE as in the monomeric simulations, if at all.

5.3.2. Interactions of the N-terminus

On one hand, methylation reduces the capacity of the N-terminus for hydrogen bonding and the terminus may be expected to be less likely to interact with the E5 carboxylate oxygen atoms. On the other hand, it is possible that methylation increases the hydrophobicity of the N-terminus and therefore enables it to more easily enter the non-polar core of the membrane to interact with E5. I employed several analyses to identify which of these interactions occurred in the MD simulations, and answer the key question regarding the interaction between residues 1 and 5. Firstly I calculated the distance between the centre-of-mass (COM) of the lipid bilayer and each of the COMs of all the phosphate atoms in the lower bilayer, the first PulG amino acid and the fifth (representative dataset shown in Figure 5.7A). The protein moved around within the lipid bilayer, with the N-terminus COM between 0-9 Å from the solvent. The N-terminus remained closer to the solvent in the methylated simulations compared to the unmethylated, whereas the hydrophobic aromatic Phe tended to remain buried more deeply in the bilayer. This analysis gave an indication that E5 may not interact extensively with MPH, potentially disproving my initial hypothesis. The distance between residues 1 and 5 is largely fixed due to the helix on which they are both located, however the residue depth can vary more, depending on the angle of the helix in the bilayer, which explains why the difference between the data sets is not constant.

Attempting to corroborate this result, I examined the proximity of E5 to the N-terminus, employing visual analysis and calculating the minimum distance between the residue 1 methyl group or amide terminus (depending on whether the system contained MPH or Phe1) with any atom of the residue 5 side-chain (see Figure 5.7B). Interestingly, in the $PulG_{WT}$ system the MPH methyl group and the Glu5 side-chain remained consistently within 3 Å of each other. These results differed notably from those obtained using the simulations with substituted residues. In the $PulG_{E5A}$ system, the Ala5 side-chain remained 6 Å from Phe1, only approaching to within 5 Å in two of the simulations – and solely in the first 30 or 70 ns. The $PulG_{MPH\text{-}E5A}$ simulations showed more variation, with the atoms approaching to within ~ 2.5 Å and moving as far as 10 Å apart; however, the side-chains mostly remained ~ 6 Å apart. Visual

analysis supported these results; Glu5 and MPH remained close when possible, and Ala5 remained embedded in lipid whilst MPH was extensively solvated in the $PulG_{E5A}$ system. These data demonstrated that the original postulation was correct and the MD simulations suggest that E5 engages with the charge on the MPH residue, possibly reducing interactions and anchoring PulG less firmly in the bilayer, primed for removal during pseudopilus assembly. Performing the same analysis on the dimer simulation data showed that the N-terminus of Phe1 did not approach the side-chain of Glu5 to within less than ~ 4.8 Å (Figure 5.7B), consistent with PulG requiring MPH for the N-terminus to interact with E5 and stabilise the N-terminus of the helix. However, these results did not invalidate the dimer interface analysis (see section 5.3.4) as the absence of MPH did not affect the extended structure of the protein.

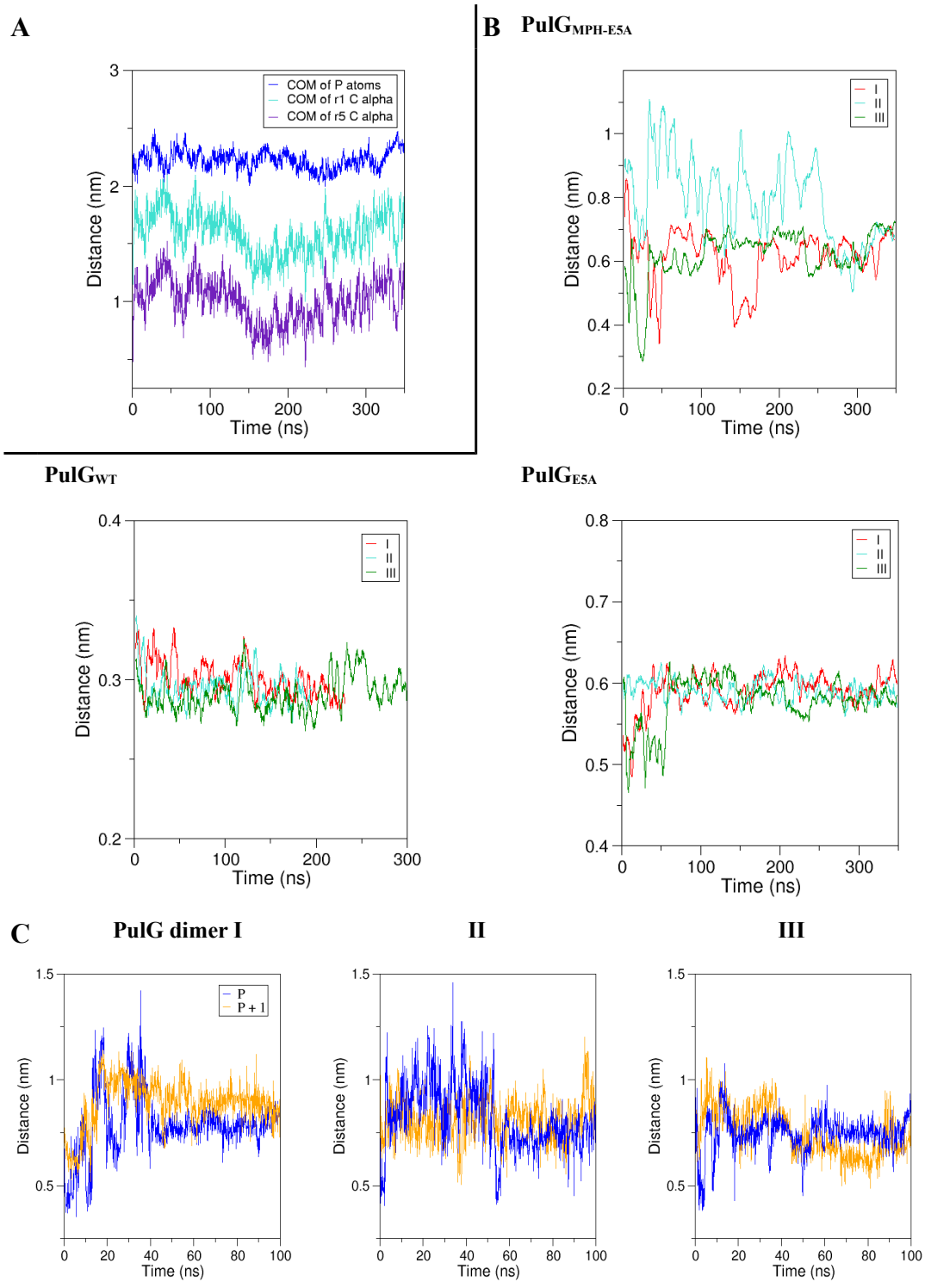


Figure 5.7 – Analyses of the intra-molecular distance between the N-terminus and residue 5

(A) Graph showing the distance between the COM of the lipid bilayer and each of the COMs of the phosphate atoms in the lower bilayer (blue), the first PulG amino acid (cyan) and the fifth (purple) in a PulG$_{E5A}$ replicate. Representative dataset. **(B)** Graphs showing the minimum distance between N-terminal methyl group of MPH and any atom of Glu5 side-chain over the course of the each monomer (note varying *y*-axis scales) and **(C)** dimer simulation.

Thirdly, I explored potential membrane disruption by studying the proximity of the solvent to the PulG terminus (Figure 5.8). Calculating the number of hydrogen bonds (length ≤ 3.5 Å) between residue 1 and solvent demonstrated that the WT terminus experienced extended periods, relative to the other systems, with no hydrogen bonds to solvent. Substitution of E5 increased the maximum number of hydrogen bonds by MPH to 4 and decreased the likelihood of a non-hydrogen-bonded state. The non-methylated E5A variant also formed a higher number of hydrogen bonds to water more frequently than the $PulG_{WT}$ system. Interestingly, relative to the other systems, the WT variant experienced extended periods with no hydrogen bonds to solvent and MPH formed 0.6 hydrogen bonds to solvent, on average. However, overall there was no statistically significant difference in the number of hydrogen bonds between each residue 1 and solvent, and the large standard deviations suggested solvation states fluctuated.

The PulG variants containing MPH formed on average ~0.9 hydrogen bonds with POPE, relative to the ~1.9 hydrogen bonds formed by F1 from $PulG_{E5A}$. When normalised relative to the number of possible hydrogen bonds from each, this analysis demonstrated that F1 engaged in hydrogen bonding to lipid 40% more than MPH. The role of methylation may therefore be to decrease the number of hydrogen bonds to lipid head groups, which may in turn reduce the energetic cost of transferring the charged terminus across the IM.

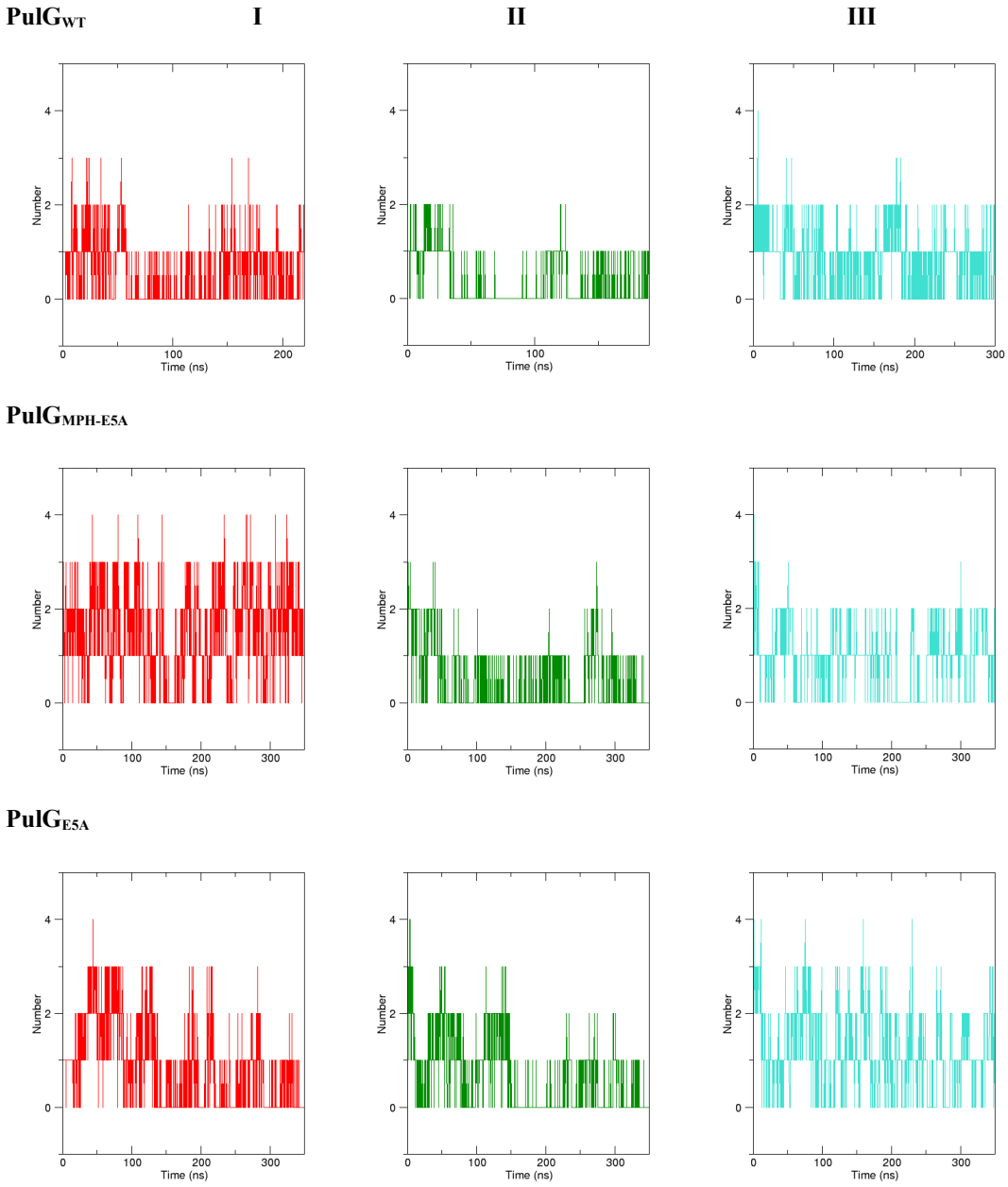


Figure 5.8 – Hydrogen bonding between residue 1 and solvent
Graphs showing the number of hydrogen bonds (cutoff radius = 0.35 nm, cutoff angle = 30 º) between PulG residue 1, whether Phe or MPH, and solvent. PulG$_{WT}$ experienced extended periods, relative to the other systems, with no hydrogen bonds to solvent. PulG$_{MPH-E5A}$ had a higher maximum number of hydrogen bonds formed by MPH (up to 4) and was less likely to inhabit a non-hydrogen-bonded state. The non-methylated E5A variant also formed a higher number of hydrogen bonds to water more frequently than the PulG$_{WT}$ system.

5.3.3 Calcium binding loop interactions

To ascertain the dynamics of the calcium binding loop, I analysed the location of the

calcium ion, included in the initial model prior to simulation and bound by Asp117 and Asp125, with Leu114, Val119 and Ser122 in close proximity. The ion remained in the binding site throughout all of the monomeric and dimeric simulations. The calcium binding loop consists of the following residues: Leu114 – Gly115 – Pro116 – Asp117 – Gly118 – Val119 – Pro120 – Glu121 – Ser122 – Asn123 – Asp124 – Asp125 (Figure 5.9). Of these, the hydroxyl and amide groups of Ser122 and Asn123 respectively can participate in hydrogen bonding. Leu114, Gly115, Pro116, Gly118, Val119 and Pro120 have hydrophobic side-chains, increasing the propensity of the loop to interact with POPE and repel solvent.

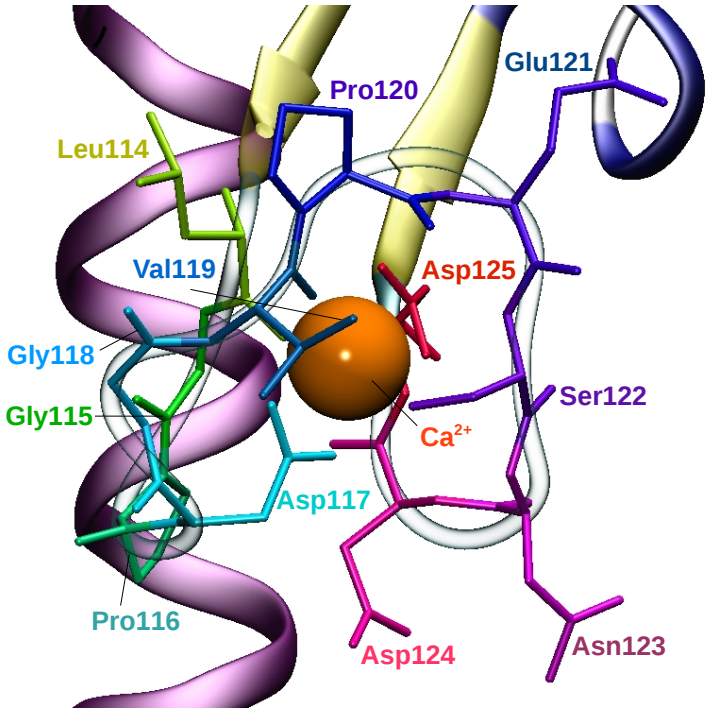


Figure 5.9 – Calcium binding loop
PulG calcium binding loop, consisting of residues 114-125 outlined as a transparent tube, coloured by residue with correspondingly coloured labels and with the Ca^{2+} ion shown as an orange sphere. The rest of the protein seen in the background is coloured by secondary structure.

The consistent number of hydrogen bonds between the side chains of Ser122 or Asn123 and POPE across the PulG simulations demonstrated that substitutions at the N-terminus do not significantly affect the interactions of these residues with the upper lipid bilayer. Serine tended to form more infrequent and sparse hydrogen bonds (maximum 2, usually 1, if any) with POPE than asparagine did, which tended

to form up to 4 bonds. There were no significant differences in the number of hydrogen bonds to POPE formed by each residue among the PulG monomer variants. Both Ser122 and Asn123 formed fewer hydrogen bonds with lipid in the dimer system, confirming previous analysis demonstrating the effect of dimerisation in keeping the globular domain upright and removed from the membrane. In the dimer system, Pro116, Asp117 and Asp124 of both monomers remained within 4 Å of the lipid throughout every simulation. In two replicates, Val119 also approached the membrane. Dimerisation, which otherwise constrained the movement of the globular domains, did not prevent interactions between the calcium-binding loop and lipid.

In all the monomer simulations, residues Pro116 and Asp117 approached the membrane consistently, as did Asp124. These residues are located at the bottom of the loop (see Figure 5.9), and polar asparagine contains COO^- and NH_3^+ groups that can interact with the lipid head groups. The proximity of Pro116, without a long charged side-chain, to POPE indicated the closeness of the entire loop to the bilayer surface. The entire loop contacted the membrane throughout one $PulG_{WT}$ replicate. However, in all other simulations, only residues 116-117/118 and 121-124 contacted the lipid, demonstrating loop relaxation onto the bilayer. These results confirmed that substitutions at the N terminus of the protein do not affect the calcium binding loop, and showed the role of polar type of residues in the interaction. This result may suggest a mechanism for PulG to remain in a favourable position for interaction with other T2SS proteins.

I compared the maintenance of the secondary structure of each protein over the course of each simulation, focusing on the N-terminus and the region surrounding residue Pro22. Pro22 is the only proline residue present in the helical domain. Figure 5.10 shows an example data set; the Timeline tool of VMD creates an interactive 2D box-plot of time *vs*. secondary structure of each amino acid, coloured by the secondary structure code listed in the legend of Figure 5.1. Examining how the secondary structure changes over time allows insight into the protein dynamics.

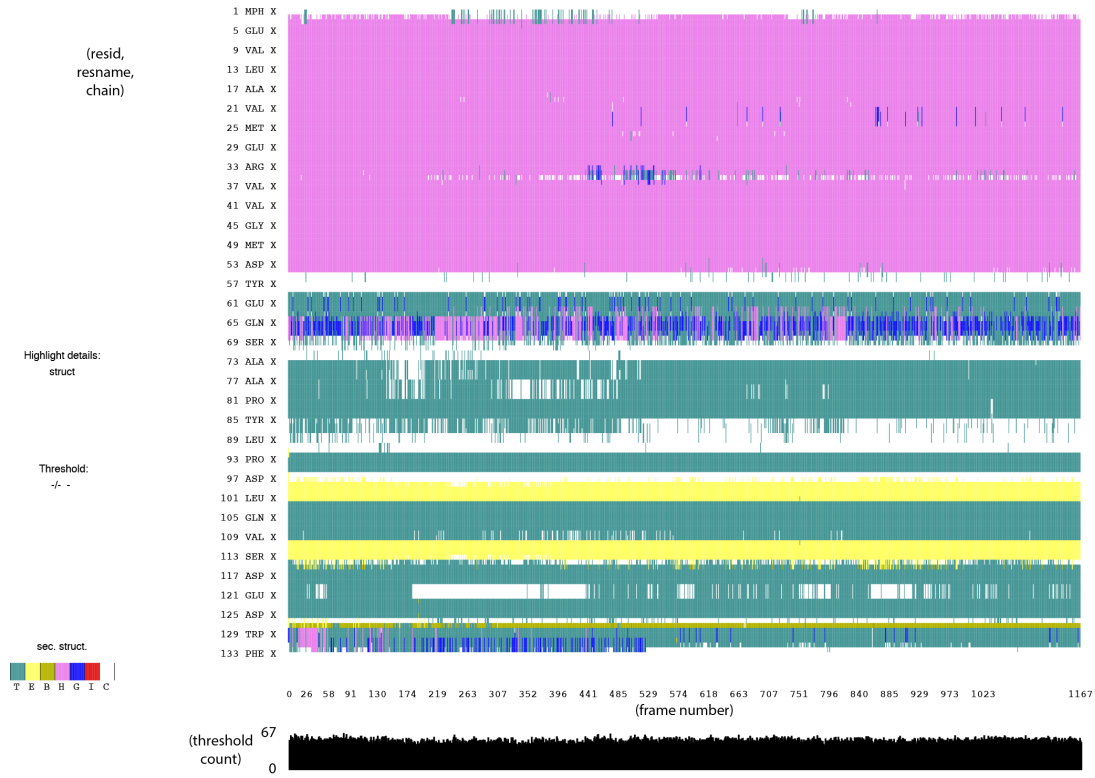


Figure 5.10 – Secondary structure maintenance
The Timeline tool of VMD creates a box-plot showing time *vs*. the secondary structure of each amino acid (α-helix – purple, 3-10 helix – blue, β-sheet – yellow, turn – cyan, coil – white). The N-terminal α-helix is almost entirely conserved throughout all the monomer simulations, as are the two β-sheets between residues 97-102 and 110-114. The area around Pro22 maintained an almost exclusively α-helical structure throughout and only fluctuated during several discontinuous frames in one $PulG_{WT}$ simulation, shown here. The first several residues demonstrated a conversion of helical structure to either turn or coil structures.

Secondary structure analysis showed that in the monomer systems, the N-terminal α-helix was almost entirely conserved throughout the simulations, as were the two β-sheets between residues 97-102 and 110-114. The largest disruptions to the α-helix were observed in two $PulG_{E5A}$ simulations (with residues 30-34, or solely residue 34, losing their helical structure) and two $PulG_{WT}$ simulations (where the helix disintegrated around either residues 27-30 or again around residue 34). These changes were transitory and in one simulation from each system, helix integrity was fully restored prior to the completion of the simulation. The structure around Pro22 did not demonstrate a systematic change in secondary structure in any monomeric simulation, maintaining almost exclusive α-helical structure throughout and only fluctuating during several discontinuous frames in one $PulG_{WT}$ simulation (this

dataset is shown in Figure 5.10), although kinking was observed visually in all three $PulG_{WT}$ replicates. Interestingly, in all $PulG_{WT}$ simulations the first several residues demonstrated a conversion from helical structure to either turns or coils, suggesting that MPH and E5 destabilised the helix terminus as they deformed to interact with each other, which may have subsequent functional implications. The interaction caused an N-terminal "loop" to form (Figure 5.11A). In the $PulG_{E5A}$ system, the five N-terminal residues maintained their helical structure throughout, and in one $PulG_{E5A\text{-}MPH}$ simulation six N-terminal residues transformed from an α-helix to occupy a 3-10 helix or turn structure between ~ 150-200 ns. These results suggest that the charged MPH residue and Glu5 play a role in N-terminal stability.

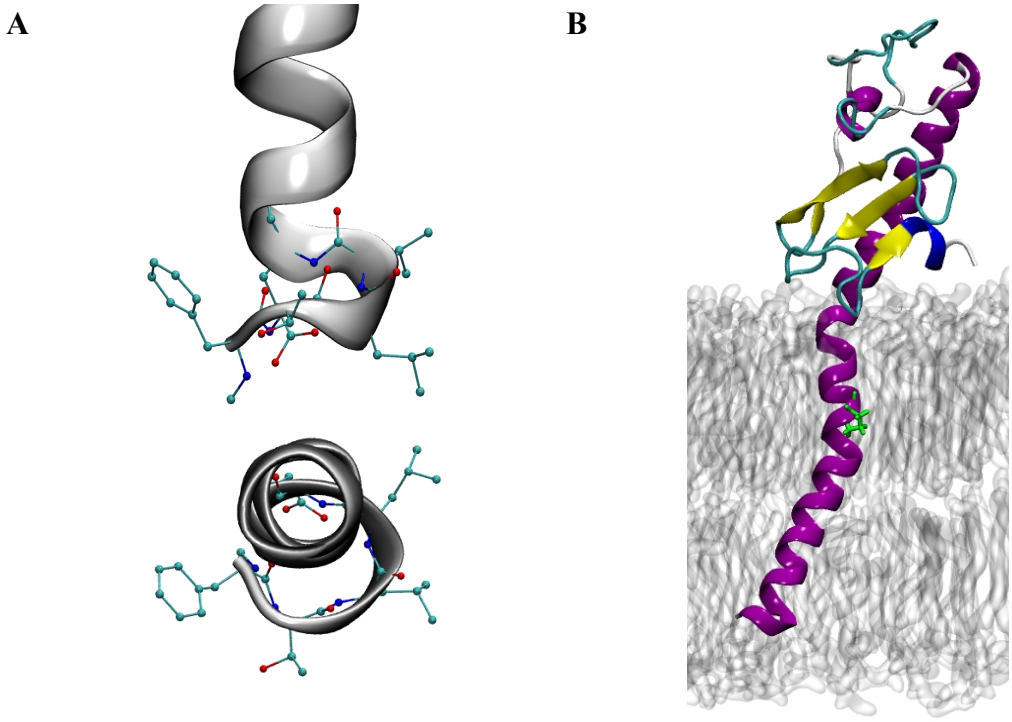


Figure 5.11 – Helix deformations
(A) The N-terminus is deformed by MPH and Glu5 interacting. This loss of α-helical structure is shown from the side and from above, with the protein shown as a silver helix and the first five residues highlighted by a stick representation coloured by atom (C – cyan, O – red, N – blue). **(B)** Kinking of the helix surrounding Pro22 (green) is shown here, with the helix bending to the left in the image whilst located within the membrane, and to the right at the membrane-solvent interface.

5.3.4 Dimer interactions

Several statements from the study by Nivaskumar *et al.* were tested to confirm the validity of the simulations. The dimer was constructed by simulating the bottom two monomers of the oligomeric structure modelled by Campos *et al.,* and represented protomers *P* and *P+1*. The number of hydrogen bonds between the two TM regions, and the two globular domains, was taken as an indication of dimer stability, and showed that, following a period of equilibration as the number of hydrogen bonds increased, the dimer conformation was indeed maintained throughout all three replicates with the number of bonds fluctuating gradually between 0 and 8 (Figure 5.12).

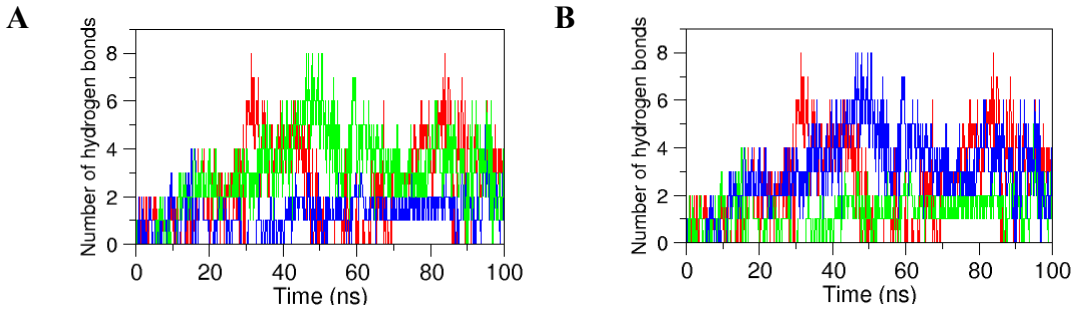


Figure 5.12 – Hydrogen bonding between dimer components indicates stability **(A)** Number of hydrogen bonds between the helical domain residues 1-54 in replicates 1 (red), 2 (green) and 3 (blue). **(B)** Number of hydrogen bonds between the globular domain residues 55-133 in replicates 1 (red), 2 (green) and 3 (blue).

Calculating the distance between each residue of one monomer and any part of the adjacent monomer gave an indication of which residues are important in the dimer interface (Figure 5.13A). The trends were very similar across all three replicates, and the positions of the interface residues described below are shown in Figure 5.13B. The first 50 residues of each monomer approached closer than 3 Å to the other protein, due to the aligned helices, with the exception of residues (*P+1*)40-48 in two replicates. The oscillating nature of the dataset reflected the helical protein structure (residues nearer the other monomer are followed by those located on the distal side of the helix, and then again by residues on the proximal side). The globular domain contacts of the *P* monomers were near identical, with residue 86, residues 105-108, and the final 6-8 residues approaching *P+1*. Residues 105-108 are located on a loop

between the two large β-sheets of the globular domain, and the C-terminus of the protein floated freely in solution, allowing inter-protomer interactions. None of the final 33 residues of *P+1* approached close to *P* as they faced the other way, and so are likely to interact with *P+2 in vivo*. However, *P+1* residues 80-90 and even up to residue 100, forming the loop closest to *P*, approached *P*, as did Val68 which is found on the second closest loop and migrated to interact with *P*.

Residues 28-35 interacted most closely with the adjacent monomer; this section of PulG contains the greatest cluster of acidic, basic and polar residues with Asn27, Lys30 and Gln34 of *P*, and Lys28, Asp32 and Lys35 of *P+1* forming the main interprotomer contacts. However, it is also possible for Lys28 to form contacts with the adjacent Glu29, reducing the dimer interface interactions, as observed during one replicate of the dimer system. In the pseudopilus structure modelled by Nivaskumar *et al*., K28 and K35 interact with (*P+3*)E5, suggesting the interactions observed in the dimer simulations may be disrupted *in vivo* by further oligomerisation, as the dimer joins the extending pseudopilus.

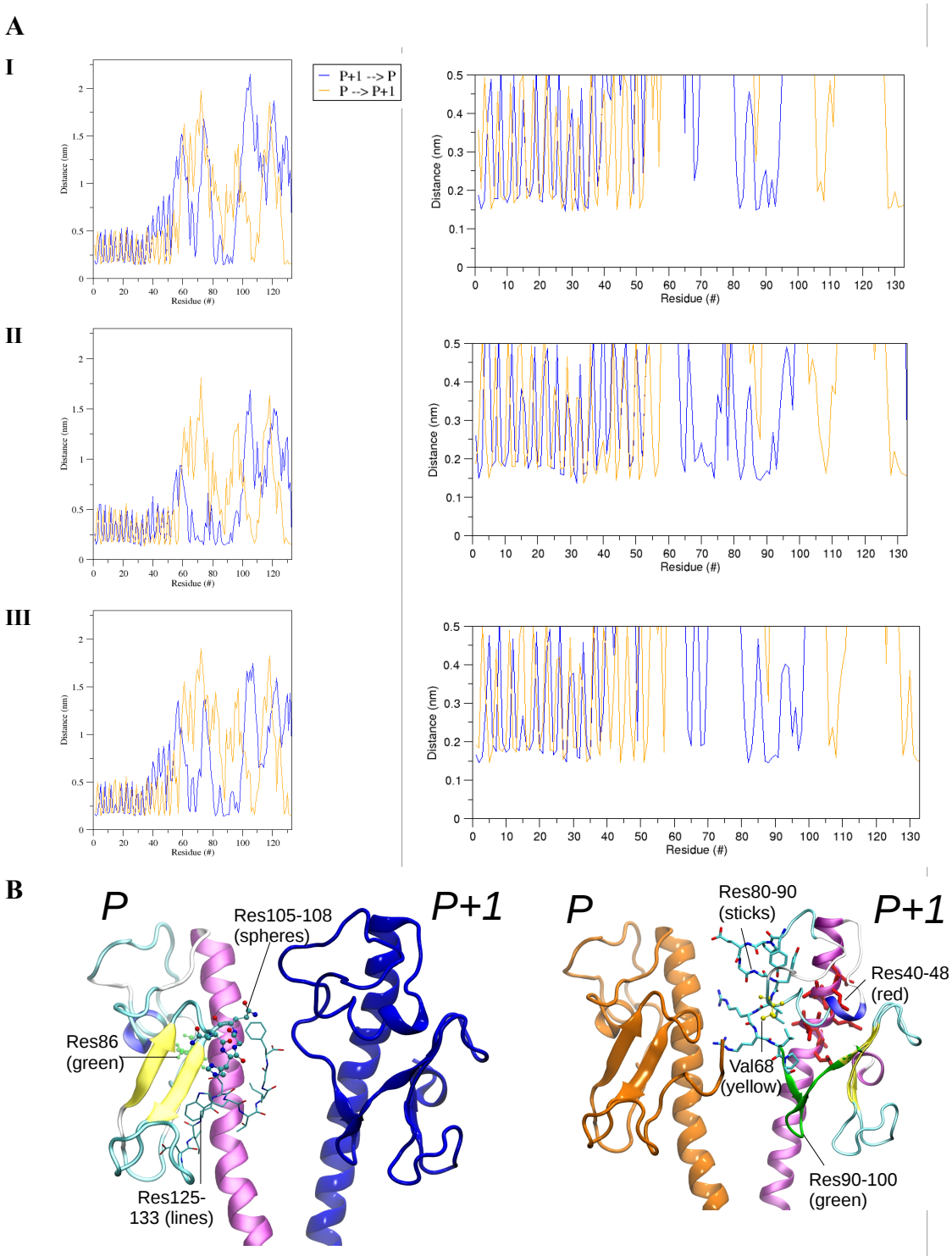


Figure 5.13 – PulG homodimer interface
(A) Graphs showing the distance between each residue of one monomer and any part of the adjacent monomer, giving an indication of which residues are important in the dimer interface. The distance between monomer *P* residues and monomer *P+1* is indicated with an orange line, and the distance between *P+1* residues and monomer *P* with a blue line. Left: full datasets. Right: enlarged datasets of distances relevant to interactions. **(B)** The positions of the described interface residues of the globular domain are highlighted and labelled.

The studies by Campos *et al.* and Nivaskumar *et al*. made several predictions regarding the dimer interface. Contacts between (*P*)16 and (*P+1*)9 or (*P*)16 and (*P+1*)11 were identified as mutually exclusive and indicated whether the protein conformation occupied the lower energy basin A or higher energy basin C, respectively. (*P*)16 and (*P+1*)10 were expected to interact regardless of which energy basin PulG inhabited. (*P*)D48 and (*P+1*)R87 were also predicted to be very close, as expected due to formation of the salt bridge. I performed a series of analyses in light of these predictions. (*P*)16 and (*P+1*)10 tended to remain within 2-3 Å of each other, occasionally moving up to 5 Å apart, particularly in the second and third replicates. In the first two replicates (*P+1*)11 usually stayed ~ 1 Å closer to P16 than (*P+1*)9 did, suggesting the dimer was more similar to the structures that Campos *et al.* calculated occupy the lower energy basin. This effect was especially pronounced in the third replicate, where (*P+1*)11 remained within ~ 3 Å of (*P*)16 throughout the final 40 ns of simulation. In contrast, during this period (*P+1*)9 remained > 5.5 Å from (*P*)16. These results clearly indicated the conformation is low energy and validated the use of MD simulations to explore this dimer structure.

Following prediction of the salt bridge between (*P*)D48 and (*P+1*)R87, Nivaskumar noted that these residues are very close in the static dimer structure. Simulation analysis (see Figure 5.14C) showed that their proximity is variable, in one case consistently maintaining contact (with small increases in distance up to 4 Å) and in another remaining more than 6 Å apart at all times after the first 25 ns of simulation. This result indicated that the salt bridges can indeed form and the residues maintain proximity, however the bonded state is likely supported by the presence of further monomers that may hold the PulG globular domains in place and facilitate salt bridge formation.

Nivaskumar states that residues (*P*)5 and (*P+1*)1 do not interact; however in the dimer simulations the E5 carboxylate atoms approached, in two of the three replicates, to within 3 Å of the Phe1 nitrogen atom. In the first, after 30 ns the residues remained approximately 4.5 Å apart. In the second, the residues remained 4-5 Å apart, except for a 20 ns period between 45-65 ns when the residues interacted and remained within 3 Å. In the third replicate, the residues remained over 6 Å apart

at all times, as predicted. The results of this analysis are shown in Figure 5.14B. As this dimer was not methylated, these results indicate that it may be possible for non-methylated Phe1 to interact with the N-terminus of the adjacent monomer, in the absence of an MPH on the same monomer.

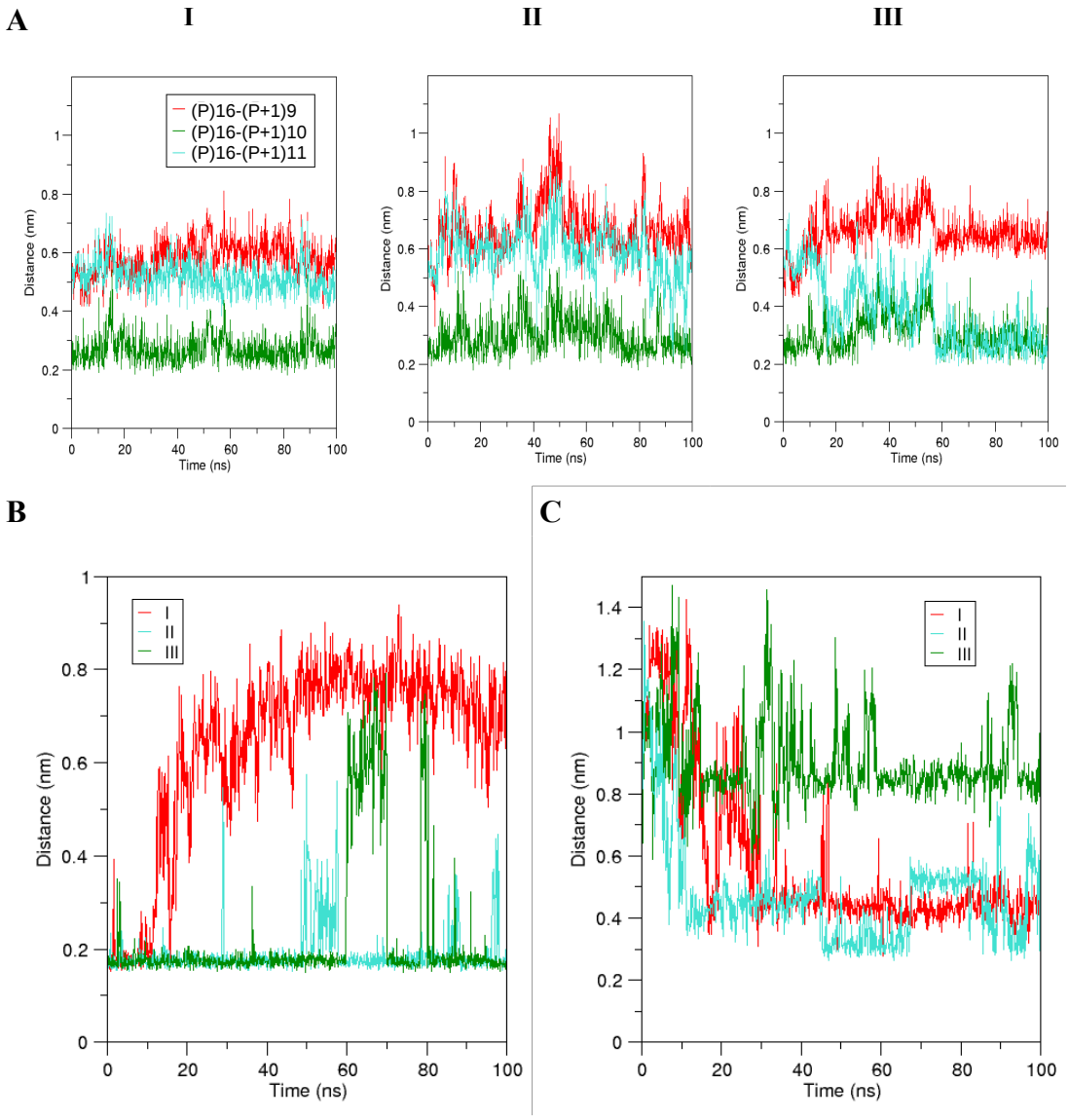


Figure 5.14 – Proximity of predicted interacting residues in the PulG α-helix
(A) Graphs showing the proximity of (*P*)16 to (*P+1*)9, (*P+1*)10 and (*P+1*)11. Aside from frames during which the helices moved apart, (*P*)16 and (*P+1*)10 remained within 2-4 Å of each other (green line). Throughout the first two replicates, both (*P+1*)9 (red line) and (*P+1*)11 (blue line) remained a similar distance (~ 4-8 Å) from *P*(16), with (*P+1*)11 usually ~ 1 Å closer. In the third replicate, after ~ 15 ns of simulation, (*P+1*)11 approached closer to (*P*)16 and fluctuated between 2-6 Å away until stabilising after 60 ns in total around at the same close proximity as (*P+1*)10. **(B)** Graph showing the proximity of residues (*P*)5 and (*P+1*)1. As shown by the distance between the atoms, in two out of three simulations the (*P*)E5 carboxylate group approached to within 2 Å of the (*P+1*)Phe1 nitrogen atom. In the third, the residues remain over 5 Å apart, as predicted. **(C)** Graph showing the proximity of residues (*P*)48 and (*P+1*)87. Variable distances were observed, suggesting the predicted salt bridges may occur but inconsistently.

5.4 Discussion

This extensive study aimed to answer a number of questions, and overall the obtained data shed light on the dynamics of the major pseudopilin in *K. oxytoca*, and its interactions both with and within a lipid bilayer. These simulations mimicked the protein periplasmic state, prior to incorporation of PulG into the T2SS pilus.

To date, only static structures of monomeric and oligomeric PulG structures had been studied; this work provided an opportunity to examine the Type II major pilin in motion. Both the monomer variants and dimer structure remained buried in the POPE bilayer into which they were placed, and no abrupt changes in structure were observed, although the globular domain tended to approach the bilayer. In all cases, both the helical and globular domains tended to deviate slightly from the initial structure, however extensive fluctuations were only observed upon free movement in solution of the loop containing residues 70-80, located on the top of the protein. This may have functional relevance, promoting availability of these loop residues for interactions with chaperone proteins (Ser72, His76, Asn79 and Tyr80 can all participate in hydrogen bonding). In contrast, the helical portion of the protein was stabilised in the presence of another PulG monomer, due to the dimeric interface, with (*P*)Lys30 and (*P+1*)Asp32 playing a key role in the middle section of the helix. Notably, the RMSF and RMSD analyses did not show that the unmethylated protein was significantly less stable than the processed, as previously indicated by experimental data. Simulation analysis suggested that this instability may be due to interaction with chaperones or removal by periplasmic proteases, rather than directly due to structural effects of the post-translational modification.

Linking the interactions of the N-terminus to function was of key interest in this study. Glu5Asp substitution conserves the negative charge, allowing partial pilus assembly, but does not complement secretion, indicating that the role of this residue amounts to more than conserving charge. The simulations presented here clarified that the presence of Glu5 allows the residue to interact with the methylated N-terminal Phe residue; in the $PulG_{WT}$ system the methyl group of MPH and the Glu5 side-chain remained consistently within 3 Å of each other, unlike in the variant

systems. This interaction caused a loop to form, affecting the secondary structure of the WT N-terminus.

The sole calcium ion present in the PulG structure remained bound in a loop by Asp117 and Asp125 throughout all simulations, despite being unconstrained. Pro116 and the calcium-binding residues consistently approached the POPE bilayer in the monomer simulations, assisted by their position in the globular domain closest to the lipid. This calcium ion has been shown to be necessary for secretion by GspG in *V. cholerae*, and a future direction for this work would be to examine the effect of calcium cation absence on the PulG globular domain fold.

Pro22 has been observed to "melt" during previous experimental work by Edward Engelman, with the surrounding area losing helicity. Secondary structure analysis of the simulations demonstrated that this section of protein did not lose helical secondary structure in the MD simulations, although kinking could be observed visually in all three $PulG_{WT}$ replicates. Research in the field now turns to the role of chaperones such as PulM in the removal of the major pseudopilin from the membrane, and it is not outside the realm of possibility that the flexibility of the α-helical section promotes binding between PulG and a chaperone. Notably supporting this, a P22A substitution has been shown to adversely affect the efficiency of oligomer assembly[89]; further research is required to analyse the interaction between PulG and PulM or PulL.

The dimer interface consists of two clear regions; the first 50 residues of each protomer, and the globular domain interface between (*P*)105-108,125-133 and (*P+1*)40-48,80-100. Analysis of the contacts between (*P*)16 and (*P+1*)9, (*P+1*)10 and (*P+1*)11 demonstrated that the dimer structure occupies a low energy state as (*P*)16-(*P+1*)10 and (*P*)16-(*P+1*)11 maintained close contacts. These results corroborated data obtained by Campos *et al*. MD analysis identified residues Asn27, Lys30 and Gln34 of *P*, and Lys28, Asp32 and Lys35 of *P+1* as providing key contacts in the central segment of PulG α-helix, and experimental substitutions of these residues could be performed to support these results. The model generated by Nivaskumar *et al.* suggests that Lys28 and Lys35 interact with (*P+3*)E5. Therefore, it

is possible that the interactions observed in this simulation are only present in the pre-assembled dimer, and are disrupted upon PulG oligomerisation. The simulated dimer contained unmethylated F1 and therefore future work would require further simulations to investigate whether methylated Phe1 on $PulG_{WT}$ in the membrane interacts with E5 and aids piliation, possibly by affecting the binding energy of the system or the energy required to extract $PulG_{WT}$ from the lipid for pseudopilus assembly.

Notably, *in vivo* the cytoplasmic side of the IM is negatively charged, therefore had no difference had been observed in the simulated lipid interactions of $PulG_{WT}$ and $PulG_{E5A}$, the explanation could have been linked to the membrane charge found in bacteria. In order to extend this study and simulate a more physiologically accurate environment, a "double-bilayer" system, in which the solute between the two bilayers is treated differently, could be employed. Extension of this study would include modelling and simulation of PulM, the proposed chaperone for PulG into the T2SS machinery, to identify residues key to the interaction and in the long term suggest a mechanism for hindering Type 2 secretion.

Chapter 6 – Coarse-grained studies of PulM interactions

Disclaimer
A portion of the non-computational experimental work described in this chapter was carried out by Olivera Francetic and colleagues, of the Institut Pasteur, Paris, and included in a manuscript accepted for publication in 2017, under the title "Polar N-terminal residues conserved in type 2 secretion pseudopilins determine subunit targeting and membrane extraction steps during fibre assembly". Parts of the text in section 6.1 were therefore written in collaboration, however all figures and remaining text are my own work, unless stated otherwise in the legend.

6.1 Introduction

As previously stated, PulG is known to oligomerise into the stem-like pseudopilus of the T2SS that subsequently secretes PulA from the bacterial cell, *via* an unknown mechanism. PulG is an IM-embedded protein that has been shown to interact directly with IM assembly platform (AP) proteins, such as PulM, PulL and PulF. Research has indicated that PulM may putatively chaperone PulG to the AP complex: interacting with the pseudopilin in the IM, moving it to the secretion machinery, and facilitating oligomerisation. The *in silico* study presented in this chapter primarily aimed to identify the likely PulG-PulM interface. The *K. oxytoca* PulM variant, composed of 161 amino acids, has not been crystallised to date, although the protein sequence has been determined[60]. As a result, the overall structure and resulting dynamics remained unknown prior to this study; MD simulations require an initial structure and therefore modelling PulM was a necessary primary step.

Homologous structures have come to light; crystal structures of *Pseudomonas* PilO[107] and EpsM from *V. cholerae*[104] have been obtained, and a cryo-electron tomographic structure of the piliated Type IVa IM complex from *M. xanthus* has recently been published[234]. PilO residues 23-45 were predicted to be TM, supported by the presence of a series of hydrophobic residues between residues 32-51, and the C-terminal domain was anticipated to be periplasmic. The PilO globular domain, crystallised from $PilO_{\Delta 68}$ and resolved to 2.2 Å (PDB ID: 2RJZ), consists of four helices joined by coiled coils, and a four-stranded anti-parallel β-sheet. The crystallised 1.7 Å resolution structure of the periplasmic domain of EpsM from *V. cholerae* (PDB ID:

1UV7), pieced together from fragmented variants crystallised separately, likewise contains an anti-parallel β-sheet with 4 strands, yet only two helices. The PilO and EpsM structures are shown in Figure 6.1A.

The landmark study using cryo-electron tomography presented for the first time a 3-4 nm resolution map of the entire T4P machinery by imaging *M. xanthus* variants with individual complex proteins missing or fused to green fluorescent protein, and combining the data with known information regarding the structures, localisation, accumulation and incorporation of the individual proteins into the T4P complex. This study demonstrated the presence of two protein rings, the lower-periplasmic and mid-periplasmic, located in the T4P assembly complex between the IM AP and OM secretin pore. Imaging a PilO variant tagged with a fluorescent fusion protein suggested that PilO localises to the lower-periplasmic ring, although higher resolution structural data could not be obtained using this method. Notably, PulG was not observed to interact with PulM in either the piliated or non-piliated structures, suggesting that either the interaction may occur during assembly of the IM platform, or dynamic changes may occur during nano-machine activity that bring the two together, or both. The predicted secondary structure of *K. oxytoca* PulM consisted of three helices of varying lengths, separated by two coils, and four β strands in the globular domain, separated by four coils and two helices. This is shown in Figure 6.1B, alongside the PulM model that was created for this study in collaboration with Dr. Peter Bond, Dr. Olivera Francetic and colleagues (see section 6.2, Methods).

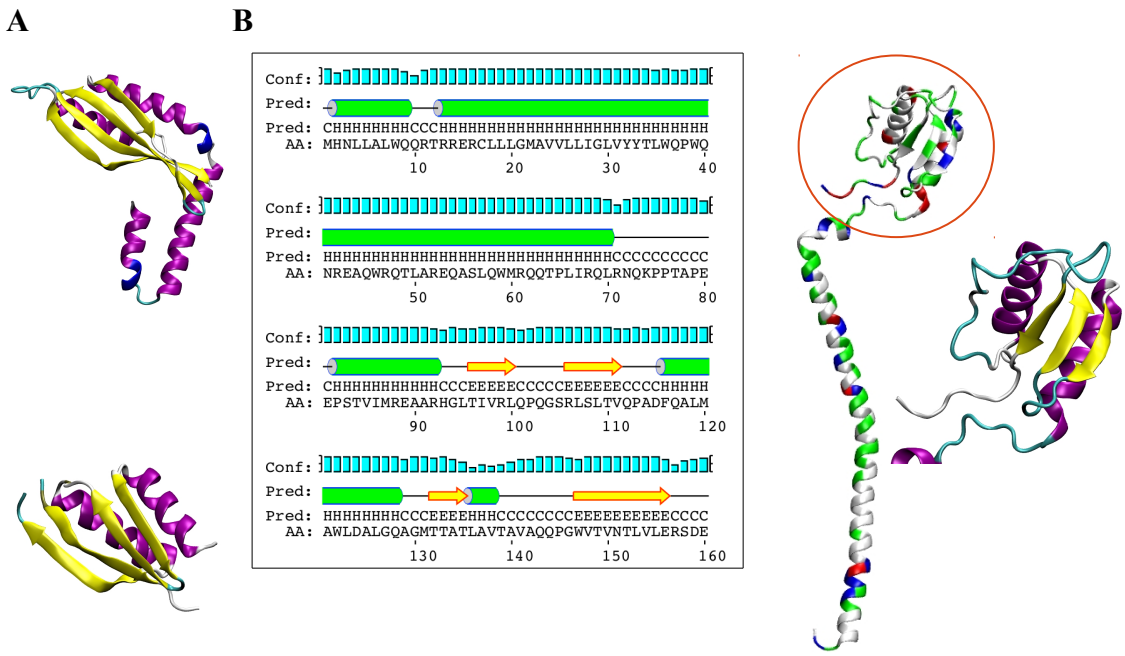


Figure 6.1 – PulM structure and homologues
(A) Crystal structures of *Pseudomonas* PilO (top) and *V. cholerae* EpsM (bottom), both coloured according to secondary structure (α-helix – purple, β-sheet – yellow, turn – cyan, coil – white). **(B)** Left: Prediction of PulM secondary structure using psipred[235]. The helices are shown as green cylinders, strands as yellow arrows, and coils as black lines. The height of the cyan bars relative to the "conf" brackets demonstrates the confidence of the prediction. "Pred" shows the predicted sequence (H – helix, E – strand, C – coil), and "AA" is the target sequence. Right: The resulting PulM model used in this study, coloured according to residue type (non-polar – white, basic – blue, acidic – red, polar – green). The globular domain is circled in red, and shown enlarged to the right, coloured according to secondary structure.

As previously mentioned, T4Ps are the best-studied of the pilial systems and research has repeatedly provided evidence of T4 major pilin interactions with AP components. For example, the *Pseudomonas* study above provided co-expression and biochemical evidence for direct interactions between the periplasmic domains of PilO (PulM homologue) and PilN (a homologue of the PulL TMS and periplasmic domain), which appear to be crucial for T4P function[107]. Likewise, a previous extensive study of *pilM/N/O/P* variants (PilM is a homologue of the PulL cytoplasmic domain) using stability and complementation analyses suggested a complex of all four is required for optimal OM secretin function[105]. Also, in *P. aeruginosa*, full-length pilin PilA co-eluted with a putative trans-envelope complex containing PilN/O/P and the PilQ secretin (homologous to PulD)[106]. Bacterial two-hybrid (BAC2H) analysis, a type of

protein-fragment complementation assay, has proven particularly useful in such studies. It has shown, for example, that the meningococcal major pilin PilE can interact individually with each of PilG (the PulF homologue), PilO (PulM homologue) and PilN (PulL homologue) – key AP components[102]. Notably, this PilE-PilO interaction is apparently mediated by the globular periplasmic domains of both proteins and the 39 N-terminal residues of PilE, with implications for the putative homologous PulG-PulM interface. A truncated version of the *Thermus thermophilus* major pilin PilA lacking the N-terminal hydrophobic α-helix interacted *in vitro* with both PilM/N (homologous to the cytoplasmic, and TM/periplasmic domains of PulL, respectively) and PilM/N/O complexes[103]. This indicates the likely key role of the globular domain in the interface.

Research into AP interactions was recently extended to the T2SS; extensive analyses performed by Dr. Olivera Francetic and colleagues demonstrated that PulM is necessary for both pilus formation and PulA secretion[236]. PulM and PulG have been shown to interact directly and independently of the other T2SS components, and BAC2H has shown that the highly conserved PulG E5 residue (see Chapter 5) is the key determinant of the PulG-PulM interaction. Biochemical and quantitative immunofluorescence studies have focused on the role of PulM, and demonstrated that *ΔpulM* variants exhibit a severe defect in pseudopilus assembly and do not secrete PulA under physiological conditions[236]. Formaldehyde cross-linking also showed that PulG and PulM interact in the context of the complete T2SS machinery (Santos-Moreno *et al.*, accepted). Following the insights outlined in Chapter 5 regarding the key N-terminal residues, interactions of PulG variants with PulM have been examined. Notably, those exhibiting impaired PulM interaction were defective in both pseudopilus assembly and PulA secretion, signalling the necessity of this step for a functional T2SS; for example, the $PulG_{E5A}$ variant was fully defective but $PulG_{T2A}$ supported partial function, providing insight into the relative importance of these N-terminal PulG residues for the interface (Santos-Moreno *et al.*, accepted).

Studying an unknown interface between two elongated proteins benefits from coarse-grained MD simulations that enable larger systems to be simulated on extended time-scales (typically multi-microseconds), relative to atomistic simulations (typically

hundreds of nanoseconds). Notably, the interaction potentials are much "softer" in CG simulations, simplifying the energy landscape and increasing the ease with which energy barriers are overcome, therefore biological time-scales are quickly reached. However, CG simulations do over-simplify specific interactions, hence the use of atomistic simulations to examine the roles of MPH and E5 previously. CG was an appropriate choice for initial analysis of PulG-PulM contacts, and this study aimed to clarify several questions surrounding this set of interactions. Firstly, what is the behaviour of monomeric PulM in the IM? Secondly, what is the likely interface between PulG and PulM, and which residues are key in maintaining these contacts? And finally, does this putative interface change in the physiologically expected trimer, namely depending on the presence of a PulG monomer or dimer adjacent to PulM? Although PulM can form homodimers, the functionally relevant state is predicted to be a heterodimer with PulL, which is favoured over homodimerisation by either protein. Functional homo- and heterodimer states may alternate, perhaps triggered by interactions with another partner (e.g. PulG). Therefore it was decided to simulate a PulM monomer, and not a homodimer, in the lipid bilayer. PulG is known to dimerise in the IM and is believed to interact with PulM in a dimerised state, explaining how both WT and E5A variants can be incorporated into pseudopili (at least one protein in the dimer carries the PulM-interacting E5, allowing incorporation). Therefore a trimeric system was also simulated, to provide physiologically relevant data regarding possible PulM-PulG interactions.

6.2 Methods and Details of Presented Coarse-grained Simulations

Five CG systems were simulated to generate data for this study: a PulG monomer, a PulM monomer, a PulG homodimer, a PulG-PulM heterodimer, and a trimer containing both a PulG dimer and a PulM monomer – each embedded in a lipid bilayer. The systems are outlined in Table 6.1 at the end of this section, for the benefit of the reader. The *E. coli* CG lipid membrane and lipid parameters were produced by Tom Piggot, University of Southampton (Hsu *et al.*, 2016, submitted to *J. Phys. Chem. B*). Structures for the PulG WT monomer and PulG dimer were taken from the studies in Chapter 5. An integrative structural model for PulM was generated in collaboration with Dr. Peter Bond, Dr. Olivera Francetic and colleagues,

using the program Modeller[237] and a combination of the following: secondary structure predictions based on the known PulM sequence; the *P. aeruginosa* homologous PilO structure (PDB ID: 2RJZ) for the globular domain; the *M. xanthus* Type IVa homologue PilO (PDB ID: 3JC8) for the TMS; and *ab initio* modelling for the linker between the two. The modelling method is represented visually in Figure 6.2A. Ten sets of 50 ns atomistic simulations of PulM in a POPE bilayer were run to equilibrate the model, using the CHARMM36 FF, and the structure for the CG simulations was chosen from a representative frame within these simulations. Secondary structure analysis demonstrated that the PulM structure remained largely stable throughout the atomistic simulations; Figure 6.2B contains a representative time-line and shows the four β-strands and three α-helices mostly retaining their structure, and the fluctuating unstructured regions. Likewise, visual analysis demonstrated the flexibility of the PulM protein (Figure 6.2B), with certain replicates showing kinking of the helical domain and significant rotational movement of the globular domain due to the unstructured linker consisting of residues 72-79.

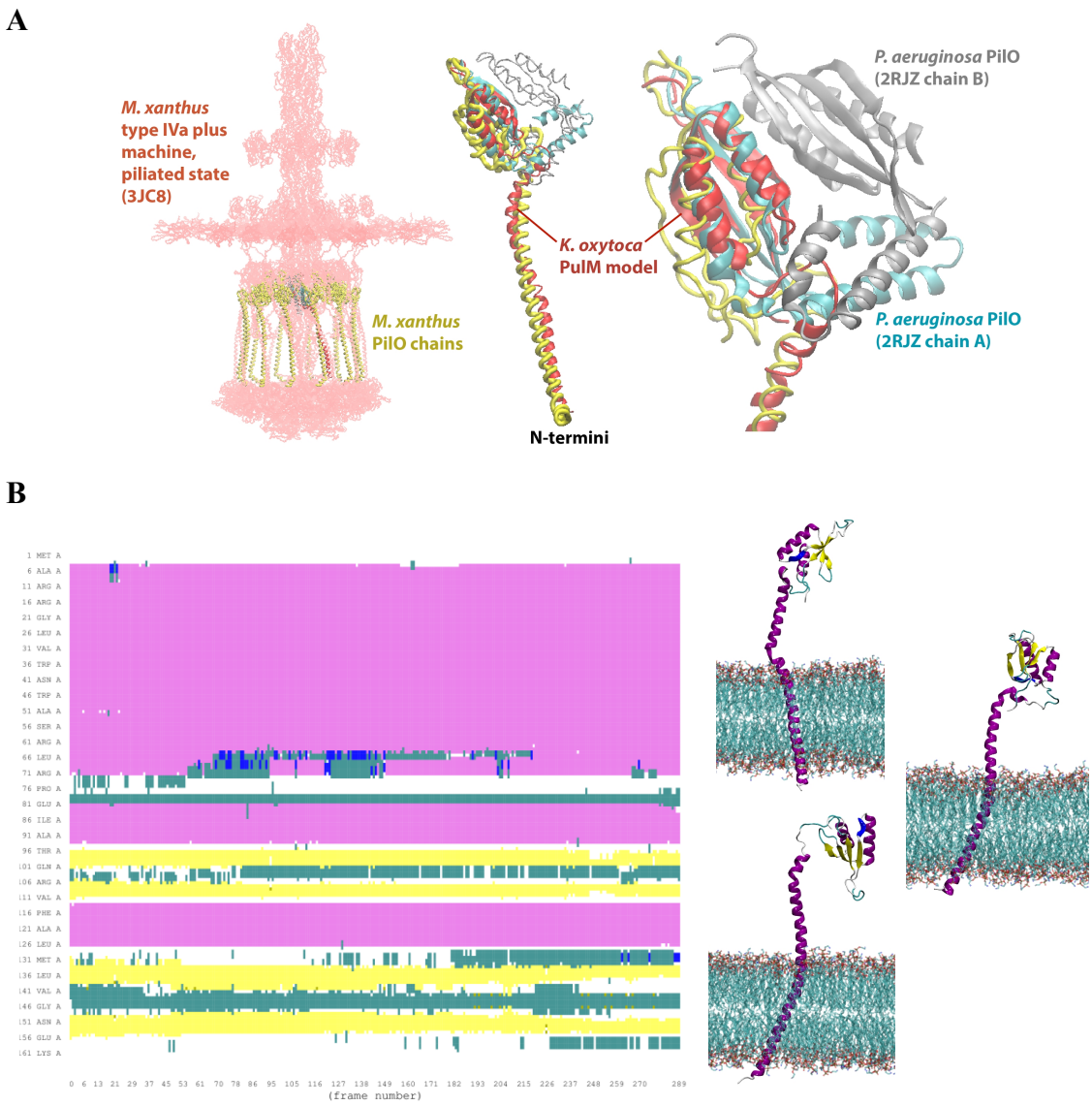


Figure 6.2 – Modelling PulM structure and secondary structure time-line from subsequent atomistic simulation
(A) Figure showing the modelling stages. Left: the Type IVa pilus machine is shown in pink with the *M. xanthus* PilO chains highlighted in yellow. Centre and right: two *P. aeruginosa* PilO chains (cyan, grey) overlaid onto *M. xanthus* PilO (yellow) and newly-modelled *K. oxytoca* PulM (red). **(B)** Left: secondary structure analysis demonstrated that the PulM structure remained largely stable throughout the atomistic simulations, plotting secondary structure of each amino acid (α-helix – purple, 3-10 helix – blue, β-sheet – yellow, turn – cyan, coil – white) against time. Right: representative snapshots show the flexible nature of the PulM protein.

I generated all the CG protein structures using the martinize.py script (http://www.cgmartini.nl/index.php/tools2/proteins-and-bilayers/204-martinize), maintaining pre-assigned protein secondary structure. An elastic network was used to

maintain higher-order protein structure in all the systems, except between PulM residues 70-80 and 150-160, to allow the domain linker and the C-terminal tail to remain flexible. The elastic network used a force constant of 500 kJ mol^{-1} nm^{-2} , and upper and lower cut-offs of 0.9 and 0.5 nm, respectively.

The initial *in silico E. coli* lipid bilayer contained POPE, 1-palmitoyl(16:0)-2-vacenoyl(18:1 *cis*-11)-phosphatidylglycerol (PVPG) and 1-palmitoyl 2-cis-vaccenic 3-palmitoyl 4-cis-vaccenic diphosphatidyl-glycerol (PVPV) in the ratio ~ 13:3:1 (569:146:43 lipids respectively were present in the original system). An *E. coli* IM model was used as the composition of the *Klebsiella* IM is unknown; also, the biochemical experiments had been carried out by Dr. Olivera Francetic in *E. coli* strain (Santos-Moreno *et al.*, accepted), validating use of this *in silico* model. In the PulG monomer and dimer systems, the membrane size was reduced to ensure appropriate PBC without simulating excessively large systems, although the lipid ratio was maintained. The initial state of every system is shown in Figure 6.3.

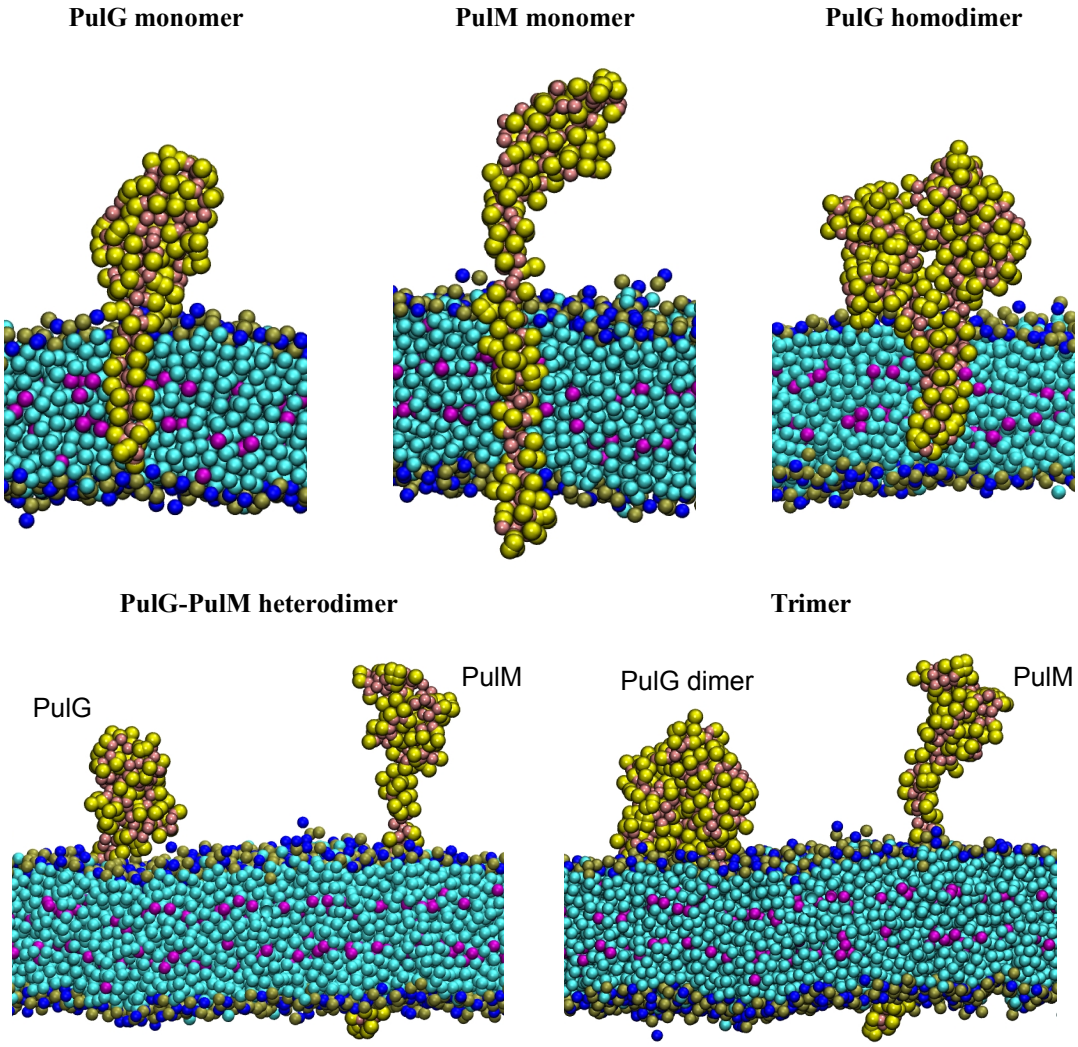


Figure 6.3 – Initial conformations of CG protein-lipid systems
Snapshots of the initial structure of each system. Beads are coloured according to type and location: charged phosphate groups – tan; charged protein groups – yellow; protein backbone beads – pink; glycerol – blue; apolar tail beads – cyan. In the bottom two images, PulG is seen on the left and PulM on the right, embedded in the bilayer.

All CG simulations were performed using GROMACS 5.0.7, with standard Martini V2.2 FF parameters used to treat all the system components. The proteins were added to the pre-equilibrated CG lipid bilayer and visually aligned to place the hydrophilic regions outside the membrane. Overlapping lipids were removed manually using VMD. Following minimisation, each system was solvated with standard (non-polarisable) CG water beads and neutralised by adding 0.15 M NaCl. Simulations were completed using the leap-frog algorithm with a 20 fs time-step, and trajectory data was collected every 200 ps. No bond length constraints were used and

the neighbours list was updated every 10 steps, with a 14 Å cut-off. The LJ interactions were slowly shifted to zero between 9 and 12 Å, and the electrostatic interactions were slowly shifted to zero between 0 and 12 Å, with a relative dielectric of 15 (thus corresponding to distance-dependent dielectric screening). Simulations were performed at an absolute temperature of 313 K coupled to separate groups for protein, lipids and solvent. The pressure was set to 1 bar (Parrinello-Rahman semi-isotropic barostat), under periodic boundary conditions. Systems containing a protein monomer were simulated for 2 μs, and those containing a dimer/trimer were simulated for 5 μs.

Table 6.1 – Systems analysed in this chapter

Name	System	Lipid no.	Time
$PulG_{CG}$	CG lipid + CG PulG monomer	232	3 x 2 μs
$PulM_{CG}$	CG lipid + CG PulM monomer	743*	3 x 2 μs
$PulG_{CG}$ dimer	CG lipid + CG PulG homodimer	350	3 x 5 μs
$PulG_{CG}$-$PulM_{CG}$	CG lipid + CG PulG monomer + CG PulM monomer	733	8 x 5 μs
Trimer	CG lipid + CG PulG homodimer + CG PulM monomer	724	3 x 5 μs

* A larger membrane bilayer was used in all systems containing PulM due to the height of the protein, ensuring that the protein was not able to interact with adjacent images.

6.3 Results

Analysis aimed to describe both the conformational dynamics of monomeric PulM in the IM and the likely interface between PulG and PulM, as well as identify the key residues that maintain interface contacts. This study also aimed to establish whether this putative interface changes in the physiologically expected trimer, namely depending on the presence of a PulG monomer or dimer adjacent to PulM.

6.3.1 Conformation of PulM and PulG monomers

To demonstrate the validity of the CG approach, the protein conformations observed in the simulations of the PulG monomer in the mixed lipid bilayer were compared to the atomistic $PulG_{WT}$ simulations presented in Chapter 5 (see Figure 5.6A). Visual observation demonstrated that the protein helix bent to allow the globular domain to approach the lipids, regardless of the simulation approach. In both cases, PulG

remained embedded in the membrane and experienced internal conformational changes. To quantify these, the RMSD values of both PulG domains in the CG system were calculated (Figure 6.4B) and showed that the RMSD of the globular domain remained stable at ~ 1 Å whereas the helical RMSD fluctuated repeatedly between 2-6 Å. This result resembled the atomistic simulation RMSD results (Figure 5.4), in which the globular domains also tended to fluctuate less (between 1-4 Å) yet the helix exhibited more varied RMSD values (between 1-7 Å). B factor values (also Figure 6.4B) showed that all three replicates exhibited very similar structural fluctuations, with the first five amino acids demonstrating higher B factors (up to ~ 700 $Å^2$). The B factors did not show large movements of the loop containing residues 70-80, as in several atomistic PulG monomer simulations. Taken together, the RMSD and RMSF results highlight that the globular domain exhibited smaller movements on a per-residue basis and over time, in comparison with the helical PulG domain.

Overall, these results demonstrated that the CG approach produced similar results to the atomistic simulations, and confirmed that this choice of method was justified.

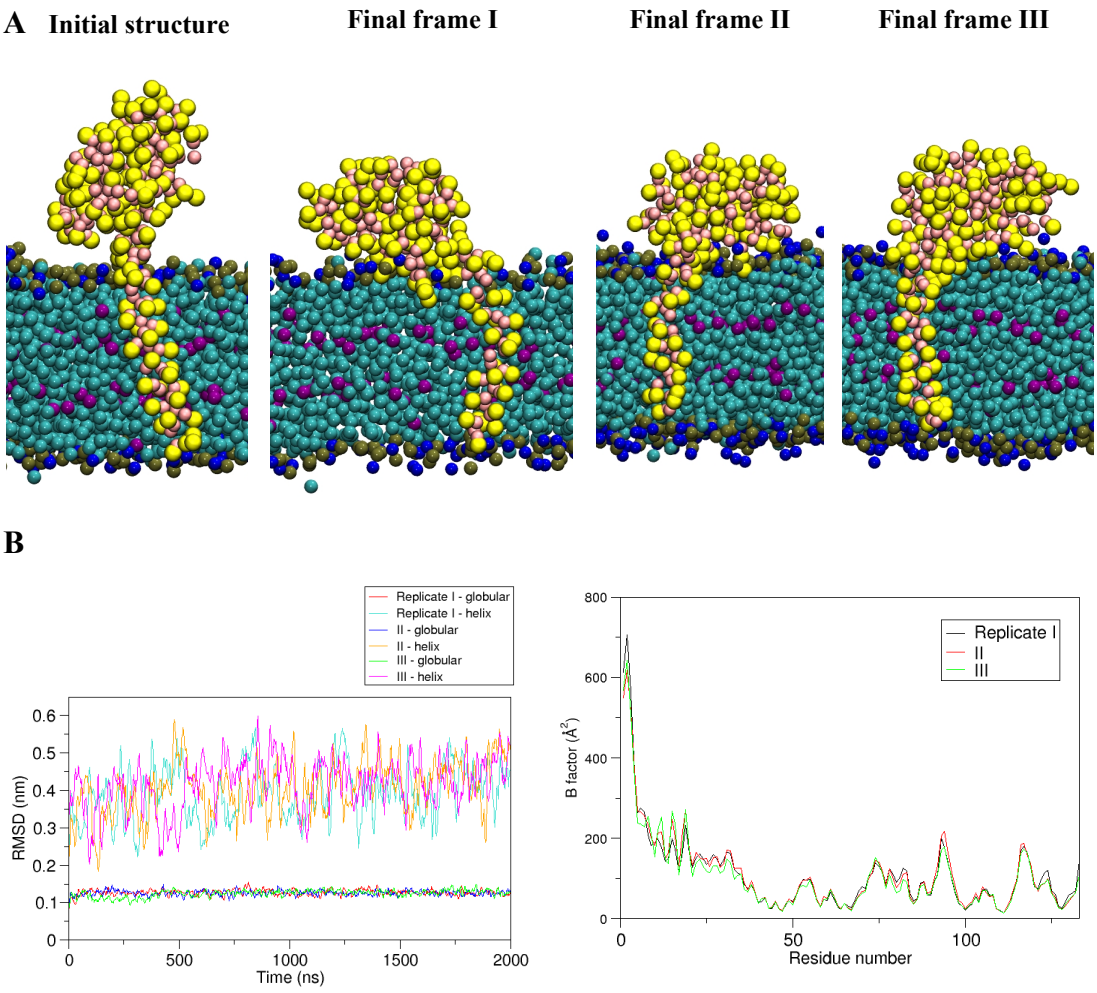


Figure 6.4 – PulG conformational changes during CG simulations
(A) Snapshots of the initial structure of the system and the final conformation of each replicate. Beads are coloured as previously noted. In all systems, the protein bends and the globular domain approaches the lipid bilayer; compare to Figure 5.6, showing the atomistic PulG system conformations. **(B)** Left: RMSD values of the PulG helix (residues 6-54) and globular domain (residues 55-133), coloured according to the legend. Right: B factor values of all three replicates over the final 500 ns of each simulation.

Snapshots of the PulM monomer simulations showed that, like PulG, the protein remained embedded in the membrane. Notably, the initial extended protein configuration adopted a more compact conformation over each 2 μs replicate, with the globular domain moving significantly closer to the membrane bilayer (Figure 6.5A). Calculating RMSD and B factor values (*via* calculation of RMSF) supported the visual analysis; see Figure 6.5B. Similarly to the PulG RMSD data, the PulM globular domain deviated from the initial structure by only 2-3 Å throughout the course of the simulations. However, the helix exhibited higher RMSD values varying

between 3-8 Å. There was no long-term variation present over the course of each simulation; the first and final values were similar for each replicate.

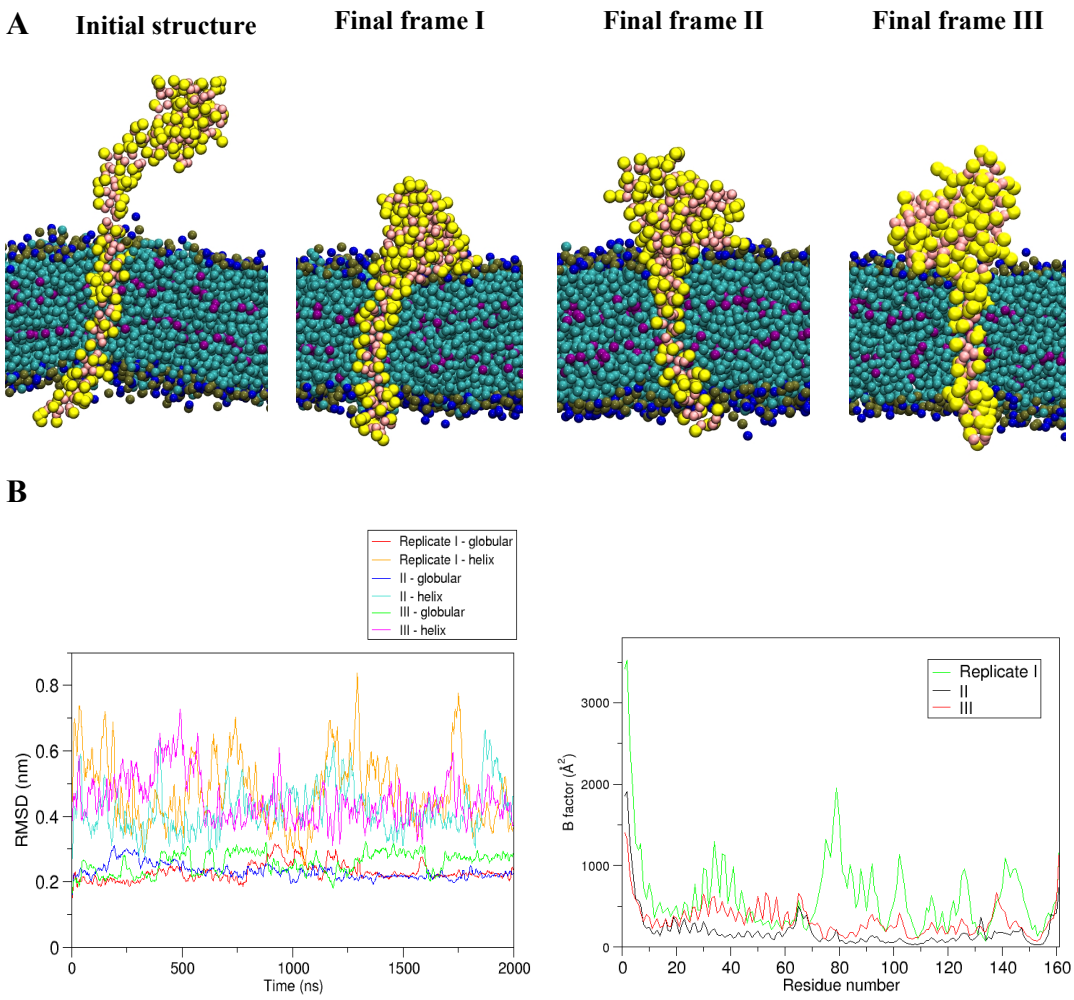


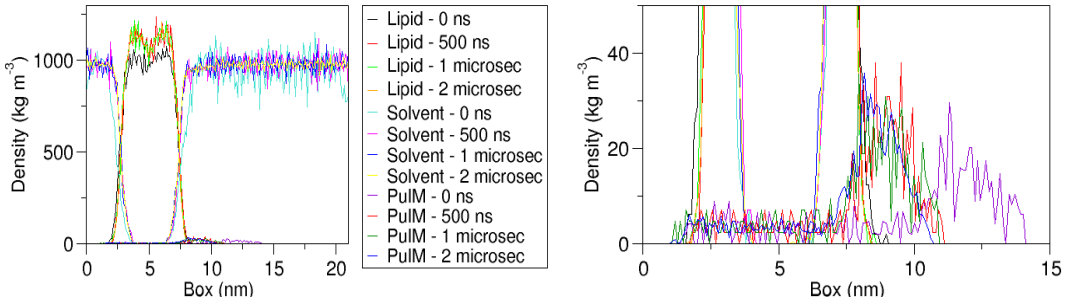
Figure 6.5 – PulM conformational changes during CG simulations
(A) Snapshots of the initial structure of the system and the final conformation of each replicate. Beads are coloured as previously noted. **(B)** Left: RMSD values of the PulM helix (residues 1-64) and globular domain (residues 80-155) coloured according to the legend. Right: B factor values of all three replicates over the final 500 ns of each simulation.

Comparing the system component densities (as a function of the *z*-axis) at 0 ns, 500 ns, 1 μs and 2 μs highlighted the mobile nature of the PulM monomer over the course of the simulations. Each replicate is represented in a graph (Figure 6.6), with the *x*-axis of the graph representing the *z*-axis of the system box. The bottom of the bilayer (representing the cytoplasmic side) and the PulM globular domain are located at ~ 0-5 nm and ~ 10-15 nm along the *x*-axis, respectively. Examining the densities at a

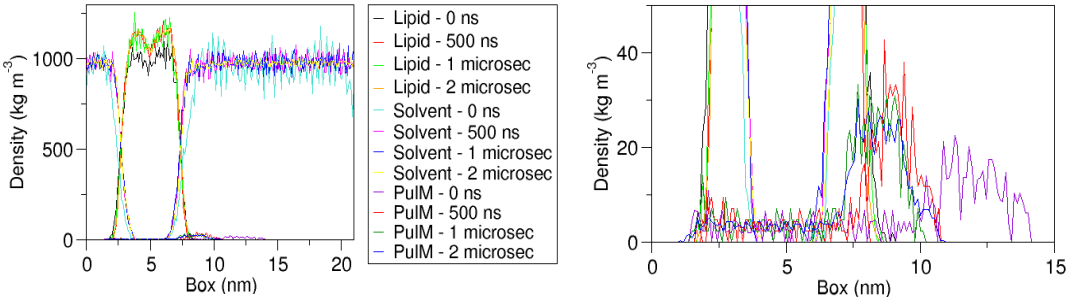
series of separate time points throughout the simulations quantified the observed protein movements. The protein globular domain exhibited a significant shift of over 40 Å towards the lipid bilayer in all three replicates, “shrinking” towards the membrane. Interestingly, the PulM helix tended to contract into/to the membrane during the first 500 ns of the simulation, as the system adjusted, and subsequently relax away from the bilayer again. PulM kinking within the membrane was observed in one replicate at 500 ns as the system equilibrated (replicate 3, red line). The maximum density of the globular domain increased during the simulations as the protein contracted and adopted a less extended conformation.

Solvent densities remained ~ 1000 kg m^{-3}, however the density became more even along the *z*-axis throughout the simulations (compare cyan and yellow lines). The lipid was found consistently at 2-8 nm along the *z*-axis. The lipid densities also became less variable, and increased by ~ 150 kg m^{-3} between 0 and 2 μs; the head groups remained more densely packed than the tail groups throughout.

Replicate I



Replicate II



Replicate III

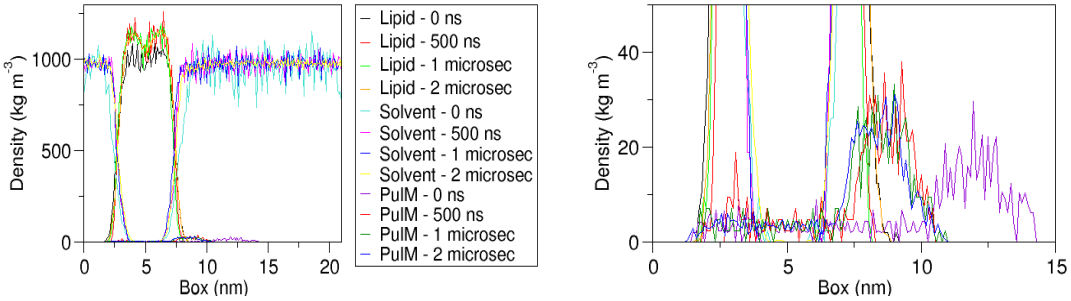


Figure 6.6 – Protein densities of PulM at CG simulation time points
Left: Graphs showing the densities of lipid, solvent and PulM as a function of the *z*-axis of the system box at 0 ns, 500 ns, 1 μs and 2 μs of each replicate, coloured according to the legend. Right: graph axes readjusted to highlight PulM density movements relative to the lipid component.

6.3.2 Identifying the PulG-PulM dimer interface

To identify the likely PulG-PulM interfaces *in vivo* and direct future experimental work, both a heterodimeric $PulG_{CG}$-$PulM_{CG}$ system and trimeric $PulG_{CG}$-$PulG_{CG}$-$PulM_{CG}$ system were simulated. In the heterodimer system, the PulG and PulM proteins were positioned ~ 140 Å apart, and the conformations into which they dimerised were subsequently analysed. Performing eight replicates aimed to allow

the most frequent, and therefore most likely, interface to be observed.

Visual observation showed that PulM once more contracted closer to the membrane, from the initial extended conformation. Replicating analyses performed using the PulG homodimer in Chapter 5, Figure 6.7A shows the final structures of the eight heterodimer replicates alongside the initial protein positioning. In three of the eight replicates, only the globular domains of the proteins interacted, with the TM helical segments remaining separated by bilayer lipids. In four of the remaining replicates, the proteins interacted fully along the helical segments. In one replicate, following 5 μs of simulation, the globular domains interacted, as well as the PulG N-terminus with PulM helix residues 16-24.

Calculating the total number of contacts closer than 6 Å between PulG and PulM throughout the simulations (Figure 6.7B) demonstrated that the proteins had dimerised following maximum 2 μs of simulation, with up to 42 contacts forming between the pair. The number of contacts that formed varied significantly between replicates; in two, the proteins did not form more than 15 contacts (black and purple lines), whereas in one replicate, the proteins quickly dimerised and subsequently did not form fewer than 20 contacts (light green line).

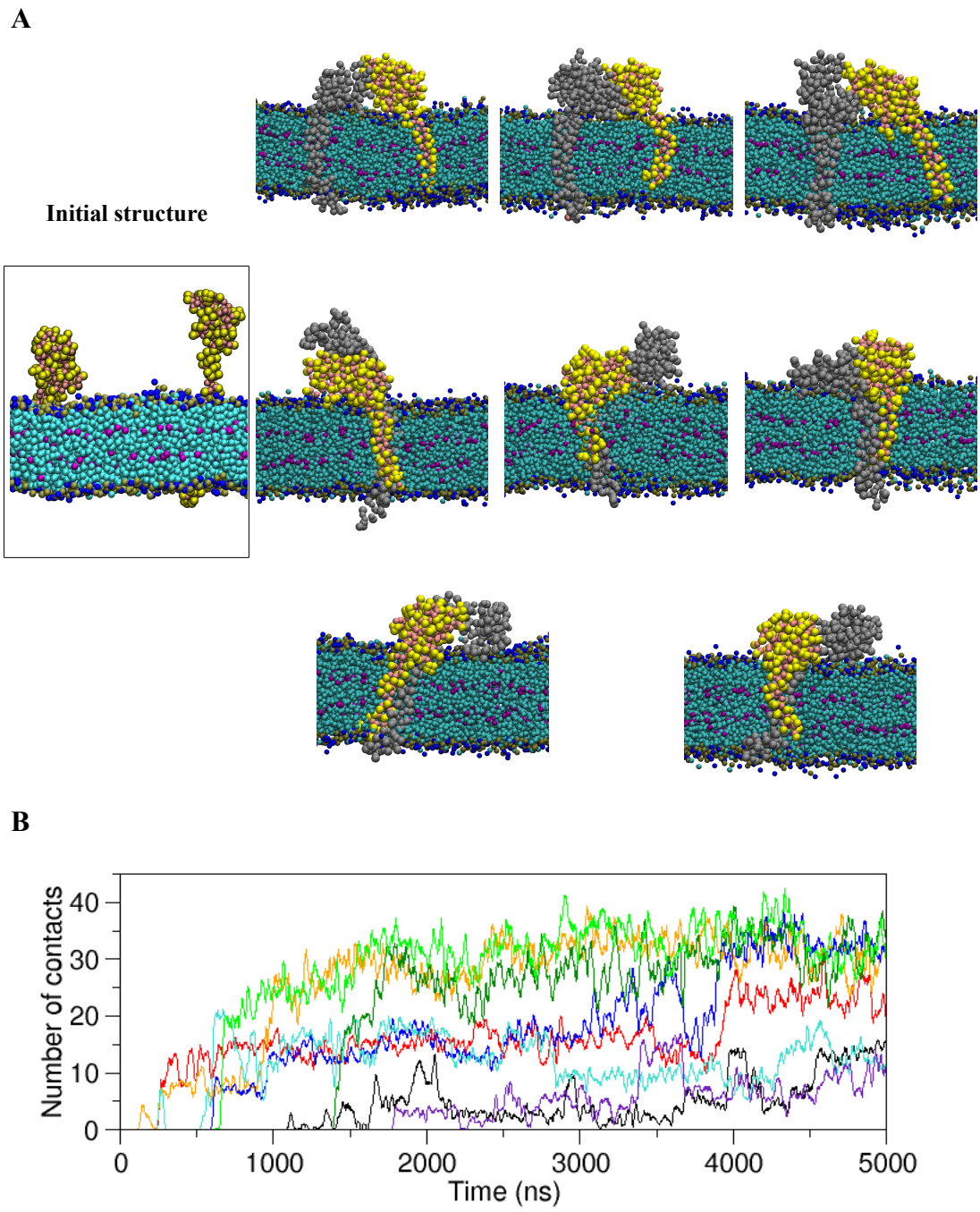


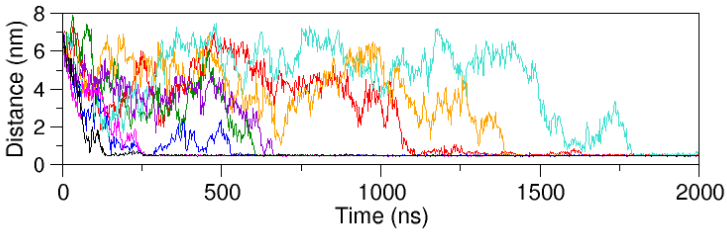
Figure 6.7 – Heterodimer structures and inter-molecular contacts
(A) Snapshots of the initial structure of the system (boxed) and the final conformation of each replicate, with PulG shown in yellow and PulM in grey. Remaining beads are coloured as previously noted. **(B)** Graph showing the total number of contacts closer than 6 Å between PulG and PulM, throughout the simulations, with each of the eight replicates coloured differently.

Calculating the minimum distance between each residue of one monomer and any part of the adjacent monomer gave an indication of which residues are important in

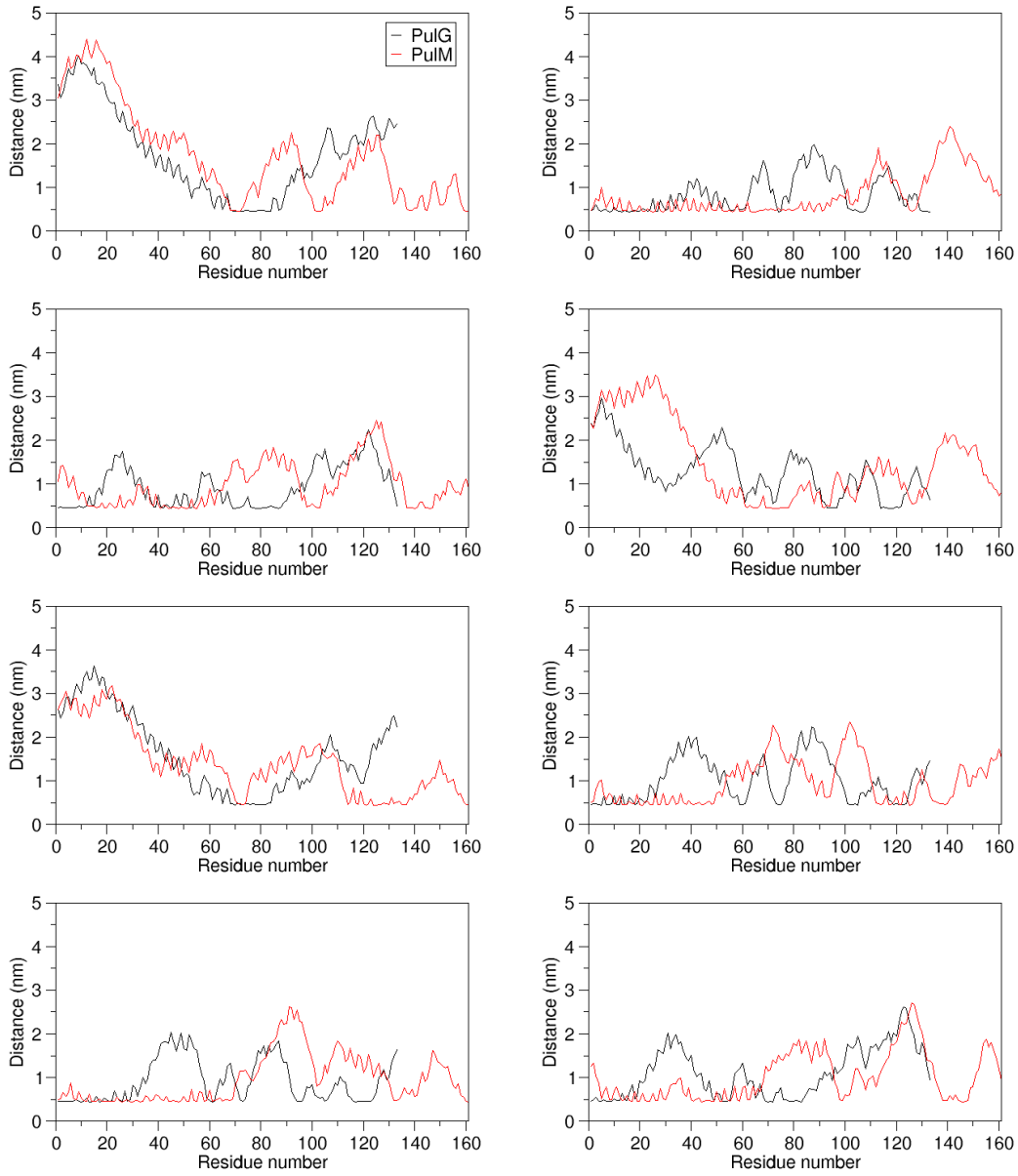
the dimer interface (Figure 6.8). Due to the size of the CG particles (~ 4.7 Å), "contacts" were defined as occurring when particles approached to within 5 Å of each other. In five replicates, the N-terminal 15 residues of PulG approached to within 5 Å of PulM, alternating as the helix rotation affected potential contacts; only PulG residues between 70-75 consistently contacted PulM in every replicate. In the replicates in which the PulM helix interacted with PulG, residues 10-50 approached to within 5 Å of PulG, exhibiting the same rotational effect due to the helical structure. There were no residues which contacted with PulG in every replicate, although residues 137-142 contacted PulG in five replicates. Overall, this analysis suggested that the PulG-PulM dimer can form a variety of interfaces, with a tendency for the helical domains to interact within the membrane, and PulG residues 70-75 and PulM residues 137-142 to engage in globular domain interactions.

Next page: Figure 6.8 – $PulG_{CG}$-$PulM_{CG}$ heterodimer interface
(A) Each coloured line represents the absolute distance between PulM and PulG throughout the first 2 μs of the simulation; the proteins approach and dimerise in every replicate after maximum ~ 1.8 μs. **(B)** The minimum distance attained between each residue of one monomer and any part of the adjacent monomer, at any point during the simulation, gives an indication of which residues are important in the dimer interface. The distance between PulG residues and PulM is indicated with an black line, and the distance between PulM residues and PulG with a red line.

A



B



The positions of these interfacial residues are shown in Figure 6.9; atomistic structures are shown, to highlight the residue locations. In both proteins, the N-terminal helix residues engage in dimerisation interactions, as do the residues on a

loop at the top of each protein, within the globular domain. The lack of a clearly reproduced dimer interface suggests that other components of the T2SS may act to constrain PulG and PulM into the correct conformations *in vivo*. Given that the globular domains interacted in all replicates (even those in which the helices did not form contacts), the initial stage of dimerisation is likely mediated by polar residues in the PulM and PulG globular domains.

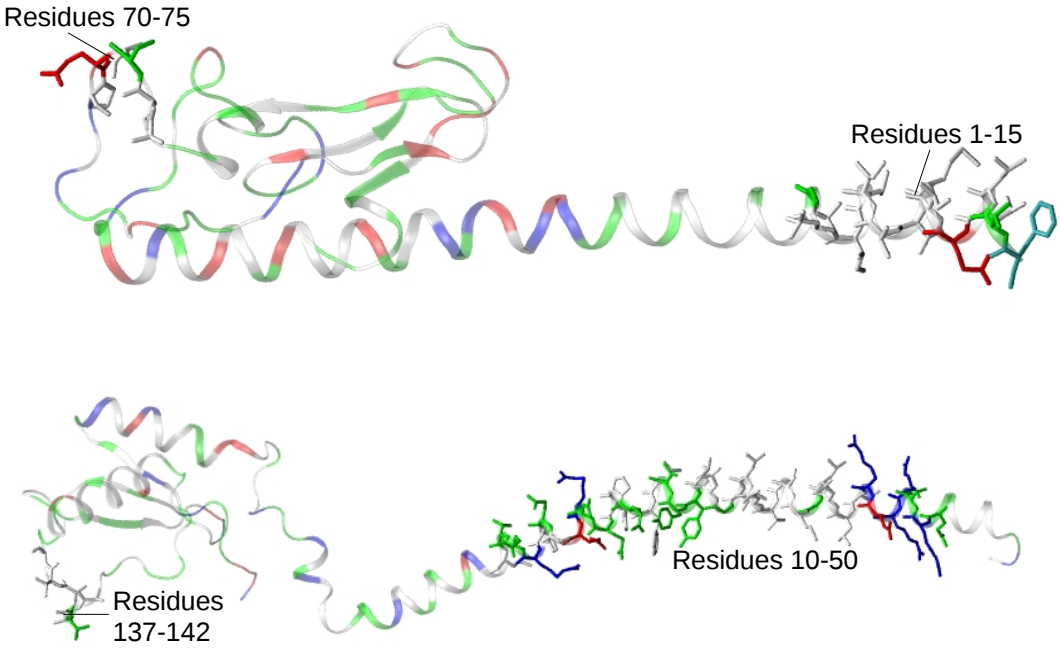


Figure 6.9 – PulG-PulM heterodimer interfacial residues
Top: PulG structure, with the N-terminus on the right, coloured by residue type (non-polar – white, basic – blue, acidic – red, polar – green), with the residues engaged in dimerisation interactions in stick representation. Bottom: PulM structure, represented likewise.

The contact matrices of the equilibrated stable dimerised states, showing the mean smallest distance between any pair of residues during the final 500 ns of each replicate, are shown in Figure 6.10, providing another representation of the observed interactions. Distances are represented using a colour scale, with dark blue showing the largest distance and red the smallest. The matrices highlight the proximity of residues within the globular domain of each protein, and the variation between the PulG-PulM interactions in each replicate; for example, the C-termini of the proteins only interacted in two replicates, and in five replicates the helices can be seen to form contacts.

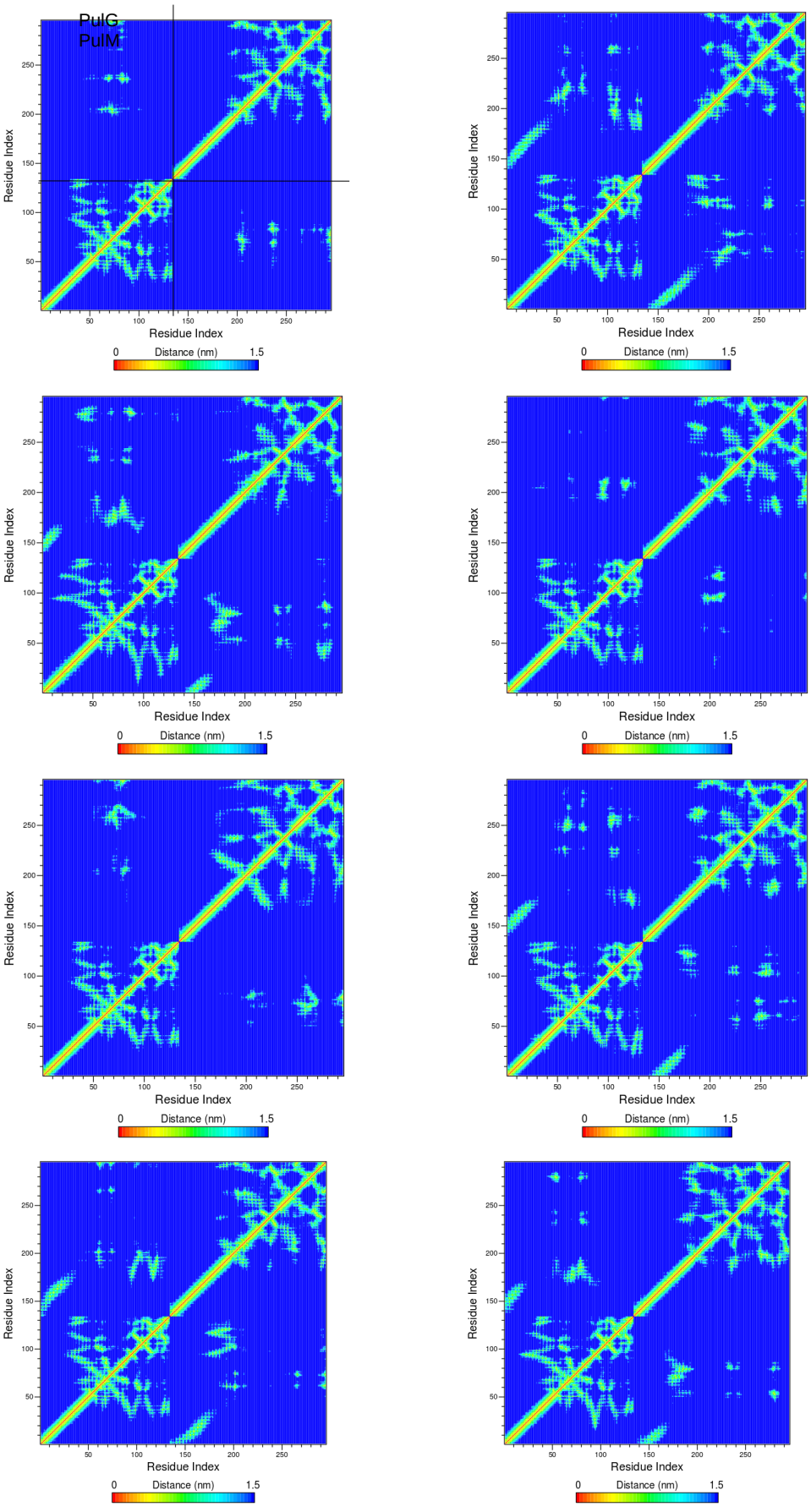

PulG
PulM
Residue Index
Residue Index
0
Distance (nm)
1.5

Figure 6.10 - Contact matrices of equilibrated stable dimerised states
The contact matrices of the eight replicates show the mean smallest distance between any pair of protein residues in each system, during the final 500 ns of each replicate, coloured according to the legend at the bottom of each matrix. The matrices highlight the interactions between PulG (residues 1 to 133) and PulM (residues 134 to 295).

6.3.3 Analysis of $PulG_{CG}$-$PulG_{CG}$-$PulM_{CG}$ trimer

To ascertain how the presence of another protein may constrain the dimeric interface, three replicates of the trimeric $PulG_{CG}$-$PulG_{CG}$-$PulM_{CG}$ system were simulated, by placing the PulG dimer sourced from the bottom of the pseudopilus model from Campos *et al.* in the same mixed lipid bilayer as a PulM monomer, ~ 130 Å apart. Figure 6.11A shows the initial and final complex structures of the trimer replicates, and Figure 6.1B shows the RMSD and B factor values of each protein within the trimer. The PulG monomers consistently displayed lower RMSD and B factor values than PulM, in line with the more stable starting structure of the dimer, and reflecting the contraction of PulM to the membrane over the equilibrating stage of each simulation.

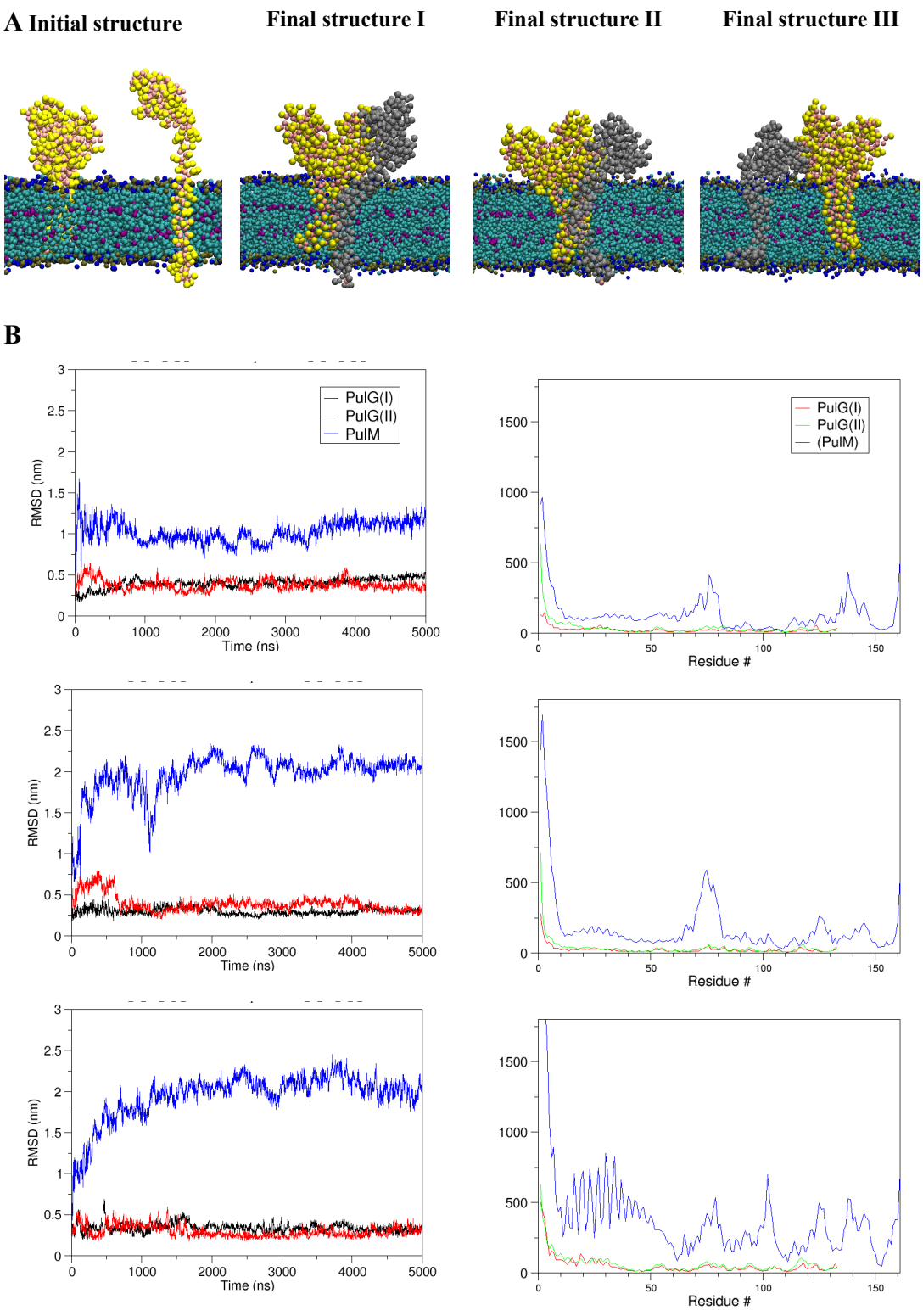


Figure 6.11 – Trimer structures and RMSD and B factor values
(A) Snapshots of the initial structure of the system and the final conformation of each replicate, coloured as previously. The PulM monomer is shown in grey in the final frame snapshot, to differentiate it from the PulG dimer. **(B)** Left: RMSD values of each PulG monomer and of PulM, coloured according to the legend. Right: B factor values of all three replicates over the final 500 ns of each simulation.

The absolute minimum distance between each constituent part of the trimer over time indicated which parts interacted, and whether any two proteins did not come into contact (Figure 6.12). Analysis showed that the PulG monomers remained within 5 Å at all times, and the dimer did not disintegrate upon addition of PulM to the system. In all three replicates, PulG(I) and PulM increased in proximity and formed contacts maintained through the remainder of the simulation, although the time point of the initial contact varied from ~ 0.5–1.3 μs. PulG(II)-PulM interactions were more variable; in one replicate, the proteins approached to within 5 Å after ~ 2 μs, and in the other two they stabilised at 10 and 25 Å apart throughout the simulations, respectively.

The distance between each residue of one protein and any part of the adjacent protein gave an indication of which residues are important in the trimer interface (Figure 6.13). PulG(I) and PulG(II) formed similar contacts with each other in every replicate, as anticipated since the dimer was present in the initial system structure. These contacts mimicked those found in the atomistic simulation analysis (Figure 5.13), although the interface engaged fewer residues than in the PulG dimer; PulG(I) residues 105-108 and 130-133 did approach PulG(II) yet residues 86/125-130 did not, and similarly only residues 88-96 of PulG(II) approached PulG(I) (in the atomistic dimer, residues 80-90 and even up to residue 100 approached PulG(I)). From the PulG dimer, it was almost exclusively PulG(I) that formed contacts with PulM. Only residues 10-20 of PulG(II) approached to within 5 Å of PulM, in one replicate. PulM contacted PulG(I) residues, however the interface varied; in one replicate, residues 75-85 and 122-128 approached within 5 Å of PulG(I), in another alternating residues between 20-64, found in the helix, as well as residue 99 and residues between 140 and the C-terminus, and in the third replicate various residues between 1-105 with the notable exception of any residues between 62-74.

Extrapolating from these data suggests that the PulM helix may interact with a PulG dimer *in vivo*, and the globular domain is extremely variable in its interactions, likely due to the flexible tether (consisting of residues 70-80) that allows a broad range of motion for the PulM “head”.

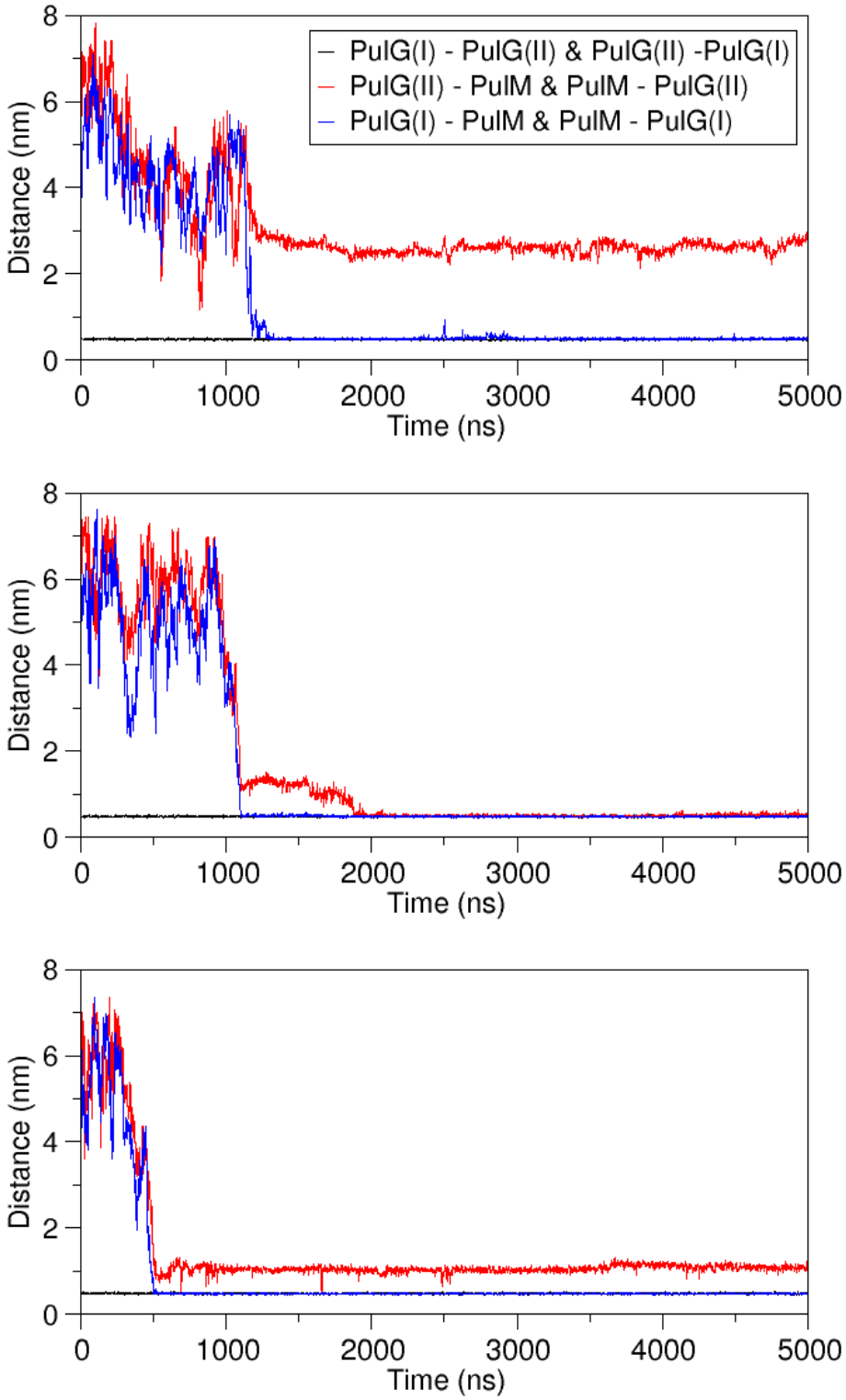


Figure 6.12 – Minimum distance between monomers in trimer over time
The absolute minimum distances between the constituent proteins of the trimer over time are shown; the PulG monomers (black line) are found adjacent from the initial frame, whereas the PulM protein either approaches the dimer gradually during the first ~ 1.5 μs, or does not contact the PulG(II) monomer at all, depending on the replicate.

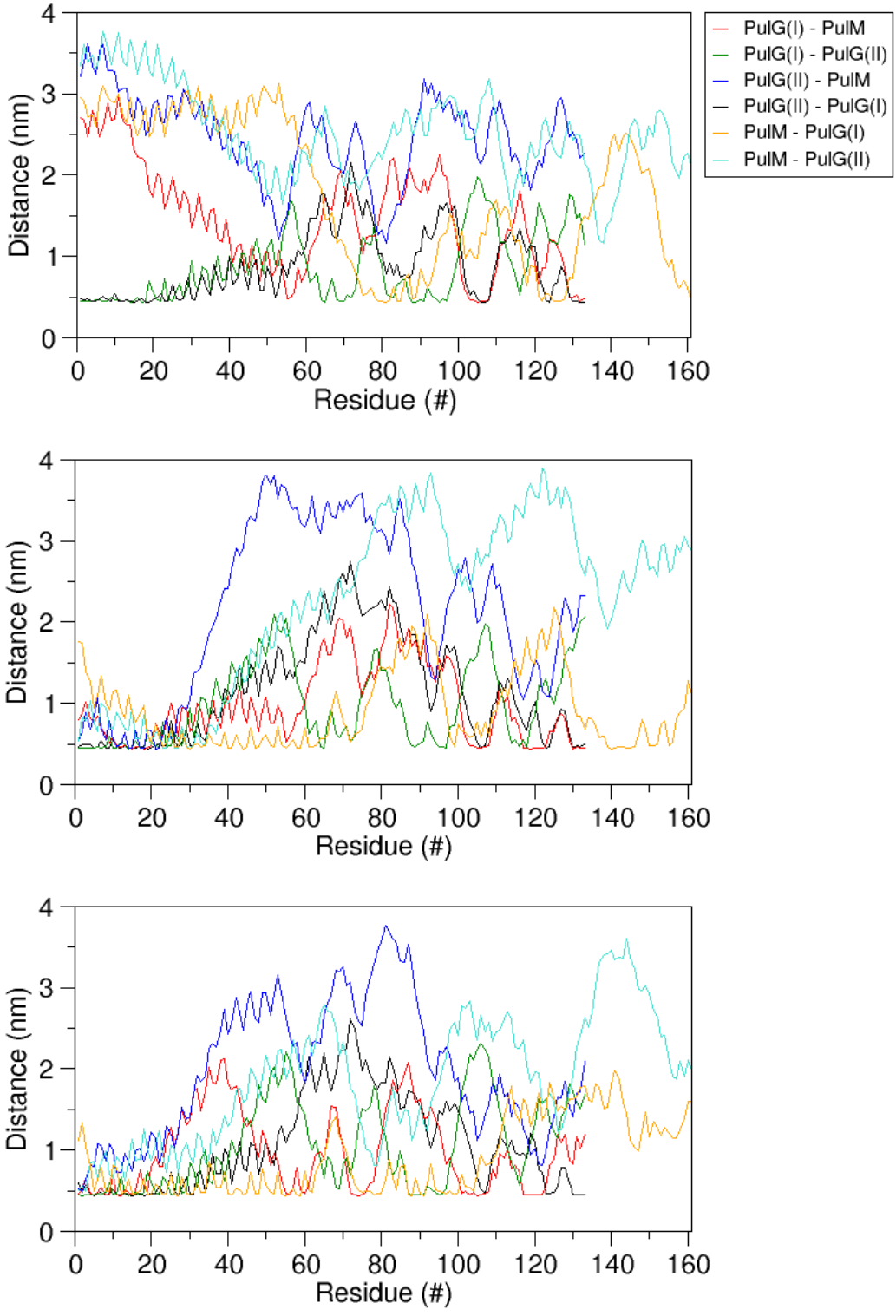
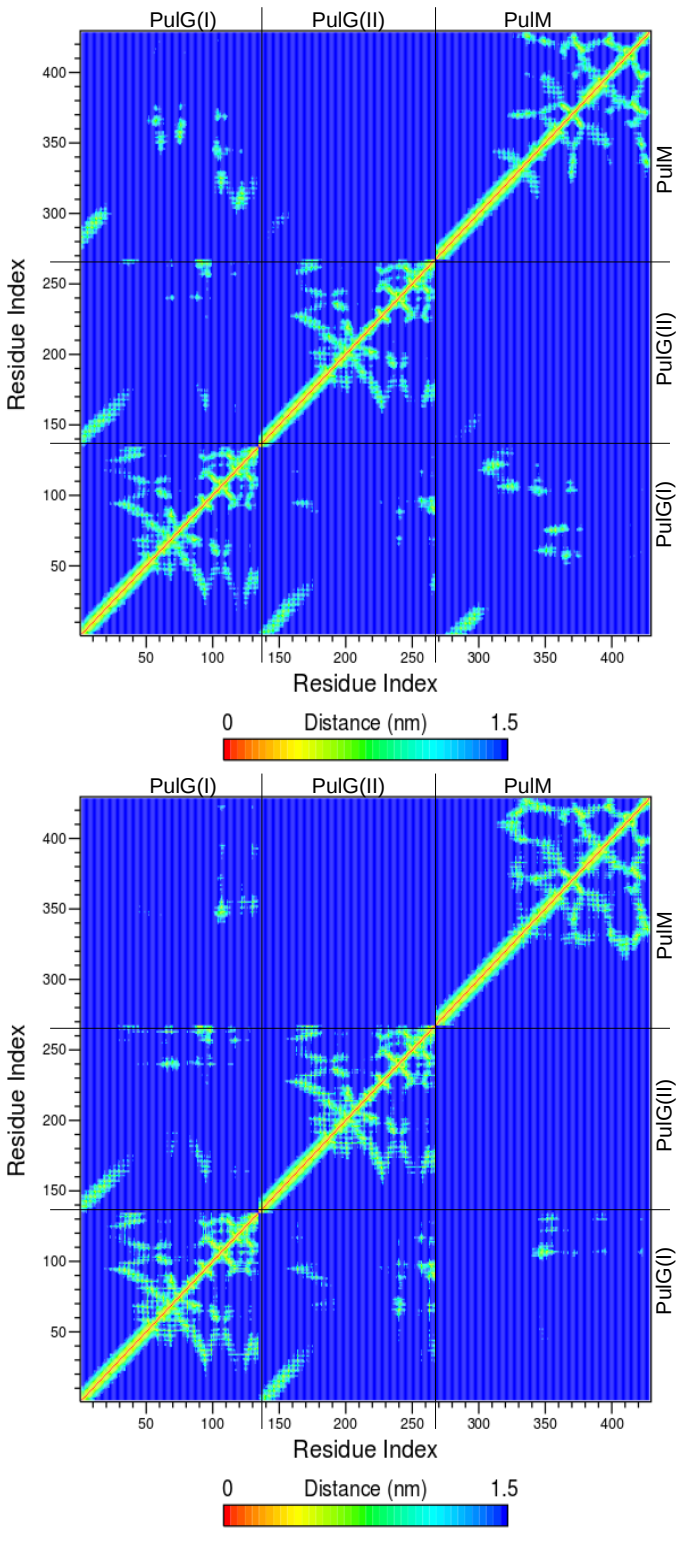


Figure 6.13 – PulG-PulG-PulM trimeric interface
(A) Graphs show the minimum distance attained between each residue of one monomer and any part of the adjacent monomer, observed throughout the simulations, giving an indication of which residues are important in the trimer interface. The lines are coloured as shown in the legend, and a detailed explanation of the data is provided in the text.

The contact matrices of the equilibrated stable trimerised states are shown in Figure 6.14 and provide another representation of the observed interactions. The matrices highlight the proximity between the PulG monomers, with at least the 30 N-terminal residues of each remaining in contact to the end of the simulations, and residues around position 70 and 90-100 interacting in two replicates with the PulG(II) C-terminal domain, similar to the contacts seen in the PulG homodimer (see Chapter 5). The matrices also show the contacts between PulG(I) and PulM, with the varied pattern of cyan pixels supporting the previously discussued analysis; in one replicate, the proteins do not interact very closely, and only PulG residues 100-110 contact PulM residues 75-85. However, in the other two replicates, the proteins engaged in closer and more extensive contacts, highlighting potential candidate residues for future experimental substitutional analyses. Notably, the matrix sections corresponding to PulG(II) to PulM residue proximity contain dark blue pixels, demonstrating the lack of interactions, and show contacts between the PulM and PulG(II) N-terminal helices in only one replicate.


PulG(I)
PulG(II)
PulM
Residue Index
Residue Index
0
Distance (nm)
1.5
PulG(I)
PulG(II)
PulM
Residue Index
Residue Index
0
Distance (nm)
1.5

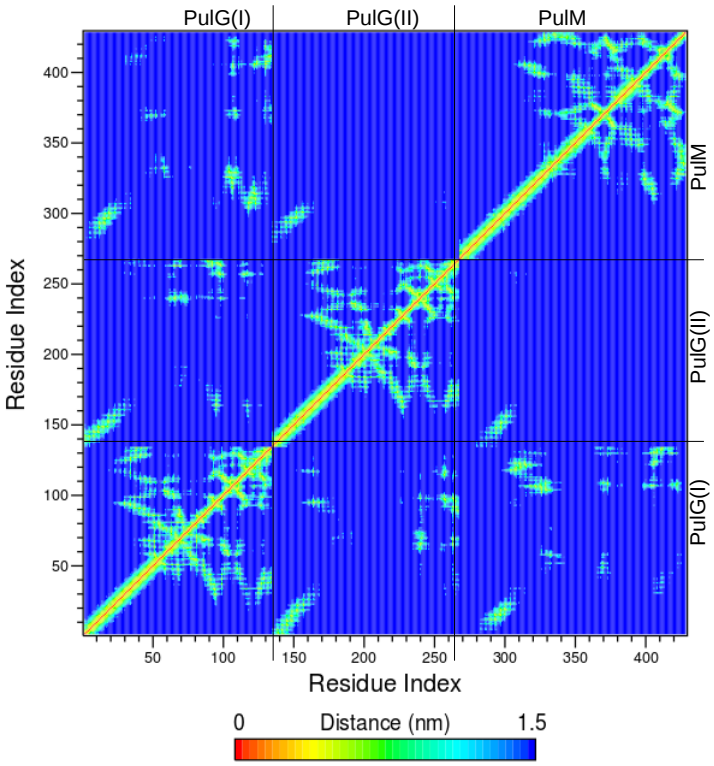


Figure 6.14 - Contact matrices of equilibrated stable trimerised state
The contact matrices of the three replicates show the mean smallest distance between any pair of protein residues in each system during the last 500 ns of each simulation, coloured according to the legend at the bottom of the matrix. The matrices highlight the interactions between the PulG monomers, and PulG(I) and PulM.

6.4 Discussion

This CG study aimed to answer three specific questions, by simulating *Klebsiella oxytoca* PulG and PulM proteins in complex. Firstly, what is the behaviour of monomeric PulM in the IM? Secondly, what do assembly simulations predict the likely interface between PulG and PulM to be, and which residues are key in maintaining these contacts? And finally, what is the putative interface of the physiologically expected PulG-PulG-PulM trimer?

Extensive CG simulations demonstrated that PulM remains located in the membrane, with the globular domain shrinking towards the lipid bilayer as the simulation progressed; most movement occurred within the first 500 ns. In one replicate, kinking of PulM was observed within the membrane, suggesting that the protein can

distort extensively to adopt the most energetically favourable conformation. The N-terminus of each replicate became more exposed to solvent as the simulations progressed, and did not remain buried in the bilayer as in the starting conformation.

There was no clear consensus among the PulG-PulM replicates regarding the heterodimeric interface; solely the globular domains interacted in three replicates, whereas in five replicates, the helices interacted. Parallel to these results, cysteine cross-linking data have been obtained by Dr. Olivera Francetic and Javier Santos-Moreno (personal correspondence). The cross-linking has shown that a C-terminal His tag in the globular head of PulG appears to reduce the amount of PulG-PulM cross-linked heterodimer, possibly indicating that the periplasmic regions of both proteins are also involved in interaction, as occurs for their *N. meningitidis* T4P counterparts[102]. The CG simulation results support this data, demonstrating that PulG residues 70-75 and PulM residues 137-142 form the most frequent inter-protein globular domain contacts, identifying these residues as candidates for future substitutional analyses. Similarly, meningococcal major pilin PilE (PulG homologue) interaction with PilO (a PulM homologue) is mediated by the globular periplasmic domains of both proteins and also the 39 N-terminal residues of PilE, as mentioned in Section 6.1. This is supported by five of the heterodimer simulation replicates, in which ~ 30 N-terminal PulG residues interact extensively with PulM. Cross-linking showed that PulM residues found close to position 17 are likely to interact with PulG, corroborated by the CG data in which PulM residue 17 interacted with PulG in the heterodimer, and with PulG(I) in two replicates and PulG(II) in one replicate, in the trimer simulations.

Only two variants, $PulM^{N3A}$ and $PulM^{Q48+MFKQQ}$ (with an inserted penta-peptide), showed a secretion defect when Pul components were produced at physiological levels in *E. coli*; for the former the defect was total, while for the latter only a partial defect was observed. Remarkably, single substitutions in PulM R14 and R16, which were good candidates to interact with PulG E5 due to their positive charge, conservation, and proximity to the IM-cytoplasm interface, did not affect the functional properties of the protein at all. SDS-PAGE and immunoblot analysis of the covalently linked proteins showed that positions 9 and 12 in PulG allow the

formation of a disulphide-bridge-linked PulG-PulM heterodimer, suggesting that they also lie close to PulM. Variant $PulG^{V9C}$ was the most efficient in forming disulphide bridges with PulM Cys-variants, especially $PulM^{L18C}$ and $PulM^{L19C}$. In the five simulation replicates in which the PulM helix interacted with PulG, residues 12-24 and 42-50 were the sole regions to approach to within 6 Å of PulG repeatedly, partially supporting the experimental data. Future work should include atomistic simulations to examine the specific interactions of the N-terminal domain residues of the two proteins, at a higher resolution.

As the structures of PulL homologues are known[238], the extension of this simulation work would include modelling the *K. oxytoca* PulL-PulM heterodimer. Likewise, preliminary BAC2H analysis suggests that the PulA Ins domain may be interacting with PulM (Olivera Francetic, personal correspondence), so the initial next step in elucidating T2SS complex interactions would be to simulate PulA and PulM together, and identify the resulting interface. The exciting new cryo-electron tomography work discussed in this chapter represents a large step towards an increasingly detailed and complex picture of secretion nano-machines, and *in silico* data can and should play a key role in the development of our understanding in this field.

Chapter 7 – Conclusions and Future Perspectives

Despite significant progress in the characterisation of the T2SS complex, there are still outstanding questions regarding interactions between the system components. In this thesis, I have attempted to provide novel insights into these questions at the molecular level, by using atomistic MD simulations of two proteins from the T2SS, the secreted substrate PulA and the major pseudopilin PulG, and coarse-grained MD simulations of PulM, the putative chaperone believed to move PulG from the bacterial IM into the centre of the IM T2SS assembly platform. The results of these simulations have been directed by and compared to experimental data where possible, helping to complement available information and formulate hypotheses regarding the T2SS.

Atomistic MD simulations allowed visualisation and rationalisation of PulA protein-lipid interactions. In all the non-substituted simulations containing a lipid bilayer with sodium chloride counter-ions, PulA approached the membrane and segments of the N1, N2 and N3 domains contacted the lipid, regardless of the presence of a Lol retention signal or acyl anchor. Notably, N1 domain rotation was observed in both systems, which may have functional relevance, possibly moving the protein into position for repeated catalytic actions. The increased hydrogen bonding and atomic contacts between domains N1/2/3 and POPE suggested a mechanism to protect the protein from proteolysis in the periplasm, in preparation for presentation of the substrate to the T2SS. In addition, MD simulations provided a molecular view of the tether region and showed that this segment plays a key role in $PulA_{NA}$ stability *in vivo*. The simulations revealed the likely extended structure of the PulA tether, which established van der Waals interactions with the membrane surface. The network of bonds between lipid and the LAS proved to be more extensive than predicted, and involved phosphate in addition to amide groups of POPE as well as protein residues 3 and 4. Interestingly, residues 470-510 and 520-530, found within the Ins domain, consistently demonstrated large movements and may act as a secretion determinant *in vivo* by interacting with other components of the T2SS although they moved freely in solution during the simulations. Future experimental and computational work would aim to fully ascertain the role of the Ins domain, for example by including the

bound calcium cations. Atomistic or CG simulations of PulA paired with components such as PulG or PulM, where crystal structures are available, may direct substitutional experiments by identifying possible key residues necessary for substrate specificity. Further analysis is necessary to examine the proximity of known sugar binding PulA residues and demonstrate the functional relevance of this data. Examining the dynamics, and notable decrease in protein-lipid interactions, of PulA in the presence of bound crystallised calcium cations and calcium chloride counter-ions demonstrated that the periplasmic, and subsequently extracellular, environment may have functional implications for the catalytic action of the protein. Further simulations and possibly experimental analyses are necessary to examine the implications of this insight; the extent to which secreted enzymes are constrained by their environment in the cell, and whether the periplasmic environment affects the interactions between the T2SS Ap components.

Extensive atomistic simulations of PulG variants provided the first data regarding the Type II major pilin in motion. Both the monomer variants and dimer structure remained buried in the POPE bilayer into which they were placed, and the globular domain tended to approach the bilayer. In all cases, both the helical and globular domains tended to deviate slightly from the initial structure, however extensive fluctuations were only observed upon free movement in solution of the loop containing residues 70-80, which may have functional relevance. The sole calcium ion present in the PulG structure remained bound in a loop by Asp117 and Asp125 throughout all simulations, despite being unconstrained; this calcium ion has been shown to be necessary for secretion by *V. cholerae* GspG, and future work would examine the effect of calcium cation absence on the PulG globular domain fold. The helical portion of the protein was stabilised in the presence of another PulG monomer, due to the dimeric interface, with (*P*)Lys30 and (*P+1*)Asp32 playing a key role in the middle section of the helix. Linking the interactions of the N-terminus to function was of key interest and simulations clarified that Glu5 interacts with the methylated N-terminal Phe residue; in the $PulG_{WT}$ system the methyl group of MPH and the Glu5 side-chain remained consistently within 3 Å of each other, unlike in the variant systems, causing a loop to form.

The PulG dimer interface consists of two clear regions; the first 50 residues of each protomer, and the globular domain interface between (*P*)105-108,125-133 and (*P+1*)40-48,80-100. The simulated dimer structure occupied a low energy state as (*P*)16-(*P+1*)10 and (*P*)16-(*P+1*)11 maintained close contacts. These results corroborated data obtained by Campos *et al*. MD analysis identified residues Asn27, Lys30 and Gln34 of *P*, and Lys28, Asp32 and Lys35 of *P+1* as providing key contacts in the central segment of PulG α-helix, and experimental substitution of these residues could be performed to support these results. The model generated by Nivaskumar *et al.* suggests that Lys28 and Lys35 interact with (*P+3*)E5. Therefore, it is possible that the interactions observed in this simulation are only present in the pre-assembled dimer, and are disrupted upon PulG oligomerisation. The simulated dimer contained unmethylated F1 and therefore future work would require further simulations to investigate whether methylated Phe1 on $PulG_{WT}$ in the membrane interacts with E5 and aids piliation, possibly by affecting the binding energy of the system or the energy required to extract $PulG_{WT}$ from the lipid for pseudopilus assembly.

Homology modelling produced a previously unknown structure for PulM and extensive CG simulations using this structure demonstrated that PulM remains located in the membrane, with the globular domain shrinking towards the lipid bilayer as the protein adopted the most energetically favourable conformation. There was no clear consensus among the PulG-PulM heterodimer replicates regarding the interface; a combination of *in silico* and *in vivo* results (in collaboration with Dr. Olivera Francetic) suggested that the globular domains form the initial dimerisation contacts and the protein helices subsequently interact within the lipid bilayer. Future work should include atomistic simulations to examine the specific interactions of the N-terminal domain residues of the two proteins, building on the work from the previous chapter regarding the key role of PulG E5 in the T2SS. Likewise, modelling a *K. oxytoca* PulL-PulM heterodimer and also a PulA-PulM heterodimer would yield valuable data.

In summary, the results presented in this thesis have highlighted the interactions which allow the T2SS to function in *K. oxytoca*. Molecular simulations represent a

genuinely useful strategy for observing and predicting protein-lipid and protein-protein interactions, enabling experimental results to be explained on a molecular level. It is hoped that the knowledge gained from this study will contribute both to understanding of the T2SS and, in the future, to other secretion systems and methods to either increase useful secretion or prevent secretion by pathogenic bacteria, thereby improving human health and industry. Further insights into the structural and thermodynamic mechanisms of secretion should undoubtedly be gained by applying similar approaches to other proteins or lipids of interest, and by simulating increasingly large and varied complexes.

Appendix

Parameters for MPH amino acid

```
[ MPH ]
 [ atoms ]
        HT1     HA      0.05       0
        HT2     HA      0.05       1
        HT3     HA      0.05       2
        CT      CT3     0.16       3
        N       NH3    -0.30       4
        H1      HC      0.33       5
        H2      HC      0.33       6
        CA      CT1     0.23       7
        HA      HB      0.10       8
        CB      CT2    -0.18       9
        HB1     HA      0.09      10
        HB2     HA      0.09      11
        CG      CA      0.00      12
        CD1     CA     -0.115     13
        HD1     HP      0.115     14
        CE1     CA     -0.115     15
        HE1     HP      0.115     16
        CZ      CA     -0.115     17
        HZ      HP      0.115     18
        CD2     CA     -0.115     19
        HD2     HP      0.115     20
        CE2     CA     -0.115     21
        HE2     HP      0.115     22
        C       C       0.51      23
        O       O      -0.51      24
 [ bonds ]
        CT      HT1
        CT      HT2
        CT      HT3
        NH3     CT
        NH3     H1
        NH3     H2
        CB      CA
        CG      CB
        CD2     CG
        CE1     CD1
        CZ      CE2
        NH3     CA
        C       CA
        C       +N
        CA      HA
        CB      HB1
        CB      HB2
```

```
        CD1     HD1
        CD2     HD2
        CE1     HE1
        O       C
        CD1     CG
        CZ      CE1
        CE2     CD2
        CE2     HE2
        CZ      HZ
[ impropers ]
        NP      -C      CA      HN
        C       CA      +N      O
[ cmap ]
        -C      NP      CA      C       +N
```

Bibliography

1. Mogensen, T. H. Pathogen Recognition and Inflammatory Signaling in Innate Immune Defenses. *Clin. Microbiol. Rev.* **22,** 240–273 (2009).
2. Kaplan, D. R. & Miller, F. D. Neurotrophin signal transduction in the nervous system. *Curr. Opin. Neurobiol.* **10,** 381–391 (2000).
3. Davidson, A. L. *et al.* Structure, Function, and Evolution of Bacterial ATP-Binding Cassette Systems. *Microbiol. Mol. Biol. Rev.* **72,** 317–364 (2008).
4. Whitman, W. B. *et al.* Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci.* **95,** 6578–6583 (1998).
5. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
6. Wingender, J. *et al.* Microbial Extracellular Polymeric Substances. (1991).
7. de Kievit, T. R. & Iglewski, B. H. Bacterial Quorum Sensing in Pathogenic Relationships. *Infect. Immun.* **68,** 4839–4849 (2000).
8. Hueck, C. J. Type III Protein Secretion Systems in Bacterial Pathogens of Animals and Plants. *Microbiol. Mol. Biol. Rev.* **62,** 379–433 (1998).
9. Porter, J. R. Antony van Leeuwenhoek: tercentenary of his discovery of bacteria. *Bacteriol. Rev.* **40,** 260–269 (1976).
10. Gram, H. C. Über die isolierte Färbung der Schizomyceten in Schnitt- und Trockenpräparaten. *Fortschr. Med.* **2,** 185–189 (1884).
11. Nikaido, H. & Vaara, M. Molecular basis of bacterial outer membrane permeability. *Microbiol. Rev.* **49,** 1–32 (1985).
12. Silhavy, T. J. *et al.* The bacterial cell envelope. *Cold Spring Harb. Perspect. Biol.* **2,** a000414 (2010).
13. Kanonenberg, K. *et al.* Type I secretion systems - a story of appendices. *Res. Microbiol.* **164,** 596–604 (2013).
14. Voulhoux, R. *et al.* Involvement of the twin-arginine translocation system in protein secretion via the type II pathway. *Embo J* **20,** 6735–6741 (2001).
15. Tosi, T. *et al.* Structural basis of eukaryotic cell targeting by type III secretion system (T3SS) effectors. *Res. Microbiol.* **164,** 605–19 (2013).
16. Cascales, E. & Christie, P. J. The versatile bacterial type IV secretion systems. *Nat Rev Micro* **1,** 137–149 (2003).
17. Jacob-Dubuisson, F. *et al.* Two-partner secretion: as simple as it sounds? *Res. Microbiol.* **164,** 583–95 (2013).
18. Henderson, I. R. *et al.* Type V Protein Secretion Pathway: the Autotransporter Story. *Microbiol. Mol. Biol. Rev.* **68,** 692–744 (2004).
19. Coulthurst, S. J. The Type VI secretion system - a widespread and versatile cell targeting system. *Res. Microbiol.* **164,** 640–54 (2013).
20. Houben, E. N. G. *et al.* Take five - Type VII secretion systems of Mycobacteria. *Biochim.*

Biophys. Acta **1843,** 1707–16 (2014).

21. Barnhart, M. M. & Chapman, M. R. Curli Biogenesis and Function. *Annu. Rev. Microbiol.* **60,** 131–147 (2006).

22. McBride, M. J. & Zhu, Y. Gliding Motility and Por Secretion System Genes Are Widespread among Members of the Phylum Bacteroidetes. *J. Bacteriol.* **195,** 270–278 (2013).

23. Sato, K. *et al.* A protein secretion system linked to bacteroidete gliding motility and pathogenesis. *Proc Natl Acad Sci U S A* **107,** 276–281 (2010).

24. Lloubes, R. *et al.* Non classical secretion systems. *Res. Microbiol.* **164,** 655–63 (2013).

25. Bleves, S. *et al.* Protein secretion systems in Pseudomonas aeruginosa: A wealth of pathogenic weapons. *Int. J. Med. Microbiol.* **300,** 534–43 (2010).

26. Filloux, A. The underlying mechanisms of type II protein secretion. *Biochim. Biophys. Acta* **1694,** 163–79 (2004).

27. Shi, L. *et al.* Direct involvement of type II secretion system in extracellular translocation of Shewanella oneidensis outer membrane cytochromes MtrC and OmcA. *J. Bacteriol.* **190,** 5512–6 (2008).

28. Sandkvist, M. Type II Secretion and Pathogenesis MINIREVIEW. *Infect Immun.* **69,** (2001).

29. Korotkov, K. V. *et al.* The type II secretion system: biogenesis, molecular architecture and mechanism. *Nat. Rev. Microbiol.* **10,** 336–51 (2012).

30. Douzi, B. *et al.* On the path to uncover the bacterial type II secretion system. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **367,** 1059–72 (2012).

31. Cisneros, D. A. *et al.* Minor pseudopilin self-assembly primes type II secretion pseudopilus elongation. *EMBO J.* **31,** 1041–53 (2012).

32. Planet, P. J. *et al.* Phylogeny of genes for secretion NTPases: identification of the widespread tadA subfamily and development of a diagnostic key for gene classification. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 2503–8 (2001).

33. Camberg, J. L. *et al.* Synergistic stimulation of EpsE ATP hydrolysis by EpsL and acidic phospholipids. *EMBO J.* **26,** 19–27 (2007).

34. Patrick, M., *et al.* Oligomerization of EpsE coordinates residues from multiple subunits to facilitate ATPase activity. *J. Biol. Chem.* **286,** 10378–86 (2011).

35. Tosi, T. *et al.* Pilotin-secretin recognition in the type II secretion system of Klebsiella oxytoca. *Mol. Microbiol.* **82,** 1422–32 (2011).

36. Collin, S. *et al.* Sorting of an integral outer membrane protein via the lipoprotein-specific Lol pathway and a dedicated lipoprotein pilotin. *Mol. Microbiol.* **80,** 655–65 (2011).

37. Hardie, K. R. *et al.* Insertion of an outer membrane protein in Escherichia coli requires a chaperone-like protein. *EMBO J.* **15,** 978–88 (1996).

38. Py, B. *et al.* An inner membrane platform in the type II secretion machinery of Gram-negative bacteria. *EMBO Rep.* **2,** 244–8 (2001).

39. Possot, O. M. *et al.* Energy requirement for pullulanase secretion by the main terminal branch of the general secretory pathway. *Mol. Microbiol.* **24,** 457–464 (1997).

40. Letellier, L. *et al.* Studies on the energetics of proaerolysin secretion across the outer membrane of Aeromonas species. Evidence for a requirement for both the protonmotive force and ATP. *J. Biol. Chem.* **272,** 11109–11113 (1997).

41. Hobbs, M. J. Common components in the assembly of type 4 fimbriae, DNA transfer systems, filamentous phage and protein-secretion apparatus: a general system for the formation of surface-associated protein complexes. *Mol Microbiol* **10,** 233–43 (1993).

42. Peabody, C. R. Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. *Microbiology* **149,** 3051–3072 (2003).

43. Campos, M. *et al.* The type II secretion system - a dynamic fiber assembly nanomachine. *Res. Microbiol.* 1–11 (2013).

44. Vignon, G. *et al.* Type IV-Like Pili Formed by the Type II Secreton: Specificity, Composition, Bundling, Polar Localization, and Surface Presentation of Peptides. **185,** 3416–3428 (2003).

45. D'Enfert, C. & Pugsley, A. P. Klebsiella pneumoniae pulS gene encodes an outer membrane lipoprotein required for pullulanase secretion. *J. Bacteriol.* **171,** 3673–9 (1989).

46. Possot, O. & Pugsley, A. P. Molecular characterization of PulE, a protein required for pullulanase secretion. *Mol. Microbiol.* 287–299 (1994).

47. Sandkvist, M. *et al.* General secretion pathway (eps) genes required for toxin secretion and outer membrane biogenesis in Vibrio cholerae. *J. Bacteriol.* **179,** 6994–7003 (1997).

48. Tauschek, M. *et al.* Identification of a protein secretory pathway for the secretion of heat-labile enterotoxin by an enterotoxigenic strain of Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 7066–71 (2002).

49. Kulkarni, R. *et al.* Roles of putative type II secretion and type IV pilus systems in the virulence of uropathogenic Escherichia coli. *PLoS One* **4,** e4752 (2009).

50. Lathem, W. W. *et al.* StcE, a metalloprotease secreted by Escherichia coli O157:H7, specifically cleaves C1 esterase inhibitor. *Mol. Microbiol.* **45,** 277–88 (2002).

51. Bally, M. *et al.* Protein secretion in Pseudomonas aeruginosa: characterization of seven xcp genes and processing of secretory apparatus components by prepilin peptidase. *Mol Microbiol.* **6,** 1121–31 (1992).

52. Jyot, J. *et al.* Type II secretion system of Pseudomonas aeruginosa: in vivo evidence of a significant role in death due to lung infection. *J. Infect. Dis.* **203,** 1369–77 (2011).

53. Rossier, O. *et al.* Legionella pneumophila Type II Protein Secretion Promotes Virulence in the A / J Mouse Model of Legionnaires ' Disease Pneumonia. *Infect. Immun.* **72,** 310-321 (2004).

54. Jiang, B. & Howard, S. P. The Aeromonas hydrophila exeE gene, required both for protein secretion and normal outer membrane biogenesis, is a member of a general secretion pathway. *Mol. Microbiol.* 1351–1361 (1992).

55. Stam, M. R. *et al.* Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Eng. Des. Sel.* **19,** 555–62 (2006).

56. Kovacs-Simon, A. *et al.* Lipoproteins of Bacterial Pathogens. *Infect. Immun.* **79,** 548–561 (2011).

57. Zückert, W. R. Secretion of bacterial lipoproteins: through the cytoplasmic membrane, the

periplasm and beyond. *Biochim. Biophys. Acta* **1843,** 1509–16 (2014).

58. Pugsley, A. P. The complete general secretory pathway in gram-negative bacteria. *Microbiol. Rev.* **57,** 50–108 (1993).

59. d'Enfert C. *et al.* Export and secretion of the lipoprotein pullulanase by Klebsiella pneumoniae. *Mol Microbiol.* **1,** 107–16 (1987).

60. d'Enfert, C. *et al.* Cloning and expression in Escherichia coli of the Klebsiella pneumoniae genes for production, surface localization and secretion of the lipoprotein pullulanase. *EMBO J.* **6,** 3531–8 (1987).

61. East, A. *et al.* Structural Basis of Pullulanase Membrane Binding and Secretion Revealed by X-Ray Crystallography, Molecular Dynamics and Biochemical Analysis. *Structure* **24,** 92–104 (2015).

62. Pugsley, A. P. *et al.* Extracellular pullulanase of Klebsiella pneumoniae is a lipoprotein. *J. Bacteriol.* **166,** 1083–8 (1986).

63. Yamaguchi, K. *et al.* A single amino acid determinant of the membrane localization of lipoproteins in E. coli. *Cell* **53,** 423–432 (2015).

64. Tokuda, H. & Matsuyama, S.-I. Sorting of lipoproteins to the outer membrane in E. coli. *Biochim. Biophys. Acta* **1693,** 5–13 (2004).

65. Okuda, S. *et al.* Lipoprotein Sorting in Bacteria. *Annu. Rev. Microbiol.* **65,** 239–259 (2011).

66. Fukuda, A. *et al.* Aminoacylation of the N-terminal cysteine is essential for Lol-dependent release of lipoproteins from membranes but does not depend on lipoprotein sorting signals. *J. Biol. Chem.* **277,** 43512–43518 (2002).

67. Masuda, K. *et al.* Elucidation of the function of lipoprotein-sorting signals that determine membrane localization. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 7390–7395 (2002).

68. Sakamoto, C. *et al.* Novel mutations of the LolCDE complex causing outer membrane localization of lipoproteins despite their inner membrane-retention signals. *Biochem. Biophys. Res. Commun.* **401,** 586–591 (2010).

69. Narita, S. *et al.* A mutation in the membrane subunit of an ABC transporter LolCDE complex causing outer membrane localization of lipoproteins against their inner membrane-specific signals. *Mol. Microbiol.* **49,** 167–177 (2003).

70. Lombard, V. *et al.* The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42,** D490-5 (2014).

71. Seydel, A. *et al.* Testing the '+2 rule' for lipoprotein sorting in the Escherichia coli cell envelope with a new genetic selection. *Mol. Microbiol.* **34,** 810–821 (1999).

72. Hara, T. *et al.* Mechanism underlying the inner membrane retention of Escherichia coli lipoproteins caused by Lol avoidance signals. *J. Biol. Chem.* **278,** 40408–14 (2003).

73. Gennity, J. M. & Inouye, M. The protein sequence responsible for lipoprotein membrane localization in Escherichia coli exhibits remarkable specificity. *J. Biol. Chem.* **266,** 16458–16464 (1991).

74. Klein, J. R. *et al.* Molecular analysis and nucleotide sequence of the envCD operon of Escherichia coli. *Mol. Gen. Genet.* **230,** 230–240 (1991).

75. Seiffer, D. *et al.* EnvC, a new lipoprotein of the cytoplasmic membrane of Escherichia coli. *FEMS Microbiol. Lett.* **107,** 175–178 (1993).

76. Terada, M. *et al.* Lipoprotein Sorting Signals Evaluated as the LolA-dependent Release of Lipoproteins from the Cytoplasmic Membrane of Escherichia coli. *J. Biol. Chem.* **276,** 47690–47694 (2001).

77. DeChavigny A. *et al.* Sequence and inactivation of the pss gene of Escherichia coli: Phosphatidylethanolamine may not be essential for cell viability. *J. Biol. Chem.* **266,** 5323–5332 (1991).

78. Abendroth, J. *et al.* The three-dimensional structure of the cytoplasmic domains of EpsF from the type 2 secretion system of Vibrio cholerae. *J. Struct. Biol.* **166,** 303–15 (2009).

79. Reyss, I. & Pugsley, A. P. Five additional genes in the pulC±O operon of the Gram-negative bacterium Klebsiella oxytoca UNF5023 that are required for pullulanase secretion. *Mol. Gen. Genet* **222,** 176 (1990).

80. Francetic, O. *et al.* Signal recognition particle-dependent inner membrane targeting of the PulG Pseudopilin component of a type II secretion system. *J. Bacteriol.* **189,** 1783–93 (2007).

81. Nunn, D. N. & Lory, S. Product of the Pseudomonas aeruginosa gene pilD is a prepilin leader peptidase. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 3281–5 (1991).

82. Strom, M. S. *et al.* A single bifunctional enzyme, PilD, catalyzes cleavage and N-methylation of proteins belonging to the type IV pilin family. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 2404–8 (1993).

83. Köhler, R. *et al.* Structure and assembly of the pseudopilin PulG. *Mol. Microbiol.* **54,** 647–64 (2004).

84. Korotkov, K. *et al.* Structure of the GspK-GspI-GspJ complex from the enterotoxigenic Escherichia coli type 2 secretion system. *Nat. Struct. Mol. Biol.* **15,** 462–8 (2008).

85. Durand, E. *et al.* XcpX controls biogenesis of the Pseudomonas aeruginosa XcpT-containing pseudopilus. *J Biol Chem.* **280,** 31378–89 (2005).

86. Possot, O. M. *et al.* Multiple interactions between pullulanase secreton components involved in stabilization and cytoplasmic membrane association of PulE. *J. Bacteriol.* **182,** 2142–52 (2000).

87. Parge, H.E. *et al.* Structure of the fibre-forming protein pilin at 2.6 Å resolution. *Nature* **378,** 32–38 (1995).

88. Nivaskumar, M. *et al.* Distinct docking and stabilization steps of the Pseudopilus conformational transition path suggest rotational assembly of type IV pilus-like fibers. *Structure* **22,** 685–696 (2014).

89. Campos, M. *et al.* Detailed structural and assembly model of the type II secretion pilus from sparse data. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 13081–6 (2010).

90. Craig, L. *et al.* Type IV pilin structure and assembly: X-ray and EM analyses of Vibrio cholerae toxin-coregulated pilus and Pseudomonas aeruginosa PAK pilin. *Mol. Cell* **11,** 1139–50 (2003).

91. Korotkov, K. V *et al.* Calcium is essential for the major pseudopilin in the type 2 secretion system. *J. Biol. Chem.* **284,** 25466–70 (2009).

92. Nunn, D. Bacterial type II protein export and pilus biogenesis: more than just homologies? *Trends Cell Biol.* **9,** 402–8 (1999).

93. Douzi, B. *et al.* Deciphering the Xcp Pseudomonas aeruginosa type II secretion machinery through multiple interactions with substrates. *J. Biol. Chem.* **286,** 40792–801 (2011).

94. Alm, R.A. & Mattick, J. S. Common architecture of type 4 fimbriae and complexes involved in macromolecular traffic. *Trends Microbiol.* 411–413 (1995).

95. Mattick, J. S. Type IV pili and twitching motility. *Annu. Rev. Microbiol.* **56,** 289–314 (2002).

96. Ghosh, A. & Albers, S.-V. Assembly and function of the archaeal flagellum. *Biochem. Soc. Trans.* **39,** 64–9 (2011).

97. Sandkvist, M. *et al.* Direct interaction of the epsl and epsm proteins of the general secretion apparatus in vibrio cholerae. *J. Bacteriol.* **181,** 3129–3135 (1999).

98. Michel, G. *et al.* Mutual stabilization of the XcpZ and XcpY components of the secretory apparatus in Pseudomonas aeruginosa. *Microbiology* **144,** 3379–3386 (1998).

99. Robert, V. *et al.* Subcomplexes from the Xcp secretion system of Pseudomonas aeruginosa. *FEMS Microbiol. Lett.* **252,** 43–50 (2005).

100. Lybarger, S. R. *et al.* Docking and assembly of the type II secretion complex of Vibrio cholerae. *J. Bacteriol.* **191,** 3149–3161 (2009).

101. Karimova, G. *et al.* A bacterial two-hybrid system based on a reconstituted signal transduction pathway. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 5752–6 (1998).

102. Georgiadou, M. *et al.* Large-scale study of the interactions between proteins involved in type IV pilus biology in Neisseria meningitidis: characterization of a subcomplex involved in pilus assembly. *Mol. Microbiol.* **84,** 857–873 (2012).

103. Karuppiah, V. *et al.* Structure and assembly of an inner membrane platform for initiation of type IV pilus biogenesis. *Proc. Natl. Acad. Sci.* **110,** E4638–E4647 (2013).

104. Abendroth, J. *et al.* The crystal structure of the periplasmic domain of the type II secretion system protein EpsM from Vibrio cholerae: the simplest version of the ferredoxin fold. *J. Mol. Biol.* **338,** 585–596 (2004).

105. Ayers, M. *et al.* PilM/N/O/P proteins form an inner membrane complex that affects the stability of the Pseudomonas aeruginosa type IV pilus secretin. *J. Mol. Biol.* **394,** 128–142 (2009).

106. Tammam, S. *et al.* PilMNOPQ from the Pseudomonas aeruginosa type IV pilus system form a transenvelope protein interaction network that interacts with PilA. *J. Bacteriol.* **195,** 2126–2135 (2013).

107. Sampaleanu, L. M. *et al.* Periplasmic domains of Pseudomonas aeruginosa PilN and PilO form a stable heterodimeric complex. *J. Mol. Biol.* **394,** 143–159 (2009).

108. Douet, V. *et al.* Systematic analysis, by the yeast two-hybrid, of protein interaction between components of the type II secretory machinery of Erwinia chrysanthemi. *Res. Microbiol.* **155,** 71–75 (2004).

109. Brisse, S. *et al.* The Genus Klebsiella Taxonomic History and Structure. *The Prokaryotes* 159–196 (2006). doi:10.1007/0-387-30746-X_8

110. Horan, T. *et al.* Pathogens causing nosocomial infections preliminary data from the national nosocomial infections surveillance system. *Antimicrob. Newsl.* **5,** 65–67 (1988).

111. Högenauer, C. *et al.* Klebsiella oxytoca as a Causative Organism of Antibiotic-Associated Hemorrhagic Colitis. *N. Engl. J. Med.* **355,** 2418–2426 (2006).

112. University of Minnesota College of Veterinary Medicine. *Klebsiella Mastitis Pathogen Factsheet #2*. (2014).

113. Wretlind, B. & Pavlovskis, O. R. Genetic mapping and characterization of Pseudomonas aeruginosa mutants defective in the formation of extracellular proteins. *J. Bacteriol.* **158,** 801–808 (1984).

114. Martinez, A. *et al.* LipC, a second lipase of Pseudomonas aeruginosa, is LipB and Xcp dependent and is transcriptionally regulated by pilus biogenesis components. *Mol. Microbiol.* **34,** 317–326 (1999).

115. Lindgren, V. & Wretlind, B. Characterization of a Pseudomonas aeruginosa transposon insertion mutant with defective release of exoenzymes. *J. Gen. Microbiol.* **133,** 675–681 (1987).

116. Davis, B. M. *et al.* Convergence of the secretory pathways for cholera toxin and the filamentous phage, CTXphi. *Science* **288,** 333–335 (2000).

117. Connell, T. D. *et al.* Endochitinase is transported to the extracellular milieu by the eps-encoded general secretory pathway of Vibrio cholerae. *J. Bacteriol.* **180,** 5591–5600 (1998).

118. Liles, M. R. *et al.* The prepilin peptidase is required for protein secretion by and the virulence of the intracellular pathogen Legionella pneumophila. *Mol. Microbiol.* **31,** 959–970 (1999).

119. Hales, L. M. & Shuman, H. A. Legionella pneumophila contains a type II general secretion pathway required for growth in amoebae as well as for secretion of the Msp protease. *Infect. Immun.* **67,** 3662–3666 (1999).

120. Aragon, V. *et al.* Secreted enzymatic activities of wild-type and pilD-deficient Legionella pneumophila. *Infect. Immun.* **68,** 1855–1863 (2000).

121. Hii, S. L. *et al.* Pullulanase: Role in Starch Hydrolysis and Potential Industrial Applications. *Enzyme Res.* **2012,** (2012).

122. Jensen, B. F. & Norman, B. E. Bacillus acidopullulyticus Pullulanase: application and regulatory aspects for use in the food industry. *Process Biochem.* **19,** (1984).

123. Shaw, J.-F. & Sheu, J.-R. Production of High-maltose Syrup and High-protein Flour from Rice by an Enzymatic Method. *Biosci. Biotechnol. Biochem.* **56,** 1071–1073 (1992).

124. Olsen, H. S. *et al.* Enzymes at Work: A Concise Guide to Industrial Enzymes and their Uses. (2000).

125. Nigam, P. & Singh, D. Enzyme and microbial systems involved in starch processing. *Enzyme Microb. Technol.* **17,** 770–778 (1995).

126. Bird, A. R. *et al.* Starches, resistant starches, the gut microflora and human health. *Curr. Issues Intest. Microbiol.* **1,** 25–37 (2000).

127. Roy, A. *et al.* Isolation and purification of an acidic pullulanase type II from newly isolated Bacillus sp. US149. *Enzyme Microb. Technol.* **33,** 720–724 (2003).

128. Rendleman, J. A. Jr. Enhancement of cyclodextrin production through use of debranching enzymes. *Biotechnol. Appl. Biochem.* **26,** 51–61 (1997).

129. Kim, Y.-K. & Robyt J. F. Enzyme modification of starch granules: Formation and retention of cyclomaltodextrins inside starch granules by reaction of cyclomaltodextrin glucanosyltransferase with solid granules. *Carbohydr. Res.* **328,** 509–515 (2000).

130. Garavito, R. M. & Ferguson-Miller, S. Detergents as Tools in Membrane Biochemistry. *J. Biol. Chem.* **276,** 32403–32406 (2001).

131. Pfleger, K. D. G. & Eidne, K. A. Illuminating insights into protein-protein interactions using bioluminescence resonance energy transfer (BRET). *Nat. Meth.* **3,** 165–174 (2006).

132. Hu, H. *et al.* Structure of the type VI secretion phospholipase effector Tle1 provides insight into its hydrolysis and membrane targeting. *Acta Crystallogr. D. Biol. Crystallogr.* **70,** 2175–2185 (2014).

133. Pei, Q. *et al.* Computational investigation of the enzymatic mechanisms of phosphothreonine lyase. *Biophys. Chem.* **157,** 16–23 (2011).

134. Dey, S. & Datta, S. Interfacial residues of SpcS chaperone affects binding of effector toxin ExoT in Pseudomonas aeruginosa: novel insights from structural and computational studies. *FEBS J.* **281,** 1267–1280 (2014).

135. Yao, Q. *et al.* Structural mechanism of ubiquitin and NEDD8 deamidation catalyzed by bacterial effectors that induce macrophage-specific apoptosis. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 20395–20400 (2012).

136. Kudryashev, M. *et al.* In situ structural analysis of the Yersinia enterocolitica injectisome. *Elife* **2,** e00792 (2013).

137. Meshcheryakov, V. A. *et al.* Inhibition of a type III secretion system by the deletion of a short loop in one of its membrane proteins. *Acta Crystallogr. Sect. D* **69,** 812–820 (2013).

138. Baker, J. L. *et al.* Steered Molecular Dynamics Simulations of a Type IV Pilus Probe Initial Stages of a Force-Induced Conformational Transition. *PLoS Comput Biol* **9,** e1003032 (2013).

139. Rathinavelan, T. *et al.* A repulsive electrostatic mechanism for protein export through the type III secretion apparatus. *Biophys. J.* **98,** 452–461 (2010).

140. Cronan, J. E. Bacterial Membrane Lipids: Where Do We Stand? *Annu. Rev. Microbiol.* **57,** 203–224 (2003).

141. Rahman, M. M. *et al.* The membrane phospholipids of Neisseria meningitidis and Neisseria gonorrhoeae as characterized by fast atom bombardment mass spectrometry. *Microbiology* **146 (Pt 8),** 1901–1911 (2000).

142. Davis, A. M. *et al.* Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew. Chem. Int. Ed. Engl.* **42,** 2718–36 (2003).

143. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9,** 646–52 (2002).

144. Durrant, J. D. & McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biol.* **9,** 71 (2011).

145. Best, R. B. Atomistic molecular simulations of protein folding. *Curr. Opin. Struct. Biol.* **22,** 52–61 (2012).

146. Bond, P. J. *et al.* Membrane simulations of OpcA: gating in the loops? *Biophys. J.* **92,** L23-5 (2007).

147. Holdbrook, D. A. *et al.* Stability and membrane orientation of the fukutin transmembrane domain: a combined multiscale molecular dynamics and circular dichroism study. *Biochemistry* **49,** 10796–802 (2010).

148. Lindorff-Larsen, K. *et al.* Simultaneous determination of protein structure and dynamics. *Nature* **433,** 128–32 (2005).

149. Shimamura, T. *et al.* Molecular basis of alternating access membrane transport by the sodium-hydantoin transporter Mhp1. *Science* **328,** 470–3 (2010).

150. Berendsen, H. J. C. *et al.* Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81,** 3684–3690 (1984).

151. Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **72,** 2384–2393 (1980).

152. Parrinello, M. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52,** 7182 (1981).

153. Leach, A. R. Molecular modelling: principles and applications. (2001).

154. Oostenbrink, C. *et al.* A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **25,** 1656–76 (2004).

155. Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **15,** 144–150 (2005).

156. Rick, S. W. *et al.* Dynamical Fluctuating Charge Force Fields: Application to Liquid Water. *J. Chem. Phys.***101,** 29 (1994).

157. Shi, Y. *et al.* The Polarizable Atomic Multipole-based AMOEBA Force Field for Proteins. *J Chem Theory Comput.* **9,** 4046–4063 (2013).

158. Brooks, B. R. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4,** 187–217 (1983).

159. Brooks, B. R. *et al.* CHARMM : The Biomolecular Simulation Program. *J Comput Chem* **30,** 1545–1614 (2009).

160. van Gunsteren, W. F. & Berendsen , H. J. C. Groningen Molecular Simulation (GROMOS) Library Manual. 1–221 (1987).

161. Best, R. B. & Hummer, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B* **113,** 9004–15 (2009).

162. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117,** 5179–5197 (1995).

163. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78,** 1950–8 (2010).

164. Kaminski, G. A. *et al.* Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **105,** 6474–6487 (2001).

165. Lopes, P. E. M. *et al.* Polarizable Empirical Force Field for Aromatic Compounds Based on the Classical Drude Oscillator. *J Phys Chem B.* **22,** 2873–2885 (2007).

166. Marrink, S. J. *et al.* The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **111,** 7812–7824 (2007).

167. Bahar, I. *et al.* Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **2,** 173–181 (1997).

168. Tama, F. *et al.* Exploring Global Distortions of Biological Macromolecules and Assemblies from Low-resolution Structural Information and Elastic Network Theory. *J. Mol. Biol.* **321,** 297–305 (2002).

169. Monticelli, L. *et al.* The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **4,** 819–834 (2008).

170. Bond, P. J. *et al.* Coarse-grained molecular dynamics simulations of membrane proteins and peptides. *J. Struct. Biol.* **157,** 593–605 (2007).

171. Russ, A. P. & Lampel, S. The druggable genome: an update. *Drug Discov. Today* **10,** 1607–10 (2005).

172. van Meer, G. *et al.* Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell Biol.* **9,** 112–124 (2008).

173. Lingwood, D. & Simons, K. Lipid rafts as a membrane-organizing principle. *Science* **327,** 46–50 (2010).

174. Lee, A. G. Biological membranes: the importance of molecular detail. *Trends Biochem. Sci.* **36,** 493–500 (2011).

175. Klauda, J. B. *et al.* Update of the CHARMM all-atom additive force field for lipids: Validation on six lipid types. *J Phys Chem B.* **114,** 7830–7843 (2011).

176. Kleinschmidt, J. H. Lipid-Protein Interactions. Methods and protocols. (2013).

177. Marrink, S. J. *et al.* Coarse Grained Model for Semiquantitative Lipid Simulations. *J. Phys. Chem. B* **108,** 750–760 (2004).

178. Shaw, D. E. *et al.* Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **51,** 91 (2008).

179. Verlet, L. Computer 'Experiments' on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **159,** 98–103 (1967).

180. Ryckaert, J.-P. *et al.* Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23,** 327–341 (1977).

181. Hess, B. *et al.* LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18,** 1463–1472 (1997).

182. Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann. Phys.* **369,** 253–287 (1921).

183. Darden, T. *et al.* Particle mesh Ewald: An Nlog(N) method for Ewald sums in Large Systems. *J. Chem. Phys.* 10089–10092 (1993).

184. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J Chem Phys* **103,** 31–34 (1995).

185. Hess, B. *et al.* GROMACS 4 : Algorithms for Highly Efficient , Load-Balanced , and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **4,** 435–447 (2008).

186. Mackerell, A. D. *et al.* Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **25,** 1400–15 (2004).

187. Wolf, M. G. *et al.* g _ membed : Efficient Insertion of a Membrane Protein into an Equilibrated Lipid Bilayer with Minimal Perturbation. *J Comput Chem* **31,** 2169–2174 (2010).

188. Jorgensen, W. L. *et al.* Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** (1983).

189. Bussi, G. *et al.* Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126,** 14101 (2007).

190. Berendsen, H. J. C. *et al.* GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91,** 43–56 (1995).

191. Humphrey, W. *et al.* VMD: visual molecular dynamics. *J. Mol. Graph.* **14,** 33–8, 27–8 (1996).

192. Francetic, O. & Pugsley, A. P. Towards the identification of type II secretion signals in a nonacylated variant of pullulanase from Klebsiella oxytoca. *J. Bacteriol.* **187,** 7045–7055 (2005).

193. Mikami, B. *et al.* Crystal structure of pullulanase: evidence for parallel binding of oligosaccharides in the active site. *J. Mol. Biol.* **359,** 690–707 (2006).

194. Duffner, F. *et al.* A new thermoactive pullulanase from Desulfurococcus mucosus: cloning, sequencing, purification, and characterization of the recombinant enzyme after expression in Bacillus subtilis. *J. Bacteriol.* **182,** 6331–8 (2000).

195. Gourlay, L. J. *et al.* Group B Streptococcus Pullulanase Crystal Structures in the Context of a Novel Strategy for Vaccine Development. *J Bacteriol* **191,** 3544–3552 (2009).

196. Hondoh, H. *et al.* Three-dimensional Structure and Substrate Binding of Bacillus stearothermophilus Neopullulanase. *J. Mol. Biol.* **326,** 177–188 (2003).

197. Ohtaki, A. *et al.* Complex structures of Thermoactinomyces vulgaris R-47 alpha-amylase 2 with acarbose and cyclodextrins demonstrate the multiple substrate recognition mechanism. *J. Biol. Chem.* **279,** 31033–40 (2004).

198. Wu, H. C. *et al.* Biogenesis of membrane lipoproteins in Escherichia coli. *Biophys. J.* **37,** 307–315 (1982).

199. Buddelmeijer, N. The molecular mechanism of bacterial lipoprotein modification--how, when and why? *FEMS Microbiol. Rev.* **39,** 246–261 (2015).

200. Leyton, D. L. *et al.* From self sufficiency to dependence: mechanisms and factors important for autotransporter biogenesis. *Nat. Rev. Microbiol.* **10,** 213–25 (2012).

201. Schulze, R. J. & Zückert, W. R. Borrelia burgdorferi lipoproteins are secreted to the outer surface by default. *Mol. Microbiol.* **59,** 1473–1484 (2006).

202. Pugsley, A. P. & Kornacker, M. G. Secretion of the cell surface lipoprotein pullulanase in Escherichia coli. *J. Biol. Chem.* **266,** 13640–13645 (1991).

203. Sauvonnet, N. & Pugsley, A. P. Identification of two regions of Klebsiella oxytoca pullulanase that together are capable of promoting beta-lactamase secretion by the general secretory pathway. *Mol. Microbiol.* **22,** 1–7 (1996).

204. Eswar, N. *et al.* Comparative protein structure modelling using Modeller. *Curr. Protoc. Bioinformatics* **5** (2006).

205. Mekalanos, J. J. Environmental signals controlling expression of virulence determinants in bacteria. *J. Bacteriol.* **174,** 1–7 (1992).

206. Clapham, D. E. Calcium signalling. *Cell* **131,** 1047–1058 (2007).

207. Pettersson, J. *et al.* Modulation of virulence factor expression by pathogen target cell contact. *Science* **273,** 1231–1233 (1996).

208. Sanders, D. *et al.* Communicating with calcium. *Plant Cell* **11,** 691–706 (1999).

209. Whitaker, M. & Larman, M. G. Calcium and mitosis. *Semin. Cell Dev. Biol.* **12,** 53–58 (2001).

210. Ikura, M. *et al.* The role of calcium-binding proteins in the control of transcription: structure to function. *Bioessays* **24,** 625–636 (2002).

211. Dominguez, D. C. Calcium signalling in bacteria. *Mol. Microbiol.* **54,** 291–297 (2004).

212. Morgan, R. O. *et al.* Deciphering function and mechanism of calcium-binding proteins from their evolutionary imprints. *Biochim. Biophys. Acta - Mol. Cell Res.* **1763,** 1238–1249 (2006).

213. Clifton, L. A. *et al.* Effect of Divalent Cation Removal on the Structure of Gram-Negative Bacterial Outer Membrane Models. *Langmuir* **31,** 404–412 (2015).

214. Gangola, P. & Rosen, B. P. Maintenance of intracellular calcium in Escherichia coli. *J. Biol. Chem.* **262,** 12570–12574 (1987).

215. Knight, M. R. *et al.* Recombinant aequorin as a probe for cytosolic free Ca2+ in Escherichia coli. *FEBS Lett.* **282,** 405–408 (1991).

216. Torrecilla, I. *et al.* Use of recombinant aequorin to study calcium homeostasis and monitor calcium transients in response to heat and cold shock in cyanobacteria. *Plant Physiol.* **123,** 161–176 (2000).

217. Jones, H. E. *et al.* Slow changes in cytosolic free Ca2+ in Escherichia coli highlight two putative influx mechanisms in response to changes in extracellular calcium. *Cell Calcium* **25,** 265–274 (1999).

218. Jones, H. E. *et al.* Direct measurement of free Ca2+ shows different regulation of Ca2+ between the periplasm and the cytosol of Escherichia coli. *Cell Calcium* **32,** 183–192 (2002).

219. Yamashita, M. *et al.* Amino acid residues specific for the catalytic action towards α-1,6-glucosidic linkages in Klebsiella pullulanase. *J. Ferment. Bioeng.* **84,** 283–290 (1997).

220. Bertoldo, C., *et al.* Pullulanase type I from Fervidobacterium pennavorans Ven5: cloning, sequencing, and expression of the gene and biochemical characterization of the recombinant enzyme. *Appl. Environ. Microbiol.* **65,** 2084–2091 (1999).

221. Saha, B. C. *et al.* Purification and characterization of a highly thermostable novel pullulanase from Clostridium thermohydrosulfuricum. *Biochem. J.* **252,** 343–348 (1988).

222. Iwamoto, H. *et al.* Interaction between pullulanase from Klebsiella pneumoniae and cyclodextrins. *J. Biochem.* **113,** 93–96 (1993).

223. Iwamoto, H. *et al.* Comparison of the binding of beta-cyclodextrin and alpha- and gamma-cyclodextrins with pullulanase from Klebsiella pneumoniae as studied by equilibrium and kinetic fluorometry. *J. Biochem.* **116,** 1264–1268 (1994).

224. MacGregor, E. A. *et al.* Relationship of sequence and structure to specificity in the alpha-amylase family of enzymes. *Biochim. Biophys. Acta* **1546,** 1–20 (2001).

225. Janecek, S. *et al.* Relation between domain evolution, specificity, and taxonomy of the alpha-amylase family members containing a C-terminal starch-binding domain. *Eur. J. Biochem.* **270,** 635–645 (2003).

226. Jespersen, H. M. *et al.* Starch- and glycogen-debranching and branching enzymes: prediction of structural features of the catalytic (beta/alpha)8-barrel domain and evolutionary relationship to other amylolytic enzymes. *J. Protein Chem.* **12,** 791–805 (1993).

227. Katsuya, Y. *et al.* Three-dimensional structure of Pseudomonas isoamylase at 2.2 A resolution. *J. Mol. Biol.* **281,** 885–897 (1998).

228. Henrissat, B. & Romeu, A. Families, superfamilies and subfamilies of glycosyl hydrolases. *Biochem. J.* **311 (Pt 1,** 350–1 (1995).

229. Rodriguez-Sanoja, R. *et al.* Microbial starch-binding domain. *Curr. Opin. Microbiol.* **8,** 260–267 (2005).

230. Yanez, M. *et al.* Calcium binding proteins. *Adv. Exp. Med. Biol.* **740,** 461–482 (2012).

231. Bond, P. J. *et al.* Membrane Simulations of OpcA: Gating in the Loops? *Biophys. J.* **92,** L23–L25 (2007).

232. Sauvonnet, N. *et al.* Pilus formation and protein secretion by the same machinery in Escherichia coli. *EMBO J.* **19,** 2221–8 (2000).

233. Campos, M. *et al.* Modeling pilus structures from sparse data. *J. Struct. Biol.* **173,** 436–44 (2011).

234. Chang, Y. *et al.* Architecture of the type IVa pilus machine. *Science* **351,** (2016).

235. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292,** 195–202 (1999).

236. Nivaskumar, M. *et al.* Pseudopilin residue E5 is essential for recruitment by the type 2 secretion system assembly platform. *Mol. Microbiol.* **101,** 924–941 (2016).

237. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815 (1993).

238. Abendroth, J. *et al.* The dimer formed by the periplasmic domain of EpsL from the Type 2 Secretion System of Vibrio parahaemolyticus. *J. Struct. Biol.* **168,** 313 (2009).