Structural analysis of SGNH domains involved in lipopolysaccharide modification

Sarah Naomi Tindall

MSc by Research

University of York
Biology

September 2017

Abstract

Gram negative bacteria have lipopolysaccharides (LPS) projecting from their outer membrane which are often acetylated by acyltransferase proteins. Some of these acyltransferase proteins have additional attached SGNH domains ($SGNH^{AT3}$); while much is known about isolated SGNH domains ($SGNH^{isol}$), little is known about those attached to acyltransferases. The aim of this research was to determine any structural or mechanistic differences between $SGNH^{AT3}$ and $SGNH^{isol}$ domains. *In silico* analysis showed that while the catalytic residues are present in $SGNH^{AT3}$ domains, the oxyanion hole residues, normally conserved in $SGNH^{isol}$, are not present. The $SGNH^{AT3}$ domains of two acyltransferases from *Salmonella* ser. Paratyphi A and *Neisseria meningitidis* were expressed and the SGNH domain from *Salmonella* ser. Paratyphi A purified. Mass spectrometry showed that two disulfide bonds were present and circular dichroism determined that reduction of these disulfide bonds had no effect on the thermal stability of the protein. The structure was determined using X-ray crystallography and, although the overall fold was the same, there were many structural differences between this $SGNH^{AT3}$ domain and $SGNH^{isol}$ domains. The linker, previously thought to be flexible, was shown to be part of the $SGNH^{AT3}$ domain structure. An additional helix was seen, not present in $SGNH^{isol}$ domains, and this could potentially be important for interaction with the acyltransferase domain. In addition, docking models highlighted residues potentially important for binding, however, further analysis will need to be carried out to investigate this.

List of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Jennifer Potts for supervising me throughout the year and providing me with many suggestions and discussions. I would also like to thank the Potts group, particularly Mike Hodgkinson, Lotte van Beek and Fiona Whelan, for helping me in the lab.

I would like thank the acyltransferase group, Caroline Pearson, Laura Clark, Marjan van der Woude and Gavin Thomas for ideas, suggestions and discussions.

For help with crystallography I would like to thank Jean Wilkinson for help at each stage, Sam Hart for help with the experimental work and Johan Turkenburg and Huw Jenkins for help solving the structure.

I would like to thank Gavin Thomas and Tony Wilkinson for being on my thesis advisory panel and helping to provide ideas and support.

I would also like to thank the Proteomics and Metabolomics lab in the Technology Facility for carrying out the mass spectrometry experiments along with Andrew Leech, head of the Molecular Interactions Laboratory for teaching me how to use the circular dichroism spectrometer and the Technology Facility for access to the SEC-MALLs and circular dichroism spectrometer.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Sarah Tindall, Sept 2017

Chapter 1. Introduction

1.1. Gram negative bacteria

1.1.1. *Salmonella enterica* subsp. *enterica* serovar Paratyphi A

Paratyphoid fever, a bacterial infection caused by *Salmonella* ser. Paratyphi, infects more than 5.4 million people each year worldwide with approximately 1% of cases resulting in death (1). Symptoms of paratyphoid fever – fever, headache, diarrhoea and abdominal pain – are very similar to other febrile diseases and therefore diagnosis is challenging (2,3). This, combined with the lack of an effective diagnostic test for paratyphoid fever, results in unnecessary or incorrect antibiotic treatment (4) leading to an increasing number of *Salmonella* ser. Paratyphi isolates showing resistance to multiple antibiotics (5,6). At present, there is no licenced vaccine for protection against paratyphoid fever, a number are in development or clinical trials with the majority comprising of purified *Salmonella* ser. Paratyphi lipopolysaccharide (LPS) structures conjugated to a carrier protein (7,8). However, despite showing promise in animal models and early stage clinical trials (9,10), none are currently available for clinical use.

1.1.2. *Neisseria meningitidis* serogroup A

Neisseria meningitidis is the cause of around 1.2 million cases of meningitis per year with around 10% resulting in death (11). Serogroup A, is responsible for meningococcal epidemics in sub-Saharan Africa and regions of Asia with up to 300,000 cases per epidemic (12), the majority of these in those younger than 15 years of age (13). Symptoms of meningitidis include headache, high fever, confusion and, if not treated rapidly, can result in severe brain damage or death within 24 hours of symptom onset (14). Due to the severity of the disease and necessity for rapid treatment, prevention using vaccination is beneficial. A commonly used vaccine, protecting against four *N. meningitidis* serogroups (A,

C, W and Y) shows limited efficacy against serogroup A meningococcal (12). More recently, a polysaccharide-conjugate vaccine was developed against serotype A specifically, which has shown to be both safe and effective (15).

1.2. Lipopolysaccharides

Gram negative bacteria, for example, Salmonella ser. Paratyphi and N. meningitidis, are characterised by a membrane consisting of a layer of peptidoglycan sandwiched between two lipid bilayers (16) (Figure 1). The cytoplasmic, or inner, membrane is a phospholipid bilayer while the outer membrane is an asymmetric membrane with phospholipids on the inside and LPS on the outside projecting away from the cell (17) (Figure 1). In between these layers is a layer of peptidoglycan and the periplasm: an aqueous environment similar to the cytoplasm, densely packed with proteins. (16) (Figure 1). Peptidoglycan protects the cell from osmotic pressure and is made up of repeating units of sugar derivatives N-acetyl glucosamine (GlcNAc) and N-acetyl muramic acid (MurNAc) crosslinked by penta-peptide chains (16,17).

1.2.1. Composition

Lipopolysaccharides, sugar structures present on the outer membrane of Gram negative bacteria, are comprised of lipid A, core oligosaccharide and O-antigen (18) (Figure 1). The lipid A portion is made up of two phosphorylated glucosamine (GlcN) residues attached to 3-hydroxyl fatty acids to anchor the LPS to the outer membrane of bacteria (19-22) (Figure 1).
A conserved core oligosaccharide region links the lipid A to the O-antigen. In Salmonella, the core oligosaccharide consists of around 15 sugar residues, with the inner core (nearest the lipid A) containing keto-deoxyoctulosonate (Kdo) and heptose (Hep) sugars and the outer core (nearest the O-antigen) containing hexose sugars (20,21) (Figure 1). *N. meningitidis* contains a similar, but shorter, core structure comprising of just Kdo and heptose sugars (23) (Figure 1).

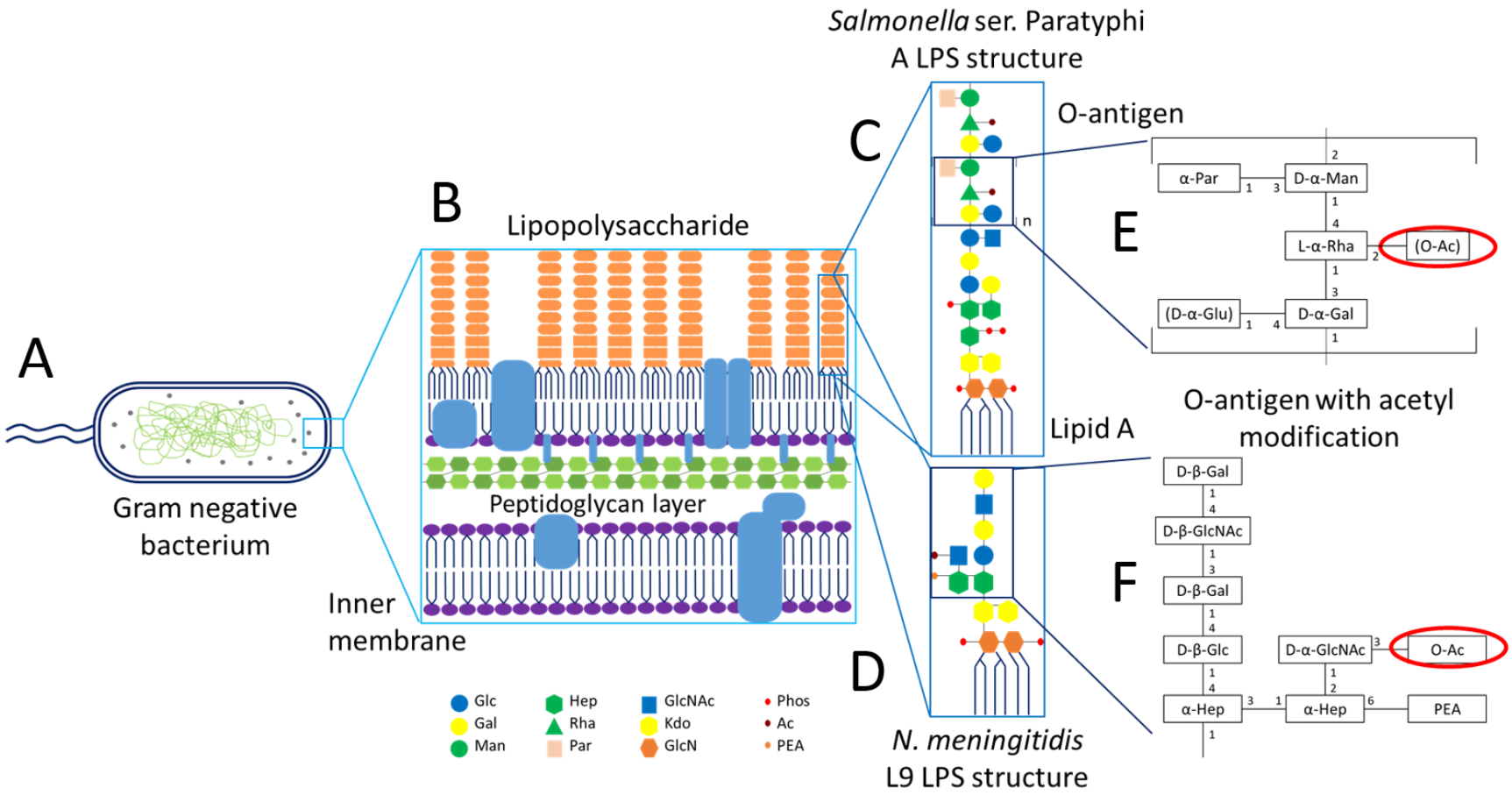


Figure 1. Structure of LPS (A) Gram negative bacteria, (B) Structure of Gram negative cell membrane, LPS structure of Salmonella ser. Paratyphi A (C) and N. meningitidis (D), structure of O-antigen repeat unit from Salmonella ser. Paratyphi A (E) and O-antigen from N. meningitidis (F) with O-acetyl groups circled in red. For abbreviations see list of abbreviations (Abbreviations, page 81)

The O-antigen is a highly variable, large sugar structure unique to the bacterial species and serotype (19). A heterogeneous structure of variable lengths (21), the O-antigen is comprised of up to 50 repeating units of 2-6 sugar residues (20,24). The O-antigen of Salmonella ser. Paratyphi A consists of a repeating unit of mannose-rhamnose-galactose with a side branch of paratose attached to galactose (7,25), and the rhamnose is partially O-acetylated (26) (Figure 1). The O-antigen structure of *N. meningitidis* differs from that of *Salmonella* ser. Paratyphi A in that it is much shorter, comprising, in total, less than 10 sugar residues (23). However, heterogeneity is introduced by having multiple structures comprised of diverse sugar residues joined by different linkages and often containing additional phospholipids (27). The L9 oligosaccharide is often found on the surface of serotype A *N. meningitidis* and consists of a more complex branched structure containing galactose, glucose and GlcNAc, with the terminal GlcNAc often O-acetylated (23) (Figure 1).

1.2.2. Biosynthesis

Bacteria synthesise LPS in units before assembling the entire structure. The individual O-antigen units are made in the cytoplasm by the transfer of nucleotide sugars on to an undecaprenyl phosphate (UndP), anchored in the cytoplasmic membrane (20,24). When the O-antigen unit is complete, the UndP is translocated across the cytoplasmic membrane to the periplasm (28) in a process requiring proton motive force to provide energy (20). The individual O-antigen units are polymerised on the periplasmic side of the cytoplasmic membrane, to create the full length O-antigen (20). In *Salmonella* ser. Paratyphi a range of O-antigen lengths are created however, it is unknown exactly how the length is regulated (29). The full length O-antigen is transferred to the core oligosaccharide attached to lipid A to form the LPS structure (28). Modification of the O-antigen, for example, O-acetylation, is thought to occur after polymerisation when the O-antigen is in the periplasm (24).

1.2.3. Function

The LPS is an important structure in enhancing the virulence and persistence of bacteria (28). In particular, it has been shown that LPS is essential for intestinal colonisation, serum resistance and invasion of host tissues (30). In addition, LPS gives protection for the bacterium against host defences for example, complement mediated killing, phagocytosis and bile acids (19,20). LPS are also important in host defences against bacteria: the majority of antibodies produced by the host to fight Gram negative bacterial infections are against the LPS, in particular the O-antigen (20). Therefore, a wider diversity of O-antigen can confer a selective advantage to the bacterium (28), one method bacteria use to increase diversity, is to modify the O-antigen, for example, by acylation and glucosylation (31), a mechanism commonly derived from phage (32). Phage often use the LPS as a receptor to enter the bacterium, and there are many examples of phage modifying, either by cleavage or by acylation, the LPS after infection to prevent superinfection by other phage (33). Modification of the LPS by acylation has been hijacked by bacteria and is used to increase diversity of the LPS (32). A common modification is O-acetylation, addition of acetyl groups to the O-antigen repeating unit. In N. meningitidis, the terminal GlcNAc is known to be O-acetylated (23) (Figure 1) and in Salmonella ser. Paratyphi A O-acetylation is found on the C2 of rhamnose (7,25) (Figure 1). It has been shown that this O-acetylation is essential for stimulating the production of bacteriocidal antibodies by the host immune system (7). Therefore, it is important to understand how O-acetylation occurs, which proteins are involved, and their mechanism.

1.3. Acyltransferase proteins

1.3.1. Structure

The O-antigen of the LPS is commonly modified by bacteria to increase diversity (34); one class of proteins known to be involved in this modification are acyltransferase family 3 proteins. These acyltransferase proteins are highly hydrophobic, integral membrane proteins with either 10 or 11 transmembrane helices (35,36). Found in the cytoplasmic membrane, these acyltransferases are predicted to transport O-acetyl groups from the cytoplasm across the membrane, where the O-acetyl group is transferred onto the receptor molecule (37). It is presumed that the O-acetyl group is donated from acetyl co-enzyme A, however this has not been verified (38). These acyltransferase family 3 proteins can be grouped into two classes, with one group consisting of only an acyltransferase domain (AT3^{isol}) of 10 transmembrane helices (Figure 2), and the other containing acyltransferase domains with 11 transmembrane helices and an attached C-terminal periplasmic SGNH domain (AT3-SGNH) (Figure 2). Currently there is no known difference in function comparing AT3^{isol} and AT3-SGNH proteins.

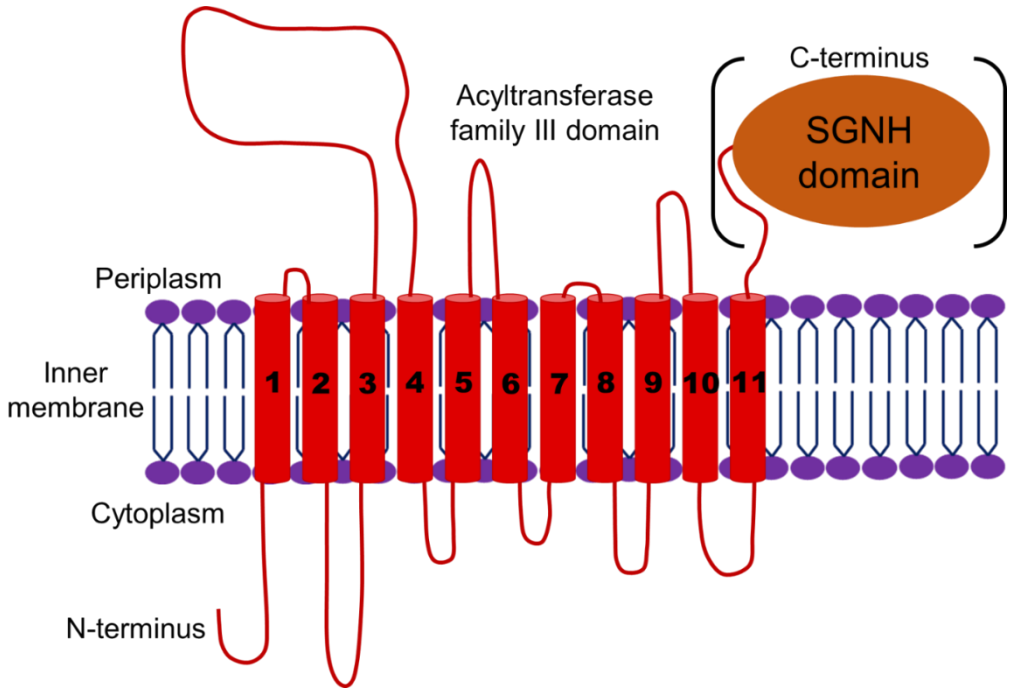


Figure 2. Predicted structure of acyltransferase family III domain located in the cytoplasmic membrane, the C-terminal, periplasmic SGNH domain is shown in brackets as it is not always present.

1.3.2. Function

Acyltransferase family 3 proteins are known to be involved in the modification of LPS and this was first observed by Davis et al. who showed that mutation of an acyltransferase protein, *NodX* from *Rhizobium leguminosarum*, inhibited nodulation of pea plants (35). This mutation was later shown to prevent O-acetylation of LPS which rendered the bacteria unable to infect the plants (39). Similarly in *Shigella flexneri*, an acyltransferase with high homology to *NodX* was found to be important in modifying the serotype by acetylation of a rhamnose residue in the LPS (40). In addition, the same family of acyltransferases have been shown to be important in the modification of many extracellular structures. For example, O-acetylation is a common modification of peptidoglycan and has been shown to give resistance to lysozyme, an enzyme produced by the body to break down bacterial cell walls (37). Similarly, O-acetylation of the exopolysaccharide of *Escherichia coli* and *N. meningitidis* has been shown to increase virulence and produce an improved antibody response (41,42). In addition, the production of macrolide antibiotics and anti-tumour compounds, by *Streptomyces mycarofaciens* and *Streptomyces griseus* respectively, involves O-acetylation to increase activity of the drug (43,44).

1.3.3. Mechanism

As previously mentioned, acyltransferase family 3 proteins have a predicted structure of 10 or 11 transmembrane helices. There are two long loops, one between helices 2 and 3, located in the cytoplasm and another, between helices 3 and 4, in the periplasm (Figure 2) (36). Thanweer et al. deleted these loops separately and showed no serotype conversion, due to a loss of function of the acyltransferase, despite no change to the location of the loops (36). Mutations of individual residues found a loss of function when three arginine residues present in the cytoplasmic loop, were mutated. These residues are also highly conserved between other acyltransferase proteins and it seems likely that they are important for catalysis (36). The location of these residues in the cytoplasm, and the loss of function of the acyltransferase when mutated, suggests that these arginine residues may be important for recognition of the substrate, presumed to be acetyl-CoA. Mutation of a range of residues in the long

periplasmic loop had no effect on the function of the acyltransferase. Little sequence conservation is seen in this loop, suggesting that it is less important for function (45). However, not all residues in this periplasmic loop were tested and while deletion of the whole periplasmic loop prevented function, this could be due to a perturbation in structure. In addition, residues in other loops were not mutated and it seems almost certain that catalytic residues are present in one of the periplasm loops to transfer the substrate on to the receptor molecule but such residues are so far unknown.

1.4. SGNH domains

1.4.1. Function

As mentioned previously, acyltransferase family 3 proteins can be classified into two groups, those with (AT3-SGNH) and those without ($AT3^{isol}$) attached SGNH domains. SGNH hydrolases can be found attached to other protein domains, as in the case here in AT3-SGNH proteins ($SGNH^{AT3}$), or as a stand-alone domain ($SGNH^{isol}$). $SGNH^{isol}$ proteins are known to catalyse a wide range of reactions with broad substrate specificity (46), even single enzymes are capable of performing multiple reactions, for example thioesterase I from *E. coli* can function as a thioesterase, protease and lysophospholipase (47). However, whilst these enzymes have the potential to react on a broad range of substrates, each has a preference for specific types of reactions or substrates (48).

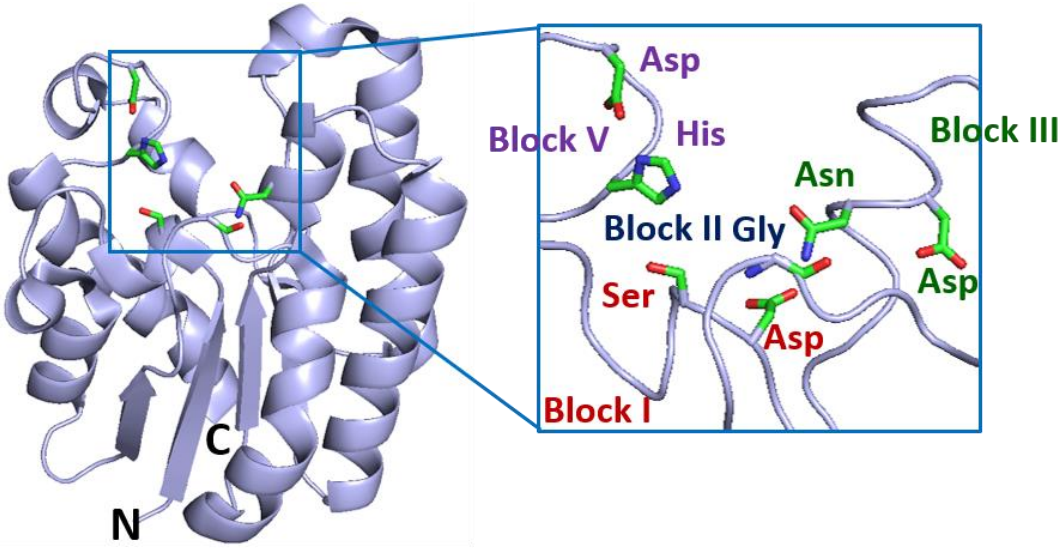


Figure 3. Structure of SGNH domain thioesterase I/protease I/lysophospholipase L1 (TAP1) from E. coli (PDB ID 1U8U, (90)), with catalytic and oxyanion hole residues highlighted.

SGNH hydrolases are often involved in bacterial virulence (46) and a common function is the addition or removal of acetyl groups from bacterial polysaccharides for example, alginate, peptidoglycan and LPS. This can aid the bacteria in avoidance of the host immune system: O-acetylation of alginate prevents clearance of biofilms (49), O-acetylation of peptidoglycan avoids degradation by lysozyme (50) and O-acetylation of LPS increases the diversity of the O-antigen (28).

1.4.2. Structure

The general structure of SGNH hydrolases is an α/β/α fold consisting of 4-5 parallel β-strands surrounded by 7-11 α-helices (46,51) (Figure 3). This core of β-sheets and α-helices is very rigid, acting as a scaffold for the substrate binding pocket (52). In contrast, the substrate binding pocket is highly flexible which enables the enzyme to process diverse substrates and reactions (46,52). While this structure is highly conserved between SGNH hydrolases, very low sequence identity is seen (48). However, four blocks of conserved sequences containing residues important for catalysis have been found: block I = GDS, block II = G, block III = GxND and block V = DxxH (where x is any amino acid) (53) (Figure 3). These blocks of sequence contain the catalytic residues: block I

Ser and block V Asp and His; and residues involved in forming an oxyanion hole: block I Ser, block II Gly and block III Asn (54). Block III Gly and Asp form a hydrogen bond which holds the oxyanion hole Asn in the correct orientation (54). Multiple studies have been carried out to examine the effects of mutating these conserved residues: mutation of block I Ser to either Ala or Cys results in no activity, similarly mutation of block V His to Ala or Phe shows no activity (51,55). In addition, mutation of block III Asn or Asp results in less than 1% activity compared to the wild type enzyme (51). Mutation in residues surrounding the catalytic residues also reduces catalytic activity and it is proposed these residues are important for positioning of the catalytic residues or involved in substrate binding (51).

1.4.3. Mechanism

Mechanisms have been proposed for reactions carried out by $SGNH^{isol}$ hydrolases, with many studies focussing on the de-acetylation reaction. However, Moynihan et al. (2014) suggested the mechanism described below for the acetylation of peptidoglycan by PatB (Figure 4) (56) and it is assumed that $SGNH^{AT3}$ domains react similarly. It has been suggested that when no substrate is bound, the side chain of block I Ser points towards the oxyanion hole and is hydrogen bonded to the carbonyl of block III Gly (46). As the acetate binds, either attached to CoA (50) or the acyltransferase domain (56), this displaces the Ser which aligns with the block V His. A salt bridge, formed between block V His and Asp, allows the imidazolium ring to remove a proton from block I Ser (Figure 4) making this residue highly nucleophilic and ready for catalysis (46). The acetate group is attacked by the nucleophilic Ser forming a negatively charged transition state complex which collapses to form an acetyl-enzyme-intermediate (Figure 4) (56). The acetyl-donor is released and, in the case of *Salmonella* ser. Paratyphi A, the C2-OH of rhamnose from the LPS enters the active site (Figure 4). The His imidazolium ring removes the protein from the C2-O, enabling it to attack the acetyl-enzyme-intermediate (56). Again, a negatively charged transition state is formed and stabilised by the oxyanion hole, formed from block I Ser, block II Gly and block III Asn (Figure 4) (54,57).

The product is released and the catalytic residues return to their original positions.

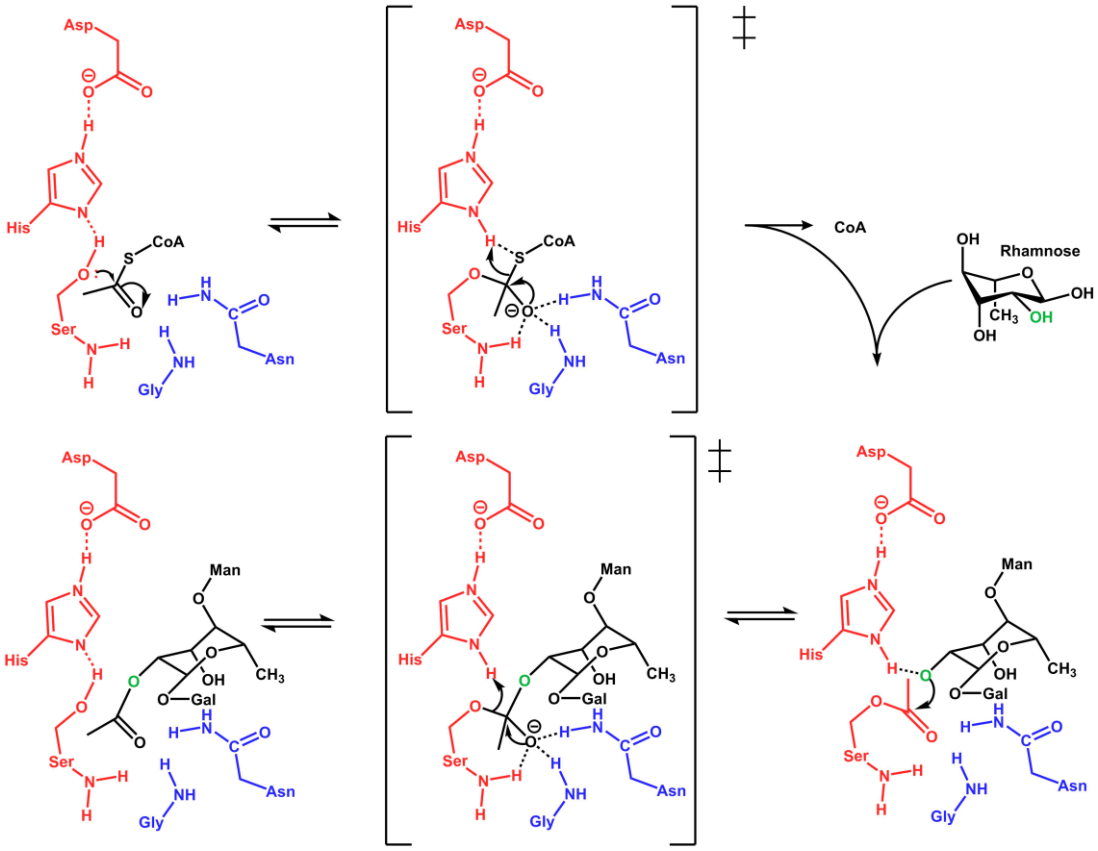


Figure 4. Proposed reaction mechanism for SGNHisol domains, acetyl group is delivered by acetyl Co-A and is added to the C2-OH position of rhamnose, as for the O-acetylation of the LPS from Salmonella ser. Paratyphi A. Catalytic residues are shown in red, and oxyanion hole residues in blue. The C2-OH group where the acetate group is added is shown in green. Adapted from Moynihan et al. (2014) (56).

1.5. Acyltransferase-SGNH domains

1.5.1. Structure, function and mechanism

Acyltransferase-SGNH proteins (AT3-SGNH) are formed of two domains: an acyltransferase family III domain consisting of 11 transmembrane helices and an attached C-terminal periplasmic SGNH domain (38,58) (Figure 2). This combination of acyltransferase and SGNH domain are known to acetylate either peptidoglycan or LPS, both of which aid bacteria in avoiding the host immune system (31,58-60). O-acetylation of MurNAc in peptidoglycan by OatA, an AT3-

SGNH, provides resistance to lysozyme (58). Similarly, O-acetylation of LPS contributes to phage avoidance, immune evasion and is also involved in many other host-bacteria interactions (27,31,60). In particular, it has been shown that O-acetylation of LPS is important in vaccine preparation. Mice vaccinated with polysaccharide containing O-acetyl groups showed a significant increase in immune response compared to polysaccharide where the O-acetyl groups had been removed (42). Many phage, for example, P22, use the LPS as a receptor to gain access to the bacteria and then cleave the LPS to prevent phage superinfection. O-acetylation of the LPS prevents adsorption of phage P22 to the LPS and therefore prevents infection (33).

$SGNH^{AT3}$ domains are required for function (33) which suggests that AT3-SGNH proteins are able to perform additional functions or act via a different mechanism to that seen in $AT3^{isol}$ (33). A similar system involved in peptidoglycan O-acetylation, consisting of an $AT3^{isol}$ protein (PatA) and $SGNH^{isol}$ domain (PatB) (not attached), has been used to predict the mechanism of AT3-SGNH proteins. PatA transports acetate (assumed to be cleaved from acetyl co-A) across the cytoplasmic membrane where PatB, present in the periplasm removes the acetate from PatA and attaches it to the peptidoglycan (61). It was initially thought that PatB was removing acetyl groups from peptidoglycan but experiments have been carried out showing PatB is involved in the addition of acetyl groups (61). Therefore, it is thought that PatA and PatB have the same mode of action as AT3-SGNH domains: the acyltransferase acts as a translocase to transport acetate, from acetyl-coA, into the periplasm where the SGNH domain attaches it to the polysaccharide (50). In the case of acyltransferase proteins, no $SGNH^{isol}$ domain genes have been found to suggest an arrangement similar to that seen for the PatA and PatB system. Therefore, it is assumed that $AT3^{isol}$ proteins perform both the translocation of the acetate and the O-acetylation reaction (36). Arginine residues, present in the cytoplasmic loops of $AT3^{isol}$ proteins, are known to be important in catalysis in both $AT3^{isol}$ proteins (36) and AT3-SGNH proteins (33) and these residues are potentially involved in recognition of the acetate substrate. In acyltransferase proteins, periplasmic loops are thought to be required for the O-acetylation reaction, however, as so little is known about this mechanism and any key residues involved, it is not possible to compare

differences between AT3^{isol} and AT3-SGNH proteins. It seems likely that key periplasmic residues are missing in the acyltransferase domain of AT3-SGNH proteins which necessitate the additional SGNH domain.

1.5.2. *Lot3* from *N. meningitidis*

O-acetylation of the LPS from *N. meningitidis* serogroup A has been shown to be vital for an immune response (42). Vaccination of mice with polysaccharide from *N. meningitidis* with and without O-acetyl groups showed an increased immune response when O-acetylated (42). Inactivation of *lot3*, a gene identified as an AT3-SGNH protein (27), resulted in a loss of O-acetylation of L9 LPS at the C3 position of the terminal GlcNAc in *N. meningitidis* serogroup A (27). Experiments carried out on a similar AT3-SGNH in *N. meningitidis* serogroup B showed that O-acetylation of the terminal GlcNAc prevents addition of phosphoethanolamine (PEA) which, when attached to LPS, is used for recognition by the host immune system for complement mediated killing (27). As O-acetylation of GlcNAc prevents addition of (PEA) this increases resistance of *N. meningitidis* to complement mediated killing (27). Whilst a similar AT3-SGNH protein is present in *N. meningitidis* serogroup A, it is unknown if it plays a similar role in resisting complement mediated killing.

1.5.3. *GtrC* family II from *Salmonella* ser. Paratyphi A

Modifications, addition of an acetyl or glucose molecule, to the O-antigen of Salmonella can alter the serotype and facilitate immune evasion. The *Salmonella enterica* subs. e*nterica* bacteria contain multiple glucosyl transferase gene clusters comprising of three genes: *gtrA, B* and *C* which are involved in transporting and transferring glucose to the LPS. GtrA is a bactoprenol-linked glucosyl translocase which transports glucose across the cytoplasmic membrane; GtrB is a bactoprenol glucosyltransferase which transfers glucose from the uracil-di-phosphate (UDP) carrier to bactroprenol;

and GtrC is a glycosyl transferase which transfers glucose from the UDP carrier to the receptor molecule (in this case LPS) (60). GtrC proteins determine the attachment residue and type of linkage and are, therefore, specific for each gene cluster (60). Davies et al. (2013) separated the GtrC genes into 10 distinct families based on the sequence similarity, with the assumption that within each family all gene clusters performed the same function. For example, all GtrC family III proteins add glucose (via a 1- linkage) to the galactose residue in the repeating unit of the O-antigen (31).

gtrC family II contains gene clusters from different *Salmonella* serovars: Typhi, Typhimurium, Gallinarum, Enteriditis and Dublin. The family II gene cluster shows many differences from other *gtr* gene clusters which suggests that they may not be involved in sugar transfer. A premature stop codon in *gtrA* and a truncation to *gtrB* from *Salmonella* ser. Typhimurium implies these are not required for the function of *gtrC*. In addition, Kintz et al. (2015) showed that the GtrC protein does not need GtrA or B to function and continues to operate when both have been removed (33). Similarly, sequence analysis showed GtrC family II proteins display similarity to OafA, an AT3-SGNH protein, and the arginine motif shown to be important for these proteins is present (33). This suggests that GtrC family II proteins are, in fact, acyltransferase family III proteins and sequence analysis reveals a C-terminal SGNH domain which is required for function (33). It has since been shown that GtrC family II from *Salmonella* ser. Typhi O-acetylates the rhamnose residue of the O-antigen at the C2 position and 99% sequence identity to the GtrC family II protein from *Salmonella* ser. Paratyphi A suggests that it also performs the same function (62).

1.6. Protein expression

E. coli is commonly used as an expression system to produce recombinant proteins for purification as it has been modified and optimised to produce high yields of the protein of interest. The gene of interest is cloned into an expression plasmid containing an inducible promoter, which is then transformed into *E. coli*. Addition of the inducer stimulates *E. coli* to transcribe and translate the gene of interest producing large quantities of the protein of interest which can then be purified.

However, while this process often produces large quantities of protein, the protein produced is not always soluble and forms aggregates or inclusion bodies, both of which are unwanted. If this is the case, many steps of the expression or purification process have to be modified and optimised to produce soluble protein. For example, the protein can be targeted for expression in the cytoplasm or periplasm; and modified *E. coli* strains have different optimised properties which may also enhance soluble expression.

1.6.1. Periplasmic vs. cytoplasmic expression

Protein expression would normally take place in the cytoplasm as this is the location of the protein production machinery. However, many $SGNH^{AT3}$ proteins contain pairs of cysteine residues predicted to form disulfide bonds, and therefore alternative expression routes must be considered. The cytoplasm is a reducing environment, maintained by thioredoxin and glutaredoxin systems, which prevents disulfide bonds from forming. Many proteins require disulfide bonds for stability, therefore, expression of these proteins in the cytoplasm, often results in the formation of insoluble protein aggregates, or inclusion bodies. Production of the protein in the periplasm, a non-reducing environment, enables disulfide bonds to form, leading to a, potentially, more stable protein. Addition of a PelB leader sequence to the start of the protein sequence targets the protein of interest to the periplasm via the Sec translocon (63), enabling the protein to fold correctly. While cytoplasmic expression is more commonly used, periplasmic is often more effective when the protein contains disulfide bonds or other post-translational modifications (64).

1.6.2. Strains

BL21 DE3 cells are the standard *E. coli* strain used for expression, however, due to the individual protein properties, not all proteins express well in these cells. Therefore, derivatives of this strain with enhanced properties have been generated which can be used to improve protein solubility.

Thioredoxin and glutaredoxin systems reduce any disulfide bonds which form in cytoplasmic proteins (65). Origami cells, derivatives of BL21 DE3 *E. coli*, have mutations in proteins involved in these systems – thioredoxin reductase and glutathione reductase – resulting in an oxidising cytoplasm enabling disulfide bond formation (66). Expression of disulfide bond containing proteins in Origami cells may aid protein folding and prevent aggregation or inclusion body formation.

pLysS cells are modified to improve control of induction; basal protein expression is inhibited by production of lysozyme which binds to RNA polymerase before protein expression is induced (67). This strain is often used where the protein of interest is toxic to *E. coli* to enable sufficient growth before protein expression is induced.

Codon bias, due to a difference in the codons used in foreign DNA compared to *E. coli* can lead to insertion of incorrect amino acids, or a premature truncation of the mRNA during production (67). Rosetta cells contain tRNAs for low frequency codons not usually found in *E. coli*. Therefore the use of Rosetta can avoid the insertion of incorrect amino acids potentially preventing the production of insoluble proteins.

Expression in the periplasm requires proteins to be targeted via the Sec translocon (68). Overexpression of periplasmic proteins can saturate the Sec-translocon leading to protein misfolding, aggregation and the formation of inclusion bodies (63). Lemo21 cells contain an additional promoter controlled by titration of rhamnose, this decreases the amount of protein produced which prevents the Sec-translocon from being overwhelmed (63). This strain has been shown to increase the yield of soluble periplasmic protein (63).

1.6.3. Lysis methods

Extraction of the expressed protein from the cells is another step which can be modified to increase soluble protein yields. BugBuster chemical lysis method uses detergents to disrupt the bacterial cell membrane (69), whilst this method

is most commonly used for small volumes, the chemicals used have the potential to cause aggregation of the protein. Alternatively, sonication uses ultrasonic vibrations to create shear forces which break down the cell membrane (69), and a range of reagents can be added to the buffer to aid protein solubility. Both chemical lysis and sonication are effective protein extraction methods. However, if insoluble protein has been produced, this is a step which can be modified to increase solubility.

1.6.4. Inducer concentration

Isopropyl β-D-1-thiogalactopyranoside (IPTG) is commonly used to induce expression of the protein of interest. However, the concentration of IPTG needs to be balanced so as to express high levels of protein without reducing cell growth. In addition, decreasing the concentration of IPTG used to induce protein expression has been shown to increase the solubility of proteins (70) and reduce the formation of inclusion bodies (71). Therefore, the concentration of inducer can be manipulated to increase solubility of the protein of interest.

1.6.5. Solubility tags

Solubility tags are proteins which fold extremely well and can be attached to insoluble proteins to increase their solubility. Maltose binding protein (MBP), in particular, is a very effective solubility tag and can also promote correct folding of the attached protein (72). It has been proposed that a large hydrophobic patch on the surface of MBP interacts with exposed hydrophobic residues on the attached protein (73). Weak binding and release interactions between the hydrophobic surfaces of both proteins can help to increase correct folding of the attached protein (73). Positioning of a cleavage site between the solubility tag and the protein of interest allows the tag to be cleaved during purification.

The use of *E. coli* as an expression system for production of recombinant proteins is invaluable for producing large quantities of protein. However, there

are multiple methods used to modify this process when the protein is forming insoluble aggregates or inclusion bodies.

1.7. Project Aims

O-acetylation by Lot3 from *N. meningitidis* has been demonstrated to be important in inducing an immune response and may also play a vital role in resistance to complement mediated killing (27). O-acetylation of the O-antigen of *Salmonella* ser. Paratyphi A by GtrC family II has also been shown to be important in immune evasion. In addition, O-acetylation of the LPS has been shown to be essential in vaccination against both *Salmonella* ser. Paratyphi A and *N. meningitidis* (7,42). Therefore, a more thorough understanding of these proteins is vital, in particular, what differentiates the AT3-SGNH proteins mentioned here from $AT3^{isol}$ or $SGNH^{isol}$ proteins.

This research aims to examine structural and mechanistic differences between $SGNH^{AT3}$ and $SGNH^{isol}$ proteins. Initially, the project analyses the sequences of AT3-SGNH proteins to explore any differences, specifically of residues important for catalysis. The project then goes on to express and purify the SGNH domains and region linking the domains from two AT3-SGNH proteins: Lot3 from *N. meningitidis* and GtrC family II from *Salmonella* ser. Paratyphi A. This is followed by biophysical analysis and structure determination. Sequence analysis and in silico structure modelling showed key residues, known to be important for catalysis, to be absent from $SGNH^{AT3}$ domains. Purification of the SGNH domain from Lot3 was unsuccessful; however, expression, purification and crystallisation of GtrC resulted in a high resolution structure. While the structures of many $SGNH^{isol}$ domains can be found in the Protein Data Bank (PDB), none are attached to acyltransferases. This is the first structure of an $SGNH^{AT3}$ protein and many structural differences have been observed.

Chapter 2. Materials and Methods

2.1. Materials

2.1.1. Bacterial strains

E. coli strains used for cloning and expression are shown in Table 1. The expression strains each have specific functions used for tuning expression of the protein of interest and these are also shown in Table 1 with more details found in section 1.6.2.

Table 1. Strains used for cloning and expression of the SGNH gene

Strain	Use	Application
BL21 DE3	Expression	Tuned for high levels of protein expression
Lemo21	Expression	Slowed expression levels for production of periplasmic proteins
Origami	Expression	Non-reducing cytoplasm for expression of proteins containing disulfide bonds
pLysS	Expression	Prevents expression prior to induction
Rosetta	Expression	Additional tRNAs for proteins containing foreign codons
XL1 blue	Cloning	Tuned for cloning of genetic material

2.1.2. Bacterial growth media

Bacterial growth media are described in Table 2. Antibiotics were added to the media dependent on the cell line and vectors used, at the concentrations stated in Table 3. Antibiotics were added only to the liquid media (and not solid media) unless stated otherwise.

Table 2. Bacterial growth media used for expression and cloning

Media	Components
Lysogeny broth (LB) liquid media	5 mg/mL yeast extract, 10 mg/mL tryptone, 10 mg/mL NaCl, autoclaved before use
LB agar	5 mg/mL yeast extract, 10 mg/mL tryptone, 10 mg/mL NaCl, 15 mg/mL agar, autoclaved before use
Super Optimal broth with Catabolite repression (SOC) media	2% tryptone, 0.5% yeast extract, 10 mM sodium chloride, 2.5 mM potassium chloride, 10 mM magnesium chloride, 10 mM magnesium sulfate 20 µL/mL 0.2 µm filtered 20% glucose added after autoclaving

Table 3. Antibiotics added to media dependent on vector or cell line used, concentrations stated are final concentrations used for selection in bacterial media.

Cell line	Antibiotic Resistance Cassette	Concentration
XL1 blue	None	
BL21 DE3	None	
pLysS	Chloramphenicol	30 µg/mL
Rosetta	Chloramphenicol	30 µg/mL
Origami	Streptomycin	100 µg/mL
	Tetracycline	10 µg/mL
Lemo21	Chloramphenicol (added to solid and liquid media)	30 µg/mL
Vectors	**Antibiotic Resistance Cassette**	**Concentration**
pETFPP_30	Ampicillin (added to solid and liquid media)	100 µg/mL
pETFPP_1/2/4	Kanamycin (added to solid a liquid media)	50 µg/mL

2.1.3. Plasmid vectors

Vectors were used to transport the desired gene sequences into the cells, multiple vectors were used with different functions – to target or tag the gene of interest to enhance expression levels (Table 4).

Table 4. Vectors used for protein expression

Vector construct	Targeting of protein of interest	Additional function
pETFPP_30	Periplasm	PelB leader sequence for targeting proteins to periplasm and C-terminal His-tag
pETFPP_1	Cytoplasm	N-terminal His-tag
pETFPP_2	Cytoplasm	N-terminal His-tag and MBP solubility tag
pETFPP_4	Cytoplasm	N-terminal His-tag and immunity protein 9 (IM9) solubility tag

pETFPP_30 provided by Caroline Pearson
pETFPP1/2/4 from University of York Bioscience Technology Facility as linearised constructs

2.1.3.1. Gene sequences

Genes of each $SGNH^{AT3}$ domain were ordered from Genewiz in pUC57-Kan vector with the gene sequence codon optimised for *E. coli*. Lyophilised vectors were reconstituted in double distilled water (ddH_2O) to a concentration of 100 ng/µL.

2.3.1.1. Primers

The primers used to amplify vectors and DNA sequences are shown in Table 5. Primers for infusion reactions (section 2.3.3) were designed so that one end was complementary to the beginning or end of the SGNH gene and the other end complementary to the linearised vector (Figure 5) to enable insertion of the SGNH gene.

Primers were ordered from Eurofins genomics and reconstituted with ddH_2O to a concentration of 100 µM.

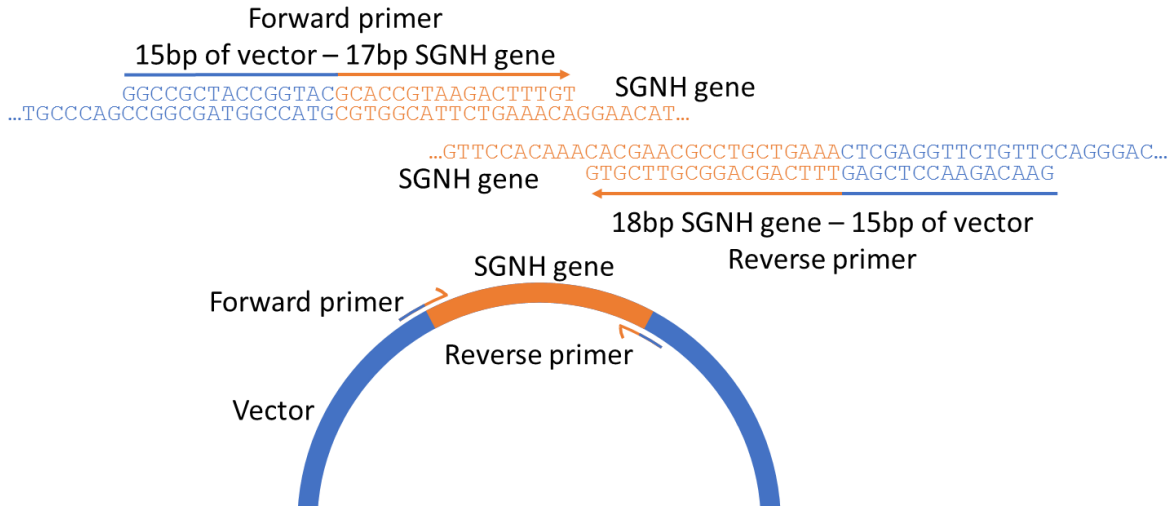


Figure 5. Design of primers for in-fusion of SGNH gene into linearised vector

Table 5. Primers used for genetic manipulation

Primer	Forward	Reverse complement
Linearisation of pETFPP_30 vector	*5' GCC GCA CTC GAG GTT CTG TTC 3'*	*5' CAT GGC CAT CGC CGG CTG 3'*
Amplification of SPA-GtrC-SGNH gene with ends complementary to pETFPP_30 linear vector	*5' CCG GCG ATG GCC ATG AAA AGC GCC GGT GAA TAC 3'*	*5' AAC CTC GAG TGC GGC TTT AAT GAT TTT ATT GCC AAT CTT G 3'*
Amplification of NM-Lot3-SGNH gene with ends complementary to pETFPP_30 linear vector	*5' CCG GCG ATG GCC ATG CGT GGC ATT CTG AAA CAG 3'*	*5' AAC CTC GAG TGC GGC TTT CAG CAG GCG TTC GTG 3'*
Amplification of SPA-GtrC-SGNH gene with ends	*5' TCC AGG GAC CAG CAA TGA AAA*	*5' TGA GGA GAA GGC GCG TCA*

complementary to pETFPP_1/2/4 linear vector	*AGC GCC GGT GAA TAC 3'*	*TTT AAT GAT TTT ATT GCC AAT CTT G 3'*
Amplification of NM-Lot3-SGNH gene with ends complementary to pETFPP_1/2/4 linear vector	*5' TCC AGG GAC CAG CAA TGC GTG GCA TTC TGA AAC AG 3'*	*5' TGA GGA GAA GGC GCG TCA TTT CAG CAG GCG TTC GTG 3'*
Amplification of SGNH genes during colony PCR (all vectors)	*5'-TAA TAC GAC TCA CTA TAG GG-3'* (T7 promoter)	*5'-GCT AGT TAT TGC TCA GCG G-3'* (T7 terminator)
Sequencing of gene insert (all vectors)		*5'-GCT AGT TAT TGC TCA GCG G-3'* (T7 terminator)

2.2. Bacterial culturing

2.2.1. Transformation

For vector purification: 1 µL vector (Table 4) and 20 µL XL1 blue super-competant cells (Agilent) were used. For protein expression: 0.5 µL vector and 50 µL cells (labelled expression in Table 1) were used.

A suitable vector was mixed with cells and incubated on ice for 30 minutes before heating at 42^{o}C for 45 seconds to stimulate the bacteria to take in the vector, then incubating on ice for a further 10 minutes. 120 µL SOC media (Table 2) was added and the culture incubated for 1 hour at 37^{o}C, with shaking at 200 rpm. The SOC culture was spread onto LB-agar (Table 2) plates containing the required antibiotics (Table 3) and incubated at 37^{o}C for 20 hours.

2.2.2. Liquid culturing

Cultures of bacteria were grown in LB liquid media (Table 2) for expression of vector or SGNH domain protein. A single transformed bacterial colony was

picked from an LB agar plate; used to inoculate liquid LB media containing the required antibiotics, and incubated at 37°C, with shaking at 200 rpm for 20 hours.

2.3. Genetic manipulation and purification

2.3.1. Polymerase Chain Reaction (PCR)

PCR was used to linearise the vectors and to amplify the SGNH genes; composition of PCR reaction mix is shown in Table 6. The PCR cycle shown in Table 7 was used with a 3 minute extension time for linearising the vectors, and 30 seconds for SGNH gene amplification. Products were run on a 1% agarose gel to ensure the correct size gene had been amplified prior to sequencing.

DpnI restriction endonuclease was used to digest cellular DNA without digesting the PCR product. 5.5 µL 10x Cutsmart buffer (New England Biolabs) and 10 units DpnI enzyme (New England Biolabs) were added to 45 µL linearised vector, before heating to 37°C for 1 hour, then to 80°C for a further 20 minutes.

Table 6. Components of PCR reaction mix used for amplification of genes and linearisation of vector

Components	Volume (µL)
5 ng/µL gene/vector (as detailed in Table 5)	8
5x HiFi buffer (Insight Biotechnology)	20
HiFi polymerase (Insight Biotechnology)	2 units
10 mM dNTPs	2
5 µM forward primer (as detailed in Table 5)	10
5 µM reverse primer (as detailed in Table 5)	10
milliQ H_2O	49

Table 7. PCR cycle for linearisation of pETFPP_30 vector and amplification of the SGNH genes.

	Step	Temperature	Time	
1.	Initial denature	95ºC	2 minutes	
2.	Denature	95ºC	30 seconds	Steps 2-4 repeated for 30 cycles
3.	Anneal	60ºC	30 seconds	
4.	Extension	72ºC	3 minutes for vector linearisation 30 seconds for gene amplification	
5.	Final extension	72ºC	5 minutes	
6.	Hold	12ºC	∞	

2.3.1.1. Colony PCR

Following transformation of the vector into XL1-blue *E. coli*, PCR was used to screen for colonies containing the SGNH gene. Transformed *E. coli* colonies were picked from the LB agar plate and mixed in the PCR mix (Table 8). The PCR cycle shown in Table 9 was used and products run on an agarose gel. Colonies which contained the inserted SGNH gene were grown in liquid culture (section 2.2.2) to harvest the SGNH gene.

Table 8. Composition of PCR reaction mix used for colony PCR.

Components	Volume per reaction (µL)
Taq green buffer (Thermoscientific)	1.5
Taq polymerase (Thermoscientific)	0.375 units
2 mM dNTPs	1.5
2 µM T7 promoter (sequence in Table 5)	0.75
2 µM T7 terminator (sequence in Table 5)	0.75
milliQ H_2O	10.4

Table 9. Colony PCR cycle used to determine which E. coli *colonies contained the SGNH gene*

	Step	Temperature	Time	
1.	Initial denature	95°C	5 minutes	
2.	Denature	95°C	30 seconds	Steps 2-4
3.	Anneal	50°C	30 seconds	repeated for 30
4.	Extension	72°C	1 minute	cycles
5.	Final extension	72°C	5 minutes	
6.	Hold	12°C	∞	

2.3.2. DNA Purification

Purification of linearised vector:

To purify the linearised vector, Nucleospin gel and PCR clean up kit (Machery-Nagel) was used and the PCR protocol followed as per the manufacturer's instructions, including an additional wash step as recommended in the protocol. An elution volume of 30 µL elution buffer was used.

Purification of DNA from agarose gel:

The desired band of DNA was cut from the agarose gel and purified using Nucleospin gel and PCR clean up kit (Machery-Nagel), following the DNA extraction from agarose gel protocol as per the manufacturer's instructions, including an additional wash step as recommended. An elution volume of 30 µL elution buffer was used, using elution buffer that had been pre-warmed to 50°C.

Purification of plasmid from bacteria:

To purify the plasmid, a single colony containing the plasmid was grown in liquid culture (Section 2.2.2). The liquid cultures were centrifuged at 5000 x g, 4°C for 10 minutes and the supernatant discarded. A Nucleospin plasmid DNA purification kit (Machery-Nagel) was used to purify the plasmid with the protocol carried out as per manufacturer's instructions. The following modifications were used, as suggested in the protocol, to increase the yield and purity of the plasmid: an additional wash step was included and the elution buffer was heated to 50°C before adding to the column. The elution buffer was incubated

on the column at 50°C for 5 minutes before elution, and an elution volume of 30 µL elution buffer was used.

2.3.3. In-fusion reaction

To insert the SGNH gene into the linearised vector an in-fusion reaction was carried out. The SGNH gene to be inserted and linearised vector were added to an infusion enzyme premix (Clontech) before incubation at 37°C for 15 minutes followed by incubation at 50°C for 15 minutes. TE (10 mM Tris and 1 mM ethylenediaminetetraacetic acid (EDTA), pH 8) buffer was added to the infusion mix to dilute the reaction before transformation into *E. coli* (section 2.2.1).

2.3.4. DNA analysis

2.3.4.1. Agarose gel electrophoresis

Agarose gels were used for both purification of DNA (preparative) and to ensure a DNA product of the correct size was present (analytical). For both, an agarose gel was prepared by addition of 1% agarose to TBE buffer (89 mM Tris, 89 mM Boric acid, 2 mM EDTA) and heating to dissolve the agarose. The agarose mix was allowed to cool then a 1 in 10,000 dilution of SybrSafe (Invitrogen) was added before pouring the gel into a mould and allowing to set. A 1 in 6 dilution of loading dye (New England BioLabs) was added to DNA samples before running on the gel at 100 V for 1 hour with a 2 log DNA ladder (New England Biolabs). Gels were imaged using a bio-imager (Syngene) using the transilluminator to display DNA bands. For preparative agarose gels, the DNA bands were cut out and purified (section 2.3.2).

Following colony PCR (section 2.3.1.1), an agarose gel was prepared as above, however, no loading dye was added to the samples and the gel run for 30 minutes. Gels were imaged as above.

2.3.4.2. DNA concentration determination

The concentration of purified DNA was determined post purification by measuring A_{260} using Nanodrop 1000 (ThermoScientific), A_{260}/A_{280} and A_{230}/A_{260} values were measured to assess purity: values greater than 1.8 were considered sufficiently pure.

2.3.4.3. Sequencing

DNA was sequenced to ensure the correct gene had been inserted into the vector. Purified vector with the SGNH domain gene inserted was sent to Eurofins genomics for sequencing, T7 terminator was used as the sequencing primer (Table 5).

2.4. Protein expression

2.4.1. Expression trials

A liquid culture was prepared (section 2.2.2) containing *E. coli* transformed with the SGNH gene in a vector. Expression cultures were prepared by diluting the liquid culture to an optical density at 600 nm (OD_{600}) of 0.1 with LB liquid media containing a suitable concentration of antibiotics. These cultures were grown at 37^{o}C, with shaking at 200 rpm until the OD_{600} reached 0.6, when 0.1 mM IPTG (unless otherwise stated) was added to induce protein expression. Cultures were incubated at different temperatures (20^{o}C, 30^{o}C or 37^{o}C) for different incubation times (1, 2, 4 or 20 hours) as stated in individual experiments. At each time point, 1 mL culture was removed, and the OD_{600} measured before centrifuging at 10,000 x g for 5 minutes and discarding the supernatant.

In expression trials with Lemo21 cells, various concentrations of rhamnose (0 µM, 50 µM, 100 µM, 250 µM, 500 µM, 750 µM, 1 mM or 2 mM) were added to LB liquid media in the expression culture and the experiment conducted as stated for other expression trials.

2.4.2. Large scale expression

As for expression trials, liquid cultures were prepared using bacterial transformants. 20 mL liquid culture was added to 1 L LB media in baffled flasks containing the relevant antibiotics. The cultures were incubated at 37^{o}C, shaken at 200 rpm until the OD_{600} reached 0.6 before induction with IPTG. After induction, cultures were incubated for 20 hours at 20^{o}C before being centrifuged at 5000 x g for 20 minutes. The supernatant was discarded.

As for expression trials, a 1 mL sample was removed at the point of induction with IPTG and after incubation for 20 hours at 20°C. The OD_{600} was measured before centrifuging at 10,000 x g for 5 minutes and discarding the supernatant.

2.4.3. Cell lysis

2.4.3.1. Chemical lysis

Table 10. Composition of resuspension buffer used for chemical lysis

Buffer	Composition
Resuspension buffer	50 mM phosphate buffer (pH 7.5), 200 mM NaCl, 10% glycerol, 5 mM Dithiothreitol (DTT), 100 µL/mL BugBuster (10 x protein extraction reagent, Millipore Novagen), 1 µL/mL DNase I (100 mg/mL stock solution), 1 µL/mL protease inhibitor
Protease inhibitor (1000 x stock solution)	1.6 ng/mL Benzamidine HCl, 1 mg/mL pepstatin, 1 mg/mL leupeptin, 1 mg/mL aprotinin)

Samples taken during expression trials were resuspended in 50 µL per OD_{600} per mL of resuspension buffer containing a chemical lysis solution (BugBuster) to lyse the bacterial cells (Table 10). Cells were mixed thoroughly to resuspend and ensure all cells were lysed.

2.4.3.2. Sonication

The cell pellets produced from large scale expression cultures were resuspended in binding buffer (Table 11) with DNase I and protease inhibitor (Table 10). The resuspended pellet was sonicated for 3 minutes, power 7 to lyse the cells before centrifuging at 30,000 x g for 45 minutes. During sonication, the resuspended pellet was cooled in an ice-water bath. The supernatant was collected and purified.

For expression trials, samples were resuspended in resuspension buffer (Table 10) before sonicating with a microtip for a total of 20 seconds on low power.

2.4.3.3. Sample preparation for sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS PAGE)

After cell lysis, a sample of the resuspended pellet was removed ('total' fraction) before the remaining samples were centrifuged for 10 minutes at 17,000 x g and a second sample removed from the supernatant ('soluble' fraction). Both the 'total' and 'soluble' fractions were run on an SDS PAGE gel.

2.5. Protein purification

2.5.1. Protein purification buffers

Table 11. Buffers used for protein purification of SPA-SGNH

Buffer	Composition
Binding buffer	20 mM TrisHCl pH 7.5, 500 mM NaCl, 20 mM imidazole (filtered and degassed)
Elution buffer	20 mM TrisHCl pH 7.5, 500 mM NaCl, 250 mM imidazole (filtered and degassed)
Dialysis buffer	20 mM TrisHCl pH 7.5, 100 mM NaCl, 20 mM imidazole, 0.5 mM DTT
Size exclusion buffer	20 mM Tris pH 7.5, 100 mM NaCl (filtered and degassed)

2.5.2. Nickel affinity purification

An Akta prime (GE healthcare) was used for all nickel affinity purifications. A His-trap column (GE healthcare) was equilibrated with binding buffer (Table 11), the cell lysate was flowed over the column and the column washed with binding buffer until a stable baseline of absorbance at 280 nm was seen. A gradient of binding buffer and elution buffer (Table 11) was flowed over the column, increasing the percentage of elution buffer over 10 column volumes until it reached 100%. Eluted fractions were collected and analysed by SDS PAGE (section 2.4.3.3).

Fractions containing protein of the expected molecular weight were pooled and 1mg 3C protease added for every 50 mg sample protein to cleave the His-tag.

The protein mixed with 3C protease was added to the dialysis membrane (5 kDa Molecular Weight Cut Off (MWCO)), the protein was dialysed into dialysis buffer (Table 11) at 4^{o}C for 20 hours.

As previously, the His-trap column was equilibrated with binding buffer, before the dialysed protein was flowed through the column, the flow through was collected and fractions run on an SDS PAGE gel. Elution buffer was flowed through the column to elute the bound His-tag.

2.5.3. Size exclusion purification

An Akta purifier (GE healthcare) was used for all size exclusion purification. After nickel affinity purification, the fractions containing the SGNH protein were pooled and concentrated in centrifugal filters (5 kDa MWCO, Millipore) until the final volume was less than 2 mL.

A Superdex 75 column (16/600, column volume 120 mL (GE healthcare)) was equilibrated with size exclusion buffer (Table 11), the protein sample injected, and eluted isocratically. The fractions were collected and run on an SDS PAGE gel to determine approximate protein yield and purity. The fractions which contained sufficiently pure protein were pooled and concentrated.

2.6. Protein analysis

2.6.1. SDS PAGE

Table 12. Buffers used for analysis of proteins

Buffer	Components
Sample buffer	4 x stock: 12 g glycerol, 3 mL H_2O, 10 mL 10% SDS, 1 mL 1 M Tris HCl pH 7.2, 0.06 g Bromophenol blue
Loading dye	750 µL 4 x sample buffer, 250 uL 1M DTT
SDS PAGE resolving gel	9.8 mL 30% acrylamide/bisacrylamide, 5 mL 1.5 M Tris (pH 8.8), 5 µL H_2O, 200 µL 10% sodium dodecyl sulfate (SDS), 200 µL 10% ammonium persulfate (APS), 20 µL tetramethylethanediamine (TEMED)
SDS PAGE stacking gel	1.3 mL 30% acrylamide/bisacrylamide, 2.5 mL 0.5M Tris (pH 6.8), 100 µL 10% SDS, 6.1 mL H_2O, 100 µL 10% APS, 10 µL TEMED
Running buffer	30 g/L Tris (pH 8.0), 140 g/L glycine, 10 g/L SDS
Coomassie stain	45% ethanol, 10% ethanoic acid, 2.5 g Brilliant Blue R
Destain	10% ethanol, 10% ethanoic acid

2.6.1.1. Coomassie gels

15% SDS PAGE gels were made by preparing the resolving gel (Table 12), pouring it into a mould and allowing it to set before addition of the stacking gel (Table 12). 4 x loading dye (Table 12) was added to each protein sample to give a final dilution of 1 x loading dye. Samples were heated at 95^{o}C for 2 minutes before loading into the gel. Precision plus protein all blue marker (Bio-Rad) was loaded in one lane of each gel to enable an estimation of molecular weight. Gels were run in running buffer (Table 12) at 200 V for 55-60 minutes or until the dye front had run off the gel. Gels were removed from the mould and stained with Coomassie stain (Table 12) for 30 minutes before washing and adding destain (Table 12) for 1-2 hours or until clear. Gels were imaged using a bio-imager (Syngene) using white light to view the protein bands.

2.6.2. Circular dichroism (CD)

Sample preparation:

The protein sample, as purified by size exclusion chromatography, was diluted in 20 mM Tris to give a final protein concentration of 0.2 mg/mL. A buffer sample was prepared in the same way, by diluting size exclusion buffer (Table 11) with 20 mM Tris.

To generate spectra under reducing conditions, Tris (2-CarboxyEthyl) Phosphine (TCEP) was used as a reducing agent. A stock solution of 100 mM TCEP was made and diluted to give the desired concentration in 20 mM Tris. This was then used to dilute the protein concentration and produce a suitable buffer to use as a baseline.

Single spectra analysis:

A Jasco J810 was used for all circular dichroism experiments. The molar ellipticity was measured at 20^{o}C with a wavelength scanning from 185 nm to 260 nm at a speed of 200 nm/minute, sensitivity of 100 mdeg, data pitch 0.5 nm, bandwidth 2 nm to produce circular dichroism spectra.

Temperature scan:

As for single spectrum analysis, circular dichroism spectra were measured over a wavelength scanning from 185 nm to 260 nm with the same parameters described above. A spectrum was recorded every 5^{o}C with a ramp rate of 2^{o}C per minute, 5 minutes equilibration time was allowed between reaching the desired temperature and reading the spectra.

Data analysis:

Photomultiplier tube (HT) voltage was also measured to give an indication of the reliability of the data, where the HT value was greater than 600, the data was discarded. Spectra of buffer and protein sample were both recorded and data points were plotted as protein spectra minus buffer spectrum. Where a temperature scan was performed, the buffer spectrum was generated at 20^{o}C

and this data was subtracted from the protein spectra recorded at each temperature (as is standard).

CD values for each temperature at 222 nm were plotted to generate a melting curve and determine the temperature of unfolding.

CDNN (74) was used to predict the percentage of α-helices and β-sheets in the structure, only data from the reliable region of the spectra (where HT < 600) was used.

2.6.3. SEC-MALLS

SEC-MALLS was performed using a Shimadzu HPLC and Wyatt Dawn HELEOS-II detector with laser wavelength set to 658 nm. 3 mg/mL and 1 mg/mL protein samples were prepared and loaded onto a Superdex 75 column (GE healthcare) equilibrated with size exclusion buffer (Table 11). Size exclusion buffer was used to elute the proteins over one column volume. A sample of Bovine Serum Albumin (BSA) was used as a standard of known molecular weight.

2.6.4. Mass spectrometry

Protein samples were dialysed (Millipore, 3.5 kDa MWCO) into 25 mM ammonium acetate for 18 hours at 4^{o}C. Mass spectrometry was performed by the Metabolomics and Proteomics lab (Bioscience Technology Facility, University of York). To determine if disulfide bonds were present in the protein, the protein sample was divided in two, half was reduced and alkylated and the other half just alkylated. Mass spectrometry was used to accurately determine the molecular weight of each sample, and the mass difference used to calculate the number of disulfide bonds in the protein.

2.6.5. X-ray crystallography

2.6.5.1. Crystal screens

Crystallisation conditions were screened using a Mosquito liquid handling robot (TTP Labtech) to prepare 96 well sitting drop plates. Index (Hampton Research) crystallisation screen and pH, anion, cation crystallisation trial (PACT) (Molecular Dimensions) were used with two concentrations of protein – 20 mg/mL and 10 mg/mL. Each sitting drop contained 300 nL of both protein

solution and reservoir and was equilibrated against 54 µL reservoir solution. Crystal trays were examined regularly using a microscope and crystals imaged.

2.6.5.2. Crystal optimisation

Conditions from crystallisation screens which yielded crystals were further optimised. 1 µL of both protein solution and reservoir were equilibrated against 1 mL reservoir solution in 24 well hanging drop plate. The buffers used for crystal optimisation are shown in Table 13. As previously, crystals were examined using a microscope and imaged.

Table 13. Buffers used for optimisation of crystals

Buffers	Precipitant	Salt
1 M BisTris pH 5.5	60% polyethylene glycol (PEG) 3350	4 M Ammonium sulfate
1 M BisTris pH 6.5		1 M Lithium sulfate
1 M 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) pH 7.5		
1 M Tris pH 8.5		

The components of each well were diluted to give a final concentration of 100 mM buffer, 25% precipitant and a range of salt concentrations: 0.2 M-1.7 M ammonium sulfate and 0-0.3 M lithium sulfate. 10 mg/mL, 15 mg/mL and 20 mg/mL protein concentrations were used.

2.6.5.3. Crystal collection

Crystals were collected 2-3 days after preparing the crystal tray. A cryo-loop was used to remove the crystal from the drop before flash freezing in liquid nitrogen, 20% glycerol was used as a cryo-protectant. Crystal diffraction was tested and those showing the highest resolution and least ice-rings were sent to the Diamond Light Source synchrotron for diffraction. A full data set was collected on beamline I04-1 with a wavelength of 0.928nm.

2.6.5.4. Data analysis and solving crystal structure

Recorded images were scaled and merged using AIMLESS (75) with data taken from 1.1-50 Å resolution. The Laue group, space group and unit cell size were determined. Matthew's co-efficient was calculated using CCP4 (76) to determine the number of protein units in an asymmetric unit, this information was also used to calculate the solvent content. FRAGON (77,78), a molecular replacement method was used to place one poly-alanine helix of 14 amino acids in length. ARP-wARP (79) carried out 3 modelling cycles and 5 ARP-RefMac cycles to model the protein structure around the helix placed by FRAGON. Coot (80) was used to manually modify the protein structure build by ARP/wARP and RefMac (81) used to refine the phases, 10 refinement cycles were used with 0.2 x-ray data weight restraint.

2.7. In silico analysis

2.7.1. Sequence analysis

Protein sequences were taken from UniProt (available at http://www.uniprot.org/).

InterPro (available at https://www.ebi.ac.uk/interpro/) was used to find the domain structures of sequences and to find other proteins with the same domain structures.

Clustal O was used to align amino acid sequences, BLOSUM62 matrix was used with gap opening penalty of 10 and gap extension penalty of 0.1, Jalview was used to view the alignments. Jalview was also used to retrieve sequences from Uniprot for alignment.

When screening large numbers of protein sequences (greater than 200) for simple motifs, 3of5 (available at http://www.dkfz.de/mga2/3of5/3of5.html (82)) was used to calculate the number of protein sequences containing a particular motif.

2.7.1.1. Generation of sample 100 sequences

For sequence analysis, a sample of 100 proteins each AT3-SGNH, AT3^{isol}, and SGNH^{isol} was used for alignment.

Interpro was used to search for the correct domain structure, SGNH^{isol} domain and AT3^{isol} sequences were filtered so that no other domains (N- or C-terminal) were included. A list was made of the Uniprot accession code for all sequences with the desired domain structure and every xth sequence was taken to give 100 sequences (Table 14). Due to the large number of sequences of AT3-SGNH domains from the same organism, these sequences were ordered alphabetically by organism before selecting sequences to give a higher diversity in the sample.

Table 14. Generation of 100 sample sequences for each domain structure, the number of sequences found on interpro was refined to leave 100 sequences

Domain structure	Sequence on interpro	Which sequences used
SGNH^{isol}	45,847	458th
AT3^{isol}	36,860	368th
AT3-SGNH	1757	17th

2.7.1.2. Phylogenetic trees

Clustal X was used to draw phylogenetic trees based on bootstrapped neighbour joining using a BLOSUM 62 matrix with gap opening penalty of 10 and gap extension penalty of 0.2. Dendroscope was used to view the phylogenetic trees (83).

2.7.2. Structure analysis

PHYRE2 (available at http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index (84)) and SWISS-MODEL (available at https://swissmodel.expasy.org/ (85-87)) were used to generate predictions of the protein structure. The amino acid sequence was inputted and known protein structures used as a template to produce a model.

Known structures of SGNH domains were taken from the PDB (available at http://www.rcsb.org/pdb/home/home.do (88)) to enable a comparison with the modelled structures.

CCP4MG and Pymol were used to view protein structures.

2.7.2.1. Ligand docking

FTMap was used to model the docking of small molecules to the protein; rhamnose, paratose, galactose, mannose and glucose were added manually to the collection of small molecules use by FTMap to probe interactions (89).

Chapter 3. *In silico* analysis

Acyltransferases with (AT3-SGNH) and without ($AT3^{isol}$) an attached SGNH domain appear to perform the same function. Therefore, it is unknown what role the $SGNH^{AT3}$ domain plays in acylation. *In silico* analysis of the sequences and structures of $SGNH^{isol}$ and $SGNH^{AT3}$ was carried out in an attempt to determine any differences which could be important in determining the function.

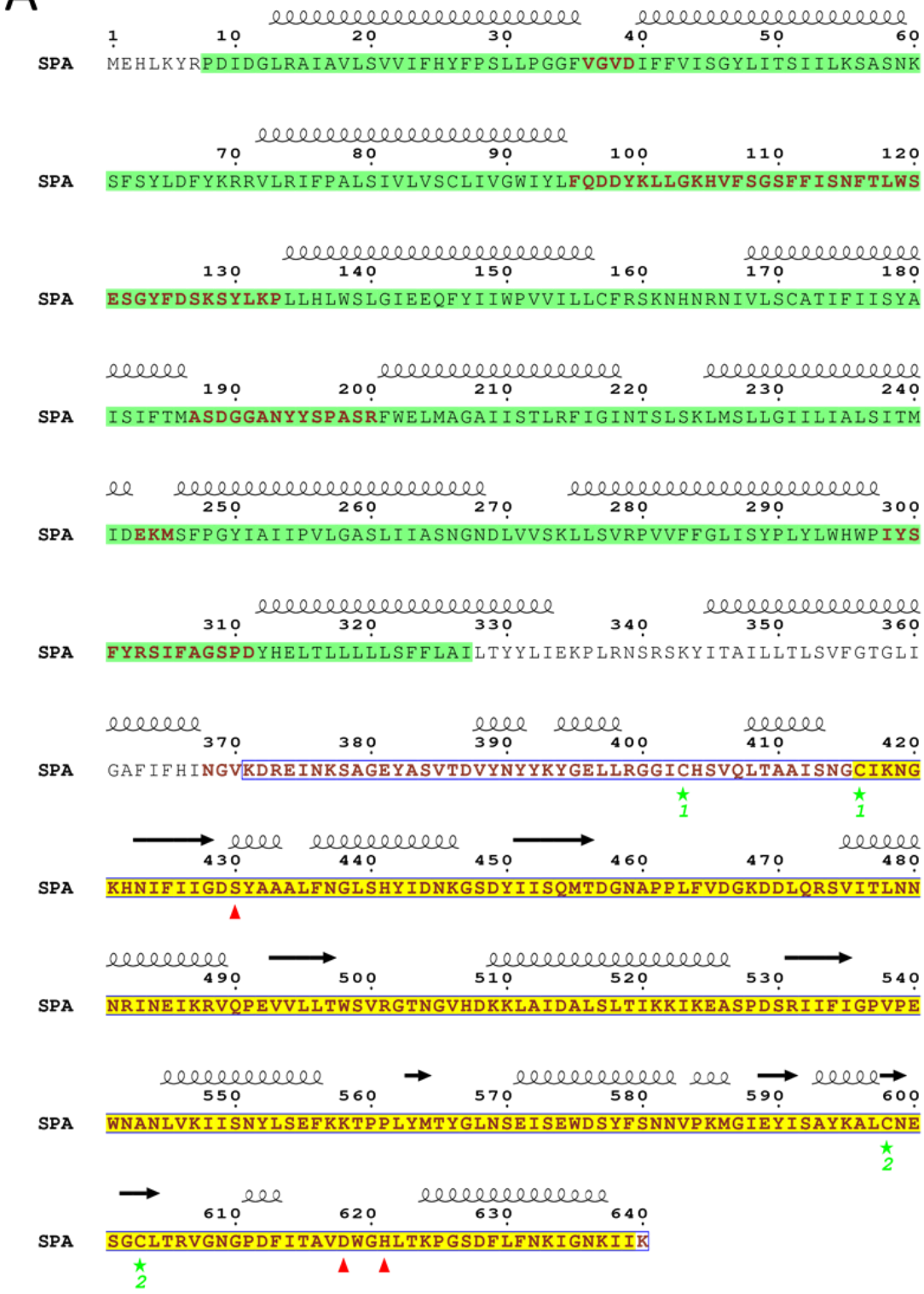
3.1. Functional domain analysis

The amino acid sequences of Lot3 from *N. meningitidis* (abbreviated to NM-Lot3) and GtrC family II from *Salmonella* ser. Paratyphi A (abbreviated to SPA-GtrC) were searched against the InterPro database for known families or functional domains. Both NM-Lot3 and SPA-GtrC are predicted to contain an acyltransferase family III domain from residues 8-338 and residues 8-327, respectively (Figure 6). In addition, SPA-GtrC is also predicted to contain an SGNH domain from residue 416-639 (Figure 6A). It was expected that NM-Lot3 would also contain an SGNH domain and therefore further analysis was carried out to determine if an SGNH domain was present. An acyltransferase protein from *Haemophilus influenzae* showing high sequence homology with NM-Lot3 (Identity = 58%, similarity = 75%) (Appendix 1) was predicted to contain an acyltransferase domain from residues 8-341 and an SGNH domain from residues 384-607. Due to the high sequence homology between the protein from *H. influenzae* and NM-Lot3, it was assumed that NM-Lot3 also contained an SGNH domain with the same domain boundaries (residues 384-607) which was not recognised by InterPro (Figure 6B). The region between the acyltransferase and SGNH domains is labelled here as the ‘linker’ region.

TMHMM was used to predict the number of transmembrane helices present in NM-Lot3 and SPA-GtrC (Figure 7). Both proteins contained 11 transmembrane helices with the C-terminal SGNH domain located in the periplasm and the N-terminus situated in the cytoplasm (Figure 6B). In both proteins, the 11^{th} transmembrane helix is not located within the acyltransferase domain but instead occurs in the ‘linker’ region. This suggests that this helix may not be

important for the function of the acyltransferase but instead important for locating the SGNH domain in the periplasm. Similarly, both proteins contain a long loop present in the periplasm, in NM-Lot3 this loop consists of 20 residues between helices 1 and 2, whereas in SPA-GtrC this loop consists of 39 residues between helices 3 and 4 (Figure 7C). Due to the length of this loop and its location in the periplasm, this loop may interact with the SGNH domain.

A


SPA MEHLKYRPDIDGLRAIAVLSVVIFHYFPSLLPGGFVGVDIFFVISGYLITSIILKSASNK
SPA SFSYLDFYKRRVLRIFPALSIVLVSCLIVGWIYLFQDDYKLLGKHVFSGSFFISNFTLWS
SPA ESGYFDSKSYLKPLLHLWSLGIEEQFYIIWPVVILLCFRSKNHNRNIVLSCATIFIISYA
SPA ISIFTMASDGGANYYSPASRFWELMAGAIISTLRFIGINTSLSKLMSLLGIILIALSITM
SPA IDEKMSFPGYIAIIPVLGASLIIASNGNDLVVSKLLSVRPVVFFGLISYPLYLWHWPIYS
SPA FYRSIFAGSPDYHELTLLLLLSFFLAILTYYLIEKPLRNSRSKYITAILLTLSVFGTGLI
SPA GAFIFHINGVKDREINKSAGEYASVTDVYNYYKYGELLRGGICHSVQLTAAISNGCIKNG
SPA KHNIFIIGDSYAAALFNGLSHYIDNKGSDYIISQMTDGNAPPLFVDGKDDLQRSVITLNN
SPA NRINEIKRVQPEVVLLTWSVRGTNGVHDKKLAIDALSLTIKKIKEASPDSRIIFIGPVPE
SPA WNANLVKIISNYLSEFKKTPPLYMTYGLNSEISEWDSYFSNNVPKMGIEYISAYKALCNE
SPA SGCLTRVGNGPDFITAVDWGHLTKPGSDFLFNKIGNKIIK

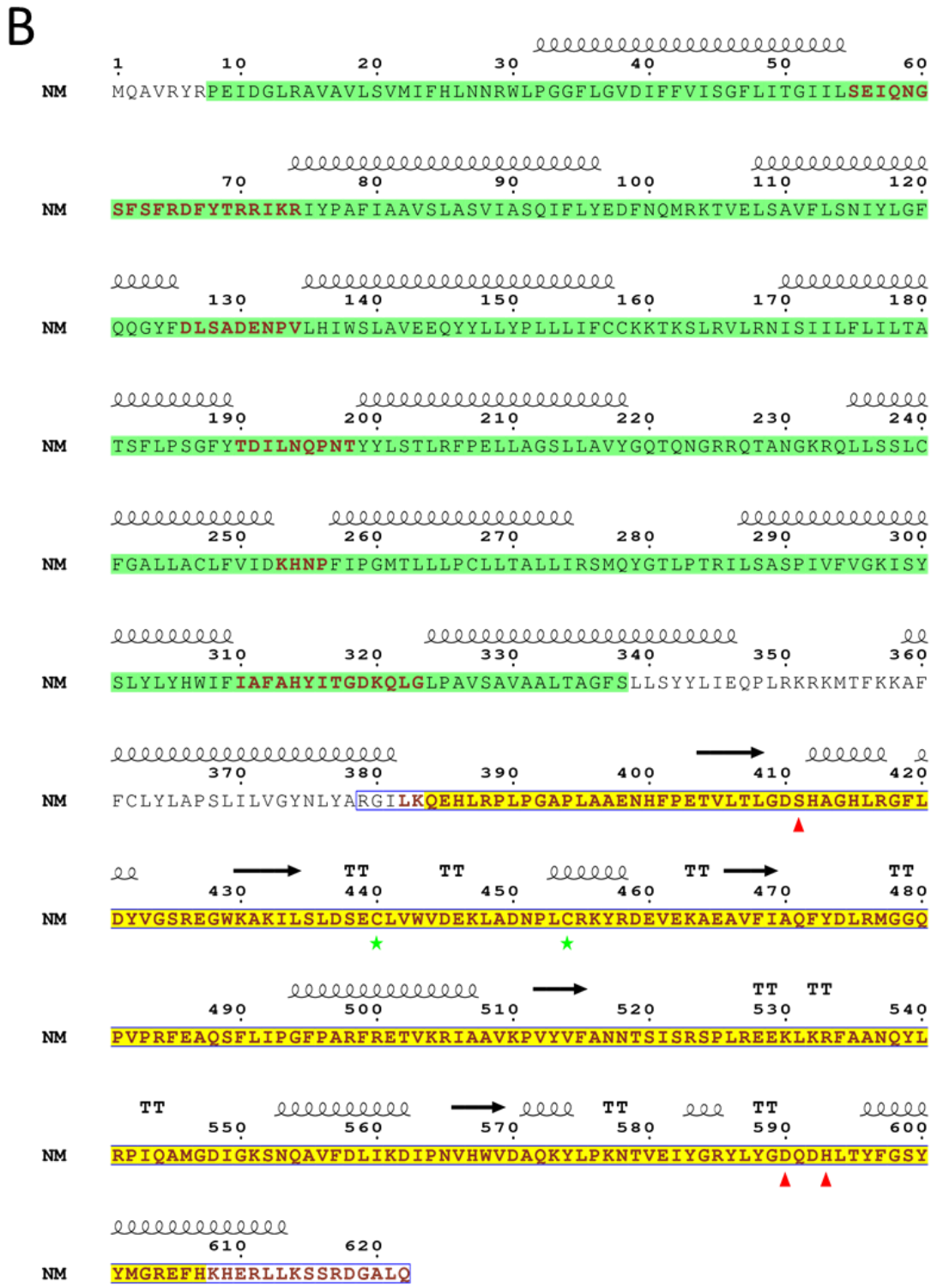


Figure 6. Annotated sequences of SPA-GtrC (A) and NM-Lot3 (B), arrow above sequence = β-sheet, coil above sequence = α-helix, brown bold font = periplasmic region, green highlighting = acyltransferase domain, yellow highlighting = SGNH domain, red triangle beneath sequence = catalytic residue, green star (numbered) beneath sequence = cysteine potentially forming disulfide bond, blue outline = region used for expression of SGNH domain.

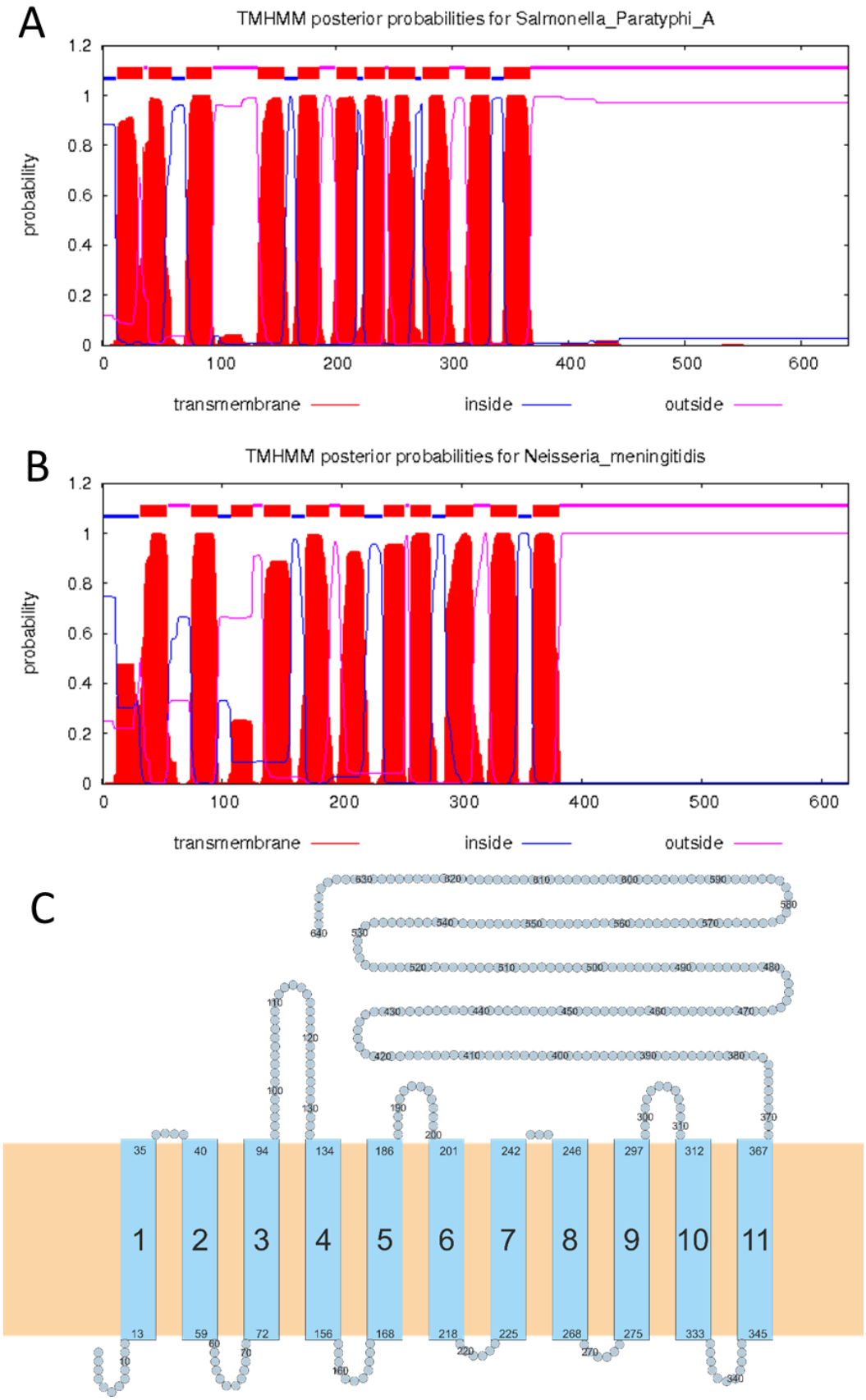


Figure 7. TMHMM plots showing probability of residues forming a transmembrane helix (A) SPA-GtrC (B) NM-Lot3. (C) Diagram of SPA-GtrC showing 11 transmembrane helices as predicted by the TMHMM plots with the N-terminus in the cytoplasm and C-terminal SGNH domain in the periplasm.

3.2. Sequence analysis of conserved residues

As described in section 1.4, SGNH domains have four blocks of conserved sequences: block I = GDS, block II = G, block III = GxND and block V = DxxH with block I Ser, block V Asp and His making up the catalytic triad. Sequence alignments show that residues involved in catalysis in block I and V are conserved in both SGNH^{isol} and SGNH^{AT3} (Figure 8). In contrast, residues in block II and III are very poorly conserved in SGNH^{AT3} and in many cases none of the residues in the consensus sequence are present (Figure 8). 3of5 pattern search (82) was used to screen large numbers of SGNH^{isol} and SGNH^{AT3} domain sequences (all proteins with correct domain structure on InterPro).

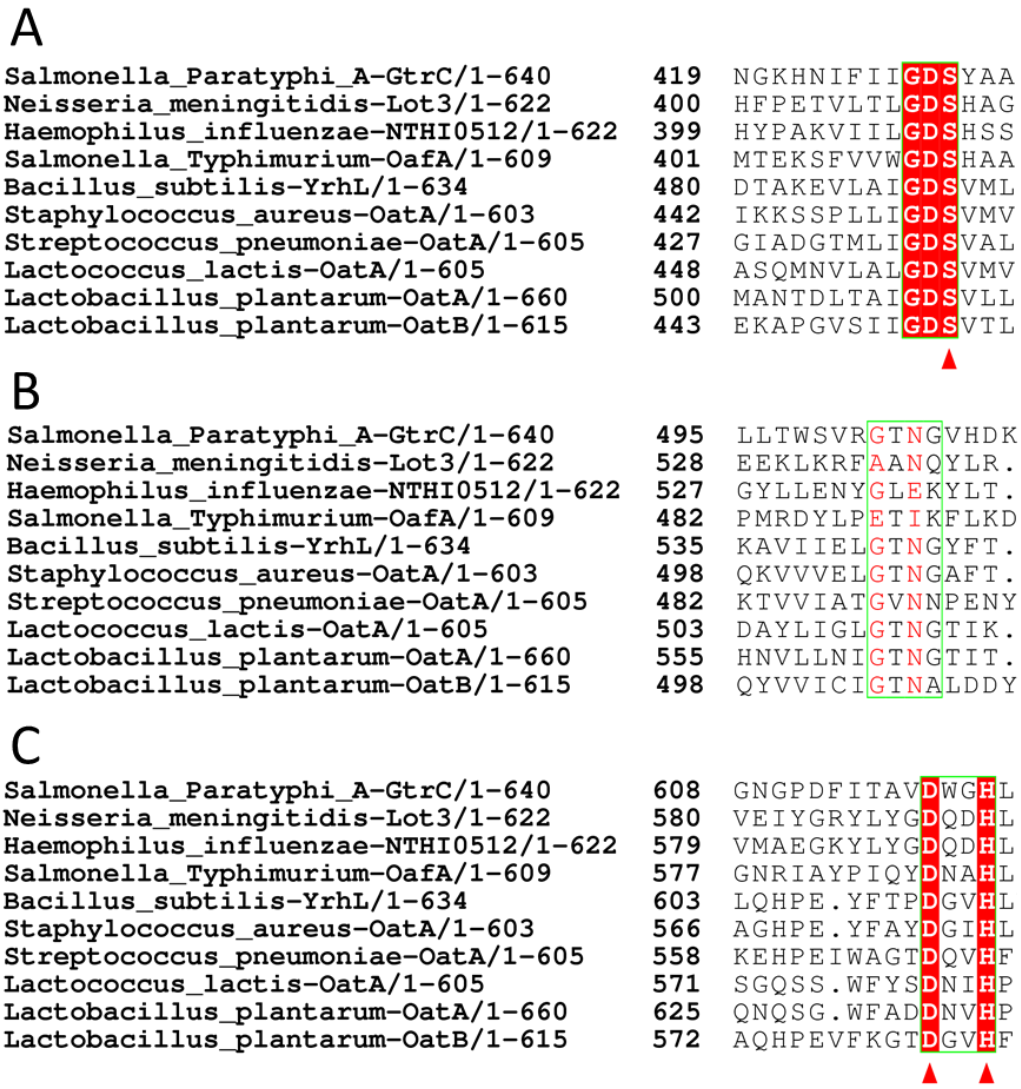


Figure 8. Sequence alignments of AT3-SGNH proteins, block I (A), block III (B) and block V (C) from 10 proteins. Red triangles indicate catalytic residues, red highlighting indicates residues that are 100% conserved, red type indicates residues that are 50% conserved, green boxes indicate the conserved sequence blocks.

73% of SGNH^{isol} contain the normal block III motif GxND compared to only 4% of SGNH^{AT3} (55,797 SGNH^{isol} and 3621 SGNH^{AT3} analysed) (Table 15). The residues in block III have been shown to be important in formation of an oxyanion hole to stabilise the transition state but are not present in SGNH^{AT3} domains.

Table 15. Percentage of sequences with each conserved sequence block, sequences screen using 3of5 to search for each consensus sequence. Total sequences: 55,797 SGNH^{isol} proteins, 3621 SGNH^{AT3} proteins.

	Block I	**Block III**		**Block V**
	% GDS	% GxND	% GTNG	% DxxH
SGNH^{isol}	75.4	73.4	0.8	84.6
SGNH^{AT3}	97.4	3.9	40.5	99.3

3.3. Structural predictions

SwissModel was used to produce models of the SGNH domains from NM-Lot3 and SPA-GtrC. Models are generated using protein structures from the PDB with the highest sequence identity as templates. NM-Lot3 was modelled on an SGNH^{isol} domain from *Bacillus subtilis* (PDB I.D. 2O14), 9% identity and 16% similarity and SPA-GtrC was modelled on an SGNH^{isol} domain from *Talaromyces cellulolyticus* (PDB I.D. 5B5S), 6% identity and 13% similarity. The sequence identity of SPA-GtrC and NM-Lot3 to the proteins used as templates is extremely low, therefore, these modelled structures can only be used to indicate the overall fold.

The modelled structures of SPA-GtrC and NM-Lot3 show an α/β/α fold with 5 parallel β-strands surrounded by α-helices, as seen for SGNH^{isol} domains (Figure 9A,C, E). In addition, the catalytic residues in the active site region of both NM-Lot3-SGNH and SPA-GtrC-SGNH are predicted to directly overlay the respective residues from SGNH^{isol} domains (Figure 9). However, as predicted from sequence analysis, the conserved residues forming an oxyanion hole in

SGNH^{isol} are not present. This is echoed in the predicted structures as the residues required to form an oxyanion hole are not present in the predicted structures of NM-Lot3-SGNH or SPA-GtrC-SGNH (Figure 9). Determining the actual structure of NM-Lot3-SGNH and SPA-GtrC-SGNH would enable more detailed analysis of the residues comprising the active site and to assess if there are amino acids which replace the block II and III residues.

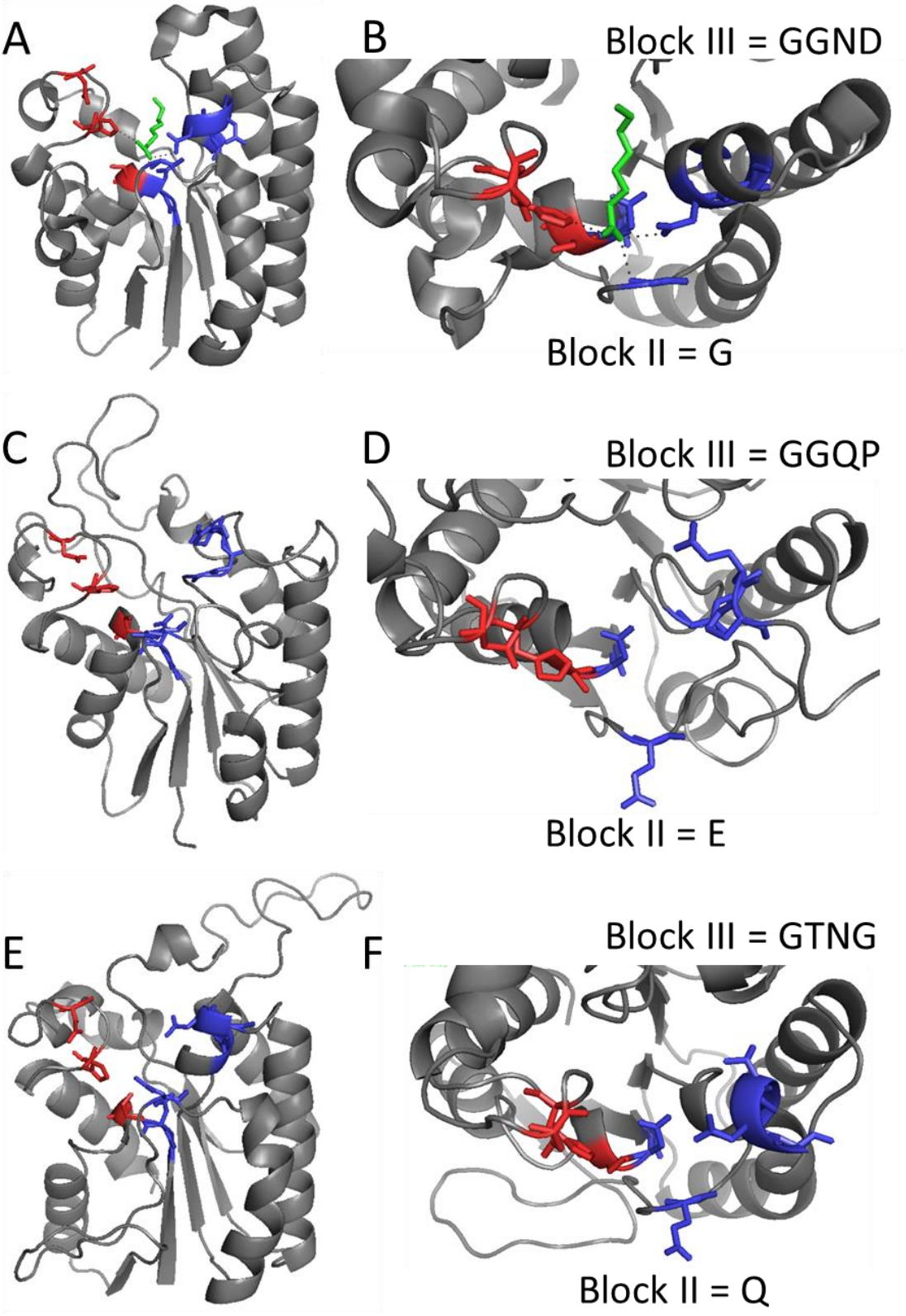


Figure 9. Modelled structure of SGNH domains and their active sites (A,B) Thioesterase I/Protease I/Lysophospholiase L1 (TAP) from E. coli in complex with octanoic acid (PDB: 1u8u (90)), (C,D) Modelled structure of SGNH domain of NM-Lot3, (E, F) Modelled structure of SGNH domain of SPA-GtrC. (B) Active site of (TAP) in complex with octanoic acid, hydrogen bonds (grey dotted lines) show interaction of substrate with catalytic and oxyanion hole residues. Catalytic residues shown in red, oxyanion hole residues in blue.

3.4. Disulfide bond formation

Many SGNH^{AT3} proteins have pairs of cysteines residues which are either present or absent in pairs (Figure 10). Not including the acyltransferase region, NM-Lot3 has one pair of cysteines residues found in the SGNH domains after block I, and SPA-GtrC has two pairs of cysteines, one pair located in the 'linker' region and the other in the SGNH domain located near the end of the sequence but before block V. When looking at a sample of 105 SGNH^{AT3} domains, 37% have at least one pair of cysteines, and of those 72% have more than one pair. The cysteine residues are either present or absent as pairs and 85% of SGNH^{AT3} domains have either an even number or no cysteine residues. These pairs of cysteines are also very well conserved in their position in the sequence (Figure 10A, B) with four possible locations for cysteine pairs – the 'linker' region, after block I, after block III or before block V (Figure 10B). A small number of proteins (15%) have 8 cysteine residues, a pair in each of the four possible locations.

The SGNH^{AT3} domains are located in the periplasm where disulfide bonds are able to form and as the cysteines are conserved in pairs, this suggests that disulfide bonds may be forming between the cysteine residues. Analysis of the modelled structures shows the pairs of cysteine residues are not in close proximity in the structure and unlikely to form a disulfide bond, in NM-Lot3 the cysteines are 11Å apart and in SPA-GtrC they are 14Å apart (Figure 10C, D). However, the sequence where the cysteine residues are located shows low sequence homology to the protein used as a template and therefore the model may not be accurate in this region. Therefore, further experimental work is required to determine if disulfide bonds are present or not.

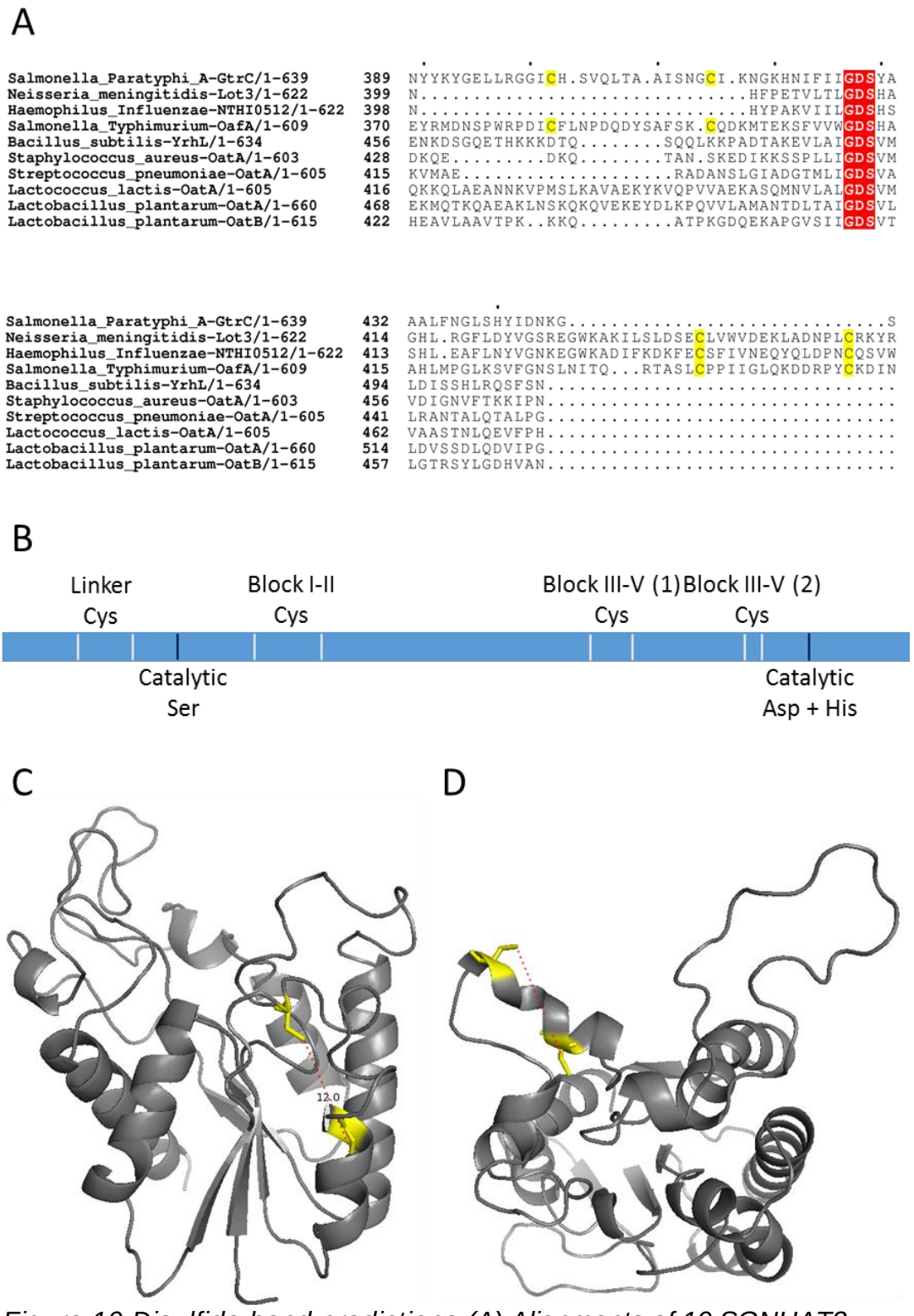


Figure 10.Disulfide bond predictions (A) Alignments of 10 SGNHAT3 domains with cysteine residues highlighted in yellow and block I residues highlighted in red. Cysteine residues are present or absent in pairs. (B) Schematic of 'linker' region and SGNHAT3 domain with the positions in the sequence of cysteine pairs and catalytic residues labelled. (C, D) Modelled structure of NM-Lot3 (C) and SPA-GtrC (D) with cysteine residues highlighted yellow.

3.5. Correlation between disulfide bond formation and block III sequence

When analysing the sequences of SGNH^{AT3} domains it was noticed that the number of cysteine pairs appeared to be linked to the sequence of block III. SGNH^{AT3} domains which had no cysteine pairs, 80% (total sequence = 65) had GTNG, or similar (GSNG, GTNN, GTNA etc,) as the block III sequence. Whereas, SGNH domains which have one or more cysteine pair, 2.5% (total sequences = 40) had GTNG as the block III sequence and the others have random sequence which show no homology e.g. GRWE, YGGD, IFLT, RLSF etc. SPA-GtrC is an exception to this rule as it has two pairs of cysteine residues and GTNG as the block III sequence. It is unknown how this may relate to the mechanism or function of the SGNH domain. The sequences of NM-Lot3 and SPA-GtrC have been shown to comprise of two domains, an acyltransferase and SGNH domain. Sequence analysis showed that block III oxyanion hole residues are absent in SGNH domains which are attached to an acyltransferase. Modelled structures were generated of both SGNH domains however, due to the low sequence homology these cannot be relied upon. Therefore, it is important to determine the structure of an AT3-SGNH to compare it to other SGNH domains. This requires a sample of protein to be expressed and purified for biophysical and structural analysis.

Chapter 4. Expression and purification

To study the structure of SGNH^{AT3} domains, two proteins were chosen for expression and purification. These are the products of the genes *lot3* from *N. meningitidis* (NM-Lot3) and g*trC* family II from *Salmonella* ser. Paratyphi A (SPA-GtrC). In both cases the amino acid sequence used for expression and purification started at the C-terminus of the last predicted acyltransferase domain transmembrane helix; and therefore included the SGNH domain and 'linker' region abbreviated NM-Lot3-SGNH and SPA-GtrC-SGNH respectively.

4.1. Periplasmic expression

In silico analysis suggests that disulfide bonds are likely to form (section 3.4) and as these are unable to form in the cytoplasm, periplasmic expression was used initially. An expression trial was carried out in BL21 DE3 *E. coli*, a standard expression strain, at a range of incubation temperatures (20, 30 and 37^{o}C) and times (0, 1, 2, 4 and 20 hours) after induction with IPTG.

The SDS PAGE gels (Figure 11A) show a large band at around 30 kDa corresponding to the approximate molecular weight of the SGNH domains in the lanes marked 'total' (section 2.4.3.3). However, in the respective 'soluble' (section 2.4.3.3) lanes, no band can be seen, indicating that the protein is expressed but is insoluble. Protein expression was seen under all conditions trialled (Figure 11A), with higher expression levels seen at the longer incubation times of 4 or 20 hours. Incubation for 20 hours at 20^{o}C, or 4 hours at 30^{o}C or 37^{o}C were the optimum expression conditions (Figure 11A). However, the expressed SGNH domains were insoluble in all cases. Therefore, a variety of strains and expression conditions were trialled with the aim of producing soluble protein.

BL21 DE3 is the standard strain used for protein expression, however, not all proteins express well in these cells; derivatives of this strain with enhanced properties can improve protein solubility. Origami and Lemo21 strain have been shown to increase soluble expression of proteins (section 1.6.2) and therefore these are most likely to produce soluble SGNH domains. Rosetta and pLysS strains were also trialled.

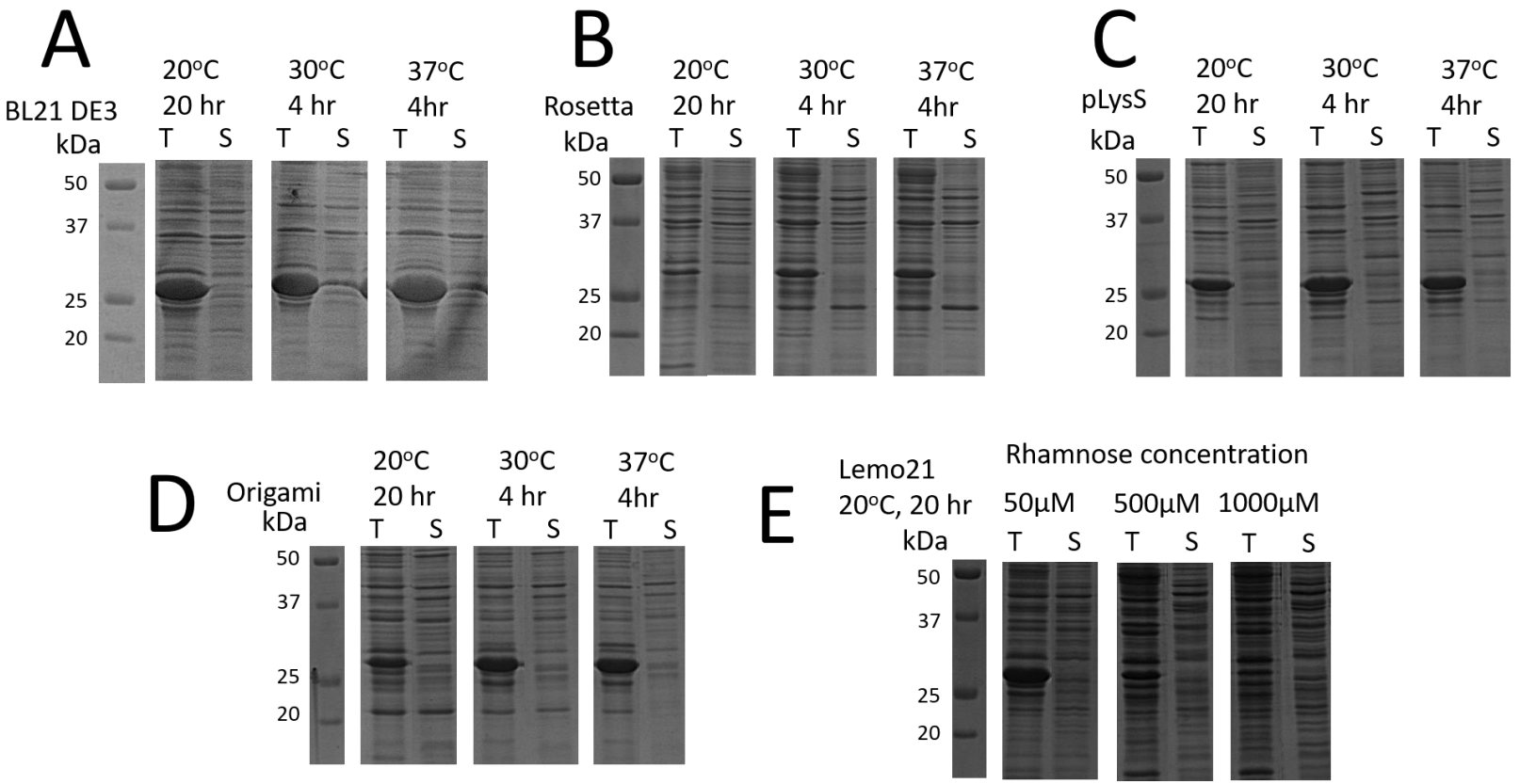


Figure 11. SDS PAGE gels showing periplasmic expression of NM-Lot3-SGNH in a range of strains (A) BL21 DE3, (B) Rosetta, (C) pLysS, (D) Origami, (E) Lemo21. A-D show expression in three incubation conditions: 20 °C for 20 hours, 30 °C for 4 hours and 37 °C for 4 hours after incubation with 0.1 mM IPTG. (E) shows expression at 20 oC for 20 hours only, and with three concentrations of rhamnose – 50 µM, 500 µM and 1000 µM. T and S refers to 'total' and 'soluble' fractions respectively (section 2.4.3.3). The SGNH domain is ~30 kDa in size which corresponds to the thick band of over-expressed protein seen on the gel at approximately this molecular weight.

The optimum conditions for expression of the SGNH domains in BL21 DE3 cells were incubation for 20 hours at 20°C, or 4 hours at 30°C or 37°C, therefore these conditions were used in this experiment. Lemo21 cells are the exception, where only one incubation time of 20 hours at 20°C was used and a range of rhamnose concentrations from 0-1000 μM.

For each strain, the amount of protein expression was similar across the range of incubation times and temperatures used (Figure 11). BL21 DE3 cells showed the most abundant protein expression. Origami, pLysS and Rosetta cell lines all showed similar expression levels (Figure 11B, C, D). However, as for BL21 DE3 cells, protein was only seen in the 'total' fraction and not the 'soluble' fraction indicating that the protein is insoluble. Lemo21 cells showed a decrease in protein expression as the concentration of rhamnose increased (Figure 11E), as expected. However, as for the other strains tested, no soluble expression was seen.

4.2. Cytoplasmic expression

In an alternative approach, both SGNH domains were expressed in the cytoplasm with either an His-tag, His-Im9 tag or His-MBP tag attached at the N-terminus. Both Origami and BL21 DE3 strains were used to compare the effect of a reducing (BL21 DE3) and non-reducing (Origami) cytoplasmic environment. As periplasmic expression experiments showed good expression when incubated for 20 hours at 20°C, this was used for cytoplasmic expression.

Soluble expression was seen for both strains in all conditions tested. Expression in Origami cells (oxidising cytoplasm) increased soluble expression in comparison to expression in BL21 DE3 cells (Figure 12). In Origami cells, 80-100% of protein produced with an MBP tag (band at about 70 kDa, (Figure 12)) was soluble in comparison to BL21 DE3 cells where only half the protein (with an MBP tag) produced was soluble (Figure 12).

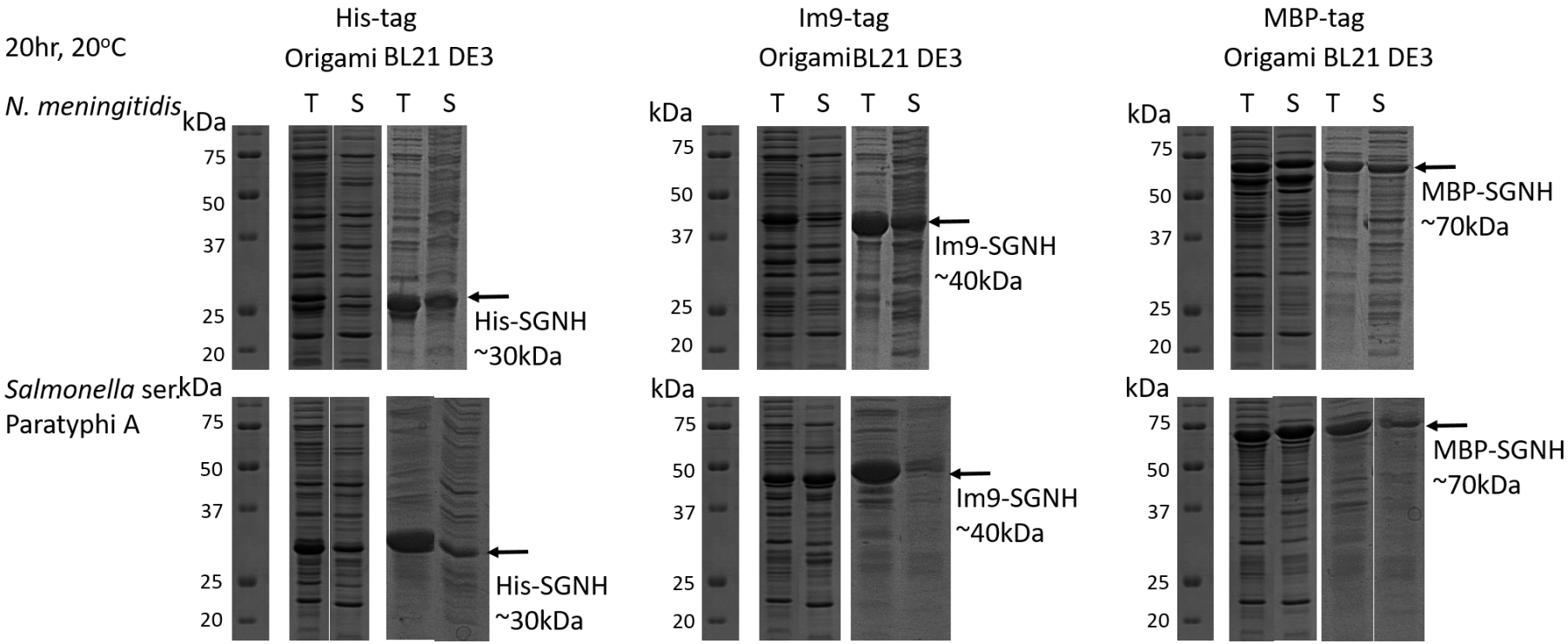


Figure 12. SDS PAGE gels showing cytoplasmic expression of NM-Lot3-SGNH and SPA-GtrC-SGNH in two strains: BL21 DE3 (reducing cytoplasm) and Origami (oxidising cytoplasm) with two solubility tags: MBP or Im9 or no solubility tag – His-tag. T and S refers to 'total' and 'soluble' fractions (section 2.4.3.3) respectively, with samples prepared as for Figure 11. The band corresponding to the SGNH domains is indicated on the figure and the difference in apparent molecular weight is due to the size of the attached tag

In both Origami and BL21 DE3 cell lines some of the expressed protein was soluble when only a His-tag was attached (band at about 30 kDa, (Figure 12)), however, soluble expression was increased when a solubility tag was connected (Figure 12). 80-100% of protein produced was soluble in Origami cells when attached to an Im9 (band at about 40 kDa, Figure 12 or MBP tag compared to only 50% of protein with just a His-tag attached (Figure 12). This suggests that the solubility tag is aiding solubility and potentially increasing stability or correct folding of the SGNH domain.

SPA-GtrC-SGNH expressed well in Origami cells when both solubility tags were attached with optimum conditions with the MBP tag (Figure 12). Therefore, these conditions – incubated for 20 hours at 20^{o}C – were used to produce large-scale cultures suitable for protein purification.

SPA-GtrC-SGNH showed higher soluble expression than NM-Lot3-SGNH in all conditions. On the SDS PAGE gels of NM-Lot3-SGNH, bands can be seen at lower molecular weights than expected when all three tags are attached (Figure 12). It seems likely that these additional bands of lower molecular weights are due to degradation. Therefore, an additional expression trial was carried out in an attempt to optimise expression and avoid potential degradation. Origami cells were used with a shorter incubation time of 4 hours at 20, 30 or 37^{o}C and as previously, solubility tags were attached. However, no soluble expression was seen at the shorter incubation time for any temperature (Figure 12). Therefore, despite potential degradation, and lower levels of soluble protein, incubation for 20 hours at 20^{o}C post induction in larger scale cultures of Origami cells with an MBP solubility tag (Figure 12) was used to produce protein for subsequent purification.

4.3. Purification

Purification was performed using immobilised metal (nickel) affinity chromatography (IMAC). Following the initial purification step, 3C protease was used to cleave the attached His-MBP tag from the SGNH domain (Figure 13B, E); a second IMAC purification step was used to separate the His-MBP tag, and

the His-tagged protease, from the SGNH domain. SPA-GtrC-SGNH separated well from the His-MBP tag leaving an almost pure sample of SGNH domain (Figure 13B). However, despite successful cleavage with 3C protease, NM-Lot3-SGNH could not be separated from the His-MBP tag and remained associated with the nickel column (Figure 13E). In addition, bands of lower molecular weight of around 16 kDa and 10k Da can be seen on the SDS PAGE gel which suggests the SGNH domain is partially degraded (Figure 13E). Size exclusion purification was used in an attempt to separate the His-MBP tag from NM-Lot3-SGNH. Size exclusion purification successfully separated SPA-GtrC-SGNH from the impurities still present after nickel affinity purification (Figure 13C). This purified protein sample was used for further structural analysis for example, mass spectrometry and x-ray crystallography.

However, size exclusion was unable to separate NM-Lot3-SGNH from the MBP-His tag (Figure 13F). Elution was faster than expected suggesting a molecular weight higher than that of the SGNH domain alone. That is, SPA-GtrC-SGNH started eluting after 65 minutes, whereas NM-Lot3-SGNH started eluting after only 45 minutes despite having very similar molecular weights (29.6 kDa for SPA-GtrC-SGNH and 28.0 kDa for NM-Lot3-SGNH). This suggests that NM-Lot3-SGNH remains associated with the MBP tag despite cleavage with 3C protease (Figure 13). As it was not possible to produce a pure sample of NM-Lot3-SGNH, this SGNH domain was not used for further analysis.

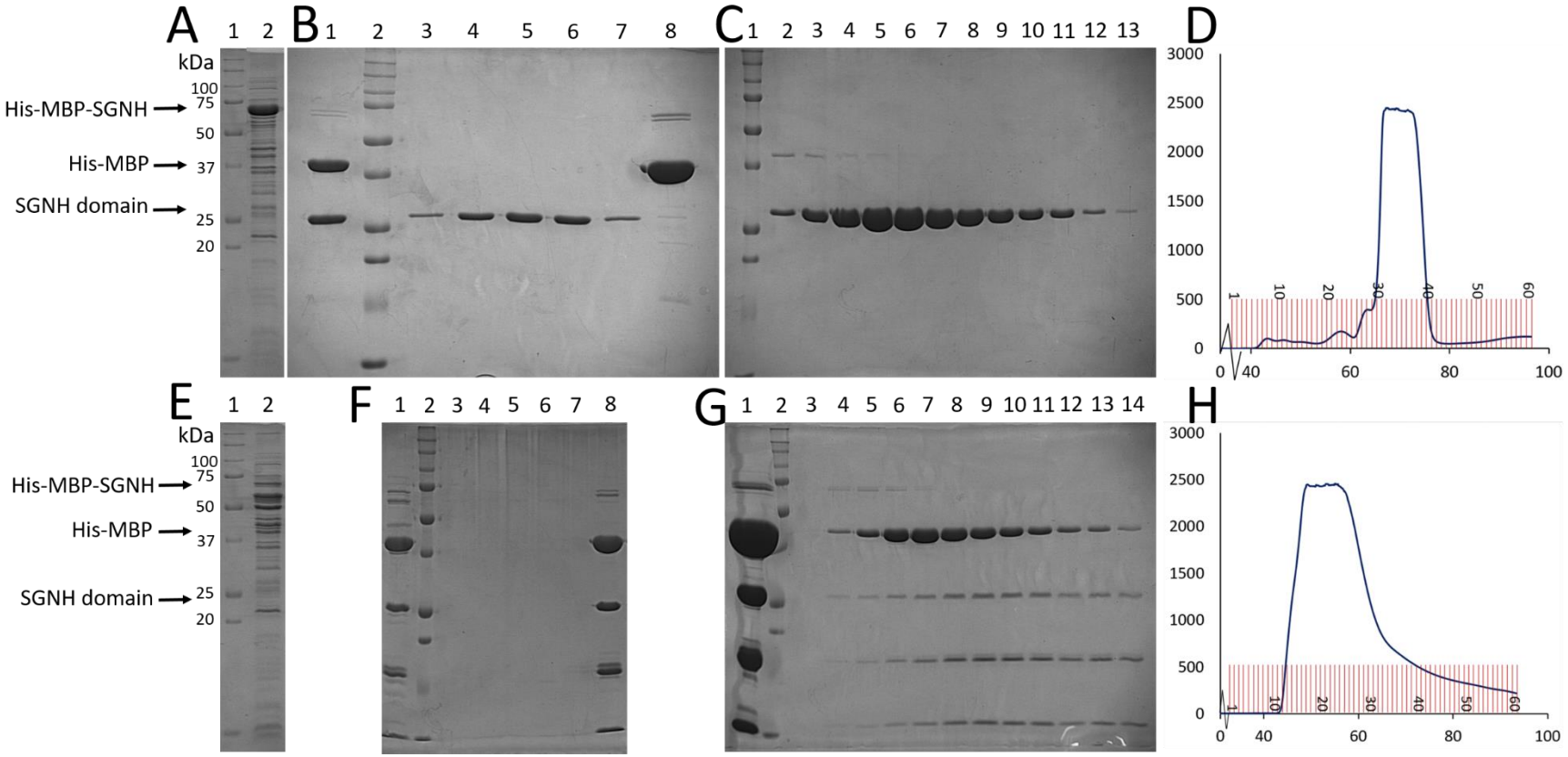


Figure 13. SDS PAGE gels showing purification of SPA-GtrC-SGNH (A-C) and NM-Lot3-SGNH (D-F). (A+E) Large scale soluble expression of SPA-GtrC-SGNH and NM-Lot3-SGNH respectively, lane 1: marker, lane 2: soluble fraction. (B+F) Post second nickel affinity purification, lane 1: after cleavage with 3C protease, lane 2: marker, lane 3-7: fractions of flowthrough, lane 8: elution after addition of imidazole. (C) Post size exclusion, lane 1: marker, lane 2-12: fractions 30-41. (G) Post size exclusion, lane 1: sample loaded onto column, lane 2: marker, lane 3-13: fractions 11, 13, 15, 17 etc. to 33. (D+H) Size exclusion chromatograms for SPA-GtrC-SGNH and NM-Lot3-SGNH respectively, fractions are indicated in red.

Chapter 5. Biophysical and structural analysis

Following expression and purification of the SPA-GtrC-SGNH domain, analysis was carried out to study the structure of the protein, in particular the potential for disulfide bond formation. X-ray crystallography was used to determine the structure of the SGNH domain to allow for comparison between $SGNH^{AT3}$ and $SGNH^{isol}$ domains.

5.1. Biophysical analysis

SEC-MALLS was used to determine if the SGNH domain is a monomer or higher oligomer in solution. Using BSA as a molecular weight standard, the approximate molecular weight was determined to be 30 kDa, very similar to the predicted molecular weight of 29.6 kDa. The presence of a single, uniform peak at this molecular weight (Figure 14A) indicated that SPA-GtrC-SGNH is monodispersed and monomeric.

Mass spectrometry was performed to assess the formation of disulfide bonds. Two samples were prepared, one was reduced and alkylated while the other was just alkylated. Mass spectrometry was used to measure the mass of both samples. The molecular weight of an untreated sample was measured as 29.606 kDa (Figure 14B), as predicted, confirming that the correct protein had been produced. A difference in mass of 187 Da between the reduced and alkylated samples (Figure 14B) indicated that two disulfide bonds were formed in SPA-GtrC-SGNH.

Circular dichroism was used to see if SPA-GtrC-SGNH was folded and to determine the melting temperature. An initial spectrum at 20^{o}C showed characteristic peaks and troughs corresponding to a folded protein containing α-helices and β-sheets (Figure 14C). Analysis of this spectra predicted the secondary structure to contain 38% α-helix, 15% β-sheet, 16% β-turn and 30% random coil. To determine a melting temperature, the temperature was increased to 90^{o}C with spectra read every 5^{o}C. The spectra showed the protein to be folded until the temperature reached 60^{o}C where the spectra became flatter, indicating that the protein unfolds between 55 and 60^{o}C (Figure 14D).

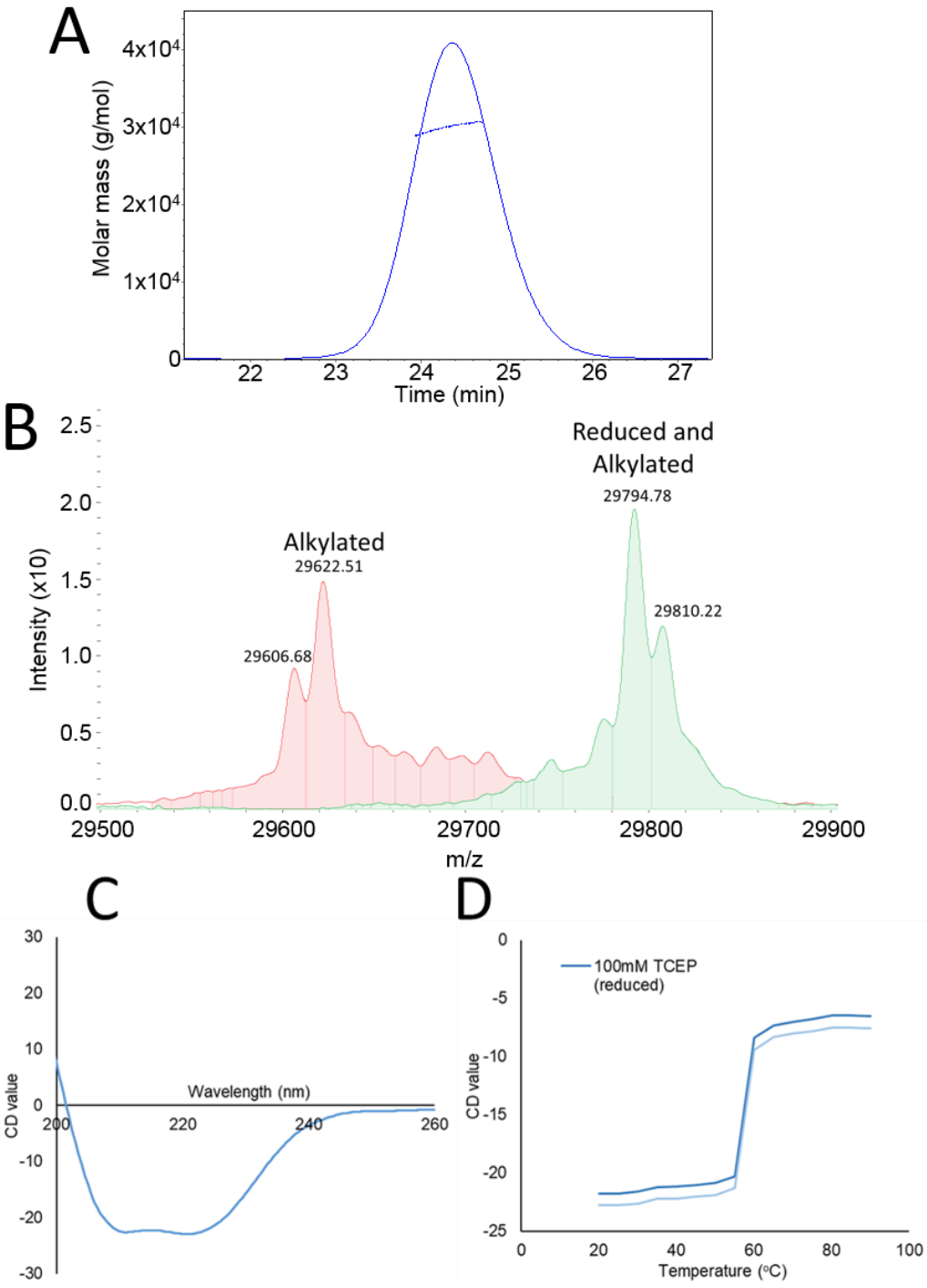


Figure 14. Structural analysis of SPA-GtrC-SGNH domain: A. SEC-MALLS data showing a monomer, and confirming molecular weight. B. Mass spectrometry, showing shift in peak size when reduced and alkylated (green) compared to just alkylated (red). C. Circular dichroism spectra of SPA-GtrC-SGNH in non-reduced conditions D. Melting curve of SPA-GtrC-SGNH in reduced (100 mM TCEP) and non-reduced conditions at 222 nm from 20 to 90 ^{o}C.

The same experiment was repeated in buffer containing 100 μM TCEP to reduce the disulfide bonds. As before, an initial spectrum at 20ºC showed SPA-GtrC-SGNH to be folded. Increasing the temperature showed that the protein unfolds between 55 and 60ºC, as seen under non-reducing conditions (Figure 14D). The spectra read in reducing and non-reducing conditions were very similar (Figure 14D), which suggested that the formation of disulfide bonds does not contribute to the thermal stability of the SGNH domain.

5.2. Structure determination

X-ray crystallography was used to determine the structure of SPA-GtrC-SGNH domain; crystallisation screens were carried out initially to find the optimum conditions.

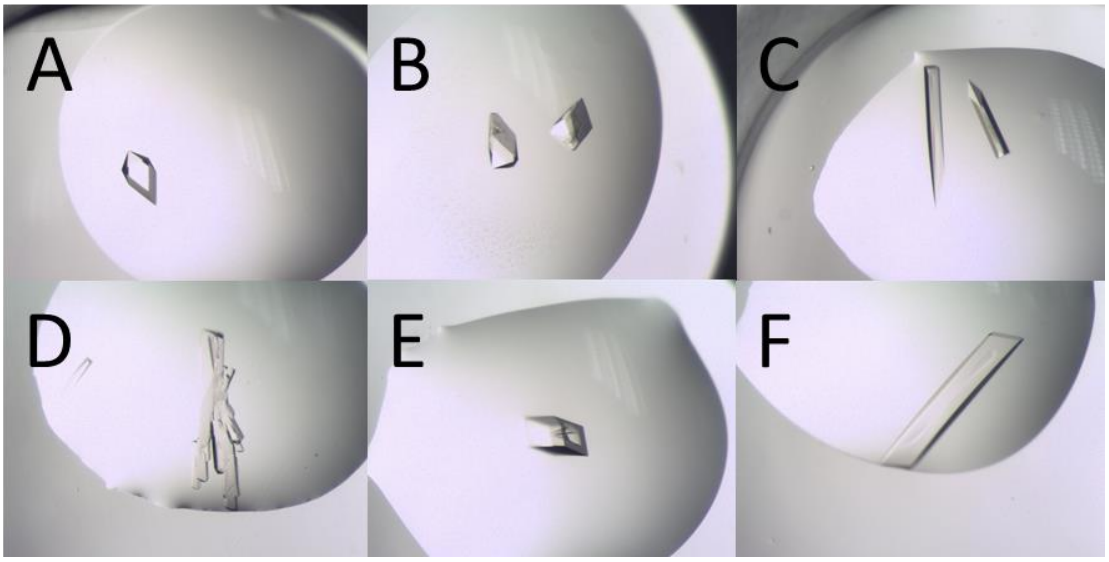


Figure 15. Crystals from initial screening, A. Index A4: 10 mg/mL SPA-GtrC-SGNH, 0.1 M BisTris pH 6.5, 2 M ammonium sulfate, B. Index A6: 20 mg/mL SPA-GtrC-SGNH, 0.1 M Tris pH 8.5, 2 M ammonium sulfate, C. Index F6: 10 mg/mL SPA-GtrC-SGNH, 0.1 M BisTris pH 5.5, 0.2 M ammonium sulfate, 25% PEG 3350, D. Index F7: 20 mg/mL SPA-GtrC-SGNH, 0.1 M BisTris pH 6.5, 0.2 M ammonium sulfate, 25% PEG 3350, E. Index G2: 20 mg/mL SPA-GtrC-SGNH, 0.1 M BisTris pH 5.5, 0.2 M lithium sulfate, 25% PEG 3350, F. Index G3: 20 mg/mL SPA-GtrC-SGNH, 0.1 M BisTris pH 6.5, 0.2 M lithium sulfate, 25% PEG 3350.

Two commercial crystallisation screens were used, Index and PACT, with two concentrations of SPA-GtrC-SGNH, 10 mg/mL and 20 mg/mL, to prepare sitting drop crystallisation trays. After incubation for 24-48 hours, crystals were seen in the Index screen in a wide range of conditions with most containing ammonium sulfate or lithium sulfate (Figure 15).

These conditions were used to prepare hanging drop optimisation trays covering a pH range from 5.5 to 8.5, 25% PEG 3350 and either 0.2-1.7 M ammonium sulfate ($(NH_4)_2SO_4$) or 0.1-0.3 M lithium sulfate (Li_2SO_4). SPA-GtrC-SGNH domain crystallised well in lithium sulfate and low concentrations of ammonium sulfate at all pH tested, with the best crystals grown in lithium sulfate (Figure 16). The best crystals were fished and tested for quality of diffraction. Two crystals diffracted well: one (crystal 4, Figure 16C, Figure 17B) showed ice rings but had good diffraction, whereas the other (crystal 2, Figure 16B, Figure 17A) had no ice rings but the diffraction was less clear due to a potential satellite crystal (Figure 17). Both crystals were sent to Diamond Light Source for data collection.

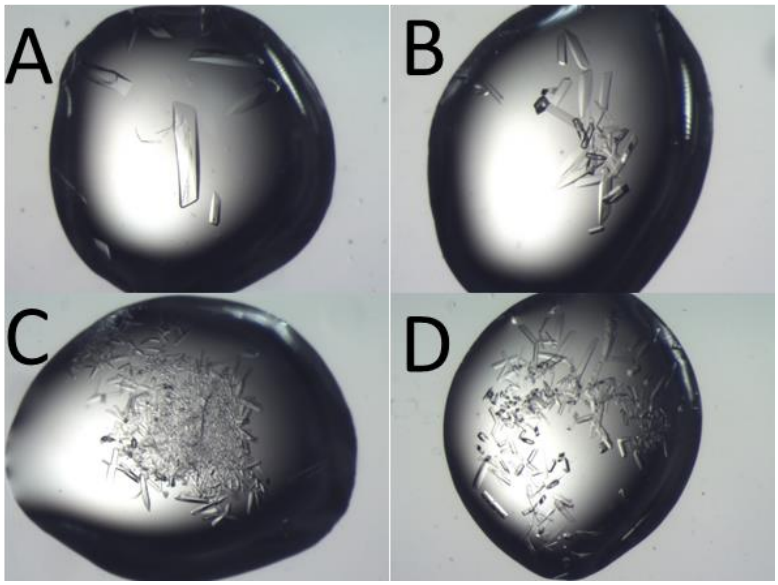


Figure 16. Crystals from optimisation screen, all wells contain 20 mg/mL SPA-GtrC-SGNH, 0.1 M BisTris pH 5.5, 25% PEG 3350, A. 0.15 M lithium sulfate, B. 0.2 M lithium sulfate (crystal 2 fished from this well), C. 0.25 M lithium sulfate (crystal 4 fished from this well), D. 0.3 M lithium sulfate.

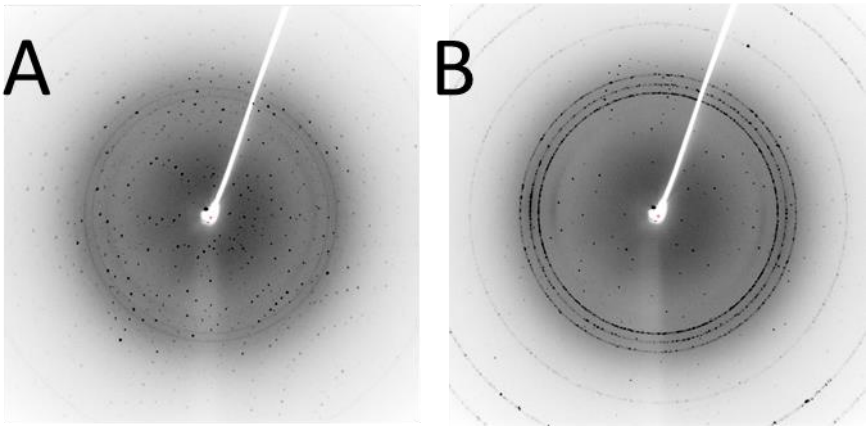


Figure 17. Diffraction from crystal 2 (A) and crystal 4 (B) in the conditions detailed in Figure 16.

Crystal 4, grown in 100 mM BisTris pH 5.5, 0.25 M lithium sulfate, 25% PEG 3350, diffracted to 1.1Å resolution with the parameters shown in Table 16. Data was scaled, merged and the space group was calculated to be $P2_12_12_1$. The Matthews coefficient was calculated to be 2.48$Å^3$/Da for a single molecule in the asymmetric unit with a solvent content of 50.37%. This suggests that there was one protein in the asymmetric unit. Fragon (77,78) was used to solve the structure using molecular replacement to place one poly-alanine helix, a correlation coefficient of 0.431 suggested that the structure had been solved and a model was built using ARP/wARP. Multiple cycles of model building using COOT and REFMAC were used to refine the structure to give an R value of 14.31% (Table 16).

Table 16. Data collection and refinement statistics for the structure of SPA-GtrC-SGNH

Data Collection Statistics	
Diamond beamline / wavelength (Å)	I04-1 / 0.92819
Spacegroup and cell dimension (Å)	P212121 (a = 55.74, b = 58.43, c = 90.04)
Resolution limits (Å)	49.01 – 1.10 (1.12 – 1.10)
No. unique reflections	119719 (5884)
Completeness (%)	99.9 (99.8)
Multiplicity	7.8 (7.0)
$Residual_{merge}$ (R_{merge})	0.074 (1.205)
Mean intensity/σIntensity (mean I/σI)	12.5 (1.4)
Pearson's correlation co-efficient ($CC_{1/2}$)	0.999 (0.697)
Structure solution	Fragon followed by Arp-wArp
Refinement	Refmac/Coot
Refinement Statistics	
R_{cryst}/R_{free} (%)	14.31 / 15.49
Root mean square deviation (RMSD) bond length (1-2) (Å)	0.010 (0.019)
RMSD angles (°)	1.277 (1.958)
RMSD chiral volumes ($Å^3$)	0.078 (0.200)
Average temperature factor (B-factor) ($Å^2$)	13.36
Ramachandran outliers (%)	0.82 %

5.3. Analysis of structure

The structure of SPA-GtrC-SGNH is similar to that of SGNH^{isol} domains found in the PDB in that the core consists of five β-sheets surrounded by six α-helices holding the catalytic residues in exactly the same location and orientation as for other SGNH domains (Figure 18). The linker region spans the length of the SGNH domain to the left of the catalytic residues ending at the top of the domain.

In comparison to SGNH^{isol} domains found in the PDB, SPA-GtrC-SGNH has an additional helix (α8 in Figure 18) and multiple loops at the top of the domain making the domain shape more elongated compared with other, more spherical, SGNH^{isol} domains. Helix α8 protrudes from the top of SPA-GtrC-SGNH domain (coloured red in Figure 18), this helix is not seen in the PDB structures of any SGNH^{isol} domains (Figure 18A).

5.3.1. Disulfide bond location

Two disulfide bonds are formed in the structure, as determined by mass spectrometry, one is within the linker region holding a small loop in place (Figure 18). The other is on the opposite side of the protein to the linker and also forms a small loop enabling Glu228 to protrude from the structure (Figure 18). In a sample of 100 sequences of AT3-SGNH domains, 26 sequences have a pair of cysteine residues, separated by four residues, in the same position as SPA-GtrC-SGNH. Of these 26 sequences, 16 have either a glutamic acid or aspartic acid as one of the middle two residues. This suggests that a negatively charged residue in this position may be important for the function of the SGNH domain. However, it is not conserved in all sequences with a pair of cysteine residues in this location, nor is it clear if there is a substitute in proteins which do not have a pair of cysteines. Further structural analysis of proteins with and without this pair of cysteines would need to be carried out to ascertain the conservation of this residue.

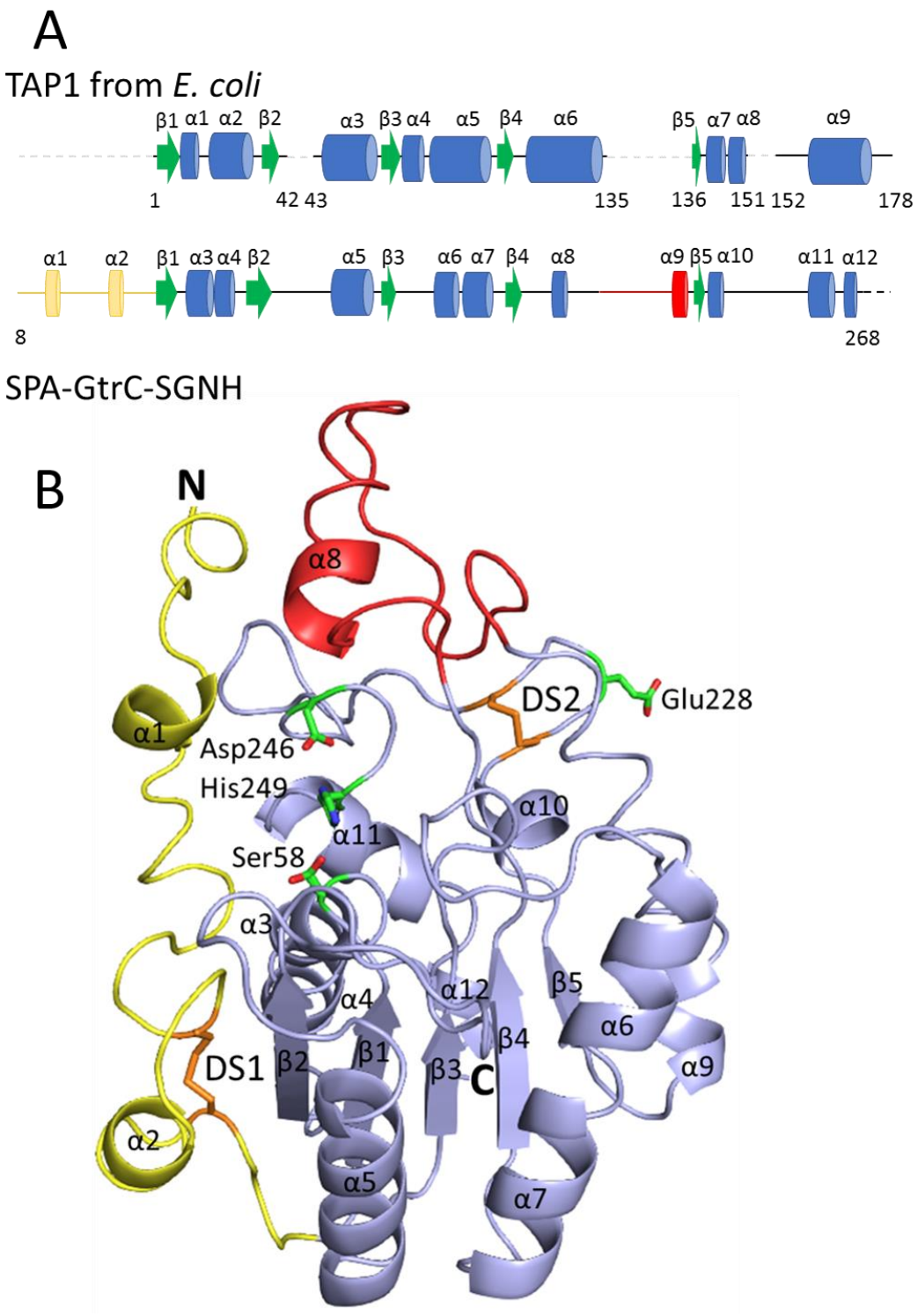


Figure 18. (A) Structural alignments of TAP1 (1U8U (90)) and SPA-GtrC-SGNH, linker shown in yellow and additional helix shown in red. (B) Structure of SPA-GtrC-SGNH, colours as for (A) disulfide bonds shown as orange sticks, catalytic residues and Glu228 shown as green sticks, with oxygen in red and nitrogen in blue. Helices labelled as (A).

5.3.2. Structure of linker region

The structure shows the linker region to be next to the SGNH domain, however, it is not clear whether this is the 'real' structure or if it has crystallised next to the SGNH domain for stability but would normally be flexible. To determine if the linker region structure is 'real' or an artefact of crystallisation the interface between the linker region and SGNH domain was examined in more detail. Many hydrophobic residues were found in close proximity suggesting that Van der Waals interactions may also stabilise this interface. In addition, a comparison of the hydrophobicity of the protein surface with and without the linker region showed the hydrophobicity to decrease significantly when the linker region was present (Figure 19). 45% of the surface accessible area was hydrophobic when the linker was removed compared to only 17% when the linker was present. In addition, the linker region was joined to the SGNH domain via 21 hydrogen bonds (Figure 20), meaning that more than half of the residues in the interface are forming hydrogen bonds. In conclusion, these analyses suggest that the structure of the linker is 'real', and rather than a flexible linker joining the SGNH domain to the acyltransferase domain, the linker region is part of the SGNH domain structure. This also implies that the distance between the SGNH domain and the acyltransferase domains is small.

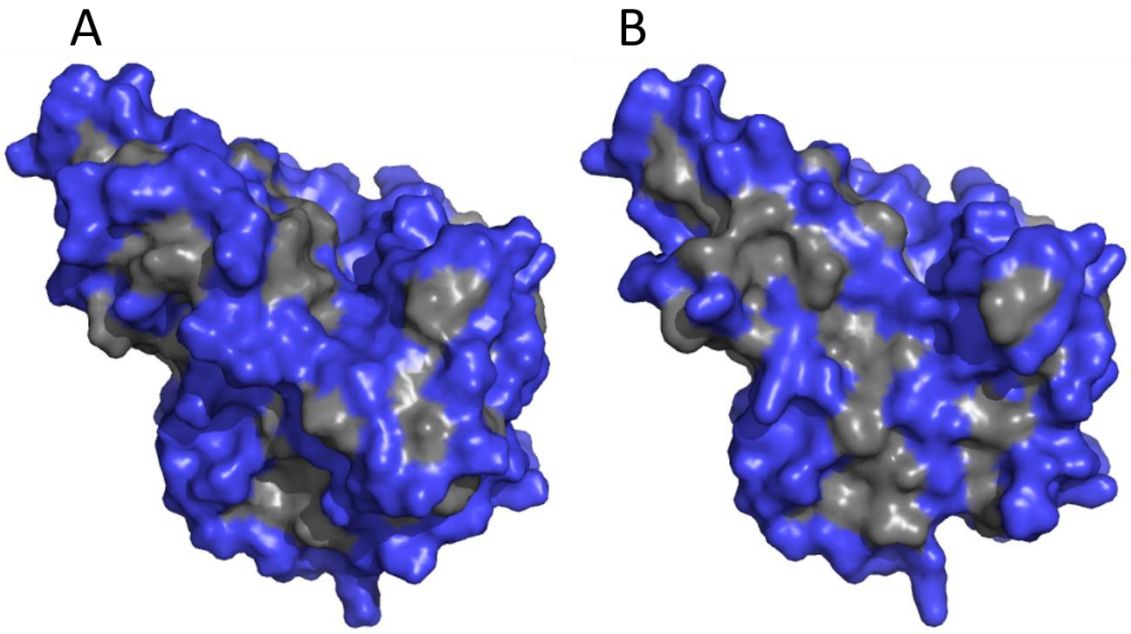


Figure 19. Surface hydrophobicity of SPA-GtrC-SGNH domain, surface coloured by residues property with hydrophobic residues shown in grey. Surface shown with (A) and without (B) the linker region

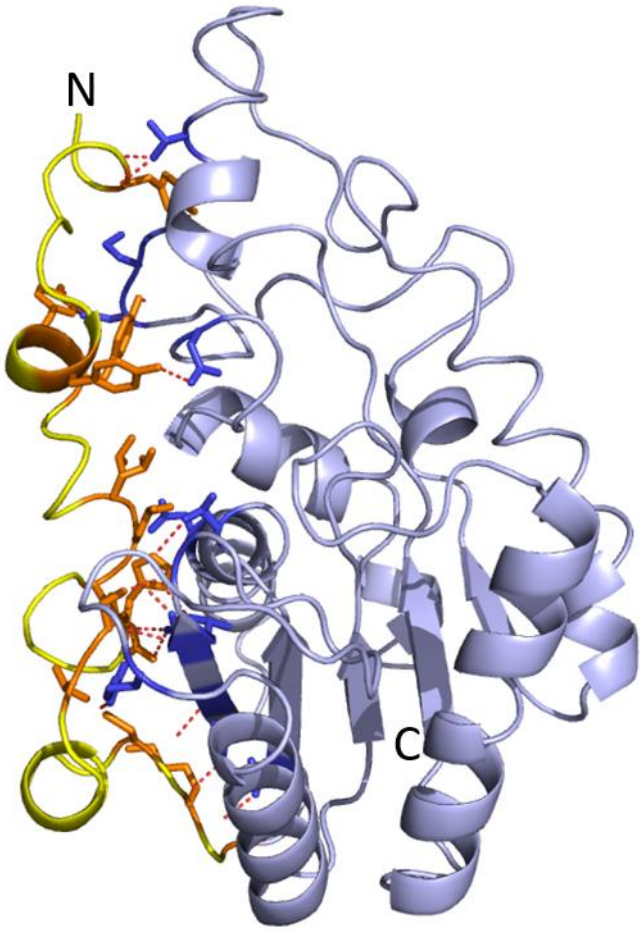


Figure 20. Hydrogen bonds from the linker to SGNH domain. Structure of SPA-GtrC-SGNH showing hydrogen bonds in red dashes, residues involved in hydrogen bonds are shown as sticks, orange in the linker and blue in the SGNH domain.

5.3.3. Additional helix present in SPA-GtrC-SGNH

SPA-GtrC-SGNH domain contains an additional helix (α8, Figure 18) in comparison to the structures of $SGNH^{isol}$ domains found in the PDB. Helix α8 protrudes from the top of the SGNH domain close to the end of the linker.

Sequence alignments of 100 $SGNH^{AT3}$ proteins suggests that some proteins have a gap in the alignment in the same region where SPA-GtrC-SGNH has the additional helix (Figure 21). This suggests that some $SGNH^{AT3}$ proteins have the additional helix seen in SPA-GtrC-SGNH whereas others do not. 35 out of 100 AT3-SGNH protein sequences contain amino acids in the same region that SPA-GtrC-SGNH has an additional helix, suggesting that these all contain this helix.

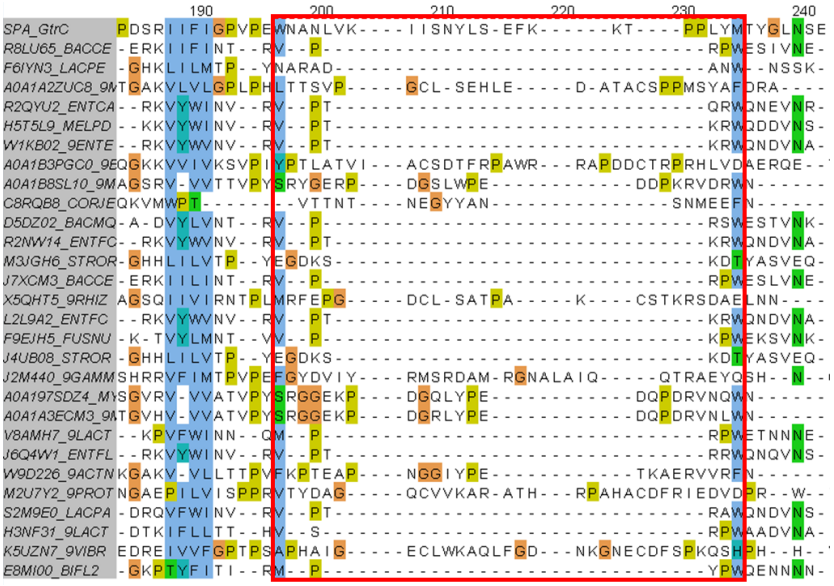


Figure 21. Sequence alignments of AT3-SGNH domains in the region where SPA-GtrC-SGNH has an additional helix, residues are seen in some proteins but not in all.

5.3.4. Structure of active site

The structure of SPA-GtrC-SGNH shows a prominent binding groove with the catalytic residues set in exactly the same location and orientation seen for other SGNH^{isol} domains. The binding site is comprised of four loops containing the four blocks of conserved residues (blocks I, II, III and V) (section 1.4). The loops holding the catalytic residues (in blocks I and V) are conserved in the sequence and structure of SPA-GtrC-SGNH domain (Figure 22). However, while the block II loop shows exactly the same structure, the sequence is not the same and an asparagine residue is in the location of the normal block II glycine (Figure 22). In addition, the block III loop is different in both sequence and structure from SGNH^{isol} domains found in the PDB. RGTNG is found in the same location as the normal block III sequence – GxND (where x is any amino acid. While the glycine and asparagine residue are conserved in sequence the structure of the block III loop in SPA-GtrC-SGNH is very different. The asparagine residue in the RGTNG motif is orientated with the side chain facing away from the active site suggesting it does not play a role in the formation of an oxyanion hole (Figure 22). This is strikingly different from the structure and orientation of the block III residues seen in SGNH^{isol} (Figure 3) and suggests the potential for an alternative catalytic mechanism.

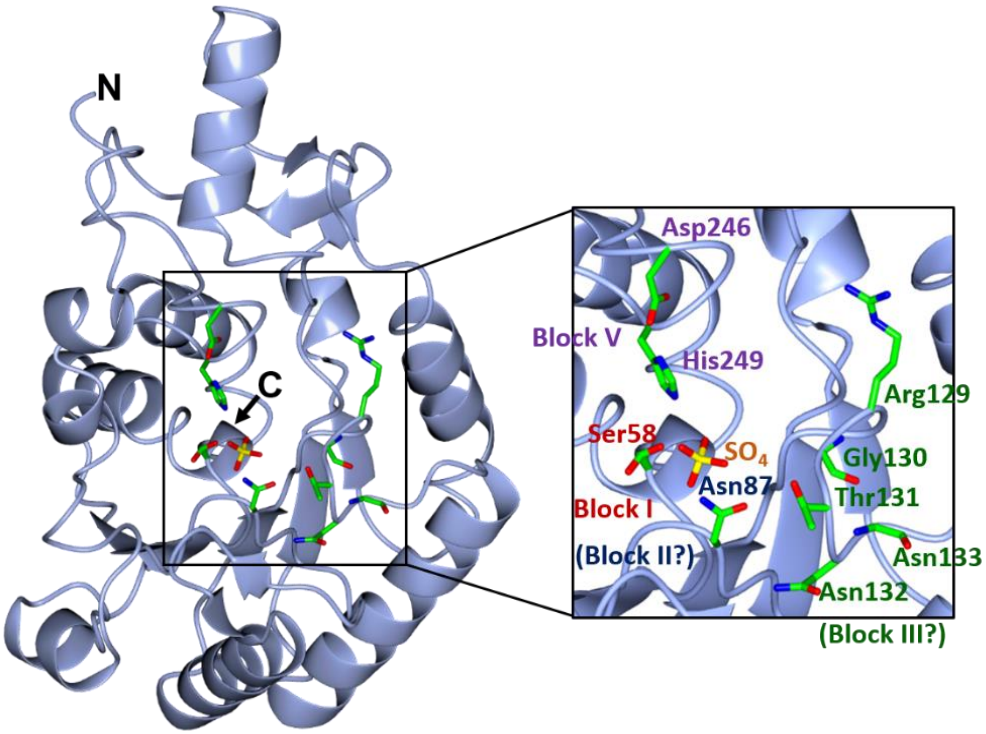


Figure 22. Active site of SPA-GtrC-SGNH domain with catalytic and potential oxyanion hole residues shown in sticks, a sulfate molecule present in the active site is also shown. Blocks II and III are shown in brackets as it is currently unknown which residues form the oxyanion hole in SPA-GtrC-SGNH.

5.3.5. Docking of small molecules into active site

FTMap models the docking of a set of small molecules with a wide range of properties and functional groups in a process mimicking ligand binding. This can be used to predict interactions between the protein and potential ligands (89). For SPA-GtrC-SGNH almost all of the small molecules formed clusters in the active site binding groove. Molecules which docked with the lowest energies of binding were cyclohexane, alcohols and esters, the functional groups which sugars contain (Figure 23A,B). Four non-overlapping cyclic molecules were predicted to dock in the binding groove with the orientation akin to four sugar residues (Figure 23A). Docking with FTMap using the sugar residues present in the LPS of *Salmonella* ser. Paratyphi A (mannose, galactose, rhamnose and paratose) also predicted sugar residues to dock within the binding groove (Figure 23C). However, FTMap appears unable to distinguish between individual sugar residues, showing galactose and mannose binding to be almost identical. This analysis was repeated using the PDB structure of TAP1 from *E. coli* (1U8U (90)), the natural substrates of which do not contain sugar molecules. The docking of sugar molecules was far more wide spread across

the protein surface (Figure 23D) – only 13 out of 48 sugar molecules docked close to the active site compared to 47 out of 48 sugar molecules docking within the binding groove in SPA-GtrC-SGNH. This suggests that FTMap is able to distinguish the binding sites of different SGNH domains despite both containing the same catalytic triad.

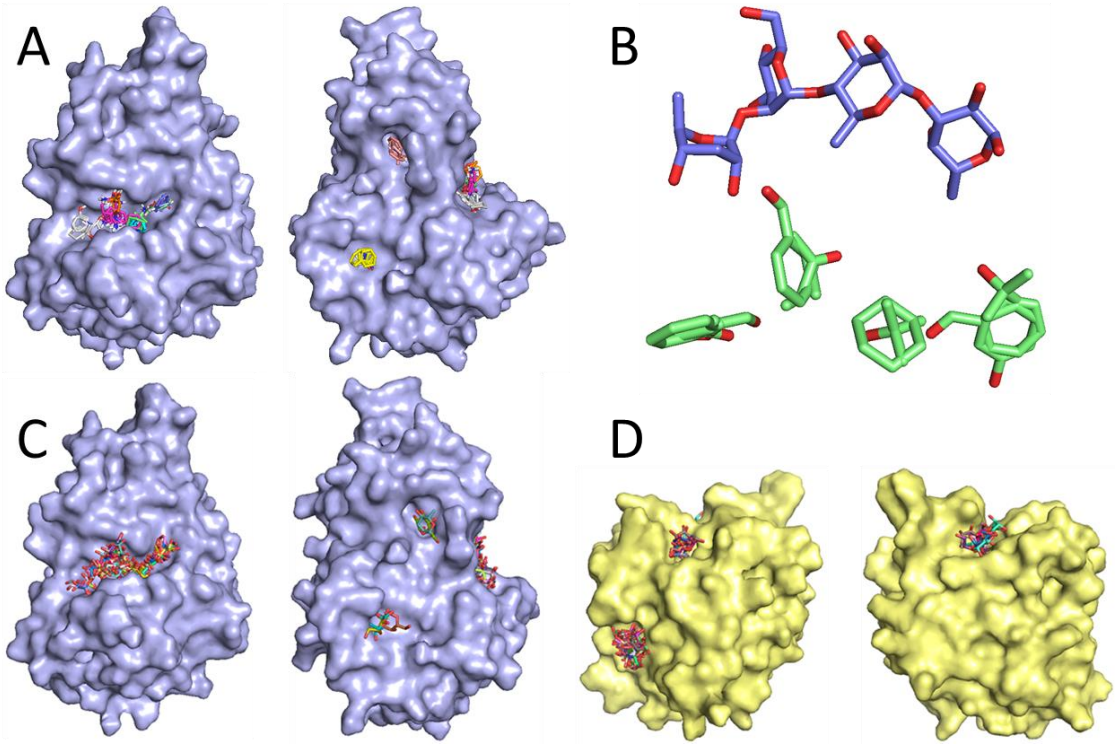


Figure 23. Docking predictions by FTMap (A) SPA-GtrC-SGNH front (left) and side (right) with small molecules bound as predicted by FTMap.(B) Single O-unit of Salmonella ser. Paratyphi A LPS (top) and small molecules bound in binding groove of SPA-GtrC-SGNH (bottom). (C, D) Individual sugars bound to SPA-GtrC-SGNH (C) front (left) and side (right), and TAP 1, E. coli (PDB 1U8U (90)) front (left) and back (right) as predicted by FTMap.

One of the lowest energy predicted binding sites for rhamnose shows the sugar to be orientated with the C2 position pointing towards the catalytic residues and within hydrogen bonding distance (2.2Å and 1.8Å) (Figure 24). In addition, the oxygens which would be involved in glycosidic bonds to mannose and galactose are pointing along the groove where the next sugar is likely to bind. These predictions by FTMap appear valid as it is known that SPA-GtrC-SGNH acetylates rhamnose at the C2 position and therefore this must be located close to the catalytic residues during catalysis.

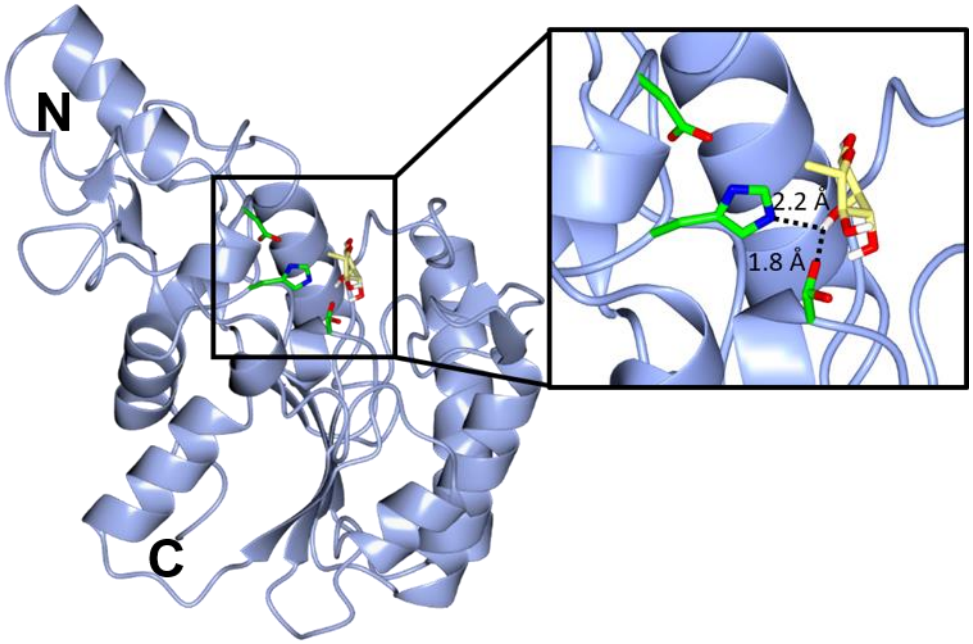


Figure 24. Docking predictions of rhamnose by FTMap, catalytic triad (sticks) of SPA-GtrC-SGNH (blue), with rhamnose residue shown in yellow, hydrogen bonds with distances between the rhamnose C2 and catalytic serine and histidine are shown.

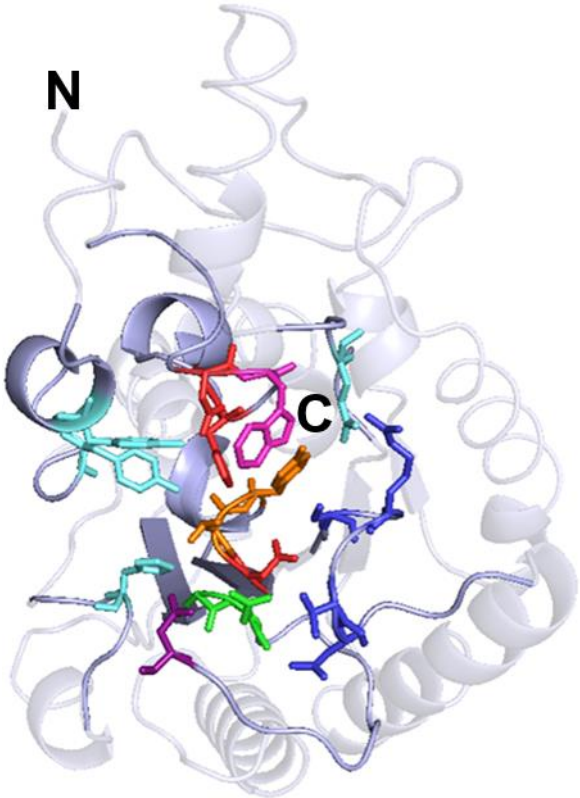


Figure 25. Residues predicted by FTMap to be involved in binding a ligand in the active site, catalytic residues are coloured red, those in block II position coloured green, and those in block III position coloured blue. Thr and Asn from block III GTNG are circled. The tryptophan predicted to bind most strongly is shown in pink.

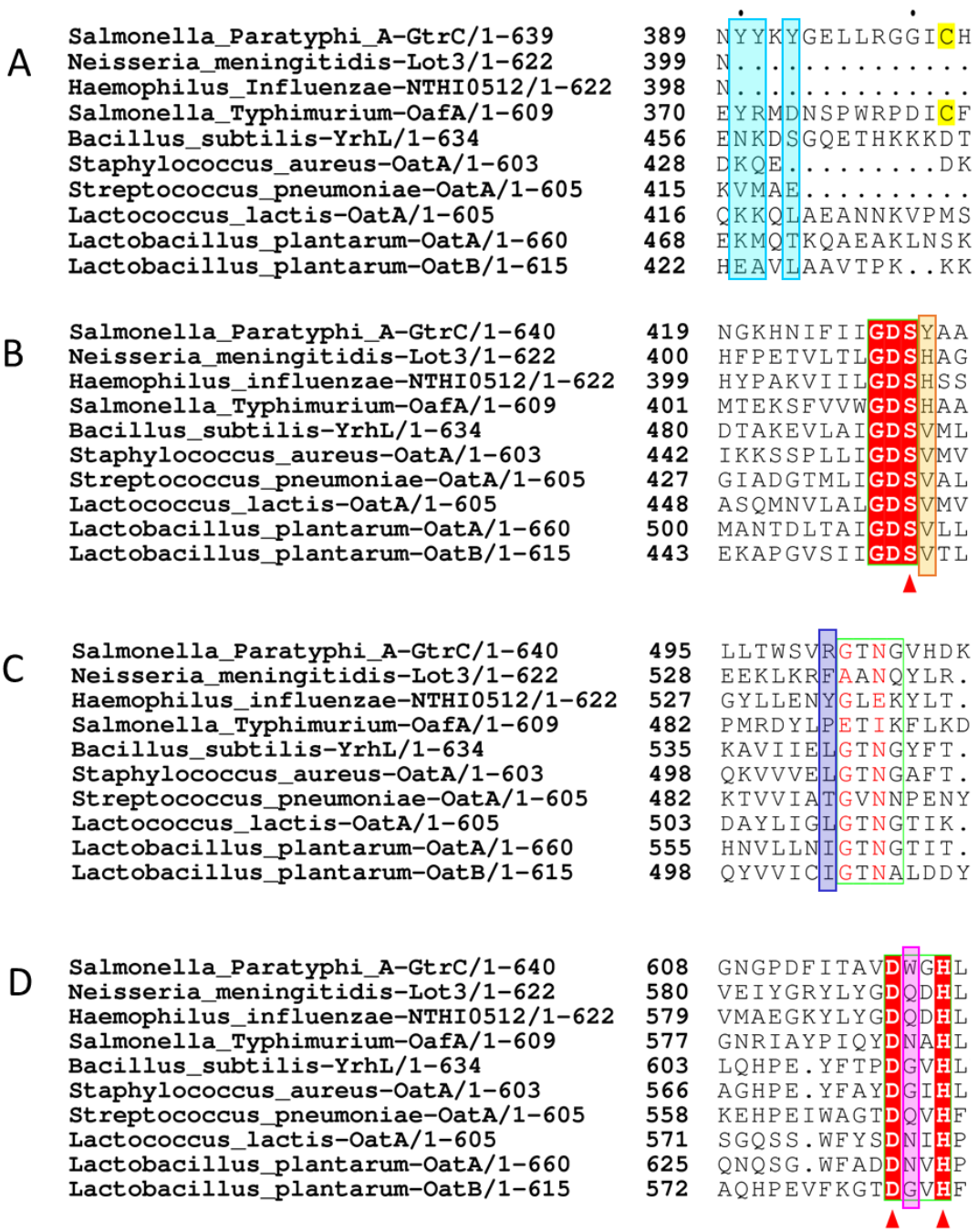


Figure 26. Sequence alignments highlighting residues predicted to be important for ligand binding. (A) 'Linker' region, (B) block I, (C) block III, (D) block V residues highlighted in colours same as those used in Figure 25.

While calculating the binding locations of small molecules, FTMap also assesses the residues which are likely to interact with ligands via either hydrogen bonds or electrostatic interactions. For SPA-GtrC-SGNH, the residues which are predicted to form the most interactions are a tryptophan close to the active site, two tyrosines located in the linker along with the catalytic residues and other residues located close to the active site (Figure 25). However, with the exception of the other Salmonella GtrC family II proteins, the residues which are strongly predicted to interact with the substrate are not conserved in sequence alignments of other AT3-SGNH domains (Figure 26).

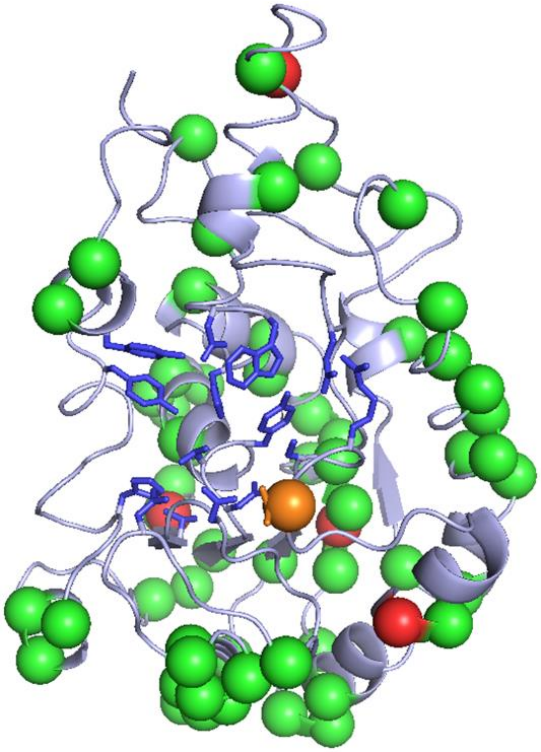


Figure 27. Mutations in residues between Salmonella ser. Paratyphi A GtrC compared to other GtrC proteins. Red spheres indicate mutations in either serovar Typhi, Dublin, Enteriditis or Gallinarum, green spheres indicate mutations in serovar Typhimurium. Blue sticks indicate residues predicted to be involved in binding in serovar Paratyphi A. Orange sphere and stick indicates residues involved in binding and mutated in Salmonella ser. Typhimurium.

The GtrC family II proteins found in other Salmonella serovars show high sequence identity to SPA-GtrC-SGNH (>96%, with the exception of Salmonella ser. Typhimurium – 77%). Any mutations between the GtrC family II proteins were mapped on to the structure of SPA-GtrC-SGNH, all mutations were found to be on the outside of the structure. Only one mutation was of a residue

predicted by FTMap to be important for binding and was located close to the active site. However, as this mutation was of a threonine to a serine it is unlikely to alter ligand binding (Figure 27). However, when comparing the sequences of other $SGNH^{AT3}$ proteins, none of the residues predicted to be important for binding were conserved. The GTNG motif observed previously (section 3.2) is shown to be less important for binding and is located further from the catalytic residues than the normal block III motif found in other SGNH domains. More research is needed to study the docking of substrates in the active site of SPA-GtrC-SGNH and to determine which residues are important for binding and forming the oxyanion hole.

Biophysical and structural analysis of SPA-GtrC-SGNH domain has determined that disulfide bonds are formed between the pairs of cysteines highlighted in previous sections. However, while these disulfide bonds are not important for protein folding, one may play a role in the function of the SGNH domain as it enables Glu228 to extend from the structure. In addition, docking analysis suggests that SPA-GtrC-SGNH is likely to bind sugar residues and has highlighted key residues involved.

Chapter 6. Discussion

6.1. Expression conditions

The cytoplasm is a reducing environment which prevents disulfide bonds from forming (65). Both NM-Lot3-SGNH and SPA-GtrC-SGNH contain pairs of cysteine residues which were predicted to form disulfide bonds (Figure 10). Therefore, expression of these proteins in the cytoplasm may result in protein misfolding, aggregation and inclusion body formation. Protein expression with the addition of a PelB leader sequence targets the protein to the periplasm (63) which enables the disulfide bonds to form and therefore is likely to result in a folded and more stable protein. However, periplasm expression of both NM-Lot3-SGNH and SPA-GtrC-SGNH resulted in only insoluble protein (Figure 11). There are a number of steps in the expression process which may have been causing this outcome: firstly, there is no evidence the proteins reached the periplasm. The proteins are transported unfolded, post-translation, therefore, there is the potential for aggregation to occur before the protein can be translocated. Secondly, transport of proteins to the periplasm occurs via the SecYEG translocon (68); it is thought that over-expression of periplasmic proteins can overwhelm the Sec-YEG translocon (63). Lemo21 cells were used to overcome this problem, however, no soluble expression was seen (Figure 11), suggesting that the SecYEG translocon was not the problem. Thirdly, the protein may have been unable to fold alone; the SGNH domains used would normally be attached to an acyltransferase domain and both domains of the protein may be required for correct folding. However, cytoplasmic expression, without the presence of a solubility tag, showed that the SGNH domains are soluble (Figure 12), and therefore able to fold alone, suggesting this was not the problem.

Cytoplasm expression combined with a solubility tag resulted in the production of soluble protein. Origami cells are optimised for the expression of proteins containing disulfide bonds (66). As expected, expression in Origami cells resulted in an increase in soluble protein production in comparison to BL21 DE3

cells where disulfide bonds are unable to form in the cytoplasm (Figure 12). This suggests that disulfide bonds are important for folding but the protein is able to fold without the presence of disulfide bonds, as a small amount of soluble expression was seen in BL21 DE3 cells (Figure 12). The addition of a solubility tag has been shown to increase the solubility of poorly soluble protein (72). In particular, the addition of an MBP tag has been shown to act as a chaperone and aid folding (73). Addition of a solubility tag increased soluble expression of both NM-Lot3-SGNH and SPA-GtrC-SGNH (Figure 12).

It is unknown the exact reasons why both SGNH domains were insoluble when expressed in the periplasm. However, a combination of cytoplasmic expression in Origami cells with the addition of an MBP solubility tag enabled a high yield of soluble SGNH proteins to be produced.

6.2. Disulfide bond formation

Disulfide bonds reduce the difference in entropy between the folded and unfolded protein and therefore commonly play a role in stability of a protein: by constraining surface loops; locking the protein into a particular conformation; or replacing the hydrophobic core (91). Mass spectrometry determined that two disulfide bonds were formed in SPA-GtrC-SGNH (Figure 14) and the structure confirmed that these both hold small loops (Figure 18). This is not uncommon, and there are many examples of protein structures where a disulfide bond is involved in stabilising the structure of a loop (92).

Circular dichroism showed that the formation of disulfide bonds has no impact on the thermal stability of unfolding (Figure 14). Both the reduced and oxidised forms of the protein unfolded at the same temperature and no major differences were seen in the spectra (Figure 14). Similar experiments show that addition of TCEP to reduce disulfide bonds alters the circular dichroism spectra in the same way as does the mutation of the cysteine pair (93). In addition, reduction of disulfide bonds commonly results in a dramatic decrease in melting temperature and thus, thermal stability (93). Whilst there is no guarantee that the disulfide bonds are reduced, as they may be inaccessible to reduction by

TCEP, the structure shows the cysteines to be close to the surface of the protein (Figure 18) and TCEP was used in molar excess. Mass spectrometry could be used to verify that the disulfide bonds in SPA-GtrC-SGNH were reduced.

Both disulfide bonds hold a small loop, consisting of 12 residues in the 'linker' and four residues close to Glu228 consistent with the evidence that these disulfide bonds do not increase stability. Studies have been carried out showing that a longer loop between the two cysteines increases the entropy involved in stabilisation of the protein (94). Additional experiments could be carried out using differential scanning calorimetry (DSC) to calculate the Gibbs free energy of unfolding to further probe the effect of disulfide bonds on stability.

Residue Glu228 is conserved in other $SGNH^{AT3}$ domains, suggesting that this residue may be important either in the function of the SGNH domain or in contacting the acyltransferase domain. Therefore, the disulfide bond close to Glu228 may be involved in stabilising this loop and enabling Glu228 to protrude from the structure (Figure 18). Further experiments involving mutating Glu228 and examining function of SPA-GtrC would need to be carried out to ascertain the involvement of this residue. In addition, a structure of the acyltransferase domain attached to the SGNH domain would enable interactions between the domains to be investigated.

6.3. Residues involved in catalysis and substrate binding

SPA-GtrC-SGNH is the first structure of an $SGNH^{AT3}$ domain and although the core structure of the protein is very similar to $SGNH^{isol}$ proteins, there are some key differences. The 'linker' region which joins the $SGNH^{AT3}$ domain to the $AT3^{SGNH}$ domain was previously thought to be flexible and therefore not change the overall structure of the $SGNH^{AT3}$. However, the structure of SPA-GtrC-SGNH shows the 'linker' region to be structured alongside the SGNH domain and this structure is not present in $SGNH^{isol}$ domains. Furthermore, SPA-GtrC-SGNH has an additional helix which is not present in structures of $SGNH^{isol}$

domains found in the PDB. The function of this additional helix is currently unknown. In addition, the loops surrounding the active site and binding groove differ in SPA-GtrC-SGNH compared to SGNH^{isol} domains.

The residues involved in catalysis by SGNH^{isol} domains are well characterised and comprise of four conserved blocks of sequence: block I = GDS, block II = G, block III = GxND and block V = DxxH (where x can be any amino acid) (95). Blocks I and V contain the catalytic residues and blocks II and III comprise of residues involved in stabilising the transition state and forming an oxyanion hole (95). As expected, sequence analysis of SGNH^{AT3} domains showed blocks I and V, the catalytic residues, to be present in the sequence (Figure 8). Molgaard et al. suggested that the block III GxND motif is completely conserved among SGNH^{isol} (53) however, both the block II and block III sequences are not present in SGNH^{AT3} (Figure 8). In addition, when examining the structure of SPA-GtrC-SGNH domain, the structure and sequence of the loops normally holding the block III residues was very different to that seen in SGNH^{isol} domains (Figure 22). This suggests that the oxyanion hole residues, normally conserved among SGNH^{isol} domains, are not present in SGNH^{AT3} domains.

Assuming that the reaction catalysed by SPA-GtrC-SGNH proceeds via a negatively charged transition state (Figure 4), as seen for other SGNH^{isol} domains (57), an oxyanion hole must be formed to stabilise the transition state. Therefore, other residues must take the place of the block III GxND motif seen in SGNH^{isol} domains. Prior to determining the structure, it was suggested that the GTNG motif (section 3.3) could be replacing the GxND motif as two out of three of the residues are the same. However, this GTNG motif is only present in around 40% of AT3-SGNH proteins (Table 15). In addition, on determining the structure, while the GTNG motif is close to the binding groove, the orientation of the asparagine and glycine residues are very different to that seen in SGNH^{isol} domains (Figure 22). For example, in SPA-GtrC-SGNH, the asparagine residue of GTNG is orientated with the side chain amide facing away from the binding groove (Figure 22), whereas in SGNH^{isol} domains the side chain amide of the asparagine in GxND is orientated towards the catalytic residues. Both the block II glycine and block III asparagine are of vital importance in stabilising the negatively charged transition state formed during the proposed mechanism of

acetylation (Figure 4). Asn87 seen in SPA-GtrC-SGNH in the same location as the block II glycine in SGNH^{isol} domains is likely to act as a replacement oxyanion hole residue, however, no replacement for the block III Asn is seen. The proposed mechanism (Figure 4) suggests that the reaction proceeds via a negatively charged transition state which must be stabilised by an oxyanion hole. A lack of obvious block III oxyanion hole residues in SPA-GtrC-SGNH suggests that either the SGNH^{AT3} domain is a pseudo-domain and does not catalyse the reaction or that the reaction does not proceed via the proposed mechanism and another mechanism is used. Previous studies have shown that the SGNH^{AT3} domain is required for function, therefore it seems highly unlikely that this domain is a pseudo-domain (33). Consequently, it is more likely that the proposed mechanism is incorrect and possibly a negatively charged transition state is not formed as SPA-GtrC-SGNH does not contain the residues required to stabilise a negatively charged transition state. However, significantly more work is required to establish the mechanism and oxyanion hole residues.

Docking experiments carried out using FTMap suggest that the residues most likely to be involved in binding substrates (aside from the catalytic residues) are aromatic residues: Tyr17, Tyr22 and Trp247 (Figure 25, Figure 26). Stacking interactions between aromatic residues and a sugar substrate are common with tryptophan being the most common residue involved (96). This is consistent with the assumption that the substrate of SPA-GtrC-SGNH is sugar based. However, sequence analysis showed that these aromatic residues are not conserved among other AT3-SGNH domains.

Docking experiments, sequence alignments and structural analysis were unable to highlight any residues which are obvious candidates to replace the oxyanion hole residues. Pfeffer et al. (2012) showed that mutation of the oxyanion hole asparagine to an alanine knocks out function of the SGNH domain as effectively as if the catalytic residues were mutated (51). This suggests that SPA-GtrC-SGNH may be acting via a different mechanism to that of SGNH^{isol} domains.

PatB, an SGNH^{isol} domain which acts in conjunction with PatA, an AT3^{isol} protein, is assumed to act via a similar mechanism to SPA-GtrC-SGNH (61). Although not connected, as is the case for AT3-SGNH proteins, this system

contains a transmembrane acyltransferase and an SGNH^{isol} hydrolase the same as AT3-SGNH proteins (61). However, unlike SPA-GtrC-SGNH, PatB does contain the block II and block III consensus sequences seen in other SGNH^{isol} domains, and although the structure is not known, it is thought to be similar to other SGNH^{isol} domains (61).

In conclusion, while the conserved oxyanion hole residues are important for catalysis in SGNH^{isol} domains, these residues are not conserved in AT3-SGNH domains and no obvious replacement residues have been uncovered. Determining the structure of SPA-GtrC-SGNH in complex with its substrate could confirm the residues important for binding and potentially find the oxyanion hole residues.

6.4. Placement of SGNH domain in relation to the acyltransferase domain

The discovery of the structured linker can be used to suggest the orientation of the acyltransferase and SGNH domains in relation to each other. It was previously assumed that the region linking the two domains was flexible however, the structure shows that the 'linker' region is an integral part of the SPA-GtrC-SGNH domain structure (Figure 20). The acyltransferase domain, and therefore, cytoplasmic membrane is likely to be situated close to the N-terminus (Figure 28). This suggests that the additional helix, not present in all AT3-SGNH proteins, may come into close contact with the acyltransferase domain (Figure 28). FTMap predicts sugar residues to bind along the groove in SPA-GtrC-SGNH, therefore, this groove is likely to be located perpendicular to the cytoplasmic membrane to allow contact with the O-antigen protruding from the membrane (Figure 28). This implies that Glu228 would also be located close to the acyltransferase domain and potentially forms a salt-bridge with a positively charged conserved lysine in the acyltransferase domain (Figure 28).

PatA and PatB can again be used as a model to look into how these domains may work together. Moynihan et al. suggested that PatA, after transporting the acetate substrate across the membrane, remains covalently linked to the acetyl

group. PatB, the SGNH domain, removes the acetate group from PatA and attaches it to peptidoglycan (61). This seems more likely than CoA being transported into the periplasm along with the acetate group and also is consistent with the theory that a strong interaction occurs between the acyltransferase and SGNH domains. Potentially the long periplasmic loop found between transmembrane helices 3 and 4 in the acyltransferase domain of SPA-GtrC (Figure 28), could be involved in transporting the acetate donor to the SGNH domain. If this is true, then the acyltransferase loop would need to interact closely with the binding groove and may even provide the oxyanion hole residues missing from the SGNH domain. However, significant further structural analysis of both the acyltransferase and SGNH domain is required to determine how these domains work together.

The point at which the acyltransferase proteins act during the biosynthesis of LPS is currently unknown. During biosynthesis, each individual O-unit, consisting of four sugar residues in *Salmonella* ser. Paratyphi A, are assembled to form a long O-antigen chain consisting of around 50 repeating O-units (20). There are two possibilities, each individual O-unit could be acetylated, or the acyltransferase could act on the assembled O-antigen chain. If each individual O-unit is O-acetylated then the $SGNH^{AT3}$ domain would need to be in close proximity to the membrane as the rhamnose residue to be acetylated would be the second sugar from the membrane (Figure 28). If the $SGNH^{AT3}$ domain is orientated as shown in Figure 28, it could potentially be close enough to the membrane to act on the individual O-unit.

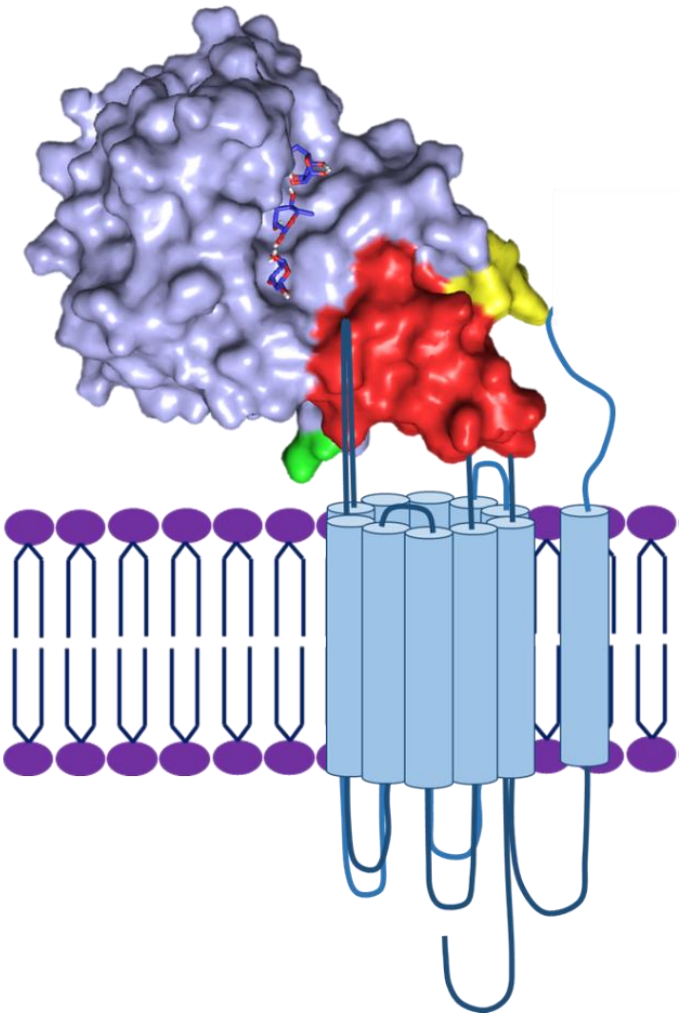


Figure 28. Proposed structure of interaction between the SGNH domain and acyltransferase domain. The additional helix is shown in red, the N-terminus of the linker in yellow and Glu228 in green. The sugar residues predicted to bind in the active site are shown as blue sticks (as predicted by FTMap).

The other possibility is that the SGNH domain acts on the long O-antigen, potentially in conjunction with the polymerisation step. However, as the new O-unit is attached to the end furthest away from the membrane this would require the SGNH domain to move further away from the membrane as the chain became longer. Therefore, it seems more likely that SPA-GtrC acts on each individual O-unit before polymerisation occurs.

The structure of SPA-GtrC-SGNH domain has given some understanding of the location of the $SGNH^{AT3}$ domain in relation to the acyltransferase domain.

However, determining the structure of the two domains together is required to gain further insight into how these domains may work together.

6.5. Further experiments

The in silico analysis and structural work carried out here has given a new insight into $SGNH^{AT3}$ domains and highlighted some important differences compared with $SGNH^{isol}$ domains. However, further work is required to gain a better understanding of the differences in function between $SGNH^{AT3}$ and $SGNH^{isol}$. Mutagenesis experiments could be carried out altering the residues surrounding the binding groove of SPA-GtrC-SGNH domain, in particular the aromatic residues highlighted previously. Combined with a functional assay this would enable a detailed understanding of which residues are important for catalysis and substrate binding. SPA-GtrC-SGNH is the first structure of an $SGNH^{AT3}$, therefore, structural work on other $SGNH^{AT3}$ domains would determine if the differences between $SGNH^{AT3}$ and $SGNH^{isol}$ seen here are consistent with other proteins.

Analysis of the sequences of GtrC family II proteins from each of these serovars and comparing the location of mutations in the sequences to the structure of SPA-GtrC-SGNH shows that none of these mutations are close to the active site or binding groove of SPA-GtrC-SGNH (Figure 27). This suggests that the mutations present are unlikely to influence substrate specificity and it therefore seems likely that each GtrC protein will be able to act on the LPS structures from other Salmonella serovars. Additional experiments could be carried out to determine if the GtrC protein, or even just the SGNH domain, from Salmonella ser. Paratyphi A was cloned into Salmonella ser. Typhimurium it would be able to O-acetylate the LPS.

In addition, structural work of the AT3-SGNH protein could be carried out as determining the complete structure would advance the understanding of how the domains combine to form a complete protein. This could also potentially help to understand the mechanism of how the domains work together to carry out the acetyltransferase reaction. Detailed analysis of the acyltranferase

domain structure comparing AT3-SGNH to $AT3^{isol}$ proteins could also enhance the understanding of how these domains differ and identify key residues involved in catalysis or at the interface of the acyltransferase and SGNH domains.

Furthermore, NMR experiments could be carried out to study the binding of sugar residues in the active site of SPA-GtrC-SGNH. This would determine the residues which are involved in substrate binding. Similarly, crystallisation with sugar monomers or the O-unit structure from *Salmonella* ser. Paratyphi A, would also enable detailed analysis of how the substrate binds.

O-acetylation carried out by NM-Lot3 and SPA-GtrC has shown to be important for immune evasion. Therefore it is vital to understand how these proteins function and their structure. The aims of this project were to examine the structural and mechanistic differences between AT3-SGNH domains and SGNH hydrolases. Detailed sequence analysis showed key residues, known to be important for catalysis to be missing from $SGNH^{AT3}$ suggesting that these domains function differently. Determining the structure of SPA-GtrC-SGNH supported this, showing the structure of the active site to be different to the structures of other SGNH hydrolases. Further work is required to gain a more detailed understanding of the mechanism of AT3-SGNH domains and how this may differ from other acyltransferase and SGNH proteins.

Abbreviations

AT3^{isol}	Acyltransferase family 3 protein not attached to an SGNH domain
AT3-SGNH	Acyltransferase family 3 protein attached to an SGNH domain
B-factor	Temperature factor
BSA	Bovine serum albumin
CC1/2	Pearson's correlation co-efficient
CD	Circular dichroism
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
Gal	Galactose
Glc	Glucose
GlcN	Glucosamine
GlcNAc	N-acetyl glucosamine
Hep	Heptose
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HT	Photomultiplier tube
I	Intensity
Im9	Immunity protein 9
IMAC	Immobilised metal affinity chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
Kdo	Keto-deoxyoctulosonate
LB	Lysogeny broth
LPS	Lipopolysaccharide
Man	Mannose
MBP	Maltose binding protein
MurNAc	N-acetyl muramic acid
MWCO	Molecular weight cut off
NM-Lot3	*Lot3* protein from *N. meningitidis*
NM-Lot3-SGNH	*Lot3* protein from *N. meningitidis,* amino acid 379-end
O-Ac	O-Acetylation

OD_{600}	Optical density measured at wavelength 600 nm
PACT	pH, anion, cation crystallisation trial
Par	Paratose
PCR	Polymerase chain reaction
PDB	Protein databank
PEA	Phosphoethanolamine
PEG	Polyethylene glycol
Phos	Phosphate
R-factor	Residual factor
Rha	Rhamnose
RMS	Root mean square
SDS PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
$SGNH^{AT3}$	SGNH domain attached to an acyltransferase
$SGNH^{isol}$	SGNH domain not attached to an acyltransferase
SOC	Super optimal broth with catabolite repression
SPA-GtrC	*GtrC* protein from *Salmonella* ser. Paratyphi A
SPA-GtrC-SGNH	*GtrC* protein from *Salmonella* ser. Paratyphi A, amino acid 371-end
TCEP	Tris(2-carboxyethyl)phosphine
UDP	Uracil diphosphate
UndP	Undecaprenyl phosphate

References

1. Crump, J. A., Luby, S. P., and Mintz, E. D. (2004) The global burden of typhoid fever. *Bull. W.H.O.* **82**, 346-353
2. Fangtham, M., and Wilde, H. (2008) Emergence of *Salmonella paratyphi A* as a major cause of enteric fever: need for early detection, preventive measures, and effective vaccines. *J. Travel Med.* **15**, 344-350
3. Maskey, A. P., Day, J. N., Tuan, P. Q., Thwaites, G. E., Campbell, J. I., Zimmerman, M., Farrar, J. J., and Basnyat, B. (2006) *Salmonella enterica* serovar Paratyphi A and *S. enterica* serovar Typhi cause indistinguishable clinical syndromes in Kathmandu, Nepal. *Clin. Infect. Dis.* **42**, 1247-1253
4. Andrews, J. R., and Ryan, E. T. (2015) Diagnostics for invasive Salmonella infections: current challenges and future directions. *Vaccine* **33**, C8-C15
5. Mohanty, S., Renuka, K., Sood, S., Das, B., and Kapil, A. (2006) Antibiogram pattern and seasonality of *Salmonella* serotypes in a North Indian tertiary care hospital. *Epidemiol. Infect.* **134**, 961-966
6. Shetty, A. K., Shetty, I. N., Furtado, Z. V., Antony, B., and Boloor, R. (2012) Antibiogram of *Salmonella* isolates from blood with an emphasis on nalidixic acid and chloramphenicol susceptibility in a tertiary care hospital in coastal Karnataka: a prospective study. *J Lab Physicians* **4**, 74
7. Konadu, E., Shiloach, J., Bryla, D. A., Robbins, J. B., and Szu, S. C. (1996) Synthesis, characterization, and immunological properties in mice of conjugates composed of detoxified lipopolysaccharide of *Salmonella* paratyphi A bound to tetanus toxoid with emphasis on the role of O acetyls. *Infect. Immun.* **64**, 2709-2715
8. Micoli, F., Rondini, S., Gavini, M., Lanzilao, L., Medaglini, D., Saul, A., and Martin, L. B. (2012) O: 2-CRM 197 conjugates against *Salmonella* paratyphi A. *PLoS One* **7**, e47039
9. Konadu, E. Y., Lin, F.-Y. C., Hó, V. A., Thuy, N. T. T., Van Bay, P., Thanh, T. C., Khiem, H. B., Trach, D. D., Karpas, A. B., and Li, J. (2000) Phase 1 and phase 2 studies of *Salmonella enterica* serovar paratyphi A O-specific polysaccharide-tetanus toxoid conjugates in adults, teenagers, and 2-to 4-year-old children in Vietnam. *Infect. Immun.* **68**, 1529-1534
10. Ali, A., An, S. J., Cui, C., Haque, A., and Carbis, R. (2014) Synthesis and immunogenicity evaluation of *Salmonella enterica* serovar Paratyphi A O-specific polysaccharide conjugated to diphtheria toxoid. *Hum. Vaccin. Immunother.* **10**, 1494-1498
11. Rouphael, N. G., and Stephens, D. S. (2012) *Neisseria meningitidis*: biology, microbiology, and epidemiology. *Neisseria meningitidis: advanced methods and protocols*, 1-20
12. Frasch, C. E., Kapre, S. V., Lee, C.-H., and Préaud, J.-M. (2015) Technical Development of a New Meningococcal Conjugate Vaccine. *Clin. Infect. Dis.* **61**, S404-S409
13. Girard, M. P., Preziosi, M.-P., Aguado, M.-T., and Kieny, M. P. (2006) A review of vaccine research and development: meningococcal disease. *Vaccine* **24**, 4692-4700

14. WHO. (2015) Meningococcal meningitis, factsheet no. 141.
15. LaForce, F. M., Ravenscroft, N., Djingarey, M., and Viviani, S. (2009) Epidemic meningitis due to Group A *Neisseria meningitidis* in the African meningitis belt: a persistent problem with an imminent solution. *Vaccine* **27**, B13-B19
16. Silhavy, T. J., Kahne, D., and Walker, S. (2010) The bacterial cell envelope. *Cold Spring Harb. Perspect. Biol.* **2**, a000414
17. Wheelis, M. (2011) *Principles of Modern Microbiology*, Jones & Bartlett Learning
18. Alexander, C., and Rietschel, E. T. (2001) Invited review: bacterial lipopolysaccharides and innate immunity. *J. Endotoxin Res.* **7**, 167-202
19. Knirel, Y. A. (2011) Structure of O-antigens. in *Bacterial Lipopolysaccharides*, Springer. pp 41-115
20. Holst, O., Ulmer, A. J., Brade, H., Flad, H.-D., and Rietschel, E. T. (1996) Biochemistry and cell biology of bacterial endotoxins. *FEMS Immunol. Med. Microbiol.* **16**, 83-104
21. Wilkinson, S. G. (1996) Bacterial lipopolysaccharides—themes and variations. *Prog. Lipid Res.* **35**, 283-343
22. Kulshin, V. A., Zähringer, U., Lindner, B., Frasch, C., Tsai, C.-M., Dmitriev, B., and Rietschel, E. T. (1992) Structural characterization of the lipid A component of pathogenic *Neisseria meningitidis*. *J. Bacteriol.* **174**, 1793-1800
23. Choudhury, B., Kahler, C. M., Datta, A., Stephens, D. S., and Carlson, R. W. (2008) The structure of the L9 immunotype lipooligosaccharide from *Neisseria meningitidis* NMA Z2491. *Carbohydr. Res.* **343**, 2971-2979
24. Samuel, G., and Reeves, P. (2003) Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. *Carbohydr. Res.* **338**, 2503-2519
25. Hellerqvist, C. G., and Lindberg, B. (1971) Structural Studies on the O-Specific Side-chains of the Cell-wall Lipopolysaccharide from *Salmonella*. *Acta Chem. Scand.* **25**, 955-961
26. Ravenscroft, N., Cescutti, P., Gavini, M., Stefanetti, G., MacLennan, C., Martin, L., and Micoli, F. (2015) Structural analysis of the O-acetylated O-polysaccharide isolated from *Salmonella* paratyphi A and used for vaccine preparation. *Carbohydr. Res.* **404**, 108-116
27. Kahler, C. M., Lyons-Schindler, S., Choudhury, B., Glushka, J., Carlson, R. W., and Stephens, D. S. (2006) O-Acetylation of the terminal N-acetylglucosamine of the lipooligosaccharide inner core in *Neisseria meningitidis* influence on inner core structure and assembly. *J. Biol. Chem.* **281**, 19939-19948
28. Liu, B., Knirel, Y. A., Feng, L., Perepelov, A. V., Sof'ya, N. S., Reeves, P. R., and Wang, L. (2014) Structural diversity in *Salmonella* O-antigens and its genetic basis. *FEMS Microbiol. Rev.* **38**, 56-89
29. Collins, R. F., Kargas, V., Clarke, B. R., Siebert, C. A., Clare, D. K., Bond, P. J., Whitfield, C., and Ford, R. C. (2017) Full-length, Oligomeric Structure of Wzz Determined by Cryoelectron Microscopy Reveals Insights into Membrane-Bound States. *Structure*
30. Kong, Q., Yang, J., Liu, Q., Alamuri, P., Roland, K. L., and Curtiss, R. (2011) Effect of deletion of genes involved in lipopolysaccharide core and O-antigen synthesis on virulence and immunogenicity of *Salmonella enterica* serovar Typhimurium. *Infect. Immun.* **79**, 4227-4239

31. Davies, M. R., Broadbent, S. E., Harris, S. R., Thomson, N. R., and van der Woude, M. W. (2013) Horizontally acquired glycosyltransferase operons drive Salmonellae lipopolysaccharide diversity. *PLoS Genet.* **9**, e1003568
32. Bogomolnaya, L. M., Santiviago, C. A., Yang, H. J., Baumler, A. J., and Andrews-Polymenis, H. L. (2008) 'Form variation' of the O12 antigen is critical for persistence of *Salmonella Typhimurium* in the murine intestine. *Mol. Microbiol.* **70**, 1105-1119
33. Kintz, E., Davies, M. R., Hammarlöf, D. L., Canals, R., Hinton, J. C., and Woude, M. W. (2015) A BTP1 prophage gene present in invasive non-typhoidal *Salmonella* determines composition and length of the O-antigen of the lipopolysaccharide. *Mol. Microbiol.* **96**, 263-275
34. Lerouge, I., and Vanderleyden, J. (2002) O-antigen structural variation: mechanisms and possible roles in animal/plant–microbe interactions. *FEMS Microbiol. Rev.* **26**, 17-47
35. Davis, E. O., Evans, I. J., and Johnston, A. W. (1988) Identification of nodX, a gene that allows *Rhizobium leguminosarum* biovar viciae strain TOM to nodulate Afghanistan peas. *Mol. Gen. Genet.* **212**, 531-535
36. Thanweer, F., Tahiliani, V., Korres, H., and Verma, N. K. (2008) Topology and identification of critical residues of the O-acetyltransferase of serotype-converting bacteriophage, SF6, of *Shigella flexneri*. *Biochem. Biophys. Res. Commun.* **375**, 581-585
37. Laaberki, M.-H., Pfeffer, J., Clarke, A. J., and Dworkin, J. (2011) O-Acetylation of peptidoglycan is required for proper cell separation and S-layer anchoring in *Bacillus anthracis*. *J. Biol. Chem.* **286**, 5278-5288
38. Slauch, J. M., Lee, A. A., Mahan, M. J., and Mekalanos, J. J. (1996) Molecular characterization of the oafA locus responsible for acetylation of *Salmonella typhimurium* O-antigen: oafA is a member of a family of integral membrane trans-acylases. *J. Bacteriol.* **178**, 5904-5909
39. Buendia, A., Enenkel, B., Köplin, R., Niehaus, K., Arnold, W., and Pünier, A. (1991) The *Rhizobium meliloti* exoZl exoB fragment of megaplasmid 2: ExoB functions as a UDP-glucose 4-epimerase and ExoZ shows homology to NodX of *Rhizobium leguminosarum* biovar viciae strain TOM. *Mol. Microbiol.* **5**, 1519-1530
40. Clark, C. A., Beltrame, J., and Manning, P. A. (1991) The oac gene encoding a lipopolysaccharide O-antigen acetylase maps adjacent to the integrase-encoding gene on the genome of *Shigella flexneri* bacteriophage Sf6. *Gene* **107**, 43-52
41. Kajimura, J., Rahman, A., Hsu, J., Evans, M. R., Gardner, K. H., and Rick, P. D. (2006) O-acetylation of the enterobacterial common antigen polysaccharide is catalyzed by the product of the yiaH gene of *Escherichia coli* K-12. *J. Bacteriol.* **188**, 7542-7550
42. Berry, D. S., Lynn, F., Lee, C.-H., Frasch, C. E., and Bash, M. C. (2002) Effect of O acetylation of *Neisseria meningitidis* serogroup A capsular polysaccharide on development of functional immune responses. *Infect. Immun.* **70**, 3707-3713
43. García, B., González-Sabín, J., Menéndez, N., Braña, A. F., Núñez, L. E., Morís, F., Salas, J. A., and Méndez, C. (2011) The chromomycin CmmA acetyltransferase: a membrane-bound enzyme as a tool for increasing structural diversity of the antitumour mithramycin. *Microb. Biotechnol.* **4**, 226-238

44. Hara, O., and Hutchinson, C. (1992) A macrolide 3-O-acyltransferase gene from the midecamycin-producing species *Streptomyces mycarofaciens*. *J. Bacteriol.* **174**, 5141-5144
45. Thanweer, F., and Verma, N. K. (2012) Identification of critical residues of the serotype modifying O-acetyltransferase of *Shigella flexneri*. *BMC Biochem.* **13**, 13
46. Williams, A. H., Veyrier, F. J., Bonis, M., Michaud, Y., Raynal, B., Taha, M.-K., White, S. W., Haouz, A., and Boneca, I. G. (2014) Visualization of a substrate-induced productive conformation of the catalytic triad of the *Neisseria meningitidis* peptidoglycan O-acetylesterase reveals mechanistic conservation in SGNH esterase family members. *Acta Crystallogr. D Biol. Crystallogr.* **70**, 2631-2639
47. Lo, Y.-C., Lin, S.-C., Shaw, J.-F., and Liaw, Y.-C. (2003) Crystal structure of *Escherichia coli* Thioesterase I/Protease I/Lysophospholipase L 1: Consensus sequence blocks constitute the catalytic center of SGNH-hydrolases through a conserved hydrogen bond network. *J. Mol. Biol.* **330**, 539-551
48. Leščić Ašler, I., Ivić, N., Kovačić, F., Schell, S., Knorr, J., Krauss, U., Wilhelm, S., Kojić-Prodić, B., and Jaeger, K. E. (2010) Probing enzyme promiscuity of SGNH hydrolases. *ChemBioChem* **11**, 2158-2167
49. Baker, P., Ricer, T., Moynihan, P. J., Kitova, E. N., Walvoort, M. T., Little, D. J., Whitney, J. C., Dawson, K., Weadge, J. T., and Robinson, H. (2014) *P. aeruginosa* SGNH hydrolase-like proteins AlgJ and AlgX have similar topology but separate and distinct roles in alginate acetylation. *PLoS Pathog.* **10**, e1004334
50. Moynihan, P. J., and Clarke, A. J. (2011) O-Acetylated peptidoglycan: controlling the activity of bacterial autolysins and lytic enzymes of innate immune systems. *Int. J. Biochem. Cell Biol.* **43**, 1655-1659
51. Pfeffer, J. M., Weadge, J. T., and Clarke, A. J. (2013) Mechanism of action of *Neisseria gonorrhoeae* O-acetylpeptidoglycan esterase, an SGNH serine esterase. *J. Biol. Chem.* **288**, 2605-2613
52. Huang, Y.-T., Liaw, Y.-C., Gorbatyuk, V. Y., and Huang, T.-H. (2001) Backbone dynamics of *Escherichia coli* thioesterase/protease I: evidence of a flexible active-site environment for a serine protease. *J. Mol. Biol.* **307**, 1075-1090
53. Mølgaard, A., Kauppinen, S., and Larsen, S. (2000) Rhamnogalacturonan acetylesterase elucidates the structure and function of a new family of hydrolases. *Structure* **8**, 373-383
54. Tyukhtenko, S. I., Litvinchuk, A. V., Chang, C.-F., Lo, Y.-C., Lee, S.-J., Shaw, J.-F., Liaw, Y.-C., and Huang, T.-H. (2003) Sequential structural changes of *Escherichia coli* thioesterase/protease I in the serial formation of Michaelis and tetrahedral complexes with diethyl p-nitrophenyl phosphate. *Biochemistry* **42**, 8289-8297
55. Chang, R.-C., Chen, J. C., and Shaw, J.-F. (1996) Site-Directed Mutagenesis of a Novel Serine Arylesterase from *Vibrio mimicus* Identifies Residues Essential for Catalysis. *Biochem. Biophys. Res. Commun.* **221**, 477-483
56. Moynihan, P. J., and Clarke, A. J. (2014) Mechanism of action of peptidoglycan O-acetyltransferase B involves a Ser-His-Asp catalytic triad. *Biochemistry* **53**, 6243-6251

57. Weadge, J. T., and Clarke, A. J. (2007) *Neisseria gonorrheae* O-acetylpeptidoglycan esterase, a serine esterase with a Ser-His-Asp catalytic triad. *Biochemistry* **46**, 4932-4941
58. Bera, A., Herbert, S., Jakob, A., Vollmer, W., and Götz, F. (2005) Why are pathogenic staphylococci so lysozyme resistant? The peptidoglycan O-acetyltransferase OatA is the major determinant for lysozyme resistance of *Staphylococcus aureus*. *Mol. Microbiol.* **55**, 778-787
59. Bernard, E., Rolain, T., Courtin, P., Guillot, A., Langella, P., Hols, P., and Chapot-Chartier, M.-P. (2011) Characterization of O-acetylation of N-acetylglucosamine a novel structural variation of bacterial peptidoglycan. *J. Biol. Chem.* **286**, 23950-23958
60. Broadbent, S., Davies, M., and Van Der Woude, M. (2010) Phase variation controls expression of Salmonella lipopolysaccharide modification genes by a DNA methylation-dependent mechanism. *Mol. Microbiol.* **77**, 337-353
61. Moynihan, P. J., and Clarke, A. J. (2010) O-acetylation of peptidoglycan in gram-negative bacteria identification and characterization of peptidoglycan O-acetyltransferase in *Neisseria gonorrhoeae*. *J. Biol. Chem.* **285**, 13264-13273
62. Kintz, E., Heiss, C., Black, I., Donohue, N., Brown, N., Davies, M. R., Azadi, P., Baker, S., Kaye, P. M., and van der Woude, M. (2017) *Salmonella enterica* serovar Typhi lipopolysaccharide O-antigen modification impact on serum resistance and antibody recognition. *Infect. Immun.* **85**, e01021-01016
63. Schlegel, S., Rujas, E., Ytterberg, A. J., Zubarev, R. A., Luirink, J., and De Gier, J.-W. (2013) Optimizing heterologous protein production in the periplasm of *E. coli* by regulating gene expression levels. *Microb. Cell. Fact* **12**, 1
64. Yoon, S. H., Kim, S. K., and Kim, J. F. (2010) Secretory production of recombinant proteins in *Escherichia coli*. *Recent Pat. Biotechnol.* **4**, 23-29
65. Prinz, W. A., Åslund, F., Holmgren, A., and Beckwith, J. (1997) The Role of the Thioredoxin and Glutaredoxin Pathways in Reducing Protein Disulfide Bonds in the *Escherichia coli* Cytoplasm. *J. Biol. Chem.* **272**, 15661-15667
66. Bessette, P. H., Åslund, F., Beckwith, J., and Georgiou, G. (1999) Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. *Proc. Natl. Acad. Sci* **96**, 13703-13708
67. Rosano, G. L., and Ceccarelli, E. A. (2014) Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.* **5**, 172
68. Steiner, D., Forrer, P., Stumpp, M. T., and Plückthun, A. (2006) Signal sequences directing cotranslational translocation expand the range of proteins amenable to phage display. *Nat. Biotechnol.* **24**, 823-831
69. De Mey, M., Lequeux, G. J., Maertens, J., De Muynck, C. I., Soetaert, W. K., and Vandamme, E. J. (2008) Comparison of protein quantification and extraction methods suitable for *E. coli* cultures. *Biologicals* **36**, 198-202
70. Takagi, H., Morinaga, Y., Tsuchiya, M., Ikemura, H., and Inouyea, M. (1988) Control of folding of proteins secreted by a high expression secretion vector, pIN-III-ompA: 16-fold increase in production of active subtilisin E in *Escherichia coli*. *Nat. Biotechnol.* **6**, 948-950

71. Chalmers, J., Kim, E., Telford, J., Wong, E., Tacon, W., Shuler, M., and Wilson, D. (1990) Effects of temperature on *Escherichia coli* overproducing beta-lactamase or human epidermal growth factor. *Appl. Environ. Microbiol.* **56**, 104-111
72. Kapust, R. B., and Waugh, D. S. (1999) *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* **8**, 1668-1674
73. Fox, J. D., Kapust, R. B., and Waugh, D. S. (2001) Single amino acid substitutions on the surface of *Escherichia coli* maltose-binding protein can have a profound impact on the solubility of fusion proteins. *Protein Sci.* **10**, 622-630
74. Böhm, G. M., Rudolf; Jaenicke, Rainer. (1992) Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Engineering, Design and Selection* **5**, 191-195
75. Evans, P. R., and Murshudov, G. N. (2013) How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1204-1214
76. Collaborative, C. P. (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **50**, 760
77. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658-674
78. Yao, J., Woolfson, M., Wilson, K., and Dodson, E. (2005) A modified ACORN to solve protein structures at resolutions of 1.7 Å or better. *Acta Crystallogr. D Biol. Crystallogr.* **61**, 1465-1475
79. Perrakis, A., Morris, R., and Lamzin, V. S. (1999) Automated protein model building combined with iterative structure refinement. *Nat. Struct. Mol. Biol.* **6**, 458-463
80. Emsley, P., and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126-2132
81. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* **53**, 240-255
82. Seiler, M., Mehrle, A., Poustka, A., and Wiemann, S. (2006) The 3of5 web application for complex and comprehensive pattern matching in protein sequences. *BMC Bioinformatics* **7**, 1
83. Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., and Rupp, R. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**, 460
84. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845-858
85. Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T. G., Bertoni, M., and Bordoli, L. (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42**, W252-W258
86. Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., and Schwede, T. (2009) Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc.* **4**, 1
87. Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**, 195-201

88. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242
89. Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., Xia, B., Beglov, D., and Vajda, S. (2015) The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat. Protoc.* **10**, 733-755
90. Lo, Y.-C., Lin, S.-C., Shaw, J.-F., and Liaw, Y.-C. (2005) Substrate specificities of *Escherichia coli* thioesterase I/protease I/lysophospholipase L1 are governed by its switch loop movement. *Biochemistry* **44**, 1971-1979
91. Fass, D. (2012) Disulfide bonding in protein biophysics. *Annu. Rev. Biophys.* **41**, 63-79
92. Thangudu, R. R., Manoharan, M., Srinivasan, N., Cadet, F., Sowdhamini, R., and Offmann, B. (2008) Analysis on conservation of disulphide bonds and their structural features in homologous protein domain families. *BMC Struct. Biol.* **8**, 55
93. Cacciapuoti, G., Fuccio, F., Petraccone, L., Del Vecchio, P., and Porcelli, M. (2012) Role of disulfide bonds in conformational stability and folding of 5′-deoxy-5′-methylthioadenosine phosphorylase II from the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1824**, 1136-1143
94. Thornton, J. (1981) Disulphide bridges in globular proteins. *J. Mol. Biol.* **151**, 261-287
95. Akoh, C. C., Lee, G.-C., Liaw, Y.-C., Huang, T.-H., and Shaw, J.-F. (2004) GDSL family of serine esterases/lipases. *Prog. Lipid Res.* **43**, 534-552
96. Vyas, N. K. (1991) Atomic features of protein-carbohydrate interactions. *Curr. Opin. Struct. Biol.* **1**, 732-740