Crystal structures of GfcC, a group 4 capsule operon protein from *Escherichia coli,* and YraM, an outer membrane lipoprotein from *Haemophilus influenzae*

by

Karthik Sathiyamoorthy

**A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biophysics)
in The University of Michigan
2011**

Doctoral Committee:

**Associate Professor Mark A. Saper, Chair
Professor Victor DiRita
Professor Ari Gafni
Associate Professor Raymond C. Trievel
Associate Professor Zhaohui Xu
Assistant Professor Oleg V. Tsodikov**

To my loving Mom and Sister

Acknowledgements

Where to begin? I should say early on I had to be blessed to have a wonderful school education under trying family circumstances. Thanks foremost to my Mom for her love and affection and for the support by being stubborn in her decision to keep me at the same school (DAV) that taught me values that I will uphold for my entire life. Thanks to my late father who was caring and helped me immensely in shaping my early education and travelling to different places as a child. Those are still fresh and loving memories that I will cherish. Thanks to my sister who was an excellent guide helping me morally and financially when it was most needed. It still burdens me that she had to forego her daytime college education to work and see me complete my undergraduation. Thanks to my friends who maintained a competitive atmosphere in school and later in college that motivated me to excel in all my subjects and research projects. I also learnt a great deal from them, science or otherwise. Thanks to the teachers and professors in school and college who helped me immensely by instilling their knowledge in me and their patience and teachings shape me to be the person that I am now. I particularly want to thank my professor in undergrad, Pennathur Gautam for inspiring me early on during my stay in Anna University.

I am especially indebted to my advisor Mark A. Saper, for being like a second father to me. His patience and optimism combined are second only to no other person I have ever seen so far. He is a perfectionist and I am fortunate to take on his footsteps and look forward to my future. Thanks to my committee members– Victor DiRita, Ari Gafni, Raymond C. Trievel, Zhaohui Xu, and Oleg V. Tsodikov who are leading experts in such diverse and interesting fields such as crystallography, microbiology and biophysics to name a few and for providing me with honest and valuable feedback on my project.

Thanks to our close collaborator Ilan Rosenshine and his lab who provided us with the *gfcC:kan* mutant construct. Thanks to his lab member Erez Mills in particular for the constructs that were used for localization and buoyancy studies to detect protein and capsule respectively. Thanks to Titus Franzmann for help in running and analyzing analytical ultracentrifugation data to detect the oligomeric state of GfcC.

Thanks to Vijayalakshmi Janakiraman, a former research scientist in Saper lab who helped me initially in transitioning to the lab and showed me around Ann Arbor. She was also instrumental for substantial progress in the YraM project solving the structures of N-domain and C-domain of YraM independently and making the constructs for full length YraM protein expression. Thanks to Austin Johnson who was kind enough to show me the works of the lab and for helping me get started with the protein purification (FPLC) system. Thanks to Bhramara Tirupati, a former post-doc in Saper lab who constructed the full-length YraM protein without the tag that was crystallized and for working out the details of purification for this protein. Thanks to Brian Akerley for initial collaboration in the YraM project.

I also want to extend my thanks to the excellent synchrotron facilities at APS, Argonne, specifically the LS-CAT beamlines in sector 21-ID (beamlines 21 ID-D and 21-ID-G). Diffraction data for GfcC and YraM was collected from 21-ID-D and 21-ID-G, SeMet data was collected from 21 ID-D beamline. Thanks to the staff and scientists at the beamline for timely help with any issues that arose during data collection.

Thanks to other crystallography labs in life sciences and medical school at U-M (Trievel lab, Smith lab, Tesmer lab, Stuckey lab, Xu lab, Young lab) who have been very kind and helpful to borrowing chemicals and troubleshooting experimental procedures. Thanks to the extensive crystallographic community at U-M who have been motivating me through the Friday morning UMXtal talks in the LSI library. They truly are a delight to the mind and I hope the tradition continues for a long long time. Thanks to the Biophysics department for providing me with a home for the past five years and special thanks to the “coffee room” in fourth floor chemistry building; without it getting through long protein preps would not have been easy.

Thanks to my professor and friend, Robert Zand, who cheered me up and always had a sense of humor. I am going to miss him a lot.

Thanks to my fellow graduate students, many of them will be my lifelong friends that I was fortunate to meet during my grad school here at U-M. Beginning the graduate school along with Akash Bhattacharya, Bharath S. Mamathambika and Saumen Chakraborty with whom I have many memories tackling endless homeworks in my initial two years. For giving good advise about courses and preliminary exams and for being such a wonderful friend I thank Chamaree de Silva. The many road trips we all made are unforgettable. Thanks to Sethuramasundaram Pitchiaya, Krishnan Raghunathan, Pavithra Aravamudhan, Vishalakshi Krishnan, and Preethi Sankaranarayanan, who joined me from the same place I did my undergraduation right through my entire grad school here in Ann Arbor. Thank you for coming along, it has been a long journey and you all made me feel at home. Thanks to Kohinoor Dasgupta and Samopriyo Maitra, I will forever cherish your friendship and admire your kind and helpful nature. Thanks to Krishna Chaitanya Suddala, Jeetender Chugh and Surma Talapatra for being good friends. We had such a memorable time together during the last few years. Thanks to Catleya Rojviriya for being a sweet friend, cooking those excellent dinners for us and for being a great colleague in the lab during her short visit.

Thanks to my undergraduate lab members who gave me good company in the lab and also taught me a great deal. Thanks to Katie Kurnit, Lisa Glass, Francis J Prael, Mrugesh Patel, Mykola Murskyj, and Louai Mokhtarm. I wish you all the good luck in whatever you decide to do in future.

Table of Contents

List of Figures

List of tables

Abstract

This thesis presents the high resolution structures of two proteins, the periplasmic GfcC from enteropathogenic *Escherichia coli* and the outer membrane lipoprotein YraM from *Haemophilus influenzae*. Both proteins have homologs in other Gram-negative bacteria. The specific functions of these two proteins are still unclear but the structures will guide further experiments to establish their critical role in the physiology of the microorganism.

Many pathogenic bacteria secrete polysaccharides that often act as a first line of defense against the host. These can be tightly associated with the outer surface of the cell and are termed capsular polysaccharide or secreted to the environment as exopolysaccharide. The group 4 capsular polysaccharide is comprised of oligosaccharide repeat units identical to the specific O-antigen present in the lipopolysaccharide (LPS). Each of the seven genes of the gfc operon in enteropathogenic *E. coli* are important for the export and assembly of this group 4 capsular polysaccharide. Three of the genes, *gfcE, etp* and *etk*, are homologous to the group 1 polysaccharide assembly genes *wza, wzb* and *wzc,* while the *gfcABCD* genes are unique to the group 4 operon but of unknown function. Osmotic shock confirmed that GfcC (~26 kDa) is a soluble, periplasmic protein but whether it is also associated with either membrane cannot be ruled out. The 1.9-Å resolution crystal structure of GfcC reveals two β-grasp domains (D2 and D3) similar to those in Wza, the proposed export channel protein for group 1 polysaccharide export. GfcC contains a C-terminal amphipathic helix, but unlike Wza, it folds to pack onto the D3 β-sheet. This helix is not long enough to span the outer membrane as it does in Wza. Inserted in D2 is a helical hairpin that fixes the location of the C-terminal helix and D3. A mostly conserved pocket, near the interface of the two β-grasp domains, contains an invariant Arg115 that hydrogen bonds to all domains of the protein and appears important for protein stability. This pocket is about the size of a sugar molecule and has available hydrogen bond donor and acceptor groups suitable for interacting with a ligand. GfcC is a monomer in solution but forms an identical noncrystallographic dimer in two different crystal forms. The partially conserved interface is comprised of mostly polar interactions between D2 and D3 of each monomer. Interestingly, in several species, the *gfcC* homolog is fused to the adjacent gene *gfcD*, predicted to be a large β-barrel protein with ~24 strands. This fact, together with published results of the alginate exopolysaccharide apparatus, which also contains a large β-barrel, suggest that GfcC may interact with GfcD.

YraM is an essential protein for the growth and viability of *H. influenzae.* Homologs of *yraM* are found in many Gram-negative genomes but appear to be not essential in organisms with larger genomes. Sequence analysis of the 575-residue outer membrane lipoprotein suggested two individual domains whose crystal structures were determined independently. The C-domain (residues 257–575) has a periplasmic binding protein fold with highly conserved residues in the expected ligand binding cleft. The N-domain (residues 33–249) has tetratricopeptide repeat-like motifs found in other proteins known to mediate protein-protein interactions. This thesis reports the full-length structure of YraM (at 1.97 Å) depicting the linker region and the relative orientation of the two domains. The N-domain and C-domain in the full-length structure are in identical conformations to the individually determined structures. The two domains are arranged like a jaw with the N-domain as upper jaw and the C-domain as the lower jaw with the linker (residues 250–256) serving as the hinge. Low frequency modes by normal mode analysis (NMA) attest to this definition of the domain's orientation. The approximate distance between the N-domain lipid anchor and the binding cleft of the C-domain (~50 Å) that is left unhindered coincide with estimates of distance of peptidoglycan from the outer membrane. Slt70, a periplasmic lytic transglycosylase is similar to YraM in its structure with an extended TPR domain followed by lysozyme like domain that has been shown capable of binding to a peptidoglycan precursor, 1,6-anhydromuropeptide. Further, the essential lipoprotein *yraP*, encoded downstream of *yraM,* was proposed to be important for outer membrane integrity in *E.coli*. YraM is also found only in Gram-negative organisms and this reinforces the idea that it is important for processes that

distinguish the cell wall structures between Gram-negative (proteobacteria) and Gram-positive (firmicutes) organisms. Identification of the ligand that binds to the C-domain will also shed more light on the protein's physiological role. The outer membrane location for YraM also makes it an attractive target for developing antimicrobials against the non-encapsulated opportunistic pathogen *H. influenzae*.

Chapter 1

Polysaccharide Capsule Assembly system is thematic: diverse sugar chemistry and branched substrates have few organized assembly machines

1.1 Introduction

1.1.1 Nature and types of Polysaccharide Capsules

Many bacteria including *E.coli* are covered with a layer of surface associated polysaccharides that can be tightly associated over the entire cell surface termed capsular polysaccharide (CPS), or be loosely associated or secreted and are referred to as exopolysaccharide (EPS) or slime polysaccharide [1-3] (Figure 1-1). Polysaccharides are polymers of oligosaccharide repeat units joined to each other through the hydroxyl groups (glycosidic linkages). They are usually of very high molecular weight (Mr>100,000). More often different monosaccharides (usually 3–5) specific for the particular type (i.e., serotype) form a basic building block that is repeated many times to form the polysaccharide chain. Diversity in these polysaccharides apart from the identity of the monosaccharides also comes from constituent branched or unbranched oligosaccharide repeat units and the incorporation of both organic and inorganic molecules in certain cases [2]. There are more than 80 different capsular serotypes identified in *E.coli* and are grouped into four distinct groups (groups 1–4) based on genetic and biochemical criteria in *E.coli*. Some strains have the group 1 (composed of K antigen) type capsule. Some other strains especially the pathogenic *E.coli* bacteria include a group 4 capsule whose repeat units are identical to the O-antigen repeat units in lipopolysaccharide (LPS) but are not anchored to a lipid A moiety like in LPS [1, 4]. A common feature of these two (group 1 and group 4) capsules is that they rely on a specific 'Wzy' polymerase that is responsible for polymerization of the respective oligosaccharide repeats. The polymerization takes place in the periplasm and is coupled with export to the outer surface [3] (see below for more description). There are other types of capsules (group 2 and 3) with different oligosaccharide repeats that are polymerized in the cytoplasm and exported through both membranes directly to the outer surface. These are further mediated by ATP-binding cassette (ABC) transporters in the inner membrane [3] and thus account for an energy driven translocation process. Chris Whitfield and Ian S. Roberts introduced this group 1–4 nomenclature for classification of polysaccharide capsules in *E.coli* back in 1999 [1]. The classification was based on organization of capsule gene clusters, details of assembly pathway and regulatory features that dictate capsule expression. During the past decade, important strides have taken place in the structural exploration of the proteins involved in export and possible feedback regulation of the group 1 capsule and the identification of genes involved in O-antigen type (group 4) capsule assembly. These advances and the similarities and differences between both these Wzy-dependent polysaccharide capsules will be further explored in this chapter.

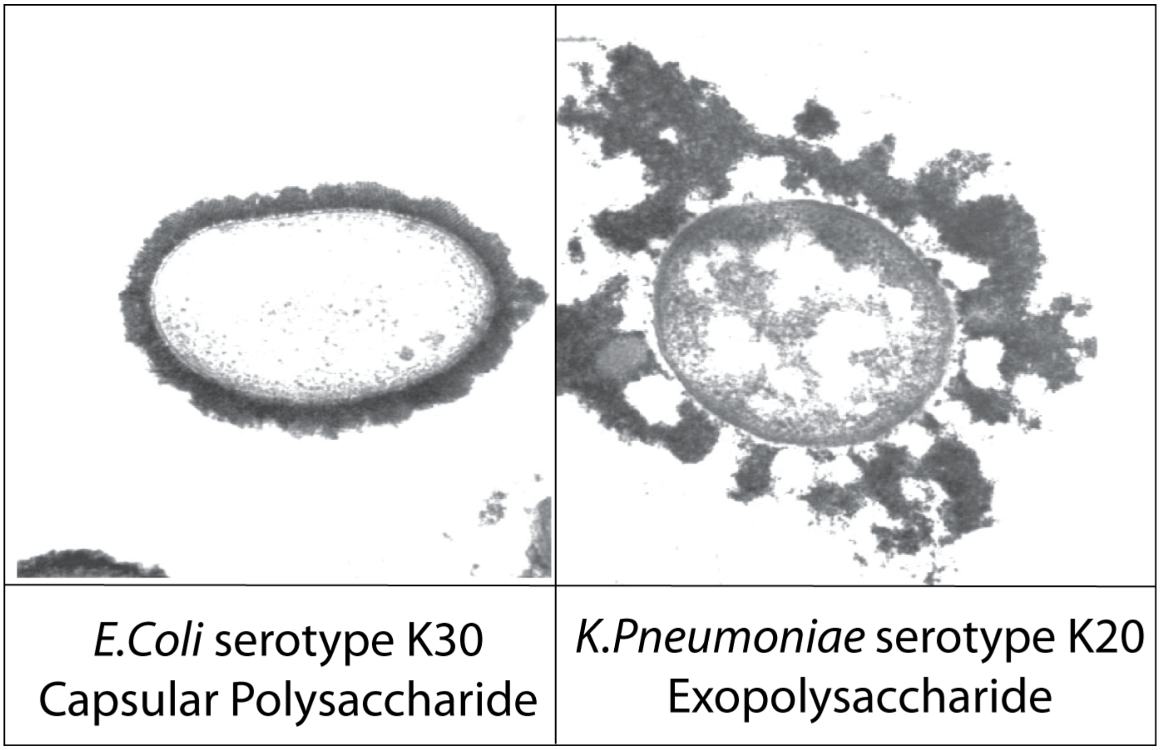


Figure 1-1 **Two different capsule types**. The left shows the capsular polysaccharide with its adherent thick stained capsule layer over the entire cell surface. The right shows loosely adhering exopolysaccharide forming a matrix that surrounds the cell and connected to the cell only sparsely at certain regions. The images are from reference [3]. Copyright 2009 American Society for Microbiology.

1.1.2 Function of polysaccharide capsules

Polysaccharide capsules are responsible for a whole range of functions including adhesion, transmission, resistances to innate host defense mechanisms and adaptive immune response, and intracellular survival [5].

In certain cases the chemistry of the polysaccharide plays a functional role, for instance, the adhesion of group A *Streptococci* to pharyngeal cells is mediated by the interaction between the hyaluronic acid capsule and CD44, the hyaluronic acid capsule binding protein [5, 6]. In certain kinds of capsule like those that have sialic acid (NeuNAc), factor H can bind to the capsule and cause inhibition of the alternative complement activation cascade thus explaining the resistance [7]. Eighty percent of blood isolates of *E.coli* K1 are resistant to opsonization and phagocytosis by normal human leukocytes [8]. Capsules also have the ability to resist being engulfed (phagocytosed) by macrophages and this property is attributed to the negatively charged capsules being repelled by the negatively charged membrane of macrophages like in case of *Pneumococcal* capsules [5, 9].

Polysaccharide capsules also protect cells from dessication by forming an impermeable outer barrier. In exopolysaccharide capsules like the colanic acid from many *E.coli* [10] or alginate from *Pseudomonas aeroginosa* [11, 12] they form hydrated polymeric matrices called biofilms that greatly improve the odds of surviving in adverse environmental conditions or to maintain persistent lung infection respectively.

1.1.3 Mechanism of group 1 (and group 4) capsule biosynthesis

Group 1 and 4 capsule biosynthesis proceed more similarly to each other with the oligosaccharide repeat units polymerized in the periplasm on the reducing end by the Wzy polymerase. This polymerization step differentiates these from the group 2 and 3 capsules. Wzy is a 12-stranded integral inner membrane protein whose glycosyl transferase activity was recently demonstrated directly [13]. Although group 1 and group 4 are thought to be identical with the involvement of Wzy, there are other proteins in both groups that do not have relatable homologs.

First, the oligosaccharide repeat units used as building blocks by the polymerase are synthesized in the cytoplasm by enzymes encoded by an operon; in O antigen polysaccharide synthesis these genes are localized to the *rfb* locus [14]. In the case of group 1 capsule the operon that synthesize the repeat units and the genes that are held responsible for assembly of the polysaccharide (*wza, wzb, wzc, wzi*) are in the same operon. These oligosaccharide building blocks are species and serotype specific. Group 4 repeat units are shared between O-antigen LPS synthesis and in high molecular weight (Mr>100,000) group 4 capsules. The genes for assembly in the group 4 capsule operon are thus distinct from the genes responsible for synthesizing the precursor moiety. In both groups 1 and 4, the respective oligosaccharide repeat unit is linked to Und-P (undecaprenyl phosphate), a C_{55} polyisoprenoid lipid derivative in the cytoplasm. Each repeat unit with the Und-P derivative is then transferred to the other side of the inner membrane by an integral inner membrane protein, Wzx (flippase).

Wzy–the polysaccharide polymerase, then polymerizes these repeat units sequentially that are now in the periplasm. The polymer is then threaded from the periplasm to the exterior through the conserved outer membrane export (OPX family) lipoprotein, Wza (homolog to GfcE in group 4 capsule) as understood currently [15]. Wza mutants (gene knockouts) do not accumulate polysaccharide in the periplasm thus supporting the notion that polymerization and translocation are coupled.

The following suggests one way by which outer membrane lipoprotein Wza could regulate the polymerization process itself with the polymerase known to reside in the inner membrane. Wza interacts (see below for structure of Wza) with the integral inner membrane Wzc (a polysaccharide co-polymerase) mediated possibly through their respective α-helical periplasmic regions. There is a cryonegatively stained electron micrograph (EM) structure of Wza-Wzc complex that reveals their direct interaction [16] [Figure 1-5]. Wza may also interact with Wzy directly but such an interaction has not been proved yet. The interaction of Wza and Wzc does however provide a scaffold that links inner membrane and outer membrane proteins. One acylation mutant of Wza where the signal sequence was replaced with one from OmpF is still localized in the outer membrane and does retain the interaction between Wza and Wzc but is incompetent in export; the polymer instead accumulates in the periplasm [17]. The details on how Wzy and Wzx interact with the Wza-Wzc complex and how the feedback regulation is controlled are still not understood.

Wzc has a cytoplasmic tyrosine kinase domain with Walker A and B motifs and a tyrosine rich C-terminal tail (7 of last 17 residues). At any time in the cell a subset of these tyrosines are phosphorylated. Wzb is the cytoplasmic cognate phosphatase responsible for dephosphorylating Wzc [18, 19]. Deletions of either *wzc* or *wzb* individually are deficient in capsule expression. Hence, the cycling of phosphorylation state of Wzc seems to be critical for polymerization of capsule and its export (hence the name co-polymerase for Wzc).

Another gene *wzi* immediately upstream of *wza* is the only member of the group 1 operon that does not affect polymerization of group 1 capsules. *wzi* knockout mutants however give a phenotype where the capsule is no longer anchored to the cell surface indicating a anchoring function for the protein encoded by *wzi* [20]. Recent strides in the crystallization of this protein (Wzi) may help address this function from a structural perspective.

1.1.4 Group 2 and 3 capsular polysaccharides

The genes for group 2 capsular assembly are organized into three principal regions (1-3). Regions 1 and 3 contain genes for export of capsular polysaccharide from the inner side of the cytoplasmic membrane–the site of synthesis– to the exterior cell surface and region 2 contain the biosynthetic genes necessary to form the capsule [2, 21]. The synthesis of group 2 is explained here by using *E.coli* K5 as the model system. The synthesis occurs in the inner membrane close to cytoplasm with four genes products KfiA–D coupled with the inner membrane export complex composed of ABC transporter KpsM and KpsT [21, 22]. KfiC and KfiA are involved in successive addition of GlcA and GlcNAc to the nonreducing terminus of the native polysaccharide. KfiB role is not known but it is presumed its role maybe structural, like other biosynthetic genes in this region *kfiB* knockout do not produce the K5 capsule. KfiD is an enzyme responsible for dehydrogenation of UDP-glucose to form UDP-GlcA, a substrate for K5 polysaccharide[21]. Neither lipid co-factors nor lipid linked intermediates are required for K5 polysaccharide (group 2) unlike the group 1 precursors transported by Wzx flippase. Two other proteins in the inner membrane KpsC and KpsS are involved in capping the finished capsular polysaccharide on the reducing terminus with phosphatidyl-Kdo (Kdo: Ketodeoxyoctanoate) before the export [21, 22]. The latter may have a role in anchoring the capsule to the cell surface. KpsE and KpsD finally export the group 2 polysaccharide across the outer membrane[21].

Group 3 polysaccharide genes are similar to group 2 and in many cases seem to have been generated by horizontal transfer of group 2 gene clusters with various insertion (IS) sequences[21]. The gene clusters are however significantly rearranged and as a result group 3 shares little sequence identity with group 2 genes. The best examples for group 3 polysaccharide capsules include K10 and K54 antigens in *E.coli* [21]. Though it is seen they share homology to group 2 polysaccharide assembly genes very little is known directly about this capsule system.

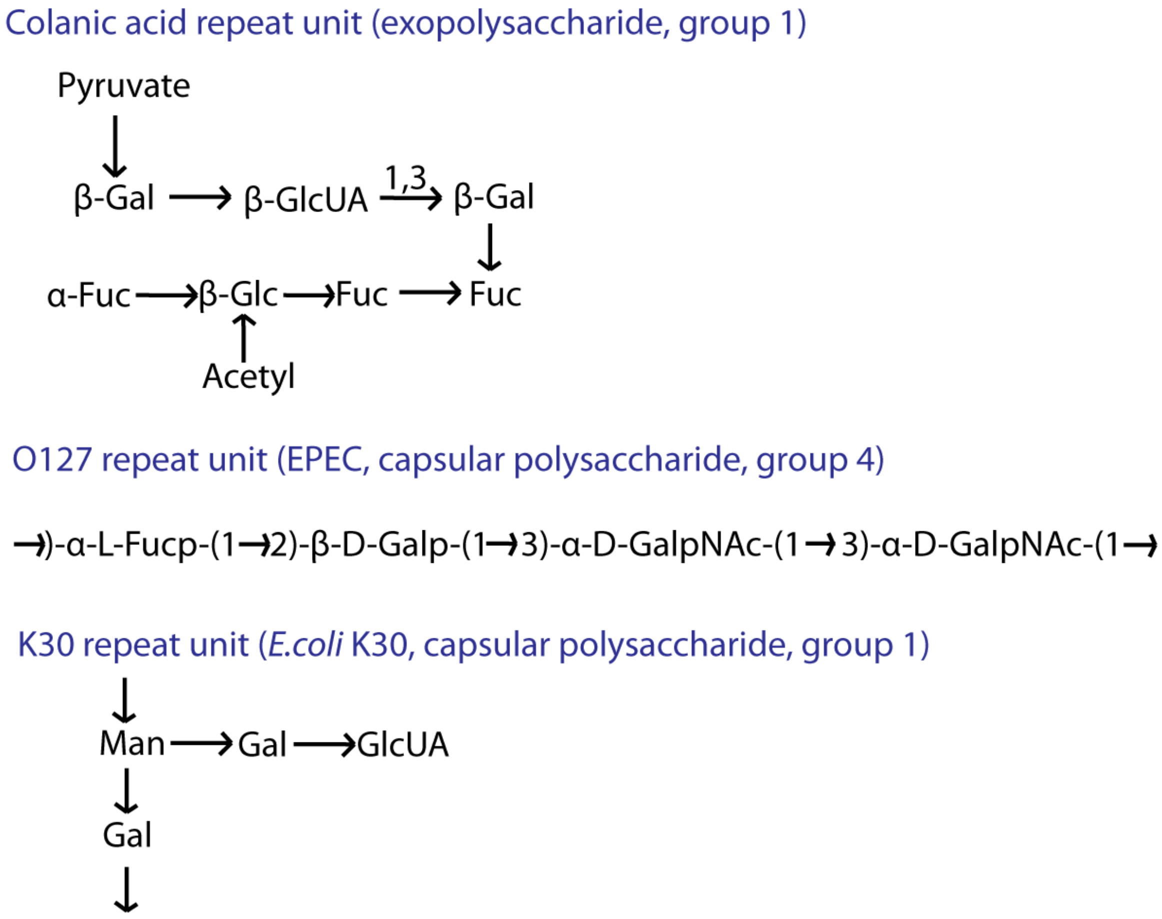

Colanic acid repeat unit (exopolysaccharide, group 1)
Pyruvate
β-Gal
β-GlcUA
1,3
β-Gal
α-Fuc
β-Glc
Fuc
Fuc
Acetyl
O127 repeat unit (EPEC, capsular polysaccharide, group 4)
→)-α-L-Fucp-(1→2)-β-D-Galp-(1→3)-α-D-GalpNAc-(1→3)-α-D-GalpNAc-(1→
K30 repeat unit (*E.coli* K30, capsular polysaccharide, group 1)
Man
Gal
GlcUA
Gal

Figure 1-2 **Oligosaccharide repeat units** in strains expressing colanic acid, group 1 or group 4 high molecular weight polysaccharide capsules; Glc: Glucose, Gal: Galactose, GluUA: Glucuronic acid, Man: Mannose, Fuc: Fucose, GalNAc: N-Acetyl Galactosamine

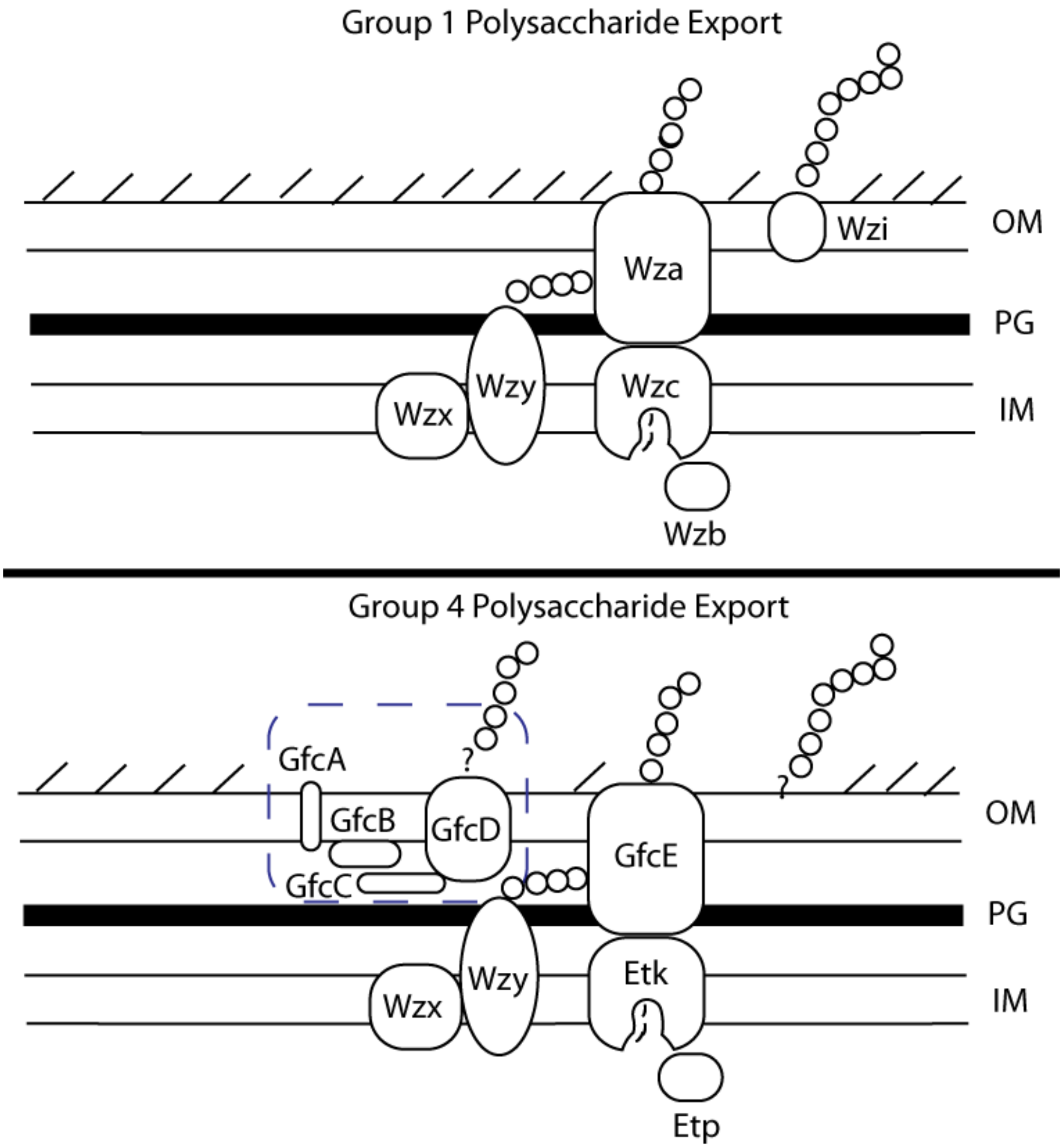


Figure 1-3 **Major players in the group 1 and group 4 polysaccharide assembly pathway with emphasis on GfcABCD and its unknown role**. Localization information has not been experimentally determined for some of the proteins above.

1.1.5 Structure of Group I polysaccharide export protein, Wza (OPX class)

The 2.26 Å crystal structure of mature acylated Wza forms an octamer (Figure 1-4), each monomer has four domains D1-D4 that in the octamer form rings R1-R4 [15]. D1 (residues 89-169) comprises the PES (Polysaccharide export sequence) that has been identified as such only by sequence conservation. D2 (residues 68-84 and 175-252) and D3 (residues 46-64 and 255-344) are similar β-grasp domains except in the octamer R2 has a smaller width (25 Å) compared to R3 (105 Å). D4 (residues 345-376) at the C-terminus of Wza is an amphipathic helix that in the octamer forms a novel α-helical pore. The structure is closed from the periplasm and is gated more likely through interaction with Wzc [15]. The putative pore formed by the octamer is ~17 Å in diameter and can thread the group 1 polysaccharide theoretically in the fully extended conformation. The residues lining the pore are not conserved and there seems to be no specific contacts possible with the growing polymer, the hydrophilic polysaccharide more likely makes water mediated contacts. This also explains why the same export apparatus can substitute in the assembly of other polysaccharide capsules with different repeat unit structures, for instance an essentially identical Wza is found in both *E.coli* K30 and *Klebsiella pneumoniae* [23, 24].

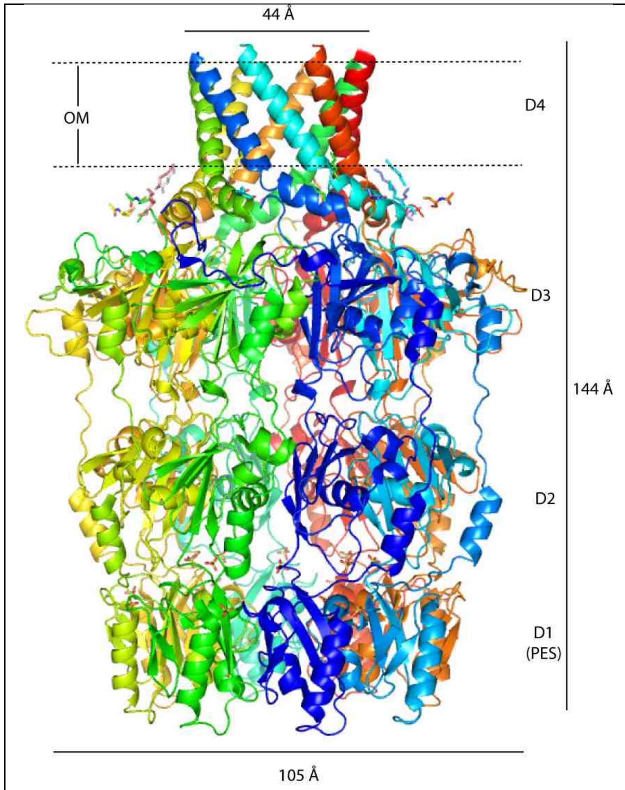


Figure 1-4 **Structure of Wza** shows the octamer with the C-terminal amphipathic region traversing the outer membrane. The structure is open from the top but closed from the periplasmic side coinciding with polysaccharide export sequence (PES) or domain 1 (D1). Gating of this pore possibly occurs through interaction with inner membrane protein, Wzc [15]. Wza PDB ID: 2J58

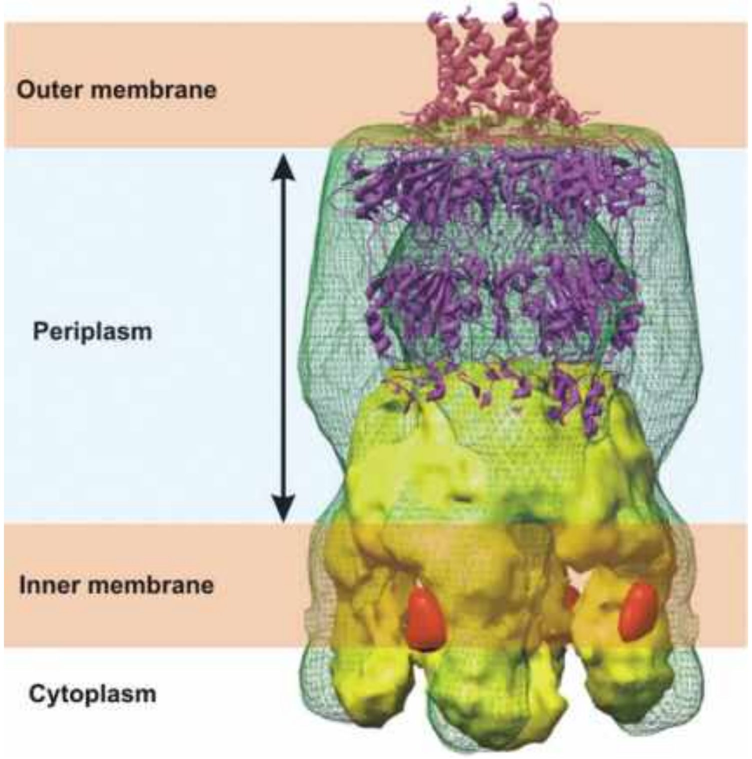


Figure 1-5 **Cryo-negative EM map of Multiprotein Wza-Wzc** complex. The complex is thought to be essential for polymerization and export of group 1 polysaccharide. The crystal structure of Wza is docked in the EM density in a cartoon representation [16]. Notice the lack of density corresponding to the α-helical pore region that has been attributed here to staining artifacts. Figure reprinted from reference [16] with permission from publisher. Copyright 2007 National Academy of Sciences, U.S.A.

1.1.6 Structure of Wzc, inner membrane kinase (PCP-2a)

There are two cryonegatively stained EM structures of the full-length Wzc, one as detergent (dodecyl maltoside, DDM) solubilized His$_6$-Wzc independently [25] at ~ 14 Å and the other in a complex with Wza at ~12 Å [16], both show this inner membrane tyrosine autokinase as a tetramer. The isolated structure of Wzc shows a tetramer with a C4 rotational symmetry and the distinct appearance of an extracted molar tooth. Viewed from the top the molecule has a diameter of ~100 Å, from the side it is ~110 Å high. The upper "crown" region forms a continuous density of ~55 Å thick and is followed by four distinctive "roots" about 65 Å high. These dimensions do not change much in the complex with Wza. Mutants lacking Wzc are unable to polymerize high molecular weight capsular polysaccharide but do not affect the production of shorter K-antigen oligosaccharides. Thus, Wzc is responsible for what is believed to be a coordinated biosynthesis and export process. There are additional pieces of evidence that again point to the tetrameric organization of Wzc irrespective of this protein's concentration [25]. First the protein eluted as a tetramer (~350-400 KDa, monomer ~82 KDa) during purification by size exclusion column. Secondly, the perfluorooctanoate (PFO)-PAGE analysis showed Wzc bands running at molecular weights corresponding to that of tetramer irrespective of concentration of Wzc [25]. PFO is a mild detergent that is known to preserve high affinity interactions in many membrane protein complexes. There were few high molecular weight aggregates but none of these aggregates were frequently observed in the raw images used for cryo-EM reconstruction. The nanogold Ni-NTA labeling established the N-terminus to be in the "roots" region of this structure.

Wzc structure consists of two distinct regions, the periplasmic region consists of predicted coiled-coil domains that mediate interactions with PES region of Wza. The exact nature of these interactions is still unresolved but it is known that these are key in gating of Wza channel as evidenced from the Wza-Wzc complex structure. The cytoplasmic or C-terminal domain of Wzc contains the tyrosine autokinase activity. This cytoplasmic autokinase domain is isolated to the "roots" region in the structure of full-length Wzc. The C-terminal 17 amino acids of Wzc contain seven tyrosines

(ASSYYRYGHNHYGYSYYDKK^{721} in Wzc$_{K30}$) which are phosphorylated to varying degrees [19]. There is also a cytoplasmic protein phosphatase Wzb that can dephosphorylate Wzc. Wzc knockouts or Wzc$^{1-704}$ mutant (lacking the C-terminal region) both deregulate high molecular weight capsule assembly (see section below for more details). In addition, Wzb mutants also deregulate high MW capsule assembly. Thus a cycling of phosphorylation and dephosphorylation seems essential for high MW capsule export. The details of how this phophorylation information is relayed to the biosynthetic machinery of Wzy/Wzx are still not clear. The kinase domain also can work in trans, this revelation came about when Wzc K540R that bears a Walker A mutation preventing ATP binding and/or hydrolysis when expressed *in-vivo* with Wzc$^{1-704}$ that lacks the C-terminus but has a functional active site is able to phosphorylate the former in its tyrosine rich tail [19]. An inactive mutant K540M of Wzc has been crystallized and its structure solved recently [26]. The protein is an octamer with the interface lined with positively charged residues (EX2RX2R motif) that is important for this oligomeric form and production of capsule. Also one of the tyrosine residues in the Y-cluster (Y-715) from one monomer is bound in the active site of the neighboring molecule indicating a possible means for autophosphorylation [26]. In contrast, the trans phosphorylation observation cannot be comprehended with the EM structure of full length Wzc where the independent Wzc kinase domains in the "roots" region do not interact. It is also not clear how the octameric Wza can interact with the presumed tetrameric Wzc, the key to this lies in seeing the periplasmic contacts between Wza and Wzc that require a high resolution structure than the one currently available.

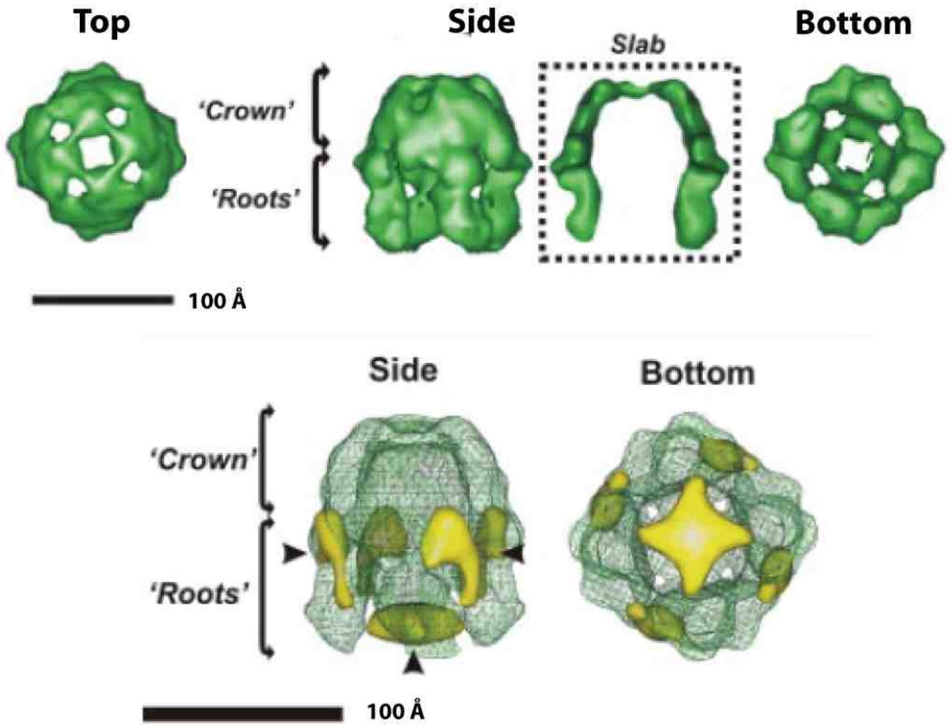


Figure 1-6 **EM structure of Wzc**. The structure has the overall appearance of an extracted molar tooth. This figure is suitably reorganized and reproduced from this publication [25] to show the EM structure of Wzc to 14 Å. Copyright 2006 The American Society for Biochemistry and Molecular Biology, Inc. U.S.A.

1.1.7 Group 4 Capsule (G4C) Operon

Group 4 polysaccharide capsule is similar to the group 1 in certain aspects that it is also a Wzy polymerase dependent capsule system. However, the building blocks of the capsule are different from group 1 and are identical to the O-antigen lipopolysaccharide building blocks (O-Ag capsule). The genes responsible for group 4 polysaccharide export in *E.coli* are localized in an operon (the *group 4 capsule* or *gfc* operon) [4]. The operon is comprised of seven genes, *gfcABCDE, etp* and *etk,* which were reported to be essential for polysaccharide expression on the outer surface of EPEC [4] and EHEC [27].

The *gfc* operon as identified by sequence analysis is also present in *E.coli* K12 but it is not expressed due to an insertion (*IS1*) in the promoter region of this operon [4]. The intact operon is also present in three strains of *Shigella* (*S. sonnei* 53G, *S. flexneri* strains 2a 301 and 2a 2475T, and *S.dysenteriae* M131649) apart from EPEC O127 E2348/69 and EHEC (EDL933 and Sakai) making a total of seven genomes with intact *gfc* operon [4].

Salmonella also produce an O-Ag capsule that is under the control of *yih* operon genes, the latter essential for translocation and assembly of the capsule [28]. *Vibrio cholerae* (serotype O139) produces the O-Ag capsule and requires *otnEFG* genes homologous to *gfcDCB* genes to function for its assembly [29]. *V.anguillarum* has two operons in opposite directions *orf1-wbfDCB* and *wzaABC* that are required for exopolysaccharide production and for its virulence and attachment to fish scales [30]. It is important to realize while homologs of *gfcABCD* may be reported to be present in some species, it does not correlate itself to the identification of group 4 capsule being present for instance, *V.anguillarum*. Also, the vice-versa i.e., group 4 capsule as defined by repeat unit similarity to O-antigen does not imply presence of *gfcABCD* genes or its homologs though it may be the case. The last three genes of the *gfc* operon *gfcE, etp* and *etk* are homologs of the group 1 counterpart *wza, wzb* and *wzc* respectively and presumed to function similarly. However, the roles of *gfcABCD* gene products in this operon are less established. Paralogs of *gfcABCD* that are *yjbEFGH* are found elsewhere in the genome and may be linked with group 1 but there have not been any studies that firmly establish the roles of these *yjbEFGH* genes. These latter genes are expressed at times of stress.

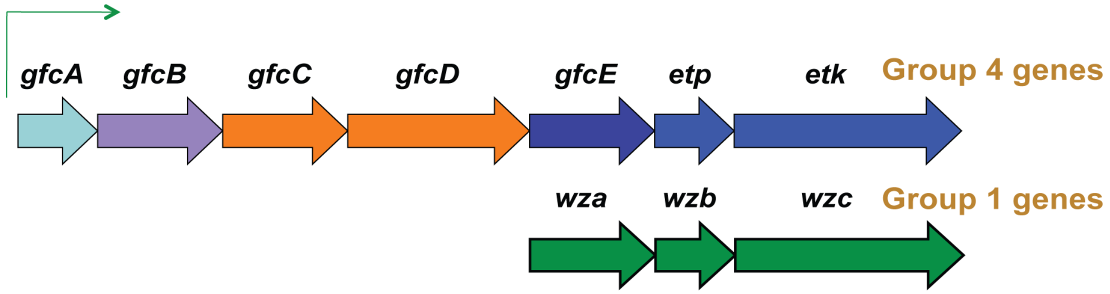


Figure 1-7 **Genes in Group 4 capsule (G4C) operon** (labeled on top) with their homologs (labeled at the bottom) from the group 1 system

Expressed proteins GfcA, GfcB, GfcC and GfcD all have the N-terminal signal sequences and should be exported to the membrane [4]. GfcB and GfcD are putative lipoproteins while GfcC is in the periplasm in a soluble form and may be associated with GfcD. Homolog in *Burkholderia sp.* has part of GfcC and GfcD coding sequences fused into one gene. GfcA is a small protein (107 residues) that has two putative transmembrane domains and is unusually rich in threonines (28/107 residues, ~one-fourth) [4]. The structure of (hypothetical) GfcB (YmcC) from *E.coli* K12 (though group 4 capsule is not expressed by this particular strain due to an *IS1* insertion in the promoter region of the operon) has been reported by the NorthEast Structural Genomics initiative (PDB code: 2IN5). The highly twisted β sheeted structure of GfcB (YmcC) though offers no insight to its probable function. It is possible that the complex structure with any of its interacting partners may shed more light to its role.

In-vivo the GFC capsule masks intimin and type III secretion system (TTSS) components required for the attachment of EHEC to intestinal cells. Further a positive regulator of TTSS i.e., Ler was also found to negatively regulate capsule related genes including *etp* and *etk* [27]. One way that this observation was rationalized [27] is that the capsule protects the cells from intimately attaching to the host (with TTSS) and triggering host defenses until the bacteria has sufficiently increased its population by colonizing the intestinal gut.

1.2 Exopolysaccharide export systems and role of β-barrel proteins in polysaccharide export

Alginate is a negatively charged exopolysaccharide (loosely adhered to the outer cell wall) that is a polymer of β-1,4-linked D-mannuronate and its C5-epimer, L-glucuronate. Alginate secretion by *Pseudomonas aeroginosa* is one of the well-studied exopolysaccharide systems and it is characterized by chronic biofilm formation in lungs of cystic fibrosis patients. Biofilms by definition are surface-attached bacteria encased in a hydrated polymeric matrix and are the cause of many persistent and chronic bacterial infections [31]. About 13 genes are implicated in the different stages of synthesis and secretion of alginate and with the exception of one member, *algC* all are located in the *algD* operon [32]. AlgK and AlgX are periplasmic proteins required for high molecular weight alginate secretion [33, 34] similar to role of Wzc or Etk (PCP-2a class) in group 1 or group 4 capsular polysaccharide or Wzz proteins (PCP-1 class) that maintain chain length of lipopolysaccharide antigens. The exact roles of AlgK and AlgX are however unclear. The operon also encodes AlgL, a lyase that degrades alginate polymers that accumulate in the periplasm [35]; in contrast there is no lyase identified in the group 1 or group 4 capsular polysaccharide assembly. AlgE is the outer membrane β-barrel protein that exports the alginate out of the periplasm [36]. The export mechanism ultimately involves proteins from both membranes assembling into a large heterooligomeric protein complex much like the group 1 (or group 4) capsular polysaccharide export is thought to function. Evidence for this fact comes from mutations that establish components from both outer and inner membrane as essential for alginate polymerization *in vitro* [37]. A similar export mechanism through β-barrel channels is found in other exopolysaccharide systems like cellulose, poly-β-1,6-GlcNAc, and Pel polysaccharide from *P.aeroginosa*.

The contrasting ways of this type of export where a β-barrel outer membrane protein is involved with the α-helical channel formed by Wza is intriguing. In contrast to periplasm spanning Wza-Wzc where Wza has a large periplasmic portion, AlgE fails to establish the precise linkage between itself in the outer membrane and inner membrane biosynthetic components (like Alg8/44). In this regard the structure of periplasmic AlgK recently solved by crystallographic methods revealing a tetratricopeptide peptide (TPR) domain with 9.5 helix-turn-helix repeats comes into focus. The extended TPR region gives the protein a superhelical twist with a convex and concave surface [31]. This domain structure is well known for mediating protein-protein interactions in diverse classes of proteins [38]. The conserved residues in AlgK outside of those required by the structural TPR motif point to regions that may be important to bind to other proteins. In AlgK such conserved patches are located primarily on the convex side and in loop regions. Further, AlgK is required for the proper localization of AlgE. Sucrose gradient fractionation studies revealed a mixture of outer and inner membrane locations for AlgE without AlgK being expressed but show up in the outer membrane when AlgK is also expressed [31]. In a similar manner to group 2 (and 3) CPS export systems where KpsE is required for the proper localization of KpsD (outer membrane export)[39]. AlgK can then be thought of as an adapter or scaffold protein that links members of the inner membrane synthetic Alg 8/44 with the outer membrane secretion apparatus of AlgE. More importantly, proteins similar to the inner and outer membrane components in alginate are also found in cellulose and poly-β-1,6-GlcNAc exopolysaccharide systems in the bcs and pga operons respectively and here the regions homologous to AlgK and AlgE are fused. BcsC and PgaA respectively from these systems are predicted to have a N-terminal TPR containing module followed by C-terminal β-barrel like module (with greater than 16 strands). Protein fusions like this are a known evolutionary phenomenon that further suggests that these independent proteins function together [40].

In analogy with the alginate system above, we observe that *gfcC* and *gfcD* homologous genes to be fused as *otnA* in *Burkholderia sp*. GfcC here though is not a lipoprotein unlike AlgK but more interestingly predicted structure of GfcD show higher propensity for a large β-barrel protein with about 24 strands (see chapter 2 discussion) similar to AlgE. The group 1 polysaccharide system lacks genes homologous to *gfcABCD* within its operon or the role of similar genes elsewhere in the genome (like *yjbEFGH*) are not linked directly to group 1 capsule assembly. Thus, knowing the function of proteins encoded by *gfcABCD* encoded within the *gfc* operon will shed more light on group 4 polysaccharide assembly.

The structural and biochemical analysis of these proteins may even make group 1 and 4 more distinct. In particular knowing the structure and function of GfcC and GfcD will clear the role of β-barrel protein in this operon and the reason for the fusion between these two proteins. The structure of GfcC may show us if its presumed role of periplasmic adapter protein is possible and if the analogy with AlgK holds true. It will also serve as an attractive alternative exit portal for the growing polysaccharide with GfcD forming the export portion and GfcC linking it with other periplasmic proteins or periplasmic region of proteins in inner membrane biosynthetic machinery (Wzy polymerase and Wzx flippase).

The large β-barrel of GfcD may also export some other factors responsible for proper assembly of group 4 polysaccharide that may or maynot facilitate anchoring of the polysaccharide. This latter hypothesis is also of interest since the GFC system lacks a protein similar to Wzi of group 1 system that is thought to anchor the capsule to the outer cell surface.

We started exploring the structure of GfcC (chapter 2) to give answers to some of the questions raised above.

Chapter 2

Structure determination of GfcC and its relevance in capsule polysaccharide assembly systems

Despite the structural advances described in the previous chapter, there were still unanswered questions about how group 1 and group 4 capsular polysaccharides were exported from periplasm. These include:

1. Is the Wza outer membrane pore or its homologs (for instance GfcE from group 4) wide enough for branched oligosaccharide repeats found in other species?
2. How does the growing polysaccharide enter the inner chamber of the periplasm spanning Wza-Wzc complex?
3. How is the polysaccharide anchored? And, How is it anchored uniformly to the outer membrane as seen in electron micrographs of whole cells?
4. Does the exported capsular polysaccharide have a protein component?

We believed that investigating the structure and function of proteins encoded by *gfcABCD* in the group 4 operon might give insight into these questions. We were intrigued that just upstream of *gfcE* (*wza* homolog), *gfcD* encoded a predicted β-barrel (~24 strands) that might serve as an export channel for capsule or accessory molecule. Moreover some species encode proteins consisting of fused GfcC and GfcD homologs. This chapter presents my studies on the structure of GfcC, a small (26 KDa) periplasmic protein.

*Portions of this chapter are from a submitted manuscript to the journal Biochemistry.

K. Sathiyamoorthy, E. Mills, I. Rosenshine, M.A. Saper, Crystal structure of *E.coli* group 4 capsule protein GfcC reveals a domain organization resembling Wza, Submitted, Biochemistry, 2010

2.1 Methods

2.1.1 Design of the GfcC (22–248) construct for structural studies

The region of the *gfcC* gene encoding amino acid residues 22–248 (i.e., without the signal peptide) was PCR amplified from genomic DNA of *Escherichia coli O127:H6 str. E2348/69* with forward primer 5'-CGACCCATGGCGCAAGGAATGGTGACT-3' and reverse primer 5'-CGTCTCGAGCTCAGGAACACGTTGCGTTA-3'. After purification by PCR cleanup the oligonucleotide insert was digested with restriction enzymes NcoI and XhoI (New England BioLabs Inc.). The appropriately cut insert was ligated to similarly cut pETBlue2 vector (EMD biosciences). The N-terminal methionine and two residues leucine and glutamate between the end of the *gfcC* gene and C-terminal His_6 tag were present from the vector. Accurate insertion and ligation of the *gfcC* sequence in frame was verified by sequencing the resultant pETBlue2 plasmid DNA (Sequencing core facility, U of M). The final expressed GfcC protein has 236 residues with a theoretical molecular weight of 26,050.6 Da and contains no cysteines.

2.1.2 Expression and Purification of GfcC

Tuner (DE3) pLacI competent cells (Novagen) transformed with the pETBlue2 construct as above were grown with shaking at 37°C in 1L terrific broth (TB). When the OD_{600} was around 0.8 the heat in the shaker was turned down and the cells allowed to shake another 30 minutes to gradually equilibrate to room temperature. At this point the cells were induced with IPTG from a 1 M stock to a final concentration of 0.2 mM in the liquid culture (200 µl of 1M IPTG was used). Protein expression was allowed to proceed for 18 hours more at room temperature after when the cells were harvested by centrifugation at 5000 rpm in JLA-10.1 rotor for 20 minutes in a Beckman J2-HS centrifuge. The cell paste was then resuspended in 40 ml of 20 mM Tris, 300 mM NaCl pH 8.0 buffer and half tablet of EDTA-free protease inhibitor cocktail (Sigma). Cells were disrupted using a sonicator (Branson) with the macro-tip operating at 40% of the maximum power amplitude with 30 s ON pulses interspaced with 15 s OFF pulses for a total ON time period of 5 minutes. Soluble fraction was collected by centrifuging this lysate at 48,400 g (20,000 rpm in a JA-20 rotor) for 40 min in a Beckman J2-HS centrifuge. The supernatant from this centrifugation step was filtered using a 0.22 μm syringe filter (Millipore) and the filtrate applied to an 8 ml column packed with Talon® Superflow resin (metal affinity chromatography) connected to an AktaFPLC (GE Biosciences). Bound protein from the column was eluted in 5 ml fractions with a linear gradient of the resuspension buffer with imidazole extending to a concentration of 300 mM with a flowrate of 3 ml/min. GfcC was eluted around 20% Imidazole in the elution buffer that corresponds to 60 mM Imidazole. The fractions containing GfcC as determined by SDS-PAGE analysis were then pooled and subject to overnight dialysis in 4L (100 fold excess) of 20 mM Tris, 300 mM NaCl, pH 8.0 buffer to remove the Imidazole. This relatively pure protein then underwent a final polishing step using size exclusion chromatography with a 120 ml column pre-packed Superdex 75 preparative column (from Amersham) with a flowrate of 1.2–1.5 ml/min. The purified protein was then concentrated with a 10 kDa cutoff membrane centrifugal filter (Amicon Ultra, Millipore) to about 23 mg/ml (determined by $A_{280\mathrm{nm}}$, theoretical extinction co-efficient of GfcC based on sequence is 41,940 $M^{-1}.cm^{-1}$ calculated by Expasy-Protparam tool [41]) to use for finding conditions favorable for crystallization.

The selenomethionine derivative of GfcC (SeMet-GfcC) was purified by a similar procedure except the cells were washed before induction with IPTG in 1 L Athena Expression Systems minimal media with excess selenomethionine (0.8 μg/ml) to favor preferential incorporation of selenomethionine instead of methionine. The cells used here were not methionine auxotrophs. Extent of selenomethionine incorporation was determined using mass spectrometry (ESI-MS). All buffers contained 2 mM TCEP (Soltec Ventures Inc.) to prevent oxidation of selenomethionine.

The native GfcC protein and SeMet-GfcC was subject to ESI-MS that gave exact molecular weights with an uncertainty of ± 1 Da. The native GfcC gave a molecular weight of 25919.0 Da and the SeMet-GfcC was 26060.0 Da. The difference 141 Da can be explained by three methionines replaced with selenomethionines (Atomic Weight of Se: 78.96, Atomic weight of S: 32.07; 3X46.89=140.67 Da). From the sequence of GfcC, this would mean three out of the four methionines in GfcC monomer (except the N-terminal start site) had been replaced to SeMet (Figure 2-2).

2.1.3 Crystallization and Structure determination of GfcC

Crystals appeared both at 22°C and 4°C under few different precipitant conditions from the various commercially available kits that were used (Wizard I and II from Emerald Biosystems; Index and Crystal screen I and II from Hampton; PACT, pHClear and Classics from Qiagen with additives I, II and III sets from Hampton Research Inc.) but the initial crystals obtained were all of needle morphology and had to be optimized for structure determination by X-ray Diffraction. Better-shaped single crystals were obtained by a combination of additives and temperature shift trials after about two weeks. The best crystals for the native protein were obtained with 1 µl of 23 mg/ml GfcC protein + 1 µl of reservoir solution having 0.1 M Tris pH 8.0, 1.6 M ammonium sulfate (E11 pHClear condition, Qiagen) and 0.2 µl of the additive 30% w/v 1,5-diaminopentane dihydrochloride (Hampton Research) equilibrated against 50 µl of reservoir solution (sitting-drop vapor diffusion method). The experiment was initially set up in 4°C but single crystals appeared within two–three days after the experiment had been shifted to 22°C on the eleventh day from its setup. Different concentration levels of precipitants (through grid screens) and different time period in 4°C and 22°C were explored simultaneously.

SeMet-GfcC crystallized at 4°C under similar conditions but the reservoir for the best looking single crystals had 0.1 M Bis-Tris pH 6.5, 0.1 M NaCl, 1.5 M Ammonium sulfate (C6 Index, Hampton Research). There were no temperature shifts necessary for crystallizing SeMet-GfcC. All crystals were harvested and frozen with original reservoir condition having 15% glycerol as the cryoprotectant. Data was collected from these frozen crystals at the LS-CAT beamline (21-ID-D and -G) at the Advanced Photon Source (APS, Argonne, IL). Three-wavelength data collection strategy was adopted for SeMet-GfcC crystals at the 21-ID-D beamline with the peak (12,660.56 eV), inflection (12,658.09 eV) and high-energy remote (12,680 eV) datasets collected in that order. The energy corresponding to these three wavelengths was calculated with a x-ray fluorescence scan using the frozen SeMet-GfcC crystal subsequently not used for actual data collection due to radiation damage. The x-ray fluorescence scan is shown below (refer Figure 2-3). For structure calculation purposes the peak dataset was used as in a single wavelength anamolous (SAD) experiment. Intensity data was indexed, integrated and scaled with HKL2000 (HKL Research Inc.)[42]. The SeMet-GfcC belonged to the $P2_1$ spacegroup and had four molecules per asymmetric unit. The point group was independently verified by *POINTLESS* (from the CCP4 package, Comprehensive Computing suite for Protein crystallography)[43] and systematic absences visualized with *HKLVIEW* (also from the CCP4 package). The structure factors were phased by single wavelength anomalous diffraction (SAD) phasing of the SeMet-GfcC peak data using the program(s) Phaser/Resolve as implemented in the AutoSol routine under Phenix [44]. Solvent flattening by Resolve resolved the phase ambiguity that otherwise result in two different phase values common to the single wavelength diffraction approach. Ninety-three per cent of the residues were built (i.e., including Ala residues) of which only 83% (or 765 out of 912 residues in 4 chains) of the main chain residues were automatically placed by Resolve having the right sidechain as in input GfcC sequence [45]. The rest of the model was fit manually with Coot v0.6 [46]. Structure refinement using the program *phenix.refine* with simulated annealing and two groups of NCS restraints for pairs of chains and TLS with one chain as one group gave a final R_{work}/R_{free} value of 0.18/0.22 with 702 water molecules. The native GfcC crystal belonged to the $P4_32_12$ space group with two molecules in the asymmetric unit. The $P4_32_12$ model was solved by molecular replacement with chain A of the SeMet-GfcC model by Phaser (CCP4) [47] and was refined to 2.1 Å with *phenix.refine* to a final R_{work}/R_{free} value of 0.24/0.28 with 119 water molecules. The electron density ($2F_o$-F_c) map in a good and bad region of $P2_1$ model of GfcC is shown in figure 2-4.

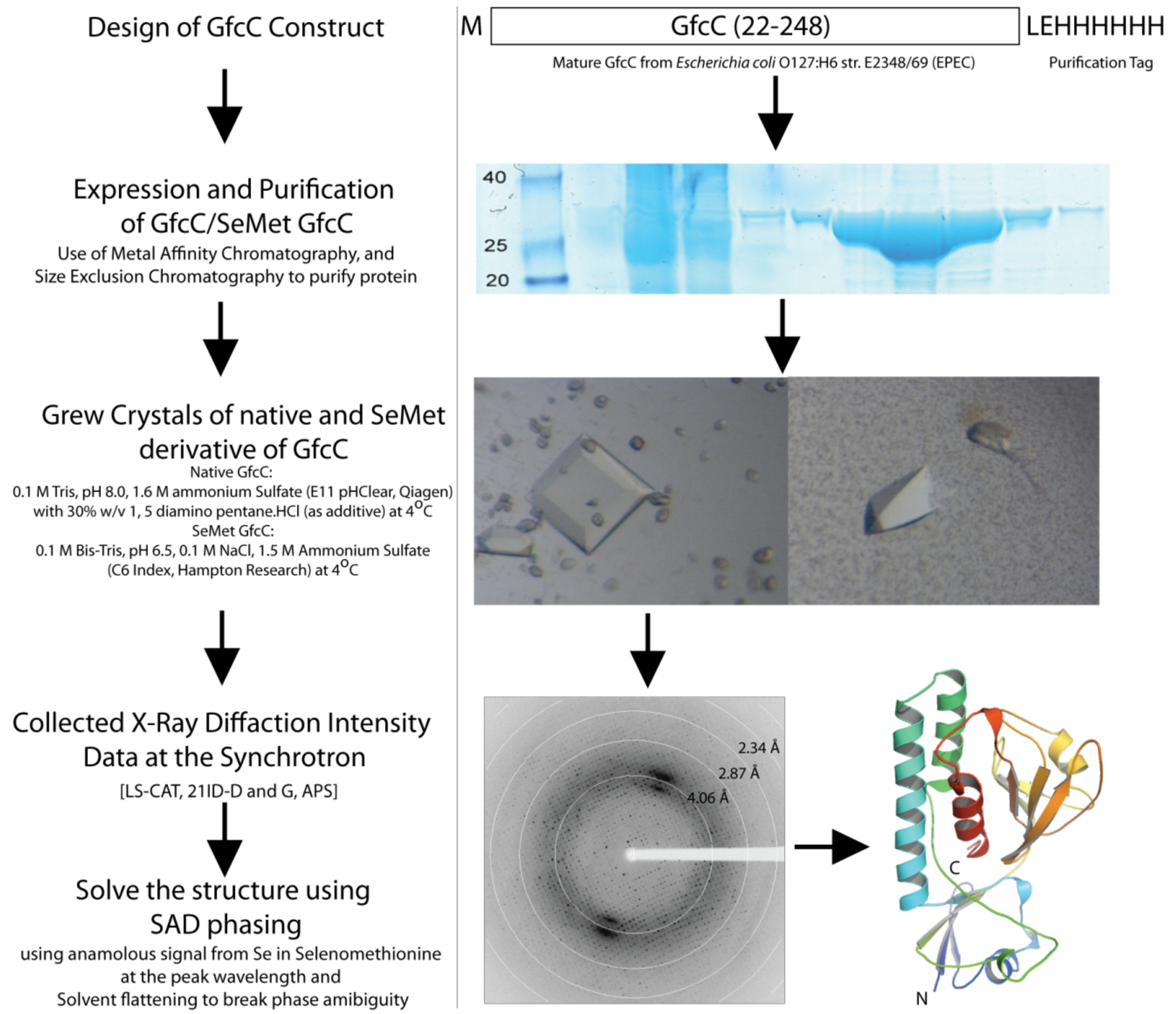

Design of GfcC Construct
M
GfcC (22-248)
LEHHHHHH
Mature GfcC from *Escherichia coli* O127:H6 str. E2348/69 (EPEC)
Purification Tag
Expression and Purification of GfcC/SeMet GfcC
Use of Metal Affinity Chromatography, and Size Exclusion Chromatography to purify protein
40
25
20
Grew Crystals of native and SeMet derivative of GfcC
Native GfcC:
0.1 M Tris, pH 8.0, 1.6 M ammonium Sulfate (E11 pHClear, Qiagen) with 30% w/v 1, 5 diamino pentane.HCl (as additive) at 4°C
SeMet GfcC:
0.1 M Bis-Tris, pH 6.5, 0.1 M NaCl, 1.5 M Ammonium Sulfate (C6 Index, Hampton Research) at 4°C
Collected X-Ray Diffraction Intensity Data at the Synchrotron
[LS-CAT, 21ID-D and G, APS]
2.34 Å
2.87 Å
4.06 Å
Solve the structure using SAD phasing
using anamolous signal from Se in Selenomethionine at the peak wavelength and Solvent flattening to break phase amibiguity
C
N

Figure 2-1 **Overview of expression, purification, crystallization and structure detrmination of GfcC**

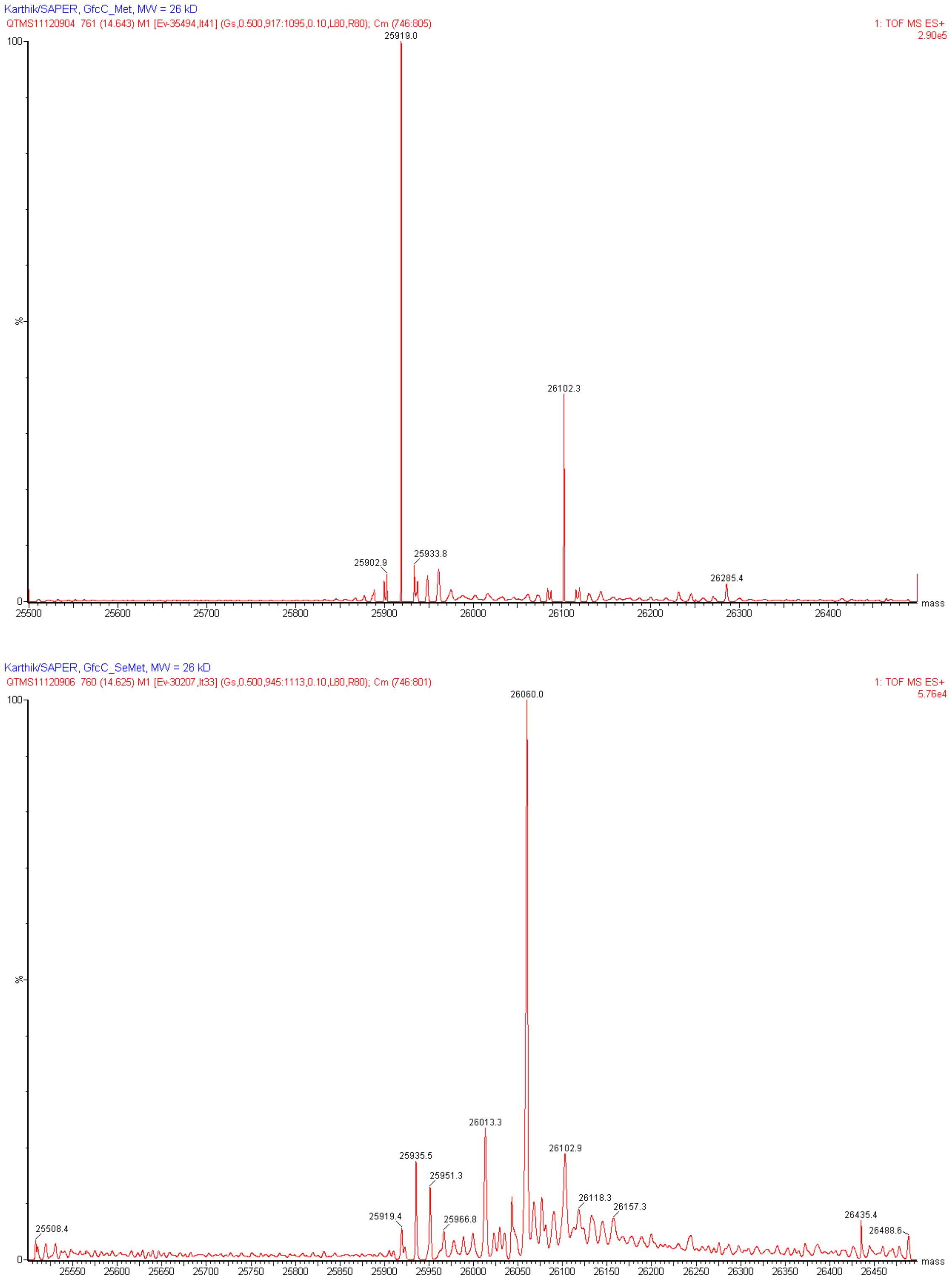


Figure 2-2 **ESI-MS spectrum for GfcC_Met and GfcC_SeMet**. The difference in molecular weight of the major peak (26060.0 Da-25919.0 Da=141 Da) confirms incorporation of 3 SeMet aminoacids instead of Met per GfcC monomer. (U-M Mass Spectrometry Core) (see Methods)

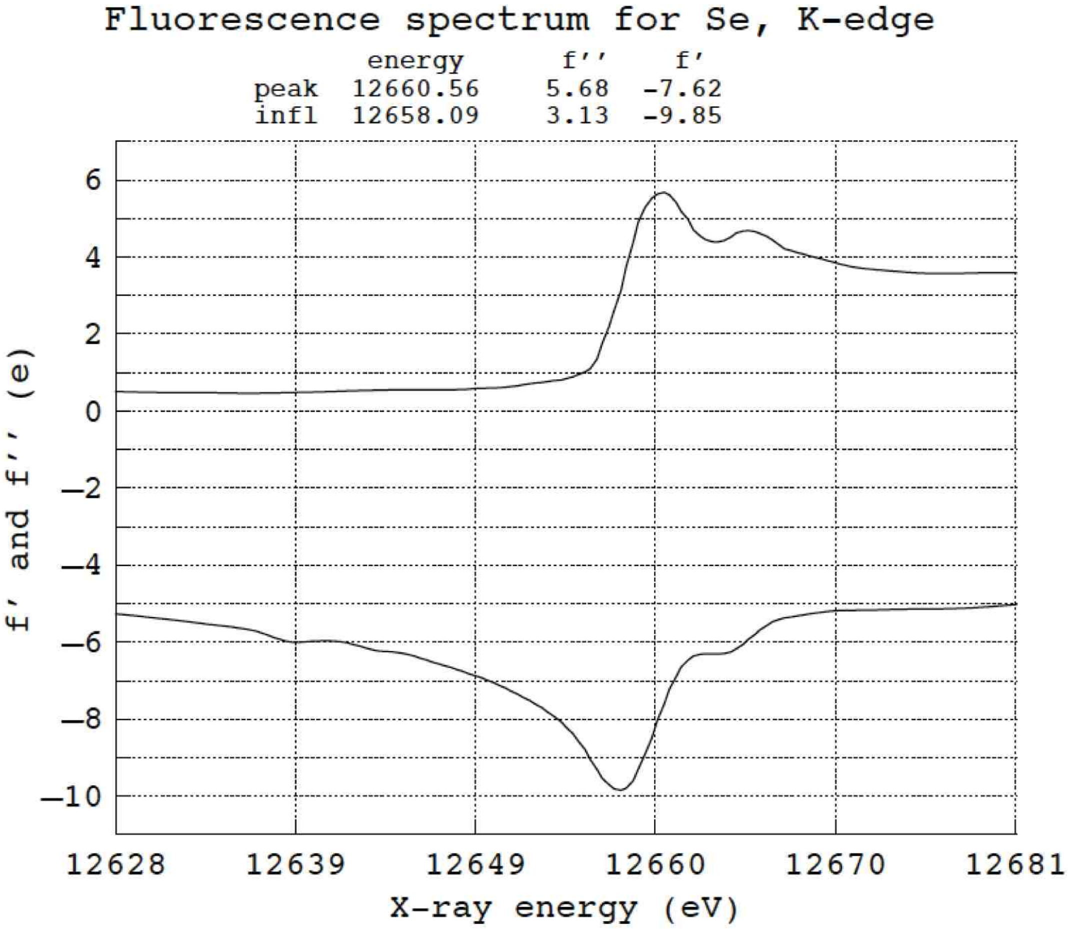


Figure 2-3 **Chooch plot** for the fluorescence spectrum of one SeMet crystal at the LS-CAT beamline 21 ID-D to determine peak and inflection energy. The energy of the beam was then tuned to these numbers for SAD/MAD data collection. The crystal used to collect this data was not used for data collection due to it capable of suffering significant radiation damage.

Table 2-1 **Data collection and refinement statistics fof SeMet-GfcC (PDB: 3P42) and native GfcC crystal structures**

Protein	**SeMet-GfcC (22-248)**	**Native GfcC (22-248)**
Data Collection and Processing		
PDB ID	*3P42*	*To be deposited*
Beamline	21 ID-D, LS-CAT, APS	21 ID-G, LS-CAT, APS
Space Group	$P2_1$	$P4_32_12$
Unit Cell (Å)	a=68.83, b=99.98, c=69.02; β=91.74^{o}	a=69.5, b=69.5, c=197.8; α=β=γ=90^{o}
Number of molecules/asu	4	2
Unique Reflections	81677 (3213)	27327 (1374)
Redundancy	6.8 (5.3)	8.5 (5.9)
Completeness (%)	98.7 (96.4)	99.8 (99.7)
R_{merge}	0.091 (0.469)	0.058 (0.404)
I/σ_I	30.88 (2.98)	51.42 (4.07)
Refinement Program	*Phenix 1.6.4-486*	*Phenix 1.6.1-357*
Resolution (Å)	34.5-1.91 (1.98-1.91)	40.3-2.15 (2.22-2.15)
R_{work}	0.17 (0.26)	0.22 (0.22)
R_{free}	0.22 (0.31)	0.28 (0.3)
Number of TLS groups	4 (one per chain)	2 (one per chain)
Number of Atoms (non-hydrogen)	7860	3577
Protein	7143	3458
Water	702	119
Sulfate	15	-
Wilson B (Å^2) (sfcheck)	33.4	49.4
Mean B_{iso} (Å^2, non-hydrogen)	37.1	45.8
Protein	36.8	45.7
Solvent	39.4	46.1
R.M.S.D		
Bonds (Å)	0.008	0.007
Angles (^{o})	1.110	1.051
Ramachandran favored (%)	98.68	96.82
Ramachandran outliers (%)	0.22	1.36

$$R_{merge} = \frac{\sum_{hkl}\sum_{i} | I_i(hkl) - \overline{I(hkl)} |}{\sum_{hkl}\sum_{i} I_i(hkl)}$$

$$R_{work} = \frac{\sum_{hkl} || F_{obs} | - | F_{calc} ||}{\sum_{hkl} | F_{obs} |}$$ calculated over all reflections used in refinement

R_{free} is similar to R_{work} but calculated from 5 % of the total number of reflections omitted in the refinement

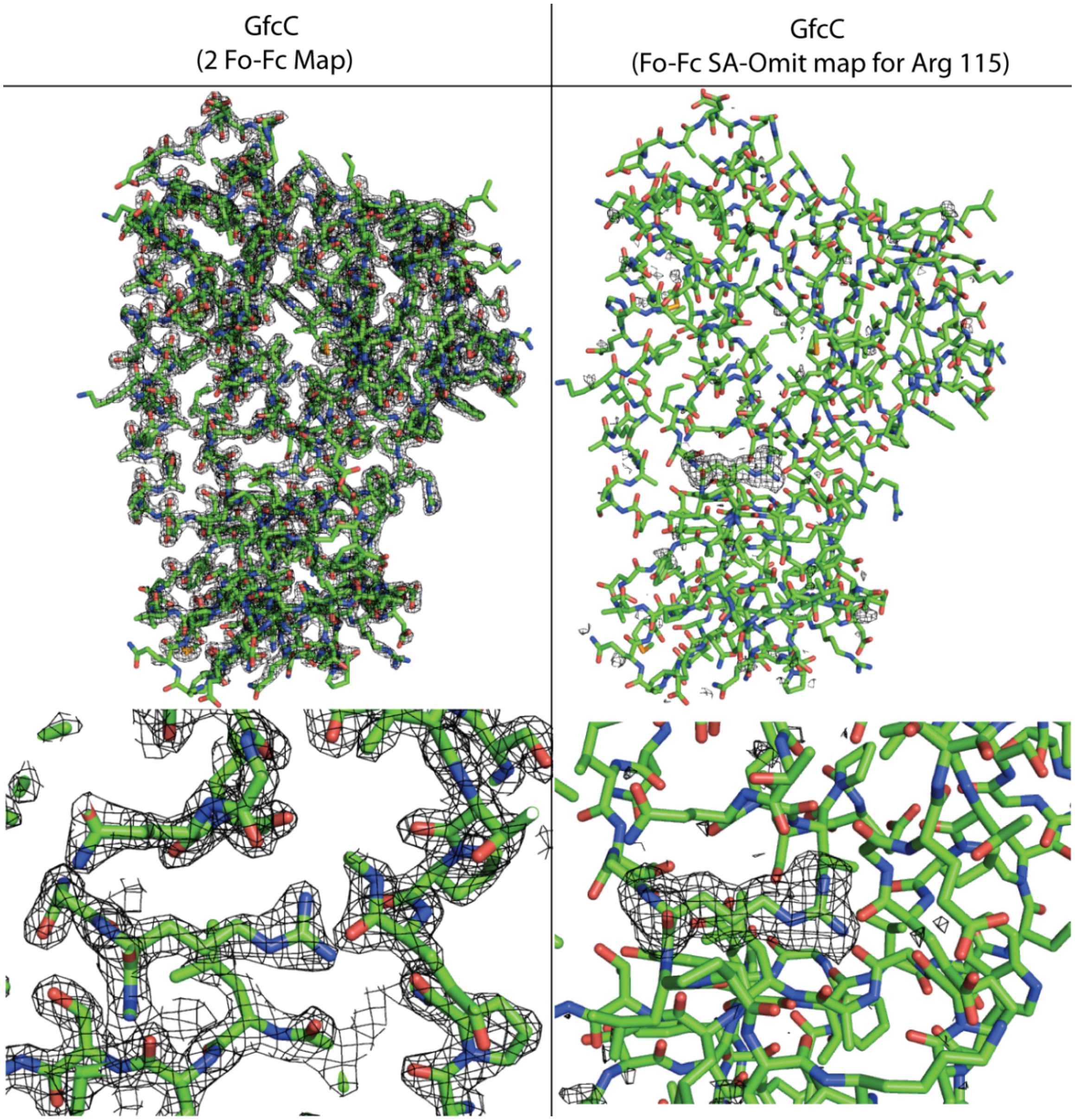
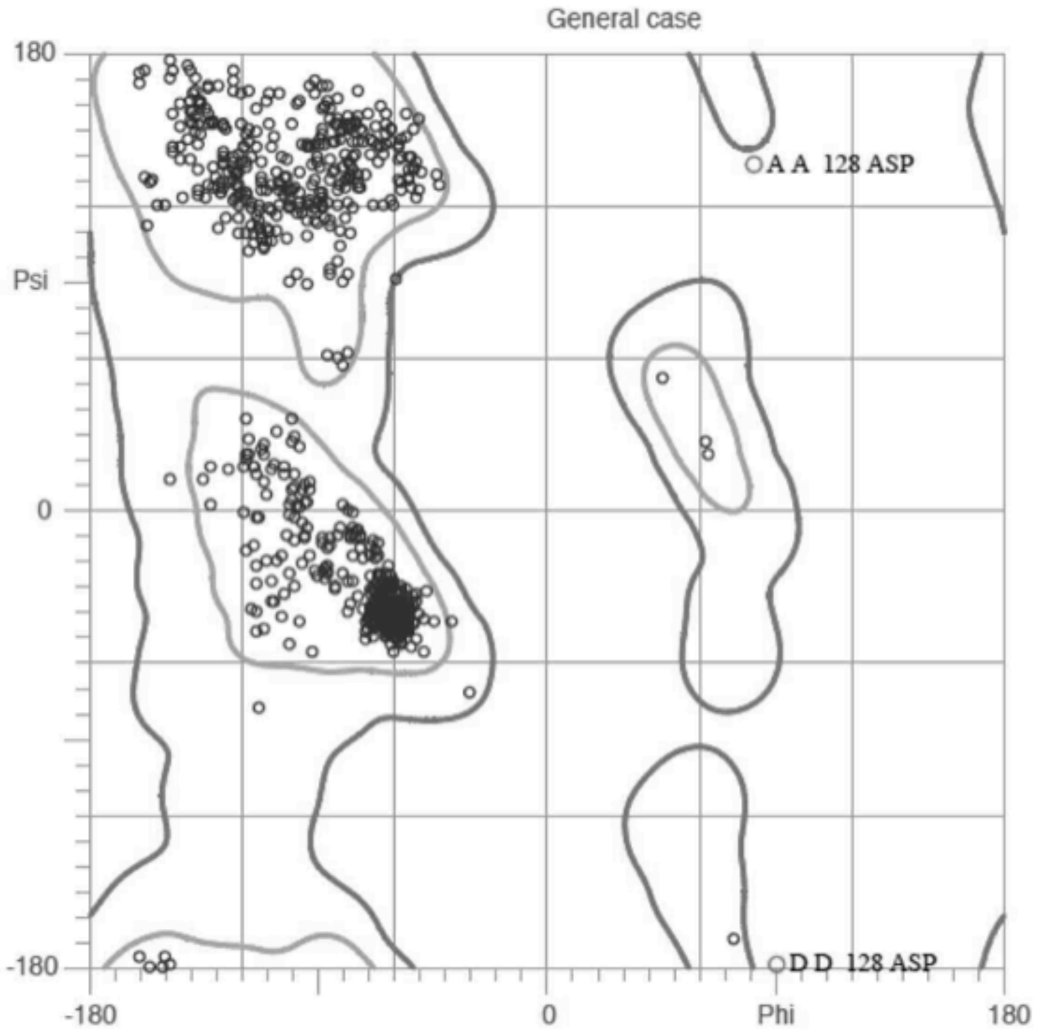


Figure 2-4 **Electron density of GfcC** *Left, Top*: ($2F_o$-F_c) map for $P2_1$ model of GfcC, bottom: shows the same map around the invariant Arg115 (see text). A clipping plane is applied in PyMol for bottom figure for clarity. *Right, Top*: F_o-F_c SA-omit map with Arg115 omitted. Maps were calculated by map utilities (fft) in CCP4 and the map was incorporated in MacPyMol at a contour level of 2.0 σ for ($2F_o$-F_c) map (left) and 2.5 σ for (F_o-F_c) SA-omit map (right) along with the GfcC, $P2_1$ structure as stick model for purposes of rendering this figure.



98.7% (903/915) of all residues were in favored (98%) regions.

99.8% (913/915) of all residues were in allowed (>99.8%) regions.

Figure 2-5 **Ramachandran plot** from Molprobity [48] showing the phi-psi values for peptide bond geometry in SeMet-GfcC structure (PDB ID: 3P42)

2.2 Results

2.2.1 Sequence analysis of GfcC shows orthologs belong to family of proteins with domains of unknown functions (DUF1017)

Proteins similar to GfcC in size and sequence homology were clustered as a subgroup in the DUF1017 pfam protein family spanning over 29 genera of proteobacteria. This subgroup belonged to a superfamily of proteins related to the ubiquitin protein fold family. This superfamily also includes subgroup that defines Wza (SLBB-type) folds. 34 sequences from the precomputed BLAST results (BLink) from NCBI were culled after removing sequences having ≥ 98% sequence identity. All these resultant sequences are predicted to express periplasmic proteins with identity ranging from 19–89% and the majority between 20–40% to GfcC. Further, majority of them are located with their locus in between *gfcB* and *gfcD* homologs. The conserved domain identifier DUF1017 starts at about residue 76 of GfcC where homology to other sequences is particularly high. Apart from proteins similarly sized to GfcC the family also includes larger sequences with domains homologous to GfcC. Figure 2-6 shows the alignment of GfcC with seven other representative GfcC homologs that were chosen based on an unrooted sequence tree with the most distinct branchpoints.

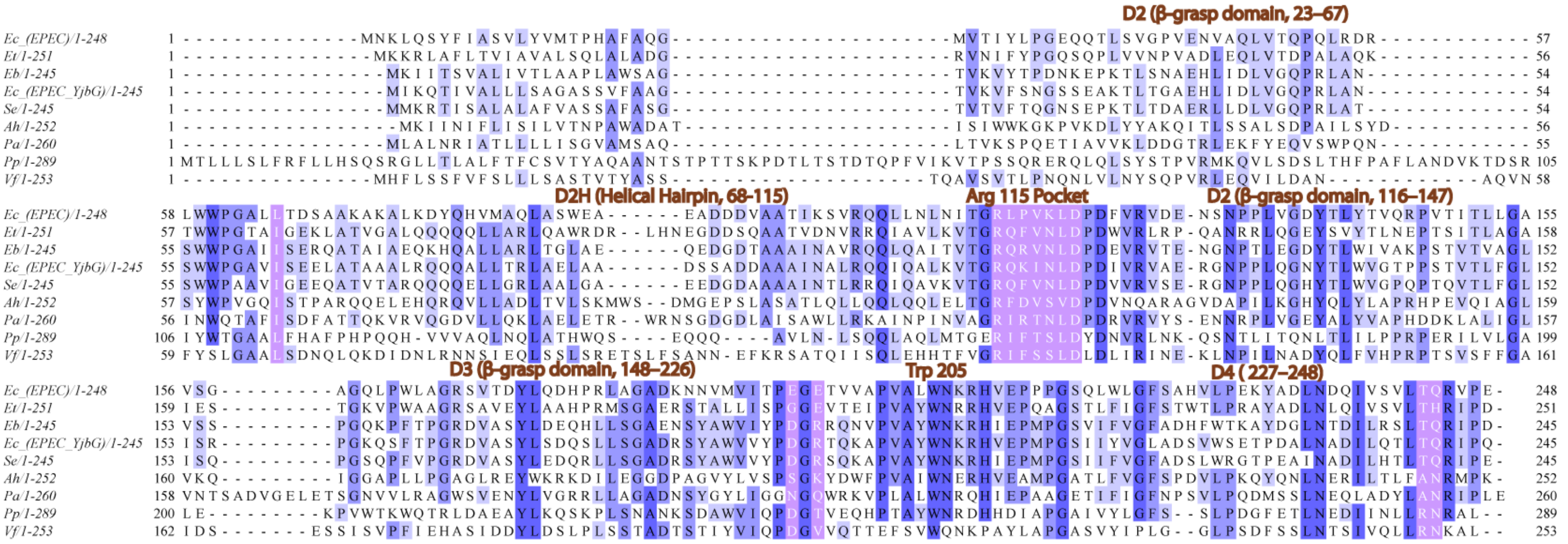


Figure 2-6 **Sequence alignment of GfcC homologs**. The figure shown here is a representation of nine homologous sequences having the DUF1017 domain and the GfcC studied here is shown as the first sequence. The identical residues are colored by degree of identity with shades of blue. Violet colored residues form the conserved pocket close to Arg 115. The sequences are colored by percentage identity using Jalview [49]. The dark blue represents residues with >80% identity, light blue >60% and the lightest blue represent >40 %. The residues in white are <40% identical. The domain ranges as found in structure of GfcC are written above in brown. The sequences are used from NCBI and are as follows along with their genbank id (gi): Ec_(EPEC): *E.coli* (gi 215486102), Et: *Erwinia tasmaniensis* (gi 188535225), Eb: *Eneterobacter sp. 638* (gi 146309900), Ec_(EPEC_YjbG): *E.coli* (gi 215489366), Sc: *Salmonella enterica* (gi 161366713), Ah: *Aeromonas hydrophila* (gi 117619422), Pa: *Pectobacterium atrosepticum* (gi 50120386), Pp: *Photobacterium profundum* (gi 90328324), Vf: *Vibrio fischeri* (gi 59710765)

2.2.2 Crystal Structure of GfcC at 1.91 Å shows two β-grasp domains

GfcC belongs to the mixed (α and β) class of proteins (Structural Classification of Proteins, SCOP) with four distinct structural motifs herein defined as domains, D2, D2H, D3 and D4 for comparison with the structure of Wza (*see below*, there is no D1 defined to illustrate the structurally analogous domains with that of Wza). D2 (residue range 23–67, 116–147) and D3 (residue range 148–226) are β-grasp domains belonging to domain of unknown function (DUF1017) protein family (Pfam: PF06251) of the ubiquitin clan that are annotated as capable of binding to diverse soluble ligands [50]. A central mixed β-sheet lined on one side by a α-helix is characteristic of these domains. D2 has only four β-strands (S1–S4), one strand less than in D3 (S5–S9). The canonical S4 strand in D2 compared to typical β-grasp (like D3) has been reduced to an extended region (residues 116-121) between strands labeled S3 and S4. This result due to a α-helical hairpin insert that is labeled D2H (residue range 68–116) that protrudes out of D2 (H2/H3 in figure 2-7) here. The extended 'S4' pseudo strand is angled 45º to the rest of the β-sheet of D2 and makes only one hydrogen bond to the neighboring S3 strand (Leu 116/N to Leu 65/O 2.9 Å). All the residues are numbered corresponding to that of the full-length protein with the signal peptide.

The D2H helical hairpin between strands S3 and S4 in D2 is over 40 Å long and consists of a regular α-helix with about 7 turns (residues 68–91) followed by a reverse turn and then a 4.5 turn helix (residues 94–109) continued by an extended chain of 5 residues before finally returning back to S4 strand in D2. The helical hairpin is a distinct structural domain not present in Wza. The presence of this insert is also reason tantamount to not annotating D2 as β-grasp domain previously in protein databases and has been revealed here for the first time only by direct structural analysis.

Domain 4, (D4) (residue range 227–248) comprises a C-terminal amphipathic helix (H5 in figure 2-7) that folds back on to the core of the protein and interacts with the β-sheet of D3 on one side and the α- helical hairpin (D2H) on the other side to give the protein overall a relatively compact tertiary structure.

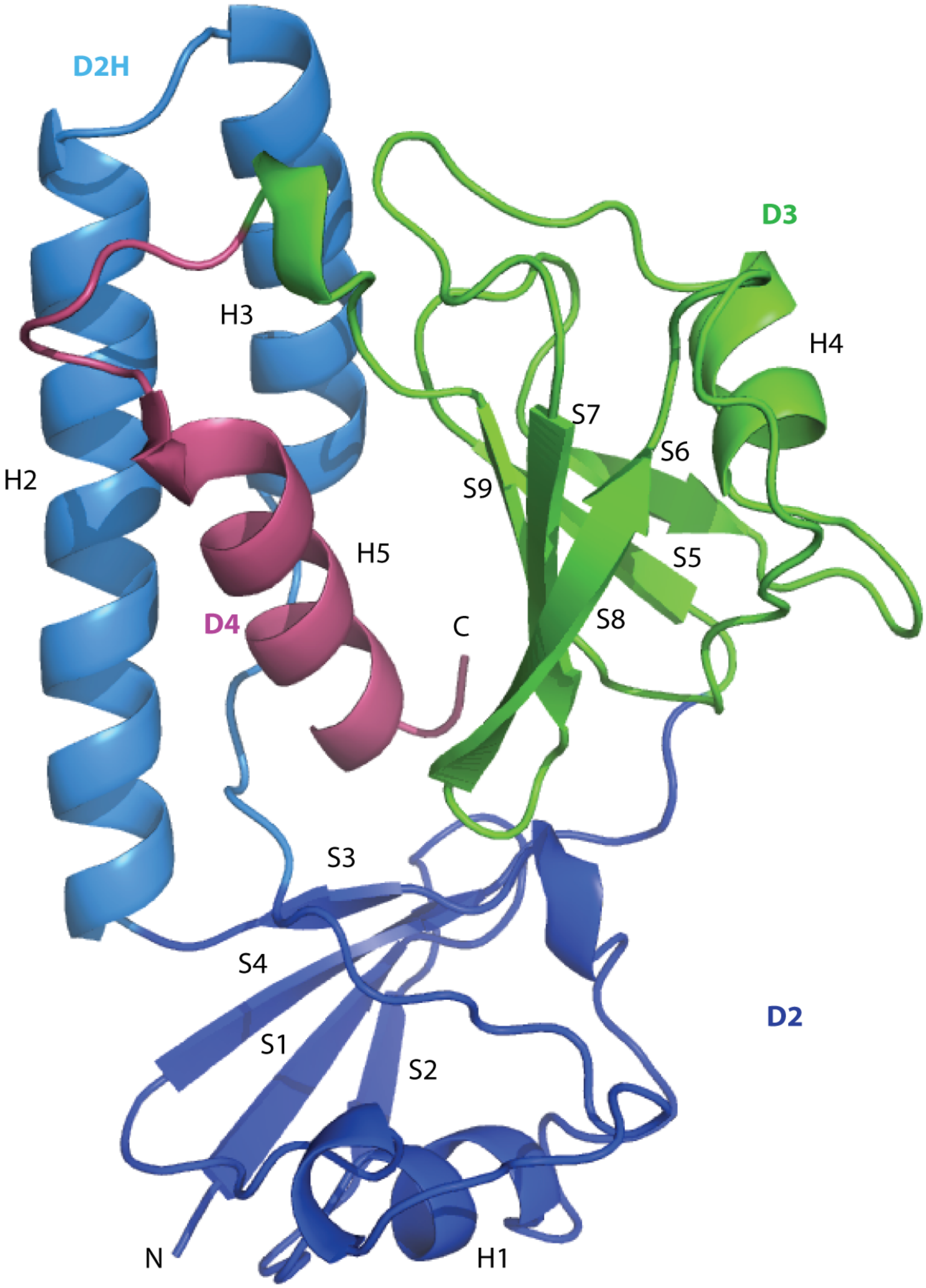


Figure 2-7 **Crystal structure of GfcC monomer** with the four domains defined as D2 (blue), D2H (marine), D3 (green) and D4 (warmpink). This figure and all subsequent and previous figures showing protein structures were generated using MacPyMol

2.2.3 Structure of GfcC closely resemble Wza, the outer membrane export protein for group 1 polysaccharide

The most striking structural feature of GfcC is its close similarity of the various domains with that of the outer membrane protein Wza involved in the export of group 1 polysaccharide [15](Figure 2-8). GfcC and Wza both have two β-grasp domains (D2 and D3) that superpose well with each other (see later). The sequence identity between GfcC and Wza is only ~11% but the similarity in their domain organization is distinct. The domains do have some differences in terms of twist of the β-strands and the number of α-helices surrounding the β-sheet. The D3 domain in Wza close to the membrane has a sixth β-strand contributed by the lipid anchored N-terminal region[15]. At this point while comparing the striking differences and similarities between GfcC and Wza it should be remembered that the *gfc* operon encodes GfcE which is 74% identical and 95% similar to Wza, and thus is expected to have the same domain structure as described for Wza.

While the β-grasp domains line up well between GfcC and Wza, the respective D4 domains have stark differences and likely reflect different oligomeric state and function for the two proteins (figure 2-8). In Wza, the D4 domain is an amphipathic α-helix (38 residues) that does not interact with any other domain in its structure and extends away from D3 to interact with a similar α-helix from a neighboring Wza molecule. Eight such monomers combine to give the octameric Wza inserted in the outer membrane[15]. Helix H5 in GfcC is also amphipathic as seen by helical wheel projections but it is much shorter (13 residues) and folds back to interact with all the other domains in GfcC to various extent. The non-polar surface interacts with both β-sheet of D3 and the D2H. While Wza has aromatic residues (for instance, tryptophan) characteristic of membrane-inserted α-helices at both ends, GfcC only has a Tyr at the beginning of helix H5. Both the helices however have charged aminoacids (Arg or Lys) at either ends of the helix. In the case of Wza it favors interactions with the phospholipid head groups.

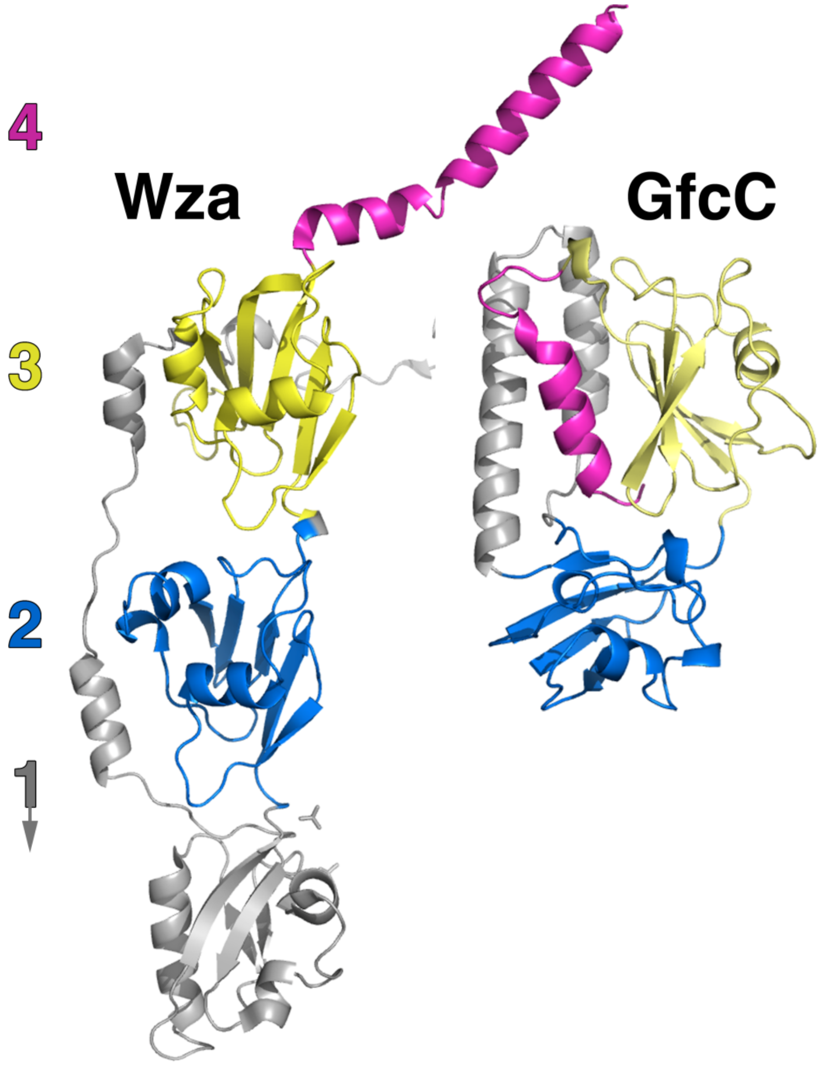


Figure 2-8 **Comparison of GfcC with Wza**. Similar domains are colored identically except for those in grey. The D2H domain in GfcC represented in grey does not have an equivalent domain in Wza. Similarly, the D1 domain of Wza is absent in GfcC. The C-terminal amphipathic helix represented in magenta is in different orientation in GfcC and Wza.

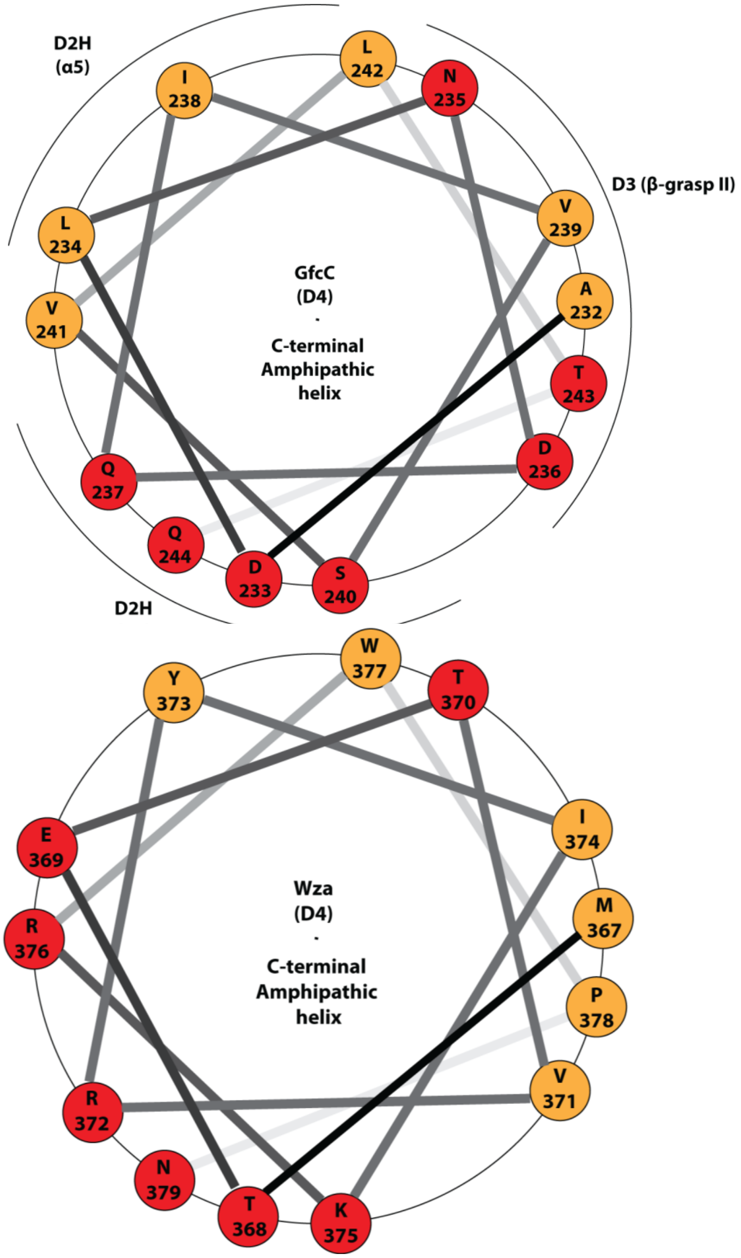


Figure 2-9 **Helical wheel for the C-terminal region of GfcC and Wza showing the amphipathic character**. Aminoacids in red circles are charged or polar and the yellow circles represent non-polar or hydrophobic aminoacids

2.2.4 Superposition of domains D2 and D3 of GfcC and Wza

It was just seen the D2 and D3 of GfcC and Wza have similar β-grasp domains. In fact DALI server [51] that compares structural similarities to all known protein structures suggests the match between GfcC D3 with either D2 or D3 of Wza that it is to any other known β-grasp domain (Z=9.1–9.5 vs z=6 for the next best fit). Based on structural alignments shown below (figure 2-10) computed by LSQMAN [52] it can be seen how similar these two domains align. In spite of the remarkable structural similarity, GfcC D3 has only 18% and 12% sequence identity to Wza D2 and D3 respectively.

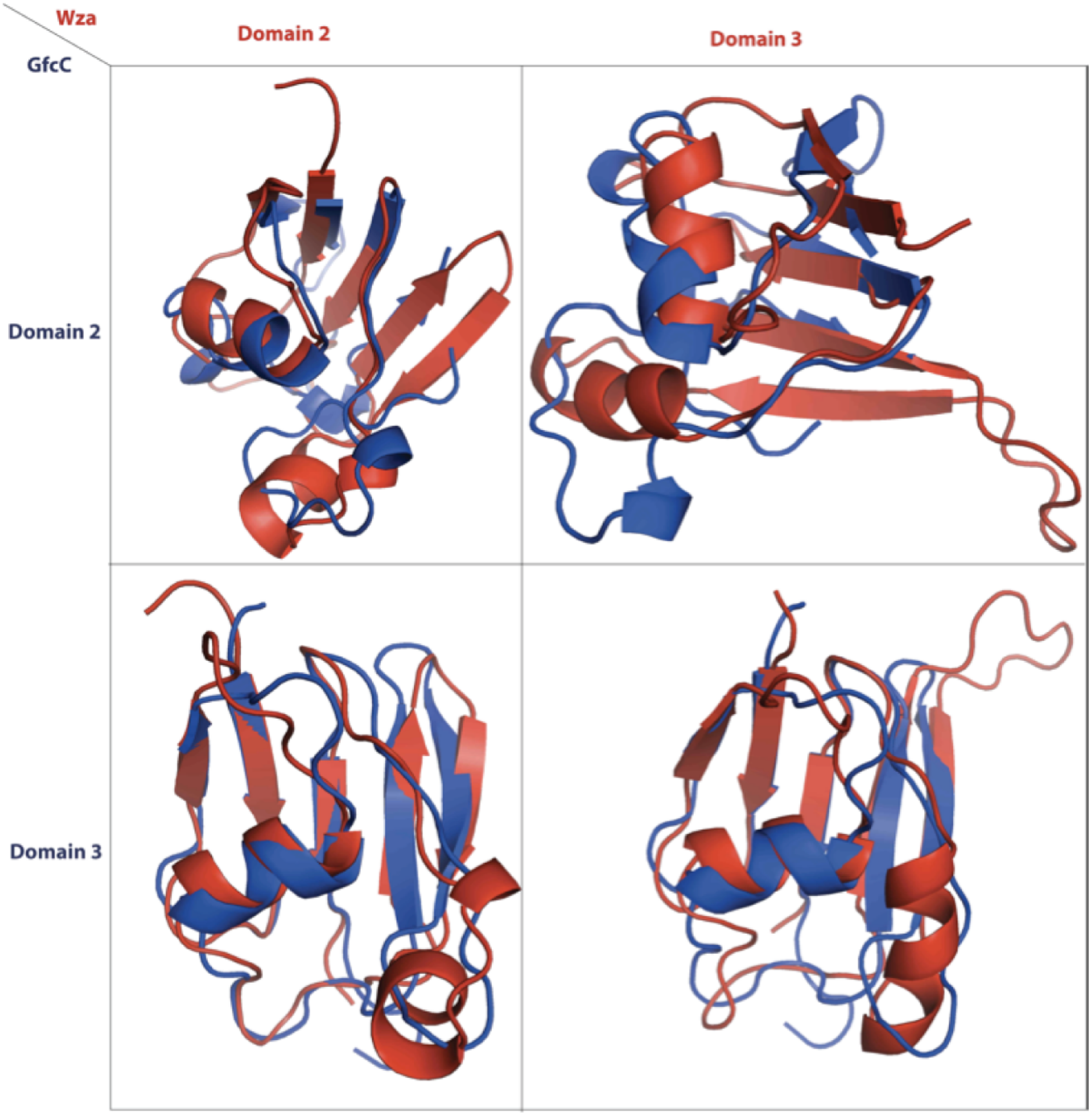


Figure 2-10 **Superposition of D2 and D3 of GfcC and Wza by LSQMAN**. Wza domains are colored in red and GfcC colored in blue.

LSQMAN Alignment	Number of Matched Residues/ Number of Aligned Residues	Levitt-Gerstein Statistic P(z>Z) [X 10^{-5}]	RMSD (Å)
GfcC D2 / Wza D2 (Top-Left)	54/104	2.647	1.97
GfcC D2 / Wza D3 (Top-Right)	52/113	2.743	2.05
GfcC D3 / Wza D2 (Bottom-Left)	71/93	0.006	1.59
GfcC D3 / Wza D3 (Bottom-Right)	68/103	0.019	1.57

2.2.5 Two different crystal forms reflect very similar monomer and observed dimer conformations

The atomic structure of GfcC was solved using diffraction data from two different crystal forms under similar precipitant conditions but slightly varying pH. The SeMet derivative of GfcC belonged to the space group $P2_1$ (number 4, monoclinic) with four molecules in the asymmetric unit with the four chains stacked in pairs aside each other- A:B and C:D and the final structure was refined to 1.91 Å (PDB code: 3P42). The protomers in the dimer are related by non-crystallographic two-fold axis and the dimers are related by approximate 4_1 screw axis parallel to the c axis of the unit cell. The chain A thus rests in the cradle formed by C:D dimer. This results in the resolved His purification tag that is usually unstructured and is not visible for the other chains in the $P2_1$ model. Native GfcC crystallized in space group $P4_32_12$ (number 96, tetragonal) with two molecules in the asymmetric unit and was solved by molecular replacement using chain A of SeMet-GfcC structure as the search model in Phaser (CCP4) and was refined to 2.1 Å. The GfcC monomer in both the crystal forms adopts very similar conformations with no major structural rearrangements. Further, the observed dimer is also identical between both the crystal forms, and the dimer chains A:B (or C:D) from SeMet-GfcC can be used as the search model in molecular replacement to solve for the observed dimer structure in the native GfcC crystal (Tetragonal). The tetragonal structure model had poorer statistics than the monoclinic model and was not used for detailed analysis.

2.2.6 The interface between the observed dimer in the crystal has mostly conserved residues

The residues in the interface for the observed dimer in the crystal structure belong predominantly to the loop regions of the two β-grasp domains of GfcC. There is an inverse relation with each half of the interface formed by interacting D2:D3 domains from different protomers. The dimer interface is identical in both the tetragonal and monoclinic crystal structures and buries a total of 1215 Å^2 of area according to PISA calculation. The interface is also mostly polar with ~17 hydrogen bonds and 14 salt bridges along with many water mediated interactions. The conserved residues here are also the most conserved residues on the surface of monomeric GfcC. The conservation scores were calculated using consurf [53] from a user input multiple sequence alignment (see above) that uses Bayesian methods to determine the residue conservation score from the input alignment. The conserved residues in the surface include Trp 59, Asp 123, Arg 126, Asn 132, His 209, Glu 211, and Pro 214. The interface does not bury enough hydrophobic residues to be stable in solution. In fact, we observed a single peak in our size exclusion chromatography corresponding to a monomer and analytical ultracentrifugation studies indicated it to be a monomer with a molecular weight ~26 kDa (see below).

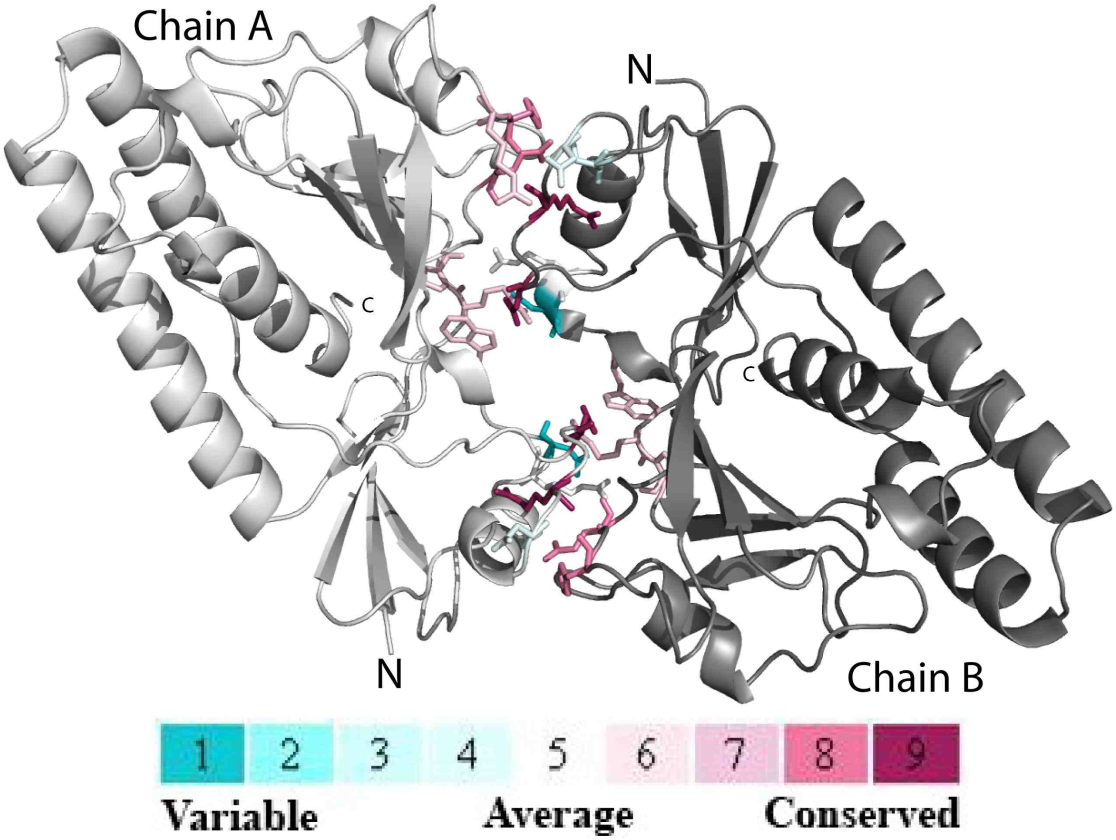


Figure 2-11 **The A:B dimer interface in cartoon representation**. The residues in sticks show the interface residues and colored according to conservation score calculated from Consurf [53] web server using user input multiple sequence alignment (see in text). The magenta indicated by numerical score 9 represent the highly conserved with the given alignment and cyan with score 1 represent the least conserved. Only the interface residues as identified by PISA are colored according to the conservation scale for clarity while the rest of the chain is indicated by different shades of grey.

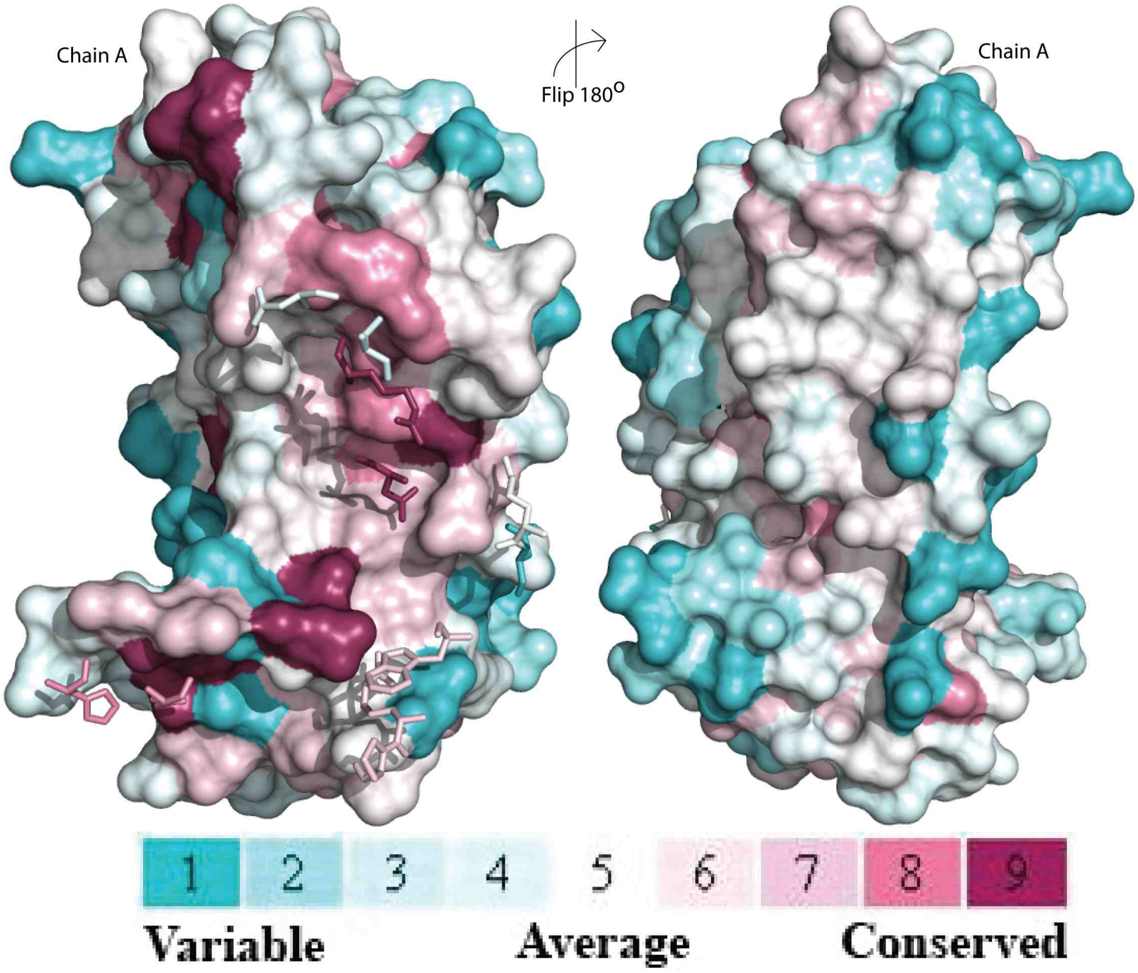


Figure 2-12 **The A:B dimer interface shown from a different perspective** *Left:* The A:B dimer interface is portrayed with the entire chain A surface colored accoring to conservation (Consurf) and the residues in the dimer interface from chain B are shown as sticks. *Right:* In comparison to the dimer interface the rest of the protein surface has lesser number of highly conserved residues.

2.2.7 Analytical Ultracentrifugation studies showed GfcC to be a monomer

Purified GfcC that was used for crystallization was run using velocity based ultracentrifugation experiment and the resulting analysis pointed to a major peak with a molecular weight of 26 kDa representative of a GfcC monomer in solution.

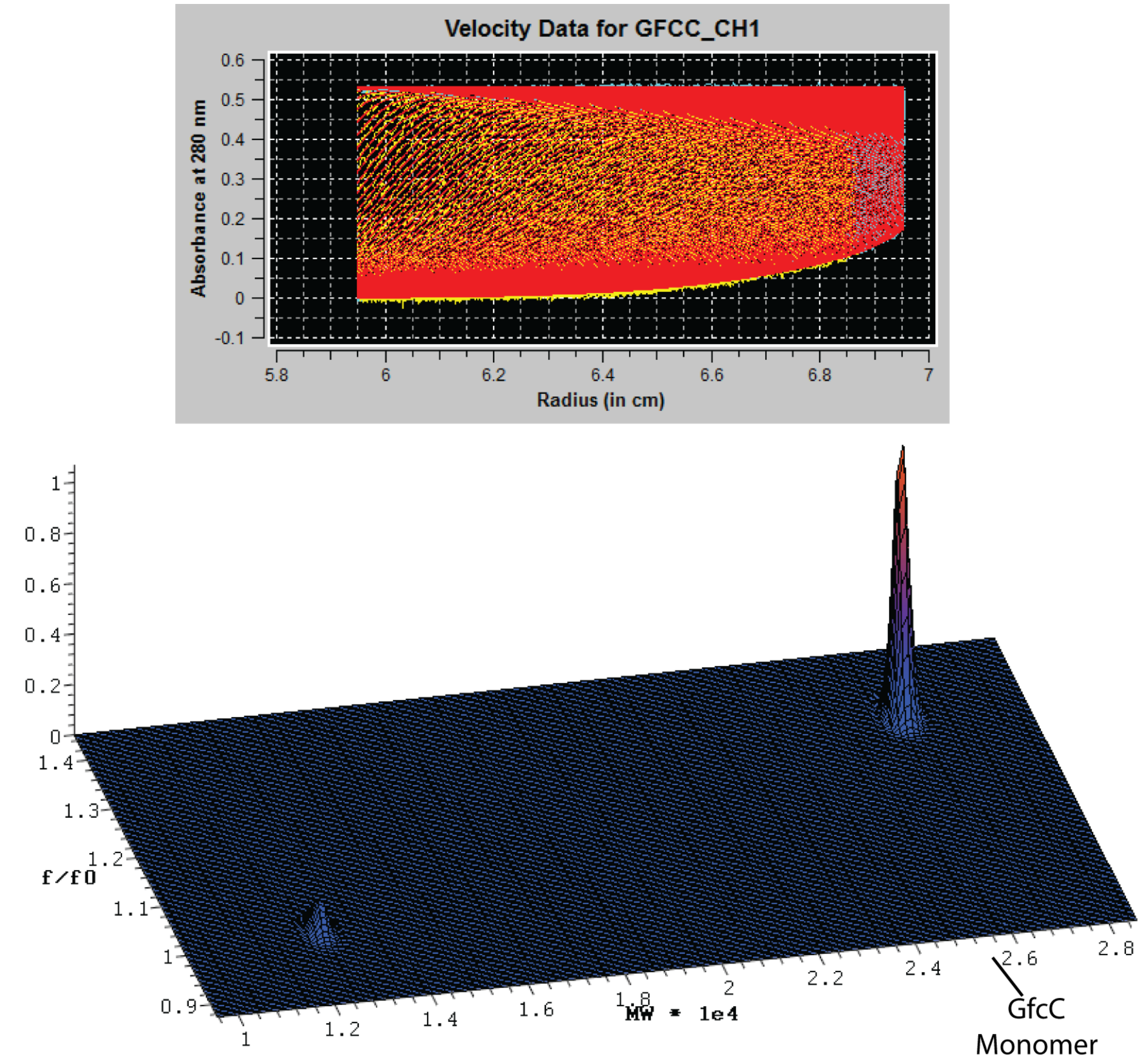
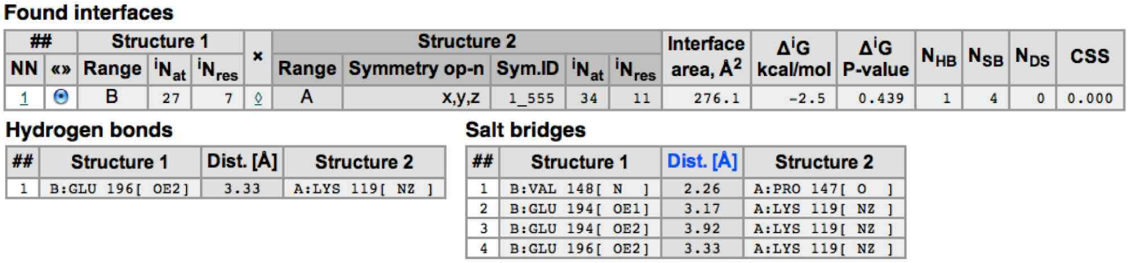


Figure 2-13 **Analytical Ultracentrifugation** experiment of purified GfcC (33–248)-His$_6$ at 1 OD (*Abs 280nm*) showing monomer in solution. (Thanks to Titus Franzmann who did the data analysis)

2.2.8 Interface between the two β-grasp domains D2 and D3 is minimal

The interface between β-grasp domains 2 and 3 of GfcC is minimal with a buried surface area of only ~280 Å^2 involving only one hydrogen bond. The B-factors for the residues in this region are also low suggesting a less dynamic interface. The hydrogen bond is between Glu 196 from domain 3 with Lys 119 from domain 2.

GfcC's Interface between the two β-grasp domains



Found interfaces

##	Structure 1				×	Structure 2					Interface	ΔⁱG	ΔⁱG	N_{HB}	N_{SB}	N_{DS}	CSS
NN	«»	Range	$^iN_{at}$	$^iN_{res}$		Range	Symmetry op-n	Sym.ID	$^iN_{at}$	$^iN_{res}$	area, Å^2	kcal/mol	P-value				
1	●	B	27	7	◊	A	x,y,z	1_555	34	11	276.1	-2.5	0.439	1	4	0	0.000

Hydrogen bonds

##	Structure 1	Dist. [Å]	Structure 2
1	B:GLU 196[OE2]	3.33	A:LYS 119[NZ]

Salt bridges

##	Structure 1	Dist. [Å]	Structure 2
1	B:VAL 148[N]	2.26	A:PRO 147[O]
2	B:GLU 194[OE1]	3.17	A:LYS 119[NZ]
3	B:GLU 194[OE2]	3.92	A:LYS 119[NZ]
4	B:GLU 196[OE2]	3.33	A:LYS 119[NZ]

Structure 1 is GfcC β-grasp domain 3 (Residues 148-226)
Structure 2 is GfcC β-grasp domain 2 (Residues 23-67 and 116-147)

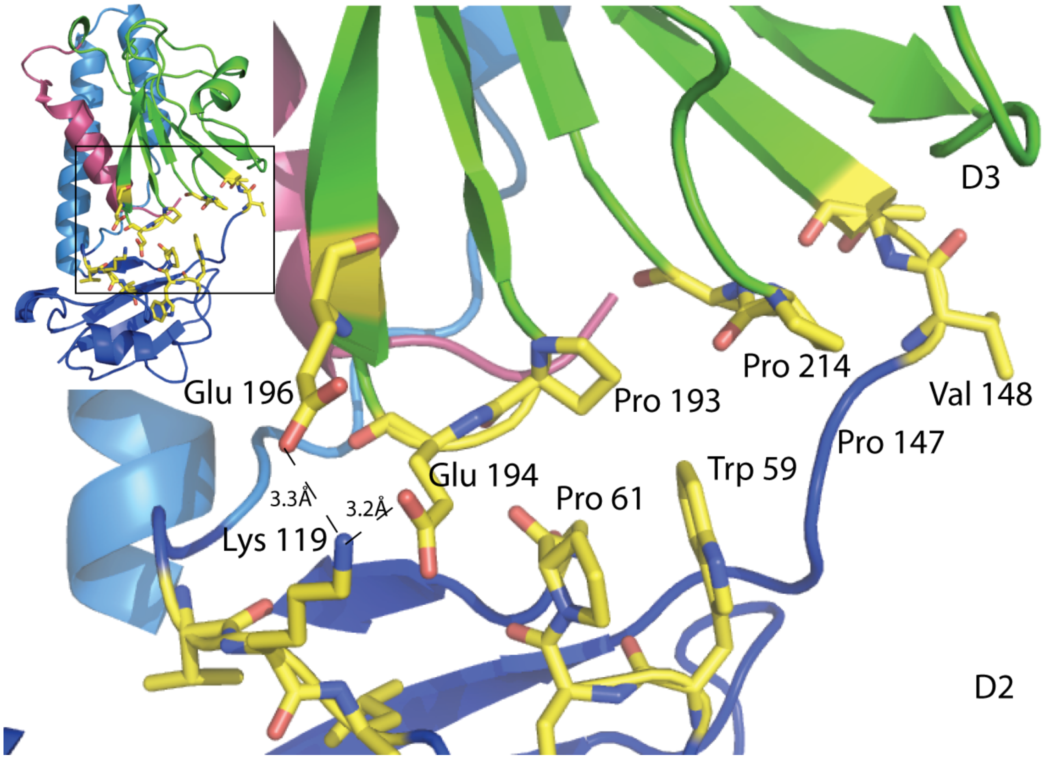


Figure 2-14 **The interface between D2 and D3 domains of GfcC**. The interface is predominantly non-polar with a buried surface of 280 Å^2. (PISA analysis, *top*) There is a weak hydrogen bond between surface exposed lysine 119 and glutamates 194 and 196 (*bottom*).

2.2.9 Interface between D2H, D4 and D3 domains form the bulk of interdomain interactions

The interfaces between D4 and D3 and between D4 and D2H are the most extensive among all four inter-domain interactions. Both the interfaces are predominantly non-polar but also have hydrogen bonds between residues that are specifically highly conserved for the ones in the D4 and D3 interface. The hydrogen-bonded interactions here are Asn235 with the backbone of Phe222 at the top and Arg245 with Gln217 at the bottom of the C-terminal helix respectively stapling the two domains D4 and D3 together. The residues inbetween the two hydrogen bonding partners are filled with residues involved in Van der Waals interactions, for instance between sidechains of Ile238 and Trp219, Leu242 and Leu153, Val239 and Mse189.

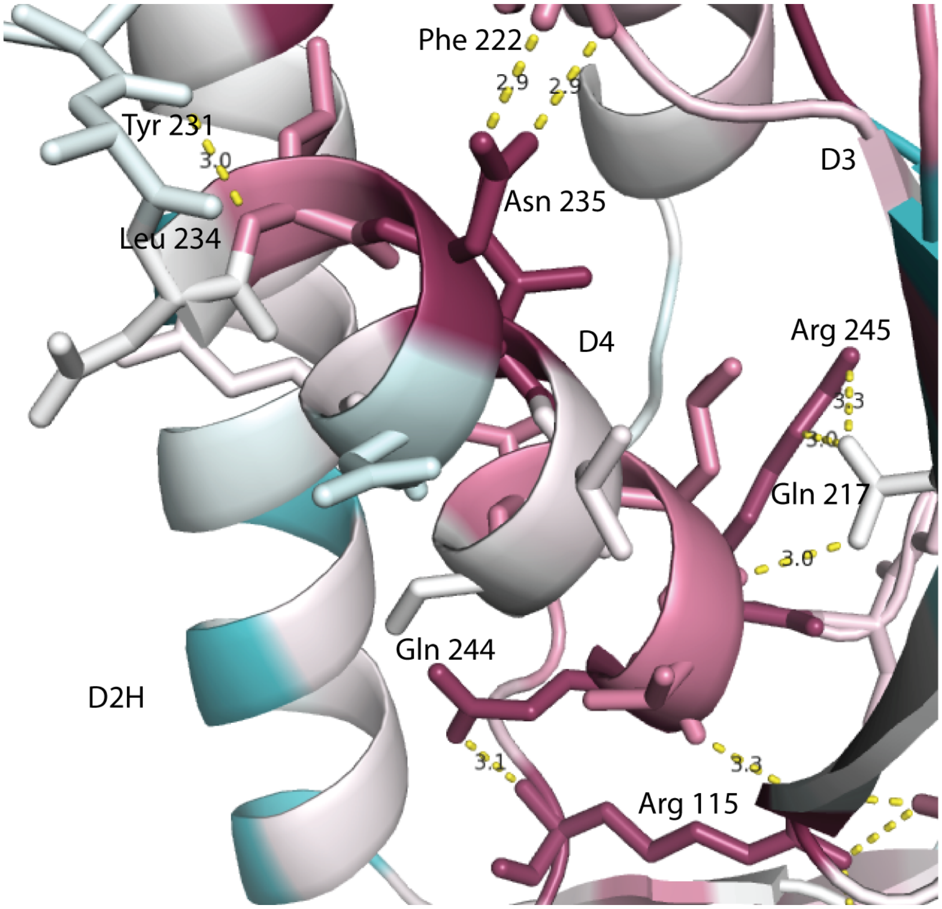


Figure 2-15 **Interaction made by C-terminal helix of D4 with D3**. The hydrogen bonded interactions and their distance are shown in yellow. The structure is colored according to conservation scores computed from Consurf with magenta showing the highly conserved residues and cyan the lowest.

2.2.10 Pockets near conserved residues in GfcC

The interior of GfcC also shows several prominent cavities (from CASTp analysis [54]). The largest ones in terms of volume (≈ 400 Å^3) include a water tunnel between the α-helical hairpin domain (D2H) and the β-sheet of D3 that is continuous with bulk water at either end. The tunnel contains seven water molecules in chain A of the P2$_1$ crystal structure. The other chains have either two or more water molecules absent reflecting the plasticity of this interaction between D2H and D4 domains.

The largest pocket after the water tunnel rests in D2 at the interface with D3 and end of α-helix from D4 (Pocket 1 in figure 2-16, ≈280 Å^3 in volume[54]). The residues surrounding the pocket include highly conserved Arg 115 from the end of D2H domain before it joins back to D2 and residues 116-121 surrounding it forming the outer lip of the surface of the pocket and Glu 194 from D3 at the top. Residues 60–65 contributed by a 3$_{10}$ helix and S3 from D2 line the backside of this pocket. All residues with side chains exposed to the pocket are conserved except for Val118 and Lys119. The volume of the pocket is filled with atleast five ordered water molecules (waters 72, 74, 164, 427 and 570) and has ≈10 hydrogen bond donors and acceptors from the main chain that can bind to a potential ligand binding to this pocket. The side chains of Arg 115, Lys 119 and Asp 121 are also available for hydrogen bonding to potential ligand that may bind in this pocket. The pocket is similar in size to a sugar molecule like galactose (one of the group 4 polysaccharide repeat units) that can be manually aligned well inside the pocket but attempts to fit it with Autodock Vina [55], an automated docking program and Dock5 [56] with a defined binding site grid were unsuccessful.

Secondly, a flatter pocket with a volume of ≈100 Å^3 (CASTp [54]) is located between D4 and β-sheet of D3 (Pocket 2 in figure 2-16) close to a highly conserved surface Trp205. The edge of this pocket is lined by residues Asp236, Asn235, Phe222 (main chain), Asp184, Asn187, Val188 (main chain), and Met189. The back side of pocket includes ring imide of Trp219, main chain of Leu220, Gly221 and Val239. Of these, Asp184 and Asn235 on one half of the pocket (and the main chain) are the most conserved.

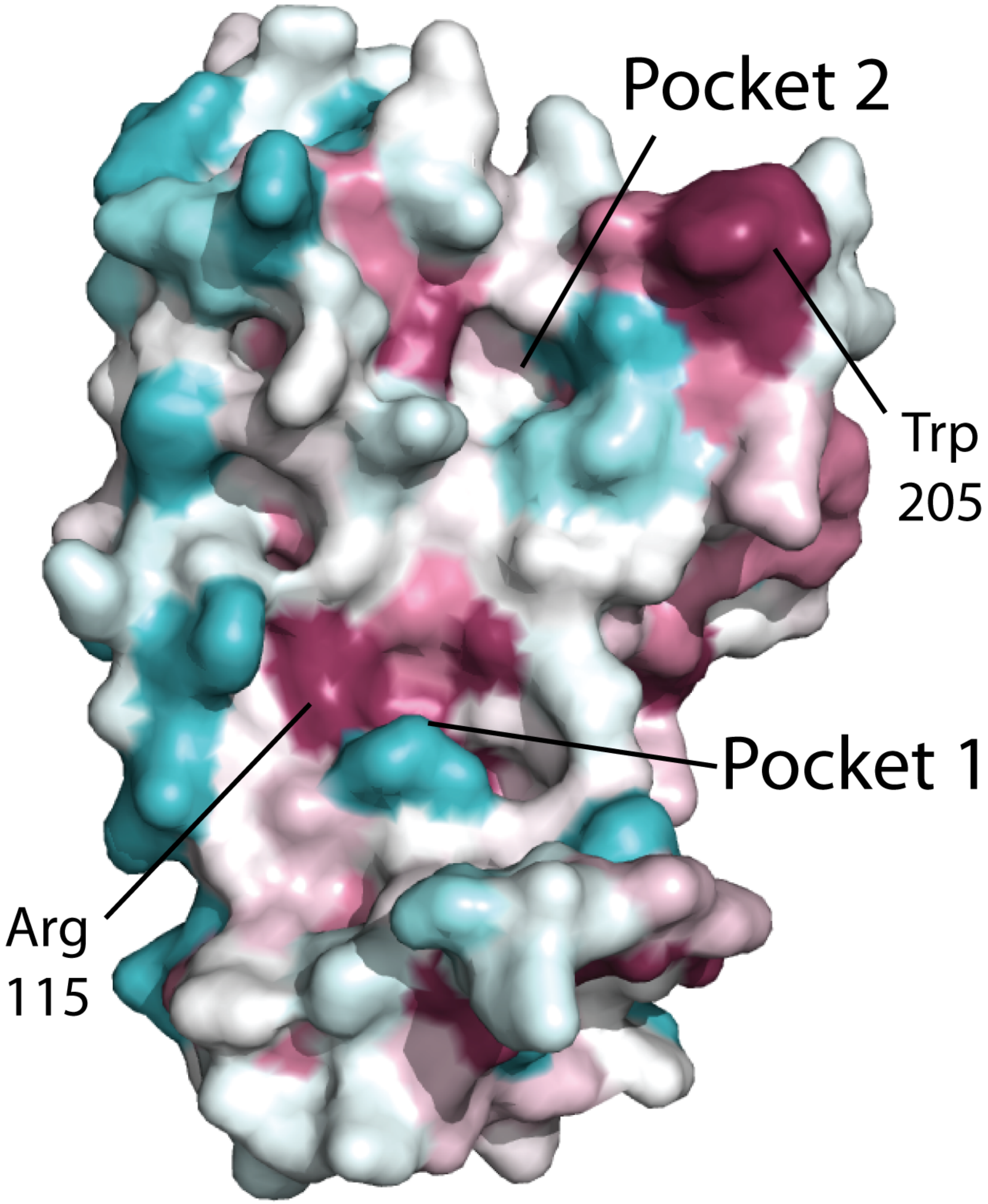


Figure 2-16 **Surface of GfcC colored according to Consurf conservation scores** [53]. Pointers show the location of Pocket 1 and Pocket 2 and the invariant Arg115 and Trp205.

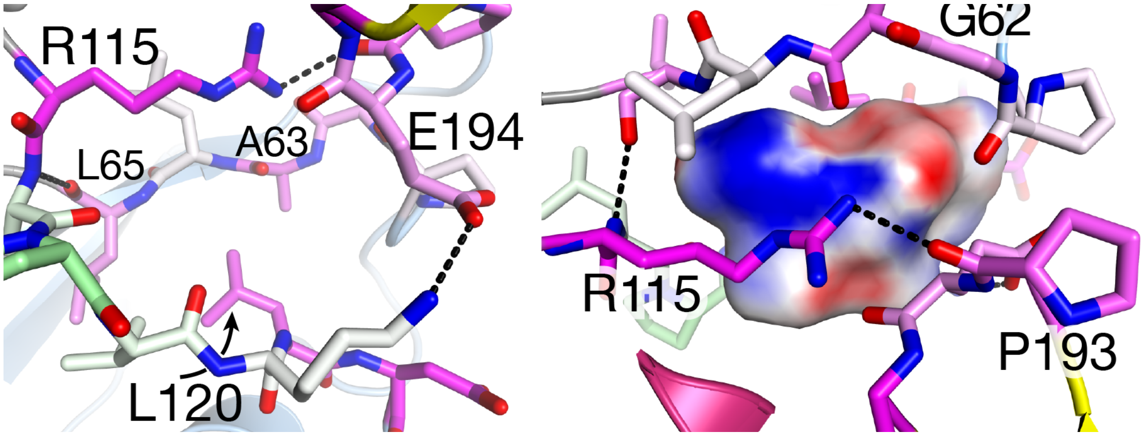


Figure 2-17 **Residues surrounding Pocket 1** *(left and right)* colored according to their conservation scores. The coloring included magenta for most conserved and black lines indicate polar contacts. *Right*: the inverse surface is shown with red and blue on surface colored according to electrostatic potential, blue for nitrogen or hydrogen bond acceptor face and red for backbone oxygen facing or the hydrogen bond donor face.

2.3 Discussion

GfcC has four domains defined (D2, D2H, D3 and D4) with three of them similar to the domains in monomeric Wza. GfcC has two β-grasp domains (D2 and D3) like in Wza and an amphipathic C-terminal helix (D4) albeit shorter and in a different conformation from Wza. The β-grasp domains in Wza (and GfcE), are predicted to be present in the entire family of outer membrane polysaccharide export (OPX) proteins [3]. It is therefore intriguing that a soluble protein like GfcC encoded upstream of *gfcE* also has the exact type of fold.

The function of lipoprotein Wza (or GfcE) as an export protein comes from the oligomeric state that brings eight identical monomers of Wza to form an octameric channel. GfcC has a shorter C-terminal amphipathic helix (233–245, 13 residues) than its Wza counterpart (345–376, 31 residues) and likely cannot traverse the membrane completely. The helix is also seen here in a conformation that attests to the protein's monomeric state in solution. In spite of observing higher order states like dimer in the asymmetric unit of the both crystal forms with certain contacts between conserved residues, the experimental methods (gel filtration chromatography and analytical ultracentrifugation) that have been used to investigate oligomeric state at possibly low concentrations (~1–3 mg/ml) all indicate GfcC to be a monomer in solution. In addition, the periplasmic location of this protein (see Appendix) suggests a completely different role for the β-grasp domains in GfcC compared to GfcE (homolog of Wza).

Wza has been established as the exit portal for the growing capsular polysaccharide in group 1 but nothing is known about how the polysaccharide enters the α-helical channel of Wza (Figure 2-18). The high-resolution crystal structure of Wza at 2.26 Å shows the octamer is closed from the periplasmic side and its interaction with Wzc (EM structure of Wza-Wzc known) likely gates the channel [16]. This explains the opening of the channel but the picture on how the growing polysaccharide threads through a stable octameric complex such as Wza still remains murky (Figure 2-18). One of the postulates for the entry is the opening up in the Ring2 (formed by one of the β-grasp domains, D2) in Wza. Theoretical free energy calculations and the evidence of the Wza structure opening up as seen by the bulge in the EM map for Wza-Wzc complex offer the only unlikely scenario for polysaccharide import into the channel [16, 57].

The adapter role of other protein components mediating the entry of growing polysaccharide into the core of Wza is another intriguing possibility. The periplasmic GfcC offers the possibility of serving as that adapter molecule. Similarities from drug efflux complexes (TolC-MexA-AcrB) that also span both membranes [58] suggest the role of adapter proteins (in this case, MexA) to link outer membrane components with the inner membrane protein.

Although Wza-Wzc (or GfcE-Etk in an analogous fashion) forms an independent membrane-spanning complex [16], it does not comprehensively describe all aspects of polysaccharide export. In particular this model does not address how the inner membrane polymerase Wzy and flippase Wzx fit in to serve the role of this coupled polymerization and export of the growing polysaccharide.

Further, the oligomeric state of Wzc is itself unclear. In the EM structures of Wza-Wzc complex [16] and Wzc alone [25], Wzc is modeled as a tetramer. However, the recent structure of non-phosphorylated cytoplasmic portion of Wzc (catalytic domain) represents an octamer whose dimensions (120 Å across) correspond closely to the width of Wza (104 Å) [26]. The location of isolated Wzc catalytic domain in the EM map (with Wzc tetramer) also fails to explain how Wzc protomers can interphophorylate. There may be other proteins and/or its relation to Wzy (polymerase) and Wzx (flippase) that offers an explanation. It is in this context the functional role of the four proteins GfcA, GfcB, GfcC, and GfcD can shed more light in understanding the mechanism of group 4 polysaccharide export and its regulation.

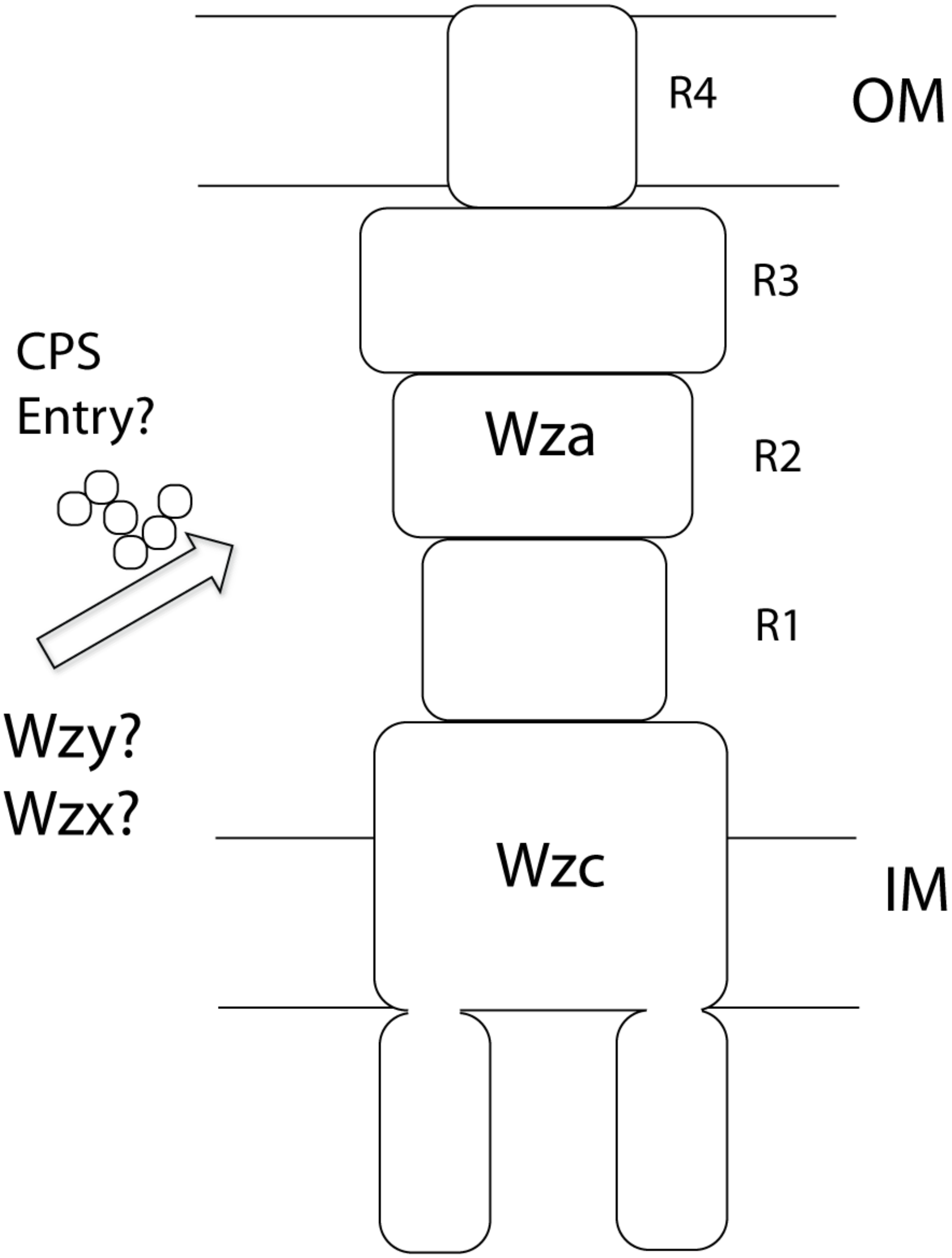


Figure 2-18 **Entry of polysaccharide capsule into Wza-Wzc is not clear**. The cartoon representation of Wzc and its orientation here with Wza is based on published EM structure of Wza-Wzc complex[16]

The surface of the large pocket seen in GfcC close to conserved Arg115 provides atleast ten hydrogen bond acceptors and donors from the main chain. This points to the possibility of GfcC being a binding protein (we managed to manually dock a Galactose molecule–a group 4 oligosacchride repeat unit precursor) but there is no experimental premise for this claim. Also a protein performing such a binding role for the oligosaccharide moiety in periplasm is not known in other capsule polysaccharide export systems. In a similar attitude the pocket may also serve to anchor or bind other proteins and polypeptide segments. The surface conserved residues in the pocket clearly suggest such a function is not improbable. Further, considering that the observed dimer interface residues are conserved but all experimental methods pointing GfcC to be monomer may suggest the interface actually capable of interacting with proteins with similar domains (for instance, GfcE). The dimer interface is identical in both the crystal forms affirming its propensity for interaction.

The α-helical outer membrane pore observed for Wza were novel not only for polysaccharide export proteins but also for any known class of bacterial outer membrane proteins until very recently [15, 59, 60]. The presence of Wza-like OMX proteins in numerous polysaccharide systems including group 2 and 3 capsule suggests that the helical spanning pore is indeed common. Although not possible for GfcC other proteins like WbfF that has similar DUF1017 domains as GfcC has predicted longer α-helical C-terminal region that may traverse the membrane like Wza. However, these lack the equivalent domain for interacting with Wzc or its equivalent. There are also other export systems are known such as in several exopolysaccharide export systems like alginate, poly-β-GlcNAc and cellulose, the outer membrane export protein is AlgE, PgaA and BcsC respectively and is a large β-barrel protein with up to 18 strands [31, 61]. The predicted structure of the outer membrane lipoprotein GfcD, the largest of the proteins encoded by *gfcABCD* set of genes with unknown protein function, is predicted to be an outer membrane β-barrel with about 24 strands plus a possible periplasmic domain [62] (see Figure 2-18). The large β-barrel architecture is more common for outer membrane proteins than the α-helical pore seen in Wza and may give a channel wide enough to pass branched polysaccharide chains.

This however does not explain the regulation of export through cycling of the phosphorylation/dephorylation states of inner membrane kinase (Etk). Herein comes the role of adapter proteins that link outer membrane export proteins with the inner membrane protein components. In the alginate polysaccharide assembly the lipoprotein AlgK with its TPR motif possibly serves as the adapter molecule between the outer membrane β-barrel protein AlgE and inner membrane biosynthetic components. Lipoprotein AlgK is also required for the proper localization of AlgE [31]. This situation may be analogous to GfcC and GfcD in the group 4 polysaccharide capsule assembly.

There is also evidence of fusion of *algK* and *algE* homologous region as a single gene in *pgaA* and *bcsC* of the PGA and cellulose operons, where the TPR domain followed by a β-barrel porin type domain [31]. GfcC and GfcD may thus be analogous to AlgK/AlgE system. There are several proteins, sometimes named OtnG, that have *gfcC* and *gfcD* gene homologs fused as one protein. These occur in several species of betaproteobacteria: *Burkholderia, Leptothrix, Variovorax, Thauera sp., and Gallionella* and one gammabacterium from a deep sea vent, *Idiomarina loihiensis.* The protein (IL0568) from this latter organism is 27% identical to GfcC and 46% identical to GfcD and annotated as fusion of these two domain families (DUF940 and DUF1017). More importantly, all these genes are encoded in operons found adjacent to *gfcB* homologs and in some cases with *gfcA*. The IL0568 locus resides within the region designated for exopolysaccharide synthesis and export [63]. By analogy to AlgK/AlgE this suggests that GfcC and GfcD interact *in vivo*. Immediate future experiments will investigate this possibility.

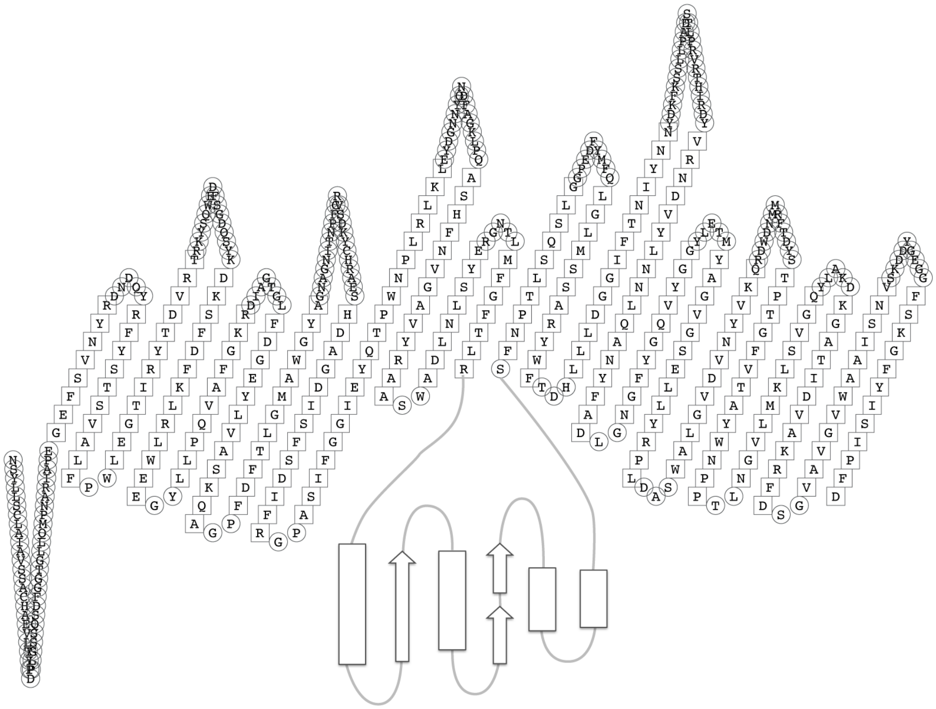


Figure 2-19 **HHomp prediction [62] for GfcD showing the membrane traversing segments in the predicted protein**. There are about 24 strands predicted capable of forming a large beta-barrel type structure. There is an insert halfway, between strand 12 and 13 that has less propensity for membrane insertion. The predicted secondary structure for this insert is shown as in this composite figure.

The initial discovery of group 4 operon with the seven contiguous genes *gfcABCDE, etk and etp* placed the importance of these in the export of group 4 polysaccharide capsule [4]. Comparing directly with the group 1 system with which it shares many characteristics also brings several differences between both these capsule systems. GfcA, GfcB, GfcC and GfcD are novel to group 4 as is Wzi to group 1 suggesting that either the four different Gfc proteins could serve to anchor the group 4 polysaccharide functioning similar to Wzi or the larger GfcD with the β-barrel architecture could serve as an alternative exit route for the capsule functioning similar to exopolysaccharide systems like the alginate. Further investigation into GfcD should throw more light on the function of GfcC and the Gfc proteins as a whole.

Chapter 3

YraM, a lipoprotein identified essential for growth of *Haemophilus influenzae* through whole genome studies and structures of its two individual domains

3.1 Introduction

3.1.1 The capsular and acapsular types of *Haemophilus influenzae*

Haemophilus influenzae is a Gram-negative rod-shaped human pathogen. There are two major strains based on the presence or absence of an outer protective polysaccharide shell called capsule. The presence of the capsule makes this microbe more virulent and the disease caused by the capsular variant of *H.influenzae* (for instance, *H.influenzae* type b)– a severe form of bacterial meningitis– can be life threatening from its onset. On the other hand, the acapsular and hence non-typeable serotype of *H.influenzae* causes primarily diseases of the mucosal membrane in children causing such diseases as middle ear infections (otitis media), eye infection and sinusitis affecting ~75% of children over the world at some point in their childhood. It also exacerbates the condition of chronic obstructive pulmonary disorder (COPD) patients being an opportunistic pathogen. The capsular variant of this microbe does however have an effective heat-treated conjugate vaccine that is protective, the non-typeable form still needs treatment and no protective measures exist so far [64].

3.1.2 Identifying essential genes in whole genomes

In-vitro transposon mutagenesis (by Akerley et.al., [65]) on the non-encapsulated *H.influenzae* Rd KW20 strain identified 478 open reading frames (covering 38% of the genome) that are essential for its growth and viability on rich media (sBHI agar at 37^{o}C). Of these genes 259 were annotated as coding for protein of unknown function and 159 of them were common to all bacteria (perhaps genes for basic cellular functions). The segments that tolerated frequent insertions were considered to harbor mostly nonessential genes and those that did not tolerate mutations were having putative essential genes [65]. One of the essential genes of unknown function determined in this fashion that was studied further was *yraM* due to it being a predicted lipoprotein sorted to the outer membrane and its essentiality for viability of the organism. These features made this gene product a possible target to developing antimicrobials.

3.1.3 *yraM* (Hi 1655) is an essential gene in *H.influenzae*

There are many criteria to defining a successful antimicrobial protein target for a given microorganism. One important criteria is to be essential for the microbe so as to prevent antigenic drift and to be selective for that microbe and not interfere with the host organism. The target present in the outer membrane simplifies this further so that the antimicrobial need not traverse the inner membrane to exert its function. One essential gene that fits these criteria is *yraM* that encodes a 575-residue lipoprotein identified as such by the conserved –LAG**C**S– lipobox sequence. Membrane fractionation and LC/MS analysis of *E.coli* proteome showed YraM to be localized in the outer membrane [66] whereas *H.influenzae* YraM is deduced to be outer membrane due to Ser following the lipidated Cys residue than Asp.

Searching for other proteins with sequence homology (PSI-BLAST) suggested that YraM has two independently folded regions, the smaller N-domain (residues 33–251) and the larger C-domain (residues 257–575). The individual domains were expressed and crystallized [64] (Vijayalakshmi. J and Saper. M unpublished). The N-domain consists of 34 residue tetratricopeptide repeats (TPR) that form helix-turn-helix motif. There are about 7.5 TPR repeats that form the N domain. TPR domains are seen in other proteins that perform varied functions as in chaperones, cell cycle proteins and so on[38]. In these proteins TPR are seen mediating protein-protein interactions. Extended TPR motifs can give a superhelical twist generating a convex side and concave side for TPR proteins. This twist is apparent in the structure of N domain (Vijayalakshmi. J and Saper. M, unpublished). The N-domain also has very few conserved residues on its surface making identifying potential binding sites difficult.

The 1.35 Å structure of the C-domain of YraM [64] is identical to a type I periplasmic binding protein fold with two nucleotide binding folds linked by a three strand cross-over. The C-domain closely resembles the periplasmic leucine/isoleucine/valine binding protein (LIV-BP) of *E. coli* in its open conformation characteristic of a substrate-free protein. The open seashell shaped structure has an amphipathic cleft that when modeled in the closed form based on leucine bound LBP or LIV-BP has dimensions of 16 Å long, 8 Å wide and 8 Å deep pocket that can potentially bind as yet unknown ligand. The region corresponding to the LIV-BP binding cleft or pocket is highly conserved amongst YraM homologs suggesting that this YraM domain may have the function of binding another molecule.

3.2 Sequence analysis and structure of YraM N- and C- domain determined individually

3.2.1 Sequence alignment of YraM orthologs

The sequences were obtained from NCBI (National Centre for Biotechnology Information) database corresponding to those that were compared earlier with the YraM C-domain as published [64]. PSI-Blast analysis of YraM (HI 1655) sequence from *Haemophilus influenzae* Rd KW20 strain with the non-redundant protein sequence library indicated presence of two different domains[64]. The amino terminal region (residues 33-251) comprised a TPR domain and carboxy-terminal domain (257-575) showed low (10–16% identity) but significant similarity to periplasmic binding type proteins (PBP) required for small molecule transport. These included *E.coli* leucine/isoleucine/valine-binding protein (LIVBP, 11% identity) whose structures have been solved with and without ligand [64, 67, 68]. There is also a linker region (residue 250-256) identified in YraM that do not belong to both these domains (Figure 3-3)

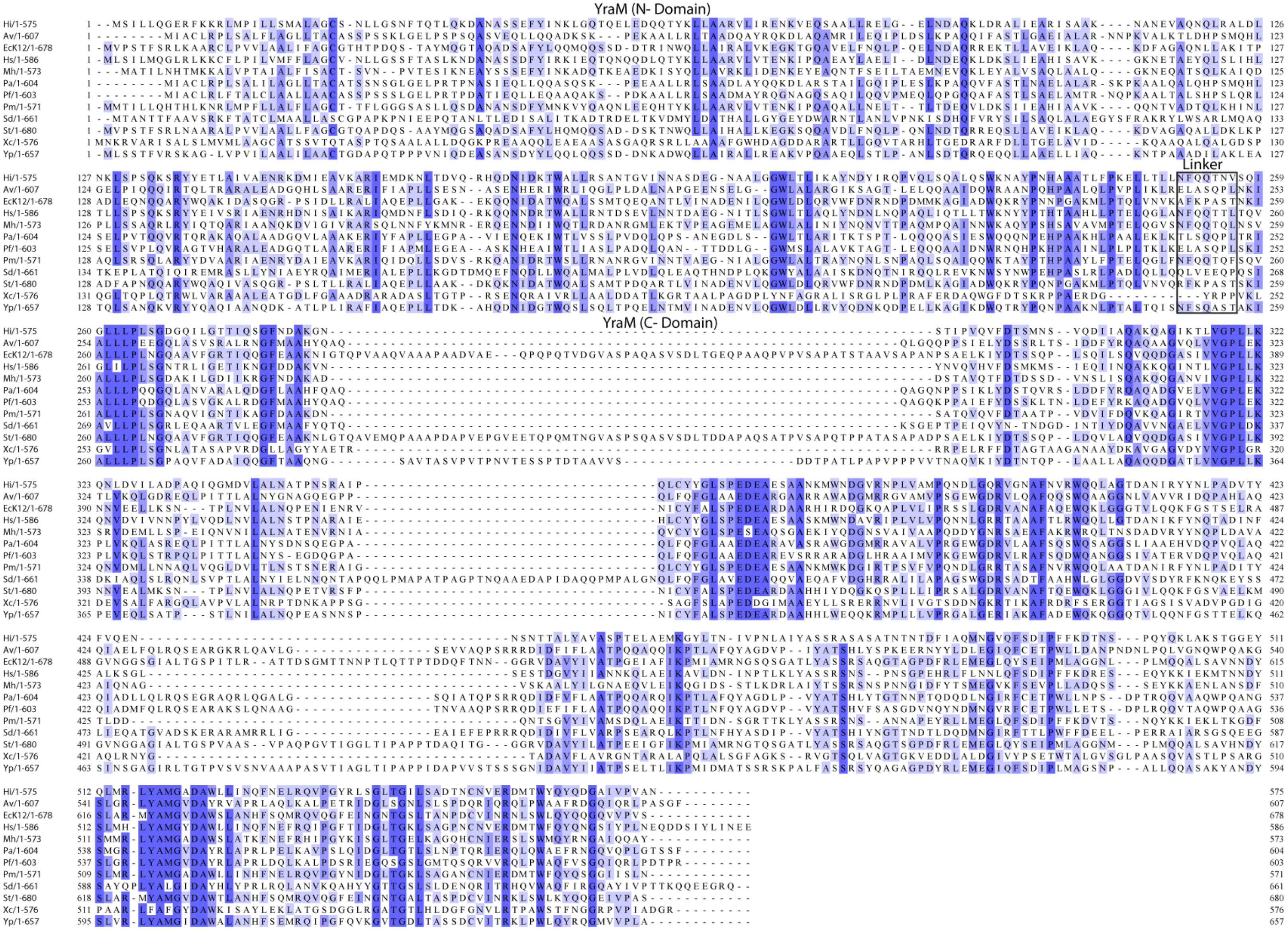

YraM (N- Domain)
Linker
YraM (C- Domain)
Hi/1-575
Av/1-607
EcK12/1-678
Hs/1-586
Mh/1-573
Pa/1-604
Pf/1-603
Pm/1-571
Sd/1-661
St/1-680
Xc/1-576
Yp/1-657

Figure 3-1 **Sequence alignment of YraM homologs from 12 representative genomes**. The sequences are same as those used in the earlier publication of C-domain [64] to keep the comparison between C-domain in individual structure and full-length structure on same scale as will be described later (in text). The sequences are colored by percentage identity using Jalview [49]. The dark blue represents residues with >80% identity, light blue >60% and the lightest blue represent >40 %. The residues in white are <40% identical. This alignment was also part of the input to calculate the full-length YraM conservation scores using Scorecons[69]. It can also be seen the beginning of C-domain has large inserts in other homologs of YraM. The sequences, protein name and genbank id (gi) for the sequences used in this alignment are as follows: Hi, *Haemophilus influenzae Rd KW20*, HI1655 (gi 16273542); Hs, *Haemophilus somnus*, LppC (gi 4096758), Pm, Pasteurella multocida, LppC (gi 15602511); Mh, *Mannheimia haemolytica*, GS60 antigen (gi 62798901); Yp, *Yersinia pestis*, YP03548 (gi 16123692); EcK12, *Escherichia coli K12*, YraM (gi 7466039); St, *Salmonella typhimurium* LT2, YraM (gi 16766562); Av, *Azotobacter vinelandii*, LppC (gi 67086486); Pf, Pseudomonas fluorescens, LppC (gi 77384908); Pa, *Pseudomonas aeruginosa* (gi 9950656); Sd, *Saccharophagus degradans* 2-40 (gi 90022787); Xc, *Xanthomonas campestris*, XCC0711 (gi 21230186).

3.2.2 Structure of the N-domain of YraM (33-249)

The N domain of YraM (33-249) had been solved earlier in our lab (J.Vijayalakshmi, unpublished) and it has about 7.5 tetratricopeptide repeat or TPR units. The N-domain of the full length YraM resembled closely the isolated structure of N domain alone. The differences in resolving certain residues and the superposition are described later. TPR's are the most versatile of the all α-helical fold. Each TPR is a 34-residue (helix-turn-helix) repeat as denoted by its name and was first identified and named in 1990. One of the first structures solved that had the TPR motif was that of protein phosphatase 5 (PP5) having three TPR motifs [38]. TPR's can also be recognized by just sequence comparisons with identifying their high consensus sequence that comprises small and large hydrophobic residues at specific locations in the 34 residue stretch. The consensus sequence reads W4-L7-G8-Y11:A20-F24-A27-P32 with some residues (e.g, Gly or Ala at position 8, and Ala at positions 20 and 27) conserved more than the others [38]. The turn between the two helices (indicated before by ':') has helix breakers and residue preference here can affect the superhelical twist that is dictated by this helix-turn-helix conformation, particularly pronounced in longer TPR domains.

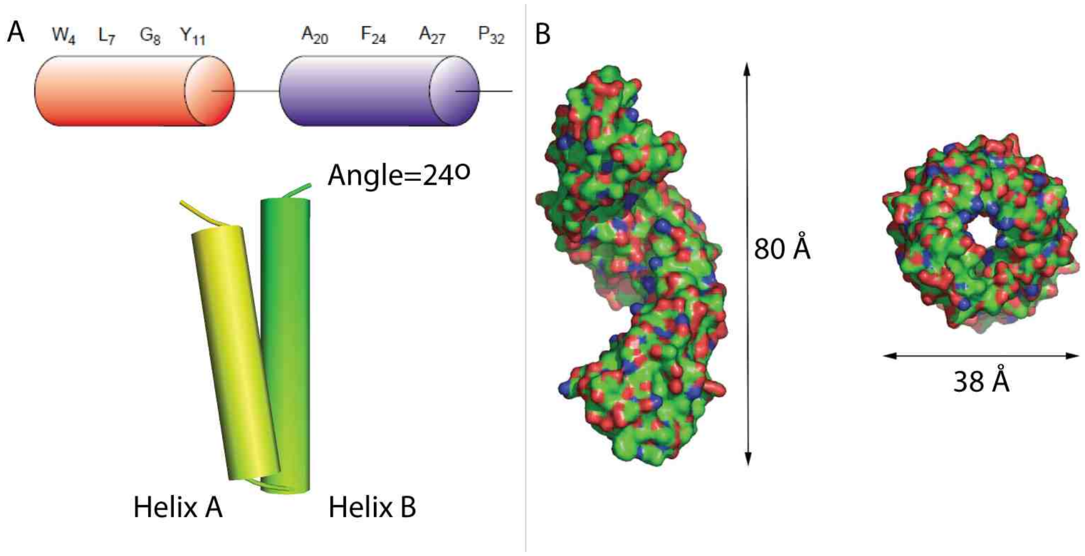


Figure 3-2 **The general consensus for TPR domain**. Left: A. The TPR consensus sequence is shown at the top and the angle between the helices is usually ~24° as shown below. *Right*: B. Extended TPR motifs induce a superhelical twist with a convex and concave side. Portions of the figure published in [70]. Copyright 2003 Elsevier Ltd.

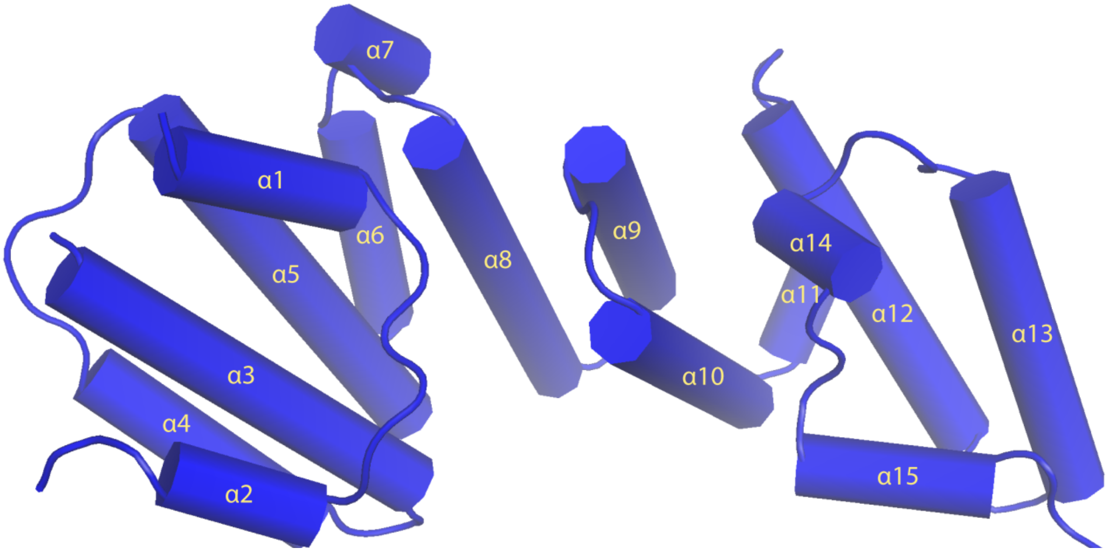


Figure 3-3 **Structure of N domain of YraM (33-249)** (Vijayalakshmi. J and Saper. M, unpublished). The figure shows the convex and concave sides for TPR domains visible. There are ~7.5 helix-turn-helix (TPR-like) motifs.

The 7.5 extended helix-turn-helix TPR-like units in YraM (33-249) are long enough to exhibit the superhelical twist due to helix arrangement and result in a concave and convex surface. In general residue conservation outside of the consensus gives an indication of residues functionally important. TPR domains are well known in mediating contacts with other proteins (protein-protein interactions). In the case of YraM (33-249) surface conserved residues were limited. Scorecons [69] was used to calculate the residue conservation score for each position based sequence alignment and the B-factor field of pdb was modified with these scores. The highly conserved residues or those that scored a fractional score greater than 0.8 were isolated to the upper convex side of the protein close to its C-terminal end. The residues in this conserved block includes unusually a trio of tryptophan residues within 7 Å of each other (Trp 179, 204 and 228) each contributed by three adjacent α-helices. The cluster also includes Leu 207 that is surrounded by the three tryptophans and Leu 245. There is also conserved His235 and Pro232. Pro 233 is in the preceding turn that carries the His residue and is surface exposed. These conserved Trp residues can be understood from the canonical conserved TPR motif sequence (Figure 3-3 A) but it should be noted that there are no tryptophans or any other residues conserved (score >0.8) in the other helices of the N-domain (TPR domain).

The structure of N domain of YraM closely resembles PilF with 13 TPR units [71] from type IV pilus biogenesis in *P.aeroginosa* and is essential for proper assembly of pilin. PilF is part a multiprotein complex PilD/F/G/T at the base in inner membrane that assemble mature pilins. Like YraM, PilF is also a lipoprotein [71].

3.2.3 Structure of C-domain of YraM (257-575)

The structure of C-domain has been individually determined and published by our lab earlier [64]. It has two Rossmann fold motifs linked by a three strand cross-over (Type I PBP-like fold). Each Rossmann fold is a parallel β-sheet lined on both sides with more than one α-helix. The full-length structure of YraM (see below) shows similar C-domain (257-573) open conformation as seen individually for the truncated C-domain protein structure [64] (Figure 3-5, 3-8 and 3-10). The conserved residues (from Scorecons calculation) were all-distinct in sequence numbers but structurally clustered between the two halves of the PBP fold indicating a binding role for C domain of YraM. The ligand is as yet unidentified but from the residues lining the surface of the pocket it can be discerned that YraM binds an amphipathic molecule. The C domain resembles closely the leucine bound LIV-BP in its open conformation. Modeling the closed form of YraM based on LIV-BP gives a pocket that is 20 Å long and 18 Å wide [64]. The residues lining the pocket are shown in figure 3-7 below.

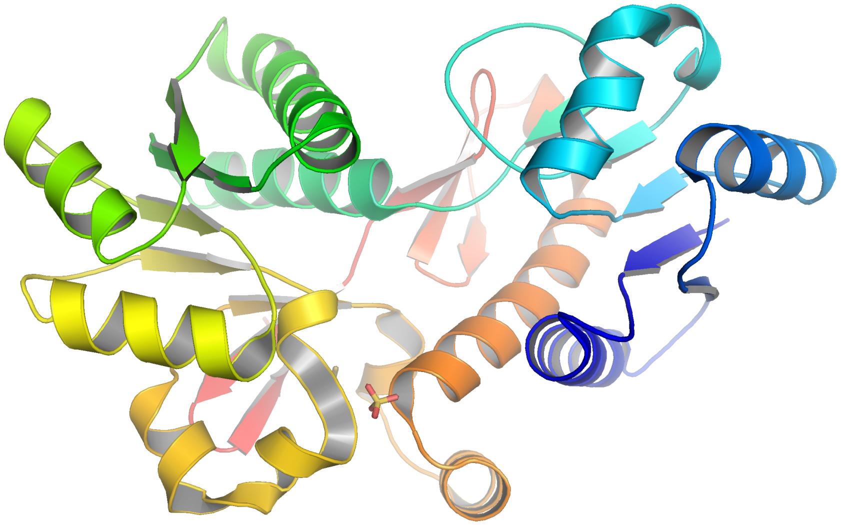


Figure 3-4 **Structure of YraM-C** showing the two Rossmann fold domains connected by a three strand crossover (type I PBP like) [64]. PDB ID: 3CKM

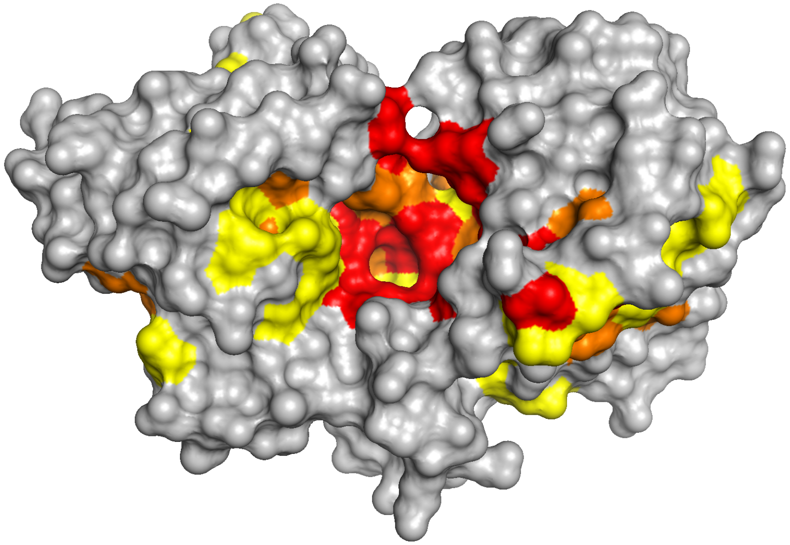


Figure 3-5 **Conserved residues cluster in the cleft between the two halves of YraM-C**. Red are highly conserved (>90 percentile), orange (80-90 percentile) and yellow (70-80 percentile); grey surface is the molecular surface drawn by PyMol. [64]. PDB ID: 3CKM

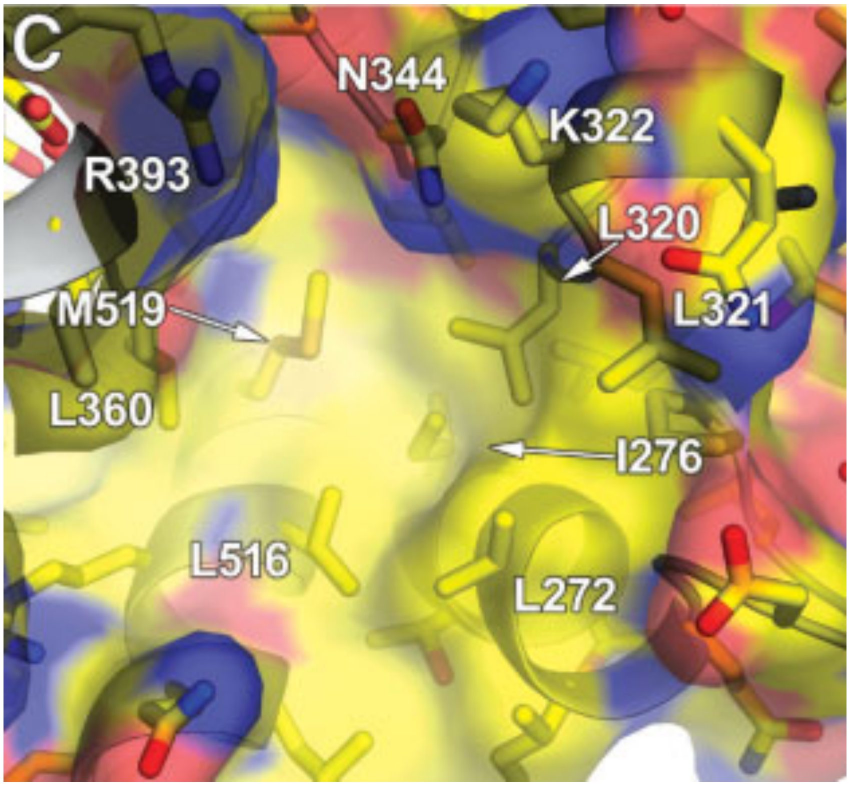


Figure 3-6 **Residues in the conserved binding cleft in YraM C- domain** [64]. The lower right half include hydrophobic aminoacids like Leu272, Ile276, Leu320, Leu321, Leu360, Leu 516 and Met 519 while the upper left include charged aminoacids Arg 393, Asn 344 and Lys 322. Figure published in reference [64]. Copyright 2008, Wiley-Liss, Inc.

The individual folds of the N- and C-domain of YraM have been observed in number of other proteins [38, 67, 72, 73]. However, the fusion of these two domains in YraM is unique. Considering the potential for ligand binding and periplasmic location, a full-length structure of YraM will surely be informative for future studies on this molecule in the search for potential ligands or antimicrobials.

Chapter 4

Structure of full length lipoprotein YraM from *Haemophilus influenzae*

The full length structure of lipoprotein YraM was thought to be key to understanding not only the orientation of the two domains but to also possibly capture the protein with a bound ligand in the C-domain. Simultaneous efforts were made to crystallize the full length protein along with experiments setup for the N-domain and C-domain crystallization (J.Vijayalakshmi). However, the full-length protein with the his$_6$-tag was not intact and disintegrated soon after purification as revealed by SDS-PAGE. The reason for this is still not understood. This chapter continues on the previous work where the full length construct was modified to express the YraM protein without the his$_6$-tag (B.Tirupati). This modification was subsequently successful in preventing the protein from degrading and in producing a diffraction quality crystal. The structure of the full length YraM (named as YraM-FL) and implications on its role based on localization are suggested and elaborated in this chapter.

4.1 Methods

The 1620 bp gene fragment corresponding to Asn^{33}–Val^{573} of YraM was amplified by polymerase chain reaction using genomic DNA isolated from *Haemophilus influenza*e Rd strain (ATCC#9008) (Akerley et. al. 2002). The start site was chosen as residue 33 from combining sequence comparisons that removed the signal II peptide and based on predicted structure of YraM that indicated unstructured regions till residue 33. Sequences of the primers were forward 5'-CATGCCATGGCGAATTTCACGCAAACCTTACAA-3' and reverse 5'-GCCGACGTCGACAACTGGTACAATTGCACCATC-3' that added NcoI and SalI sites to the 5' and 3' respectively. The amplified gene fragment was cut with NcoI and SalI and then ligated with T4 DNA ligase into pETBlue™-2 vector (Novagen/EMD) that had been opened with NcoI and XhoI. The resulting plasmid was transformed into expression host Tuner (DE3) pLacI (Novagen) that produced the soluble protein with sequence Met–Val–YraM (33–573)–Val–Asp–6(His). The protein was purified by nickel metal affinity chromatography followed by gel filtration (Superdex S200) but initial screening for crystallization conditions at both 22^{o}C and 4^{o}C was unsuccessful (done by J.Vijayalakshmi).

As part of an effort to generate antibodies for YraM, the YraM expression plasmid was modified by Quikchange to restore the native carboxyl-terminus of the protein. Mutagenesis to insert the native two amino acids and stop codon were made with the forward primer 5´-GTGGTGGTGTTAGTTGGCAACTGGTACAATTGCACCATC-3´ and reverse 5´-GATGGTGCAATTGTACCAGTTGCCAACTAACACCACCAC-3´. The resulting plasmid was sequenced with the pETBlue™-2-specific primers nt746 and nt1139 at the University of Michigan DNA sequencing core facility (This work was done by Bhramara Tirupati). The resulting plasmid was transformed into Origami (DE3) pLacI (Novagen) and expressed the protein Met–Val–YraM (33-573)–Ala–Asn, herein referred to as YraM-FL. YraM-FL has one disulfide bridge (Cys 356–Cys554).

Flasks containing 1 L terrific broth with 1 ml 100 mg/ml ampicillin and 1ml of 33 mg/ml chloramphenicol were inoculated with a 10 ml overnight culture in LB medium. Cells were grown at 37°C with shaking at 250 rpm until absorbance at 600 nm was 0.6-0.8, then allowed cool to room temperature (25ºC). IPTG was added (0.2 mM final concentration) to induce protein expression and the culture incubated with shaking at 25°C. Cells were harvested after 16 hours and used for purification. Further, the media was always supplemented with ampicillin (1µl of a 100 mg/ml per ml of media) and chloramphenicol (1 µl of a 33 mg/ml per ml of media) to select for pETBlue™-2 vector and pLacI helper plasmid, respectively including overnight starter cultures.

YraM-FL was purified in three steps. An initial ammonium sulfate precipitation was carried out using 0–30% and 30–50% saturation. The pellet from 30-50% was resuspended in 50mM Tris, pH 8.0 and dialyzed overnight to remove excess salt. The protein was then passed through anion-exchange columns, SourceQ and MonoQ (GE) successively with 0–1 M gradient of NaCl. The protein from the SourceQ column was dialyzed to reduce the ionic strength before adding to the MonoQ column. YraM-FL eluted at peak corresponding to 150-200mM NaCl. The fractions were selected after running on SDS-PAGE and finally polished by gel filtration using a Hiprep™ Superdex 75 column (Amersham/GE Biosciences) where YraM-FL eluted as a single peak with an apparent molecular weight of 60KDa.

The protein was concentrated to 35mg/ml as determined by absorbance at 280 nm and the theoretical extinction coefficient of 69,915 $M^{-1}.cm^{-1}$ (Protparam, Expasy [41]). Crystallization was straightforward and produced many needle shaped crystals within days at both 22°C and 4°C under a wide variety of conditions. It was not clear whether absence of the hexahistidine tag or expression in Origami cells provided a protein more conducive to crystallization. Single crystals suitable for X-ray diffraction were hard to obtain reproducibly. Grid screens to optimize different salt concentrations, variation in pH, microseeding and additives were all tried to improve crystal morphology. One additive Xylitol (Hampton Research Inc.) with 0.1M MMT Buffer, pH 4.0 and 25% w/v PEG 1500 (PACT, Qiagen) as the reservoir in a 1:1 sitting drop 96-well vapor diffusion setup was successful in preventing needle clusters and gave a nice rod-shaped crystal that diffracted to a resolution of 1.97 Å at the synchrotron (21ID-G, LS-CAT, Advanced Photon source, Argonne National Laboratory). The structure was phased by molecular replacement using Phaser as implemented in CCP4 [47]. First the C-terminal domain structure served as the search model, and once its position was fixed, the N-terminal domain (Vijayalakshmi and Saper, unpublished) was located. Missing residues were built manually using Coot [46] and refined with phenix.refine [44].

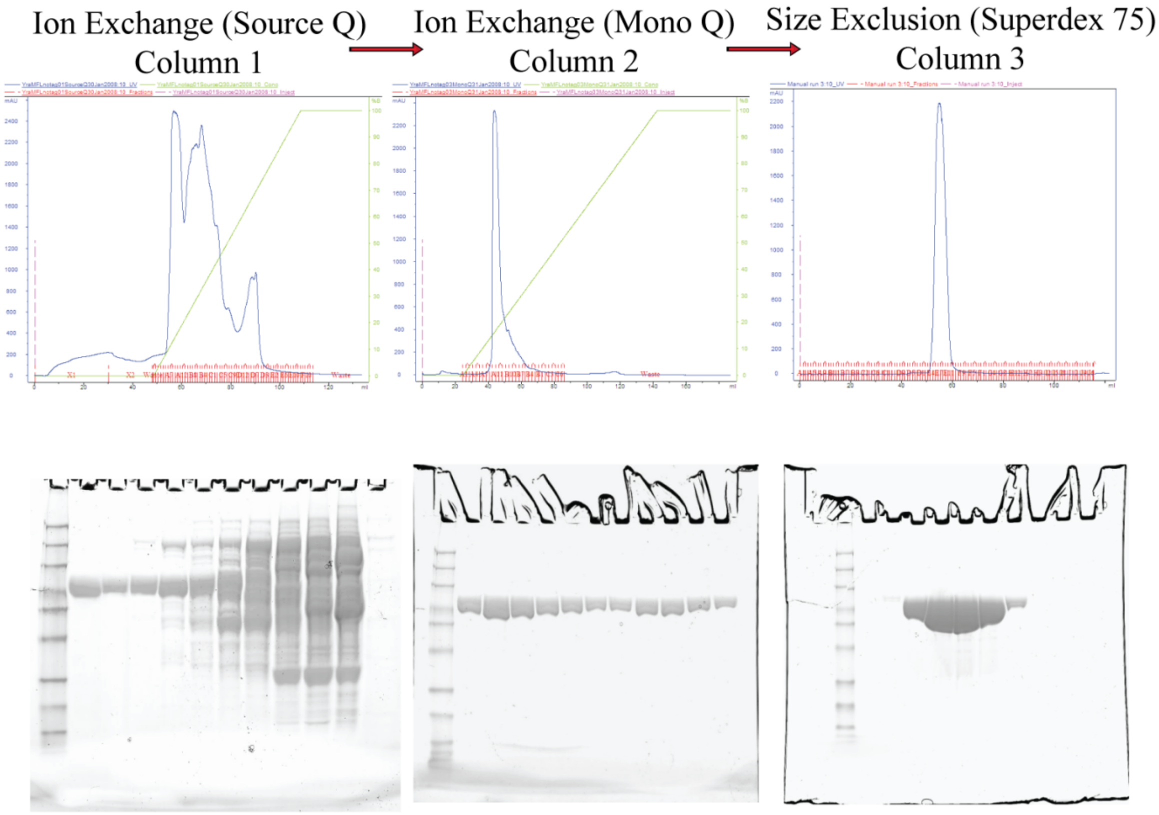


Figure 4-1 **Three stage purification for tag-less YraM full-length protein (33-575)**. The construct expressing YraM (33-575) was cloned in pETBlue2 plasmid in Origami (DE3) pLacI without a purification tag. Ammonium sulfate precipitation in two stages (0–30% and 30–50%) followed. The resuspended and dialyzed pellet from second fractionation was subject to two ion-exchange columns (Source Q and MonoQ and then a final polishing step with Superdex 75 and dialyzed to a final buffer with 50 mM Tris, 150 mM NaCl, pH 8.0 (Purification protocol worked out by B. Tirupati)

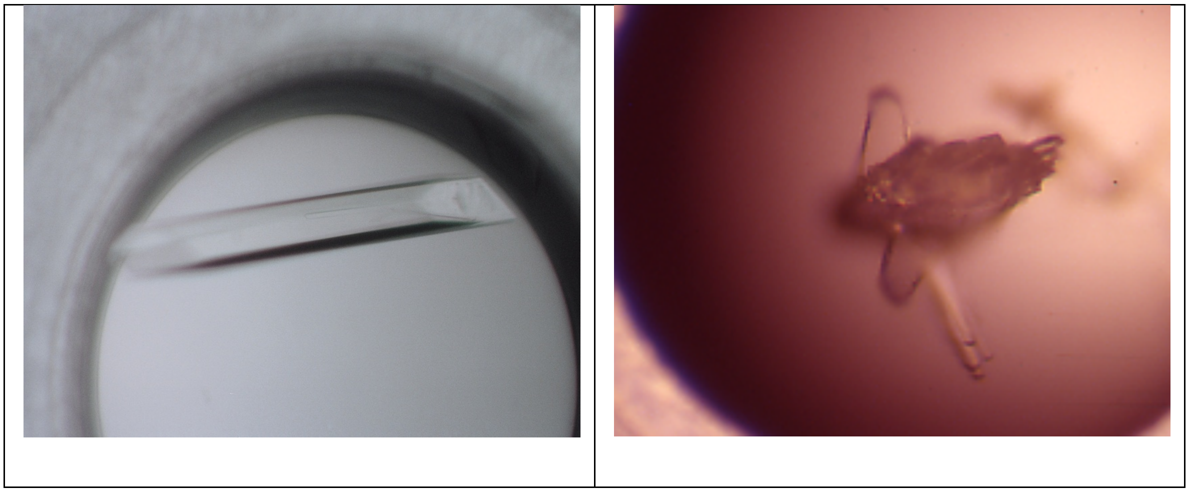


Figure 4-2 **Crystals of full-length YraM**. *Left*: Rod-shaped crystal grown in 4ºC after several months with a reservoir containing 0.1 M MMT buffer pH 4.0, 25 % PEG 1500 (50 µl) and a drop containing 1 µl of reservoir+1 µl of YraM protein (33 mg/ml in 50 mM Tris, 150 mM NaCl, pH 8.0)+30 % Xylitol (additive, 0.2 µl). *Right*: Stacked plates of YraM crystal grown at 4ºC that were relatively easy to obtain under few different conditions but diffracted poorly and the structure solution was not successful.

4.2 Results

4.2.1 Structure of YraM (33–575) at 1.97 Å resolution

The full-length structure of YraM at 1.97 Å (Figure 3-6) shows both the N- and C-domain in a similar open conformation as previously observed in the individual crystal structures (J.Vijayalakshmi and Saper. M, unpublished and [64]) The N- and C-domain together forms a open ended structure resembling a clamp (Figure 3-9). The open cleft of C-domain is not obstructed in anyway by the N-domain in the current observed conformation and can potentially bind any ligand in the periplasm or outside of cellular milieu depending on YraM's orientation in the outer membrane. The linker region ~251-256 residues (figure 3-12) may offer certain flexibility (as evidenced from normal mode analysis, see later) making the observed conformation perhaps one of many that is physiologically relevant.

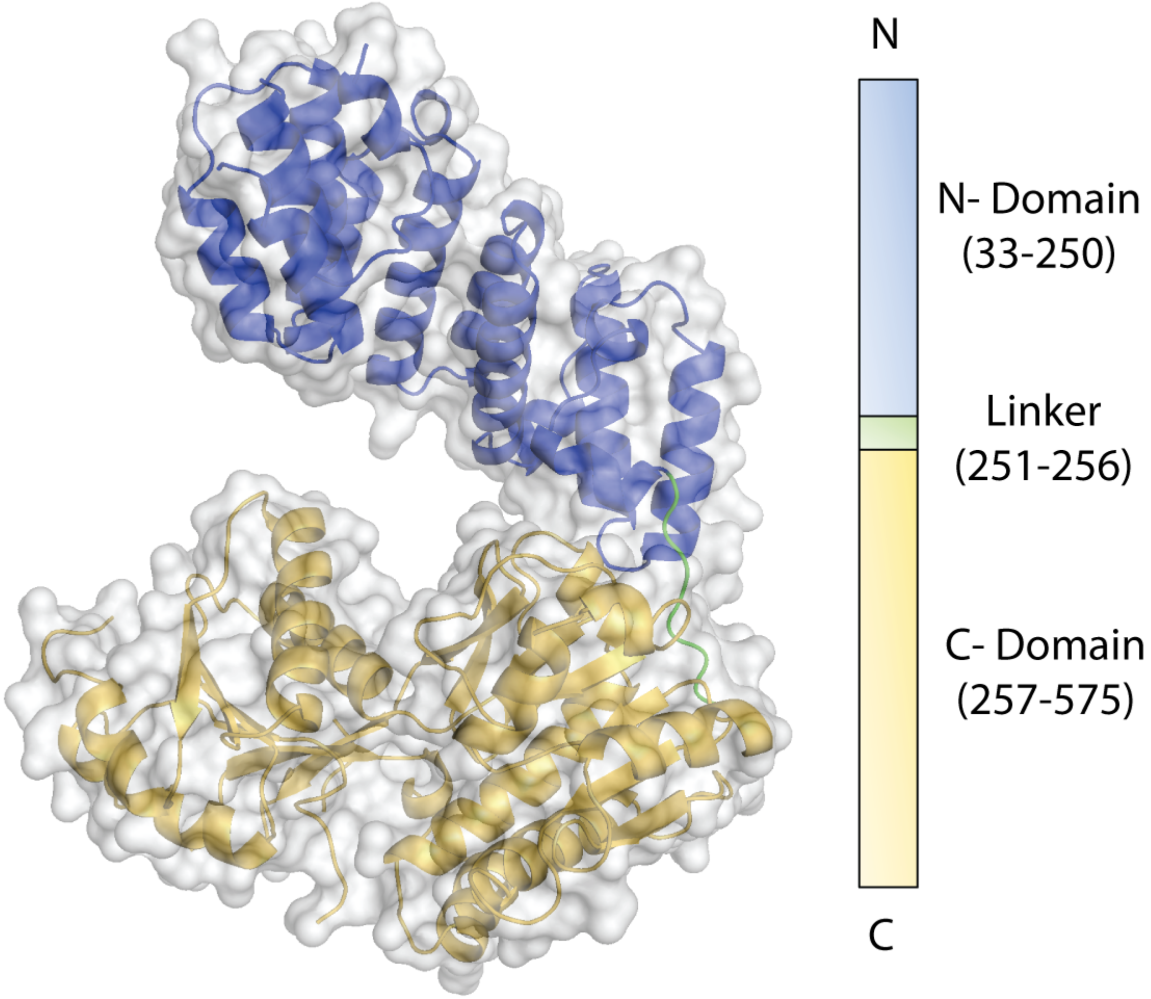


Figure 4-3 **Structure of YraM at 1.97 Å showing two different domains**. The amino terminal domain has a TPR fold and the carboxy terminal resemble type I periplasmic binding protein (PBP)-like fold.

Table 4-1 **Data collection and Refinement statistics for mature full length structure of YraM**

	YraM-FL
Beamline	21 ID-G, LS-CAT, APS
Space Group	$P2_12_12_1$
Unit Cell (Å)	a=66.04, b=68.47, c=128.57; $\alpha=\beta=\gamma=90°$
Number of molecules/asu	1
Unique reflections	40907 (1976)
Redundancy	7.2 (6.1)
Completeness (%)	99.2 (100)
R_{merge}	0.05 (0.2)
I/σ_I	47.2 (11.6)
Refinement program	*Phenix 1.6.4-486*
Resolution (Å)	30.4–1.97 (2.02–1.97)
R_{work}	0.22
R_{free}	0.28
Number of atoms (non-hydrogen)	4456
Protein	4190
Water	266
Wilson B (Å^2) (sfcheck)	29.76
Mean B_{iso} (Å^2, all non-hydrogen)	40.07
Protein	40.21
Solvent	37.91
r.m.s.d. from ideality	
bonds (Å)	0.008
angles (°)	0.968
Ramachandran favored (%)	95.51
Ramachandran outliers (%)	0.37

$$R_{merge} = \frac{\sum_{hkl}\sum_{i} | I_i(hkl) - \overline{I(hkl)} |}{\sum_{hkl}\sum_{i} I_i(hkl)}$$

$$R_{work} = \frac{\sum_{hkl} || F_{obs} | - | F_{calc} ||}{\sum_{hkl} | F_{obs} |}$$ calculated over all reflections used in refinement

R_{free} is similar to R_{work} but calculated from 5 % of the total number of reflections omitted in the refinement

4.2.2 Superposition of N-domain from YraM (33-249) structure solved earlier and YraM-FL (33-575)

The N-domain of YraM (33-249) solved independently earlier and the N domain in the full length mature YraM superpose well indicating no major conformational changes have occurred. The difference in resolution was not much between the two structures (2.0 Å to 1.97 Å) but residues 57-59, Asn 112 and Asp 195 were resolved in YraM-FL more likely due to different packing interactions due to presence of the larger C-domain (both also belong to the spacegroup $P2_12_12_1$). The r.m.s.d deviation is 0.56 Å.

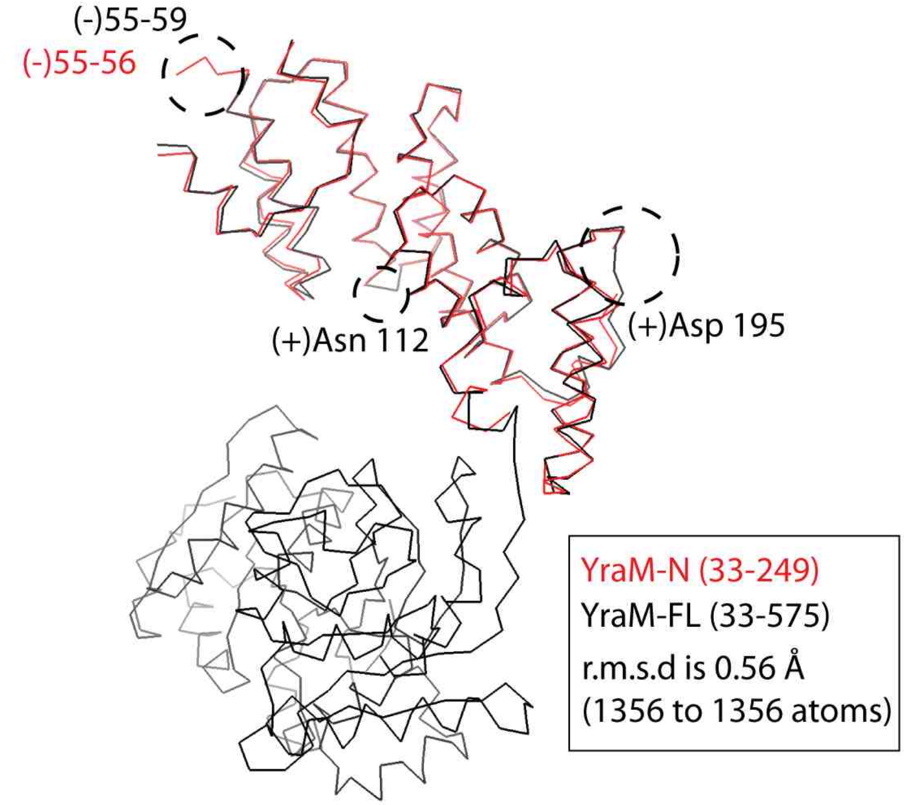


Figure 4-4 **Superposition of YraM N-domain (33–249) with the full length of YraM (33–575)**

4.2.3 Superposition of C-domain (257-575) with C-domain of YraM (33-575)

The C-domain with its type I PBP like fold is in the similar open conformation between the truncated and full length versions. The r.m.s.d between the two structures is 0.8 Å. This implies we do not observe the elusive ligand that was hoped to bind in the putative binding site and the subsequent closure of the two halves of the C-domain as suggested earlier in comparison to proteins that exhibit the periplasmic binding protein-like fold (PBP-like fold). It is possible that the purification method or the expression system did not have the cognate ligand resulting in this observation. The full-length structure has poor electron density corresponding to residues 470-479. The conformation in this region was copied from the 1.35 Å C-domain structure of YraM solved earlier [64] and used for refinement of YraM-FL. This was further justified since there were no large conformational changes seen in other parts of the molecule between C-domain alone and YraM-FL (Figure 3-10).

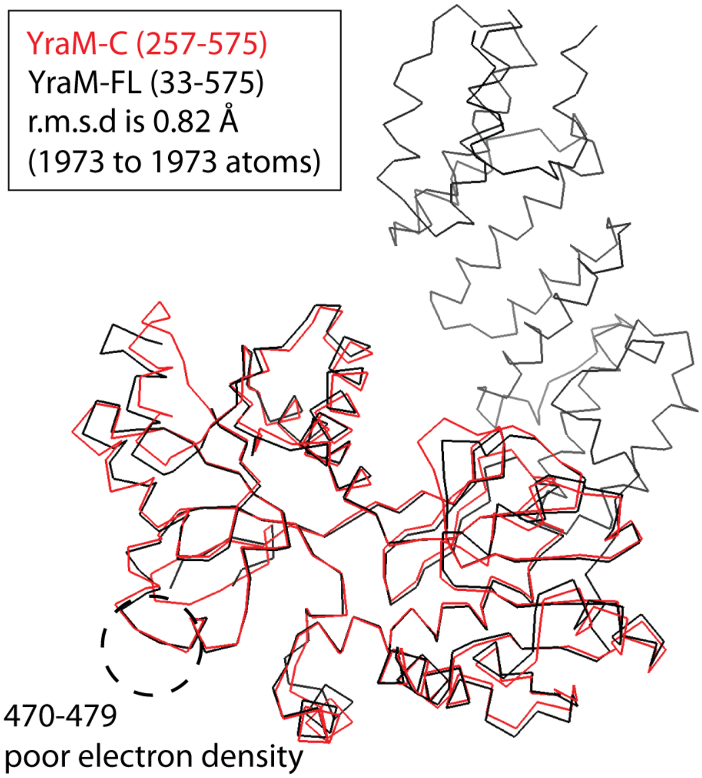


Figure 4-5 **Superposition of YraM C-domain (257–575) with the full length of YraM (33–575)**

4.2.4 Structure of the linker region (residues 250–256) in YraM-FL and how it relates the N- and C-domains

The linker (residues 251-256) has the sequence ^{251}Phe-Gln-Gln-Thr-Asn-Val^{256} in *H.influenzae*. The sidechains of three residues including the two glutamines and the asparagine have poor electron density. The linker assumes an extended conformation with three hydrogen bonded interactions to the long helix (511–535) in the C-domain with an odd 120° kink at Asn 528. The three hydrogen bonds include Asn531/ND2 to Gln 252/O (2.8 Å), Asn 531/N to Thr254/OG1 (2.9 Å) and Glu532/OE1 to Thr/OG1 (2.5 Å). The interactions with the N domain are minimal except for a long (3.2 Å) hydrogen bond between backbone amide hydrogen and oxygen of Phe251 with Leu248 respectively. Further, none of the residues in the linker are highly conserved based on conservation scores calculated by Scorecons [69] based on our alignment. Phe251 scores a fractional score of 0.5 and Val256 has 0.68 but the residues within this range all have a fractional score less than 0.4.

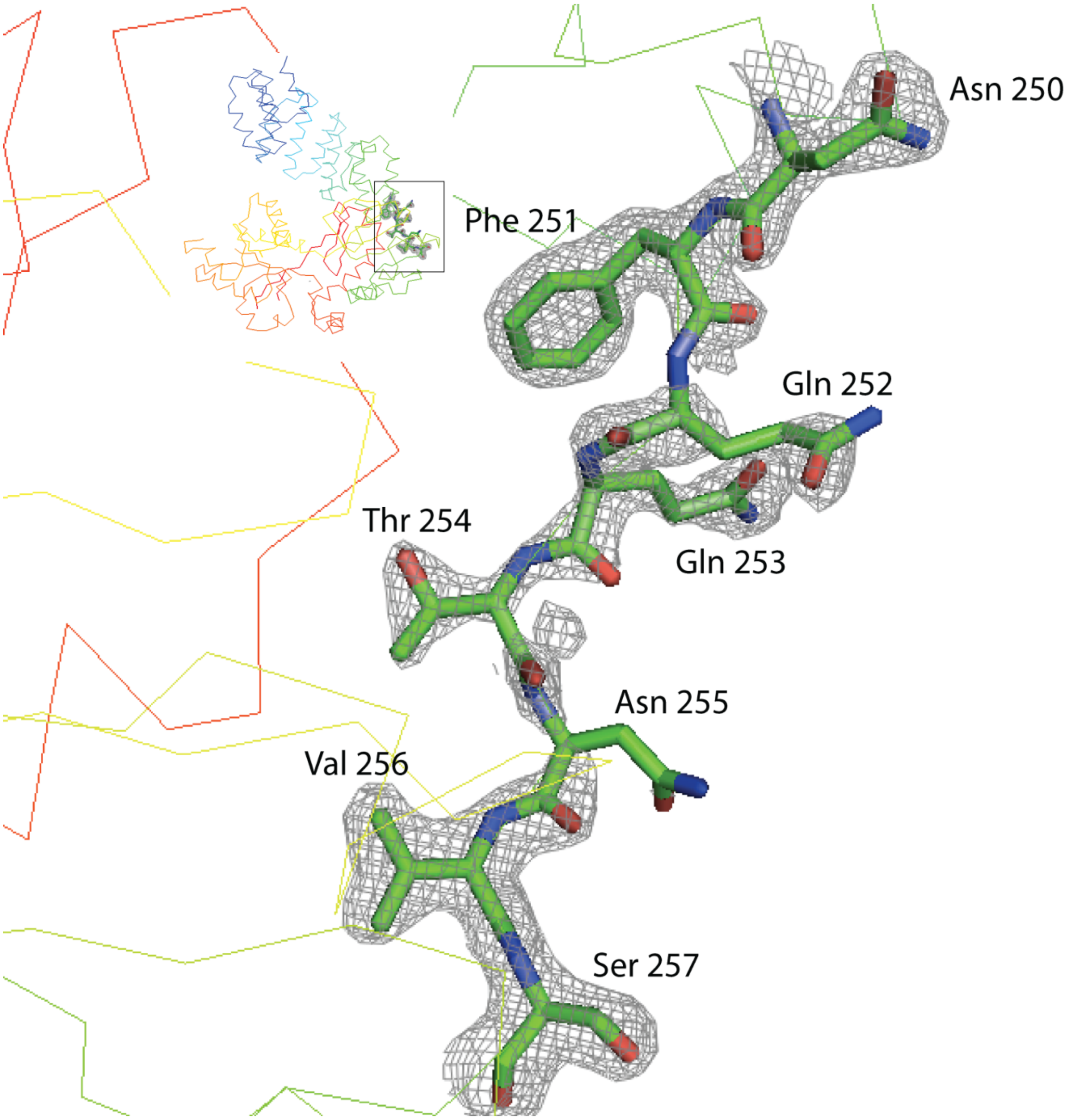


Figure 4-6 **The linker region of YraM (250–256) with the (2Fo-Fc) map rendered using MacPyMol**. B-factors for the linker region are higher than in rest of the protein ranging 40–55 while the rest of the protein is below 30.

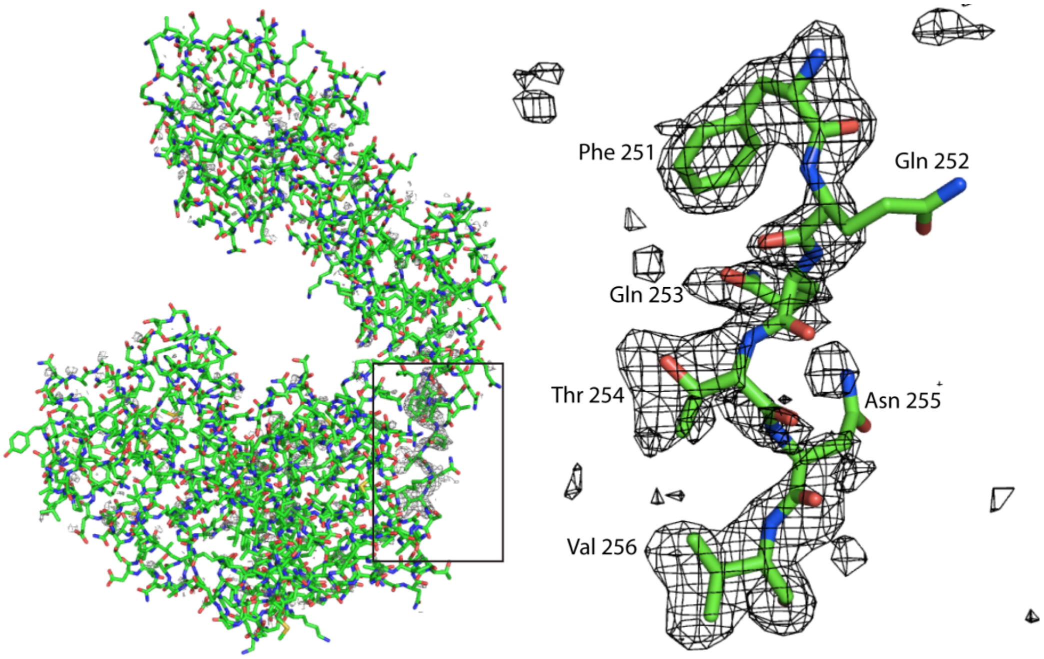


Figure 4-7 **(F_o-F_c) SA-Omit map of the linker region** (residues 251–256) of YraM-FL (*left*). Linker shown with (F_o-F_c) SA-omit map density contoured at 2.5 σ (*right*)

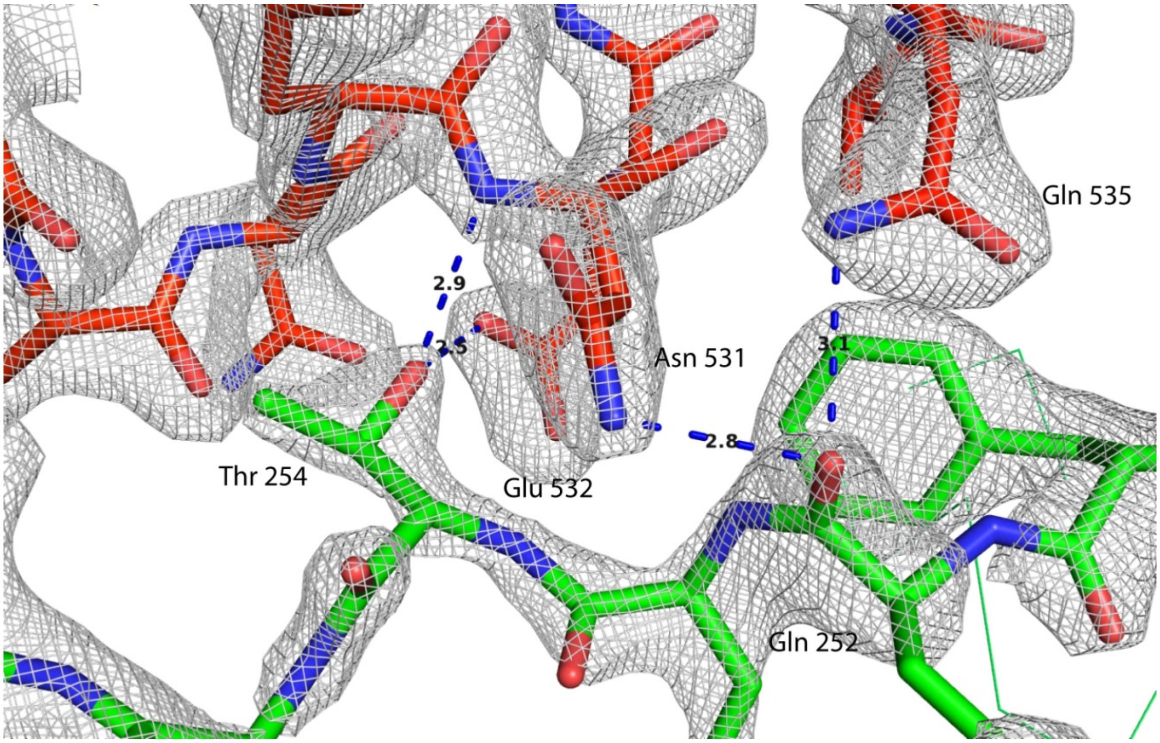


Figure 4-8 **Polar contacts between linker (250–256) (green carbons) with the C-domain YraM-FL (red carbons)**

4.2.5 YraM-FL has few surface conserved residues except as seen earlier with the proposed binding site in the C-domain

The lack of surface conserved residues other than the binding pocket in C domain discussed earlier can be seen again with the full-length structure. Further, the fewer conserved residues on the surface scattered in the N- and C- domain are random and do not fall on the same side of the molecule. The conserved patch on the convex face of the N-domain close to the linker is minimal but knowing the typical role of TPR domain is to mediate protein-protein interactions [38], this may function in binding as yet unknown partner molecule *in-vivo*. The orientation of this site is also opposite the proposed binding cleft of the C-domain [64] that is free to bind a small molecule ligand.

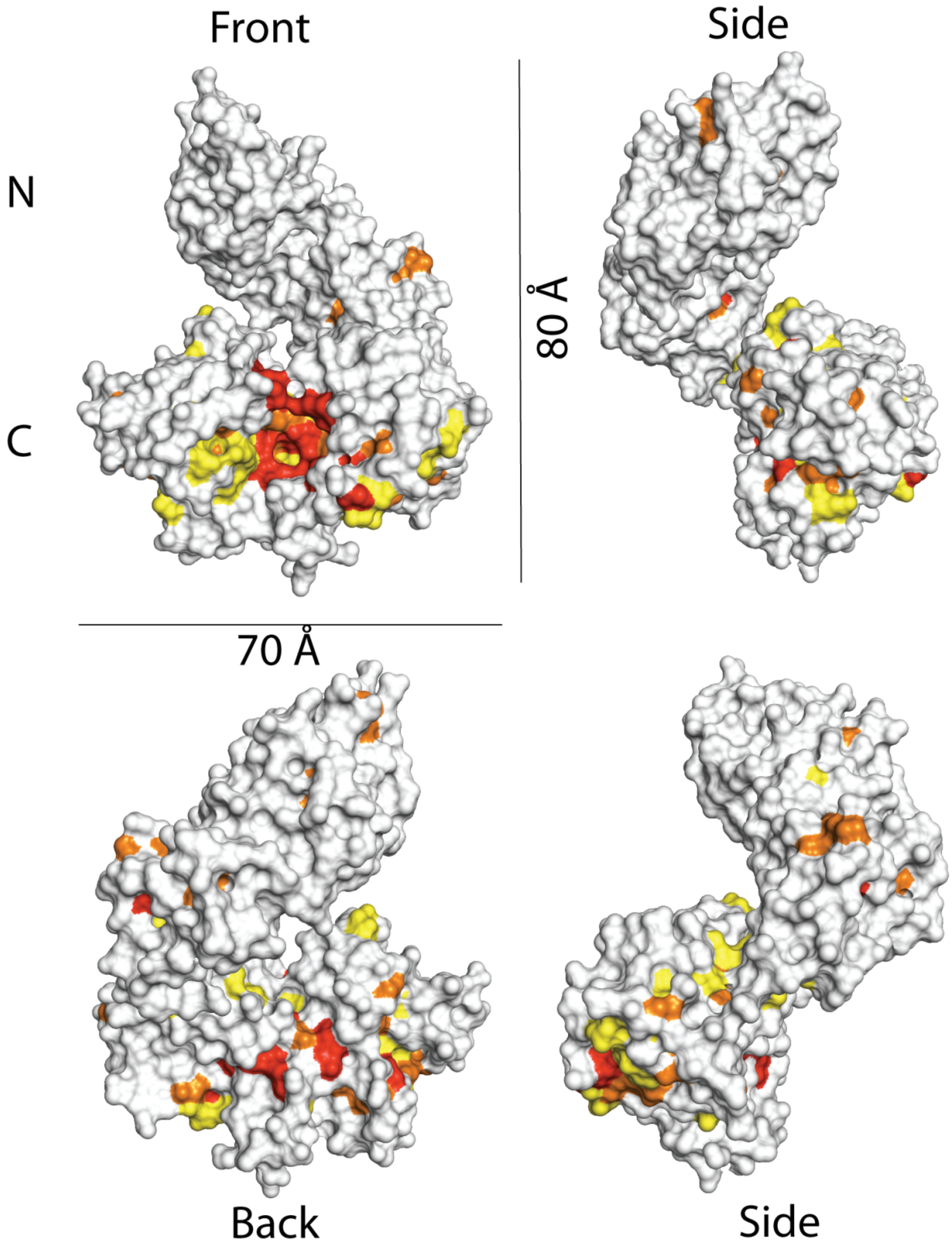


Figure 4-9 **YraM-FL Surface colored according to residue conservation by Scorecons** [69]. Red indicate residues with fractional score >0.9, orange for score>0.8 and yellow for score>0.7. Residues in gray have a score below 0.7. The figure shows the YraM-FL molecule from four different directions each related by 90º rotation with respect to vertical axis of the page. Top-left also shows the dimensions of the molecule.

4.2.6 The binding pocket of C- domain is obstructed in the full-length crystal structure by a symmetry related molecule

The conserved residue analysis earlier for YraM [64] clearly indicated a possible binding site in the C-domain and it was thought the full length YraM may stabilize the closed form of this PBP-like domain trapping the ligand. The $P2_12_12_1$ crystal structure however obliterates any such binding even if the ligand were co-purified due to crystal contacts with the N- domain of symmetry related molecule contacting the site of entry to the proposed binding cavity in the C-domain. The usual closing of the two halves of the PBP fold of C-domain seen with bound ligand is thus not possible.

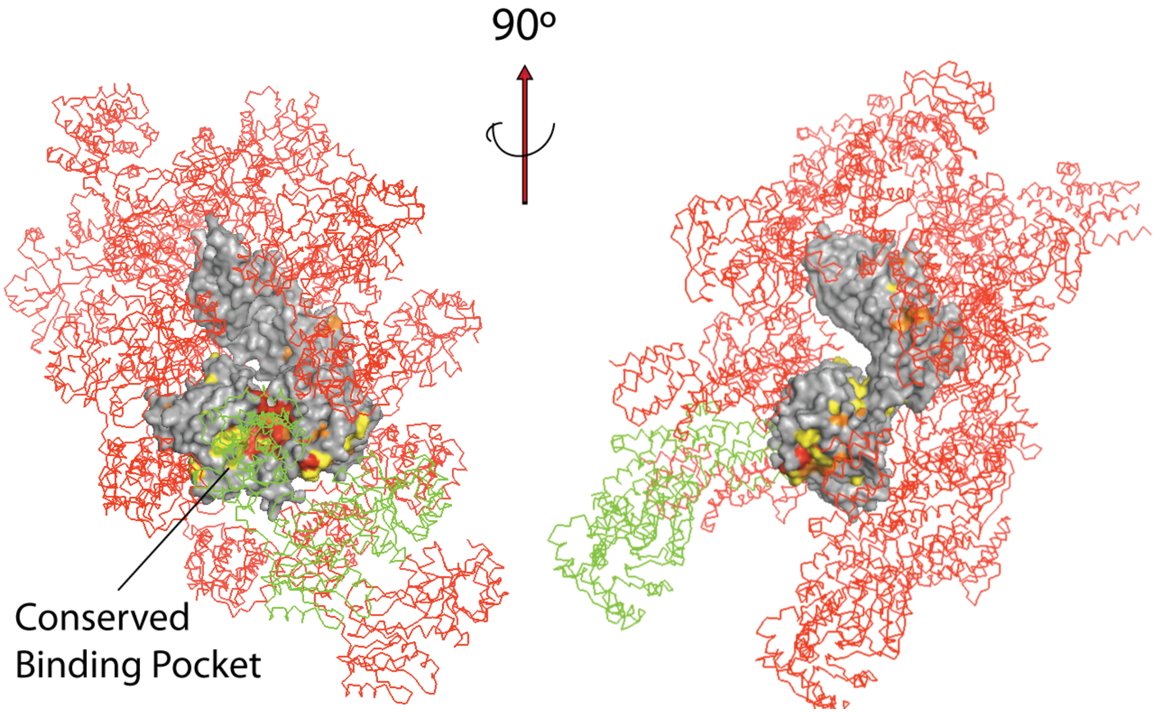


Figure 4-10 **Symmetry-related molecule of YraM occludes the proposed binding site on the C-domain**. The YraM molecule in center is shown in surface colored according to conservation scores from Scorecons [69]. The green molecule in wireframe is the symmetry related YraM that occludes the binding site in C-domain.

4.2.7 Normal mode analysis reveals flexible linker region in full-length YraM

Normal mode analysis (NMA) can be powerful to detect real motion in proteins. It has been shown that low frequency modes correlate well with motions in observed crystal structures in cases where more than one structural form is available [74]. We used an implementation of normal mode analysis the elastic network model: ElNémo to calculate the low frequency modes in YraM. The lowest frequency mode or most probable motion came to be the liberating motion between N- and C-domain where they come close to each other with the linker region (residues 251-256) serving as a hinge (Mode 7).

4.2.8 Interaction region between the linker and C-terminal domain of YraM

The residues in the region of linker that interact with the C-terminal domain are all not highly conserved whereas the residues on the opposite face of this interaction helix are highly conserved (residue conservation score from scorecons >0.7). This highlights a groove in which the extended linker region is able to interact freely (Figure 3-17). This observation together with the normal modes suggests the domains may move towards each other as suggested by the lowest frequency mode.

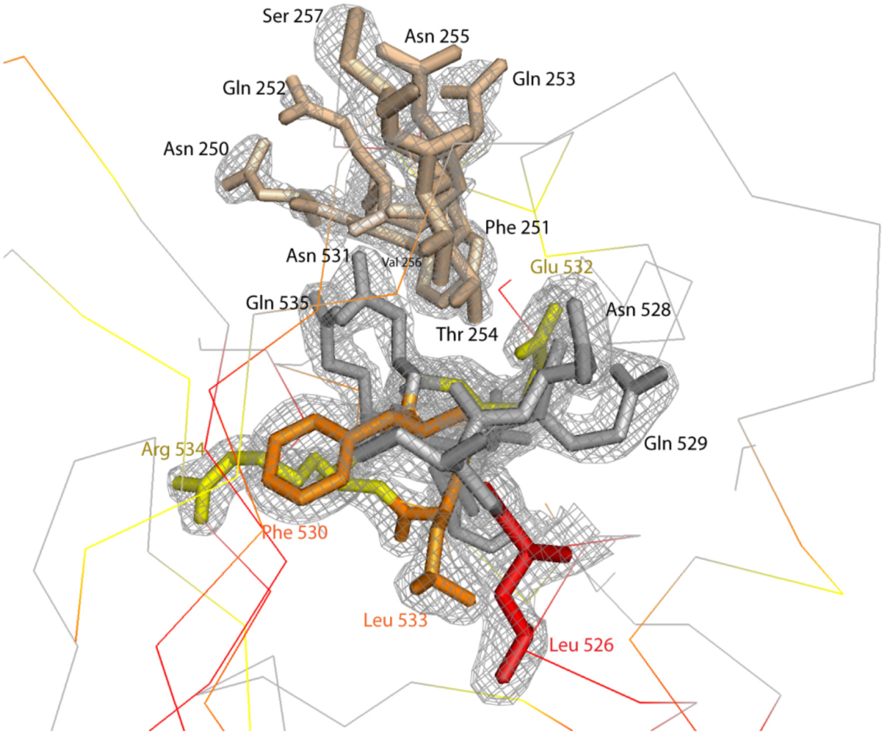


Figure 4-11 **The interaction region between linker and C-domain helix (End on view).** Residues in the C-domain are colored according to conservation scores from Scorecons. Red are residues with >0.9 score, orange >0.8 and yellow>0.7. The linker region is colored uniformly (light brown). The view is oriented in such a way that the C-domain is pointed out from plane of this page.

4.3 Discussion

YraM represents the first structure of a unique fusion of a tetratricopeptide repeat (TPR) domain with a periplasmic binding like protein (PBP). There has been little information published on functional characterization of YraM except for structure of C-domain published by our lab [64] earlier and the characterization of the gene to be essential for growth and viability of *H.influenzae* [65]. It is a predicted outer membrane lipoprotein in this organism based on the lipobox sequence but its localization has been experimentally verified in *E.coli* [66]. The conserved cleft in C-domain strongly indicates a binding role for the protein but the binding ligand partner is yet to be identified.

Considering *yraM* in its genomic context is one way that can provide us with some clues to its function [64]. *yraM* is found second as part of a five gene cluster in most Gram-negative bacteria (Figure 3-18). The analysis with the current annotation and published biochemical evidence all suggests why YraM should be essential but fail to offer coherent evidence for YraM's role in any one physiological function. The last gene in this cluster *HI1658* (*E.coli yraP*), also encoding an outer membrane lipoprotein like YraM was shown to be essential for *H.influenzae* viability [65]. This protein is under the regulation of σ^E factor in *E.coli* (shock response) and was proposed to be involved in outer membrane integrity [75]. *HI1657* (*E.coli yraO or diaA*) expresses a cytoplasmic protein DiaA that in *E.coli* is binds with DnaA and unwinds the origin of replication (oriC) [76]. If *yraM* is corregulated with *diaA* (like if in an operon) it would explain its role in cell division and hence its essentiality to growth and viability. HI1654 (*E.coli yraL*) encodes a putative tetrapyrrole methylase. Tetrapyrroles are precursors of porphyrins and the latter bind iron and thus are important for physiology of *H.influenzae.* *yraL* is essential in *H.influenzae* but not in *E.coli* similar to *yraM* [65].

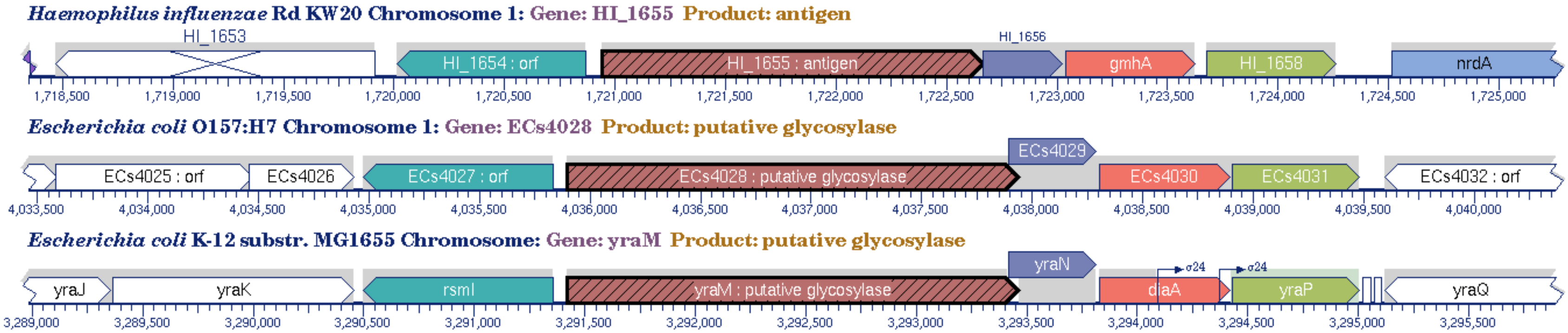


Figure 4-12 **Comparison of *yraM* five gene cluster between *H.influenzae* and *E.coli*.** Figure drawn by biocyc.org.

Approaching the function of YraM from a structure point of view however gives another interesting possibility. Analysis of proteins with TPR domain followed by a binding module or any other functional domain was searched for in the structure database [64]. Slt70, a periplasmic lytic transglycosylase that functions to break glycosidic bonds and involved in the turnover of peptidoglycan has a similar topology as YraM (Figure 4-13) [77]. The amino terminal domain of this protein has 11 TPR units followed by a catalytic domain that structurally resembles lysozyme [77]. Slt70 has infact been crystallized with 1,6-anhydromuropeptide (a precursor of peptidoglycan biosynthesis) bound to the catalytic domain (Figure 3-19).

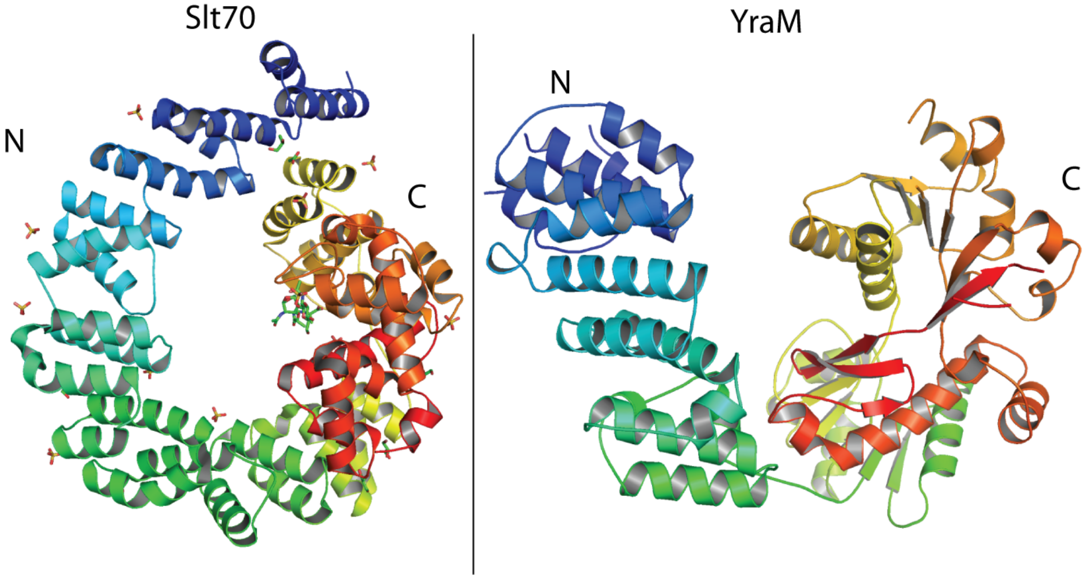


Figure 4-13 **Comparison of YraM with Slt70**. Both have amino terminal extended TPR fold followed by a functional domain. Slt70 has a ligand, 1,6-anhydromuropeptide, (PDB ID: 1QTE) a precursor of peptidoglycan biosynthesis bound in the carboxy terminal domain[77].

The role of YraM in binding peptidoglycan is not improbable as YraM is only found in Gram-negative microbes and the dimensions of the full-length YraM protein determined here reinforce the idea. Assuming the more likely orientation for YraM that is the inner leaflet of the outer membrane it would extend 80 Å (Figure 4-9) from one end of the molecule anchored to the membrane to the other end free in the periplasm. These dimensions are based on the observed crystal structure and would put the carboxy terminal binding pocket within reach of the peptidoglycan layer. The peptidoglycan layer is estimated to be about ~50 Å from the membrane [64, 78]. Further, there is precedence for binding protein folds relevant in peptidoglycan biosynthesis, for instance, MppA [72] is a PBP essential for importing murein tripeptide and MurG is a cytoplasmic PBP-like glycosyltransferase required for the biosynthesis of peptidoglycan precursors [79].

It is also possible based on the orientation of the two domains and the flexibility between them as evidenced by normal modes that the binding pocket in C-domain could be more extensive. The two domains together can close on binding peptidoglycan precursors involving the conserved cleft in the C-domain and the concave surface of the N domain. Investigating the unique possibilities for YraM’s function discussed considering its location and the search for ligand can be exciting and useful future explorations.

Chapter 5

Summary and Future Directions

Chapters 1 and 2 of this thesis present the background information on *Escherichia coli* polysaccharide capsule assembly and the experimental determination of the three dimensional structure of GfcC. The latter is a soluble periplasmic protein encoded by the group 4 capsule operon shown to be important for capsule expression in pathogenic *E.coli* (EPEC). Chapters 3 and 4 delve into an unrelated project with the identification of a gene encoding an essential lipoprotein named YraM discovered through whole genome *in-vitro* mutagenesis studies of *Haemophilus influenzae.* This protein is critical for the growth and survival of this organism. Determination of the three dimensional structure of YraM revealed a unique fusion of two well-studied domain folds whose individual domain structures had been previously determined in our lab (J.Vijayalakshmi). The single common attribute that ties both these proteins (GfcC and YraM) is the absence of specific functional information for their physiological role. The most important results from this thesis are therefore all structural in nature. In this regard, this thesis presented two exciting structures from two different systems that will complement future experiments into the proteins' function.

5.1 Pondering over the structure of GfcC and its relevance with respect to other capsule proteins

The structure of GfcC is now the second capsule assembly protein structure determined (the other being Wza) where both contain β-grasp domains. The amphipathic C-terminal helix from GfcC is shorter and folds back onto the protein core compared to the one in Wza that is extended and contributes to the octameric functional form for Wza. GfcC is therefore monomeric in solution as is also shown through gel filtration and analytical ultracentrifugation studies here. More importantly the different conformation of the C-terminal helix in GfcC is maintained due to the helical hairpin domain (D2H) that seems to lock the helix in place. Thus, the two proteins possibly have entirely different roles within the polysaccharide assembly process. This observation stated above however lends credence to one possible idea: that Wza could chaperone its C-terminal helix in a GfcC-like form before it oligomerizes into its functional octameric state in the outer membrane. We have seen from GfcC that such an interaction is possible between a similar C-terminal amphipathic helix with the β-sheet of the β-grasp domain (D3). It should be pointed out in this context that Wza is a lipoprotein whereas GfcC is not and that a D2H-like domain is absent in Wza.

The group 4 operon further encodes an identically sized (379 residues) lipoprotein, GfcE, that is 74% identical and 95% similar to Wza leading to the conclusion that their structures are very much alike. GfcE also has β-grasp domains and so does GfcC like we have seen in this thesis. What is the role of β-grasp domains from these two different proteins within the same operon? Given the propensity of β-grasp domains to interact with each other (like we saw in Wza forming the octamer), is it possible the domains of GfcC interact with periplasmic portion of GfcE? This taken together with the hypothesis that GfcC may interact with GfcD based on presence of fusion genes found in other organisms (like *Burkholderia sp.*), the polysaccharide assembly apparatus seems to involve a more extensive complex involving not just both inner and outer membranes but the periplasm too than what is currently appreciated. Studies of group 1 revealed a minimal periplasmic spanning complex of Wza-Wzc linking both the outer and inner membrane but immediate analogs for GfcABCD proteins were not apparent in group 1 system. In this context it should be mentioned that another operon, *yjbEFGH,* located elsewhere in *E.coli K30* genome are paralogs of *gfcABCD* and possibly take a similar role for group 1 capsule but their involvement is not proven and they are also not contiguous with the *wzi, wza, wzb, wzc* group 1 genes. It is also known that the *yjbEFGH* genes are stress-regulated and may also function in an entirely different capsule system. These predictions and observations throw more light on a more extensive supramolecular complex involved in assembly or possibly anchoring the group 4 polysaccharide capsule.

These also seem to suggest that group 1 and group 4 capsule assembly may be more dissimilar to each other than currently presumed.

The hypothesis that the proteins encoded by *gfcABCD* genes may be involved in anchoring the polysaccharide to the outer membrane is suggested by comparing the group 1 operon of *E. coli* K30 with the group 4 operon genes. The group I operon encodes *wzi* that encodes a monomeric β-barrel membrane protein. Mutants that lack *wzi* give a phenotype where the polysaccharide capsule is not tightly associated with the outer surface as visulaized through electron micrographs [20]. Does GfcABCD proteins and particularly GfcD with the predicted large β-barrel architecture fulfill a function similar to Wzi?

Revisiting the observation that GfcC and GfcD coding genes are fused in some organisms also parallels interestingly to what is seen in some exopolysaccharide assembly systems like cellulose, polyglucosamine (pga) and alginate exopolysaccharides. Here unlike an α-helical outer membrane barrel protein like Wza, the export of polymerized polysaccharide occurs through a larger integral outer membrane β-barrel protein. In alginate export, the β-barrel protein AlgE associates with the lipoprotein AlgK to do this function, the latter also required for the proper localization of AlgE in the outer membrane. AlgK and AlgE homologs are again fused as single proteins named PgaA for poly-β-1, 6-N-acetyl-D-glucosamine (pga) export and BcsC for cellulose export. Therefore it is attractive to think of an analogous situation with GfcC and GfcD functioning similar to AlgK and AlgE.

5.2 Proposed Experiments to determine function of GfcC

The first step in determining the function of GfcC is to study the phenotype changes or differences in capsule expression that lack of *gfcC* gene would produce in pathogenic *E.coli* cells. The second step is to map these phenotypic changes onto the available structure of GfcC through site-directed mutagenesis. For instance, what is the role of Arg 115? Is it just structural? Does mutation of residues in the potential probable galactose like binding site (pocket 1) affect the phenotype? Is the capsule still produced and is it intact? And finally, mutate residues in the observed (and conserved) dimer interface and see if it affects capsule expression. Also, do residues at the end of the C-terminal helix affect an interaction with GfcD and thus affect capsule formation?

We already know that this GfcC protein is being encoded by the seven-gene operon that was shown to be linked with group 4 capsule assembly. Before dwelling on phenotype studies it is also of particular interest to identify the localization of this GfcC protein in the *E.coli* cell. It is a small (~26 KDa) periplasmic protein as evidenced by sequence and preliminary experiments performed with osmotic shock of whole EPEC (EM4462) cells but association of GfcC with either membrane cannot be dismissed at this point. There is precedence that GfcC could in fact be associated with the membrane or a membrane integral protein based on two points of evidence. One is the observation of homologous regions of GfcC and GfcD encoding genes fused as one in some *Burkholderia* and other species. GfcD is further predicted to be a large β-barrel membrane integral protein. The second is based on some initial experiments with sucrose gradient ultracentrifugation and separation of the membrane fraction of EPEC cells (EM4462) expressing native GfcC with a C-terminal His tag and identification of the latter with a specific antibody on a western blot. These experiments do not distinguish between the inner and outer membrane and controls to identify inner and outer membrane marker proteins or enzyme (esterase or NADH oxidase) activity were not conclusive. This experiment needs to be repeated and then followed by a "pull-down" assay to reveal other proteins that copurify with GfcC. The confirmation of cellular location and its co-localization with GfcD would indeed be a very important result in verifying the role of GfcC and its partner protein in capsule assembly subsequently.

A related subject of interest with reference to GfcC's localization would also be to determine if GfcC aids in the proper localization of some other protein say GfcD. The latter is due to some strong similarities between the organization of the gfcC and gfcD genes with *algK* and *algE* genes from the alginate exopolysaccharide export system. It was reported that AlgK (a lipoprotein) is required for the proper localization of AlgE to the outer membrane; its absence results in more or less equal distribution of AlgE to both membranes. Likewise, is GfcC essential for the proper localization of GfcD? Membrane fractionation studies followed by western blot visualization of known markers of inner and outer membrane proteins in the presence and absence of *gfcC* gene without affecting *gfcD* should also put this question to rest.

The second experiment is to determine the effect on capsule expression upon deleting *gfcC* (deletion mutant) and then complementing the gene on a plasmid. This should complement the loss of function i.e., the phenotype should return back to being identical to the wildtype. To study changes to expressed capsule we have used the similar buoyancy assay that had been used earlier [4]. The premise is that cells with their intact capsule float on high density media (like the Percoll reagent) whereas cells that lack capsule are more dense and settle to the tube bottom upon centrifugation. This proposed experiment has been done many times in our lab but although we see a slight change in buoyancy the capsule is still present in the *gfcC* deletion mutant. Further the complementation does not restore wildtype phenotype completely. In contrast, the buoyancy assay with an *etk* (kinase) deletion mutant seems to abolish all capsule synthesis as evidenced by the cells settling to the bottom of the centrifuge tube.

Since the above buoyancy experiments to determine phenotype are inconclusive, it is now clear we need more direct methods on observing the changes in capsule expression. We turn to imaging for this purpose. Electron micrographs of cells lacking *gfcC* and cells that encode *gfcC* on a plasmid can be performed. Alternatively, fluorescently tagged anti-O127 (O127 is the EPEC capsule serogroup) antibodies should be used to visualize the cell surface for evidence of capsule. These direct visualization methods will shed more light on what exactly happens to the *E.coli* cell and the capsular surface when *gfcC* is deleted. If the immunofluorescence experiments show the absence of O-antigen in the *gfcC* deletion mutant, but buoyancy suggests the presence of capsule, there are two possibilities: either the group 4 capsule is made but trapped in the periplasm, or a different capsule is made that has not been modified by O-antigen. The former can be tested by electron microscopy, while the latter requires careful chemical analysis of the capsule in the *gfcC* deletion mutant.

It may also be possible that *yjbEFGH,* and particularly *yjbG* being a paralog of *gfcC* could substitute for its function and mask the true phenotypic changes of the *gfcC* deletion mutant. The experiments must therefore be performed in a background where these genes (*yjbEFGH*) are deleted to prevent any interference. There is also the possibility that any single gene altercation can only have a mild effect on the overall phenotype that may not be detected. Then deletion of two genes at a time, for instance *gfcC* and *gfcD,* should provide an answer if the two corresponding proteins function together to bring about a more drastic change to the wildtype capsule expression.

Considering the specific role of GfcC, there are at least two scenarios on which the above experiments speculate:

1) GfcC and GfcD function together and GfcD is an outer membrane pore that serves as the alternative exit route for branched polysaccharides like the group 4 polysaccharide. GfcC could serve here as the *'adapter molecule'* that links GfcD to other periplasmic and inner membrane components for instance, the kinase (Etk) and the polymerase (Wzy).
2) GfcC and GfcD are replacing the role played by Wzi in group 1 and are required for capsule attachment to cell surface.

5.3 Pondering over the structure of full-length YraM

The structure of lipoprotein YraM had already been described adequately by separate structures of its two individual domains, the N-domain (residues 33–253) and C-domain (residues 257–575) (J.Vijayalakshmi and M.A. Saper). The N-domain has a TPR-like motif that in other proteins with the same fold has been identified to mediate protein-protein interactions. The C-domain has a periplasmic binding protein type fold similar to the amino acid binding proteins like Leucine, Isoeucine, Valine binding protein (LIV-BP) in the open unliganded conformation. The binding partners for both the N-domain that could be a large protein and the C-domain that based on the conserved cleft similar to related proteins suggests an amphipathic small molecule binding site are both yet to be discovered.

The structure of full-length lipoprotein YraM does reveal the conformation of N-domain and C-domain together and it shows how the two domains are related to each other. Importantly, the binding cleft in the C-domain is left unhindered for a small molecule ligand to bind, possibly one found in the periplasm. The extended linker and normal mode analysis also point to flexibility in the linker. The observed conformation may be only one of the possible conformations that this protein takes *in-vivo*. It can be reasoned, the fact that the two domains are linked in one polypeptide suggests a binding event in one domain can relay that message effectively enabling or preventing the relationship with a partner molecule in the other domain. This fusion of two distinct domains is a big advantage from the point of view of relaying a signal within the cell from one protein to another in a reaction or signal transduction cascade.

The closing of the two halves of C-domain upon ligand binding follows from similar observation in other related PBP proteins. However, in the case of YraM such closing need not occur. Further the N-domain itself may close upon the ligand bound C-domain like a lid domain encasing the bound ligand in place. The open conformations of periplasmic binding proteins have been captured in crystal structures sometimes in differing degrees of openness (open vs super-open and so on) essentially due to the flexibility between two halves of the PBP fold without the ligand bound. However, in the case of YraM, individually determined C-domain structure (1.35 Å) and the C-domain in full-length YraM (1.97Å) have exactly identical open conformations.

Since the Lipobox sequence suggests an outer membrane anchored location for YraM (*E.coli* YraM is outer membrane as also shown in a proteomics study) and since the majority of lipoproteins in outer membrane are preferably facing the periplasm, the full length structure of YraM and the dimensions of the molecule also suggest the context for this protein's function. It is with reference to this orientation and the organization of the gram-negative peptidoglycan in the outer wall, that it is proposed YraM may be involved in peptidoglycan assembly or disassembly. This function certainly makes it essential and also confirms why YraM might be found exclusively in gram-negative bacteria and not in gram-positives. Further, searching for two distinct domains joined together by a linker also revealed Slt70, a murein transglycosylase involved in peptidoglycan breakdown that also has an extensive TPR domain followed by an enzymatic domain resembling lysozyme.

5.4 Proposed experiments to determine function of YraM

The first experiment required is to determine whether YraM is exposed to the periplasm or the cell exterior. Next, to elucidate YraM's function is to find binding partners for both the N-domain and the C-domain. Pull-down assays or crosslinking can identify large molecule binding parterners, while high throughput ligand binding assays may help find a small molecule ligand. It must be remembered that the protein in this study was made recombinantly in *E.coli* as a soluble protein. Eventually, it may be required to isolate the protein from its native environment (*Haemophilus influenzae* outer membrane or the *E.coli* variant) using mild detergents to use in functional studies. The ligand that binds to the C-domain may also have not been present in *E.coli* cytoplasm for it to be co-purified and crystallized in our studies here.

In terms of essentiality of YraM in *H. influenzae*, mutants lacking either the N-domain or C-domain individually while maintaining the proper localization and expression levels should indicate which of the two domains are critical to cell survival. Mutations in the active site or proposed binding site in the C-domain should further validate the protein's function.

Appendix

A.1 Localization studies of GfcC by osmotic shock

Methods

Wild-type EPEC cells (strain EM4462 *gfcc::kan*) were electroporated with pAP1720 (a modified pSA10 vector) expressing wildtype GfcC with the intact signal sequence and C-terminal hexahistidine detection tag (plasmid kindly provided by Prof. Ilan Rosenshine, Hebrew University Faculty of Medicine). Cells without the plasmid were the control for this experiment. 100 µl of overnight cultures of both these strains were used as inoculum in a 10ml LB culture with the appropriate antibiotics (1% inoculum). Expression of GfcC from the plasmid was induced with a final of 0.02mM IPTG (Soltec Ventures) and allowed to proceed for 4 hours. Periplasmic fractions were then extracted by the osmotic shock procedure. First, the preweighed cells were washed in buffer containing Tris with 20% sucrose according to the wet weight of cell paste (8ml of 30 mM Tris.Cl/20%sucrose, pH 8.0 for every 0.1 gm cells). EDTA was added from a 0.5M stock to make the final concentration 1mM. The suspension was then incubated for 10 minutes at room temperature with shaking. The cells were then centrifuged at 4500g on a tabletop centrifuge (eppendorf, 5804 R). The pellet was resuspended in 2.5 ml of 5mM ice-cold $MgSO_4$ and centrifuged for 10 minutes at 14,000 g at 4^{o}C. The pellet from this step would have the spheroplasts and unbroken cells and membranes whereas the supernatant constitutes the cold osmotic shock fluid (periplasmic fraction). The latter was then concentrated with Talon beads and run on SDS-PAGE. The western blot of the polyacrylamide gel was then carried out according to standard protocol for wet transfer (tank) and probed with anti-C-His primary antibody from Sigma. The antibody was detected by the chemifluorescence signal from the ECL+ (GE Healthcare) reagent.

Results

The *gfcC::kan* EM4422 EPEC cells were grown with and without pAP1720 expressing the native GfcC. The osmotic shockate (corresponding to the soluble fraction of the periplasm) was probed by Western blots with anti-His antibody. Control cells lacking *gfcC* (lacking the pAP1720 plasmid) showed no signal with the anti-His antibody, while cells transformed with pAP1720 showed a band co-migrating with recombinant GfcC. When indiced with IPTG the amount of GfcC in the periplasm increased (Figure A-1). The invariant Arg115 close to Pocket 1 (see results, chapter 2 and alignment, figure 2-6) was mutated to Lys (R115K, conservative) or Ala (R115A) to probe the function of resulting GfcC. These mutants in a similar experiment showed no detectable GfcC in the periplasm but in this case the overall expression levels also falls dramatically compared to wild-type GfcC. It is possible that the substitution of this Arg residue affected the folding of the protein and/or made it more susceptible to proteolytic degradation by destabilizing the structure of GfcC (Figure A-2).

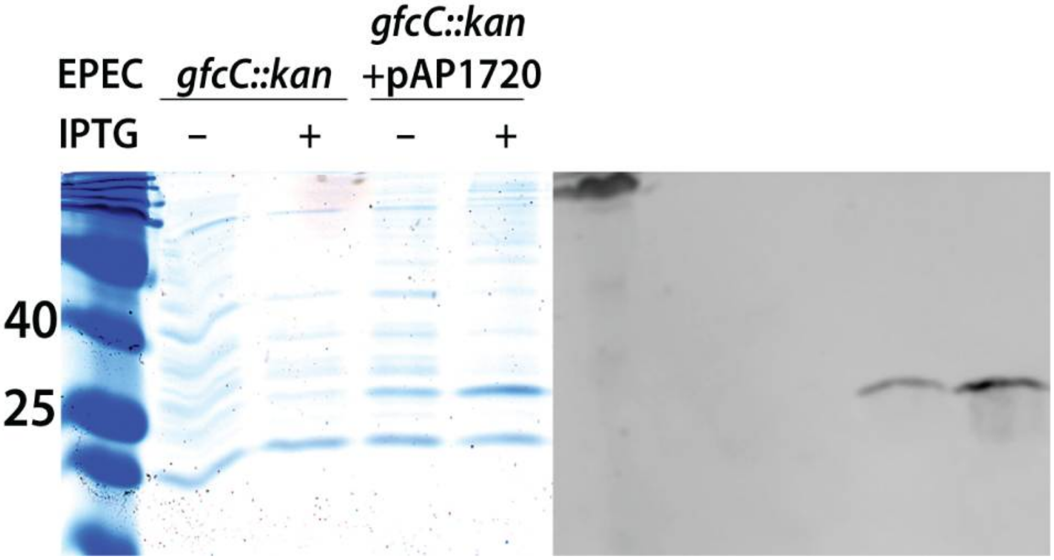


Figure A-1 **GfcC is localizaed in the periplasm as seen from the anti-his antibody labeled blots of the osmotic shock fluid**. The increase in expression with IPTG induction can also be seen from the western blot to the right.

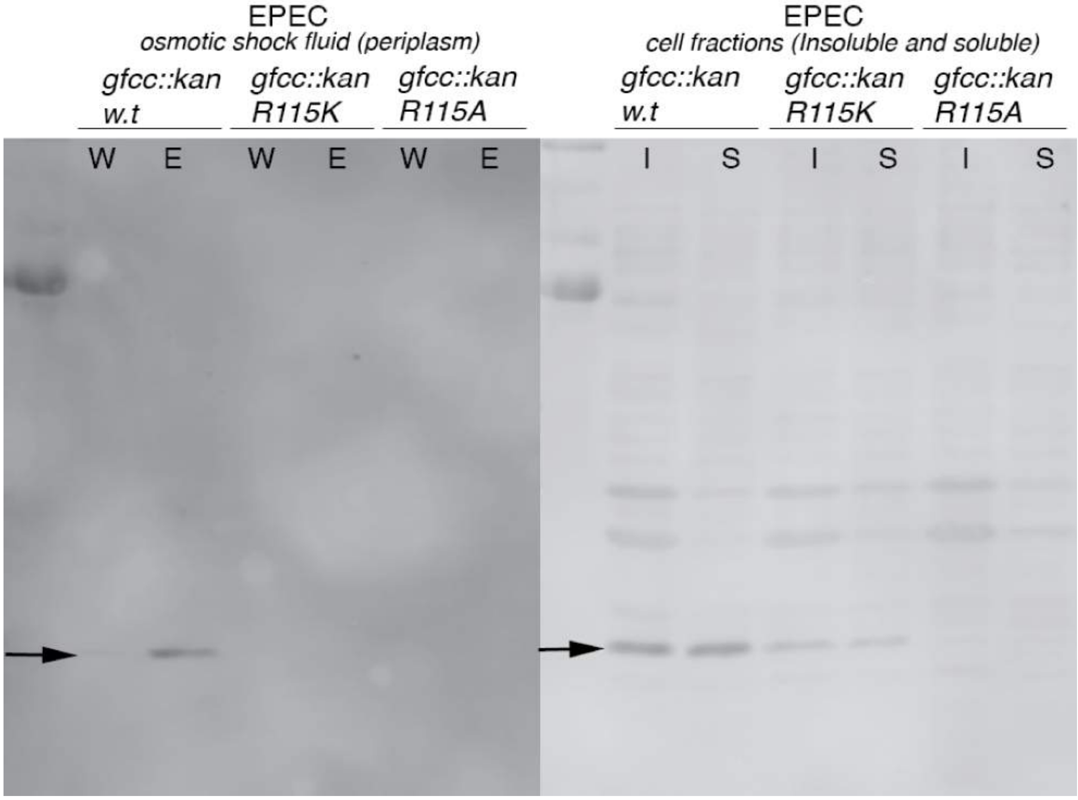


Figure A-2 **R115K and R115A mutants of GfcC show no detectable periplasmic signal** (*left*). This is because the mutation of the conserved invariant Arg affects the folding and stability or expressed protein levels of GfcC and the Arg115 mutants as can be seen from lower amounts expressed from the cellular fractions (*right*). Abbreviations used, W: washes and E: eluted fraction from Talon beads concentration step and I: Insoluble and S: soluble fraction.

References

1. Whitfield, C. and I.S. Roberts, Structure, assembly and regulation of expression of capsules in Escherichia coli. Mol Microbiol, 1999. 31(5): p. 1307-19.

2. Whitfield, C., Biosynthesis and assembly of capsular polysaccharides in Escherichia coli. Annu Rev Biochem, 2006. 75: p. 39-68.

3. Cuthbertson, L., et al., Pivotal roles of the outer membrane polysaccharide export and polysaccharide copolymerase protein families in export of extracellular polysaccharides in gram-negative bacteria. Microbiol Mol Biol Rev, 2009. 73(1): p. 155-77.

4. Peleg, A., et al., Identification of an Escherichia coli operon required for formation of the O-antigen capsule. J Bacteriol, 2005. 187(15): p. 5259-66.

5. Roberts, I.S., The biochemistry and genetics of capsular polysaccharide production in bacteria. Annu Rev Microbiol, 1996. 50: p. 285-315.

6. Cywes, C. and M.R. Wessels, Group A Streptococcus tissue invasion by CD44-mediated cell signalling. Nature, 2001. 414(6864): p. 648-52.

7. Michalek, M.T., C. Mold, and E.G. Bremer, Inhibition of the alternative pathway of human complement by structural analogues of sialic acid. J Immunol, 1988. 140(5): p. 1588-94.

8. Stevens, P., et al., Restricted complement activation by Escherichia coli with the K-1 capsular serotype: a possible role in pathogenicity. J Immunol, 1978. 121(6): p. 2174-80.

9. Brown, E.J., et al., The interaction of C3b bound to pneumococci with factor H (beta 1H globulin), factor I (C3b/C4b inactivator), and properdin factor B of the human complement system. J Immunol, 1983. 131(1): p. 409-15.

10. Danese, P.N., L.A. Pratt, and R. Kolter, Exopolysaccharide production is required for development of Escherichia coli K-12 biofilm architecture. J Bacteriol, 2000. 182(12): p. 3593-6.

11. O'Toole, G., H.B. Kaplan, and R. Kolter, Biofilm formation as microbial development. Annu Rev Microbiol, 2000. 54: p. 49-79.

12. Govan, J.R. and V. Deretic, Microbial pathogenesis in cystic fibrosis: mucoid Pseudomonas aeruginosa and Burkholderia cepacia. Microbiol Rev, 1996. 60(3): p. 539-74.

13. Woodward, R., et al., In vitro bacterial polysaccharide biosynthesis: defining the functions of Wzy and Wzz. Nat Chem Biol, 2010. 6(6): p. 418-23.

14. Raetz, C.R. and C. Whitfield, Lipopolysaccharide endotoxins. Annu Rev Biochem, 2002. 71: p. 635-700.

15. Dong, C., et al., Wza the translocon for E. coli capsular polysaccharides defines a new class of membrane protein. Nature, 2006. 444(7116): p. 226-9.

16. Collins, R.F., et al., The 3D structure of a periplasm-spanning platform required for assembly of group 1 capsular polysaccharides in Escherichia coli. Proc Natl Acad Sci U S A, 2007. 104(7): p. 2390-5.

17. Nesper, J., et al., Translocation of group 1 capsular polysaccharide in Escherichia coli serotype K30. Structural and functional analysis of the outer membrane lipoprotein Wza. J Biol Chem, 2003. 278(50): p. 49763-72.

18. Wugeditsch, T., et al., Phosphorylation of Wzc, a tyrosine autokinase, is essential for assembly of group 1 capsular polysaccharides in Escherichia coli. J Biol Chem, 2001. 276(4): p. 2361-71.

19. Paiment, A., J. Hocking, and C. Whitfield, Impact of phosphorylation of specific residues in the tyrosine autokinase, Wzc, on its activity in assembly of group 1 capsules in Escherichia coli. J Bacteriol, 2002. 184(23): p. 6437-47.

20. Rahn, A., et al., A novel outer membrane protein, Wzi, is involved in surface assembly of the Escherichia coli K30 group 1 capsule. J Bacteriol, 2003. 185(19): p. 5882-90.

21. Corbett, D. and I.S. Roberts, Capsular polysaccharides in Escherichia coli. Adv Appl Microbiol, 2008. 65: p. 1-26.

22. Rigg, G.P., B. Barrett, and I.S. Roberts, The localization of KpsC, S and T, and KfiA, C and D proteins involved in the biosynthesis of the Escherichia coli K5 capsular polysaccharide: evidence for a membrane-bound complex. Microbiology, 1998. 144 (Pt 10): p. 2905-14.

23. Drummelsmith, J. and C. Whitfield, Translocation of group 1 capsular polysaccharide to the surface of Escherichia coli requires a multimeric complex in the outer membrane. Embo J, 2000. 19(1): p. 57-66.

24. Reid, A.N. and C. Whitfield, functional analysis of conserved gene products involved in assembly of Escherichia coli capsules and exopolysaccharides: evidence for molecular recognition between Wza and Wzc for colanic acid biosynthesis. J Bacteriol, 2005. 187(15): p. 5470-81.

25. Collins, R.F., et al., Periplasmic protein-protein contacts in the inner membrane protein Wzc form a tetrameric complex required for the assembly of Escherichia coli group 1 capsules. J Biol Chem, 2006. 281(4): p. 2144-50.

26. Bechet, E., et al., Identification of structural and molecular determinants of the tyrosine-kinase Wzc and implications in capsular polysaccharide export. Mol Microbiol, 2010. 77(5): p. 1315-25.

27. Shifrin, Y., et al., Transient shielding of intimin and the type III secretion system of enterohemorrhagic and enteropathogenic Escherichia coli by a group 4 capsule. J Bacteriol, 2008. 190(14): p. 5063-74.

28. Gibson, D.L., et al., Salmonella produces an O-antigen capsule regulated by AgfD and important for environmental persistence. J Bacteriol, 2006. 188(22): p. 7722-30.

29. Nesper, J., et al., Role of Vibrio cholerae O139 surface polysaccharides in intestinal colonization. Infect Immun, 2002. 70(11): p. 5990-6.

30. Croxatto, A., et al., Vibrio anguillarum colonization of rainbow trout integument requires a DNA locus involved in exopolysaccharide transport and biosynthesis. Environ Microbiol, 2007. 9(2): p. 370-82.

31. Keiski, C.L., et al., AlgK is a TPR-containing protein and the periplasmic component of a novel exopolysaccharide secretin. Structure, 2010. 18(2): p. 265-73.

32. Jain, S., and Ohman, D.E., Alginate biosynthesis. In Pseudomonas, J.-L. Ramos, ed. (New York: Kluwer Academic/Plenum Publishers), 2004: p. pp 53-81.

33. Jain, S. and D.E. Ohman, Deletion of algK in mucoid Pseudomonas aeruginosa blocks alginate polymer formation and results in uronic acid secretion. J Bacteriol, 1998. 180(3): p. 634-41.

34. Robles-Price, A., et al., AlgX is a periplasmic protein required for alginate biosynthesis in Pseudomonas aeruginosa. J Bacteriol, 2004. 186(21): p. 7369-77.

35. Jain, S. and D.E. Ohman, Role of an alginate lyase for alginate transport in mucoid Pseudomonas aeruginosa. Infect Immun, 2005. 73(10): p. 6429-36.

36. Rehm, B.H., et al., Overexpression of algE in Escherichia coli: subcellular localization, purification, and ion channel properties. J Bacteriol, 1994. 176(18): p. 5639-47.

37. Remminghorst, U. and B.H. Rehm, In vitro alginate polymerization and the functional role of Alg8 in alginate production by Pseudomonas aeruginosa. Appl Environ Microbiol, 2006. 72(1): p. 298-305.

38. D'Andrea, L.D. and L. Regan, TPR proteins: the versatile helix. Trends Biochem Sci, 2003. 28(12): p. 655-62.

39. Arrecubieta, C., et al., The transport of group 2 capsular polysaccharides across the periplasmic space in Escherichia coli. Roles for the KpsE and KpsD proteins. J Biol Chem, 2001. 276(6): p. 4245-50.

40. Marcotte, E.M., et al., A combined algorithm for genome-wide prediction of protein function. Nature, 1999. 402(6757): p. 83-6.

41. Wilkins, M.R., et al., Protein identification and analysis tools in the ExPASy server. Methods Mol Biol, 1999. 112: p. 531-52.

42. Minor, Z.O.a.W., Processing of X-ray Diffraction Data Collected in Oscillation Mode Methods in Enzymology, 1997. 276: Macromolecular Crystallography, Part A: p. 307-326.

43. Potterton, E., et al., The CCP4 molecular-graphics project. Acta Crystallogr D Biol Crystallogr, 2002. 58(Pt 11): p. 1955-7.

44. Adams, P.D., et al., PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr, 2010. 66(Pt 2): p. 213-21.

45. Terwilliger, T., SOLVE and RESOLVE: automated structure solution, density modification and model building. J Synchrotron Radiat, 2004. 11(Pt 1): p. 49-52.

46. Emsley, P., et al., Features and development of Coot. Acta Crystallogr D Biol Crystallogr, 2010. 66(Pt 4): p. 486-501.

47. McCoy, A.J., et al., Phaser crystallographic software. J Appl Crystallogr, 2007. 40(Pt 4): p. 658-674.

48. Chen, V.B., et al., MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr, 2010. 66(Pt 1): p. 12-21.

49. Waterhouse, A.M., et al., Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics, 2009. 25(9): p. 1189-91.

50. Burroughs, A.M., et al., A novel superfamily containing the beta-grasp fold involved in binding diverse soluble ligands. Biol Direct, 2007. 2: p. 4.

51. Holm, L. and P. Rosenstrom, Dali server: conservation mapping in 3D. Nucleic Acids Res, 2010. 38 Suppl: p. W545-9.

52. Kleywegt, G.J. and T.A. Jones, Detecting folding motifs and similarities in protein structures. Methods Enzymol, 1997. 277: p. 525-45.

53. Ashkenazy, H., et al., ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res, 2010. 38 Suppl: p. W529-33.

54. Dundas, J., et al., CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res, 2006. 34(Web Server issue): p. W116-8.

55. Trott, O. and A.J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem, 2010. 31(2): p. 455-61.

56. Moustakas, D.T., et al., Development and validation of a modular, extensible docking program: DOCK 5. J Comput Aided Mol Des, 2006. 20(10-11): p. 601-19.

57. Ford, R.C., et al., Structure-function relationships of the outer membrane translocon Wza investigated by cryo-electron microscopy and mutagenesis. J Struct Biol, 2009. 166(2): p. 172-82.

58. Higgins, M.K., et al., Structure of the periplasmic component of a bacterial drug efflux pump. Proc Natl Acad Sci U S A, 2004. 101(27): p. 9994-9.

59. Ziegler, K., R. Benz, and G.E. Schulz, A putative alpha-helical porin from Corynebacterium glutamicum. J Mol Biol, 2008. 379(3): p. 482-91.

60. Chandran, V., et al., Structure of the outer membrane complex of a type IV secretion system. Nature, 2009. 462(7276): p. 1011-5.

61. Hay, I.D., Z.U. Rehman, and B.H. Rehm, Membrane topology of outer membrane protein AlgE, which is required for alginate production in Pseudomonas aeruginosa. Appl Environ Microbiol, 2010. 76(6): p. 1806-12.

62. Remmert, M., et al., HHomp--prediction and classification of outer membrane proteins. Nucleic Acids Res, 2009. 37(Web Server issue): p. W446-51.

63. Hou, S., et al., Genome sequence of the deep-sea gamma-proteobacterium Idiomarina loihiensis reveals amino acid fermentation as a source of carbon and energy. Proc Natl Acad Sci U S A, 2004. 101(52): p. 18036-41.

64. Vijayalakshmi, J., B.J. Akerley, and M.A. Saper, Structure of YraM, a protein essential for growth of Haemophilus influenzae. Proteins, 2008. 73(1): p. 204-17.

65. Akerley, B.J., et al., A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. Proc Natl Acad Sci U S A, 2002. 99(2): p. 966-71.

66. Lopez-Campistrous, A., et al., Localization, annotation, and comparison of the Escherichia coli K-12 proteome under two states of growth. Mol Cell Proteomics, 2005. 4(8): p. 1205-9.

67. Trakhanov, S., et al., Ligand-free and -bound structures of the binding protein (LivJ) of the Escherichia coli ABC leucine/isoleucine/valine transport system: trajectory and dynamics of the interdomain rotation and ligand specificity. Biochemistry, 2005. 44(17): p. 6597-608.

68. Sack, J.S., M.A. Saper, and F.A. Quiocho, Periplasmic binding protein structure and function. Refined X-ray structures of the leucine/isoleucine/valine-binding protein and its complex with leucine. J Mol Biol, 1989. 206(1): p. 171-91.

69. Valdar, W.S., Scoring residue conservation. Proteins, 2002. 48(2): p. 227-41.

70. Kajander, T., et al., Structure and stability of designed TPR protein superhelices: unusual crystal packing and implications for natural TPR proteins. Acta Crystallogr D Biol Crystallogr, 2007. 63(Pt 7): p. 800-11.

71. Kim, K., et al., Crystal structure of PilF: functional implication in the type 4 pilus biogenesis in Pseudomonas aeruginosa. Biochem Biophys Res Commun, 2006. 340(4): p. 1028-38.

72. Park, J.T., et al., MppA, a periplasmic binding protein essential for import of the bacterial cell wall peptide L-alanyl-gamma-D-glutamyl-meso-diaminopimelate. J Bacteriol, 1998. 180(5): p. 1215-23.

73. Hansson, M. and L. Hederstedt, Bacillus subtilis HemY is a peripheral membrane protein essential for protoheme IX synthesis which can oxidize coproporphyrinogen III and protoporphyrinogen IX. J Bacteriol, 1994. 176(19): p. 5962-70.

74. Krebs, W.G., et al., Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. Proteins, 2002. 48(4): p. 682-95.

75. Onufryk, C., et al., Characterization of six lipoproteins in the sigmaE regulon. J Bacteriol, 2005. 187(13): p. 4552-61.

76. Keyamura, K., et al., The interaction of DiaA and DnaA regulates the replication cycle in E. coli by directly promoting ATP DnaA-specific initiation complexes. Genes Dev, 2007. 21(16): p. 2083-99.

77. van Asselt, E.J., A.M. Thunnissen, and B.W. Dijkstra, High resolution crystal structures of the Escherichia coli lytic transglycosylase Slt70 and its complex with a peptidoglycan fragment. J Mol Biol, 1999. 291(4): p. 877-98.

78. Matias, V.R., et al., Cryo-transmission electron microscopy of frozen-hydrated sections of Escherichia coli and Pseudomonas aeruginosa. J Bacteriol, 2003. 185(20): p. 6112-8.

79. Hu Y, C.L., Ha S, Gross B, Falcone B, Walker D, Mokhtarzadeh M, Walker S, Crystal structure of the MurG:UDP-GlcNAc complex reveals common structural principles of a superfamily of glycosyltransferases. Proc Natl Acad Sci USA, 2003. 100: p. 845-849.