# Structural Analysis of Galactofuranose-Binding Lectins and Biosynthetic Enzymes

By

Kittikhun Wangkanont

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

UNIVERSITY OF WISCONSIN-MADISON

2016

Date of final oral examination: 12/14/2015

The dissertation is approved by the following members of the Final Oral Committee:
Laura L. Kiessling, Professor, Chemistry and Biochemistry
Katrina T. Forest, Professor, Bacteriology
Helen E. Blackwell, Professor, Chemistry
Judith N. Burstyn, Professor, Chemistry
Douglas B. Weibel, Associate Professor, Chemistry, Biochemistry, and Biomedical Engineering

## Abstract

Galactofuranose (Gal*f*) is a carbohydrate residue essential for pathogenic organisms, such as *Mycobacterium tuberculosis*, that is not found in mammalian systems. Therefore, tools that recognize Gal*f* specifically would be useful for diagnostic purposes and the development of targeted therapeutics. In addition, the biosynthetic enzyme UDP-galactopyranose mutase (UGM) essential for Gal*f* incorporation is an attractive antibiotic target.

To examine how Gal*f* is recognized, the structures of microbe-binding intelectins were investigated. Because no structure of any protein in the intelectin family was available, a crystal structure of *Xenopus* embryonic epidermal lectin (XEEL) was determined using experimental phasing. The XEEL structure enabled solution of the structure of human intelectin-1 (hIntL-1) as well as structures of XEEL and hIntL-1 bound to their carbohydrate ligands. These structures revealed intelectins to be structurally distinct from other lectins. They also showed that intelectins use an unusual ligand recognition mechanism, a protein-bound calcium ion is chelated by the exocyclic vicinal diol on the ligands. These structures expanded our appreciation of lectin diversity, and could aid in rational engineering of intelectins for microbe detection and elimination.

In addition to Gal*f* detection, inhibition of the essential Gal*f* biosynthesis enzyme UGM was investigated. Triazolothiadiazine inhibitors discovered through virtual screening showed low micromolar inhibitory constants and were effective against virulent strains of *Mycobacterium tuberculosis*. A structure of a triazolothiadiazine inhibitor bound to *Corynebacterium diphtheriae* UGM (CdUGM) was determined. The

structure not only shows the first small molecule, non-substrate inhibitor bound to UGM, but also suggests the role of UGM conformation in inhibitor design. Additional CdUGM and *Mycobacterium smegmatis* UGM (MsUGM) structures were also determined to examine structural dynamics and rationalize inhibitor affinity across UGM homologs. The inhibitor-bound CdUGM structure, and various forms of CdUGMs and MsUGMs, should enable development of next generation inhibitors that could target a wide range of UGMs.

**Acknowledgements**

Many people have helped me getting through graduate school. First, I would like to thank my advisor Prof. Laura L. Kiessling. I started in her research group as an undergraduate student and she also welcome me back as a graduate student. Laura has helped me tremendously to become a good scientist and a better writer. I am also indebted to my collaborator, Prof. Katrina T. Forest. She helped me get started in structural biology and also welcome me to be a part of her research group. I would also like to thank Dr. Kenneth A. Satyshur for his assistant with diffraction data collection and maintenance of the computer system that make much of my thesis work possible.

All X-ray diffraction data in this thesis were collected at various synchrotron light sources. Several beamline scientists and administrators have helped me over the years. Dr. Keith Brister and Nancy Brennan at the Life Science Collaborative Access Team (LS-CAT, Advanced Photon Source, Argonne National Laboratory) have been generous with me requesting extra beam time. Thanks to Dr. Elena Kondrashkina, Dr. David Smith, Dr. Zdzislaw Wawrzak, Dr. Spencer Anderson, and Dr. Joseph Brunzelle at LS-CAT who helped me learn good data collection practice and basic beamline operations. At the National Synchrotron Light Source (Brookhaven National Laboratory), Dr. Annie Héroux screened and collected several data sets for the UGM inhibitor project. Without her dedication and her superb beamline, the UGM-inhibitor complex structure would not have been possible. Several crystals were also screened at the Stanford Synchrotron Radiation Light Source (SLAC National Accelerator Laboratory). I would like to thank Lisa Dunn for promptly handling my time request, and Dr. Tzanko Doukov and Dr. Silvia Russi for hosting me both on site and remotely.

In addition to the synchrotron staffs, several crystallographers have helped advanced my career in numerous ways. I am grateful to know Dr. Matthew Benning (Bruker AXS) who assisted me with Burker softwares and also helped in SAXS data collection and processing. Thanks to Dr. Bernhard Lohkamp and Dr. Andrey Lebedev for insightful discussion on lattice-translocation disorder and help processing difficult data.

These projects would not have succeeded without the support of facility managers at the University of Wisconsin-Madison. Dr. Darrell McCaslin helped with centrifugations experiments. Grzegorz Sabat and Dr. Greg Barrett-Wilt gave practical advice regarding mass spectrometry and processed numerous samples for several projects. Edman degradation experiments were carried out by Dr. Michael Berne (Tufts University Core Facility).

Other group members, past and present, are also instrumental in success of many projects. Dr. Lingyin Li and Prof. Andrew Hinck has given valuable advice for the TBRII project. Darryl A. Wesener has worked closely with me on the intelectin projects resulting in fruitful publications. Virginia A. Kincaid and Valerie Winton have been productive collaborators on the UGM project. In addition, I would like to thank my undergraduate assistant Jack A. Vidani who had been working with me for almost the entire duration of his undergraduate career. Thanks to Dr. Sayaka Masuko, Dr. Amanda Dugan, Alex M. Justen, Dr. Paul J. Wrighton, Daniel B. Zwick, and Anna W. Baker for proofreading various manuscripts and thesis chapters.

I would not have made it through graduate school without several great friends. Ian Windsor, Ian Norden, Brandon Kilduff, Floriana Foarta, Dr. John Lukesh have provided continuing support throughout my time here.

I would like to thank the Development and Promotion of Science and Technology Talents Project of Thailand for financial support. Most importantly, I would like to thank my parents and grandmother for their unconditional love and understanding, especially while I spend many years abroad.

# Table of Contents

## List of Figures

## List of Tables

## Abbreviations

| | |
|---|---|
| AAA | *Anguilla anguilla* agglutinin |
| AUC | Analytical ultracentrifugation |
| BAC | Bacteria artificial chromosome |
| BES | N,N-Bis(2-hydroxyethyl)-2-aminoethanesulfonic acid |
| Bis-Tris | Bis(2-hydroxyethyl)amino-tris(hydroxymethyl)methane |
| BSA | Bovine serum albumin |
| Cα | Alpha carbon of an amino acid |
| CdUGM | *Corynebacterium diphtheriae* UGM |
| CeUGM | *Caenorhabditis elegans* UGM |
| CRD | Carbohydrate recognition domain |
| DIGS | Demodulation of intensities by grouping and scaling |
| DrIntL | *Danio rerio* intelectin |
| DTT | Dithiothreitol |
| ECD | Extracellular domain |
| EDTA | Ethylenediaminetetraacetic acid |
| ELISA | Enzyme-linked immunosorbent assay |
| ER | Endoplasmic reticulum |
| FAD | Flavin adenine dinucleotide |
| FBD | Fibrinogen-like domain |
| Gal*f* | Galactofuranose |
| GalNAc | N-acetyl galactosamine |

| | |
|---|---|
| Gal*p* | Galactopyranose |
| GlcNAc | N-acetyl glucosamine |
| GroP | D-glycerol 1-phosphate |
| GSH | Reduced glutathione |
| GSSG | Oxidized glutathione |
| HEK293T | Human embryonic kidney 293T cells |
| HEPES | 4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid |
| hIntL | Human intelectin |
| HRP | Horseradish peroxidase |
| IPTG | Isopropyl β-D-1-thiogalactopyranoside |
| KDO | 3-deoxy-D-*manno*-oct-2-ulosonic acid |
| KO | D-glycero-D-*talo*-oct-2-ulosonic acid |
| KpUGM | *Klebsiella pneumoniae* UGM |
| mIntL | Mouse intelectin |
| MPD | 2-Methyl-2,4-pentanediol |
| MsUGM | *Mycobacterium smegmatis* UGM |
| MtUGM | *Mycobacterium tuberculosis* UGM |
| NCS | Non-crystallographic symmetry |
| NDSB | Non-detergent sulfobetaine |
| Neu5Ac | N-acetylneuraminic acid |
| Ni-NTA | Nickel nitrilotriacetic acid |
| $OD_{600}$ | Optical density at 600 nm |
| PAGE | Polyacrylamide gel electrophoresis |

| | |
|---|---|
| PBS | Phosphate buffer saline |
| PBST | Phosphate buffer saline with 0.05% Tween-20 |
| PDB | Protein data bank |
| PEG | Polyethylene glycol |
| PMSF | Phenylmethylsulfonyl fluoride |
| Rib*f* | Ribofuranose |
| RMSD | Root-mean-square deviation |
| SAD | Single-wavelength anomalous diffraction |
| SAXS | Small-angel X-ray scattering |
| SDS | Sodium dodecyl sulfate |
| Se-Met | Selenomethionine |
| SPR | Surface plasmon resonance |
| TBRI | Type I TGF-β receptor |
| TBRII | Type II TGF-β receptor |
| TEV | Tobacco etch virus |
| TGF-β | Transforming growth factor- β |
| TLS | Translation-libration-screw |
| TMB | 3,3′,5,5′-Tetramethylbenzidine |
| tNCS | Translational non-crystallographic symmetry |
| Tris | Tris(hydroxymethyl)aminomethane |
| Trx | Thioredoxin |
| XCGL-1 | *Xenopus laevis* cortical granule lectin 1 |
| XCGL-2 | *Xenopus laevis* cortical granule lectin 2 |

| | |
|---|---|
| XEEL | *Xenopus laevis* embryonic epidermal lectin |
| XSL-1 | *Xenopus laevis* serum lectin 1 |
| XSL-2 | *Xenopus laevis* serum lectin 2 |
| UDP | Uridine 5′-diphosphate |
| UGM | UDP-galactopyranose mutase |
| UGM6 | 2-[(7S)-3-(2-furyl)-6-phenyl-7H-[1,2,4]triazolo[3,4-b][1,3,4]thiadiazin-7-yl]acetic acid |
| UGM6.7 | 2-[4-(4-chlorophenyl)-7-(2-thienyl)-2-thia-5,6,8,9-tetrazabicyclo[4.3.0]nona-4,7,9-trien-3-yl]acetic acid |
| UV | Ultraviolet light |

# Chapter 1

# Structure and function of animal lectins

## 1.1 Abstract

Virtually all cells are covered in a carbohydrate coat composed of polysaccharides, glycoproteins, and glycolipids (glycans). Surface glycans are involved in several crucial biological processes such as cellular attachment, embryonic development, and immune defense. Molecular recognition events of these surface glycans are mediated through carbohydrate binding proteins or lectins. Because 35 and over 700 monosaccharide building blocks are used in mammalian and microbial glycan synthesis, respectively[1, 2], there is also great diversity of lectins that recognize these carbohydrate residues. In animals, lectins are generally classified according to structural similarities, which usually reflect the scope of their carbohydrate ligands. Knowledge of lectin structures and ligand specificity is not only crucial in understanding their biological function but lectin-specific recognition could also be used as tools for cell identification and targeting. This introductory chapter summarizes the biological roles of glycans in animals with emphasis on the recognition of glycans by lectins. The background information highlights recent progress in understanding structures, ligand recognition mechanisms, and biological function of animal lectins.

## 1.2 Biological roles of glycans

Glycans are carbohydrate polymers attached to proteins or lipid molecules on the cell surface. These surface glycoconjugates participate in crucial biological processes such as embryonic development and recognition of foreign cells[3]. The biological function of glycans could be broadly divided into three overlapping categories: structural roles, regulatory functions, and molecular recognition markers.

Unlike plant or bacteria, animal cells lack rigid cell walls. However, animal cells are covered with the glycocalyx which provide structural integrity[4]. In addition, heavily glycosylated proteins called proteoglycans are part of the extracellular matrix that help maintain tissue integrity[5]. Mutations in proteoglycans are associated with connective tissue diseases. For example, mutations in the proteoglycan aggrecan causes cartilage matrix deficiency and aberrant skeletal development[6, 7]. In addition to providing structural framework, these glycan also provide binding sites for cell adhesion molecules for tissue organization. Neural cell adhesion molecule (NCAM) binds heparan sulfate proteoglycan and is crucial for cell-cell adhesion and formation of neural tissue[8], further highlighting the structural roles of glycans.

In addition to structural roles, glycans also provide regulatory function for cells that interact with them. For example, basic fibroblast growth factor (bFGF) signaling is enhanced when heparan sulfate proteoglycan is present as a co-receptor[9]. Extracellular matrix glycans are also attachment sites for various growth factors. These growth factors are important for pattern formation by modulating morphogen gradient during embryonic development. It is proposed that proteoglycan controls diffusion of Wg morphogen gradient essential in wing disc formation in *Drosophila*[10]. In addition to regulatory function by spatially modulating signaling pathways, attachment of glycans to protein can also help in proper protein folding and localization. Several glycosyltransferases and glycohydrolases modify N-glycan structure in the endoplasmic reticulum (ER) to direct the folding pathway or divert the misfolded protein for degradation pathways[11]. Thus, the interaction of glycans to other macromolecules or cells provides regulatory function in addition to structural support.

The structural and regulatory roles of glycans depend also on their ability to participate in molecular recognition of glycans. For example, calnexin and calreticulins are ER chaperones that bind monoglucosylated intermediate of N-linked glycan and recruit other chaperones for proper protein folding[12]. Such molecular recognition of glycans are mediated by carbohydrate binding proteins or lectins. In a broader context, lectins are key players in vital biological processes that involve interactions between glycan and cells or other biological molecules. For instance, the adhesion of lymphocytes to epithelial cells requires selectins as adhesion molecules[13]. Oligomeric C-type lectins in viper and cobra venoms crosslink platelet receptors and induce blood clots in the prey[14]. However, the most studied aspect of animal lectins is their role in the immune system, both adaptive and innate. Antigen-presenting cells such as dendritic cells express several receptors containing C-type lectin domains[15]. These receptors are known to interact with pathogens[16]. Galectins bind bacteria in the gut and modulate mucosal immune response[17]. Ficolins and serum mannose-binding lectins recognize microbes and initiate the lectin complement pathway[18]. Given the diverse biological roles of lectins, lectins are attractive targets for scientific investigation not only because of their intriguing structural and ligand-specificity diversity but also their potential as biological probes or therapeutics.

### 1.3 Animal lectin classification and carbohydrate recognition mechanism

The diversity of lectins reflects the variety of biological processes that they are involved in and of the glycan structures present in each system. Mammals use approximately 35 monosaccharide building blocks for glycan biosynthesis[2]; however, bacteria utilize over 700 monosaccharide[1]. Recognition of these glycans by animal

lectins are crucial in both normal physiological processes and recognition of foreign epitopes. Animal lectins are classified based on sequence and structural homology. Generally, the homology within each lectin family implies the recognition of a similar ligand. For example, galectins share eight highly conserved residues at the ligand binding site[19]. As a result, galectins generally bind β-galactosides, hence the name galectins. There have been at least 16 families of animal lectins characterized to date. Most lectins are well-characterized structurally with the exception of the X-type lectin family, in which no 3-dimensional structure is available prior to this thesis. The remainder of this chapter provides background information for the structure and function of known animal lectin families.

### 1.3.1 Annexins

Annexins are scaffolding proteins involved in membrane organization through interactions with negatively charged phospholipids[20]. Thus, annexins regulate endocytosis, exocytosis, and membrane homeostasis in various organelles[21]. However, annexin A2 and annexin A4 are secreted proteins that interact with heparin and heparan sulfate. The complex acts as a docking site for tissue plasminogen activator to regulate coagulation and fibrinolysis[22, 23]. Annexin A2 binds its ligand in a calcium dependent manner[22]. The structure of annexin A2 consists of four homologous repeats, each containing five helices[24]. The structure of annexin A2 with heparin tetrasaccharide reveals that it binds its ligand on the convex face (Figure 1.1 A)[24]. The sulfate groups of heparin interact with a protein-bound calcium ion (Figure 1.1 B). Negatively charged protein residues around the binding site also form additional salt bridges to the heparin ligand.

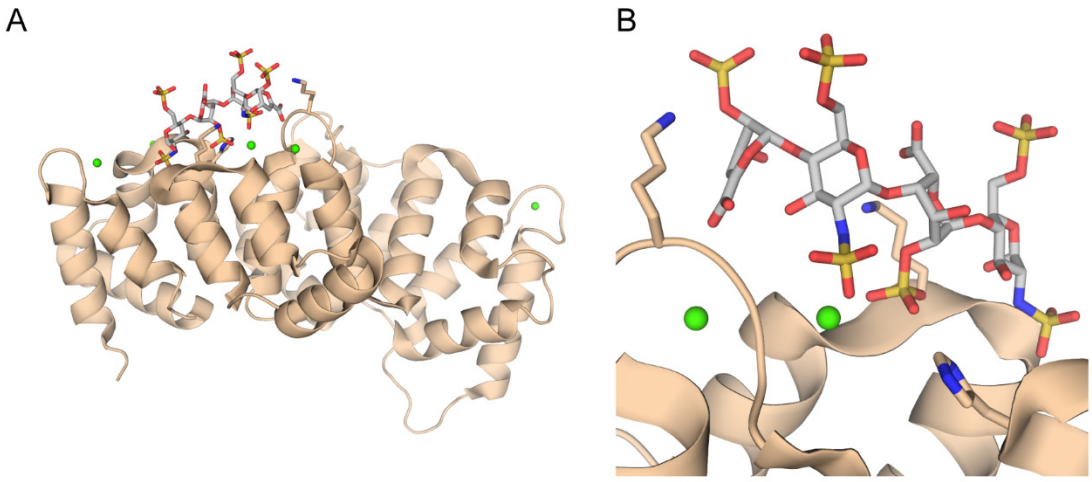



**Figure 1.1**: (A) Overall structure of annexin A2. (B) Ligand binding site of annexin A2 with bound calcium ions (green).

### 1.3.2 Calnexin and Calreticulin

Calnexin and calreticulin are chaperones that assist the folding of N-linked glycoproteins in the ER[12]. Calnexin is a type I membrane protein with a single transmembrane helix whereas calreticulin is a soluble protein[25]. Both proteins share about 40% sequence identity and recognize monoglucosylated oligosaccharide of N-liked glycoproteins[26]. The monoglucosylated oligosaccharide is a signature of improperly folded protein. Thus interaction of calnexin and calreticulin results in ER retention and recruitment of other chaperone to assist folding or disposal of misfolded protein[27]. Both calnexin and calreticulin contain a globular lectin domain and an elongated arm domain (Figure 1.2 A)[28, 29]. The globular lectin domain of calreticulin contains a shallow carbohydrate binding site with hydrogen bonding between ring hydroxyl groups and polar side chains (Figure 1.2 B)[29]. Hydrophobic protein side chains also interact with the hydrophobic face of the carbohydrate ring. Although a calcium ion is present in the globular lectin domain, the calcium ion is remote from the ligand binding site, thus presumably has structural role.

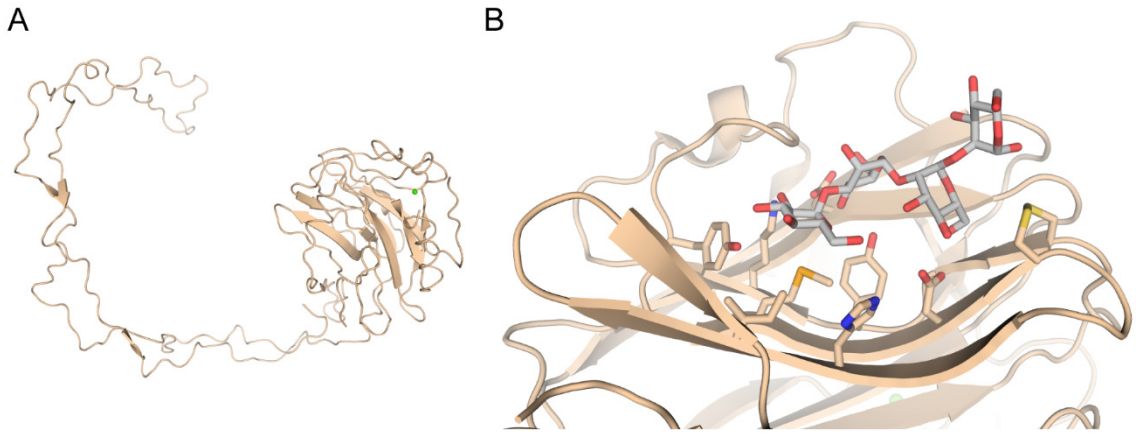



**Figure 1.2**: (A) Structure of calnexin showing globular and arm domains. (B) Ligand binding site of calreticulin in the globular lectin domain.

### 1.3.3 Chitinase-like lectins

Chitinase-like lectins are expressed in various animal tissues[30]. However, their biological function is poorly understood. Human cartilage glycoprotein-39 (chitinase 3-like protein 1 or YKL-40) is proposed to participate in tissue remodeling and inflammation[31]. Ym1 is a murine protein secreted by macrophages upon nematode infection, and may have anti-inflammatory properties[32]. Oviductin is secreted by mannalian oviductal epithelial cells and may play a role in fertilization and early embryonic development[33]. Although the exact physiological function of these lectins is not known, structural information is available for YKL-40 and Ym1. These chitinase-like lectins are catalytically inactive members of the glycoside hydrolase family 18 with a triose-phosphate isomerase fold[34]. The structure of YKL-40 in complex with GlcNAc oligosaccharide reveals a deep groove with extensive hydrogen bonding from the protein residue side chain to the carbohydrate hydroxyl groups[35]. There are also several tryptophan residues that stack onto the CH face of the carbohydrate residues.

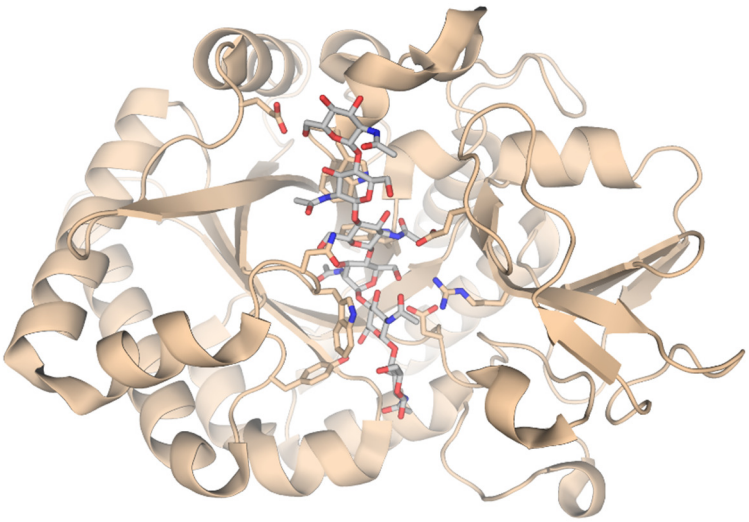


**Figure 1.3**: Structure of YKL-40 with GlcNAc oligosaccharide bound.

### 1.3.4 C-type lectins

C-type lectins require calcium ion for ligand binding[36]. The C-type lectin domain is ubiquitous in nature, usually in combination with other protein domains. Classification of C-type lectin, as well as C-type lectin-like domains, into at least 14 groups based on structural organization has been proposed[37, 38]. C-type lectins are involved in several biological processes, notably in the immune system[39]. Serum mannose-binding lectin binds microbes and activates the complement system[18]. The selectin family is involved in cell adhesion and localization of immune cells to the site of inflammation[13]. Antigen-presenting cells express several C-type lectin receptors which are responsible for pathogen recognition, or sometimes exploited by pathogens for dissemination[15]. The C-type lectin domain of a dendritic cell C-type lectin receptor DC-SIGN in complex with tetramannose shows three calcium ions (Figure 1.4 A)[40]. One of these calcium ions is held in a loop and is coordinated by the equatorial O(3) and (O4) of a terminal mannose residue (Figure 1.4 B)[40]. The protein residues that coordinate the calcium ion in the

ligand binding site also form hydrogen bonds with mannose ring hydroxyl groups. Other mannose residues on tetramannose interact with DC-SIGN through hydrogen bonds.

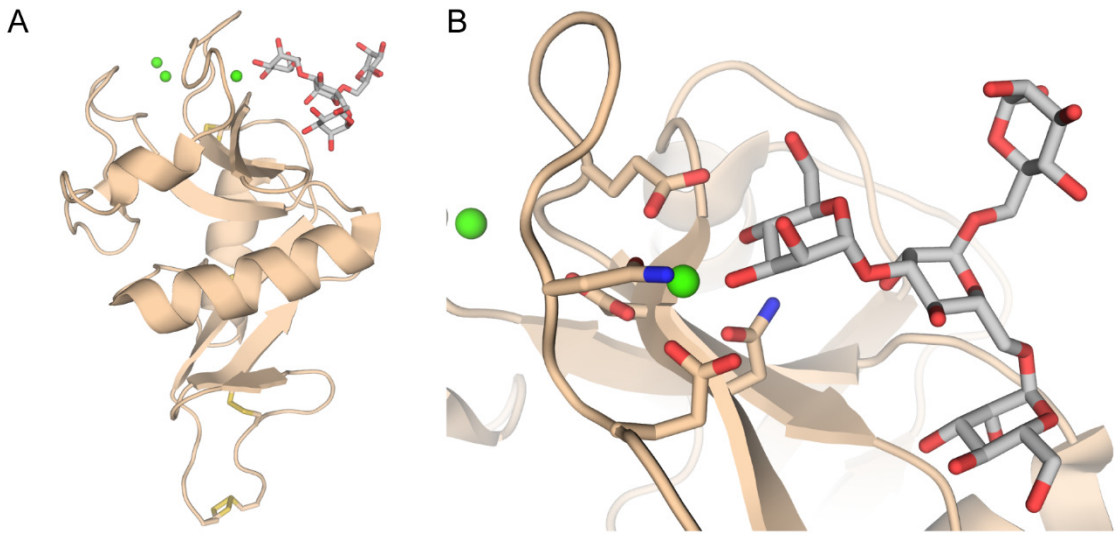



**Figure 1.4**: (A) Structure of a C-type lectin DC-SIGN with calcium ions (green). (B) Ligand binding site of DC-SIGN with tetramannose bound.

### 1.3.5 F-box lectins

The F-box lectin domain is found in several proteins in the ubiquitin-proteasome system involved in glycoprotein degradation[41]. The domain recognize GlcNAc disaccharide which is directly attached to glycoproteins and is usually covered in the folded protein. However, this disaccharide motif is accessible in unfolded glycoprotein exported from the ER[42]. Recognition of the GlcNAc disaccharide leads to ubiquitin attachment and degradation of misfolded protein[43]. The structure of Fbs1 shows a β-sandwich fold with ligand binding loop on one side and two short capping helices on the opposite side[44]. Unlike ligand binding sites in other lectins that are defined pockets, the ligand binding site of Fbs1 is a loop that fits in between the monosacchride unit of GlcNAc disaccharide[44]. The backbone amide of the loop forms hydrogen bonds with the

carbohydrate hydroxyl groups. Stacking of the carbohydrate CH onto the π face of a tryptophan indole ring is observed.

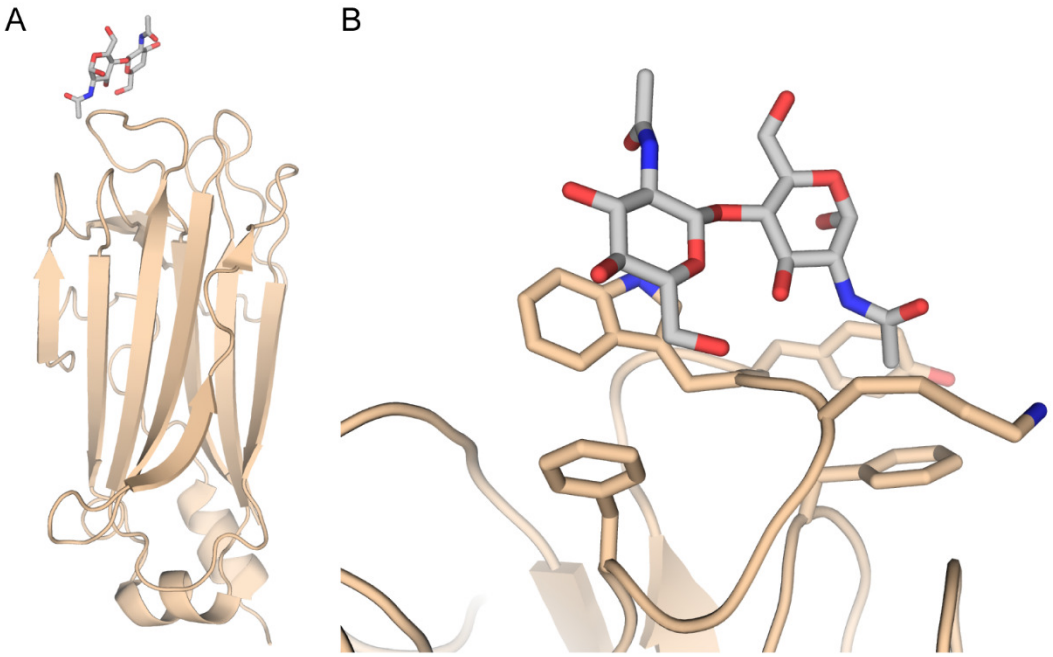



**Figure 1.5**: (A) Structure of an F-box lectin domain of Fbs1. (B) Ligand binding site of Fbs1 with bound GlcNAc disaccharide.

### 1.3.6 Ficolins

Ficolins are soluble immune lectin that bind N-acetyl sugars, as well as galactose and fucose, on microbes[45]. Upon microbe binding, the complex either activate the complement system or acts as an opsonin for phagocytosis of targeted microbe[46]. Humans have three ficolins: L-ficolin, H-ficolin, and M-ficolin[47]. L- and H-ficolins are serum proteins[48, 49], whereas M-ficolin is found in the secretory granules of monocytes and neutrophils[50, 51]. Their function and mode of action are similar. The collagen-like domain in ficolins mediates assembly of the carbohydrate recognition domains into higher-order oligomers[52]. The carbohydrate recognition domain contains a fibrinogen-like domain (FBD) proposed to be a signature of innate immune defense protein[53]. The

carbohydrate recognition domain of L-ficolin contains a structural calcium ion (Figure 1.6 A)[54]. The ligand binding site appear to vary depending on the ligand identity[54]. The galactose binding site is the cleft formed by the twisted β-sheet of the FBD[54]. The ring hydroxyls of galactose is recognized by several acidic residues as well as an arginine and a serine.

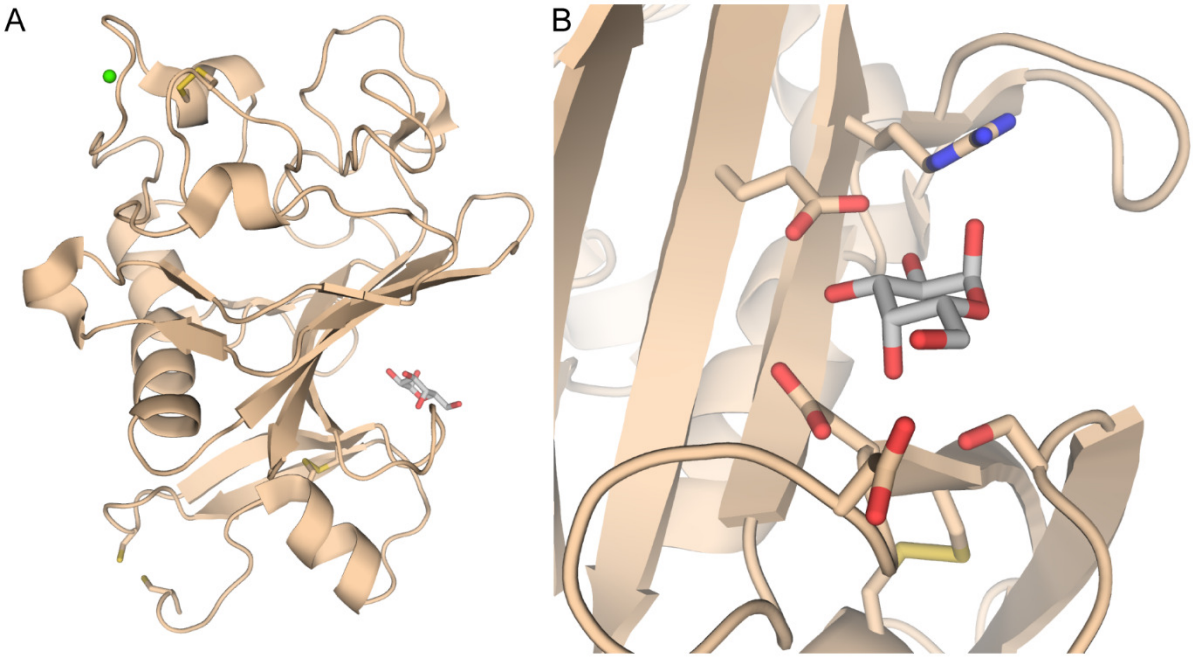



**Figure 1.6**: (A) Structure of L-ficolin with a structural calcium ion (green). (B) A putative ligand binding site with galactose bound.

### 1.3.7 F-type lectins

F-type lectins or fucolectins bind a fucose residue[55]. The lectin family scatter throughout evolutionary tree from bacteria to fishes, but seems to have lost its expression in mammals. The F-type lectin domain is often found in multi-domain proteins with diverse domain architectures[56]. These proteins are implicated in microbe binding and the innate immune response[57]. A striped bass F-type lectin, FBP32, is upregulated during inflammation[58, 59]. F-type lectin was first characterized in eels[60].

*Anguilla anguilla* agglutinin (AAA) is an F-type lectin from the European eel that binds a fucose residue in the H blood group antigen, thus AAA is used in blood typing[61]. The structure of AAA shows a β jellyroll fold with a structural calcium ion (Figure 1.7 A)[62]. The oligomeric state of the protein is controversial, but it is likely trimeric based on the crystal structure[62]. The ligand binding site is highly positively charged with histidine and arginine residues forming hydrogen bonds with hydroxyl groups of the fucose ring[62]. The binding site also features a disulfide bond formed by two consecutive cysteines.

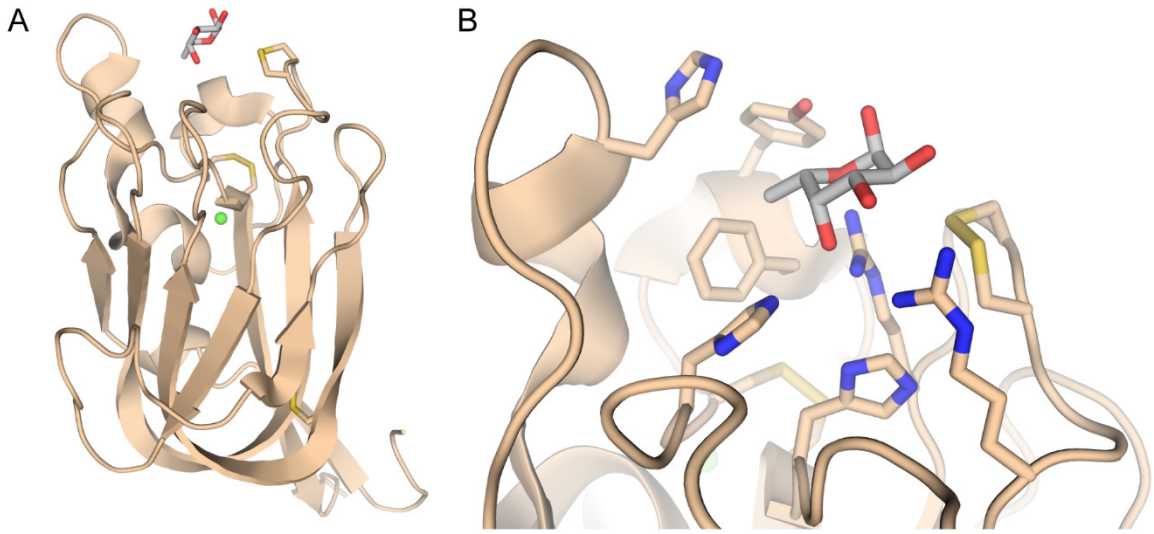



**Figure 1.7**: (A) Structure of AAA monomer with a structural calcium ion (green). (B) Ligand binding site of AAA with bound fucose.

### 1.3.8 Galectins

Galectin is a large galactoside-binding lectin family that is ubiquitous among animals[19]. Galectins are involved in various biological processes such as cell signaling, cell adhesion, development, and cell death[63-66]. Thus, galectins are important targets for immunomodulation and cancer therapy[67, 68]. Galectins have various molecular arrangements and are categorized into three classes: prototypical, chimeric, and tandem-repeat[69]. Prototypical galectins (e.g. galectin-1) contain a single carbohydrate

recognition domain that may be homodimeric. Chimeric galectins (e.g. galectin-3) contain other unrelated domains in addition to the carbohydrate recognition domain. Tandem-repeat galectins (e.g. galectin-1) have multiple carbohydrate recognition domains in a single polypeptide chain. The carbohydrate recognition domain adopts a β-sandwich structure (Figure 1.8 A)[70]. The ligand binding site contains eight highly conserved residues that interact with galactosides[19]. The galactose ring stack onto the π face of a tryptophan residue. Other conserved binding site residues form a shallow pocket around the galactoside and form hydrogen bonding network with the carbohydrate ring hydroxyl groups.

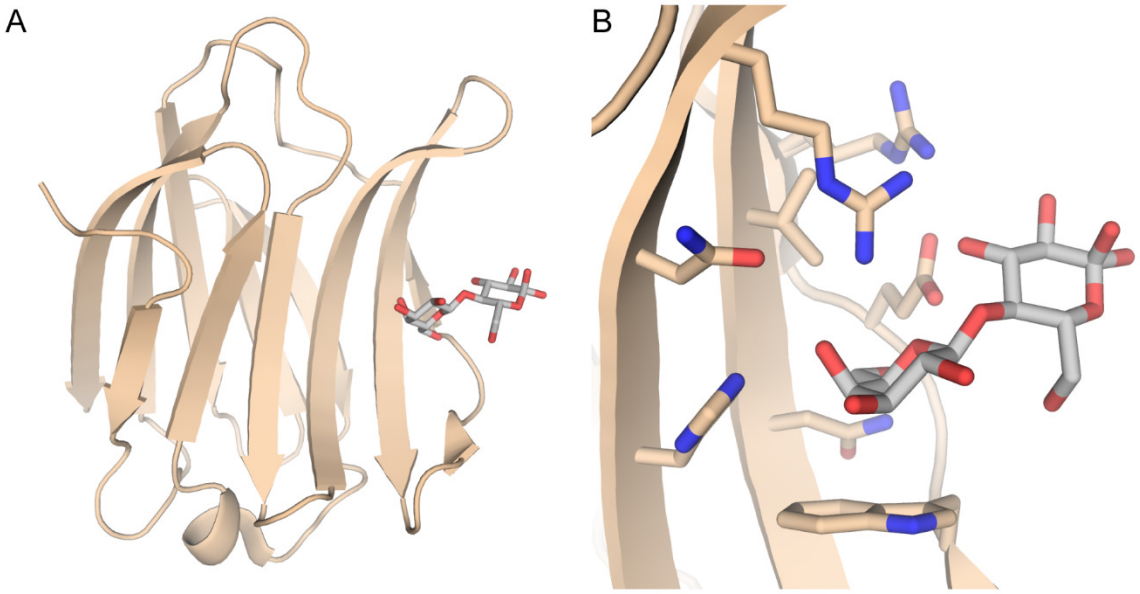



**Figure 1.8**: (A) Structure of galectin-3. (B) Ligand binding site with bound lactose. The eight protein side chains that are conserved among most galectins are shown.

### 1.3.9 H-type lectins

H-type lectins or hyaladherins binds hyaluronan, a polysaccharide in the extracellular matrix composed of repeating [D-glucuronic acid β(1-3) D-GlcNAc β(1-4)] units[71]. H-type lectins are involved in matrix adhesion, cell-cell interaction, and cell

migration[72]. However, structural information of H-type lectins is limited. An NMR structure of TSG-6, which is expressed in inflammatory diseases and may have anti-inflammatory roles[73], reveal a globular lectin domain containing two β sheets and two α helices[74]. Previous mutagenesis studies[75] and NMR titration[74] reveal the hyaluronan binding site in a groove between protein loops.

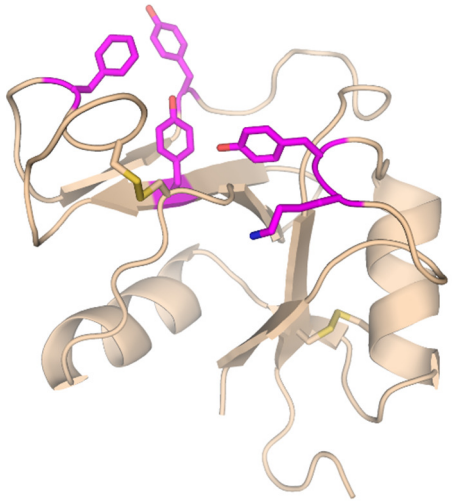


**Figure 1.9**: H-type lectin domain of TSG-6 with hyaluronan binding residues identified by site-directed mutagenesis shown in magenta. Additional hyaluronan binding residues identified by NMR spectroscopy are not shown for clarity.

### 1.3.10 I-type lectins

I-type lectins have an immunoglobulin-like fold[76]. The best characterized I-type lectins are membrane-bound siglecs that recognize sialic acid. Most siglecs are expressed in immune cells[77]. CD33 (siglec-3) and related proteins are likely involved in regulation of leukocyte function[78]. CD22 (siglec-2) is found on B cells, where it act as both an adhesin and a modulator of B cell signaling[79]. Siglec-7, which recognize α(2,8)-linked disialic acid[80], is expressed on natural killer cells where it suppresses cell-mediated cytolytic activity[81]. The structure of siglec-7 shows a β-sandwich

immunoglobulin fold (Figure 1.10 A)[82]. The carboxylate of the sialic acid forms a salt bridge to a conserved arginine residue[82] (Figure 1.10 B). The acetamide NH forms a hydrogen bond with the protein backbone whereas the acetamide carbonyl is recognized by a tyrosine side chain. The exocyclic triol moiety of sialic acid stacks on to the indole ring of a tryptophan residue with additional hydrogen bonding with a lysine side chain.

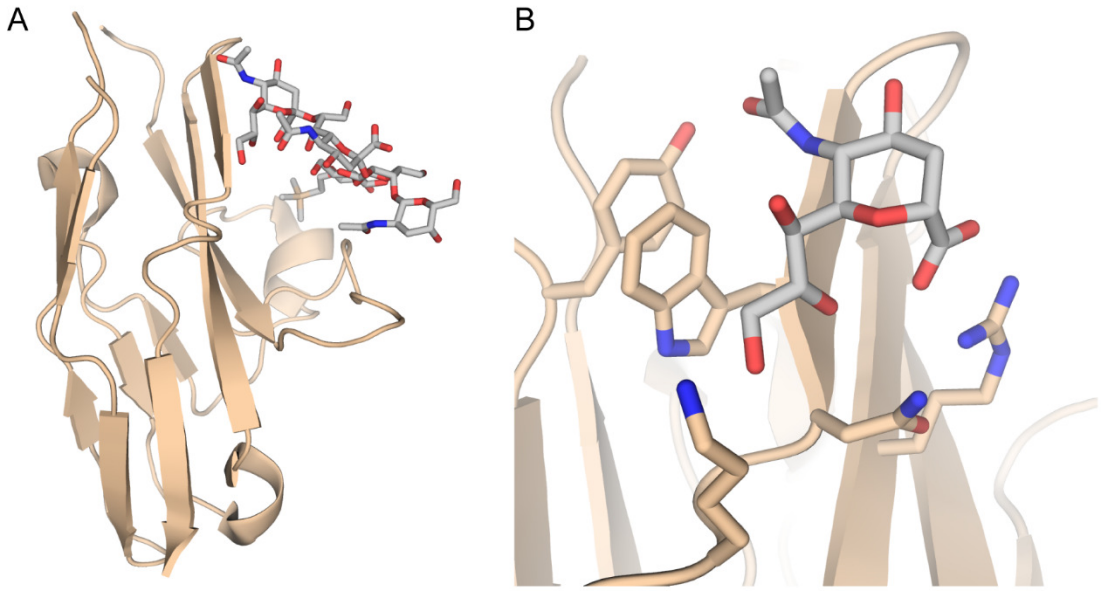



**Figure 1.10**: (A) Structure of siglec-7 with GT1b oligosaccharide bound. (B) Ligand binding site showing only the terminal sialic acid in GT1b.

### 1.3.11 L-type lectins

L-type lectin domain are structurally similar to leguminous plant lectin such as concanavalin A[83]. Mammals have four L-type lectins: ERGIC-53, ERGL, VIP36, and VIPL[84-87]. All of these proteins are membrane-bound, with a single transmembrane helix. The lectin domain are found in the lumen of the ER and the Golgi apparatus. L-type lectins are proposed to participate in glycoprotein secretion and trafficking[88]. The structure of rat ERGIC-53 shows a β-sandwich fold with two bound calcium ions (Figure

1.11). Although a ligand-bound structure is not available, it is known that L-type lectins bind mannose like their plant counterpart. Thus, the ligand-binding mechanism is proposed to be similar.

**Figure 1.11**: Structure of rat ERGIC-53 with bound calcium ions (green) surrounded by protein residues and ordered water molecules (red).

### 1.3.12 M-type lectins

M-type lectins are related to α-mannosidase which belongs to the glycoside hydrolase family 47[89]. These lectins are found in the ER and the Golgi apparatus where they are involved in degradation of incorrectly folded proteins[90]. When a protein fails to fold from interactions with calnexin and calreticulin, the terminal mannose residue is trimmed and the protein is recognized by M-type lectins. The misfolded protein is then exported from the ER for degradation by the proteasome[91-93]. No structure of M-type lectins is available, but they are likely to adopt a barrel-like structure as seen in closely related α-1,2-mannosidase (Figure 1.12 A). Complex with a thiodisaccharide substrate analog reveal a deep ligand binding site in the middle of the helical barrel[94]. Vicinal ring hydroxyls on mannose (O(2) and O(3)) coordinate a protein bound calcium ion. The

binding site is hydrophilic with multiple hydrogen bonds between ring hydroxyls and protein side chains (Figure 1.12 B).

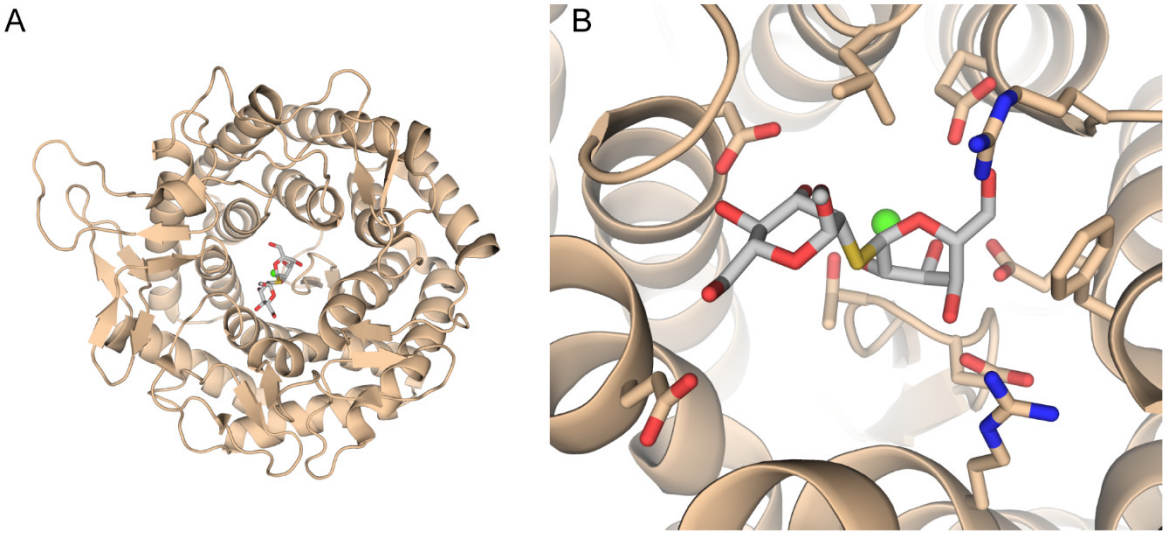



**Figure 1.12**: (A) Helical barrel-like structure of human α-1,2-mannosidase. (B) Ligand binding site containing a substrate analog.

### 1.3.13 P-type lectins

P-type lectins are specific for mannose 6-phosphate[95]. The most studied P-type lectins are mannose 6-phosphate receptors which bind mannose 6-phosphate tagged glycoproteins and direct the proteins to the lysosome[96]. Cation-dependent mannose 6-phosphate receptor uses a manganese (II) ion for ligand binding[97], although some cation-independent mannose 6-phosphate receptors do not require a metal ion for ligand binding[98]. The structure of bovine cation-dependent mannose 6-phosphate receptor reveals a binding site in a cleft between two β sheets (Figure 1.13 A)[97]. All ring hydroxyls are recognized by polar protein side chains. The phosphate group is recognized by the manganese (II) ion and the histidine side chain (Figure 1.13 B).

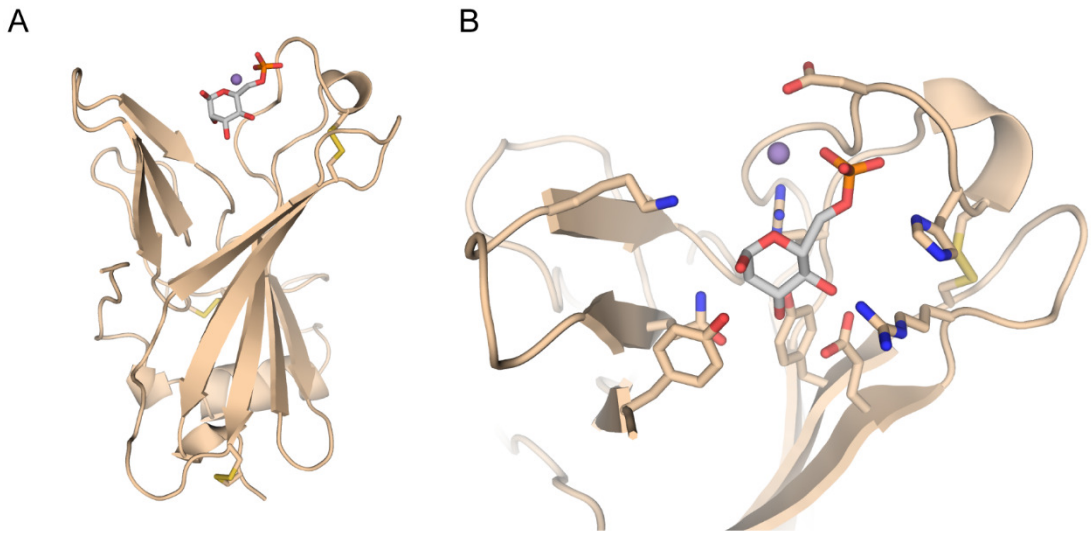



**Figure 1.13**: (A) Overall structure of bovine cation-dependent mannose 6-phosphate receptor. (B) Ligand binding site with bound mannose 6-phosphate and a manganese (II) ion (purple).

### 1.3.14 R-type lectins

The structure of R-type lectin domain resembles, and is named after, the toxic protein ricin[99]. The R-type lectin domain generally recognizes galactose or its derivative and is found in numerous animal lectins. GalNAc transferase enzymes contain an R-type domain that may recognize the glycan and help with processivity of the enzyme[100]. The domain is also found in combination with C-type lectin as a part of the multi-domain mannose receptor family[101]. The R-type domain has a β-trefoil structure[102] (Figure 1.14 A). The structure GalNAc-4-sulfate bound R-type domain reveal stacking of the galactose ring into the π face of a tryptophan residue (Figure 1.14 B)[102]. The ring hydroxyl groups form hydrogen bonds with the polar protein residue side chains. The sulfate group of GalNAc is recognized by asparagine residues and the backbone amides.

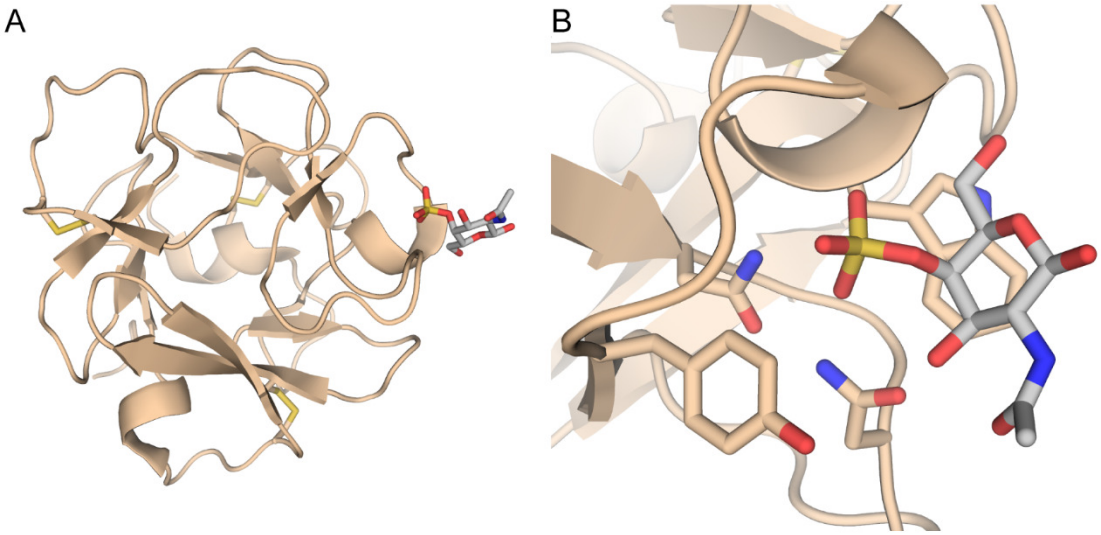



**Figure 1.14**: (A) Overall structure of the cysteine-rich R-type domain of mouse mannose receptor. (B) Ligand binding site with GalNAc-4-sulfate bound.

### 1.3.15 Tachylectins

Tachylectins are found in the hemolymph of the Japanese horseshoe crab (*Tachypleus tridentatus*)[103]. Unlike conventional lectin classification based on sequence and structural homology, each tachylectin is distinct. However, they all recognize foreign epitopes and are suggested to be a part of the innate immune system of the horseshoe crab[103]. At least 5 tachylectins have been identified. Tachylectin-1 binds lipopolysaccharide (LPS) through the 2-keto-3-deoxyoctonate (KDO) residue[104]. Tachylectin-2 binds GlcNAc and GalNAc on LPS and staphylococcal lipoteichoic acids[105]. Tachylectin-3 and -4 (F-type lectin) are specific to S-type LPS, thus they likely interact with carbohydrate residues in the O antigen[106]. Tachylectin-5 recognize acetyl group-containing ligands[107]. Interestingly, tachylectin-5 is similar in sequence and structure to mammalian ficolins[108]. Other than tachylectin-5, tachylectin-2 is the only tachylectin with available structural information. The structure of tachylectin-2 displays a five-bladed β-propeller architecture (Figure 1.15 A)[105]. The ligand binding site is formed

by two large loops with amino acid side chains forming a pocket (Figure 1.15 B). The N-acetyl group stacks onto the π system of a tryptophan residue. Ring hydroxyls of the ligand is recognized by the amide carbonyls or NHs of the backbone.

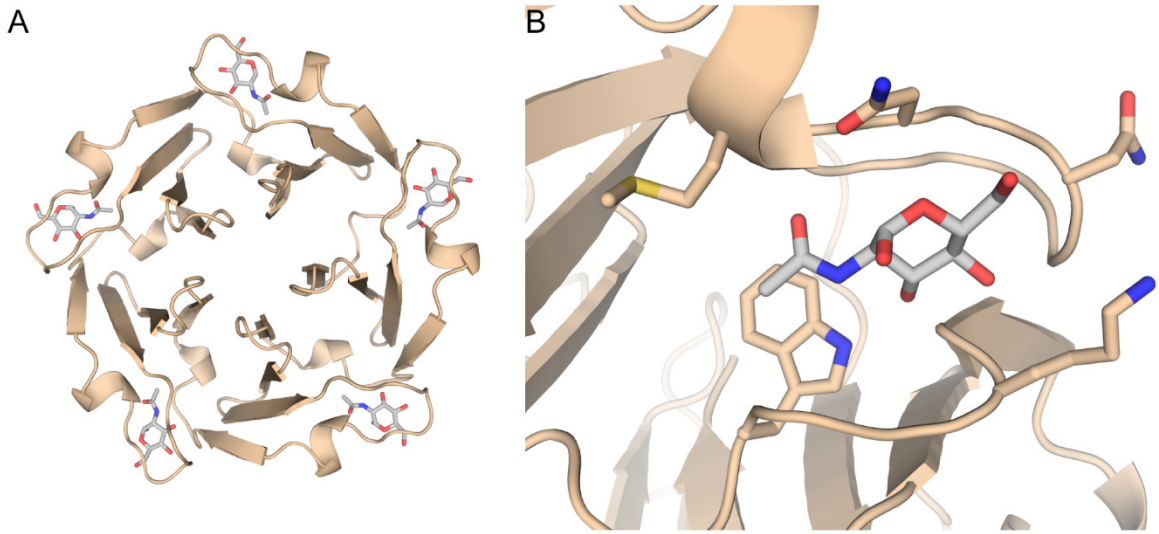



**Figure 1.15**: (A) Five-bladed β-propeller structure of tachylectin-2. (B) Carbohydrate binding site of tachylectin-2 with GlcNAc bound.

### 1.3.16 X-type lectins

X-type lectins, or intelectins, were first isolated from *Xenopus laevis* egg as calcium-dependent galactose-binding lectins[109, 110]. Homologs of intelectins are found in chordates, from tunicates to mammals[111, 112]. Like ficolins, intelectins contain a fibrinogen-like domain (FBD)[113]. Because intelectins are upregulated upon microbial exposure and bind microbes, they are proposed to have immune defense roles[114-118]. Specifically, human intelectin-1 (hIntL-1)[119] does not bind mammalian glycans, but interacts with microbe-specific glycans[120]. Biological consequence of intelectin-microbe interaction is not known. In addition, only the first 50 amino acids of intelectin, corresponding to FBD, align with any known protein. Thus, very little is known about the structure of intelectins. The first part of this thesis is devoted to the *de novo* structure

determination of *Xenopus* and human intelectin as well as the discovery of a unique ligand binding mechanism of intelectins compared to other lectins.

### 1.4 Conclusions

Animals possess lectins with diverse structure and function. Most lectins are secreted proteins, many of which contain disulfide bonds. A variety of mechanisms is employed for carbohydrate-ligand recognition. Lectins recognize their carbohydrate ligands using their protein structure alone without additional non-protein factors. Metal ions, such as calcium ion, is occasionally used for coordination of hydroxyl groups on the carbohydrate ligand. All lectins recognize hydroxyl groups on carbohydrate rings, often with extensive interactions on the entire carbohydrate residue. Therefore, a given lectin is generally specific for limited carbohydrate structures as the ring hydroxyls determine carbohydrate identity and linkage type. Examination of lectin structures not only reveal the molecular basis for ligand recognition but also guide lectin engineering for specific targeting of cargo in biological systems.

# Chapter 2

## Structural and biophysical characterization of *Xenopus* embryonic epidermal lectin

Portions of this chapter have been published in:

Wangkanont, K., Wesener, D. A., Vidani, J. A., Kiessling, L. L., Forest, K. T. Structure of *Xenopus* embryonic epidermal lectin reveals a conserved mechanism of microbial glycan recognition. *In preparation* (2015)

## 2.1 Abstract

Intelectins (X-type lectins) are widely distributed throughout chordates, and they have been implicated in innate immunity. *Xenopus laevis* embryonic epidermal lectin (XEEL) is an intelectin secreted by developing *X. laevis* embryo into environmental water, and is postulated to function as a defense against microbes. XEEL is a homolog (64% identical) of human intelectin-1 (hIntL-1) previously shown to bind microbial glycans and may function as an innate immune defense protein. However, the structure and ligand binding mechanism of this protein family is unknown. This chapter details structure elucidation of XEEL, which was then used as a molecular replacement search model to solve the structure of hIntL-1. Despite lacking the intermolecular disulfide bond found in hIntL-1, the carbohydrate recognition domain of XEEL ($XEEL_{CRD}$) is trimeric in solution. The full length XEEL is a disulfide-linked dimer of trimeric $XEEL_{CRD}$ mediated by an N-terminal peptide. XEEL contains three bound calcium ions in each monomer: two in a structural site and one in the ligand binding site. The structure of D-glycerol 1-phosphate-bound XEEL reveals direct coordination of a vicinal diol to the calcium ion in the ligand binding site. The ligand specificity and quaternary structure of XEEL predicts an agglutination activity, which was confirmed by agglutination of *Streptococcus pneumoniae*. These results provide a foundation for the structural characterization of intelectins.

## 2.2 Introduction

Cells are covered in a carbohydrate coat that is essential for distinguishing self from non-self[121]. Recognition of carbohydrate residues is mediated by carbohydrate-binding proteins or lectins. C-type lectins and ficolins are immune lectins found

throughout the animal kingdom, highlighting the importance of carbohydrate recognition in the immune system[18, 122]. Intelectins, also known as X-type lectins[111], have been proposed to bind microbes and function as an immune lectin[113]. However, little is known about the structure and function of intelectins.

Homologs of intelectin are found across diverse species from tunicates to human[112]. Data from various intelectin homologs support their roles in host defense. For example, tunicate intelectin functions as an opsonin for phagocytosis by hemocytes[123]. Amphioxus intelectin agglutinate bacteria *in vitro*[124, 125]. In organisms where intelectin function has not been examined, there is often a correlation between intelectin expression levels and microbial exposure. For example, rainbow trout and zebrafish intelectins are upregulated upon infection[117, 118, 126]. Sheep and mice increase intelectin expression after infection by intestinal parasitic nematodes[114, 127, 128]. Because several intelectins bind microbes and are upregulated upon microbial exposure, it is likely that intelectins function as a conserved immune defense lectins.

To elucidate the conserved function of intelectins, characterization of intelectins from multiple organisms is required. Since intelectin was first discovered in *Xenopus laevis*[109, 110, 129], at least five intelectins from *X. laevis* has have since been described. *X. laevis* cortical granule lectins (XCGL-1 and XCGL-2) are found in oocytes and developing embryos[130], where they participate in formation of the fertilization envelope to block polyspermy[109, 129]. *X. laevis* serum lectins (XSL-1 and XSL-2) are expressed upon lipopolysaccharide exposure, suggesting their roles in immune response[116, 131]. *X. laevis* embryonic epidermal lectin (XEEL) is secreted during various developmental stages, highest around hatching, and has been proposed to recognize microbes[132, 133].

XEEL is produced by the goblet cells of the epidermis, which has similar property as the human mucosal epithelia[134]. These parallels suggests XEEL may also function in innate immunity.

Humans express two intelectins, hIntL-1 and hIntL-2. hIntL-1 is found at mucosal barriers such as the lung while hIntL-2 expression is limited to the intestine[119]. Although the biological function of hIntLs has not been fully elucidated, several lines of evidence suggest that hIntL-1 is involved in the innate immune defense system. For example, it was suggested that hIntL-1 binds furanoses such as ribose and galactofuranose (Gal*f*) found on microbial glycans[113]. Additionally, we recently found that hIntL-1 does not bind any human glycan in an array of over 600, but binds a variety of microbial glycans[120]. Specifically, hIntL-1 interacts with microbe-specific carbohydrate residues such as β-Gal*f*, D-glycerol 1-phosphate (GroP), and 3-deoxy-D-*manno*-oct-2-ulosonic acid (KDO). In addition, hIntL-1 binds *Streptococcus pneumoniae* cell surfaces, supporting the role of hIntL-1 in microbial recognition and innate immunity.

Little is known about the structure of intelectins and their ligand binding mechanism. XEEL and other intelectins share high sequence similarity, including a conserved fibrinogen-like domain (FBD) also found in ficolins[112]. However, sequence alignment suggests that FBD only correspond to merely 45 of the 300 amino acid residues. Similar to a C-type lectin, intelectins require a calcium ion for ligand binding, but the primary sequences of intelectins lack the C-type lectin motif, suggesting that intelectins are structurally distinct[109, 119, 135]. Moreover, the oligomeric state and relative orientation of intelectin carbohydrate recognition domain (CRD) are not well characterized. Thus, structural investigation of XEEL will not only elucidate how

intelectins bind diverse ligands, but will also enable analysis of structural and functional conservation of intelectins across species. In addition, biophysical characterizations of XEEL in solution could provide clues to domain organization of intelectins.

This chapter describes the biophysical properties and crystal structure of the complete CRD of XEEL ($XEEL_{CRD}$) with and without bound GroP. XEEL binds GroP by direct calcium ion coordination of the vicinal diol in GroP. Despite lacking the intermolecular disulfide bonds present in full length XEEL, $XEEL_{CRD}$ is trimeric in solution. In light of the previously described disulfide-linked, hexameric state of full length XEEL[132], we propose that XEEL is a dimer of trimers. This molecular arrangement suggests that XEEL may possess agglutination activity, a hypothesis supported by *Streptococcus pneumoniae* agglutination. Along with the structure of hIntL-1[120] (Chapter 3), structural characterization of XEEL will guide further research into biological function and ligand specificity of intelectins.

## 2.3 Material and Methods

### 2.3.1 Expression and purification of *Xenopus laevis* embryonic epidermal lectin (XEEL)

The cDNA for XEEL (GenBank accession number BC087616) was purchased (Source BioScience) and amplified with primer A (5'-TTTTTGCACTTGCATTTCCAGCAGGGCACGCTGGTTCATGGAGCCATCCGC-AGTTTGAAAAGGGTTCATGTGAACAAGCTTCAATTTCTG-3') and primer B (5'-GGTACCAAGCTCATTAACGGTAGAAGAGCATCACAGCTGCC-3'). The PCR product was diluted and further amplified with primers B and C (5'-GCGCGGATCCATGTTGTCATATAGCCTGTTGCTTTTTGCACTTGCATTTCCAGCAG

GG-3'). The resulting product was digested with BamHI and KpnI then ligated into a digested pFastBac1 vector (Life Technologies). The N-terminus of the protein product is expected to be (residue 4) MLSYSLLLFALAFPAGHA (residue 22) GS**WSHPQFEK** (residue 22) GSCEQASIS; where the expressed protein begins with residue 4, the second methionine in the annotated sequence. The underlined amino acids are the predicted secretion signal peptide and the bold amino acids denote the Strep-tag® II used for purification. The vector was then transformed into DH10Bac to generate a recombinant baculovirus according to the manufacturer protocol (Bac-to-Bac Baculovirus Expression System, Life Technologies). The resulting baculovirus genomic DNA was transfected (Insect GeneJuice, Novagen) into Sf21 insect cells to produce the first generation (P1) of recombinant baculovirus. The amplified P2 baculovirus was produced using P1 virus to infect suspension culture of Sf21 cells grown in SF900-II-SFM (Life Technologies). The amplified virus supernatant was harvested by centrifugation and fetal bovine serum was added to 2% to stabilize the baculovirus.

XEEL was produced as a secreted protein using High Five cells (Life Technologies), a derivative of *Trichopulsia ni*. For expression of selenomethionine (Se-Met) labeled XEEL used for experimental phasing, High Five cells were suspension cultured in 921 Delta Series, methionine deficient medium (Expression Systems, cat. no. 96-200, 200 mL) supplemented with 1x antibiotic-antimycotic (Life Technologies) and 10 µg/mL gentamicin (Life Technologies). Cells were infected when they reached a density of at least $2X10^6$ cells/mL by addition of 0.5 µL of P2 baculovirus per $1X10^6$ viable cells. After the first addition of L-SeMet (10 mg) at 12 hours post infection, additional portions (10 mg) were added every 24 hours until medium harvest. Cells

producing XEEL were cultured for 5 days at 22 ºC in a baffled flask with shaking at 90 rpm. Conditioned culture medium was harvested by centrifugation and filtered through a 0.22 μM filter then stored at 4 °C for at least one week. This incubation period is crucial for successful crystallization of XEEL. Unlabeled XEEL was produced using the same conditions except using Express Five SFM media (Life Technologies) supplemented with 1x antibiotic-antimycotic, 10 μg/mL gentamicin, and 4 mM glutamine.

Conditioned media was dialyzed extensively against 20 mM Bis-Tris (pH 6.7), 150 mM NaCl, and 1 mM EDTA. The media was slowly adjusted to pH=6.7, and $CaCl_2$ was added to 10 mM from a 1 M stock. In addition, avidin (Calbiochem) was added to the conditioned media (28 μg/mL) to absorb excess biotin, per the manufacturers protocol (IBA GmbH). The solution was then cleared by centrifugation. Strep-tag® II XEEL was purified using Strep-Tactin® Superflow resin (IBA GmbH, cat. no. 2-1206-002). The column was washed with 20 mM Bis-Tris (pH 6.7), 150 mM NaCl, 10 mM $CaCl_2$ and then 20 mM Bis-Tris (pH 6.7), 150 mM NaCl, 0.5 mM EDTA. Protein was eluted with 5 mM d-desthiobiotin (Sigma Aldrich) in 20 mM Bis-Tris (pH 6.7), 150 mM NaCl, 0.5 mM EDTA then concentrated using a 10,000 MWCO Amicon Ultra Centrifugal Filter. During concentration, large sheet-like crystals began to form. Crystals were harvested by centrifugation at 2,000 RPM and washed 2 times with 20 mM Bis-Tris (pH 6.7), 150 mM NaCl, 0.5 mM EDTA. The crystals were resuspended in 20 mM Bis-Tris (pH 6.7), 150 mM NaCl, 0.5 mM EDTA prior to addition of $CaCl_2$ to 5 mM. Within one minute, the crystals completely redissolved. Purity of the re-crystallized protein assessed by SDS-PAGE and Coomassie blue staining and was >95 %. The concentration of XEEL was determined using absorbance at 280 nm with an

estimated[136] ε=75,455 $cm^{-1}M^{-1}$ for the monomer and a calculated molecular mass of 36,258 Da, post signal peptide removal. Typical yields were 0.5 mg per 50 mL of conditioned media.

### 2.3.2 Expression and purification of thioredoxin-fused XEEL (residues 22-47)

The expression construct was synthesized as a gBlock (Integrated DNA Technologies) then cloned into the MfeI and BamHI sites of pMAL-c5x (New England Biolabs). The sequence coding for maltose-binding protein was excised in the process and replaced with the expression construct. The expected protein sequence corresponds to full-length thioredoxin (underlined) followed by a linker, a histidine tag, and an enterokinase cleavage site (bold) fused to residues 22-47 of XEEL (italics):

<u>MSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLN IDQNPGTAPKYGIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDANLA</u>GSGSGHMHH HHHHSSGT**DDDDK***GSCEQASISEKKEKILNLLACWTEGN*

The plasmid was transformed into SHuffle Express *Escherichia coli* (New England Biolabs). Bacteria were grown in Terrific Broth at 30 ºC until $OD_{600}$ reached 0.6 when the temperature was lowered to 15 ºC. Induction of protein expression was achieved by addition of IPTG to 0.1 mM and fermentation was allowed to proceed for 16 hours. The cells were lysed and the protein purified using Ni-NTA using a previously described protocol[137].

### 2.3.3 X-ray crystallography

SeMet-labeled $XEEL_{CRD}$ was concentrated to 2 mg/mL and crystallized (hanging-drop vapor-diffusion) by mixing 1 µL of the protein solution and 1 µL of well solution

(100 mM Tris, pH 7.0, 20-24% PEG 400). Crystals appeared in 2-3 hours and grew to full size within 2-3 days. For cryoprotection, crystals were briefly dipped in well solution containing 30% PEG 400 and vitrified in liquid nitrogen.

Unlabeled $XEEL_{CRD}$ protein was used to obtain the crystal structure with bound GroP (Sigma Aldrich catalog number G7886). Crystals were grown under similar conditions to Se-Met-labeled XEEL, with 50 mM GroP present during crystal growth and cryoprotection.

Single crystal X-ray diffraction experiments were performed at the Life Sciences Collaborative Access Team beamline 21-ID-D at the Advanced Photon Source, Argonne, IL. For experimental phasing and structure determination of Se-Met $XEEL_{CRD}$, diffraction data were collected at the Se K-edge (12661.38 eV). Integration, scaling, and merging were performed with HKL2000 [138]. The structure was solved using PHENIX by Se-SAD phasing with AutoSol[139]. The substructure search yielded 39 out of the 42 expected sites. Figures of merit were 0.28 and 0.66 before and after density modification, respectively. The GroP-bound $XEEL_{CRD}$ structure was solved by molecular replacement using Phaser using monomeric Se-Met $XEEL_{CRD}$ as a search model [140]. Both structures were adjusted and refined with Coot and phenix.refine respectively [141, 142]. The structural restraint for GroP was generated using PRODRG [143]. MolProbity was used for validation [144]. Data collection and refinement statistics are presented in Table 2.1. Structure figures were generated using PyMOL[145].

**Table 2.1**: Data collection and refinement statistics for Se-Met $XEEL_{CRD}$ and $XEEL_{CRD}$-GroP complex.

| Data collection statistics | Se-Met $XEEL_{CRD}$ | $XEEL_{CRD}$-GroP complex |
|---|---|---|
| Wavelength (Å) | 0.97924 | 0.97856 |
| Resolution range (Å)* | 30.10 - 2.30 (2.38 - 2.30) | 30.2 - 2.20 (2.28 - 2.20) |
| Space group | P $2_1$ | P 6 |
| Unit cell (Å) | 123.6 111.1 123.6 90 119.7 90 | 124.6 124.6 55.6 90 90 120 |
| Total reflections | 991049 | 285681 |
| Unique reflections | 129043 (12885) | 25366 (2526) |
| Multiplicity | 7.7 (7.7) | 11.3 (11.1) |
| Completeness (%) | 100 (100) | 100 (100) |
| Mean I/σ(I) | 13.7 (2.6) | 19.4 (3.2) |
| Wilson B-factor ($Å^2$) | 21.0 | 29.6 |
| R-merge | 0.168 (0.625) | 0.127 (0.755) |
| R-meas | 0.180 (0.670) | 0.133 (0.792) |
| R-pim | 0.065 (0.241) | 0.040 (0.237) |
| Refinement Statistics | | |
| Resolution range (Å) | 30.10 - 2.30 (2.38 - 2.30) | 29.93 - 2.20 (2.25 - 2.20) |
| R-factor | 0.1597 (0.2152) | 0.1515 (0.1977) |
| R-free (5%) | 0.1881 (0.2673) | 0.1692 (0.2363) |
| Number of atoms | | |
| Protein | 12927 | 2151 |
| Calcium | 18 | 3 |
| GroP | - | 10 |
| Water | 1470 | 166 |
| Protein residues | 1662 | 277 |
| RMSD (bonds, Å) | 0.011 | 0.010 |
| RMSD (angles, °) | 1.170 | 1.089 |
| Est. coordinate error (ML, Å) | 0.21 | 0.17 |
| Ramachandran favored (%) | 96.1 | 95.7 |
| Ramachandran outliers (%) | 0 | 0 |
| Average B-factor ($Å^2$) | | |
| Protein | 13.01 | 24.18 |
| Calcium | 7.83 | 17.28 |
| GroP | - | 38.03 |
| Solvent | 24.36 | 33.71 |

*Statistics for the highest-resolution shell are shown in parentheses. Coordinates and structure factors of Se-Met $XEEL_{CRD}$ and $XEEL_{CRD}$-GroP complex were deposited at the Protein Data Bank under accession code 4WMO and 4WN0, respectively.

### 2.3.4 Surface plasmon resonance

Analysis of $XEEL_{CRD}$ using surface plasmon resonance (SPR) was conducted on a ProteOn XPR36 (Bio-Rad) by Darryl A. Wesener at the University of Wisconsin–Madison Department of Biochemistry Biophysics Instrumentation Facility. To

measure $XEEL_{CRD}$ binding, ProteOn NLC sensor chips (NeutrAvidin coated sensor chip) (Bio-Rad, cat. no. 176-5021) were used to capture biotinylated carbohydrate ligands. All experiments presented here were conducted at surface saturated levels of ligand, ~200 response units (R.U.). In all experiments, captured biotin was used in flow cell one as a control. Samples containing purified $XEEL_{CRD}$ were prepared by serial dilution into SPR running buffer (20 mM HEPES (pH 7.4), 150 mM NaCl, 1 mM $CaCl_2$, and 0.005 % Tween-20). Surfaces were regenerated with short injections of solutions of 10 mM hydrochloric acid (HCl). Data were referenced using the biotin reference channel, and processed using the Bio-Rad ProteOn software package. Glycerol competition experiments were performed by adding glycerol to the SPR running buffer that purified $XEEL_{CRD}$ was diluted in. Data were analyzed on Prism6 (GraphPad). For determination of the glycerol $IC_{50}$, SPR data were fit to a one site competition model in Prism6. The top of the fit was constrained to 110 R.U. based on the equilibrium response in the presence of 1 mM $CaCl_2$.

### 2.3.5 Chemical crosslinking

$XEEL_{CRD}$ aliquots (1 mg/mL, 2 µL) were mixed with 1 µL of bis(sulfosuccinimidyl)suberate crosslinker (Pierce) stock solutions to achieve final crosslinker concentration of 0 – 5 mM. The buffer for all components was 20 mM Bis-Tris (pH 6.7), 150 mM NaCl, 0.5 mM EDTA, and 5 mM $CaCl_2$. Crosslinking was performed for 30 minutes at room temperature. Each reaction was diluted by adding 7 µL of the reaction buffer and denatured by adding 2 µL of 6X SDS loading buffer (350 mM Tris pH 6.8, 30% glycerol, 10% SDS, 9.3% DTT, and 0.06% bromophenol blue)

then heated at 95 ºC for 3 minutes. The crosslinking products were analyzed by SDS-PAGE stained with Coomassie blue.

### 2.3.6 Small-angle X-ray scattering

Scattering data of a 2 mg/mL (63 μM monomer) solution of $XEEL_{CRD}$ were collected on a Bruker NANOSTAR SAXS system. Background scattering (20 mM Bis-Tris (pH 6.7), 150 mM NaCl, 0.5 mM EDTA, and 5 mM $CaCl_2$) was subtracted from the data. Preliminary analysis was performed using the ATSAS suite (PRIMUS and GNOM)[146-148]. Predicted scattering curves from the crystal structure of $XEEL_{CRD}$ were generated using CRYSOL[149]. DAMMIF was used to generate 15 dummy atom models in the fast mode, which were then averaged by DAMAVER [150, 151]. The resulting model was refined with DAMMIF in a slow mode. The final dummy atom model was aligned with the trimeric $XEEL_{CRD}$ crystal structure by SUPCOMB[152].

### 2.3.7 Sedimentation equilibrium analytical ultracentrifugation

Recrystallized $XEEL_{CRD}$ samples were prepared by dilution to concentrations of 9.0 μM (trimer), 5.2 μM, and 2.6 μM for analysis by sedimentation equilibrium analytical ultracentrifugation. Equilibrium data were collected at 20 °C in a Beckman Optima XLA Analytical Ultracentrifuge using 1.2-cm double sector charcoal-filled Epon centerpieces. Protein gradients were recorded at 276 nm every 2-3 h until two or more were superimposable. Equilibrium data were collected at speeds of 6,000, 8,000, 9,600, 11,500, 13,200, and 15,900 rotations per minute (rpm). After the 15,900 rpm run, the contribution from non-sedimenting absorbance was determined by high speed depletion of the protein; this absorbance was ≤0.006 for the three samples. After depletion, subsequent re-equilibration at 11,500 rpm resulted in a gradient essentially

superimposable on the original gradient at this speed, indicating no significant loss of protein to irreversible aggregation during the course of the experiment.

The solvent density ($\rho$) of 1.004 g/mL was computed from density increments [153], except that the BIS-TRIS buffer was not included as it is not in the available table. The calculated partial specific volumes ($\upsilon$) of intact XEEL and $XEEL_{CRD}$ are the same and based on the amino acid sequences, are 0.719 mL/g. The monomeric masses based on sequence ($M_s$) are 36,258 and 31,652, respectively.

Analysis of the equilibrium data followed an approach similar to that previously described[154] and was performed by Darrell R. McCaslin (University of Wisconsin–Madison, Department of Biochemistry, Biophysics Instrumentation Facility). The analysis utilized programs developed in Igor Pro (Wavemetrics Inc., Lake Oswega, OR) by Darrell R. McCaslin. The measured non-sedimenting absorbance for each sample was included as a fixed parameter. The data from the three concentrations and five speeds were globally fit to models consisting of one or two macromolecular species. However, the single species model describes the complete data set adequately, and thus a single species model was used. The reduced molecular weight ($M_r$) was used as the fitting parameter as this removes the impact of ambiguities in $\upsilon$ and $\rho$ on the fit. The reduced molecular weight is defined as $M_r = M_w(1-\upsilon\rho)$, where $M_w$ is, in the single species case, the molecular weight of the macromolecular complex; and furthermore, it must be an integral multiple of the sequence weight $M_s$ thus $M_r = nM_s(1-\upsilon\rho)$, where $n$ is equal to the oligomeric state of the protein.

### 2.3.8 *Streptococcus pneumoniae* agglutination assay

Formaldehyde-fixed *S. pnemoniae* serotypes were prepared as previously described[120]. Hexameric XEEL was expressed in High Five cells in a similar manner as the monomeric $XEEL_{CRD}$ except that fetal bovine serum was added to 1% during expression and SIGMAFAST protease inhibitor (Sigma Aldrich, S8820) was added after media collection to suppress proteolysis of XEEL. Each bacterial suspension was mixed with hexameric XEEL and imaged using a 40X objective lens on a Nikon TE2000 inverted microscope equipped with an Andor IXON camera. Images were acquired with *MetaMorph. Scale bar was added using ImageJ with the scale of 0.162 µm/pixel.*

## 2.4 Results

### 2.4.1 Mapping of intermolecular disulfide bonds in XEEL

Insect cell expression system was used to produce XEEL both because of its ability to yield large quantities of protein needed for biophysical characterization and because it can be used to incorporate selenomethionine (Se-Met) into protein for crystallographic structure determination. The expression construct was engineered to contain a Strep-tag II purification tag inserted between the predicted signal peptide[155] and the mature XEEL. Cleavage of the signal peptide should yield XEEL with a two amino acid linker followed by an eight amino acid Strep-tag II at the N-terminus. After purification using Strep-tactin affinity chromatography[156], the protein was further characterized by mass spectrometry. Although the expected mass of the expressed construct is 36230 Da, MALDI and ESI-TOF analysis returned masses of 32884 Da and 32802 Da, respectively. Without glycosylation taken into account, the observed mass difference indicated a truncation by secondary proteolysis of at least 30 residues. N-

terminal sequencing (ABI 494, Tufts University Core Facility) revealed the first five amino acids to be RSGGS. Therefore, the XEEL construct used in this study corresponds to residue 54 to 342. We refer to this as $XEEL_{CRD}$ to denote the carbohydrate recognition domain. Because $XEEL_{CRD}$ lacks the Strep-tag II purification tag, we speculate that $XEEL_{CRD}$ has significant affinity to agarose resin, and was eluted by EDTA present in wash and elution buffers. The production of this truncated protein provided the opportunity to examine the role of the N-terminal sequence in oligomerization.

The N-terminal sequence of intelectins are not conserved, and previous studies suggest that this region controls intelectin oligomeric state[112]. $XEEL_{CRD}$ lacks the intermolecular disulfide bonds present in hIntL-1. Consistently, $XEEL_{CRD}$ is monomeric in molecular weight on reducing and non-reducing SDS-PAGE (Figure 2.1 A). However, expression of the equivalent XEEL construct in HEK293T cells yielded a full length protein that migrates as a monomer under reducing conditions, but as a disulfide-linked hexamer under non-reducing conditions, as previously observed[132]. These results suggest that the N-terminal peptide, truncated during expression and conditioning of the spent insect cell media, is responsible for the disulfide-linked oligomerization of the full-length XEEL. Cysteines C24 and C42 are located within this peptide and are likely to form the intermolecular disulfide bonds. To test this hypothesis, we created a recombinant construct in which residues 22-47 of XEEL were fused to the C-terminus of thioredoxin. When expressed and purified under non-reducing conditions, the recombinant protein forms disulfide-linked dimers and hexamers (Figure 2.1 B). Thus, this 26 amino acid peptide is an oligomerization domain of $XEEL_{CRD}$.

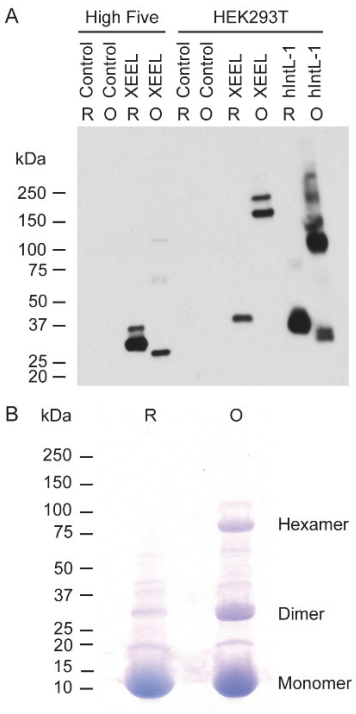



**Figure 2.1**: (A) Western blot of XEEL$_{CRD}$ expressed by High Five and HEK293T cells in both reducing (R) and non-reducing (oxidized, O). hIntL-1 is included for comparison. The hIntL-1 antibody used (Proteintech, 11770-1-AP) cross-reacts with XEEL. (B) SDS-PAGE of thioredoxin-fused XEEL (residues 22-47) under reducing and oxidizing conditions.

### 2.4.2 Crystal structure of XEEL

To analyze the structure and ligand binding mechanism of intelectins, we pursued crystal structure determination of XEEL$_{CRD}$. Crystals of native XEEL$_{CRD}$, that diffracted to 2 - 2.3 Å, were readily obtained. However, molecular replacement attempts using ficolins, FBD-containing proteins, as search models did not yield a solution. Because no structure of other intelectin homologs were available, the phase problem had to be solved experimentally. Se-Met labeled XEEL$_{CRD}$ was produced and Se-Met substitution was confirmed by mass spectrometry. Both native and Se-Met labeled XEEL$_{CRD}$ crystals grew under similar conditions. The structure of monoclinic Se-Met labeled XEEL$_{CRD}$ crystal was successfully solved by Se-SAD (with 67.5% solvent

content) and refined to 2.3 Å resolution. The asymmetric unit consists of two trimers related to one another by a pseudotranslation vector (55.6 Å) approximately parallel to the b axis (Figure 2.2 A). The first resolvable N-terminal residue is G66. The rest of the protein chain was traceable without a gap. The C terminus is clearly resolved in the electron density map.

The XEEL$_{CRD}$ is a globular structure that does not resemble other known protein families. The fibrinogen-like lobe contains a twisted 7-stranded β-sheet. The inserted carbohydrate-binding, intelectin-specific region is predominantly a random coil with a 3-stranded curved β-sheet, and three distinct calcium ion binding sites. XEEL$_{CRD}$ is also encircled by 12 short solvent-exposed α helices.

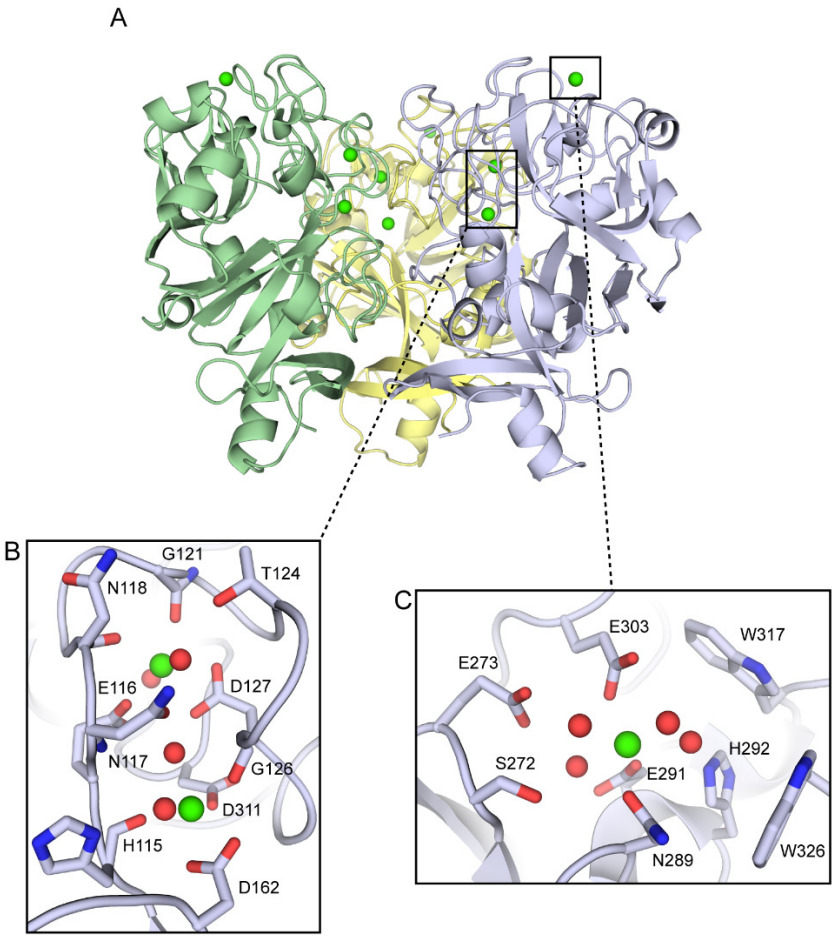


**Figure 2.2**: (A) Crystal structure of Se-Met-labeled trimeric $XEEL_{CRD}$. The second trimer in the asymmetric unit is removed for clarity. (B) Structural calcium site with two calcium ions (green) and four ordered water molecules (red). (C) Ligand binding site with one calcium ion (green) and four ordered water molecules (red).

Three well-ordered calcium ions are organized into two groups. Two calcium ions are buried deep in the protein core and are presumed to play a structural role (Figure 2.2 B). The structural calcium ion closer to the protein surface is coordinated by two water molecules and five protein-derived interactions: the side chains of E116 and D127, and the backbone carbonyls of N118 and G121. N117 forms additional hydrogen bonds to calcium-bound water molecules through both its side chain and backbone amide NH. The structural calcium ion located farther inside the trimeric protein is coordinated directly by the side chains of D162 and D311, backbone amide carbonyls of H115 and G126, and two water molecules. The side chain of H115 also forms a hydrogen bond with a calcium-bound water molecule. No single amino acid nor water molecule coordinates both calcium ions directly

The remaining calcium ion is surface exposed and is later identified as a ligand binding site (Figure 2.2 C). This calcium ion is directly coordinated by N289, E291, E303, and four ordered water molecules. The pocket geometry is characterized by a dense network of hydrogen bonds among the calcium ion, the aforementioned residues, the four water molecules, S272, E273, and H292. The side chains of W306 and W326 form the back wall of the ligand binding site.

### 2.4.3 $XEEL_{CRD}$ binds ligands containing a 1,2 terminal diol

The ability of hIntL-1 to bind β-Gal*f* has been reported[113, 120]. However, it is not known whether this ligand specificity is conserved between XEEL and hIntL-1. Therefore we examined the $XEEL_{CRD}$ ligand binding properties using SPR (Figure 2.3

A). XEEL$_{CRD}$ exhibit concentration-dependent binding to immobilized β-Gal*f* with rapid on and off rates, indicating minimal structural reorganization upon ligand binding. To quantify the binding affinity of XEEL$_{CRD}$ towards immobilized β-Gal*f*, we performed equilibrium binding analysis of the SPR data (Figure 2.3 B). The apparent affinity of the trimeric XEEL$_{CRD}$ for β-Gal*f* is 4.1±0.5 μM. Minimal binding was observed to β-galactopyranose (β-Gal*p*) or β-ribofuranose (β-Rib*f*) suggesting that, like hIntL-1, XEEL does not bind furanosides, but instead recognize a specific epitope on β-Gal*f*. Because hIntL-1 also binds GroP, it is likely that the vicinal diol on β-Gal*f* is the epitope recognized by XEEL$_{CRD}$.

To test whether the exocyclic vicinal diol on β-Gal*f* is the epitope recognized by XEEL$_{CRD}$, we demonstrated that soluble glycerol is an inhibitor of β-Gal*f* binding to XEEL$_{CRD}$ in a concentration-dependent manner (Figure 2.3 C). These experiments suggest that XEEL binds β-Gal*f* and GroP through their common exocyclic vicinal diol epitope. From the inhibition data, the IC$_{50}$ value of glycerol was determined to be 0.6±0.6 mM. The glycerol inhibition data fit well with a one-site competition model suggesting that XEEL$_{CRD}$ binds as a single species, with no cooperative assembly of the monomers. Therefore, the IC50 value can be used as an estimate of the $K_d$ of a exocyclic vicinal diol-containing ligand for a single XEEL binding site.

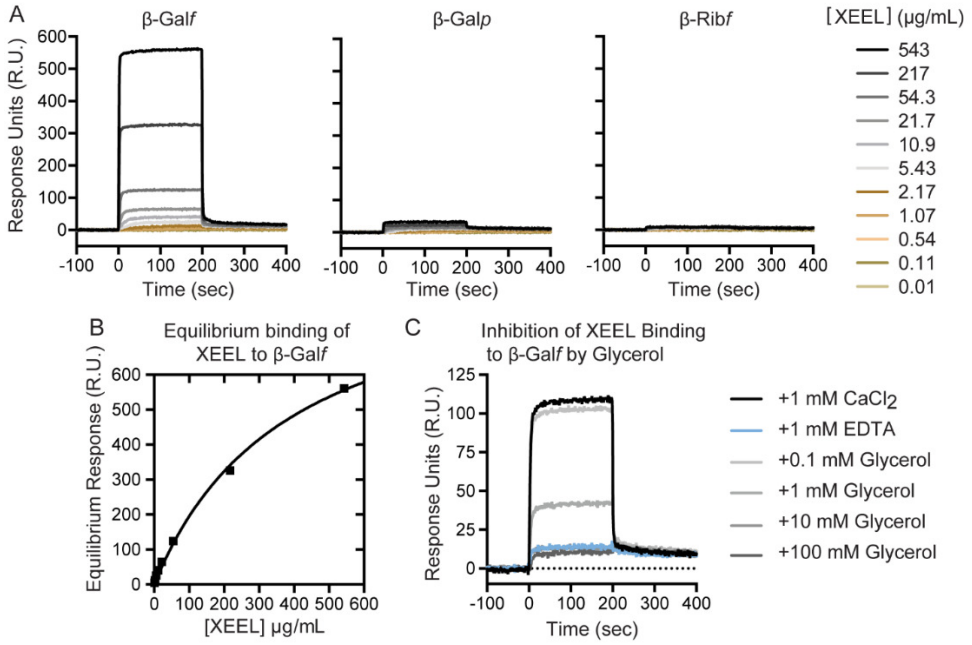



**Figure 2.3**: (A) SPR sensograms of XEEL$_{CRD}$ binding to immobilized carbohydrates. (B) Equilibrium binding analysis of XEEL$_{CRD}$ to immobilized β-Gal*f*. (C) Inhibition of XEEL$_{CRD}$ binding to β-Gal*f* by glycerol.

### 2.4.4 Glycerol phosphate recognition by XEEL$_{CRD}$

The crystal structure of Gro-P bound XEEL$_{CRD}$ was obtained by co-crystallization. The structure reveals direct calcium ion coordination by GroP hydroxyl groups, displacing two ordered water molecules in the ligand-binding site (Figure 2.4). This ligand binding mechanism is identical to what has been observed in the hIntL-1-β-Gal*f* complex structure[120]. The diol binds in a pocket formed by the side chain of W317 and W326. The indole NH of W326 forms a hydrogen bond with the phosphate group of GroP. Electron density around the phosphate group elongates toward the indole NH of W317, suggesting dynamic equilibrium of hydrogen bonding between the phosphate group and the side chain of W317 and W326. The primary hydroxyl group of GroP also forms a hydrogen bond with the imidazole side chain of H292. In addition to calcium ion coordination, the side chain of E303 forms a hydrogen bond with the secondary

hydroxyl group of GroP. Alignment of Se-Met $XEEL_{CRD}$ and Gro-P bound $XEEL_{CRD}$ reveals an RMSD value of 0.17 Å over 276 Cα atoms, suggesting no significant structural changes upon either Se-Met labeling or ligand binding. No significant movement of residues in the ligand binding site is observed upon GroP binding. This observation is consistent with rapid binding kinetics observed from SPR.

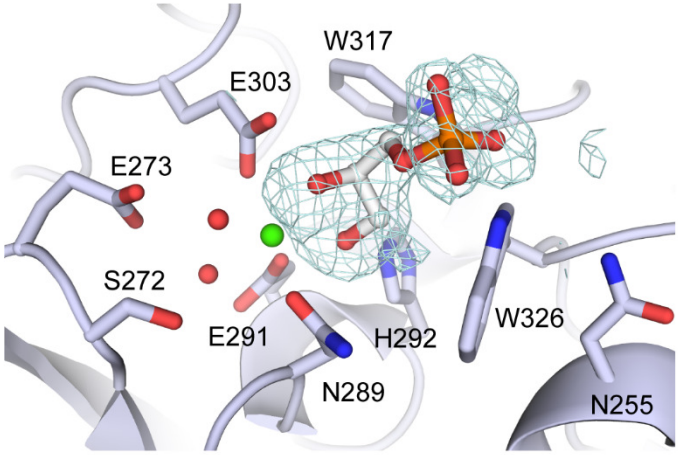



**Figure 2.4**: Crystal structure of $XEEL_{CRD}$-GroP complex. Calcium ion and ordered water molecules are shown in green and red respectively. The cyan mesh represents the $mF_o$-$DF_c$ map contoured at 3σ after refinement without GroP.

### 2.4.5 Intelectins and ficolins are structurally divergent

Intelectins have been proposed to be part of the ficolins structural family due to the presence of FBD. The structure of $XEEL_{CRD}$ shows significant divergence between intelectin and ficolins despite the shared FBD. Topology diagrams of $XEEL_{CRD}$ and L-ficolin[54] (PDB ID 2J3U) highlight these differences (Figure 2.5) and explain why our attempts at molecular replacement using ficolin structures as search models did not yield interpretable maps. Compared to $XEEL_{CRD}$, L-ficolin has three deletions: one between strands 5 and 6 of XEEL, one between strands 6 and 7 of XEEL, and the XEEL ligand-binding site (XEEL strands 9 and 10 as well as a large loop). In L-ficolin, a single structural

calcium site lies between helices E and F, which differ in location and structure to helices K and L in XEEL$_{CRD}$. The structural calcium site residues in XEEL$_{CRD}$ are clustered between strands 3 and 4 with the exception of D311, located in the ligand binding site. Strand 4' in L-ficolin does not have an equivalence in the XEEL$_{CRD}$ structure because D162 in XEEL$_{CRD}$ is a part of the structural calcium site and this interactions disrupt the secondary structure found in L-ficolin. The ligand binding residues of XEEL$_{CRD}$ are located between strands 9 and 11, whereas in L-ficolin the residues in the ligand binding site are dispersed. L-ficolin does not contain calcium ion in the ligand binding site. Therefore, while XEEL$_{CRD}$ contains a FBD, the ligand binding and structural calcium sites are distinct from other lectins. Thus, intelectins represent a novel class of lectins with unique structure and function.

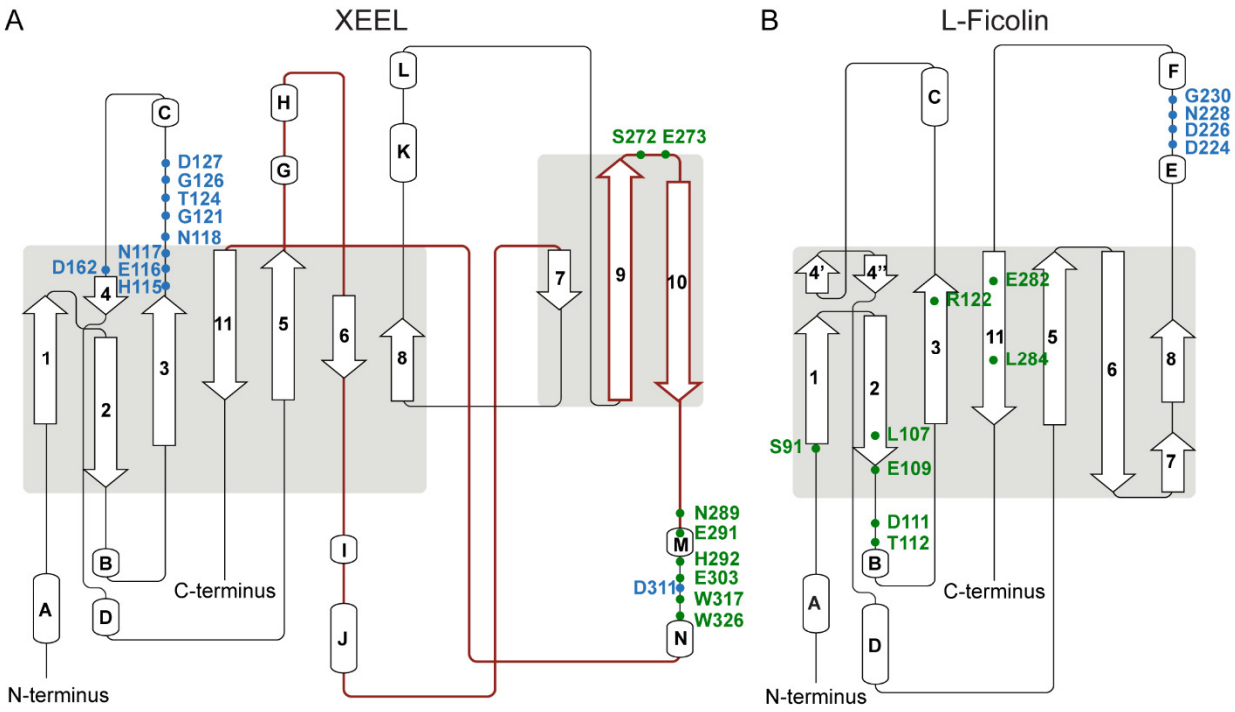



**Figure 2.5**: Topology diagram of (A) XEEL$_{CRD}$ and (B) L-ficolin (PDB ID 2J3U). The residues in the structural calcium sites are labeled in blue and the ligand binding sites

are in green. Regions found in $XEEL_{CRD}$ but not in L-ficolin correspond roughly to brown traces in $XEEL_{CRD}$. The diagrams were generated by Pro-origami[157].

### 2.4.6 $XEEL_{CRD}$ is trimeric in solution

Despite lacking the N-terminal region that engages in disulfide bond formation, the structure of $XEEL_{CRD}$ suggests it forms a non-covalent trimer. Although two head-to-tail trimers are observed in the asymmetric unit of the $XEEL_{CRD}$ structure, there are only a few contacts between the trimers. This weak crystal packing is consistent with the macroscopic observation that $XEEL_{CRD}$ crystals tends to separate into thin sheets when mishandled. Thus, this particular hexameric arrangement is unlikely to be biologically relevant. We next examined whether trimeric $XEEL_{CRD}$ could be observed in solution.

To determine whether the trimeric $XEEL_{CRD}$ persists in solution, chemical crosslinking was performed using bis(sulfosuccinimidyl)suberate (Figure 2.6). No species larger than trimeric $XEEL_{CRD}$ were observed by SDS-PAGE. Because the crosslinking agent only targets amino groups, this chemical limitation may have prevented crosslinking of higher order species. Therefore, small angle x-ray scattering (SAXS) was performed on $XEEL_{CRD}$ solution to obtain additional structural information. A Gunier plot of the data was linear, indicating monodispersity with no interactions between scattering particles (Figure 2.7 A). The radius of gyration (Rg) was 27.9±0.2 Å which is consistent with the calculated Rg of 27.2 Å from the trimeric $XEEL_{CRD}$ crystal structure. The calculated Rg for monomeric and hexameric $XEEL_{CRD}$ (18.7 and 39.1, respectively) do not agree with the experimental data. Comparison of the simulated scattering curve for trimeric or head-to-tail hexameric $XEEL_{CRD}$ with the experimental data revealed a superior fit for trimeric $XEEL_{CRD}$ (Figure 2.7 B). Therefore, SAXS indicates that $XEEL_{CRD}$ is trimeric in solution.

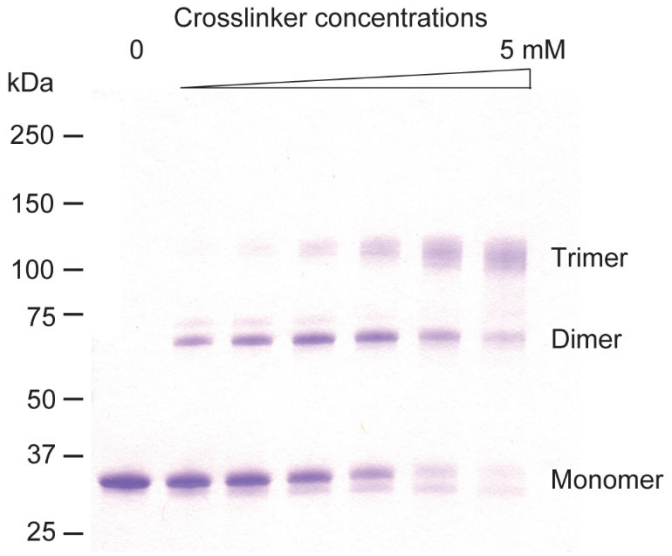



**Figure 2.6**: Crosslinking of XEEL$_{CRD}$ with bis(sulfosuccinimidyl)suberate. The final crosslinker concentrations are 0, 0.1, 0.25, 0.5, 1, 2.5, and 5 mM respectively. The expected masses are 31.6, 63.2, and 94.8 kDa for monomer, dimer, and trimer, respectively.

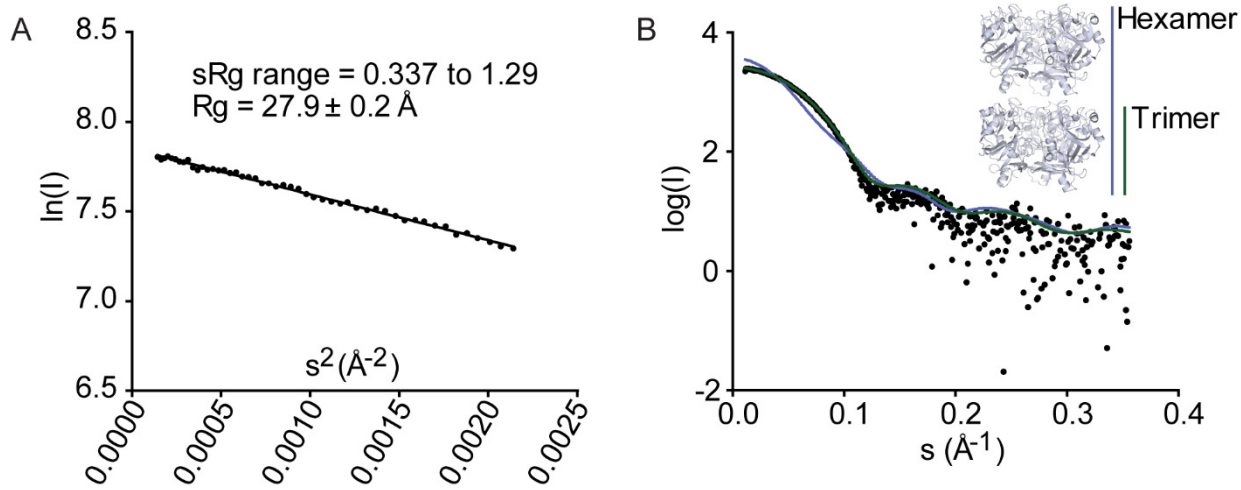



**Figure 2.7**: SAXS of XEEL$_{CRD}$ confirms trimeric assembly. (A) Gunier plot. (B) Simulated scattering curve of timer (green) and hexamer (blue) in the crystal structure compared with the experimental data (black). (C) Pair-distance distribution plot with total probability normalized to unity. (D) Molecular envelope based on dummy atom model with docked trimeric XEEL$_{CRD}$.

Because the protein concentration used in SAXS is relatively high, it is not clear whether the interaction between monomers is high in affinity. XEEL$_{CRD}$ was further

characterized by equilibrium analytical ultracentrifugation. The absence of curvature in the ln(Absorbance) vs radial position squared plots indicated that only a single molecular species is present (Figure 2.8). The reduced mass ($M_r$) obtained was 27,120 Da over $XEEL_{CRD}$ trimer concentrations ranging from 0 to 28.5 μM. The calculated molecular mass was 96,811 Da, which corresponds to an aggregation number of 3.08, based on the molecular mass of the $XEEL_{CRD}$. This results is consistent with the SAXS data suggesting trimeric $XEEL_{CRD}$. Moreover, the centrifugation data indicate that assembly of the $XEEL_{CRD}$ trimer proceeds with high affinity because no free monomeric $XEEL_{CRD}$ was observed at any point during the experiment. Therefore, $XEEL_{CRD}$ forms a stable trimer in solution despite lacking intermolecular disulfide bonds.

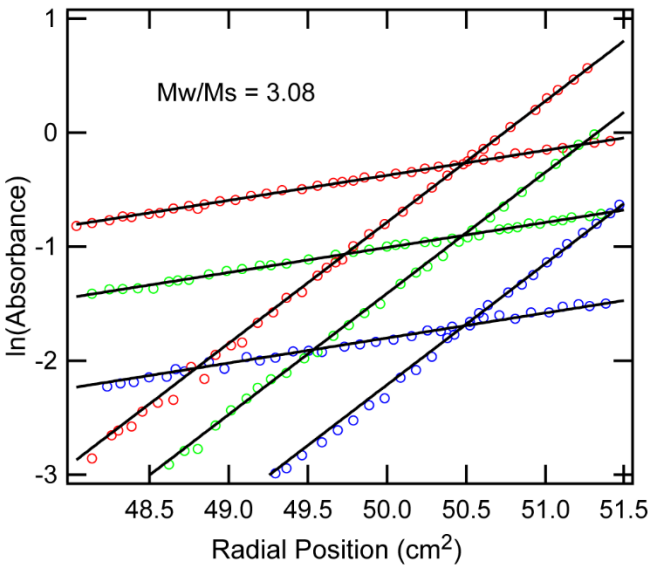



**Figure 2.8**: Sedimentation equilibrium analytical centrifugation of $XEEL_{CRD}$. The ln(Absorbance) vs radial position squared ($cm^2$) plot of $XEEL_{CRD}$ at 6,000 and 13,200 RPM for each $XEEL_{CRD}$ concentration (2.6, 5.2, and 9.0 μM). Only every third raw data point is shown for clarity. The solid lines represent the single species fit of the data that was used to calculate the reduced mass.

Analysis of the interfaces in the $XEEL_{CRD}$ structure indicates that the trimer has the total buried surface area of 4,715 $Å^2$. Each monomer has an average buried surface

area of 1,5712 Å$^2$. The interface surfaces are polar, but not highly charged. Interactions between each monomer are include 18-19 hydrogen bonds and 1-3 salt bridges. Amino acid sequences at the interface are not highly conserved as determined by ConSurf[158]. With the large buried surface area and the only two unique intelectin structures available to date, the relationship between sequence variation in the monomer interface and the oligomeric state of intelectins remains to be explored.

### 2.4.7 Full length XEEL is a dimer of trimers, suggesting agglutination activity

Because full length XEEL is a disulfide-linked hexamer but XEEL$_{CRD}$ is trimeric, the structure of full length XEEL likely exists as a tail-to-tail dimer of trimeric XEEL$_{CRD}$. We hypothesize that residues 22-47 function as a dimerization domain. Secondary structure predictions of this sequence strongly suggest a helical structure[159]. Therefore, we propose that the dimerization domain is an anti-parallel six-helix bundle held together by six disulfide bonds. A consequence of this molecular arrangement is the presentation of three ligand-binding sites on once face, and another three on the opposite face (Figure 2.9). This barbell-like model suggests that XEEL may bind two microbial surfaces simultaneously, thus possessing agglutination activity. To validate this hypothesis, we tested whether full length XEEL can agglutinate serotypes of *S. pneumoniae* with and without XEEL ligands in their capsular polysaccharide (Figure 2.10). Indeed, *S. pneumoniae* serotype 43 (which contains GroP) and serotypes 20 and 70 (which contain β-Gal*f*) were agglutinated by full length XEEL. Furthermore, this agglutination activity requires calcium ion, as inclusion of excess EDTA abolished agglutination. No agglutination was observed with serotype 8 which lacks an exocyclic vicinal diol in its capsular polysaccharide.

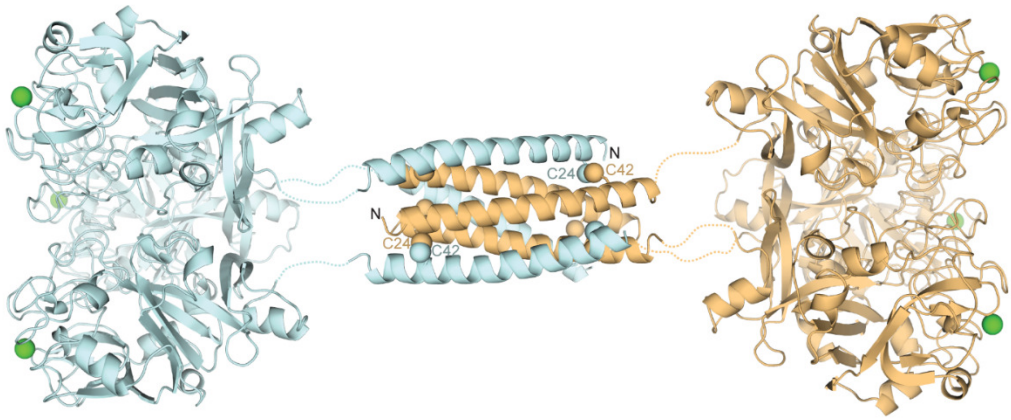



**Figure 2.9**: The predicted structure of full length XEEL. The ligand binding site on each monomer (derived from the crystal structure of $XEEL_{CRD}$) is marked by a calcium ion (green). The predicted structure of the disulfide-linked dimerization domain (derived from PDB ID 2SIV[160]) is shown as a helical bindle with two of the six C24-C42 disulfide bonds labeled.

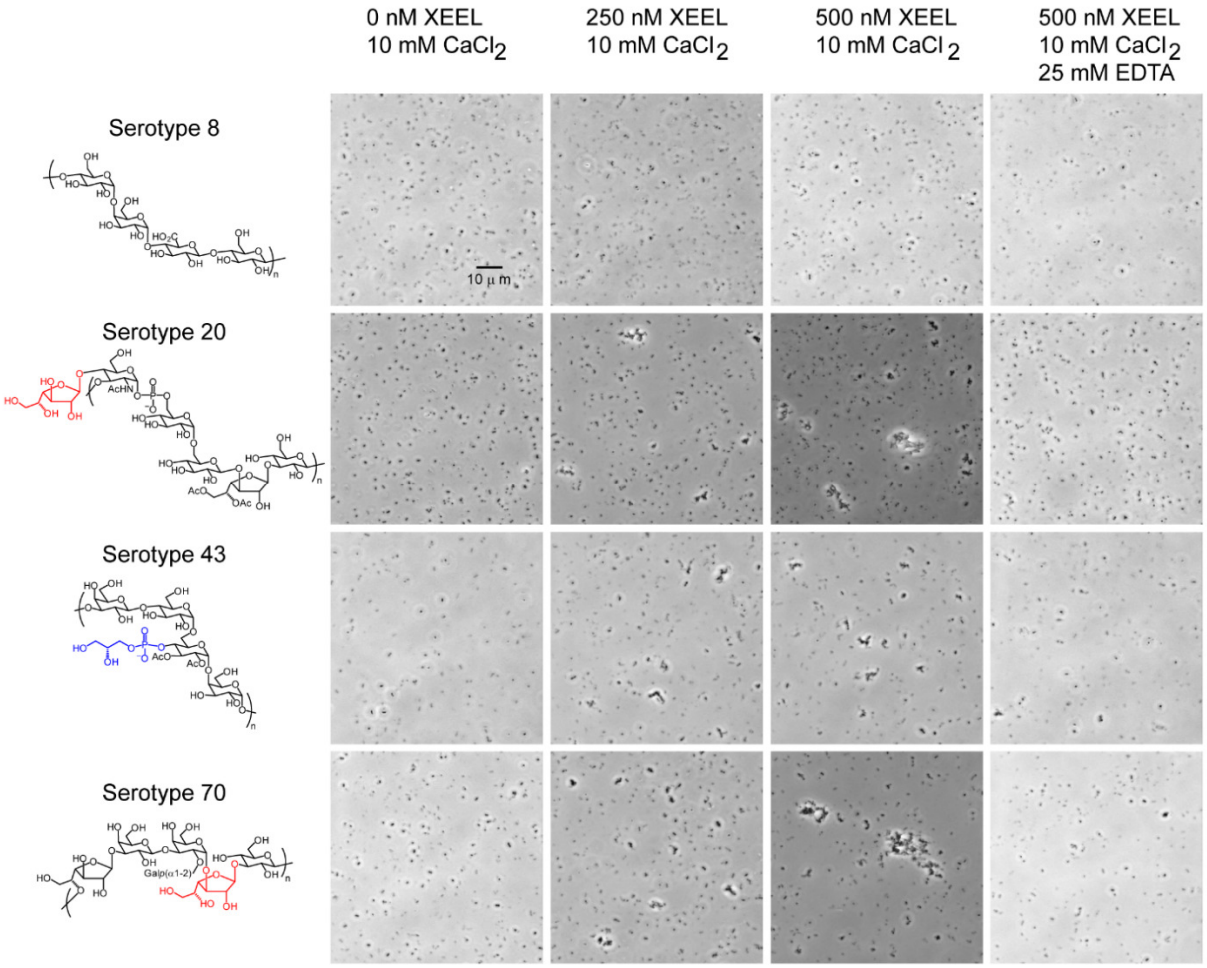


**Figure 2.10**: Agglutination of *Streptococcus pneumoniae* serotypes by full length hexameric XEEL. Structure of glycan epitopes for each strain[161] are depicted with β-Gal*f* highlighted in red and GroP in blue.

## 2.5 Discussion and Conclusions

Intelectins are soluble lectin found in organisms ranging from tunicates to humans, and they have been suggested to participate in innate immunity. Despite this proposal and the high sequence similarity in the intelectin family, little is known regarding intelectin function. An understanding of structure and carbohydrate-binding specificity can illuminate intelectin functional roles. The human intelectin hIntL-1 binds a range of microbe-specific epitopes[120]. To examine whether other homologs share similar binding modes and oligomeric states, we determined the structure of the carbohydrate-recognition domain of the *Xenopus* homolog, XEEL, alone and complexed to GroP.

The X-ray crystal structure of $XEEL_{CRD}$ indicates that the intelectins are a discrete lectin class. Though the intelectins and ficolins share an FBD, our X-ray structures show that they are structurally divergent. Intelectins, ficolins and many C-type lectins possess structural calcium ions. Still, the structural calcium site of XEEL bears no resemblance to the single structural calcium site present in ficolins nor to the two structural calcium ions in the C-type lectin DC-SIGN (Figure 2.11). The structural calcium ions of intelectins are buried, which contrasts with the solvent-exposed site in L-ficolin. In addition, intelectins use protein residues from three loops (Figure 2.5) to coordinate their structural calcium ions while ficolins employ side chains within a short single loop [54]. The intelectin site is also distinct from the two non-conserved structural calcium ions in the C-type lectin DC-SIGN, which are solvent-exposed and are coordinated by two protein loops [40, 162].

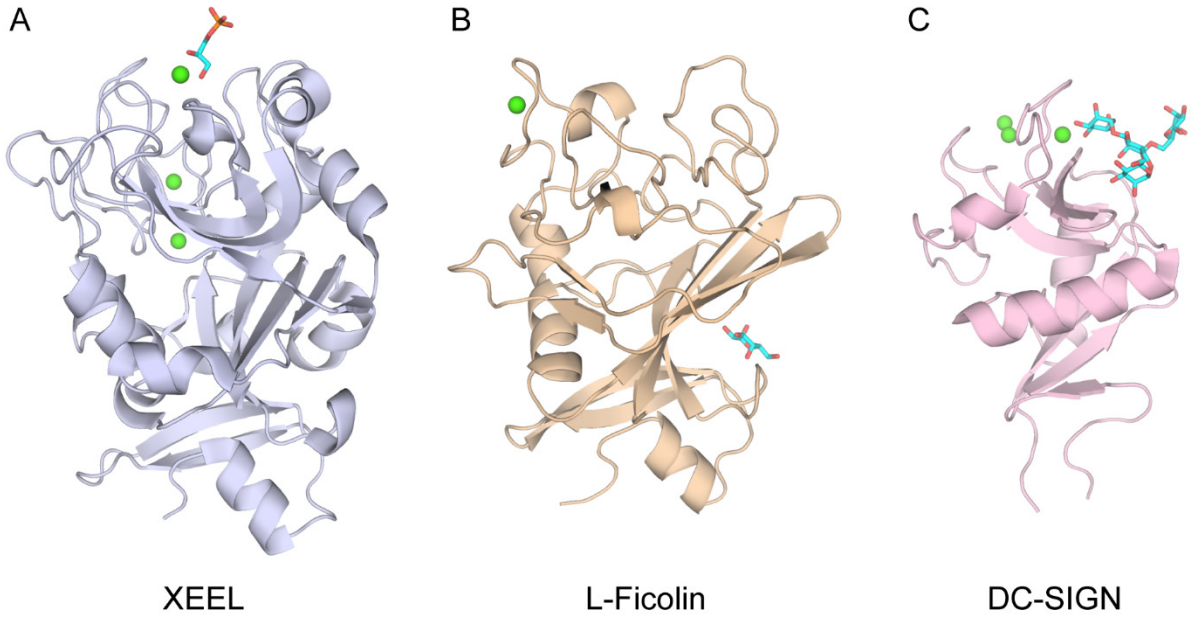



**Figure 2.11**: Comparison of lectin structures. A) XEEL from this work (PDB ID 4WN0). B) L-ficolin CRD (PDB ID 2J3U) [54]. C) DC-SIGN (PDB ID 1SL4) [40]. Calcium ions are represented by green spheres and the carbohydrate ligands are shown in cyan.

The ligand-binding site of the intelectins is also different from that of other lectins. In the intelectin structures solved to date, a calcium ion directly coordinates an exocyclic vicinal diol present in the carbohydrate epitopes. Ficolin structures show no direct interactions between the calcium ion and the ligand. Alternatively, C-type lectins, such as DC-SIGN, recognize their carbohydrate ligands through calcium coordination to adjacent secondary hydroxyl groups within a pyranose ring (Figure 2.11). As would be anticipated from its ability to coordinate to two secondary hydroxyl groups, the ligand-binding site of DC-SIGN contains no electron-rich aromatic residues that surround the calcium-coordinated ligand, as observed in hIntL-1 and the $XEEL_{CRD}$. These observations underscore that intelectins are a distinct lectin class.

We attribute the ligand binding specificity of $XEEL_{CRD}$ to W317 and W326, which form a tight box around the vicinal diol and impose steric constraints during ligand binding (Figure 2.4). The structure of GroP-bound $XEEL_{CRD}$ confirms that GroP does

indeed bind in the same site as β-Gal*f* on hIntL-1. Both hIntL-1 and XEEL contain an aromatic residue (W326 in XEEL, Y297 in hIntL-1) in the binding site that serves an aromatic wall, a steric barrier that prevents the binding of more substituted diols. Moreover, the placement of the aromatic residue suggests it can engage in a CH-π interaction with the methylene group of the 1,2-diol-containing ligands [163]. Coordination of the diol to a calcium ion should increase the strength of the CH-π interaction [164]. We anticipate that intelectins with this dual aromatic residue signature will show similar ligand-binding modes and specificities.

The oligomeric state of intelectins appears to vary among family members. For example, hIntL-1 is a disulfide-linked trimer, whereas XEEL is a disulfide-linked hexamer. We show that $XEEL_{CRD}$ is trimeric in solution despite lacking intermolecular disulfide bonds. The intermolecular disulfide bonds that exist in full length XEEL can be mapped to C24 and C42, the only two cysteine residue missing from the $XEEL_{CRD}$ studied for much of this work. Thus in XEEL, the intermolecular disulfide bonds may be required for forming stable high-order hexamers, but is not required for trimerization of the $XEEL_{CRD}$. Previous studies have explored the significance of intelectin oligomeric states in biological activities by altering these N-terminal cysteines, and subsequently confirming the monomeric state by denaturing SDS-PAGE[165, 166]. This experimental design cannot reveal high affinity, non-covalent oligomeric states. The work presented here suggests that additional solution phase techniques such as crosslinking, SAXS, or equilibrium analytical ultracentrifugation are required to establish the true oligomeric state of an intelectin prior to biochemical studies.

Intermolecular disulfide bonds are found in several important proteins such as antibodies, hormones, and growth factors [167-169]. Their presence is shown to provide stability and is often crucial for the biological function [170]. For example, TGF-β3, in which the intermolecular disulfide bond was mutated, is monomeric and has diminished biological activity [169]. It has been suggested that disulfide bonds are conserved once acquired through evolution, and the acquisition of disulfide bond is associated with organismal complexity [171]. Whether this trend is observed for all intermolecular disulfide bonds is not clear, as we did not find a systematic evolutionary study relating the acquisition or loss of an intermolecular disulfide bond to protein oligomeric state. With regard to intelectins, mouse intelectins do not contain intermolecular disulfide bonds, but their oligomeric state is not yet known. Given our work on $XEEL_{CRD}$ and the large sequence variation in the N-terminus of intelectins, we suggest that biophysical characterization of intelectin oligomeric state in solution in combination with sequence analysis could provide insight into intermolecular disulfide bond evolution.

The structure of $XEEL_{CRD}$ presented here allows prediction of the structure of full length XEEL. We hypothesize that the structure is a dimer of trimers held together by the N-terminal helical bundle that was removed by proteolysis in our experimental system. Because the bundle would align the two $XEEL_{CRD}$ faces in opposite directions, this barbell-like structural features suggests that XEEL could agglutinate bacteria. This hypothesis was validated by showing that full length XEEL agglutinates *S. pneumoniae* in a ligand- and calcium ion-dependent manner. This agglutination activity could be further optimized and utilized for pathogen detection applications. The work presented

in this chapter establishes a foundation for crystallographic analysis of hIntL-1 (Chapter 3) and evolutionary analysis of intelectins (Chapter 4).

# Chapter 3

## Structural basis of carbohydrate recognition by human intelectin-1

Portions of this chapter have been published in:

Wesener, D. A., Wangkanont, K., McBride. R., Song, X., Kraft, M. B., Hodges, H. L., Zarling, L. C., Splain, R. A., Smith, D. F., Cummings, R. D., Paulson, J. C., Forest, K. T., Kiessling, L. L. Recognition of microbial glycans by human intelectin-1. *Nat Struct Mol Biol*. 2015, 22, 603-610.

## 3.1 Abstract

Human intelectin-1 (hIntL-1) is expressed in mucosal barriers and has been proposed to function in the innate immune system. We used glycan arrays to show that hIntL-1 did not interact with human glycan epitopes but bound microbe-specific glycan containing β-D-galactofuranose (β-Gal*f*), D-glycerol 1-phosphate (GroP), heptoses, D-*glycero*-D-*talo*-oct-2-ulosonic acid (KO) and 3-*deoxy*-D-*manno*-oct-2-ulosonic acid (KDO) residues. Structural information was needed to understand how hIntL-1 can recognize ligands with such diverse structures. X-ray crystallography of hIntL-1 in complex with β-Gal*f* revealed that hIntL-1 used a bound calcium ion to coordinate exocyclic vicinal diol, an epitope common to all hIntL-1 ligands. Modeling studies suggested that steric clashes and electrostatic repulsion likely excluded N-acetyl neuraminic acid (Neu5Ac), an exocyclic vicinal diol-containing residue prevalent in the human glycans, from the hIntL-1 binding site. Structural insight into the oligomeric assembly and ligand recognition mechanism of hIntL-1 provides a platform that not only suggests the biological function of hIntL-1, but will also enable intelectin engineering for microbial detection and targeted therapeutics development.

## 3.2 Introduction

Cell surface glycans serve as identity markers in numerous biological events, especially in distinguishing self from non-self[172, 173]. These glycans are recognized by carbohydrate-binding proteins or lectins[174]. For example, mannose-binding lectin binds mannosylated glycans on invading microbes and activates the complement system, leading to pathogen clearance[18, 175]. Ficolins, galectins, and collectins have also been

demonstrated as immune lectins[122, 172, 176]. One group of lectins whose immune function has been proposed but remain poorly characterized is intelectins.

The first intelectin described was isolated from *Xenopus laevis* oocyte, thus the intelectin family is also called the X-type lectin family[110, 111]. In the years following, homologs of this intelectin have been identified throughout chordates, including tunicates, fishes, and mammals[112]. Intelectins are homo-oligomers of 35 kDa monomers that, like C-type lectins, require calcium ion to bind carbohydrates despite lacking a C-type lectin sequence motif[135]. However, intelectins contain a fibrinogen-like domain (FBD) also present in immune lectins such as ficolins[113]. Thus, based on sequence similarity intelectin has been proposed as an immune lectin. In support of this, mammalian intelectins are expressed at mucosal barriers such as lung and intestine where they encounter microbes[113, 177, 178]. In mice and sheep, intelectin is upregulated upon parasitic nematode infection, further supporting the role of the intelectin family in immune defense[115, 127]. Moreover, human intelectin-1 (hIntL-1) is found in asthmatic mucus[179, 180]. In addition, hIntL-1 was shown to be a lactoferrin receptor[181] and may also be involved in modulating insulin sensitivity[182].

Given that hIntL-1 has diverse functions, we hypothesize that thorough examination of hIntL-1 ligand specificity might provide clues to hIntL-1 function. hIntL-1 was shown to bind furanose residues such as ribofuranose (Rib*f*) and β-galactofuranose (β-Gal*f*)-containing disaccharides[113, 183]. Using an ELISA (Figure 3.1 A), we showed that hIntL-1 binds β-Gal*f* (Figure 3.1 B) with apparent $K_d$ of 85±14 nM (8.0±1.3 μg/mL) (Figure 3.1 C). However, binding to galactopyranoside (Gal*p*) and Rib*f* was not detected using this approach. These results suggest that hIntL-1 is not a general furanoside

binder, but is rather specific for β-Gal*f*. Given the limited scopes of ligand tested, a more extensive screen is required to truly assess ligand specificity.

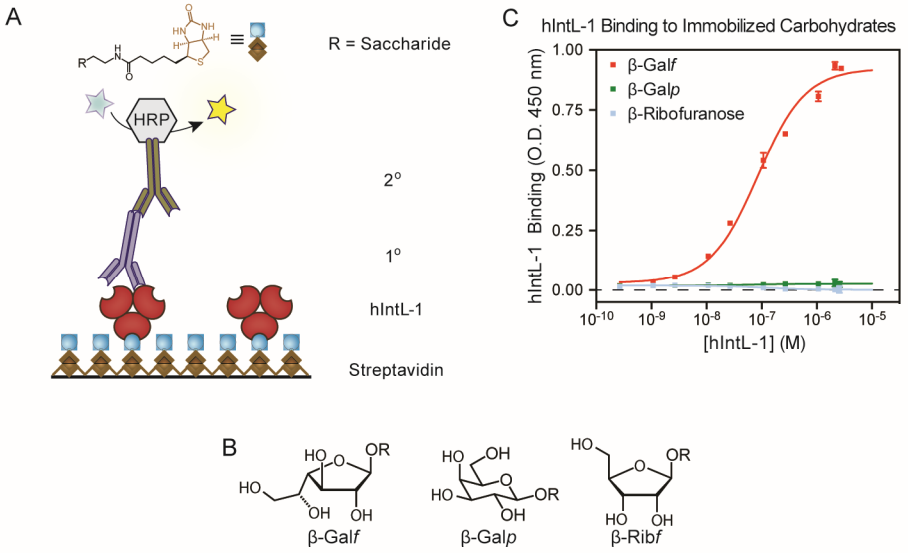



**Figure 3.1**: (A) Schematic for an ELISA used to detect binding of hIntL-1 to glycosides. (B) Structure of biotin-linked carbohydrate used in the ELISA assay. (C) Binding curve from the ELISA showing specific interaction between hIntL-1 and β-Gal*f*.

Using the mammalian glycan microarray v5.1 generated by the Consortium for Functional Glycomics[184], we discover that hIntL-1 does not bind any of the more than 600 glycans present on the array (Figure 3.2 A). However, hIntL-1 binds β-Gal*f*-containing glycans printed on a control array (Figure 3.2 B). These results suggest that hIntL-1 is functional and the detection method is valid, but no intelectin ligand is present in the mammalian glycan. Given the proposed role of intelectin in recognizing microbes, it is conceivable why hIntL-1 may not interact with mammalian glycans. Therefore, we further explore the ligand scope of hIntL-1 using a pathogen glycan array consisting of more than 300 oligosaccharides[185]. In contrast to the mammalian array, the results from the pathogen glycan array reveal several ligands for hIntL-1 (Figure 3.2 C). Considering the top 15 ligands, four ligands from *Klebsiella pneumoniae* and *Streptococcus*

*pneumoniae* contain β-Gal*f* residue. Surprisingly, other ligands do not contain the β-Gal*f* epitope. The top five hits are glycans from *S. pneumoniae*, *Proteus mirabilis*, and *Proteus vulgaris*, sharing a common D-glycerol 1-phosphate residue. Other ligands from the top 15 hits contain heptose, KO, and KDO residues representing glycans from *Yersinia pestis* and *S. pneumoniae*. Structure of all these ligand residues, which are widely distributed among bacteria[2], are shown in Figure 3.2 D. The results from the pathogen glycan array indeed suggest that hIntL-1 binds microbe-specific glycan. However, the structures of hIntL-1 ligands are diverse. It is not clear how hIntL-1 can recognize such disparate ligands. Therefore, structural information is needed to understand the molecular basis of ligand recognition. Such knowledge will additionally shed light on the structure and function of intelectins in general.

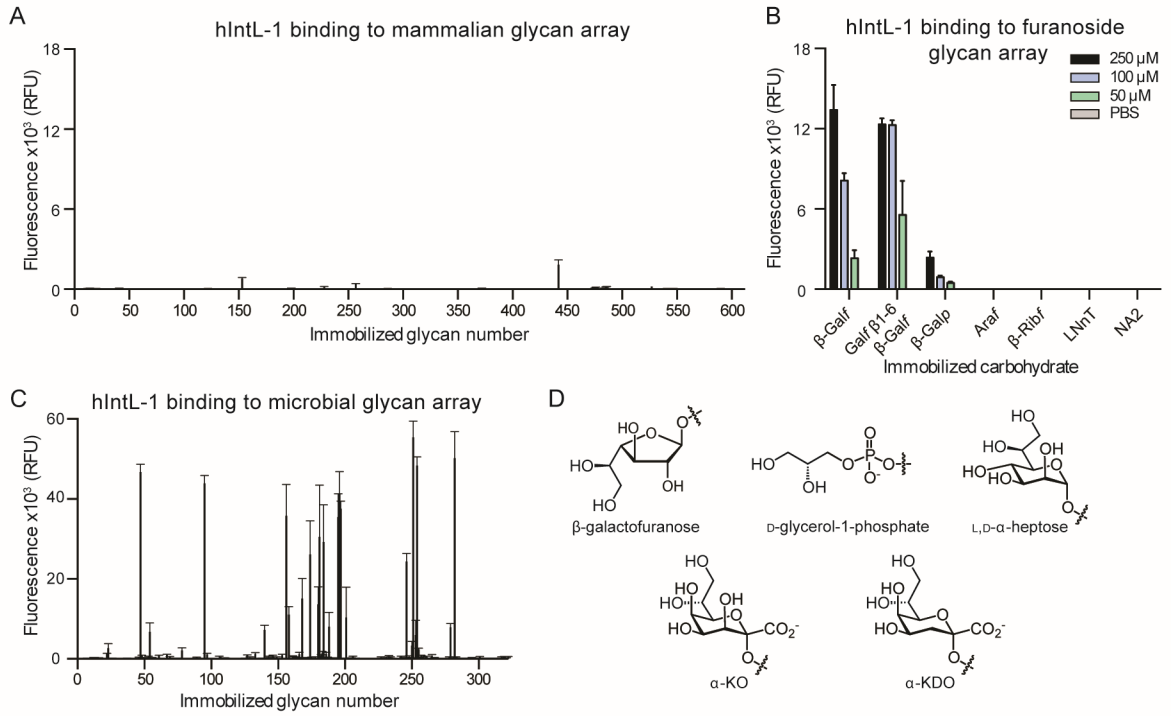



**Figure 3.2**: (A) Mammalian glycan array showed no signal from hIntL-1 binding. (B) Control array displaying positive controls (β-Gal*f*-containing ligands) and negative controls (arabinofuranose (Araf), ribofuranose (Ribf), lacto-N-neotertraose (LNnT), and

asialo, glycosylated biantennary N-linked glycan (NA2)). (C) Microbial glycan array results showed multiple hits for hIntL-1 binding. (D) Structure of common residues shared among hIntL-1 ligands.

### 3.3 Material and Methods

hIntL-1 was expressed with an N-terminal Strep-tag II and purified using a similar procedure as $XEEL_{CRD}$ (Chapter 2). The protein was concentrated to 1.5 mg/mL and $CaCl_2$ added to 10 mM. Crystallization was achieved using hanging-drop vapor diffusion by mixing 1 µL of the protein solution with 1 µL of well solution (100 mM Bis-Tris pH 6.0 and 25% PEG 3350). Crystals grew to full size in two weeks. Crystals of apo-hIntL-1 were cryoprotected by transferring to well solution containing 35% PEG 3350 for 1 min then vitrified in liquid nitrogen. The allyl-β-Gal*f*-hIntL-1 complex was formed by soaking of apo-hIntL-1 crystals in the cryoprotection solution supplemented with 50 mM allyl-β-D-Gal*f* for two weeks.

Single crystal X-ray diffraction experiments were performed at beamline 21-ID-D (Life Sciences Collaborative Access Team), Advanced Photon Source, Argonne National Laboratory. Integration, scaling, and merging were performed with HKL2000[138] (Table 3.1). The structure was solved using the PHENIX suite[186]. The monomeric structure of Se-Met $XEEL_{CRD}$ was used as a search model to determine the structure of apo-hIntL by molecular replacement using Phaser[140]. Because the apo-hIntL and allyl-β-D-Gal*f*-bound hIntL-1 data are isomorphous, the structure of A-Galf-bound hIntL-1 was solved by a difference Fourier method using apo-hIntL as a starting model for rigid-body refinement with phenix.refine[141]. Chemical restraint for allyl-β-D-Gal*f* was generated using PRODRG[143]. Crystallographic refinement was performed using phenix.refine.

Model adjustment was made in Coot[142] and validation was performed with MolProbity[144]. Structure figures were generated using PyMOL[145].

**Table 3.1**: Data collection and refinement statistics for apo-hIntL-1 and Gal*f*-bound hIntL-1.

| Data collection statistics | Apo-hIntL-1 | Gal*f*-bound hIntL-1 |
|---|---|---|
| Wavelength (Å) | 0.97924 | 1.00394 |
| Resolution range (Å)* | 22.00-1.80 (1.86-1.80) | 28.59-1.60 (1.66-1.60) |
| Space group | P $2_1$ 3 | P $2_1$ 3 |
| Unit cell (Å) | 118.4 | 117.9 |
| Total reflections | 573173 | 801024 |
| Unique reflections | 51379 (5062) | 71869 (7134) |
| Multiplicity | 11.2 (10.1) | 11.1 (10.9) |
| Completeness (%) | 100 (100) | 100 (100) |
| Mean I/σ(I) | 19.6 (3.7) | 29.4 (3.0) |
| Wilson B-factor (Å$^2$) | 12.89 | 17.63 |
| R-merge | 0.119 (0.495) | 0.078 (0.773) |
| R-meas | 0.125 (0.521) | 0.082 (0.811) |
| R-pim | 0.037 (0.162) | 0.024 (0.245) |
| Refinement Statistics | | |
| Resolution range (Å) | 22.00-1.80 (1.86-1.80) | 28.59-1.60 (1.66-1.60) |
| R-factor | 0.1331 | 0.1544 |
| R-free (5%) | 0.1664 | 0.1793 |
| Number of atoms | | |
| Protein | 4562 | 4606 |
| Calcium | 6 | 6 |
| Gal*f* | 0 | 30 |
| Water | 670 | 619 |
| Protein residues | 564 | 564 |
| RMSD (bonds, Å) | 0.010 | 0.010 |
| RMSD (angles, °) | 1.102 | 1.126 |
| Ramachandran favored (%) | 96.3 | 96.9 |
| Ramachandran outliers (%) | 0 | 0 |
| Average B-factor (Å$^2$) | | |
| Protein | 13.26 | 19.39 |
| Calcium | 9.39 | 14.66 |
| Gal*f* | 0 | 32.73 |
| Solvent | 25.69 | 31.95 |

*Statistics for the highest-resolution shell are shown in parentheses. Coordinates and structure factors of apo-hIntL-1 and Gal*f*-bound hIntL-1 were deposited at the Protein Data Bank under accession code 4WMQ and 4WMY, respectively.

## 3.4 Results and Discussion

### 3.4.1 Structure of apo-hIntL-1

An apo-hIntL-1 crystal diffracted X-ray to 1.8 Å. The cubic apo-hIntL-1 crystal has solvent content of 38.9% with Matthews coefficient[187] of 2.01. Two chains (monomers) of hIntL-1 were identified in the asymmetric unit. Each chain is a part of a distinct trimer that could be generated by crystallographic symmetry. The two chains are similar in conformation, but not identical (Cα RMSD of 0.65 Å). In chain A, the peptide connecting each monomer and forming an intermolecular disulfide bond is resolved (Figure 3.3). The intermolecular disulfide bond consists of C31 from one monomer and C48 from the neighboring monomer forming a cyclic disulfide-linked trimer. C31 is the first resolvable residue in the electron density map in chain A. In chain B, the peptide linking each monomer is not visible. However, W29-C31 is traceable and the rest of the monomer stars from S36. In both chains, the C terminal arginine is entirely resolved.

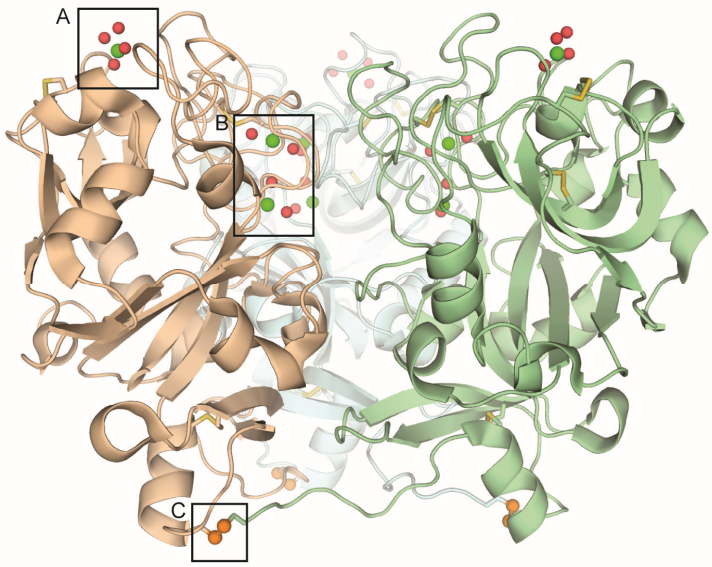



**Figure 3.3**: Structure of disulfide-linked trimeric hIntL-1. Box A indicates ligand binding site and Box B indicates structural calcium site. The disulfide linking between monomers is shown in Box C.

Each monomer of hIntL-1 is similar in conformation to monomeric XEEL$_{CRD}$ after refinement. (Cα RMSD of 0.48 Å). Like XEEL$_{CRD}$, each monomer of hIntL-1 contains three bound calcium ions: one in the ligand binding site and two in the structural calcium site (Figure 3.3). The overall coordination environment of these calcium ions are identical to those in XEEL$_{CRD}$. Four ordered, calcium-bound water molecules in the ligand binding site are clearly resolved.

### 3.4.2 Structure of hIntL-1 in complex with β-allyl-Gal*f*

To determine the mechanism of ligand binding by hIntL-1, we determined the structure of allyl-β-D-Gal*f*-hIntL-1 complex by soaking an apo-hIntL-1 crystal in the ligand solution. Allyl-β-D-Gal*f* molecules were identified in the binding site of both monomers in the asymmetric unit. Each allyl-β-D-Gal*f* sits in the ligand binding site on the same face of the hIntL-1 trimer (Figure 3.4). This mode of binding indicates that hIntL-1 can bind multiple ligand simultaneously. The multivalent binding mode explains high avidity of hIntL-1 for immobilized ligand shown in ELISA. The Cα RMSD of 0.12 Å between the asymmetric unit of apo- and allyl-β-D-Gal*f*-bound hIntL-1 suggests minimal structural changes upon ligand binding.

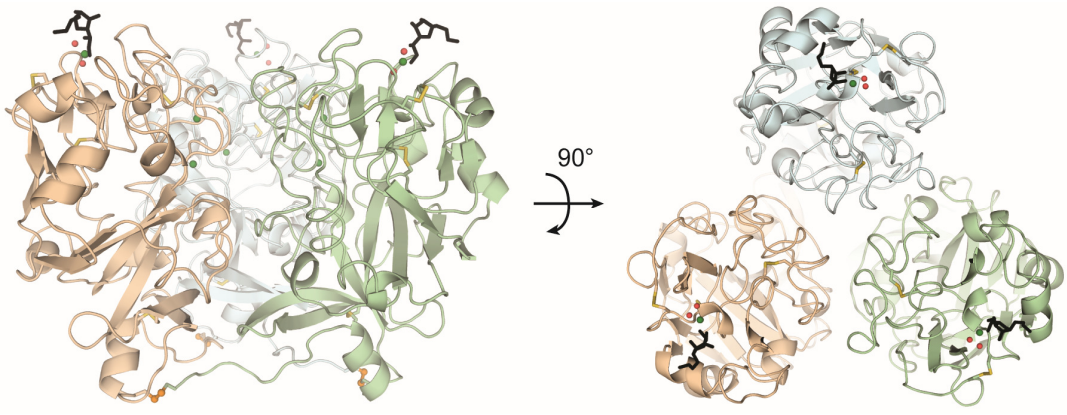



**Figure 3.4**: Structure of trimeric hIntL-1 in complex with allyl-β-D-Gal*f* (black).

Binding of allyl-β-D-Gal*f* displaces two calcium-bound water molecules. The O(5) and O(6) hydroxyls of Gal*f* coordinate directly to the binding site calcium ion (Figure 3.5). The primary O(6) hydroxyl also hydrogen bonds to H263. No direct interactions between other oxygen atoms on allyl-β-D-Gal*f* and hIntL-1 is observed. The exocyclic vicinal diol sits in a hydrophobic box formed by W288 and Y297. The resolution of the structure (1.6 Å) and quality of the map is high enough to assign the conformation of allyl-β-D-Gal*f* unambiguously (Figure 3.6). The exocyclic vicinal diol chelating the calcium ion is a gauche conformation with dihedral angles of 45° and 51° for chain A and B, respectively. Using the Altona-Sundaralingam pseudorotation model[188, 189], the calculated pseudorotational phase angles for allyl-β-D-Gal*f* in chain A and B are 105° ($^1T_0$) and 57° ($^4E$) respectively. The conformational variation is consistent with the flexibility of furanosides[190].

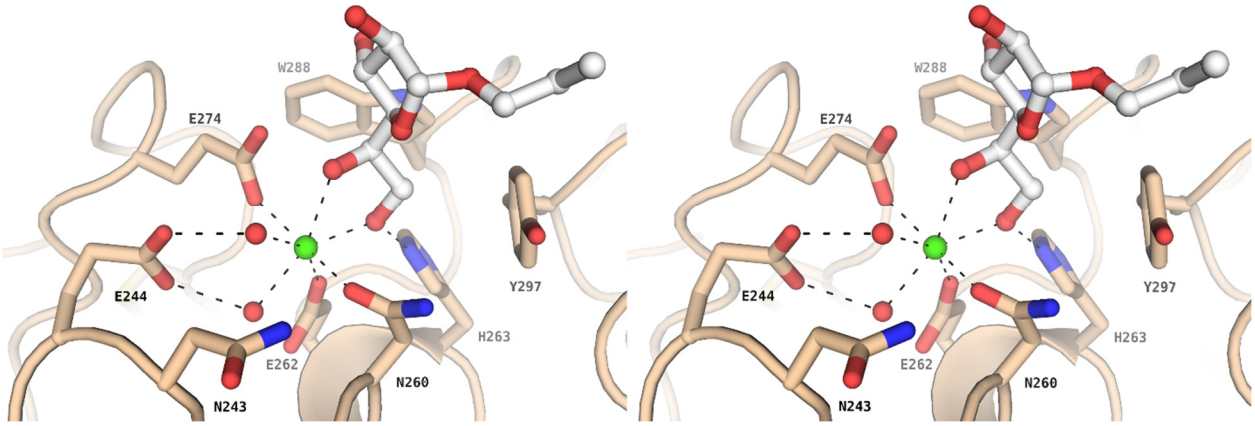
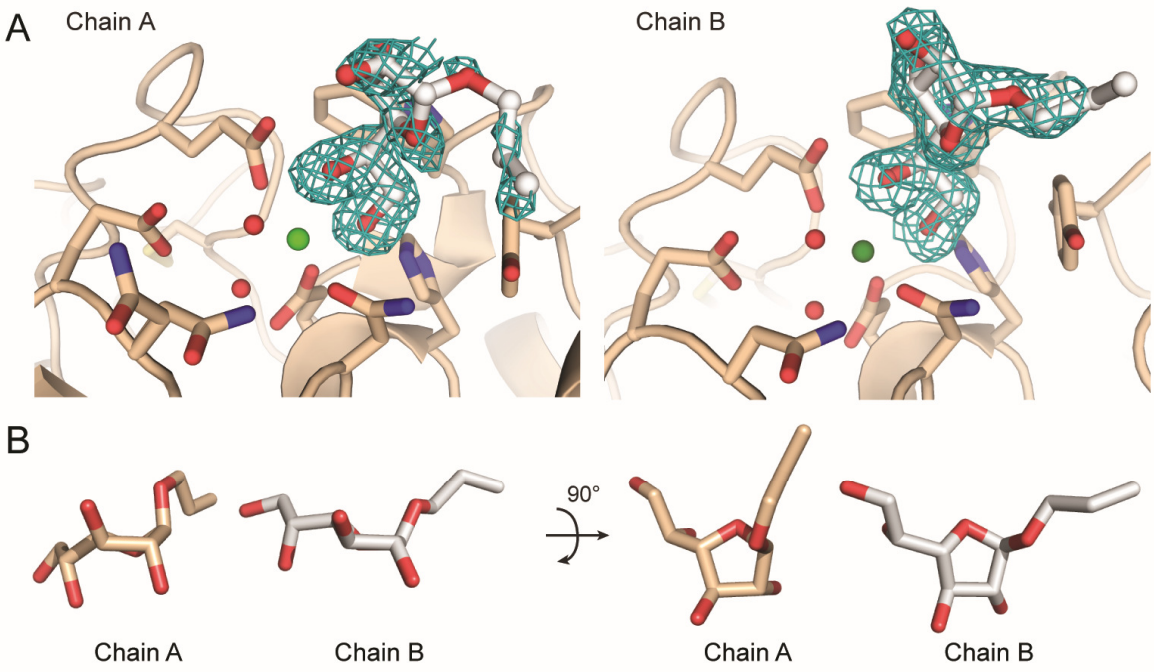



**Figure 3.5**: Wall-eyed stereo image of hIntL-1 ligand binding site containing allyl-β-D-Gal*f*.




C

| | Chain A | Chain B |
|---|---|---|
| Furanoside Conformation | $^1T_0$; Calculated P = 105° | $^4E$; Calculated P = 57° |
| Exocyclic Diol Dihedral | 45° | 51° |
| Gal*f* Conformation | $^1T_0$-gg-gt | $^4E$-gg-gt |

**Figure 3.6**: (A) Electron density map ($mF_o$-$DF_c$, 3σ) calculated after allyl-β-D-Gal*f* residues were removed and structure re-refined. (B) Comparison of allyl-β-D-Gal*f* structure in chain A and B. (C) Comparison of conformational parameters for the furanoside ring of allyl-β-D-Gal*f* in chain A and B.

### 3.4.3 Alignment of ligand-bound hIntL-1 and trimeric XEEL$_{CRD}$ structures

Monomeric XEEL$_{CRD}$ was used successfully as a search model for solving hIntL-1 structure by molecular replacement, thus the tertiary structures are highly conserved between the two proteins. Indeed, the Cα RMSD of the refined monomeric XEEL$_{CRD}$ and hIntL-1 is 0.54 Å confirming that these two proteins are similar in structure. However, due to the lack of amino acid conservation at the trimeric interfaces as mentioned in Chapter 2, it is unclear whether the trimeric architecture and the position

of the ligand binding site are conserved in three dimensional space. Superposition of the trimeric ligand-bound hIntL-1 and XEEL$_{CRD}$ reveals no significant structural differences (Cα RMSD = 0.71 Å) (Figure 3.7 A). The ligand binding sites are almost identical in position (Figure 3.7 B). These results highlight structural conservation among intelectins which are discussed in greater detail in Chapter 4.

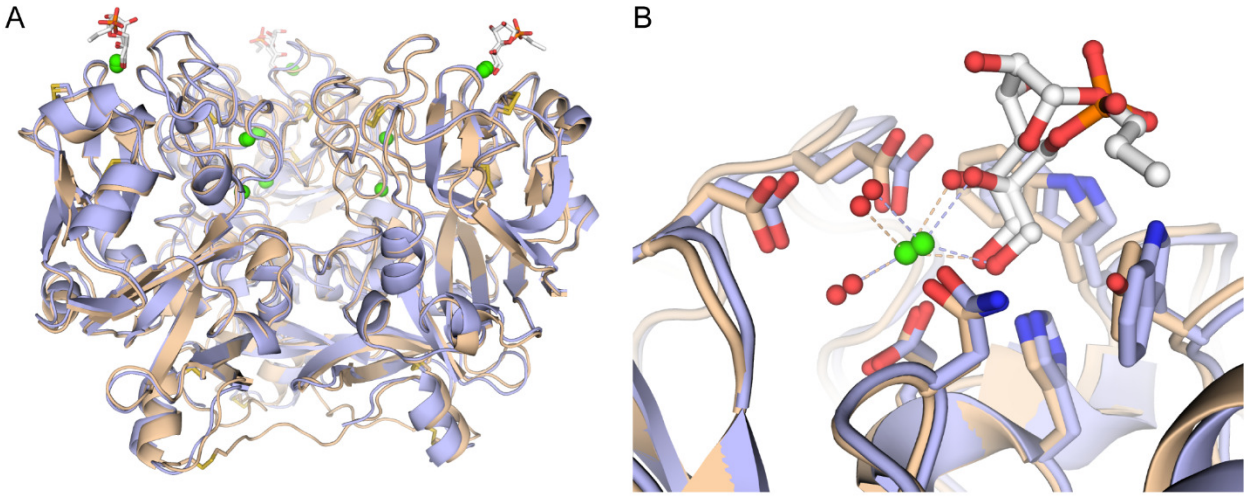



**Figure 3.7**: (A) Alignment of trimeric hIntL-1 (wheat) and XEEL$_{CRD}$ (light blue). (B) Superposition of the ligand binding sites from trimeric hIntL-1 and XEEL$_{CRD}$.

### 3.5 Conclusions

The structure of trimeric hIntL-1 reveals disulfide linkage between monomers. This observation is consistent with non-reducing SDS-PAGE analysis indicating trimeric molecular weight. The ligand binding sites are on the same face of the trimer suggesting multivalent mode of binding. The results of this study explain how hIntL-1 binds carbohydrates with various structures by recognizing the exocyclic vicinal diol epitope common among Gal*f*, GroP, heptoses, KO, and KDO. This recognition mode is conserved between hIntL-1 and XEEL$_{CRD}$. Direct calcium ion coordination of hydroxyls is also utilized by C-type lectins[135]. However, C-type lectin recognizes hydroxyl groups on the pyranose rings of mannose, fucose, and galactose[191]. The binding pocket of

hIntL-1 is selective for exocyclic vicinal diol containing a terminal primary hydroxyl group because W288 and Y297 forms a tight hydrophobic box excluding more substituted diols from the binding site. These features suggests that ligand binding mechanism of intelectins is unique among other known lectins.

Given that hIntL-1 recognizes exocyclic vicinal diol, it is intriguing why hIntL-1 does not bind any Neu5Ac-containing glycans[192] in the mammalian glycan array. Using the electron density of the exocyclic diol as an anchor point, the lowest energy conformation of Methyl-α-Neu5Ac (as found in PDB ID 2BAT, 2P3I, 2P3J, 2P3K, 2I2S, 1KQR, 1HGE, and 1HGH)[193-196] was docked into the ligand binding site of hIntL-1 (Figure 3.8 A). In addition, methyl-α-KDO, a known hIntL-1 ligand, was docked as a control (Figure 3.8 B). Methyl-α-KDO is readily accommodated on the binding site. However, methyl-α-Neu5Ac shows both steric clashes and electrostatic repulsion between its carboxylate and the carboxylate of E274 side chain. Conformational change in methyl-α-Neu5Ac to relieve these interactions likely results in steric clashes of the bulky C(5) N-acetyl group with hIntL-1 surface. Thus, modeling studies provide compelling rationale for selectivity against Neu5Ac binding. Further experiments with hIntL-1 mutagenesis and Neu5Ac analogs will help validate these hypotheses. The ability of hIntL-1 to discriminate between microbial vs human exocyclic vicinal diol-containing saccharides provide further support for hIntL-1 as an innate immune defense protein.

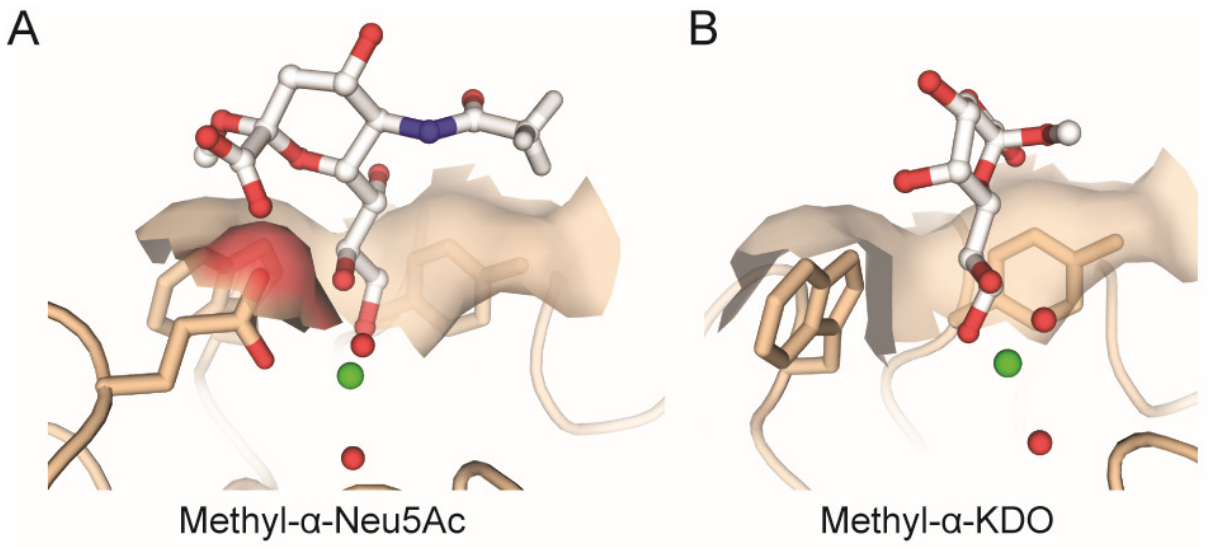



**Figure 3.8**: Model of methyl-α-Neu5Ac (A) and methyl-α-KDO (B) docked into the ligand binding site of hIntL-1.

The structure of hIntL-1 is crucial for elucidation of its biological function. Genome-wide association studies (GWAS) have linked hIntL-1 to asthma[197] and Crohn's disease[198]. The hIntL-1 structure revealed that the V109D mutation associated with asthma[197] is located at the monomer-monomer interface, remote from the ligand binding site. The V109D mutation likely affects the interaction of hIntL-1 with other proteins, but may not affect ligand binding properties. Genome-wide association studies (GWAS) have yet to link hIntL-1 to acute infection. However, many GWAS studies to date have focused on chronic diseases[199] which may not be able to correlate hIntL-1 with acute infectious diseases. Although GWAS were performed to examine susceptibility to viral infections[200-203]. Viruses still use host cell machinery to glycosylate their proteins. Thus, further GWAS examining microbes that contain hIntL-1 ligand may shed more light into the role of hIntL-1 in infectious diseases.

Exocyclic vicinal diols are a small epitope found in several carbohydrate residues, thus affording hIntL-1 a broad residue specificity and likely allowing it to bind a

broad range of microbes. These properties further support the role of hIntL-1 as a microbe sensor. The orientation of the ligand binding sites on the same face of hIntL-1 may also provide the available remaining surfaces for recruiting other proteins to the microbial surface. One such candidate protein is lactoferrin which was previously shown to bind hIntL-1 and to have antimicrobial properties[204-207]. Any synergistic effects of hIntL-1 and lactoferrin in eliminating microbes remain to be explored. The structure of hIntL-1 also suggests that the N terminus, which faces the opposite direction of the ligand binding site, could be engineered for tag insertion or attachment of other biologically active moieties. Therefore, it may be possible to utilize microbe-specific binding of hIntL-1 for immobilization and detection of microbes, or for targeted therapy of microbial infection.

# Chapter 4

## Preliminary characterization of intelectins from various model organisms

Portions of this chapter have been published in:

Wangkanont, K., Wesener, D. A., Vidani, J. A., Kiessling, L. L., Forest, K. T. Structure of *Xenopus* embryonic epidermal lectin reveals a conserved mechanism of microbial glycan recognition. *In preparation* (2015)

Wesener, D. A., Wangkanont, K., McBride. R., Song, X., Kraft, M. B., Hodges, H. L., Zarling, L. C., Splain, R. A., Smith, D. F., Cummings, R. D., Paulson, J. C., Forest, K. T., Kiessling, L. L. Recognition of microbial glycans by human intelectin-1. *Nat Struct Mol Biol*. 2015, 22, 603-610.

## 4.1 Abstract

The structures of $XEEL_{CRD}$ and hIntL-1 enable assignment of residue function and allow evolutionary analysis of intelectins. Sequence alignment shows that residues in the structural calcium site is highly conserved among intelectins, suggesting that all intelectins contains two calcium ions in the structural calcium site. In contrast, significant sequence variation is observed in the ligand binding site, both in the calcium binding residues and in the aromatic box proposed to determine ligand specificity in $XEEL_{CRD}$ and hIntL-1. This observation identifies intelectin homologs that could potentially have unique ligand specificity and novel function. To investigate biochemical properties of these intelectins, efficient expression strategies need to be established. However, intelectins can only be expressed in higher eukaryotic systems, such as insect or mammalian cells. In addition, stability and yield differences were observed with hIntL-1 in these expression systems. Thus, it is necessary to screen different expression constructs of intelectin homologs in both insect and mammalian protein expression systems. This chapter presents preliminary expression and purification of various intelectins in common model organisms. Signal peptide cleavage sites were also determined experimentally, so that purification tag could be inserted. These results should serve as a platform for further investigation of intelectin structure and function in human and various model organisms.

## 4.2 Introduction

Intelectins are found throughout the evolutionary tree of chordates[112]. Previous analysis showed a highly conserved carbohydrate recognition domain containing a fibrinogen-like domain[111, 112]. However, lack of structural information prevent further

analysis of residue function. With the availability of $XEEL_{CRD}$ and hIntL-1 structures, sequence alignment and conservation can be examine in much greater detail. Analysis of sequence alignment in structural context will not only reveal the evolution of functional residues, but will also enable identification of suitable homologs in model organisms for further studies to reveal physiological roles of intelectins.

Intelectins from Xenopus laevis (XEEL, XCGL-1, XCGL-2, XSL-1, and XSL-2), human (hIntL-1 and hIntL-2), mouse (mIntL-1 and mIntL-2), and zebrafish (DrIntL-1 and DrIntL-2) were aligned using Clustal W2 and important residues from the structure were mapped (Figure 4.1). With the exception of N118, the structural calcium site is completely conserved among all the intelectins examined, compared to XEEL. However, N118 coordinates the calcium ion through its backbone amide, thus the identity of the amino acid may not be crucial. The majority of the structural calcium site residues cluster between amino acids 115-127 with the consensus sequence of HENXXXGXCTXGD. C123 is a conserved cysteine in the intelectin fold, but does not participate in calcium ion binding. D162 and D311 are also completely conserved in the structural calcium site, but these residues are further down toward the C-terminus from the signature motif. The highly conserved nature of the structural calcium site suggests that most, if not all, intelectins contain two calcium ions in the structural calcium site.


**Figure 4.1**: Sequence alignment of intelectins found in common model organisms highlighting residues in the structural calcium site (purple) and residues in the ligand binding site (green).GenBank accession number is provided in parentheses.

In contrast to the structural calcium site, the ligand binding site residue displays lower sequence conservation within and across species. Among the five *X. laevis* intelectins, residues that directly coordinate the ligand binding site calcium ions are completely conserved. W317 and W326, which form a box around the exocyclic vicinal diol ligand, are conserved among XEEL, XSL-1, and XSL-1. These observation suggests that these intelectins share similar ligand specificity. Intriguingly, XEEL, XSL-1, and XSL-2 are proposed to participate in microbe recognition. By contrast, the corresponding residues in XCGL-1 and XCGL-2 are phenylalanine and asparagine, respectively. Because XCGL-1 and XCGL-2 are involved in fertilization membrane formation, they likely recognize self carbohydrate epitopes[109, 110]. Indeed, an unpublished glycan array results (Consortium for Functional Glycomics, primscreen_758) suggest that XCGL-1 binds Galα(1-3)GalNAc, although the biological relevance of the interaction has yet to be established[208]. Further evidence for specificity determination by W317 and W326 can be found when comparing XEEL and hIntL-1. W326 in XEEL is replaced by a tyrosine residue in hIntL-1. This conservative substitution preserves the β-Gal*f* and GroP binding activity of these proteins. These

results suggests that the W317 and W326 equivalence in intelectins are major determinants of carbohydrate specificity. In addition, they might be indicators for intelectins that function as innate immune proteins versus intelectins with other biological functions.

Several other poorly studied intelectins have more dramatic substitutions at positions W317 and W326. For example in hIntL-2, W326 is substituted with a polar serine residue. Furthermore negatively charged E303, which directly coordinates the calcium ion in the ligand binding site of XEEL and hIntL-1, is a glutamine residue in hIntL-1. This substitution results in one fewer acidic residue for calcium ion coordination. The effect of these substitutions on metal ion and ligand specificity of hIntL-2 remains to be explored. Non-conservative substitution of W317 is also observed in mIntL-2. Further sequence variations in the ligand binding site are seen in DrIntLs. Ligand specificity of these poorly studied intelectins remains unknown. Thus at present, it is not possible to understand the effect of these non-conservative substitutions.

Sequence alignment predicts that intelectins may have much broader ligand scope and metal ion diversity in the ligand binding site than previously assumed. To examine the ligand scope and biochemical properties of other intelectin homologs, it is necessary to identify appropriate expression construct and expression systems. Thus far, active XEEL and hIntL-1 can only be obtained when expressed in higher eukaryotic expression systems such as insect and mammalian cells. Better stability of hIntL-1 is observed when expressed from HEK293T cells, whereas superior yield is observed when XEEL is expressed in insect cells. Therefore, it is necessary to explore both insect and HEK293T expression systems for other intelectin homologs. In addition because

intelectins are secreted proteins. It is crucial to determine the signal peptide cleavage site for further purification tag insertion or protein engineering. This chapter presents preliminary expression and purification of intelectins from various model organisms. This study will not only enable discovery of novel intelectin properties, but will also allow identification of a suitable model organism in which biological function of intelectin could be elucidated.

## 4.3 Material and Methods

### 4.3.1 Expression constructs

Open reading frames containing wild-type intelectins including the native signal peptide (Source Biosystems or Genscript) were cloned into pFastBac1 (Table 4.1) to create recombinant baculovirus for expression in High Five insect cells or pcDNA4 (Table 4.2) for expression in HEK293T cells. pcDNA4 mIntL-1 was previously made by Darryl A. Wesener. Proteins were expressed according to manufacturer instruction for each systems (Life Technologies). In case of Strep-tag II insertion, the tag was added by site-direct mutagenesis two amino acids after the signal peptide cleavage site.

**Table 4.1**: Primers for cloning intelectins into pFastBac1

| Construct | Forward Primer | Reverse Primer |
|---|---|---|
| BamHI-XCGL-1-SalI | GCGCGGATCCATGCTGGTGCACATTCTTCTCCTGCTG | CTACGTCGACTCATTATAGATAGAAAAGTAATACAGCGGCCTCAGTTATCTCTAT |
| BamHI-XSL-1-SalI | GCGCGGATCCATGTTTGTGCACAGTCTGGTTCTTCTGTCC | CTACGTCGACTCATTAGCGGTAAAAGAGGAGTACAGCAGCCTCTG |
| BamHI-mIntL-1-HindIII | GCGCGGATCCATGACCCAACTGGGATTCCTGCTGTTTATCATGGTTGCTACCAGAGGTTGCAGT | TACCAAGCTTTCATTAGCGATAAAACAGAAGCACAGCTGCTTCAG |
| BamHI-mIntL-2-SalI | GCGCGGATCCATGACCCAACTGGGCTTCCTGCTG | CTACGTCGACTCATTAGCGATAAAACAGAAGCACAGCTGCTTCAG |

| | | |
|---|---|---|
| BamHI-hIntL-2-SalI | GCGCGGATCCATGCTGTCCATGC TGAGGACAATGACC | CTACGTCGACTCATTATCTATAGA ACAAGAGTACAGCCGCCTCCGTT |
| BamHI-DrIntL-1-KpnI | GCGCGGATCCATGATGCAGTCAG CTGGTTTTCTTCTGTTGTGTGTGC CACTCATTAGTCTACTCTGTGAAT CGATATTGTCATTGCC | GGTACCAAGCTCATTAGCGGTAG AAAATAAATACAGCCGCATT |
| BamHI-DrIntL-2-SalI | GCGCGGATCCATGGGAATGGCTG CCTTTCTTACAAAAATG | CTACGTCGACTCATTAGCGATAAA ATAGAAGCACAGCTGCTTCAATT |

**Table 4.2**: Primers for cloning intelectins into pcDNA4

| Construct | Forward Primer | Reverse Primer |
|---|---|---|
| KpnI-XCGL-1-BamHI | GCTTGGTACCATGCTGGTGCACA TTCTTCTCCTGC | GCTCGGATCCTCATTATAGATAGA AAAGTAATACAGCGGCCTCAGTTA TCTC |
| KpnI-XSL-1-BamHI | GCTTGGTACCATGTTTGTGCACAG TCTGGTTCTTCTGTC | GCTCGGATCCTCATTAGCGGTAAA AGAGGAGTACAGCAGCCT |
| HindIII-mIntL-2-BamHI | AGTTAAGCTTCACCATGACCCAAC TGGGCTTCCTGCTG | GCTCGGATCCTCATTAGCGATAAA ACAGAAGCACAGCTGCTTCAG |
| HindIII-hIntL-2-BamHI | AGTTAAGCTTCACCATGCTGTCCA TGCTGAGGACAATGACC | GCTCGGATCCTCATTATCTATAGA ACAAGAGTACAGCCGCCTCCGTT |
| KpnI-DrIntL-1-BamHI | GCTTGGTACCATGATGCAGTCAG CTGGTTTTCTTCTG | GCTCGGATCCTCATTAGCGGTAG AAAATAAATACAGCCGC |
| KpnI-DrIntL-2-BamHI | GCTTGGTACCATGGGAATGGCTG CCTTTCTTACAAAAATG | GCTCGGATCCTCATTAGCGATAAA ATAGAAGCACAGCTGCTTCAATTA TC |

### 4.3.2 Purification of intelectins

Intelectins that were examined for resin binding were dialyzed against 20 mM Bis-Tris pH 6.5, 150 mM NaCl, and 0.5 mM EDTA. The media was then dialyzed against 20 mM HEPES pH 7.5, 150 mM NaCl, and 0.5 mM EDTA. Prior to resin binding, $CaCl_2$ was added to 10 mM. After loading of intelectin solution on the an appropriate resin, the column was washed with 20 mM HEPES pH 7.5, 150 mM NaCl, and 10 mM

$CaCl_2$. Intelectins were then eluted with 20 mM HEPES pH 7.5, 150 mM NaCl, and 10 mM EDTA. Western blot analysis of intelectin was performed using a pan-intelectin antibody (Proteintech, 11770-1-AP). Edman degradation was performed at Tuft University Core Facility.

Galactose and sorbitol resins were produced by immobilization of the carbohydrate onto divinylsulfone-activated Sepharose 6B according to a published protocol for lactosyl-Sepharose[209]. Common in carbohydrate literature, and in contrast to macromolecule immobilization, large excess of ligand is used so that quenching of unreacted sites is not necessary. Therefore, quencher that contains hydroxyl groups, such as Tris, were avoided to eliminate potential secondary binding events from their carbohydrate-like structures that can obscure binding results.

## 4.4 Results and Discussion

### 4.4.1 XCGL-1 and XSL-1

XCGL-1 and XSL-1 expresses as ladder of multiple oligomeric species in insect cells as evident in non-reducing SDS-PAGE, consistent with previous reports[116, 129] (Figure 4.2). Abnormal-shaped bands that do not enter the SDS-PAGE gel well were also observed for both XCGL-1 and XSL-1. All the oligomeric states collapse into a single band in the reducing conditions, suggesting that the oligomeric species observed under non-reducing conditions are disulfide-linked. XCGL-1 expression in HEK293T cells also yields similar results. However, no expression of XSL-1 was observed in HEK293T.

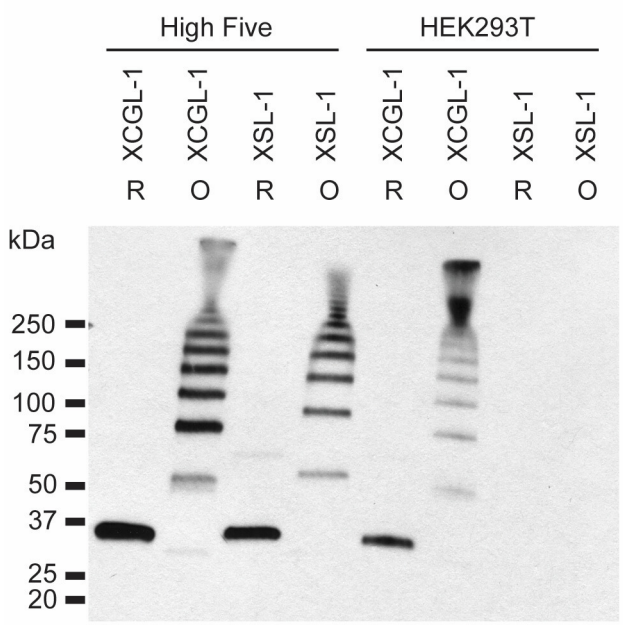



**Figure 4.2**: Expression of XCGL-1 and XSL-1 in High Five insect cells and HEK293T cells. R = reducing, O = oxidizing (non-reducing).

XCGL-1 expressed in High Five cells was further tested for binding to various resins. Galactose-Sepharose 6B was selected because most intelectins to date were discovered using galactose-functionalized resin[113, 116]. Sorbitol-Sepharose 6B was included as hIntL-1 and XEEL were shown to bind exocyclic vicinal diols. XCGL-1 binds galactose and sorbitol functionalized column, but not the resin itself (Figure 4.3). Because the protein was eluted with a quarter of the volume of the input protein, enrichment was expected and this was indeed observed after purification. These results suggest that galactose and sorbitol resin could be used to purify XCGL-1. Following larger scale purification on galactose resin, purified XCGL-1 was subjected to 5 cycles of Edman degradation to determine its N-terminal sequence. The first degradation cycle did not reveal a product, but the subsequent cycles revealed EPVV as the sequence. These results are consistent with CEPVV as the N-terminal sequence. Edman

degradation product of cysteine is known to be unstable[210], thus giving rise to the blank result in the first sequencing cycle.

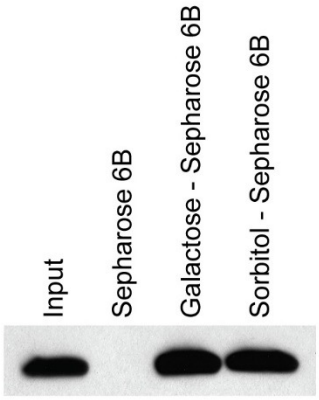



**Figure 4.3**: EDTA elution fraction of XCGL-1 trial purification using galactose- and sorbitol-Sepharose 6B. Protein input and Sepharose 6B were included as controls.

### 4.4.2 mIntL-1 and mIntL-2

Expression of mIntL-1 is detectable in High Five cells, but an insignificant amount was observed in HEK293T cells. In contrast, mIntL-2 expresses well in both High Five and HEK293T cells. Both mIntL-1 and mIntL-2 travel as monomers in non-reducing SDS-PAGE suggesting that if they are oligomeric in solution, they are not disulfide-linked.

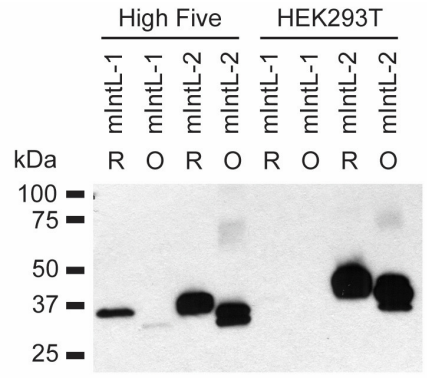



**Figure 4.4**: Expression of mIntL-1 and mIntL-2 in High Five insect cells and HEK293T cells. R = reducing, O = oxidizing (non-reducing).

A Strep-tag II was inserted two amino acids after the predicted signal peptide in mItnL-1, and the protein was purified in a similar fashion as hIntL-1. Because all the residues in the binding site of mIntL-1 and hIntL-1 are identical, mIntL-1 is expected to have the same ligand specificity as hIntL-1. Indeed, of the limited saccharides examined for mIntL-1 binding by SPR (Figure 4.5), only β-Gal*f* showed significant interaction. At present, the low protein yield of mIntL-1 precludes more extensive analysis.

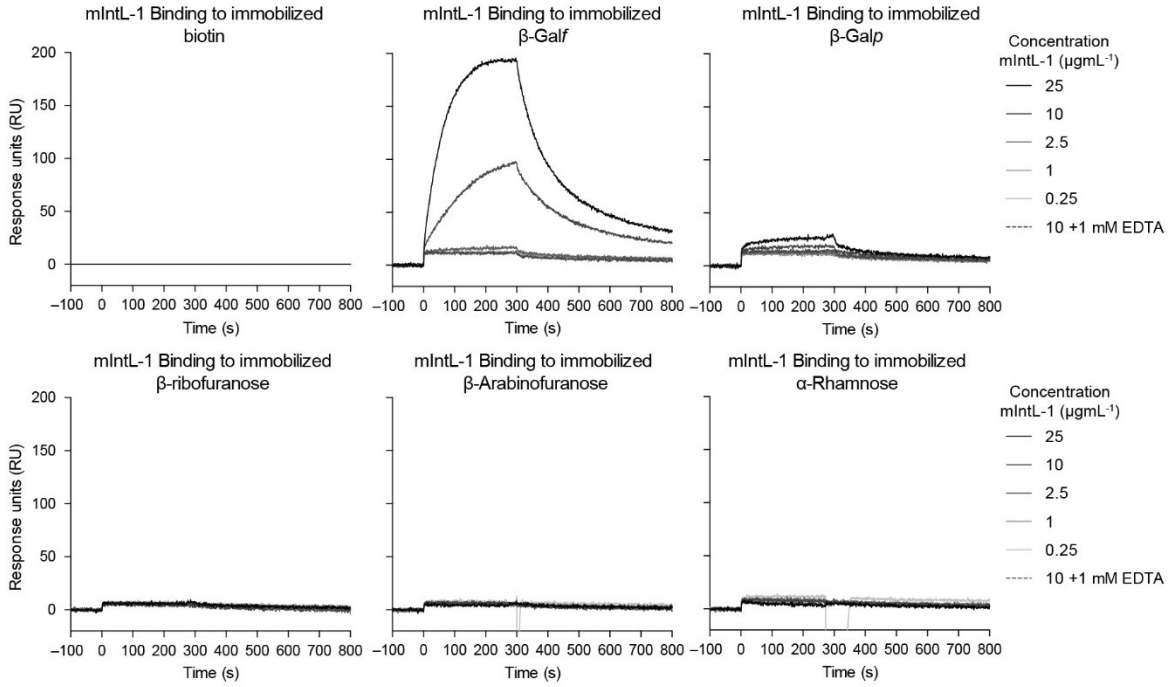



**Figure 4.5**: SPR analysis of mIntL-1 binding to surfaces with immobilized monosaccharide. SPR analysis was performed by Darryl A. Wesener.

As mentioned above the XEEL W317 equivalent in mIntL-2 is an alanine. Thus one might hypothesize that such a drastic substitution would likely perturb ligand specificity. To purify mIntL-2 for further investigation, binding of mIntL-2 to various resins was examined (Figure 4.6). Surprisingly, mIntL-2 could be purified using galactose, sorbitol, and Gal*f*-functionalized resin. Larger scale purification was

performed with Gal*f* agarose. Edman degradation of the purified mIntL-2 revealed AEENL as the first five amino acids. Knowledge of the signal peptide cleavage site enables insertion of the Strep-tag II for further purification and detection of mIntL-2 in various binding assays. Insertion of the Strep-tag II does not alter the expression levels of mIntL-2 (Figure 4.7).

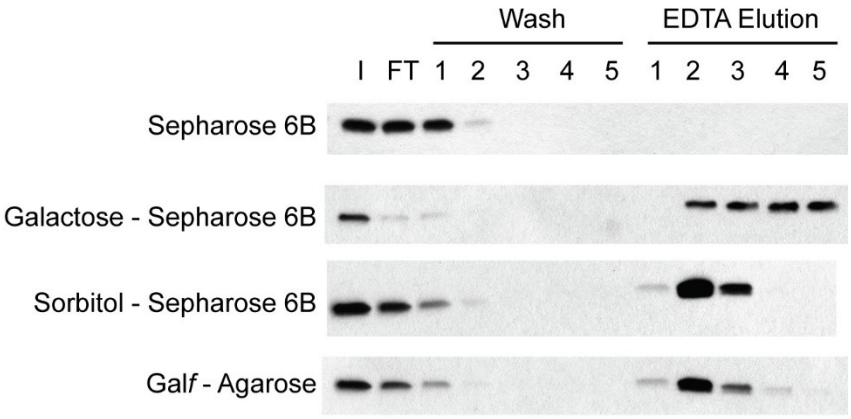



**Figure 4.6**: Fractions of mIntL-2 trial purification using galactose- and sorbitol-Sepharose 6B, and Gal*f*-agarose. I = input control, FT = column flow-through.

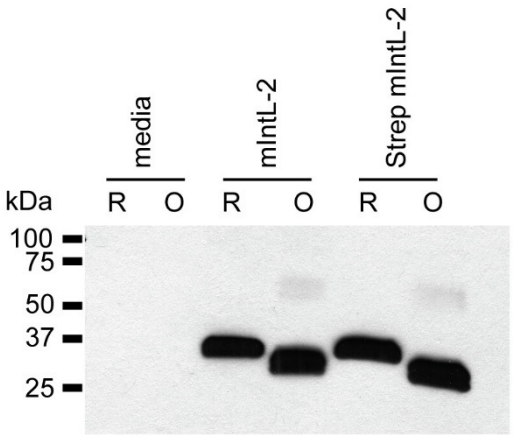



**Figure 4.7**: Expression of mIntL-2 and Strep-mIntL-2 in High Five insect cells. R = reducing, O = oxidizing (non-reducing).

### 4.4.3 hIntL-2

Expression of hIntL-2 is observed in both High Five and HEK293T cells (Figure 4.8). However, the expression levels is less in HEK293T. Disulfide-linked trimer was observed when hIntL-2 was expressed in High Five insect cells. The disulfide-linked trimeric state of hIntL-2 is expected because the responsible cysteines are conserved between hIntL-1 and hIntL-2. Signal peptide cleavage site predicted by Phobius[155] is …TSGCSA/AAASSLEMLSR… (cleavage site indicated by /). However, insertion of the Strep-tag II two residues after the cleavage site results in reduced expression.

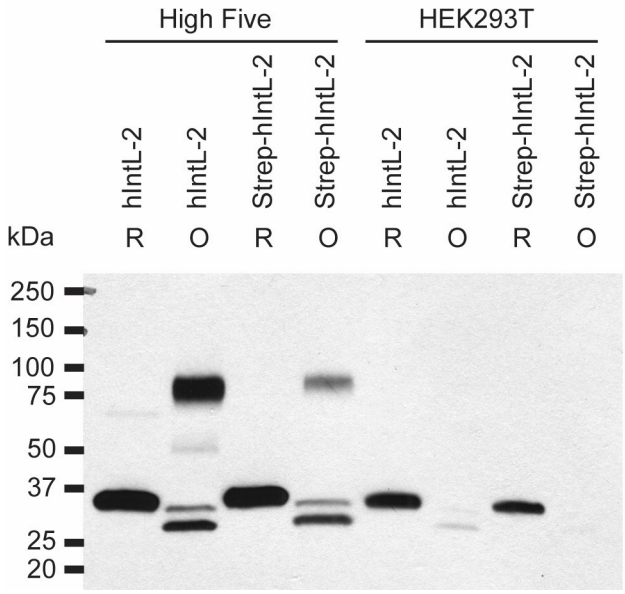



**Figure 4.8**: Expression of hIntL-2 and Strep-hIntL-2 in High Five insect cells and HEK293T cells. R = reducing, O = oxidizing (non-reducing).

To determine whether reduced expression is due to suboptimal tag insertion site, it is necessary to determine the signal peptide cleavage site. Therefore, trial purification to find suitable resin was performed (Figure 4.9). Binding to galactose and sorbitol resins was observed. However, EDTA elution with 25% of the input volume did not enrich for hIntL-2; instead less protein was obtained. Under the conditions tested, these

results suggest that these two carbohydrates are not likely ligands for calcium-bound hIntL-2. As mentioned above, there is a possibility that hIntL-2 does not use a calcium ion in the ligand binding site. Thus, more metal ions need to be tested. Larger scale purification yielded enough hIntL-2 for Edman degradation. The N-terminal sequence was SSLEM. Thus, the experimentally determine signal peptide cleavage site is three amino acids further into the protein than the predicted site. These results should guide further construct optimization.

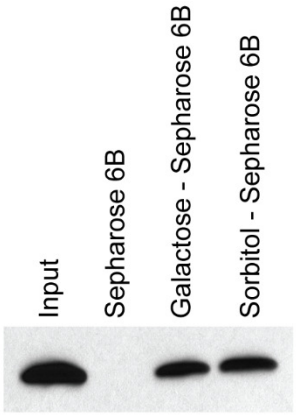



**Figure 4.9**: EDTA elution fraction of hIntL-2 trial purification using galactose- and sorbitol-Sepharose 6B. Protein input and Sepharose 6B were included as controls.

### 4.4.4 DrIntL-1 and DrIntL-2

Robust expression of DrIntL-1 and DrIntL-2 in High Five cells was observed (Figure 4.10). However, these protein do not express well in HEK293T cells. DrIntL-1 appears predominantly as a monomer under both reducing and non-reducing SDS-PAGE. However, most of DrIntL-2 is a disulfide-linked dimer; although some higher disulfide-linked oligomeric states may exist as well.

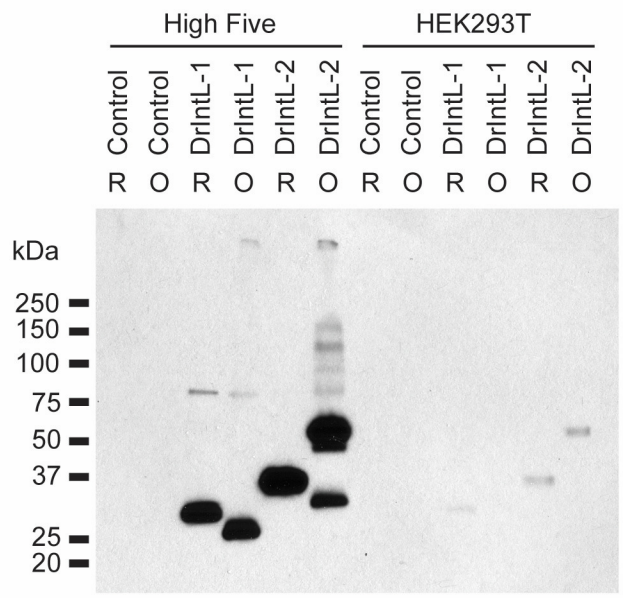



**Figure 4.10**: Expression of DrIntL-1 and DrIntL-2 in High Five insect cells and HEK293T cells. R = reducing, O = oxidizing (non-reducing).

Because disulfide-linked dimeric intelectin has not been previously reported, the cysteines responsible for forming the intermolecular disulfide bond in DrIntL-2 were investigated. Two cysteines are present near the N terminus: C25 and C43. The sequence in this region is …LFSVTLS|LWF**C**EA/KSMGAMYVIQETETP**C**INI…, where the two cysteines are in bold. The predicted signal peptide cleavage site by SignalP4.1[211] and Phobius are indicated with | and /, respectively. Because C25 falls in between the two contradicting predicted signal peptide cleavage site, it was unclear whether one or both cysteines are involved in the intermolecular disulfide bond. Therefore mutagenesis and Strep-tag II insertion at both sites were performed (Figure 4.11). Mutation of C25A does not affect the oligomeric state whereas C43A renders DrIntL-2 monomeric under non-reducing, denaturing SDS-PAGE. Therefore, the C43A is the only cysteine responsible for forming disulfide-linked dimer in DrIntL-2. Insertion of Step-II tag two residues after the first site predicted by SignalP4.1 results in reduced

protein expression whereas tag insertion after the second site predicted by Phobius retained expression levels. Collectively these results suggest that C25 is a part of the signal peptide and does not participate in disulfide-linked oligomerization. In addition the signal peptide cleavage site predicted by Phobius is likely correct for DrIntL-2.

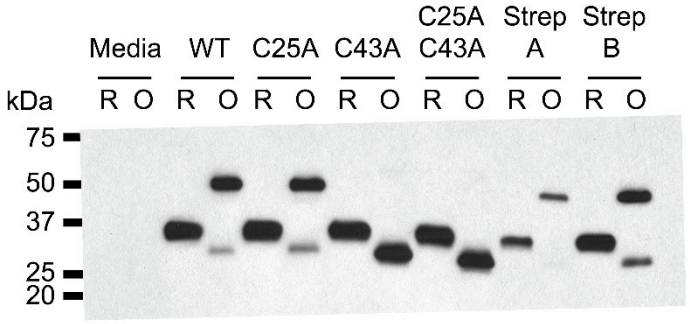



**Figure 4.11**: Mutagenesis of DrIntL-2 to identify the cysteine involved in disulfide-linked dimerization. WT = wide-type, Strep A = Strep-tag II insertion after the signal peptide cleavage site predicted by SignalP 4.1. Strep B = Strep-tag II insertion after the signal peptide cleavage site predicted by Phobius.

**4.5 Conclusions**

Examination of intelectin expression in both insect and human cells reveals dramatic differences in several cases. mIntL-1 does not express well in HEK293T compared to High Five cells. Trimeric hIntL-2 is only observed when expressed in High Five cells. Both DrIntL-1 and DrIntL-2 express better in High Five cells. However, higher expression levels do not necessary indicate that the proteins are properly folded and correctly posttranslationally modified. Previous work with hIntL-1 suggest that although hIntL-1 could be expressed in High Five cells, the protein was more prone to precipitation than hIntL-1 expressed from HEK293T cells. Therefore for a given intelectin, it may be necessary to test both systems for both expression and purification.

Despite limited ligand specificity information, several intelectins binds to galactose and sorbitol resins. These properties could be utilized for purification of intelectins for initial characterizations. This method allows N terminal sequencing of intelectins. The signal peptide cleavage site of intelectins are summarized in Table 4.3. Because N terminal tagging of intelectin will be required for detection and functionalization in multiple assays, signal peptide cleavage sites identified in this chapter will aid construct design and engineering of intelectins.

**Table 4.3**: Summary of signal peptide cleavage site of intelectins (indicated by /)

| Intelectins | Predicted by Phobius | Observed |
|---|---|---|
| XEEL | MLSYSLLLLALAFPAGH**A/G**SCE | Phobius is likely correct. Strep-tag II insertion two amino acids after the predicted site gave the expected Edman degradation results. |
| XCGL-1 | MLVHILLLLVTGGLSQ**S/C**EPV | MLVHILLLLVTGGLSQ**S/C**EPV |
| XSL-1 | MFVHSLVLLSILITVRF**S/A**GCD | Not determined |
| hIntL-1 | MNQLSFLLFLIATTR**G/W**STDEAN | MNQLSFLLFLIATTRGWS**T/D**EAN[113] |
| hIntL-2 | MLSMLRTMTRLCFLLFFSVATSGCS**A/A**AASSLE | MLSMLRTMTRLCFLLFFSVATSGCSAAA**A/S**SLE |
| mIntL-1 | MTQLGFLLFIMVATRGCS**A/A**EEN | Assumed to be the same as mIntL-2 |
| mIntL-2 | MTQLGFLLFIMIATRVCS**A/A**EEN | MTQLGFLLFIMIATRVCS**A/A**EEN |
| DrIntL-1 | MMQSAGFLLLCVPLISLLCESIL**S/L**PTG | Not determined |
| DrIntL-2 | MGMAAFLTKMLLGFLFSVTLSLWFCE**A/K**SMG | Not determined, but Phobius is likely correct |

Characterization of intelectins from multiple species can help elucidate intelectin function and evolution. For example, binding site residues in mIntL-1 is identical to hIntL-1. Among saccharides tested by SPR, mIntL-1 has the same ligand specificity as hIntL-1. It remains to be determined whether mIntL-1 and hIntL-1 share the same ligand

specificity using glycan arrays. Sequence variation in the ligand binding site of mIntL-2 suggest that it may have a unique ligand specificity. Ligand and metal ion specificity of hIntL-2 has also yet to be explored. Preliminary characterization of DrIntL-2 revealed a unique disulfide-linked dimeric state. Given the diversity of intelectin structure and function predicted by sequence alignment, preliminary expression and characterization should enable further studies of intelectin biochemistry. Characterization of ligand specificity of an intelectin coupled with genetic manipulation in a model organism will likely provide conclusive evidence for biological function of intelectins.

# Chapter 5

## Structural characterization of *Klebsiella pneumoniae* and *Corynebacterium diphtheriae* UGM in complex with a triazolothiadiazine inhibitor

Portions of this chapter have been published in:

Kincaid, V. A., London, N., Wangkanont, K., Wesener, D. A., Marcus, S. A., Héroux, A., Nedyalkova, L., Talaat, A. M., Forest, K. T., Shoichet, B. K., Kiessling, L. L. Virtual screening for UDP-galactopyranose mutase ligands indentifies a new class of antimycobacterial agents. ACS Chem Biol. 2015, 10, 2209-2218

## 5.1 Abstract

Galactofuranose residue (Gal*f*) is found in pathogenic microbes and parasites, however the residue is not found in mammalian systems. In addition, the biosynthetic enzyme UDP-galactopyranose mutase (UGM) is essential for these pathogens. Thus, UGM is an attractive antibiotic target. Through virtual screening, a class of triazolothiadiazine inhibitors was identified as a promising UGM inhibitor scaffold. To validate the proposed binding mode for further inhibitor development, X-ray crystallography was employed. Crystals of *Klebsiella pneumoniae* UGM (KpUGM) and *Corynebacterium diphtheriae* (CdUGM) grown with the UGM6.7 inhibitor were examined. The KpUGM crystal showed pseudotranslation coupled with lattice-translocation disorder. After intensity corrections, electron densities that fit UGM6.7 were observed, but could not be correctly modeled. The active site of CdUGM contains UGM6.7 bound in a shallow hydrophobic pocket. The conformation of UGM and the inhibitor is distinct from the structure predicted by virtual screening. These results highlight conformational flexibility of UGM and allow further development of next generation inhibitors.

## 5.2 Introduction

D-galactofuranose (Gal*f*) is an essential cell wall and virulence factor component of several pathogens such as *Mycobacterium tuberculosis*[212, 213], *Klebsiella pneumoniae*[214], *Aspergillus fumigatus*, *Leishmania major*[215], and parasitic nematodes[216-218]. Galf glycoconjugates are synthesized using the sugar donor UDP-galactofuranose (UDP-Gal*f*) synthesized from UDP-galactopyranose (UDP-Gal*p*) by UDP-galactopyranose mutase (UGM)[219-221]. Because UGM is not found in mammalian

systems, the enzyme is an attractive target for developing antibiotics that could be useful for several classes of microbes[222].

Multiple efforts have been made to develop UGM inhibitors[223-234]. Substrate analogs that show inhibitory effects in vitro have been identified[227-234]. Although they function well as mechanistic probes, these UDP analogs cannot cross the cell membrane and, thus show low growth inhibition activity. In addition, these compounds are likely to be non-specific and can potentially inhibit other enzymes utilizing UDP sugars. Non-substrate inhibitors of UGM have been identified using high-throughput screens. However, thiazolidinone[225] hits from the high-throughput screen are prone to conjugate addition in biological systems[235]. Optimizations yielded a 2-aminothiazole inhibitor which had $IC_{50}$ value of 7.2 μM against *K. pneumoniae* UGM (KpUGM) and 37 μM against M. tuberculosis UGM (MtUGM)[224]. However, the inhibitor is poorly soluble and shows toxicity towards mammalian cells[223]. An additional screen for new scaffolds yielded one lead with low activity ($IC_{50}$ > 250 μM) against MtUGM (PubChem AID 504439). Because high-throughput screens only explore limited chemical space, new approaches were needed to obtain potent inhibitors that retained potency in biological systems.

To expand the chemical space and discover novel UGM inhibitor scaffolds, we employed virtual screening with DOCK3.6[236, 237]. After screening a library of 4.6 million commercially available compounds[238, 239] against a structure of KpUGM in the catalytically active, closed conformation, a class of competitive triazolothiadiazine inhibitors was identified (Figure 5.1 A). UGM6 was the initial hit from virtual screen (Figure 5.1 B). The inhibitor was expected to bind in the active site space below the

closed lid and form ionic interactions with R280, previously shown to be essential for catalytic activities[240]. A commercial analog, UGM6.7, was identified to be more potent than UGM6 in KpUGM inhibition studies (Table 5.1). UGM6.7 was also a potent inhibitor against UGM from *Corynebacterium diphtheriae* (CdUGM), MtUGM, and *Caenorhabditis elegans* (CeUGM). Subsequently, other triazolothiadiazine analogs that were active against *M. tuberculosis* in cell culture were identified. These results suggest that virtual screening is a valuable tool for identifying novel inhibitor scaffolds for UGMs.

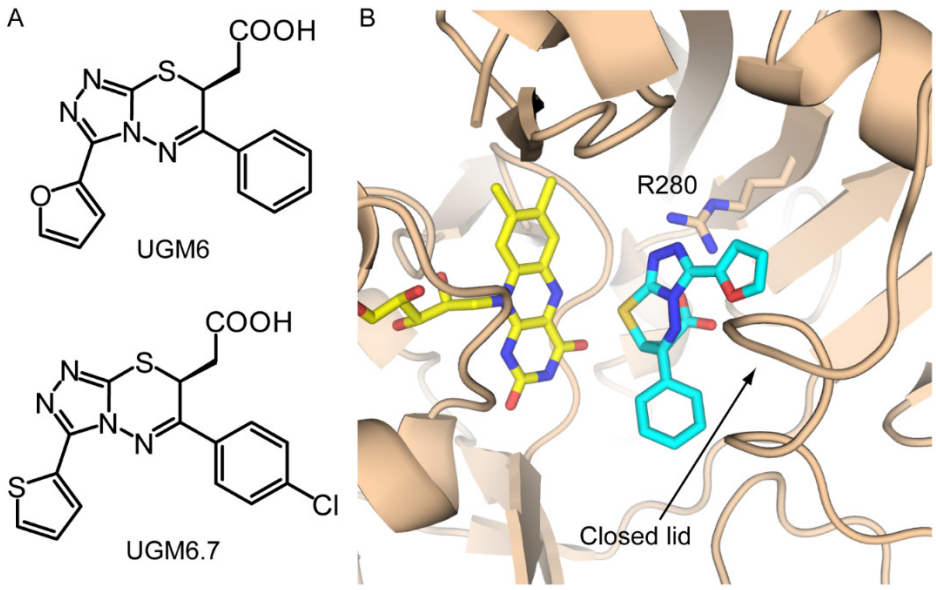



**Figure 5.1**: (A) Chemical structures of UGM6 and UGM6.7. (B) Complex of UGM6 and KpUGM from virtual screening.

**Table 5.1**: Affinity ($K_i$, μM) for competitive inhibition by UGM6 and UGM6.7 towards UGM homologs.

| | UGM6 | UGM6.7 |
|---|---|---|
| KpUGM | 78±31 | 8±3 |
| CdUGM | | 77±37 |
| MtUGM | | 31±18 |
| CeUGM | | 25±8 |

Because no structure of non-substrate inhibitors bound to UGM is available, we envisioned that the new scaffold identified through virtual screening might provide an opportunity to obtain a structure of UGM-inhibitor complex. Not only would such a structure help in understanding the molecular basis of UGM-inhibitor interactions, the structure would also enable rational design of more potent inhibitors. Focusing on UGM6.7, co-crystallization trials were performed against KpUGM and CdUGM. KpUGM crystals exhibited translational non-crystallographic symmetry coupled with lattice-translocation disorder, obscuring model building. After intensity corrections[241], a putative inhibitor location was identified, but the data was not of sufficient quality for accurate modeling. CdUGM crystals do not contain pathologies observed in the KpnUGM crystals. Thus, electron density of UGM6.7 could be identified and modeled. The UGM6.7-bound CdUGM is in an open conformation, in contrast to the KpUGM used in virtual screening. Although UGM6.7 occupies the same general pocket of the active site, the ligand pose is quite different in each case. The only exception is the ionic interactions between the carboxylate group from UGM6.7 and R288 (R280 in KpUGM). These results suggest that conformational dynamics of the enzyme might play an important role in inhibitor affinity. The first structure of nonsubstrate inhibitor bound to UGM presented in this chapter serves as a foundation for designing next generations of potent and selective antibiotics.

## 5.3 Material and Methods

### 5.3.1 Expression, purification, and crystallization of KpUGM with UGM6.7

Sequence encoding residues 1-384 of KpUGM was amplified from a pGEM-T Easy cloning vector[219] and cloned into a modified pMALc5x vector. The resulting

construct contained an N-terminal His tag that is cleavable by TEV protease. The cleavage product should only contain a single extra glycine residue at the N terminus. Expression was performed in *E. coli* ER2523 by induction with 0.1 mM IPTG at 15 °C after the culture reached $OD_{600}$ of 0.6. A previously reported protocol for purification of KpUGM was employed[137]. The protein was dialyzed against 20 mM Tris (pH 7.0) and captured onto a HiTrap Q HP anion exchange column. The protein was eluted with a linear gradient of 0 - 600 mM NaCl. Fractions containing pure KpUGM, as judged by SDS-PAGE, were pooled together and dialyzed against 20 mM Tris (pH 7.0).

KpUGM was concentrated to 10 mg/mL based on the FAD content (absorbance at 450 nm using an extinction coefficient of 11300 $M^{-1}cm^{-1}$) and UGM6.7 was added to 5 mM from a 50 mM stock in DMSO. The final protein concentration was 9 mg/mL. Crystallization was achieved by hanging drop vapor diffusion using 2 µL of protein solution and 2 µL of well solution (100 mM BES (pH 6.5), 200 mM ammonium sulfate, 24% PEG 3350, 15% glycerol). The drop was equilibrated against 1000 µL of well solution at room temperature. Crystals appeared and grew to full size in 2-5 days. The crystals were vitrified without further cryoprotection and stored in liquid nitrogen.

Single crystal diffraction data was acquired using a MAR 300 CCD at beamline 21-ID-G of Advanced Photon Source, Argonne National Laboratory. Precession photography was reconstructed using labelit.precession_photo[242]. Indexing and integration was performed with MOSFLM[243]. AIMLESS[244] and POINTLESS[245, 246] were used for scaling and preliminary space group determination, respectively (Table 5.2). Molecular replacement was carried out with Phaser[140] using a previously reported KpUGM structure[247] (PDB ID 3INT) as a search model. Refinement and model

adjustment were performed using REFMAC5[248] or phenix.refine[141] and Coot[142], respectively. A previously reported demodulation script[241] was used for intensity correction as a result of lattice translocation disorder. The Patterson map was calculated using CCP4[249].

**Table 5.2**: Data collection and refinement statistics for a KpUGM crystal grown with UGM6.7

| Data collection statistics | | |
|---|---|---|
| Wavelength (Å) | 0.97856 | |
| Resolution range (Å)* | 80.04 - 2.50 (2.58 - 2.50) | |
| Space group | P $4_3$ $2_1$ 2 | |
| Unit cell (Å) | 157.7 157.7 114.9 | |
| Total reflections | 1,500,118 (137,679) | |
| Unique reflections | 50674 (4597) | |
| Multiplicity | 29.6 (29.9) | |
| Completeness (%) | 100.0 (100.0) | |
| Mean I/σ(I) | 15.1 (4.5) | |
| Wilson B-factor (Å$^2$) | 34.07 | |
| R-merge | 0.155 (0.867) | |
| R-meas | 0.158 (0.881) | |
| R-pim | 0.029 (0.161) | |
| **Refinement statistics** | **Before intensity correction** | **After intensity correction** |
| Resolution range (Å) | 78.91 – 2.50 (2.55 – 2.50) | 78.91 – 2.5 (2.57 – 2.50) |
| R-factor | 0.2618 (0.3166) | 0.2098 (0.2810) |
| R-free (5%) | 0.3116 (0.3920) | 0.2525 (0.3640) |
| Number of atoms | | |
| Protein | 6246 | 6248 |
| FAD | 106 | 106 |
| Water | 421 | 210 |
| Sulfate | | 10 |
| Protein residues | 766 | 766 |
| RMSD (bonds, Å) | 0.009 | 0.0086 |
| RMSD (angles, °) | 1.179 | 1.248 |
| ML est. coordinate error (Å) | 0.33 | 0.164 |
| Ramachandran favored (%) | 96.84 | 97.8 |
| Ramachandran outliers (%) | 0.53 | 0.0 |
| Average isotropic B-factor (Å$^2$) | | |
| Protein chain A | 36.19 | 41.07 |
| Protein chain B | 34.30 | 38.31 |
| FAD chain A | 36.46 | 38.88 |
| FAD chain B | 33.85 | 35.55 |
| Water | 36.93 | 34.08 |
| Sulfate | | 59.26 |

*Statistics for the highest-resolution shell are shown in parentheses.

### 5.3.2 Expression, purification, and crystallization of CdUGM with UGM6.7

Sequence encoding residues 1-387 of CdUGM was amplified from genomic DNA of *C. diphtheriae* NCTC 13129 (ATCC). The PCR product was cloned into a modified pMALc5x vector, resulting in an N-terminally His-tagged UGM cleavable with a TEV protease. The TEV cleavage product should only leave a single glycine residue at the N terminus. The expression plasmid was transformed into *E. coli* ER2523. After culturing at 37 °C until $OD_{600}$ reached 0.6, the temperature was lowered to 15 °C and protein expression induced with 0.1 mM IPTG. CdUGM was purified using the same protocol as KpUGM described earlier.

Purified CdUGM was dialyzed against 20 mM Tris-HCl (pH 7.0) and concentrated to 10 mg/mL based on the FAD content (absorbance at 450 nm using an extinction coefficient of 11300 $M^{-1}cm^{-1}$). UGM6.7 (50 mM stock in DMSO) was added to the concentrated CdUGM solution to achieve final concentration of 5 mM and 9 mg/mL CdUGM. Crystallization was achieved by hanging drop vapor diffusion using 2 µL of protein solution and 2 µL of well solution (100 mM Bis-Tris (pH 6.5), 2 M ammonium sulfate). The drop was equilibrated against 500 µL of well solution at room temperature. Crystals appeared and grew to full size in 2 weeks. The crystals were cryoprotected by a brief swipe through 100 mM Bis-Tris (pH 6.5), 2.5 M ammonium sulfate, and 15% glycerol, then vitrified and stored under liquid nitrogen. Single crystal diffraction data were collected on a DECTRIS PILATUS 6M detector at beamline X25 of the National Synchrotron Light Source, Brookhaven National Laboratory. Data reduction was performed with HKL2000 (Table 5.3)[138]. Molecular replacement was performed with Phaser[140, 186] using a previously reported E. coli UGM structure[221] (PDB ID 1I8T) as a

search model. FAD and UGM6.7 coordinates and restraints were generated using eLBOW[250]. Model adjustment and refinement were performed in Coot[142] and phenix.refine[141], respectively (Table 5.3). The model was validated using MolProbity[144]. Structure figures were generated with PyMOL[145].

**Table 5.3**: Data collection and refinement statistics for a CdUGM crystal grown with UGM6.7

| Data collection statistics | |
|---|---|
| Wavelength (Å) | 1.1 |
| Resolution range (Å)* | 48.77 – 2.65 (2.74 – 2.65) |
| Space group | P $6_4$ 2 2 |
| Unit cell (Å) | 179.4 179.4 145.2 |
| Total reflections | 391794 |
| Unique reflections | 40032 (3895) |
| Multiplicity | 9.8 (9.7) |
| Completeness (%) | 99.2 (98.6) |
| Mean I/σ(I) | 25.9 (2.1) |
| Wilson B-factor (Å$^2$) | 70.74 |
| R-merge | 0.076 (0.717) |
| R-meas | 0.080 (0.834) |
| R-pim | 0.025 (0.265) |
| **Refinement Statistics** | |
| Resolution range (Å) | 48.77 – 2.65 (2.74 - 2.65) |
| R-factor | 0.1821 (0.3073) |
| R-free (5%) | 0.2269 (0.3908) |
| Number of atoms | |
| Protein | 6365 |
| FAD | 106 |
| UGM6.7 | 50 |
| Water | 79 |
| Protein residues | 772 |
| RMSD (bonds, Å) | 0.009 |
| RMSD (angles, °) | 1.300 |
| ML est. coordinate error (Å) | 0.45 |
| Ramachandran favored (%) | 95.2 |
| Ramachandran outliers (%) | 1.3 |
| Average isotropic B-factor (Å$^2$) | |
| Protein | 76.63 |
| FAD | 84.38 |
| UGM6.7 Chain A | 106.29 |
| UGM6.7 Chain B | 97.15 |
| Solvent | 98.96 |
| UGM6.7 occupancy | |
| Chain A | 0.81 |
| Chain B | 0.87 |

| | |
|---|---|
| UGM6.7 real space correlation | |
| Chain A | 0.791 |
| Chain B | 0.897 |

*Statistics for the highest-resolution shell are shown in parentheses. The coordinate and structure factors have been deposited to the Protein Data Bank under accession code 4XGK

## 5.4 Results and Discussion

### 5.4.1 Pseudotranslation and lattice translocation disorder in KpUGM crystals grown with UGM6.7

Tetragonal KpUGM crystals diffracted to around 2.5 Å. However, the diffraction pattern showed directional streaking. After initial indexing, precession photographs were reconstructed from oscillation photographs to identify the crystal pathologies. Diffraction pattern viewed along c* axis showed normal reflection shape (Figure 5.2 A). However, reflections showed periodic streaking in c* direction with maximum streaking roughly at l = 8n (Figure 5.2 B). Periodic streaking suggests lattice-translocation disorder with directional domain shift along the c direction[251].

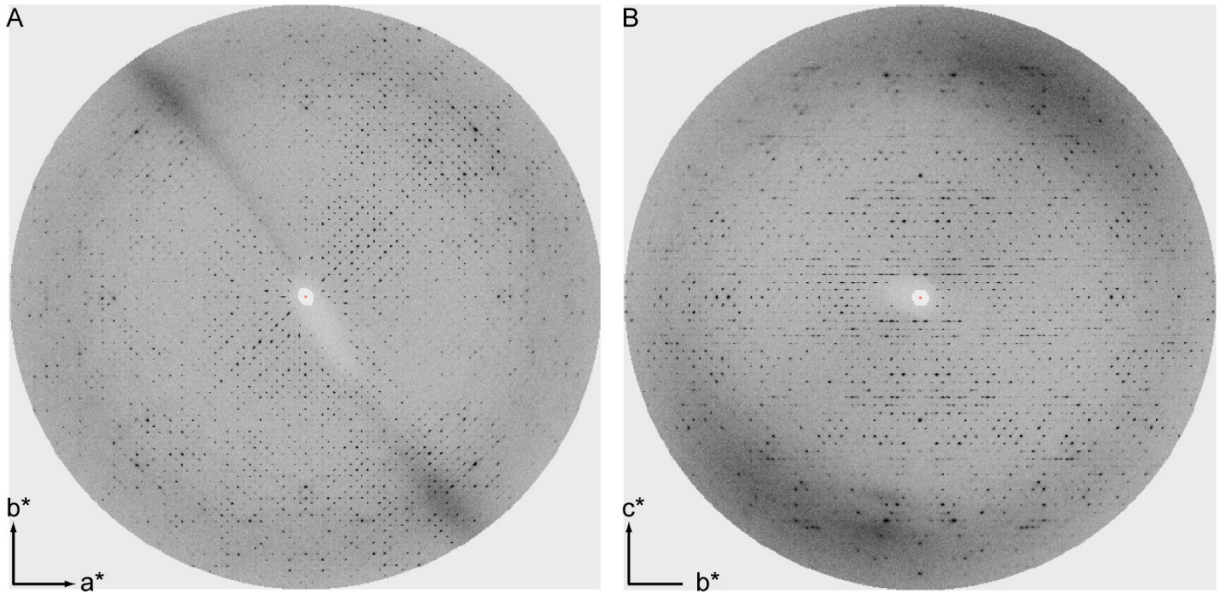



**Figure 5.2**: Reconstructed precession photographs. (A) hk0 plane (B) 0kl plane.

Patterson map calculated from scaled data revealed two sets of significant off-origin peaks (Figure 5.3). The peak closer to origin is weaker (26.5% of origin height) and is 13.6 Å along the w direction. Because the vector length is relatively short compared to the size of a protein, this origin satellite was attributed to lattice-translocation disorder along the c direction in real space[252]. Another set of peaks are stronger (66.4% of origin height) and are a 0.5, 0.5, 0.44 vector (122.5 Å). These strong peaks are a result of pseudotranslation or translational non-crystallographic symmetry (tNCS).

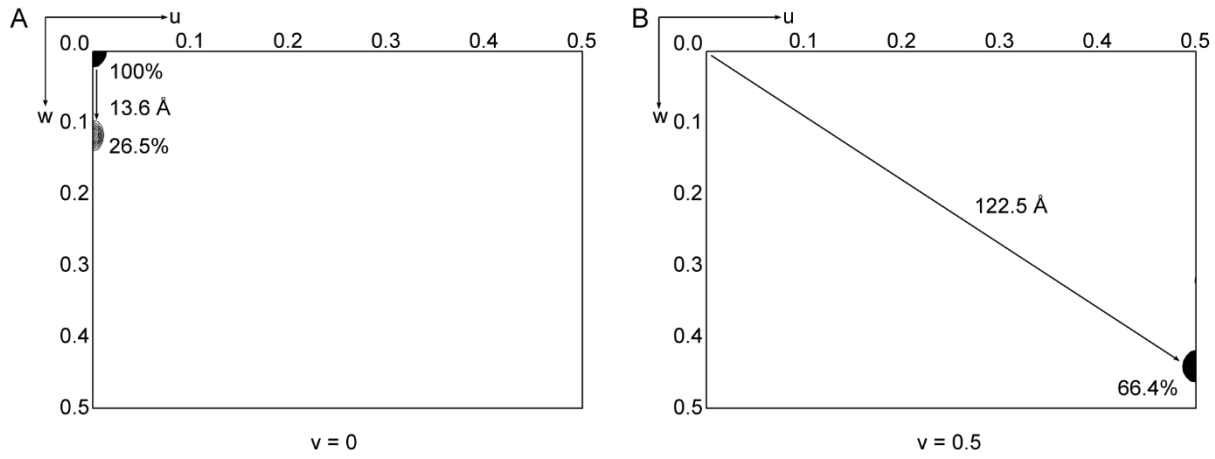



**Figure 5.3**: Native Patterson map of KpUGM plotted with 25-250σ and contoured at 5σ intervals. (A) v = 0 section (B) v = 0.5 section.

Molecular replacement was performed with tNCS taken into account. There are two molecules in the asymmetric unit with solvent content of 70.2% and Matthews coefficient[187] of 4.13. Although the two molecules are related by tNCS, they also correspond to the biological dimer related by a two-fold non-crystallographic symmetry (NCS) (Figure 5.4 A). The refined structure shows significant $mF_o$-$DF_c$ map peaks with no recognizable shapes throughout the unit cells (Figure 5.4 B). These peaks are a result of lattice-translocation disorder. The peaks in the active site cavity obscure

interpretation of the map, thus it was uncertain whether UGM6.7 was present in the active site. Therefore, intensity correction to remove the effect of lattice-translocation disorder needed to be applied to allow map interpretation.

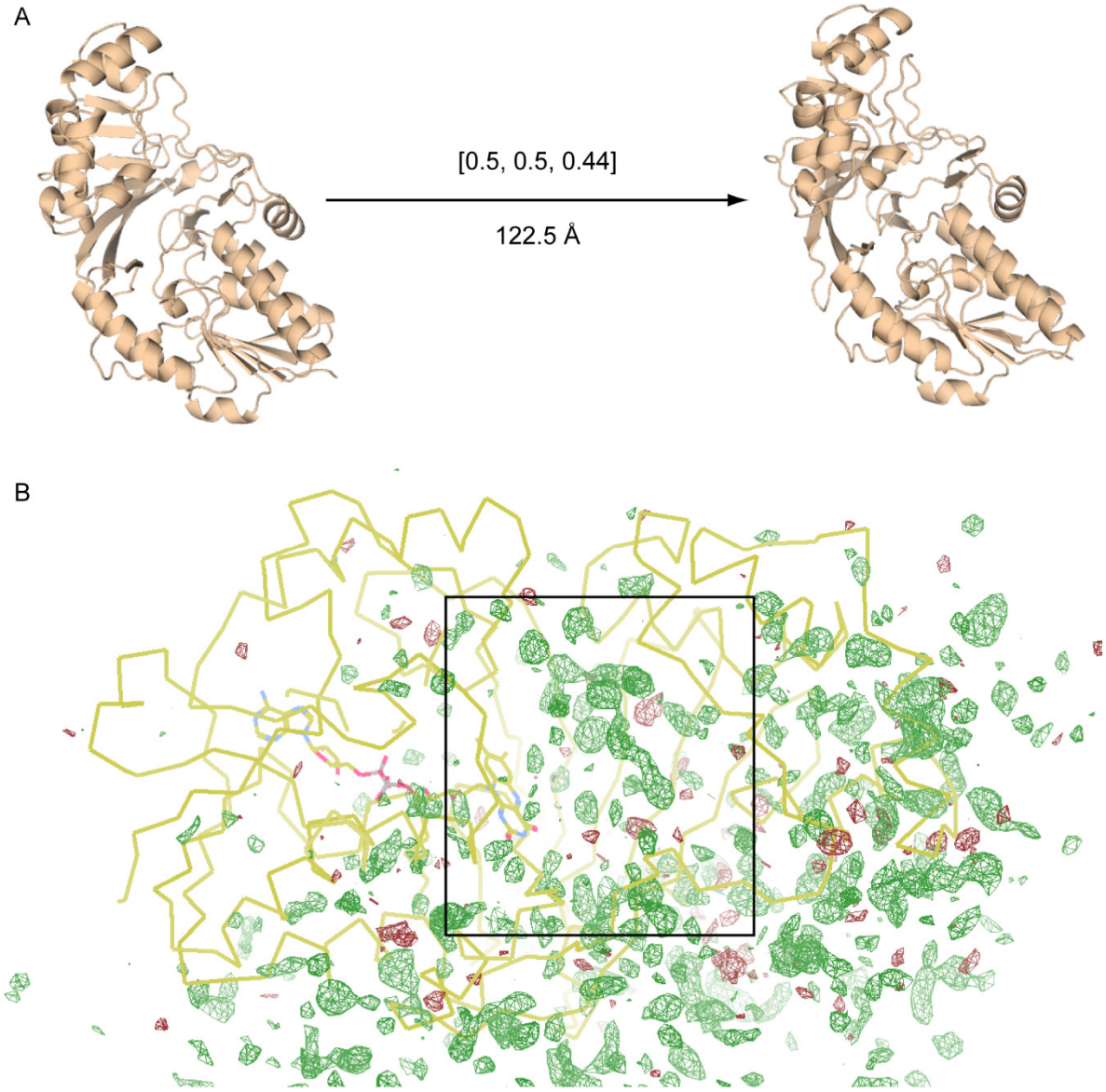



**Figure 5.4**: (A) Molecular replacement solution of KpUGM featuring tNCS between each subunit of the biological dimer. (B) Refined structure of KpUGM. The $mF_o$-$DF_c$ map at 3 σ shows significant peaks (green = positive, red = negative) throughout the unit cell. The active site cavity (shown in box) also contains numerous peaks obscuring map interpretation.

Intensity correction based on a two-lattice model has been developed and previously applied successfully[252, 253]. However, lattice-translocation disorder was the sole problem, with no coupling with tNCS, in these previously reported cases. Demodulation of intensities a by grouping and scaling (DIGS) algorithm was previously shown to be effective when both tNCS and lattice translocation disorder are present[241]. This protocol relies on an accurate reference model for recalculation and scaling of structure factors. Using the refined KpUGM as a reference model, DIGS was performed to remove spurious peaks from the electron density map. The resulting $mF_o$-$DF_c$ map has fewer strong peaks. The active site still contained a set of peaks that could correspond to UGM6.7 (Figure 5.5 A). However, the low quality the map did not allow unambiguous fitting of UGM6.7, which could be due to low ligand occupancy. Interestingly, KpUGM was crystallized in an open conformation, not the closed conformation used for virtual screening. The crystallized enzyme conformation could influence the binding affinity of the ligand. To determine whether the DIGS algorithm produced a reliable map, other peaks in the map were examined as independent evidence. In many regions where no positive difference peaks were previously observed, DIGS produced peaks that could correspond to water with the expected peak shape and hydrogen bond lengths (Figure 5.5 B). These results suggested that DIGS corrected the intensities appropriately.

With low quality of data and several crystal pathologies, another crystal form of KpUGM is desirable. However, additive screens yielded the same crystal form. Additionally, mutagenesis of KpUGM at the packing interface did not produce crystals,

even when a new crystallization screen was performed. Therefore, CdUGM was employed in further structural characterization as an alternative homolog.

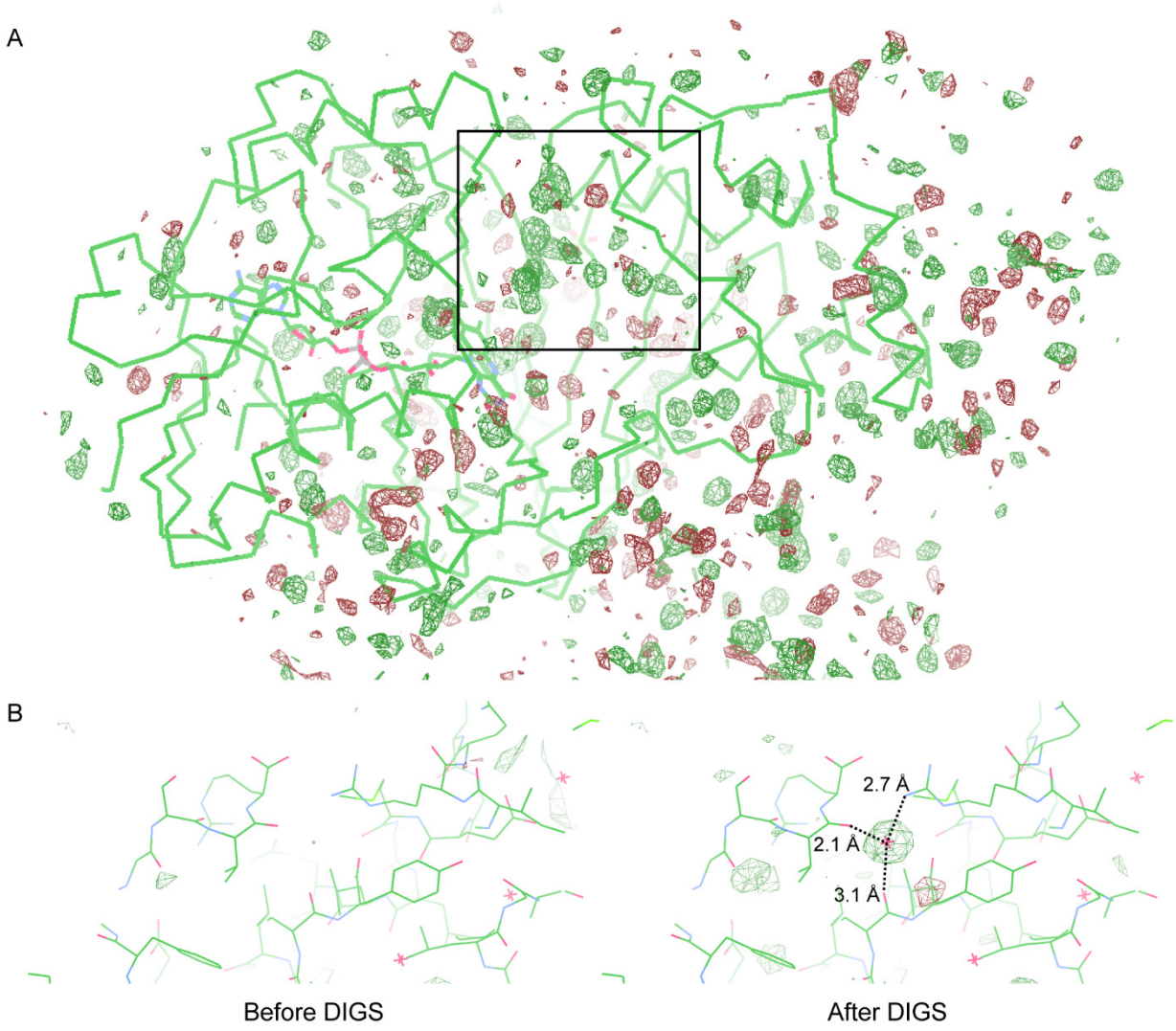



**Figure 5.5**: (A) The $mF_o$-$DF_c$ map (3 σ) after DIGS. The active site (box) retained a set of strong peaks that could correspond to UGM6.7. (B) Appearance of a water peak with appropriate shape and hydrogen bond lengths after DIGS suggests the algorithm was properly applied.

### 5.4.2 Structure of UGM6.7-bound CdUGM

Unlike KpUGM, CdUGM diffraction and data collection statistics are normal, with no indication of twinning. Molecular replacement yielded two molecules in the

asymmetric unit with solvent content of 67.3% and Matthews coefficient[187] of 3.76. The asymmetric unit corresponds to the biological dimer of CdUGM (Figure 5.6). Both monomers are in the open conformation and each monomer contains one bound UGM6.7.

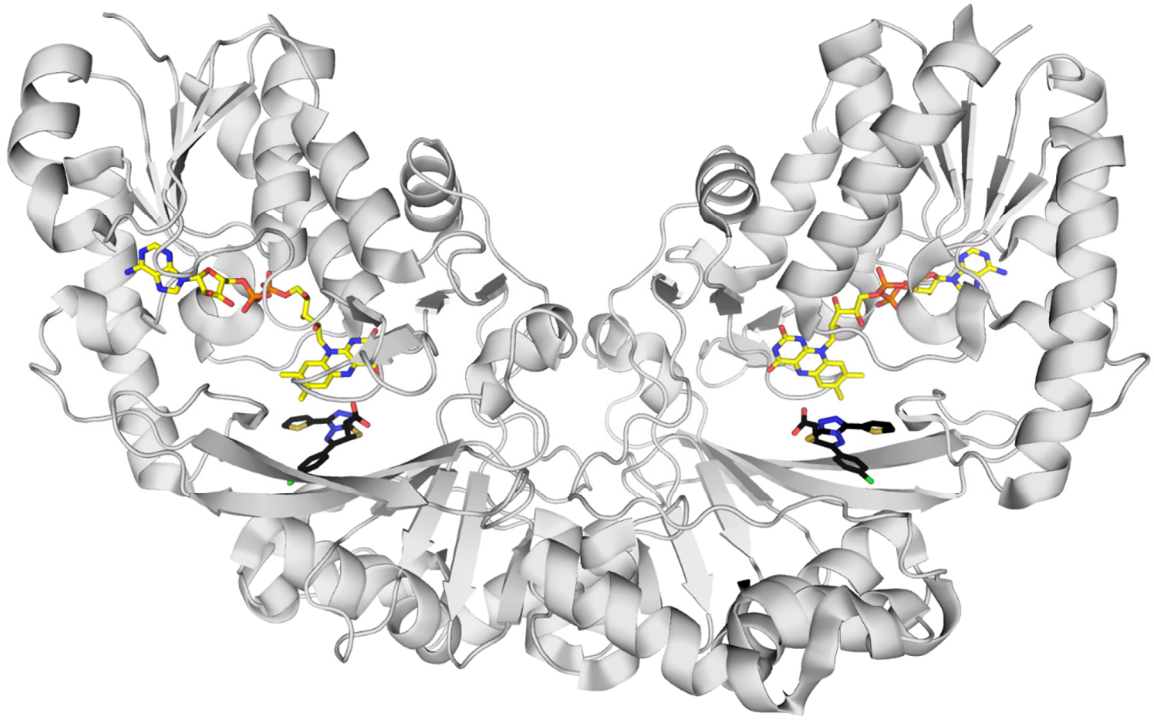


**Figure 5.6**: Structure of dimeric CdUGM. FADs are shown in yellow and UGM6.7 in black.

In chain A, UGM6.7 occupancy was refined to 0.81 with real space correlation of 0.791. The presence of an out-of-plane, forked density served as an anchor point to model the carboxylate group of UGM6.7 (Figure 5.7). In chain B, UGM6.7 occupancy was refined to 0.87 with real space correlation of 0.897. The density of the carboxylate in this chain is not clear, but the ring density is more defined. Conformations of UGM6.7 in both chains are similar. The triazolothiadiazine ring stacks onto Y364. The

chlorophenyl ring stacks onto Y326 with chlorine occupying a pocket formed by W162, S315, and methylenes of N161 and R257 side chains. The thiophene ring occupies an open space above the active site. The carboxylate group form ionic interactions with R288 in chain A, but the interaction is not obvious in chain B.

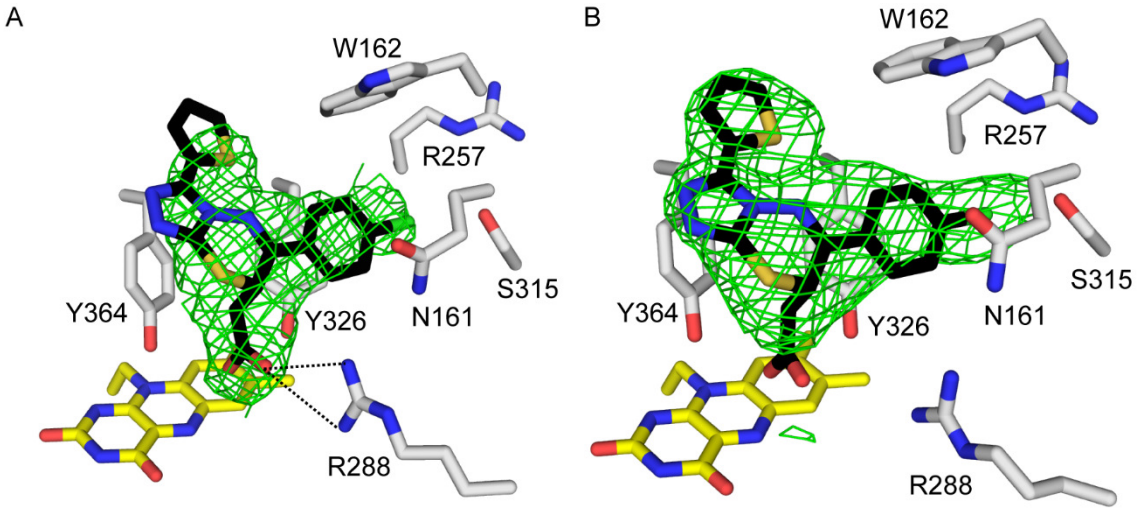



**Figure 5.7**: Structure of UGM6.7 bound in the active site of CdUGM (A) chain A. (B) chain B. The meshes represent simulated annealing $mF_o$-$DF_c$ map with ligands removed (2 σ).

**5.5 Discussion and Conclusions**

Although UGM inhibitors generally have higher affinity towards KpUGM, no structure of inhibitor-bound KpUGM has been obtained thus far. In our case, lattice-translocation disorder obscured map interpretation. After intensity corrections, electron densities that could correspond to UGM6.7 were identified, but was not of high enough quality to allow accurate model building. CdUGM was found to be a suitable homolog that yielded a structure of the complex. The inhibitor densities found in CdUGM appeared in the same location of the active site as in the demodulated KpUGM map. These results suggest that demodulation was mostly successful, but current technologies do not yet allow complete treatment of such data.

The CdUGM structure obtained is in an open conformation, whereas the KpUGM structure used in virtual screening was in a closed conformation[247]. The conformation and pose of the inhibitor was also significantly different (Figure 5.8 A). The carboxylate group of the docked ligand is in an endo conformation, while the crystal structure shows an exo conformation. The rings of the inhibitor occupy distinct spaces of the active site in each case. The only common feature is the ionic bond between the carboxylate and R288. Because UGM6.7 has higher affinity towards KpUGM, these structural discrepancies could explain the affinity difference. Superposition of the closed KpUGM structure reveals that UGM could also accommodate the inhibitor in the closed conformation (Figure 5.8 B). Even though there is a steric clash between a lid arginine, the interactions could be relieved by flexibility of the side chain. Therefore, it is possible that UGM6.7 binds KpUGM in a closed conformation with higher affinity than other UGMs. This hypothesis is further supported by the obscured ligand density in the structure of opened KpUGM. Because UGM have been crystallized in both opened and closed forms[233], it is possible that the free energy difference between the opened and closed forms is low. Indeed, positive peaks above the inhibitor in the CdUGM structure could indicate a residual closed lid conformation (Figure 5.9). Therefore, both the docked KpUGM structure and the CdUGM crystal structure could be relevant. Even though several nonsubstrate UGM inhibitors have been reported[224-226, 254, 255], no structure of a complex has been reported until this work. Thus, the structures presented could provide starting points for inhibitor optimization.

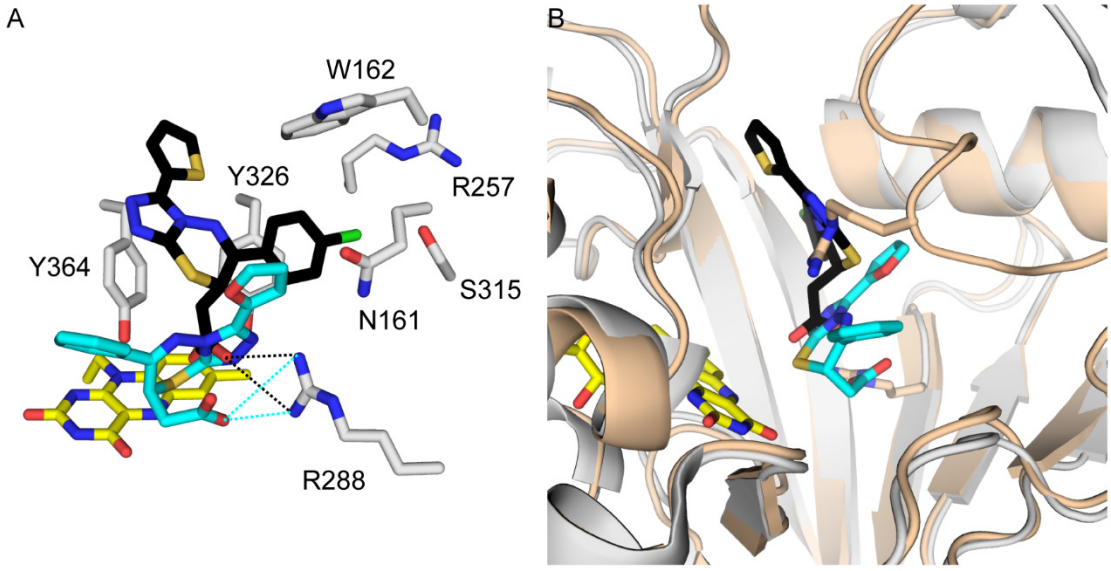



**Figure 5.8**: (A) Overlay of UGM6 (cyan) from the docked KpUGM structure and UGM6.7 (black) in the active site of CdUGM. (B) Alignment of KpUGM (wheat) and CdUGM (grey) structures with ligands bound.

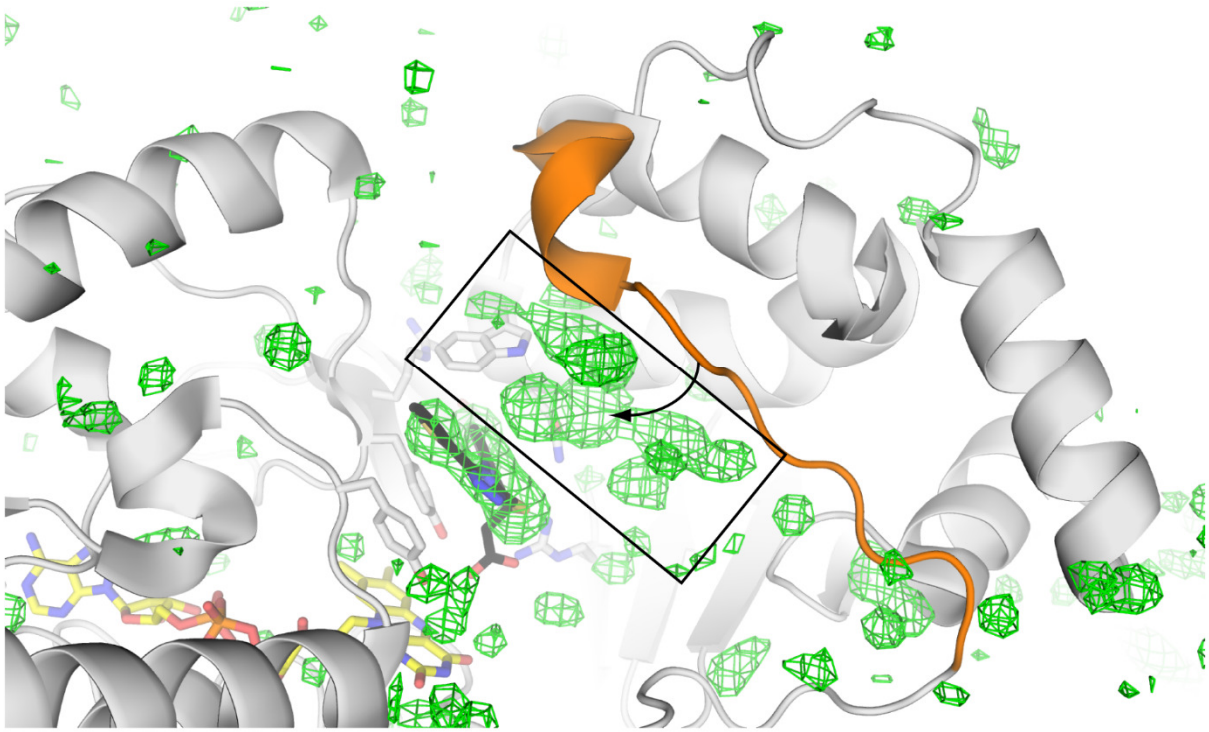


**Figure 5.9**: Simulated annealing $mF_o$-$DF_c$ map (3 σ) with ligand removed. The box indicated positive peaks that could correspond to partial closing of the lid (orange).

# Chapter 6

## Conformational states of *Corynebacterium diphtheriae* UGM and potential implications on affinity of inhibitors

## 6.1 Abstract

UDP-galactopyranose mutase (UGM) is a key biosynthetic enzyme for generating galactofuranose (Galf)-containing glycoconjugates. Because UGM is an essential protein in various pathogens and is not found in mammalian systems, UGM is an ideal antibiotic target. We recently reported a structure of *Corynebacterium diphtheriae* UGM (CdUGM) in complex with a triazolothiadiazine inhibitor. The structure suggests an alternative conformation of the protein-inhibitor complex. Therefore, we screened additional protein constructs (CdUGM-His6 and GSG-CdUGM) and inhibitors to reveal the inhibitor complex structure in greater detail. Two additional crystals forms were identified. Tetragonal CdUGM-His6 (1.95 Å) was crystallized in an opened conformation and contains a sodium and a citrate ion from buffer bound in the active site. Monoclinic GSG-CdUGM (2.15 Å) was obtained in a closed conformation and contains a UDP in each active site. Although these structures do not contain inhibitors, the tetragonal CdUGM-His6 has vastly improved structure geometry and the citrate binding site could be exploited for inhibitor design. The monoclinic GSG-CdUGM provides additional insight into the active site conformation and helps explain inhibitor affinity. Together with the previously reported inhibitor-bound CdUGM structure, these additional CdUGM structures will help guide design of next generation inhibitors.

## 6.2 Introduction

UDP-galactoyranose mutase (UGM) catalyzes interconversion of UDP-galactopyranose (UDP-Gal*p*) and UDP-galactofuranose (UDP-Gal*f*)[219-221]. The latter serve as an activated sugar donor used by various glycosyltransferases to synthesize Gal*f* glycoconjugates. Gal*f* forms a galactan layer found in Mycobacterial cell wall[212, 213].

Virulence factors of Leishmania contain Gal*f* residues[215]. Nematodes also express UGM[216-218], but Gal*f* glycans in nematodes have yet to be identified. Despite its ubiquity in pathogenic organisms, Gal*f* residues and UGM are not found in mammalian systems[222]. In addition, UGM knock downs or knock outs are often lethal[212, 213]. These observations suggest UGM is an attractive antibiotic target.

Several attempts have been made to discover potent UGM inhibitors. Many characterized inhibitors are charged UDP-Gal*f* analogs which exhibit potency in vitro, but are ineffective in vivo[227-234]. Our group has identified a potent 2-aminothiazole inhibitor[224] (Figure 6.1 A), but the inhibitor suffers from poor solubility and toxicity to mammalian cells[223]. Through virtual screening, we recently discovered a family of triazolothiadiazines (Figure 6.1 B) which have better physical properties and are active against live *Mycobacterium tuberculosis*[256]*.* Therefore, structural insight into how these small molecules inhibit UGMs will guide the design of more effective inhibitors.

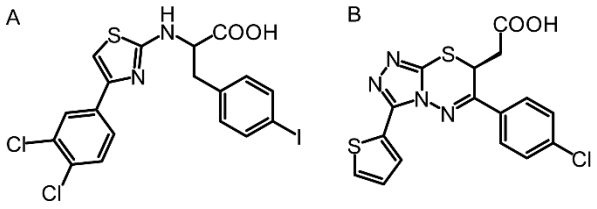



**Figure 6.1**: Structures of (A) 2-aminothiazole and (B) triazolothiadiazine inhibitors.

We have obtained a crystal structure of a triazolothiadiazine inhibitor bound to *Corynebacterium diphtheriae* UGM (CdUGM)[256]. The triazolothiadiazine inhibitor binds in the same general area of the active site as predicted by virtual screening, although the pose is different than the docked pose. Interestingly, the CdUGM is in an opened conformation in contrast to the structure of *Klebsiella pneumoniae* UGM (KpUGM) used

in virtual screening. In addition, there are unexplained electron densities under the opened lid that could close over the active site. These results suggest that there could be multiple conformations of UGM-inhibitor complexes. Furthermore, all known small molecule inhibitors appear to be more potent towards KpUGM compared to other UGMs[224, 256]. For example, the $K_i$ values of the triazolothiadiazine inhibitor against KpUGM, CdUGM and *Mycobacterium tuberculosis* (MtUGM), are 8±3, 77±37, and 31±18μM, respectively. However, it is not obvious from the triazolothiadiazine inhibitor-bound CdUGM structure why such discrepancies exist. It is possible that the difference in inhibitor affinities could be explained if multiple conformations of CdUGM could be compared.

We screened for additional crystal forms of CdUGM using two different protein constructs (CdUGM-His6 and GSG-CdUGM). In addition to the triazolothiadiazine inhibitor, the 2-aminothiazole inhibitor was also included in the screen because no structural information of the 2-aminothizole inhibitor-bound UGM is available. If successful, fragments of the 2-aminothiazole inhibitor could later be combined with the triazolothiadiazine inhibitor to generate a next generation of inhibitors. Co-crystallization of CdUGM with UDP-Gal*p* was also performed to explore the possibility of exchanging UDP-Gal*p* with inhibitors by soaking. This effort yielded two new crystal forms of CdUGM.

Co-crystallization of the 2-aminothizole inhibitor with CdUGM-His6 yielded a tetragonal CdUGM. Instead of the inhibitor bound in the active site, a citrate ion from the crystallization buffer was observed. A sodium ion was also identified in the active site. This structure suggest that the pocket occupied by the citrate ion could be further

exploited for inhibitor design. The monoclinic GSG-CdUGM was in a closed conformation, explaining the inability to exchange the ligand, with UDP from hydrolyzed UDP-Gal*p* bound. Alignment of the closed GSG-CdUGM with the opened structure of G-CdUGM with the triazolothiadiazine inhibitor bound reveals a residue movement not observed in KpUGM. The residue movement upon lid closure could generate steric clashes that may explain the differential affinity of the triazolothiadiazine inhibitor toward UGMs. The two CdUGMs in different conformations reported in this work will not only aid in designing more potent inhibitors, but also shed light on structural dynamics of UGM which could be crucial in inhibition strategies.

### 6.3 Material and Methods

Both CdUGM constructs were purified using a previously reported protocol[137, 256]. HKL2000 was used for data reduction[138]. Molecular replacement was performed with Phaser[140]. Restraints for small molecule ligands were generated using eLBOW[250]. Model adjustment and refinement were performed with Coot and phenix.refine, respectively[141, 142]. The model was validated using MolProbity[144]. Figures were generated with PyMOL[145].

#### 6.3.1 CdUGM-His6 in complex with sodium citrate

Sequence encoding residues 1-387 of CdUGM was amplified using primers: CGAGCAATTGACCAACAAGGACCATAGATTATGTCTGACTTTGATCTGATCGTGGT AGGT and ATTCGAGCTCTCATTAATGGTGATGGTGGTGATGTTTCAGGGCGTCGACAAGCTTG TTAT. The PCR product was then cloned into the MfeI and SalI sites of pMALc5x. The resulting construct coded for CdUGM with a C-terminal His6 tag without any linker

(CdUGM-His6). Protein was dialyzed against 20 mM Tris pH 7.0 and concentrated to 10 mg/mL. A 2-aminothiazole inhibitor[224] was added to a final concentration of 1 mM (from 40 mM stock in isopropanol). Crystals were grown by hanging drop vapor diffusion using 2 µL of protein solution and 2 µL of well solution (100 mM sodium citrate pH 5.6, 15% isopropanol, and 16-18% PEG 5000 MME). Crystals grew to full size within a few days. Crystal were cryoprotected by briefly soaking in well solution containing 20% ethylene glycol before vitrification and storage in liquid nitrogen. Single crystal X-ray diffraction data were collected at beamline 21-ID-F (LS-CAT) at Argonne National Laboratory using a MAR225 CCD detector. Molecular replacement was performed using a previously reported *Klebsiella pneumoniae* UGM (KpUGM) structure (PDB ID 2BI7)[257].

For potassium citrate soaking, crystals were grown as described above. The crystal was then briefly washed and soaked for 24 hours in 100 mM potassium citrate pH 5.6, 15% isopropanol, and 18% PEG 5000 MME. Cryoprotection was achieved as described above, but sodium citrate was substituted with potassium citrate. Single crystal X-ray diffraction data were collected at beamline 14-1 at Stanford Synchrotron Radiation Lightsource using a MAR325 CCD detector. Data collection and refinement statistics are shown in Table 6.1.

**Table 6.1**: Data collection and refinement statistics for CdUGM-His6 in complex with sodium citrate

| Data collection statistics | Sodium Citrate | Potassium Citrate Soak |
|---|---|---|
| Wavelength (Å) | 0.97872 | 1.18080 |
| Resolution range (Å)* | 46.72 - 1.95 (2.02 - 1.95) | 49.01 – 2.35 (2.43 - 2.35) |
| Space group | P $4_1$ $2_1$ 2 | P $4_1$ $2_1$ 2 |
| Unit cell (Å) | 98.3 98.3 126.2 | 98.0 98.0 126.3 |
| Total reflections | 1,075,063 (89,525) | 378,033 (33,333) |
| Unique reflections | 45,492 (4,454) | 26,343 (2,584) |
| Multiplicity | 23.6 (20.1) | 14.4 (12.9) |
| Completeness (%) | 99.9 (99.6) | 100.0 (100.0) |
| Mean I/σ(I) | 36.3 (3.6) | 24.0 (3.0) |
| Wilson B-factor (Å$^2$) | 29.30 | 37.05 |
| R-merge | 0.103 (0.756) | 0.127 (0.971) |
| R-meas | 0.105 (0.775) | 0.123 (>1) |
| R-pim | 0.020 (0.164) | 0.035 (0.280) |
| **Refinement Statistics** | | |
| Resolution range (Å) | 46.73 – 1.95 (2.02 – 1.95) | |
| R-factor | 0.1528 (0.1982) | |
| R-free (5%) | 0.1895 (0.2267) | |
| Number of atoms | | |
| Protein | 3302 | |
| FAD | 53 | |
| Citrate | 13 | |
| $Na^+$ | 2 | |
| Water | 458 | |
| Protein residues | 386 | |
| RMSD (bonds, Å) | 0.008 | |
| RMSD (angles, °) | 1.05 | |
| Est. coordinate error (ML, Å) | 0.17 | |
| Ramachandran favored (%) | 98.8 | |
| Ramachandran outliers (%) | 0 | |
| Average Isotropic B-factor (Å$^2$) | | |
| Protein | 31.61 | |
| FAD | 26.51 | |
| Citrate | 32.31 | |
| $Na^+$ | 30.19 | |
| Water | 41.19 | |

*Statistics for the highest-resolution shell are shown in parentheses. The coordinates and structure factors have been deposited to the Protein Data Bank under accession code 5BR7.

### 6.3.2 GSG-CdUGM in complex with UDP

Site directed mutagenesis was performed on a previously reported CdUGM construct (G-CdUGM)[256] to add a two amino acids (SG) after the glycine of TEV

cleavage site for more efficient cleavage (GSG-CdUGM). The primers used were AGCGAAAACCTGTATTTTCAGGGTAGCGGCATGTCTGACTTTGATCTGATCGTGGTAG and CTACCACGATCAGATCAAAGTCAGACATGCCGCTACCCTGAAAATACAGGTTTTCGCT. Protein was dialyzed against 20 mM Tris pH 7.0 and concentrated to 10 mg/mL. UDP-galactopyranose (UDP-Gal*p*) was added to 10 mM (from 100 mM stock in water). Crystals were grown by hanging drop vapor diffusion using 2 µL of protein solution and 2 µL of well solution (100 mM Bis-Tris pH 5.5, 200 mM lithium sulfate, and 21% PEG 3350). Crystal grow to full size in 1-2 weeks. Cryoprotection was achieved by brief soaking in well solution with PEG 3350 supplemented to 35%. Crystals were then vitrified and stored in liquid nitrogen. Single crystal X-ray diffraction data were collected at beamline 21-ID-G (LS-CAT) at Argonne National Laboratory using a MAR300 CCD detector. Molecular replacement was performed using a previously reported *Mycobacterium smegmatis* UGM (MsUGM) structure (PDB ID 5EQD). Data collection and refinement statistics are shown in Table 6.2.

**Table 6.2**: Data collection and refinement statistics for GSG-CdUGM in complex with UDP

| Data collection statistics | |
|---|---|
| Wavelength (Å) | 0.97856 |
| Resolution range (Å)* | 29.1 - 2.15 (2.23 – 2.15) |
| Space group | C2 |
| Unit cell | 115.6 82.7 108.6 90 112.6 90 |
| Total reflections | 376,968 (26,894) |
| Unique reflections | 50,316 (4,637) |
| Multiplicity | 7.5 (5.8) |
| Completeness (%) | 97.5 (90.5) |
| Mean I/sigma(I) | 27.4 (2.8) |
| Wilson B-factor (Å$^2$) | 41.90 |
| R-merge | 0.064 (0.419) |
| R-meas | 0.069 (0.458) |
| R-pim | 0.025 (0.181) |
| **Refinement Statistics** | |
| Resolution range (Å) | 28.25 – 2.15 (2.22 – 2.15) |
| R-factor | 0.1765 (0.2485) |
| R-free (5%) | 0.2281 (0.3009) |
| Number of atoms | |
| protein | 6,344 |
| FAD | 106 |
| UDP | 50 |
| water | 228 |
| Protein residues | 771 |
| RMSD (bonds, Å) | 1.192 |
| RMSD (angles, Å) | 0.008 |
| Est. coordinate error (ML, Å) | 0.25 |
| Ramachandran favored (%) | 98.4 |
| Ramachandran outliers (%) | 0 |
| Average Isotropic B-factor (Å$^2$) | |
| Chain A | |
| protein | 48.14 |
| FAD | 52.05 |
| UDP | 37.35 |
| water | 49.55 |
| Chain B | |
| protein | 69.60 |
| FAD | 85.26 |
| UDP | 56.75 |
| water | 51.71 |

*Statistics for the highest-resolution shell are shown in parentheses. The coordinates and structure factors have been deposited to the Protein Data Bank under accession code 5EQF.

## 6.4 Results and Discussion

### 6.4.1 CdUGM-His6 in complex with sodium citrate

To obtained structures of CdUGM in complex with the 2-aminothiazole inhibitor, crystallization screen was performed with CdUGM-His6 pre-incubated with the inhibitor. Tetragonal CdUGM-His6 crystals were obtained. This crystal form has one molecule in the asymmetric unit (Figure 6.2), corresponding to a monomer of the biological UGM dimer, with solvent content of 64.3% and Matthews coefficient[187] of 3.45. The molecule is in an open conformation (Figure 6.3). A sodium ion is observed at a site remote from the active site surrounded by residues S126-S131 and S143. The sodium ion is assigned based on the bond length of around 2.4 Å and the near perfect octahedral coordination geometry[258]. The $Na^+$-O bond length is restrained to 2.43 Å with tolerance of 0.25 Å. The axial ligands are a water molecule and the backbone carbonyl of S126. The axial water is also hydrogen bonded to the side chain of S131, which in turn is positioned by hydrogen bonding to the side chain of S143. The equatorial ligands are two water molecules and backbone carbonyls of I129 and S143. One equatorial water is hydrogen bonded to the side chain of S126. The other equatorial water hydrogen bond to both the backbone carbonyl of A127 and the side chain of D130.

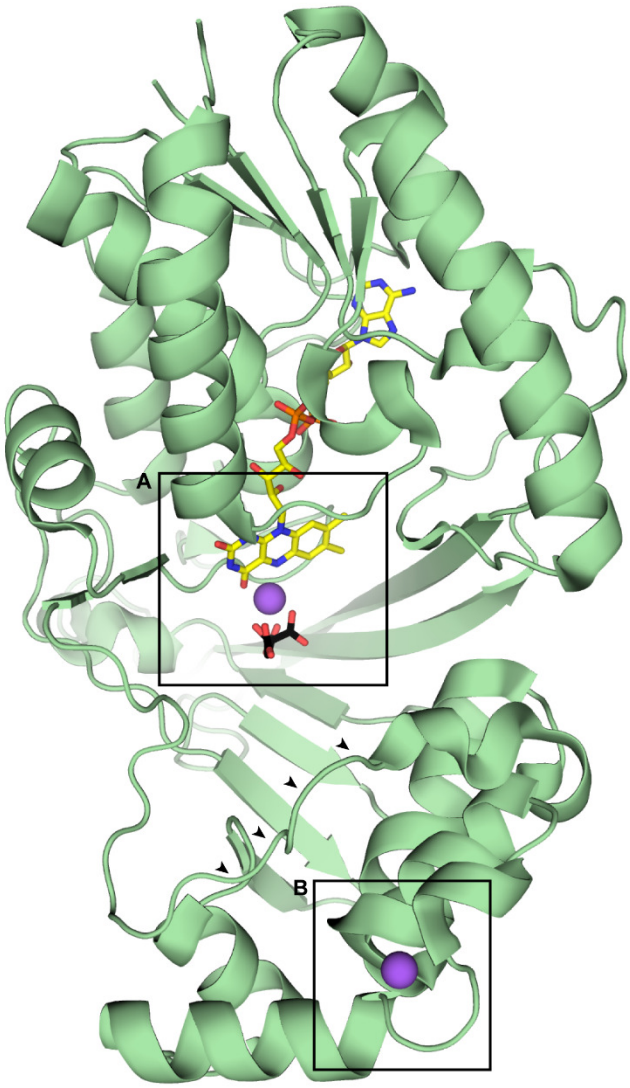



**Figure 6.2**: Overall structure of CdUGM-His6 in complex with sodium citrate. The lid (arrows) is in an opened conformation and does not cover the FAD-containing active site (Box A). Box B indicates a position of a remote sodium ion binding site.

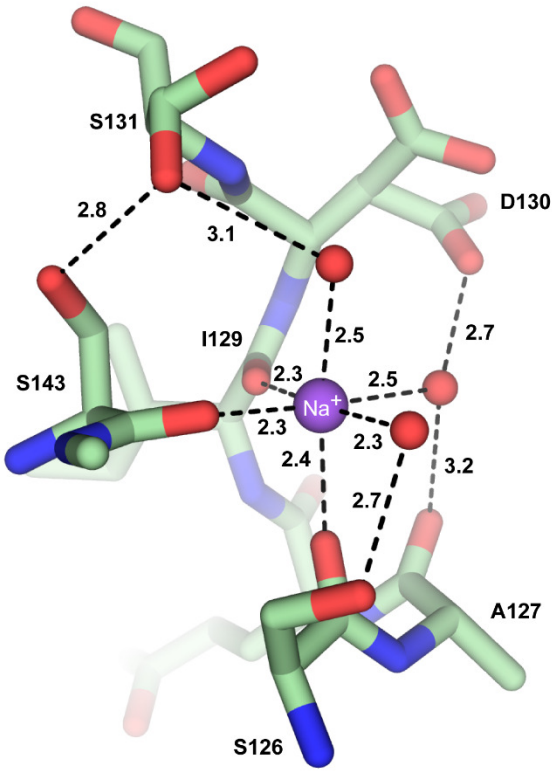



**Figure 6.3**: Structure of the remote sodium ion binding site in CdUGM-His6 with $Na^+$-O bond distances from restrained refinement.

The active site of CdUGM-His6 does not contain a density corresponding to the 2-aminothiazole inhibitor. Instead, citrate ion from the crystallization buffer is observed (Figure 6.4). The citrate ion forms salt bridges with the side chains of R288 and H290. The citrate ion also interacts with numerous ordered water molecules, forming a complex hydrogen bonding network with the active site residues. Given that citrate buffer is ubiquitous in crystallization screens, this UGM-citrate complex is a surprising discovery. When the occupancy of the citrate ion is set to one, residual negative peaks are observed in the $mF_o$-$DF_c$ map. Therefore, the occupancy was refined to 0.92. Because the citrate concentration was 100 mM and citrate does not bind to full occupancy, the dissociation constant ($K_d$) of citrate is estimated to be around 10 mM

using a simple equilibrium model. Interestingly, the estimated $K_d$ is about the same order of magnitude as the intracellular citrate concentration in *Escherichia coli* (roughly 2-20 mM depending on growth media)[259]. It is possible that citrate could be a regulator of UGM activity. However, preliminary experiments on the effect of citrate on in vitro activity of UGM have been inconclusive (data not shown).

In addition to the citrate ion, an initially unidentified peak was observed bridging the citrate ion and the oxidized FAD. Initial refinement with an ordered water molecule did not reveal peaks in the $mF_o$-$DF_c$, suggesting that the scattering factor was roughly correct. Due to its irregular coordination geometry, bond length of roughly 2.4 Å[258], and proximity to negatively charged citrate ion, it is possible that the peak corresponds to a sodium ion. Sodium ions and water have 10 total electrons, thus they can be difficult to distinguish in electron density maps, especially with the data quality commonly encountered in macromolecular crystallography. With limited prior knowledge about coordination of sodium ion between citrate and oxidized FAD, we gathered more experimental evidence for sodium ion identification. No crystallization is observed if potassium citrate buffer is used instead of sodium citrate buffer. Soaking of crystals grown with sodium citrate buffer in potassium citrate buffer results in a loss of sodium and citrate electron density judged by the difference Fourier map (Figure 6.4 C). Therefore, we concluded that sodium ion is coordinated by FAD, citrate ion, and the backbone carbonyl of A60. The structure of CdUGM-His6 in complex with sodium citrate not only features an unusual sodium ion complex, but it also identifies an active site pocket that could be targeted by small molecules. In combination with the structure of

the triazolothidiazine inhibitor already available, analogs that extend into the citrate-binding site could be further explored.

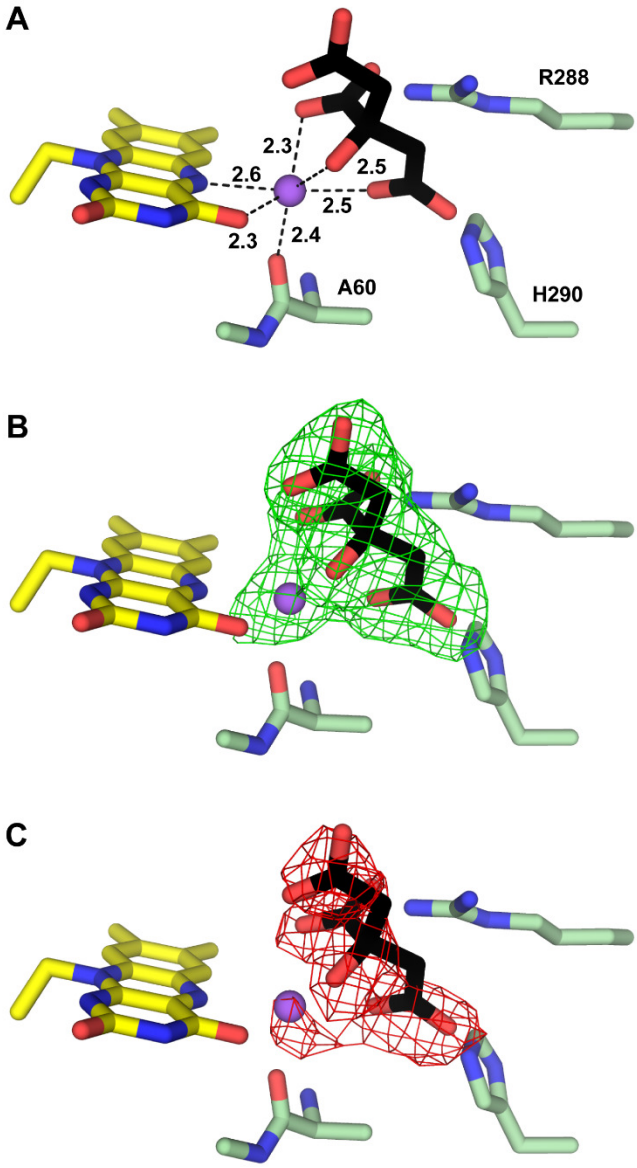



**Figure 6.4**: (A) Active site of CdUGM-His6 containing one sodium ion and one citrate ion. (B) The mesh corresponding to the $mF_o$-$DF_c$ map at 3 σ after the citrate ion was removed and structure re-refined with simulated annealing. (B) Disappearance of the citrate ion and the sodium ion upon soaking of a crystal in potassium citrate-containing buffer. The mesh corresponding to a negative $mF_o$-$DF_c$ difference Fourier map at -3 σ.

### 6.4.2 GSG-CdUGM in complex with UDP

The GSG-UGM construct was crystallized in the presence UDP-Gal*p* for subsequent inhibitor soaking to exchange the ligand. However, the crystals cracked instantly upon soaking with inhibitor solutions. GSG-UGM is monoclinic with two molecules in the asymmetric unit (Figure 6.5), corresponding to the biological dimer, with solvent content of 54.9% and Matthews coefficient of 2.73. Both dimers are in a closed conformation, providing rationale for crystal cracking upon inhibitor soaking. The structure also revealed that during the course of crystallization UDP-Gal*p* had hydrolyzed to UDP. Electron density of chain A is well-defined. In contrast, chain B has significantly elevated B factors and side chain densities in several regions are weak. However, the backbone density is strong enough for unambiguous tracing. During refinement, torsional NCS restraints were applied to improve the geometry of chain B. The optimal refinement strategy involves splitting chain B into four TLS groups (residues 2-62, 63-235, 236-316, and 317-386) and chain A as one TLS group.

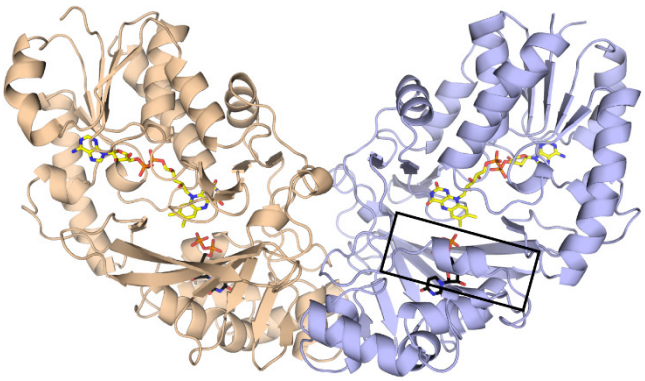


**Figure 6.5**: Overall structure of dimeric GSG-CdUGM. Both subunits are in the closed conformation with the lid forming a short helix (box) covering the active site.

Upon lid closing, UDP occupies a tight space in the active site (Figure 6.6). The uracil ring stacks onto the π face of Y157. L174 and L177, along with F98 and I154,

form a hydrophobic box around the uracil. The N3H of uracil forms a hydrogen bond with the backbone carbonyl of F153. Uracil O8 is recognized by side chains of N278 and N280. T158 hydrogen bonds to both O7 of uracil and 2'OH of ribose. W162 hydrogen bonds to the 3'OH of ribose. The diphosphate group is recognized by the side chains of R176 (from the lid), Y187, R288, Y326 and Y364. Positions of these residues are conserved when the GSG-CdUGM structure is superimposed with the catalytically active, reduced, closed KpUGM[247] (Cα RMSD = 1.27 Å) (Figure 6.7). UDP-Gal*p* from the KpUGM structure fits well into the pocket of closed GSG-CdUGM. Moreover, the superimposed anomeric carbon of Gal*p* is about 3.4 Å away from the nucleophilic nitrogen of FAD. This distance is comparable to the value observed in KpUGM (3.6 Å). These results suggest that even though GSG-CdUGM has oxidized FAD and is not catalytically competent, the protein can adopt the catalytically active conformation. Thus, the energy barrier for conformational change upon FAD reduction is small and might largely be dictated by crystal contact.

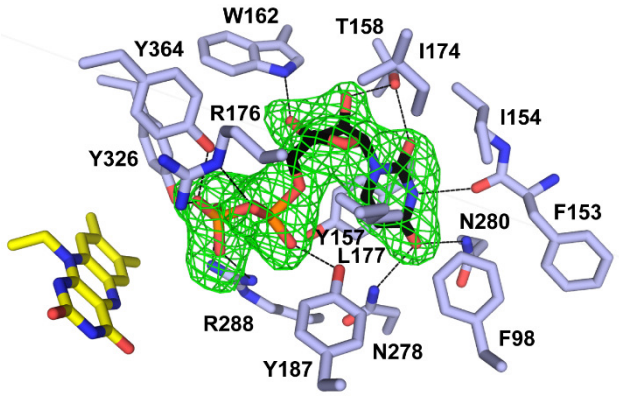



**Figure 6.6**: Active site of GSG-CdUGM (chain A) with bound UDP. The mesh correspond to the $mF_o$-$DF_c$ map at 3σ after UDP was removed and the structure re-refined with simulated annealing.

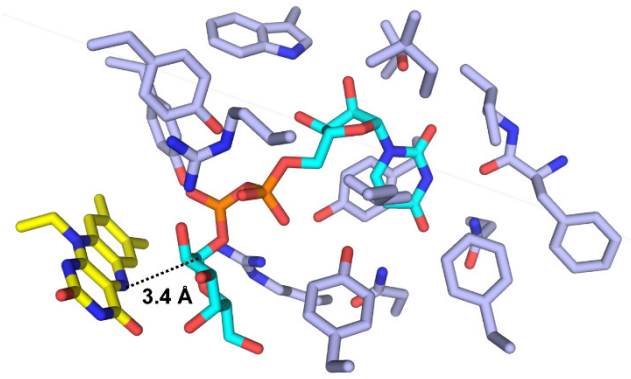



**Figure 6.7**: Position of UDP-Gal*p* from a structure of catalytically active KpUGM (PDB ID 3INT) aligned to GSG-CdUGM with bound UDP.

Previously, the structure of CdUGM with the triazolothiadiazine suggest that the lid-closed conformation of CdUGM might be relevant for inhibition[256]. With the availability of the closed GSG-CdUGM structure from this work, structure superposition was performed (Figure 6.8). The docked inhibitor could fit into both the structure of opened and closed CdUGM without any obvious clashes. These results suggest that closed CdUGM could accommodate the inhibitor adopting the docked pose, as predicted from the docked KpUGM structure. However, the chemical environment of the closed active site around the docked ligand is conserved between CdUGM and KpUGM. Thus, the docked ligand pose does not explain why the inhibitor is more effective against KpUGM compared to CdUGM.

Because the docked inhibitor pose is different from the inhibitor pose from our previously reported crystal structure, we superimposed the crystal structure of open CdUGM containing bound inhibitor with the closed GSG-CdUGM. The superposition reveals steric clashes. R176 from the lid clash with the inhibitor, but flexibility of the R176 side chain and the available space may alleviate this steric clash. Interestingly, Y364 in the closed GSG-CdUGM structure moves into the active site and the side chain

phenol ring is twisted perpendicular compared to Y364 in the opened structure. This Y364 movement clashes into the inhibitor found in opened CdUGM. These results suggest that the inhibitor may not bind CdUGM in the closed conformation effectively. Moreover, the same position of Y364 equivalent is observed in the closed structure of MtUGM[233]. However, this movement is not observed in KpUGM[247]. In addition, N173 equivalent in KpUGM is an isoleucine and can thus interact with higher affinity to the hydrophobic chlorophenyl ring of the inhibitor. Collectively, these results suggest that the pose of triazolothiadiazine inhibitor observed previously with X-ray crystallography is likely correct. The inhibitor could bind to both the closed and opened conformation of KpUGM, with the closed form being a higher affinity form compared to the opened structure. In contrast, the inhibitor may only bind CdUGM in opened, lower affinity conformation. These results explain why the inhibitors are more potent against KpUGM compared to CdUGM and MtUGM, and also highlight the role of UGM conformational flexibility in inhibitor design.

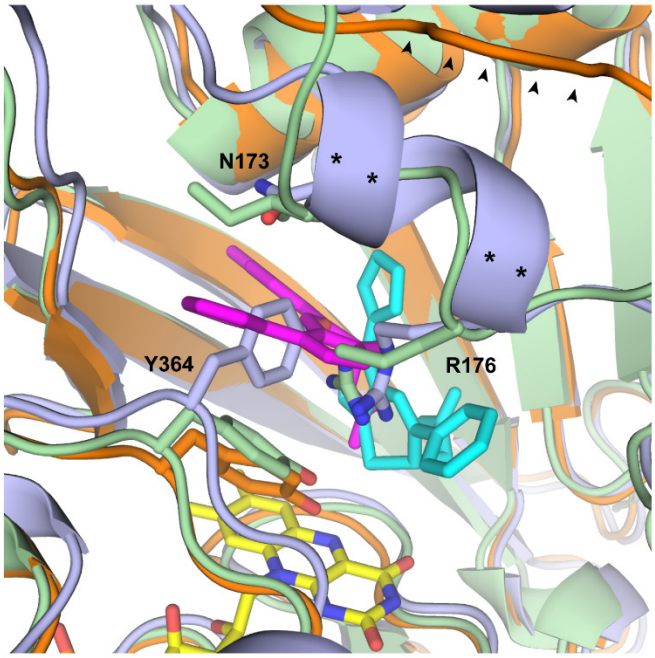



**Figure 6.8**: Alignment of GSG-CdUGM (blue) in complex with UDP, G-CdUGM (orange) with a triazolothiadiazine inhibitor (magenta), and KpUGM (green) with a docked triazolothiadiazine inhibitor (cyan). The opened lid is marked with a row of arrows. The closed helical lid is marked with asterisks. The residue numbering of CdUGM is used.

## 6.5 Conclusions

Despite availability of UGM structures, to date there is only one UGM structure with non-substrate analog bound. Because UGM can be in either FAD reduced or oxidized form in combination with closed and opened lid conformation, it is not clear how these structural dynamics contribute to inhibitor affinities. Our work provides two additional structures of CdUGM for comparison. The opened CdUGM-His6 in complex with sodium citrate has improved geometry over the reported structure and identifies an additional pocket in the active site that could be exploited in inhibitor design. The GSG-CdUGM in complex with UDP provides the first structure of closed CdUGM. Analysis of these structure in combination of previously reported CdUGM and KpUGM structure suggests that the closed form of UGM might be the high affinity form targeted by

existing inhibitors. In addition, this work underscores the importance of understanding UGM structural dynamics for not only designing novel inhibitors, but also for obtaining structures of UGM-inhibitor complexes.

# Chapter 7

## Structural dynamics of UGM from *Mycobacterium smegmatis*

## 7.1 Abstract

UDP-galactopyranose mutase is central to biosynthesis of galactofuranose (Gal*f*) and is an essential enzyme for several pathogenic microorganisms. Several classes of UGM inhibitor have been identified, but structural knowledge of the inhibition mechanism is limited. We previously reported the first structure of a small molecule, non-substrate mimic inhibitor bound to *Corynebacterium diphtheriae* UGM (CdUGM). The structure suggests that conformational dynamics are likely a significant affinity determinant and may explain differential affinity of inhibitors among UGM homologs. Although structural dynamics of eukaryotic UGM has gained significant attention recently, studies on prokaryotic UGM structural dynamics were performed when structural knowledge were limited. Here, we present three structures of UGMs from *Mycobacterium smegmatis* (MsUGM) in various conformational stages. The UDP bound MsUGM (1.83 Å) shows a dimeric structure with one lid closed and the other opened. The same crystal form with partial UDP occupancy (1.69 Å) shows a mixture of opened and closed conformation within the same subunit. The $Mg^{2+}$-bound MsUGM at 2.05 Å is in an opened conformation and structural movements were inferred from TLS refinement. These structures are not only the highest resolution UGM structure reported to date, but also provide insights into conformational dynamics of UGM and offer additional venue for UGM inhibitor design.

## 7.2 Introduction

Galactofuranose (Gal*f*) residue is found in virulence factors[215], surface polysaccharides[161], and cell wall components[212, 213] of several pathogenic organisms. These Gal*f* glycoconjugates are synthesized using UDP-Gal*f* as an activated sugar

donor. UDP-Gal*f* is in turn synthesized from UDP-galactopyranose (UDP-Gal*p*) by UDP-galactopyranose mutase (UGM)[219-221]. UGM is an essential gene in Mycobacteria[212], in which Gal*f* is a component of the galactan layer in the cell wall[213]. Because UGM is not found in mammalian systems[222], UGM is an attractive target for antibiotic development.

Substrate mimics have been described as UGM inhibitors[227-234]. However, these UDP-based inhibitors are not cell permeable, thus do not exhibit antimicrobial activity. We have reported small molecule, non-substrate inhibitors of UGM[224, 225, 256]. Specifically, triazolothiadiazine inhibitors discovered through virtual screening have low μM affinity toward UGMs and also exhibit anti-mycobacterial activity[256]. The crystal structure of the triazolothiadiazine inhibitor bound to *Corynebacterium diphtheriae* UGM (CdUGM) reveal that the inhibitor binds UGM in an opened conformation, in contrast to the closed conformation of *Klebsiella pneumoniae* UGM (KpUGM) used in virtual screening[256]. However, unexplained electron density below the active site lid suggests that the inhibitor could potentially bind to multiple conformations of UGM. Differential affinity of inhibitors toward different conformations of UGM may also explain why inhibitors are more effective against KpUGM compared to other homologs. Specifically, the $K_i$ values of a triazolothiadiazine inhibitor against KpUGM, CdUGM and *Mycobacterium tuberculosis* (MtUGM), are 8±3, 77±37, and 31±18μM, respectively[256]. Recent structures of CdUGM in different conformations (Chapter 6) suggest that CdUGM changes conformation upon lid closing, which sterically clash with the inhibitor. The clash is not observed in the structure of KpUGM, thus may explain higher affinity of inhibitors towards KpUGM. These observations prompt us to obtain additional structural information on conformational dynamics of UGMs.

Although molecular dynamics of eukaryotic UGM is well described[260, 261], studies of prokaryotic UGM were performed with early crystal structures without experimentally determined substrate binding pose[262-264]. Therefore, it was unclear whether molecular dynamic simulations using computationally docked ligand are biochemically relevant. With the availability of substrate-bound crystal structures[247, 265], it is evident that a mobile lid closes down on the active site and the anomeric carbon of UDP-Gal*p* moves into close proximity of N5 of FAD for catalysis. However, this movement has only been observed experimentally in KpUGM crystals upon reduction of FAD[247]. To date, other prokaryotic UGM structures in complex with UDP-Gal*p* show a closed lid conformation even before FAD reduction[233, 265]. These results suggest that crystal contact could be a major determinant of enzyme conformation regardless of FAD redox state and presence of the substrate. Because evidence for enzyme movement within the same crystal form is limited to KpUGM, more information is needed in order to elucidate the role of structural dynamics in the context of inhibitor design. Information from homologs other than KpUGM would be desirable to explain the differential affinity of inhibitors[224, 256] and would also provide additional evidence for the role of enzyme conformation and inhibitor affinity suggested by CdUGM structures.

Here we present three structures of *Mycobacterium smegmatis* UGM (MsUGM) in various conformations. The first high resolution structure (MsUGM-SS, 1.83 Å) contain UDP in one subunit. Interestingly, the subunit could adopt the catalytically active closed conformation. The other subunit is completely opened. The second high resolution structure (MsUGM-QS, 1.69 Å) is the same crystal form, but the crystal was vitrified immediately without prolonged cryoprotection. This crystal shows partial

occupancy of UDP in the formerly opened subunit. Intriguingly, the subunit with partial UDP occupancy exhibits a mixture of both closed and opened conformation that could be modeled unambiguously. Therefore, these two structures provide valuable information regarding dynamics of UGM structures. The third structure (MsUGM-His6) is a magnesium ion complex of MsUGM. The magnesium ion bridges UGM molecules to pack into crystals. The coordinate is refined with TLS (translation, libration, screw) parameters, which gives general insights into movement of each segments. These structures not only provide information of enzyme dynamics, but also suggest strategies for inhibitor design to target various UGM homologs.

### 7.3 Material and Methods

Both MsUGM constructs were purified and concentrated to 10 mg/mL using a previously reported protocol[137, 256]. Indexing, integration, and scaling were performed with HKL2000[138]. Molecular replacement was performed with Phaser[140]. Coordinate restraints for small molecule ligands were generated using eLBOW[250]. Model adjustment and refinement were performed with Coot and phenix.refine respectively[141, 142]. The model was validated using MolProbity[144]. Figures were generated with PyMOL[145].

#### 7.3.1 MsUGM (13-412) (MsUGM-SS and MsUGM-QS) in complex with UDP

Sequence coding residues 13-412 were amplified with primers CTCGCATATGCACCATCACCATAGCGAAAACCTGTATTTTCAGTCCACCGGTAATTTCGACCTATTCGTC and TATCGTCGACTCATTATGCTTTGCTGCTTTCGTCTTCGGTG using a pET24a plasmid containing *Mycobacterium smegmatis* $MC^2$ 155 UGM with a C-terminal His6 tag as a

template. The PCR product was then cloned into a modified pMALc5x that contains an N-terminal His6 tag in place of the MBP sequence. The resulting construct has an N-terminal His6 followed by a TEV cleavage site. The C-terminal glycine of the TEV recognition site is replaced with a serine so that, after TEV cleavage, the product has native MsUGM (13-412) protein sequence. Crystallization was achieved with hanging drop vapor diffusion by mixing 2 µL of protein solution and 2 µL of well solution (100 mM Tris pH 8.5, 100 mM $NaNO_3$, 1.8 M $(NH_4)_2SO_4$). Numerous small rod-shaped crystals appear within a few days. These rods disappear over the course of 1-2 months and is replaced by a few larger crystal, presumably of a different crystal form. Two cryoprotection strategies were employed. First, the crystals were soaked in 100 mM Tris pH 8.5, 100 mM $NaNO_3$, 2.5 M $(NH_4)_2SO_4$, 5% glycerol for 12 hours then soaked in an equivalent solution containing 10% and 15% glycerol, each for 3 hours. For the second strategy, crystals were briefly soaked in a solution of 100 mM Tris pH 8.5, 100 mM $NaNO_3$, 2.5 M $(NH_4)_2SO_4$, 15% glycerol. After cryoprotection, the crystals were vitrified and stored in liquid nitrogen. Single crystal X-ray diffraction data were collected at beamline 21-ID-D (sequential soak) and 21-ID-F (quick soak) (LS-CAT) at Advanced Photon Source, Argonne National Laboratory. For the sequential soak data (refer to as MsUGM-SS, molecular replacement was performed using a previously reported structure of *Escherichia coli* UGM as a search model (PDB ID 1I8T)[221]. The quick soak structure (refer to as MsUGM-QS) was solved with a difference Fourier method using the sequential soak structure as the starting model with phenix.refine rigid body refinement.

**Table 7.1**: Data collection and refinement statistics of MsUGM (13-412) in complex with UDP

| Data collection statistics | MsUGM-SS | MsUGM-QS |
|---|---|---|
| Wavelength (Å) | 0.97854 | 0.97872 |
| Resolution range (Å)* | 47.34 – 1.85 (1.92 – 1.83) | 42.80 – 1.70 (1.76 – 1.69) |
| Space group | I 2 2 2 | I 2 2 2 |
| Unit cell (Å) | 124.2 132.2 135.7 | 123.7 131.4 135.8 |
| Total reflections | 699,321 (65,062) | 914,126 (89,451) |
| Unique reflections | 96,510 (9,568) | 122,178 (12,088) |
| Multiplicity | 7.2 (6.8) | 7.5 (7.4) |
| Completeness (%) | 100.0 (100.0) | 100.0 (100.0) |
| Mean I/σ(I) | 23.3 (2.57) | 23.4 (2.0) |
| Wilson B-factor (Å$^2$) | 25.06 | 21.68 |
| R-merge | 0.088 (0.694) | 0.093 (0.810) |
| R-meas | 0.095 (0.752) | 0.099 (0.871) |
| R-pim | 0.035 (0.286) | 0.036 (0.319) |
| Refinement Statistics | | |
| Resolution range (Å) | 45.81 – 1.83 (1.90 – 1.83) | 42.53 – 1.69 (1.75 – 1.69) |
| R-factor | 0.1682 (0.2312) | 0.1866 (0.2514) |
| R-free (5%) | 0.1958 (0.2646) | 0.2203 (0.3064) |
| Number of atoms | | |
| Protein | 6458 | 7031 |
| FAD | 106 | 106 |
| UDP | 25 | 50 |
| Water | 740 | 932 |
| Protein residues | 783 | 787 |
| RMSD (bonds, Å) | 0.007 | 0.007 |
| RMSD (angles, °) | 1.059 | 1.082 |
| Est. coordinate error (ML, Å) | 0.18 | 0.21 |
| Ramachandran favored (%) | 98.6 | 98.1 |
| Ramachandran outliers (%) | 0 | 0.47 |
| Average B-factor (Å$^2$) | | |
| Chain A | | |
| Protein | 28.15 | 25.60 |
| FAD | 28.81 | 27.60 |
| UDP | - | 28.34 (0.65 occ) |
| Water | 35.98 | 35.10 |
| Chain B | | |
| Protein | 28.81 | 27.23 |
| FAD | 31.44 | 30.93 |
| UDP | 26.38 | 22.95 |
| Water | 36.68 | 36.26 |

*Statistics for the highest-resolution shell are shown in parentheses. The coordinates and structure factors have been deposited to the Protein Data Bank under accession code 5EQD (MsUGM-SS) and 5ER9 (MsUGM-QS).

### 7.3.2 MsUGM-His6 in complex with magnesium ion

Sequence coding residues 1-412 were amplified from the genomic DNA of *Mycobacterium smegmatis* $MC^2$ 155 with primers TATACATATGACGTCCATATCGCACCCGGTCGC and CGACAAGCTTTCATTAATGGTGGTGATGATGGTGTGCTTTGCTGCTTTCGTCTTCG GTG. The PCR product was then cloned into the NdeI and HindIII of pET24a. The resulting construct contains a C-terminal His6 tag without any linker residue (MsUGM-His6). Crystallization was achieved with hanging drop vapor diffusion by mixing 2 µL of protein solution and 2 µL of well solution (50 mM Tris pH 8.5, 25 mM $MgCl_2$, 18% PEG 3350, and 20% glycerol). Crystal appears instantly and grow to full size within 24 hours. Crystals were then soaked in the well solution supplemented with 5 mM of a triazolothiadiazine inhibitor (2-[(7S)-3-(2-furyl)-6-phenyl-7H-[1,2,4]triazolo[3,4-b][1,3,4]thiadiazin-7-yl]acetic acid) (from a 50 mM stock in DMSO) overnight. Crystals were then vitrified and stored in liquid nitrogen. Single crystal X-ray diffraction data were collected at beamline 21-ID-F (LS-CAT) at Advanced Photon Source, Argonne National Laboratory. Molecular replacement was performed using an opened structure of MsUGM (13-412) (PDB ID 5EQD) as a search model.

**Table 7.2**: Data collection and refinement statistics for MsUGM-His6 in complex with $Mg^{2+}$

| Data collection statistics | |
|---|---|
| Wavelength (Å) | 0.97872 |
| Resolution range (Å)* | 40.70 – 2.05 (2.12 – 2.04) |
| Space group | P $4_1$ $2_1$ 2 |
| Unit cell | 86.0 86.0 126.2 |
| Total reflections | 874,713 (82,858) |
| Unique reflections | 30,831 (3,024) |
| Multiplicity | 28.4 (27.4) |
| Completeness (%) | 100.0 (100.0) |
| Mean I/sigma(I) | 47.1 (3.6) |
| Wilson B-factor ($Å^2$) | 41.80 |
| R-merge | 0.074 (0.923) |
| R-meas | 0.075 (0.940) |
| R-pim | 0.014 (0.176) |
| **Refinement Statistics** | |
| Resolution range (Å) | 40.70 – 2.04 (2.11 – 2.04) |
| R-factor | 0.1731 (0.2466) |
| R-free (5%) | 0.2396 (0.3449) |
| Number of atoms | |
| protein | 3209 |
| FAD | 53 |
| $Mg^{2+}$ | 1 |
| Water | 198 |
| Protein residues | 391 |
| RMSD (bonds, Å) | 0.007 |
| RMSD (angles, Å) | 1.063 |
| Est. coordinate error (ML, Å) | 0.22 |
| Ramachandran favored (%) | 98.2 |
| Ramachandran outliers (%) | 0 |
| Average B-factor | |
| protein | 52.12 |
| FAD | 48.54 |
| $Mg^{2+}$ | 34.89 |
| Water | 52.18 |

*Statistics for the highest-resolution shell are shown in parentheses. The coordinates and structure factors have been deposited to the Protein Data Bank under accession code 5F3R.

## 7.4 Results

### 7.4.1 MsUGM-SS and MsUGM-QS in complex with UDP

MsUGM (13-412) crystals were cryoprotected with two different strategies: sequential soak (MsUGM-SS) and quick soak (MsUGM-QS). These two crystals are of

the same form and contains two molecules in the asymmetric unit with solvent content of 59.7% and Matthews coefficient of 3.05.

Each subunit of MsUGM-SS are in different conformations (Figure 7.1). Chain A is in an opened conformation with the mobile lid lacking a secondary structure, not covering the active site cavity. Chain B is in a closed conformation with the mobile lid adopting a helical structure covering the active site containing UDP produced from degradation of UDP-Gal*p*.

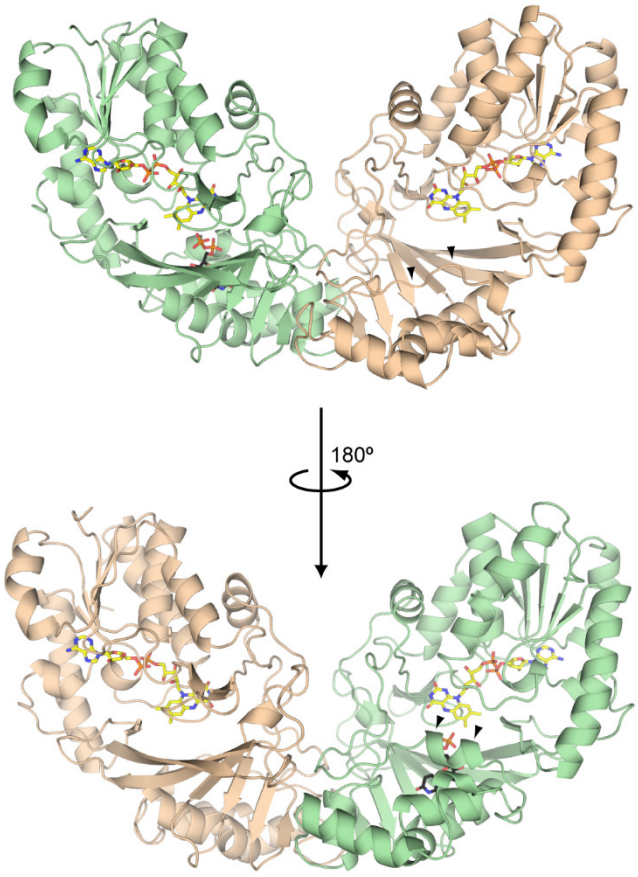



**Figure 7.1**: Structure of dimeric MsUGM-SS with chain A opened (wheat) and chain B closed (green). The position of the lid is marked with arrows.

The UDP in chain B occupies a tight cavity of the active site (Figure 7.2 A). The uracil portion occupies a hydrophobic box forms by F111, I167, I187, and I190. The

uracil ring stacks onto the π system of Y170. N3H forms a hydrogen bond with the backbone carbonyl of F166. O8 is recognized by the side chain of N291 and N293. O7 forms a hydrogen bond with T171, which also forms a hydrogen ond with the 2'OH of the ribose ring. The 3'OH is recognized by the indole NH of W175. The diphosphate moiety forms salt bridges with R189 from the mobile lid and also R301 at the bottom of the active site. Additional hydrogen bonds between the diphosphate groups and Y200, Y337, and Y375 is also observed. The conformation of the closed MsUGM-SS is the same as found in the catalytically active KpUGM[247]. Superposition of KpUGM onto MsUGM-SS allows examination of the expected UDP-Gal*p* pose (Figure 7.2 B)[247]. The anomeric carbon of UDP-Gal*p* is at the same distance (3.6 Å) away from N5 of FAD as found in the KpUGM structure. These results suggest that the protein conformation of chain B is the same as catalytically competent UGM even though FAD is not reduced.

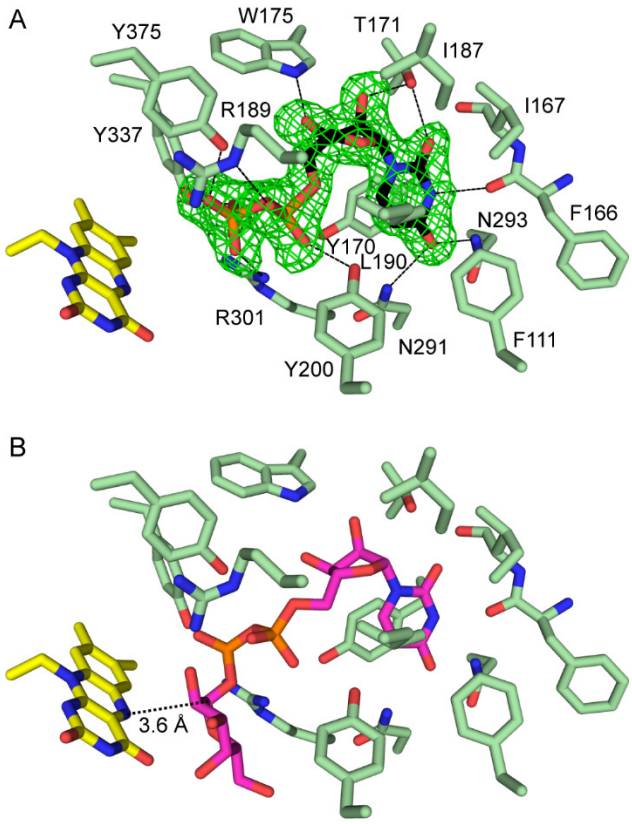



**Figure 7.2**: (A) UDP-bound active site of chain B. The mesh represents a mFo-DFc map at 3σ calculated after UDP was removed and the structure re-refined with simulated annealing. (B) Superposition of UDP-Galp from a catalytically competent structure of KpUGM (PDB ID 3INT).

Vitrification after quick soak in cryoprotectant, not allowing extensive back-soaking of UDP, yielded the MsUGM-QS structure. Chain B remains closed with full UDP occupancy. However, chain A also has a weaker density of UDP. The occupancy of UDP was refined to 65%. Intriguingly, chain A is a mixture of both opened and closed conformations. The alternative conformations could be modeled unambiguously (Figure 7.3 A). Occupancy of contiguous fragments were constrained to be the same, while no occupancy constraints were applied across fragments. Residues that were split are 137-164 (41% occupancy) and 174-193 (53% occupancy). Residues 137-164 move toward the active site mostly as a rigid body with little change in secondary structure (Figure 7.3

B). In contrast, residues 174-193, close over the active site and form a helical lid structure. In addition to these segments, Y375 also moves into the active site to hydrogen bond with the diphosphate group of UDP.

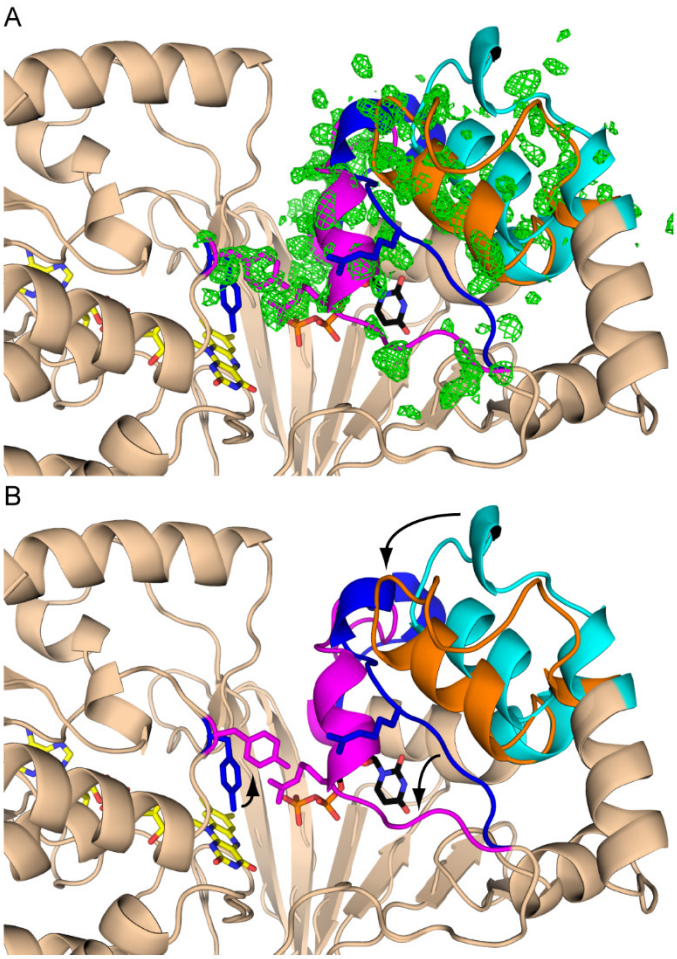



**Figure 7.3**: (A) Structure of chain A in the dimeric MsUGM-QS with 65% UDP occupancy. The monomer is in a mixed opened and closed conformation. The green mesh represents $mF_o$-$DF_c$ at 2σ before modeling of the closed conformation. Residues 137-164 are in cyan (opened) and orange (closed). Residues 174-193 are in blue (opened) and magenta (closed). Position of Y375 is also shown in blue (opened) and magenta (closed). (B) Arrows indicate putative movement paths of different segments.

### 7.4.2 MsUGM-His6 in complex with magnesium ion

Tetragonal MsUGM-His6 contains one molecule in the asymmetric unit with solvent content of 50% and Matthews coefficient of 2.48. Thus, both subunit of the

biological dimer are identical. MsUGM-His6 is in an opened conformation (Figure 7.4). The mobile lid does not adopt a secondary structure as observed in chain A of MsUGM-SS. Initial refinement attempts yield $R_{work}$ and $R_{free}$ of 20.3% and 27.0% respectively, which is relatively high for a 2.05 Å structure. Because the Wilson B factor is also high (41.8 $Å^2$), we reason that thermal motion is not sufficiently modeled by isotropic B factor refinement and TLS refinement would be more appropriate[266, 267]. Using the TLS Motion Determination (TLSMD) server[268, 269], several potential TLS groups were identified. One to nine TLS fragments were used in trial refinement runs to identify an optimum fragment. The optimal TLS refinement strategy consists of 4 TLS groups: residues 15-98, 99-199, 200-325, and 326-405. The resulting $R_{work}$ and $R_{free}$ are 17.4% and 23.7% respectively.

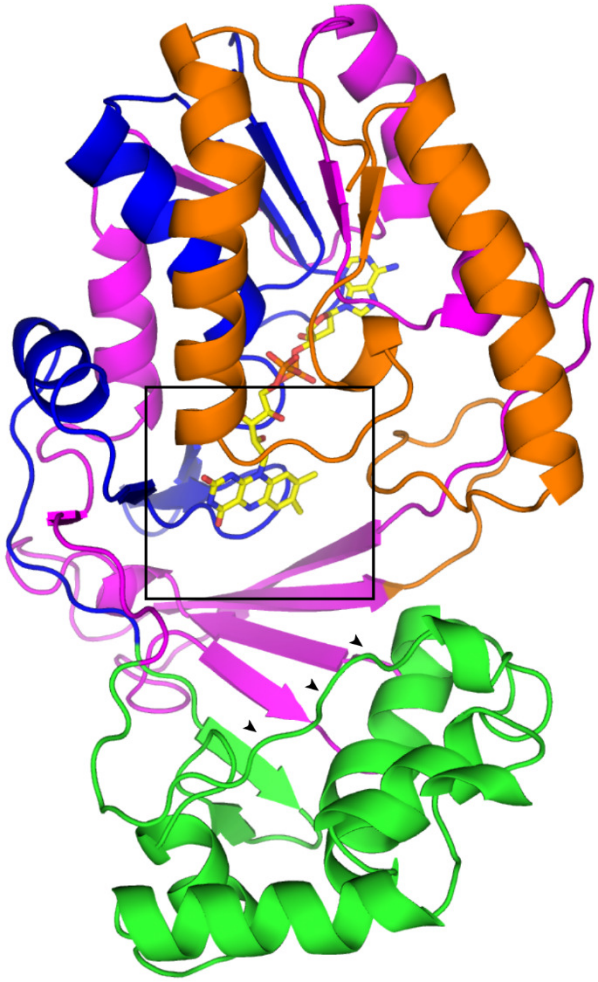


**Figure 7.4**: Structure of MsUGM-His6 with magnesium ion. TLS groups are labeled with colors: 15-98 (blue), 99-199 (green), 200-325 (magenta), and 326-405 (orange). The lid (arrow) is in an opened conformation, not covering the FAD-containing active site (box).

An octahedral set of peaks are present on the two-fold axis relating two subunits that are not biological dimers (Figure 7.5 A). Substantial evidence indicates that this set of peaks corresponds to magnesium ion coordination complex. Magnesium ion is required for crystallization, while a variety of PEG could be used as a precipitant. High concentration of magnesium ion results in showers of microcrystals. The octahedral geometry and bond length (2.1 Å) is also consistent with a magnesium ion complex[270]. In addition, placement of the magnesium ion does not result in positive or negative

peaks in the $mF_o$-$DF_c$ map after refinement. The magnesium ion has D330 from each UGM subunit as its axial ligands (Figure 7.5 B). Four ordered water molecules are equatorial ligands of the magnesium ion. Two of these water molecules, which are symmetry equivalent, form hydrogen bonds with side chain of R327 and backbone carbonyl of F328. The remaining oxygen of D330 that does not coordinate the magnesium ion also forms a hydrogen bond with these equatorial water molecules. The two remaining symmetry equivalent water molecules do not participate in other interactions.

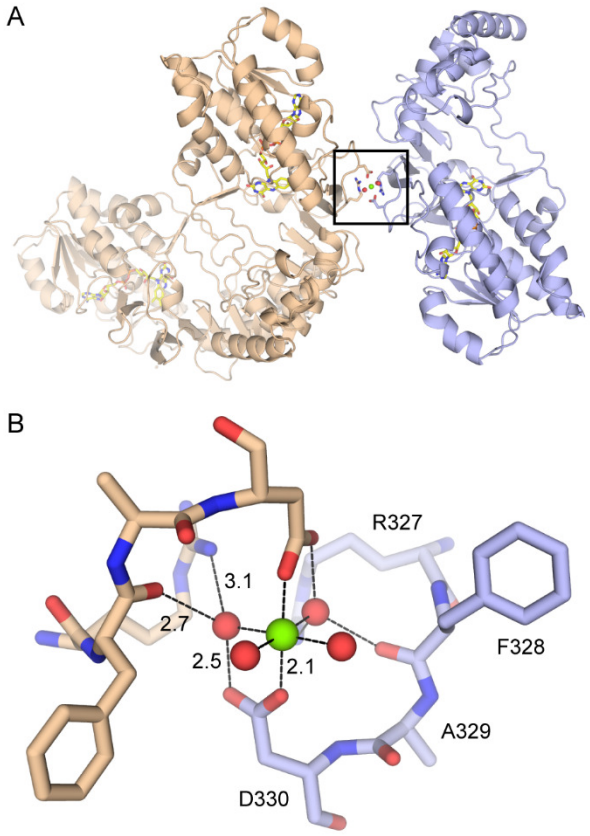



**Figure 7.5**: (A) Biological dimers (wheat) are linked together to form a crystal by magnesium ion coordination (box) on a crystallographic two-fold axis. The symmetry equivalent molecules are shown in wheat and blue for clarity. (B) Coordination environment around the magnesium ion (green). Numbers indicate bond distance in angstrom (Å).

### 7.5 Discussion and Conclusions

Previously reported prokaryotic UGM crystals, with the exception of KpUGM, are trapped in distinct opened or closed conformations that do not change conformation upon FAD reduction[233, 257, 265]. These observations suggest that crystal packing could be a major determinant of UGM conformation and that the energy barrier to conformational change in UGM is small. The structure of KpUGM in complex with UDP-Gal*p* previously reported by our group showed closing of the mobile lid onto the active site upon UGM reduction in the same crystal form[247]. These results suggest that, at least in KpUGM, reduction of FAD is the trigger for lid closing. The structures of MsUGM-SS and MsUGM-QS reported here also reveal movements of UGM segments within the same crystal form. However, in contrast to the KpUGM structure, FAD is oxidized in both structures and the ligand is UDP. Partial occupancy of UDP results in a mixed opened and closed conformation of UGM. The closed conformation is identical to those observed in KpUGM. In addition, back-soaking of UDP results in complete loss of UDP, which also leave the active site completely opened. These results not only suggest that oxidized MsUGM is capable of adopting catalytically active conformation, but non-substrate like UDP is also capable of inducing conformational change. Thus, it is likely that the energy barrier to conformational change is small regardless of FAD redox state. Therefore, it is possible that non-substrate, small molecule inhibitor could induce lid closing as well. The availability of the structure for both opened and closed conformation will provide valuable information to design anti-mycobacterial compound that target residues or pockets that are invariant upon conformational change. These

inhibitors should not lose affinity upon lid closing and should be more potent than inhibitors that target a single conformation.

Additional evidence for directional movement of UGM is found in the structure of magnesium ion-bound MsUGM-His6. TLS refinement dividing MsUGM-His6 into four segments results in significantly improved refinement statistics. Interestingly, residues 99-199, containing residues 137-164 (with mobile lid) and 174-193 that moves in MsUGM-QS, is in the same TLS group as identified by the TLSMD server. Therefore, it is likely that residues 99-199 have propensity to move in the same direction. These prediction form the opened MsUGM-His6 structure is in agreement with the mixed conformation of MsUGM-QS. With recent development in structural analysis with diffuse scattering, it would be beneficial to examine the agreement between current TLS groups and the observed diffuse scattering once more advanced detectors become commonly available[271, 272].

MsUGM-His6 crystal forms readily in contrast to MsUGM-QS. A variety of precipitants can be used, but presence of magnesium ion is absolutely required. The structure revealed that the magnesium ion bridges UGM together to form crystals. The magnesium ion itself occupy the crystallographic two-fold axis. Because of the ease in crystallization and possible dynamics of the structure, this crystal form could be useful for obtaining structures of inhibitor complex. Moreover, if magnesium ion coordination is the main driving force behind crystallization, the MsUGM-His6 structure could be used for protein surface engineering to obtain crystal structure of other UGMs that are recalcitrant to crystallization.

Three MsUGM structures presented here complement the existing knowledge of mycobacterial UGM structure. Although several structures of MtUGM are available, structural dynamics of MtUGM is not well understood. Specifically, it is not known whether structural dynamics of mycobacterial UGMs is the same as other homologs such as KpUGM or eukaryotic UGMs. These MsUGM structures highlighting experimentally determined conformational dynamics within the same crystal form. Given the potential role of UGM conformational changes in inhibitor affinity as seen in CdUGM-inhibitor complex structure, these MsUGM structures could aid in further development of potent anti-mycobacterial both biochemically and computationally.

# Appendix A

## Crystallization Screen for Intelectins

Intelectins require calcium ion for their function. However, large proportion of conditions in commercial crystallization screens contain calcium-incompatible reagents such as phosphate, sulfate, and citrate. Thus, crystallization screen of intelectin using commercial screens will result in chelation of calcium ion or false positive hits from insoluble calcium salt crystals. Therefore, a targeted screen was developed based on previously reported conditions for calcium containing proteins and lectins found in the Protein Data Bank (PDB). The table below shows the 48-conditions screen used to crystallize hIntL-1 and XEEL without further optimization. The screen is designed to be compatible with commercial screens. Therefore, conditions that are exact overlap with commercial screens were slightly modified, such as changing buffer salt or pH, and noted accordingly.

**Table A.1**: Crystallization screen for intelectins

| Number | Buffer | pH | Salt/Additive | Precipitant | PDB ID | Notes |
|---|---|---|---|---|---|---|
| 1 | | | 200 mM $NH_4Cl$ | 15% PEG 3350 | 3KQG | |
| 2 | 100 mM NaOAc | 4.5 | | 2 M NaCl | 2BPE | Published pH = 4.6 |
| 3 | 100 mM NaOAc | 4.5 | 200 mM $Zn(OAc)_2$ | 20% PEG 1000 | 1FM5 | |
| 4 | 100 mM MES | 6 | 200 mM $Ca(OAc)_2$ | 15% PEG 400 | | |
| 5 | 100 mM Bis-Tris | 6 | | 25% PEG 3350 | | |
| 6 | 100 mM Na Cacodylate | 6 | | 15% PEG 4000 | | |
| 7 | 100 mM MES | 6 | 200 mM NaCl | 15% PEG 6000 | | |
| 8 | 100 mM Na Cacodylate | 6 | 200 mM $Zn(OAc)_2$ | 7% PEG 8000 | 1TLG | |
| 9 | 100 mM Bis-Tris | 6 | | 12% PEG 10000 | 1PWB | Published using Tris |
| 10 | 100 mM Bis-Tris | 6.5 | 200 mM $Ca(OAc)_2$ | 45% MPD | | |
| 11 | 100 mM Imidazole | 6.5 | | 30% PEG 2000 | | |
| 12 | 100 mM Na Cacodylate | 6.5 | 100 mM $MgCl_2$ | 12% PEG 3000 | 1XPH | |
| 13 | 100 mM Na Cacodylate | 6.5 | 200 mM NaCl | 20% PEG 3350 | 2XR5 | |
| 14 | 100 mM Na Cacodylate | 6.5 | 200 mM NaCl | 35% PEG 3350 | 2XR6 | |
| 15 | 100 mM MES | 6.5 | 300 mM $MgC_{l2}$ | 5% PEG 4000 | 2DUR | |
| 16 | 100 mM Bis-Tris | 6.5 | | 20% PEG 5000 MME | 2C6U | |
| 17 | 100 mM MES | 6.5 | | 20% PEG 6000 | | |
| 18 | 100 mM Na Cacodylate | 6.5 | 200 mM KCl<br>100 mM $Mg(OAc)_2$ | 10% PEG 8000 | 2BPH | |
| 19 | 100 mM Na Cacodylate | 6.5 | 200 mM $Ca(OAc)_2$ | 18% PEG 8000 | 1K9I | |
| 20 | 100 mM Na Cacodylate | 6.5 | 200 mM NaOAc | 30% PEG 8000 | 2HRL | |
| 21 | 100 mM HEPES | 7 | 10% Glycerol | 5 M Na Formate | 1JZN | |
| 22 | 100 mM Tris-HCl | 7 | | 20% PEG 400 | 1SL4 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 23 | 100 mM Tris-HCl | 7 | 200 mM NaCl | 30% PEG 3000 | 2IT5<br>2IT6 | |
| 24 | 100 mM HEPES | 7 | 400 mM $MgCl_2$ | 25% PEG 3350 | 3C22 | |
| 25 | 100 mM HEPES | 7 | 100 mM $MgCl_2$ | 25% PEG 4000 | 3P5F<br>3P5I | |
| 26 | 100 mM Na Cacodylate | 7 | 100 mM $MgCl_2$<br>5 mM $CaCl_2$ | 13% PEG 4000 | 3P7F<br>3P7G<br>3P7H | Published pH = 6.9 |
| 27 | 100 mM HEPES | 7 | | 20% PEG 6000 | | |
| 28 | 100 mM Tris-HCl | 7 | 200 mM $Zn(OAc)_2$ | 8% PEG 8000 | 2OX8 | |
| 29 | 100 mM HEPES | 7.5 | | 4.3 M NaCl | 3ALS | |
| 30 | 100 mM HEPES | 7.5 | 200 mM NaCl | 30% PEG 300 | 1SL6 | |
| 31 | 100 mM HEPES | 7.5 | 100 mM LiCl | 20% PEG 400 | | |
| 32 | 100 mM HEPES | 7.5 | | 25% PEG 2000 MME | | |
| 33 | 100 mM HEPES | 7.5 | | 20% PEG 3350 | | |
| 34 | 100 mM HEPES | 7.5 | 15% Glycerol | 17% PEG 4000<br>8.5% Isopropanol | 2DF3 | |
| 35 | 100 mM HEPES | 7.5 | 100 mM KCl | 20% PEG 6000 | | |
| 36 | 100 mM HEPES | 7.5 | | 20% PEG 8000 | 1K9J | |
| 37 | 100 mM Tris-HCl | 8 | 10 mM $CaCl_2$ | 60% MPD | 1WMY | |
| 38 | 100 mM Tris-HCl | 8 | | 25% PEG 400 | | |
| 39 | 100 mM Tris-HCl | 8 | 100 mM KCl | 15% PEG 2000 MME | | |
| 40 | 100 mM Tris-HCl | 8 | 200 mM NaCl | 25% PEG 3350 | | |
| 41 | 100 mM Tris-HCl | 8 | | 20% PEG 4000 | 1PW9 | Published using 10 mM Tris |
| 42 | 100 mM Tris-HCl | 8 | | 30% PEG 6000 | 2YHF | |
| 43 | 100 mM Tris-HCl | 8 | 20 mM $CaCl_2$<br>10 mM NaCl | 15% PEG 8000 | 1FIF | Published using 13.5% PEG |
| 44 | 100 mM Tris-HCl | 8 | 200 mM MgCl2 | 30% PEG 8000 | 3VPP | |
| 45 | 50 mM Tris-HCl | 8.5 | | 3.75 Na Formate | 1XAR | |
| 46 | 100 mM Tris-HCl | 8.5 | 200 mM MgCl2 | 30% PEG 4000 | 1SL5 | |
| 47 | 100 mM Tris-HCl | 8.5 | | 25% PEG 6000 | | |
| 48 | 100 mM Tris-HCl | 8.5 | 200 mM NaCl | 30% PEG 8000 | 2OX9 | Published using imidazole |

# Appendix B

## Facilitation of folding and crystallization of the type II TGF-β receptor extracellular domain by non-detergent sulfobetaine-201

Portions of this chapter have been published in:

Wangkanont, K., Forest, K. T., Kiessling, L. L., The non-detergent sulfobetaine-201 acts as a pharmacological chaperone to promote folding and crystallization of the type II TGF-β receptor extracellular domain. *Protein Expr Purif*. 2015, 115, 19-25.

## B.1 Abstract

Signaling through the type II TGF-β receptor (TBRII) is crucial for numerous biological processes such as development and cancer metastasis. Therefore, chemical probes that bind and modulate TBRII activities could have therapeutic potentials. To develop such probes, large amounts of TBRII extracellular domain (TBRII-ECD) are needed for biophysical studies. However, the active TBRII-ECD yield obtained from existing folding protocols are highly variable and limited. To overcome the protein production barrier, we developed an on-plate folding screen to discover new folding conditions. This assay identified non-detergent sulfobetaine 201 (NDSB-201) as an effective additive. Because not all NDSBs are effective additives, we postulated that TBRII-ECD possesses a specific binding site for NDSB-201. X-ray crystallography using TBRII-ECD crystals grown with NDSB-201 revealed a hydrophobic binding pocket for NDSB-201. The ability of NDSB-201 to bind TBRII-ECD suggests stabilization of the folded state as a molecular mechanism for TBRII-ECD folding, in addition to minimization of protein aggregation by NDSB-201 previously reported for other systems. NDSB-201 also accelerates TBRII-ECD crystallization. Thus, NDSB-201 might be a useful crystallization additive if a protein has been folded with it. Moreover, our results suggest that TBRII contains a binding pocket that could be targeted by small molecule modulators.

## B.2 Introduction

Transforming growth factor beta (TGF-β) signaling is indispensable for cell growth and differentiation[273]. Signal transduction is initiated by binding of TGF-β to type II TGF-β receptor (TBRII)[274]. The TGF-β-TBRII complex then binds and phosphorylates

type I receptor (TBRI). This activated complex then phosphorylates SMAD proteins which then regulate expression of numerous genes[275]. The biological outcome of TGF-β signaling is modulated by numerous factors such as cell age, presence of other growth factors, and the cellular environment[276]. Thus, spatial and temporal control of TGF-β signaling is crucial to direct the physiological response for wound healing or cell differentiation[277].

Because of the countless roles of TGF-β signaling, chemical tools that modulate the receptor activities are valuable. For example, an inhibitor of TBRI kinase, SB431542, prevents cell migration crucial for cancer metastasis[278]. These diverse signaling outcomes led us to screen for peptide ligands that bind and modulate TBR activities. Using phage display, we discovered two peptides (LTGKNFPMFHRN and MHRMPSFLPTTL) that can bind both TBRI and TBRII[279]. Pre-organization of the receptors using immobilized peptide reduced the amount of required TGF-β for receptor activation to the pM range, and also allowed spatial control of cellular TGF-β signaling activation[280, 281]. To design the next generation of probes, the knowledge of peptide binding site and binding mode is required. These biophysical characterizations requires large amounts of correctly folded TBRs. However, existing protocols to produce active TBRII result in highly variable receptor yield[282-284]. Therefore, we developed an on-plate ELISA-based screen to optimize the folding conditions.

Using a thioredoxin-fused TBRII-ECD (Trx-TBRII-ECD) in the folding screen, non-detergent sulfobetaine-201 (NDSB-201) was found to be an effective additive for increasing active protein yield. NDSB-201 was also able to assist folding of protease-resistant, untagged TBRII-ECD (TBRII-ECD-PR) in solution. Because the mechanism of

NDSB-201-assisted folding is not well understood, we sought to further characterize the interactions. X-ray crystallography revealed the presence of a binding pocket for NDSB-201 on TBRII-ECD. These results suggest that NDSB-201 acts as a pharmacological chaperone that binds and stabilizes the folded state in addition to preventing protein aggregation[285-289].

## B.3 Material and Methods

### B.3.1 Expression of thioredoxin-fused TBRII (20-136) (Trx-TBRII-ECD)

The DNA sequence encoding TBRII-ECD (20-136) was amplified from a previously reported TBRII (15-136) expression plasmid[284] and ligated into the KpnI and NcoI sites of pET32b. The resulting construct contains a thioredoxin (Trx) tag at the N terminus of TBRII-ECD. The plasmid was transformed into Tuner (DE3) and grown at 37 °C to $OD_{600}$ of 0.6 in Terrific Broth (Research Products International Corp.). Temperature was reduced to 15 °C and protein expression induced by addition of isopropyl β-thiogalactopyranoside (IPTG) to 0.1 mM. Bacteria were harvested 14 h later by centrifugation. The pellet (25 g) was then resuspended in 50 mL of 50 mM sodium phosphate (pH 7.4), 300 mM NaCl, 0.1 mg/mL lysozyme, 2.5 mM benzamidine, 1 mM phenylmethylsulfonyl fluoride (PMSF) and 1 tablet of SIGMA*FAST*™ Protease Inhibitor Cocktail (EDTA-free), and lyzed by sonication. The lysate was clarified by centrifugation at 100,000xg for 1 h and filtration through a 0.45 μm membrane.

### B.3.2 On-plate folding assay

Lysate (200 μL) containing Trx-TBRII-ECD was added to Ni-NTA HisSorb plate (Qiagen) and incubated for 2 h. The plate was then washed 3 times for 5 min each with 200 μL of 50 mM sodium phosphate (pH 7.4), 300 mM NaCl, and 25 mM imidazole.

Folding solutions (200 µL) were added to the plate and incubated at 4 °C for 40 h. The plate was then washed with phosphate-buffered saline (PBS) containing 0.05% Tween 20 (PBST). Detection of correctly folded TBRII-ECD employed a previously reported assay to monitor TGF-β1 binding to the receptor[283]. Briefly, the plate was blocked for 1.5 h with 5% BSA in PBST (200 µL) at 37 °C. TGF-β1 solution (100 µL, 5 ng/mL, Cell Signaling Technology) was added and incubated for 2 h at room temperature. After washing, a solution of biotinylated anti- TGF-β1 (100 µL, 100 ng/mL, BAF240, R&D Systems) was added and incubated at room temperature for 1 hour. After washing, streptavidin-HRP solution (100 µL, 200 ng/mL, Jackson Immunoresearch) was added. After incubation for 20 min, the colorimetric substrate 1-Step Ultra TMB (100 µL, Pierce) was added. The reaction was stopped by addition of 2 M sulfuric acid (100 µL). Absorbance at 450 nm was measured as proportional to the amount of correctly folded TBRII-ECD.

### B.3.3 Folding and purification of TBRII (26-136) Q26A K97T (TBRII-ECD-PR)

The sequence encoding untagged TBRII (26-136) with protease-resistant mutations (Q26A and K97T) was cloned in between the NdeI and BamHI sites of pET24a. The expression plasmid was transformed into Tuner (DE3) and grown to $OD_{600}$ of 0.6 at 37 °C in Terrific Broth. Protein expression was induced by addition of IPTG to 1 mM. Cells were harvested after 6-8 h by centrifugation and kept frozen at -80 °C until use.

Extraction of TBRII-ECD-PR inclusion bodies was accomplished using a similar protocol used for TGF-β3 inclusion bodies[290]. Briefly, the bacterial pellets were resuspended in 100 mM Tris (pH 8.0), 10 mM EDTA, 0.1 mg/mL lysozyme, 2.5 mM

benzamidine, and 1 mM PMSF, and lysed by sonication. After centrifugation, the insoluble pellet was resuspended in 100 mM Tris (pH 8.0), 10 mM EDTA, and 1% Triton X-100 by sonication. The insoluble material was isolated by centrifugation and resuspended in 100 mM Tris (pH 8.0), 10 mM EDTA, and 1 M NaCl by sonication. The washed inclusion bodies was then pelleted and solubilized in 20 mM Tris (pH 8.0), 20 mM Bis-Tris, 100 mM DTT, and 8 M urea. The insoluble materials were removed by centrifugation, and the pH of the solution was adjusted to 6.0. The denatured protein was purified using an anion exchange column (HiTrap Q FF, GE Healthcare) equilibrated with 20 mM Bis-Tris (pH 6.0) and 8 M urea. TBRII-ECD-PR was eluted with a linear gradient of 0-300 mM NaCl. The protein concentration was measured by UV absorption at 280 nm using an extinction coefficient of 8480 $M^{-1}cm^{-1}$ [136]. The protein solution was stored at -80 °C until use.

To initiate folding of TBRII-ECD-PR, the urea solution containing TBRII-ECD-PR was added dropwise to an ice-cold solution of 75 mM Tris (pH 8.0), 1 M NDSB-201, 2 mM reduced glutathione (GSH), and 0.5 mM oxidized glutathione (GSSG) keeping the final protein concentration of 0.1 mg/mL. After stirring at 4 °C for 2 days, the solution was concentrated under nitrogen gas using Amicon stirred cell equipped with a 3 kDa cut-off membrane. The concentrated protein was dialyzed against 20 mM Bis-Tris (pH 6.0) and captured with an anion exchange column (HiTrap Q HP, GE Healthcare). TBRII-ECD-PR was eluted with a linear gradient of 0-300 mM NaCl (Figure B.1). Fractions containing correctly folded TBRII-ECD-PR was identified using the same TGF-β1 binding assay used in the on-plate folding screen but with TBRII-ECD-PR absorbed directly onto a Nunc MaxiSorp ELISA plate (Thermo Scientific). The desired

fractions were pooled and dialyzed against 50 mM sodium phosphate (pH 7.0). Ammonium sulfate was added to 1 M and minor impurities were captured with HiTrap Butyl HP column (GE Healthcare). The flow-through containing active TBRII-ECD-PR was further purified using size exclusion chromatography (Superdex 75 16/60) with 20 mM Tris (pH 7.4) and 300 mM NaCl as a running buffer.

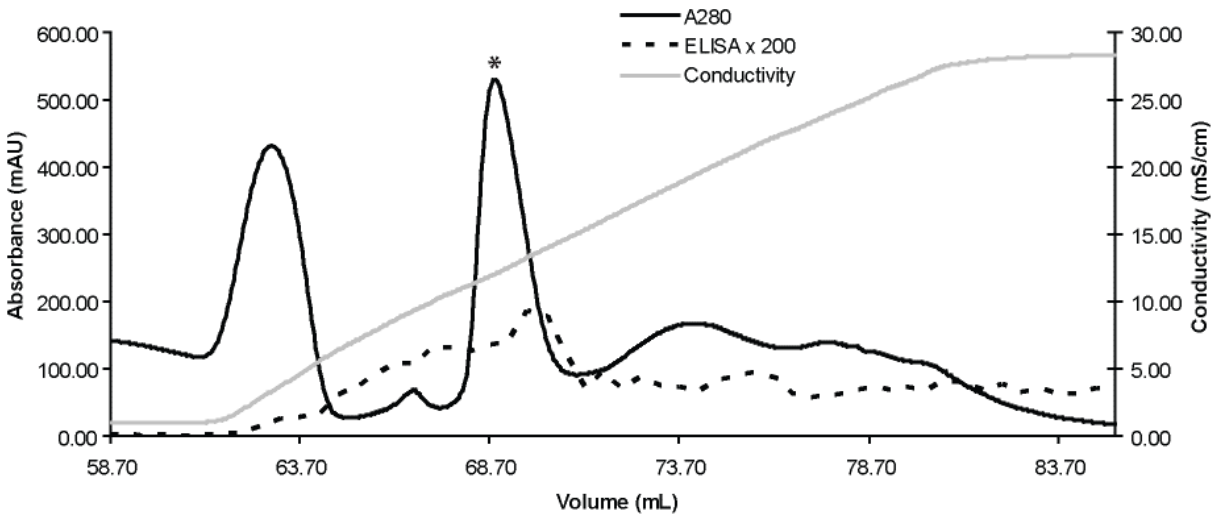



**Figure B.1**: Anion exchange chromatogram (HiTrap Q HP) of TBRII-ECD-PR. The peak corresponding to correctly folded TBRII-ECD-PR is marked with an asterisk.

### B.3.4 X-ray crystallography of TBRII-ECD-PR

Purified TBRII-ECD-PR was dialyzed against water and concentrated to 20 mg/mL. The final concentration was determined by UV absorption at 280 nm using an extinction coefficient of 9230 $M^{-1}cm^{-1}$ [136]. NDSB-201 was added to 50 mM and the protein crystallized using the previously reported conditions (100 mM sodium citrate (pH 5.0) and 30% PEG 2000)[291]. Crystals grew to full size in 2-3 days as opposed to 7-14 days without NDSB-201, consistent with that reported previously[291]. The crystals were cryoprotected with paratone, as opposed to glycerol in the previously reported procedure, and vitrified in liquid nitrogen. Single crystal diffraction data were collected

using a MAR 300 CCD detector at beamline 21-ID-D (LS-CAT), Advanced Photon Source, Argonne National Laboratory. Data reduction was performed with HKL2000[138] (Table B.1). Molecular replacement was performed with Phaser[140, 186] using a previously reported TBRII-ECD-PR structure[292] (PDB ID 1M9Z) as a search model. NDSB-201 coordinates and restraints were generated using eLBOW[250]. Refinement and model adjustment were performed with phenix.refine[141] and Coot[142] respectively (Table B.1). The structure was validated using MolProbity[144], and figures of protein structures were generated using PyMOL[145].

**Table B.1**: Data collection and refinement statistics for TBRII-ECD-PR in complex with NDSB-201.

| Data collection statistics | |
|---|---|
| Wavelength (Å) | 0.97919 |
| Resolution range (Å)* | 22.1 - 1.5 (1.55 - 1.5) |
| Space group | P 21 21 21 |
| Unit cell | 33.6 40.5 75.8 |
| Total reflections | 104,863 (10,083) |
| Unique reflections | 17,064 (1,653) |
| Multiplicity | 6.1 (6.1) |
| Completeness (%) | 99.5 (98.5) |
| Mean I/sigma(I) | 18.7 (3.9) |
| Wilson B-factor | 17.9 |
| R-merge | 0.054 (0.419) |
| R-meas | 0.060 (0.458) |
| R-pim | 0.024 (0.182) |
| Refinement Statistics | 21.4 - 1.5 (1.55 - 1.5) |
| R-factor | 0.172 (0.195) |
| R-free (10%) | 0.200 (0.240) |
| Number of atoms | 1002 |
| protein | 899 |
| ligand | 13 |
| water | 90 |
| Protein residues | 102 |
| RMSD (bonds, Å) | 0.006 |
| RMSD (angles, Å) | 1.17 |
| Ramachandran favored (%) | 98 |
| Ramachandran outliers (%) | 0 |
| Average B-factor | 22.3 |
| protein | 20.7 |
| ligand | 39.1 |
| solvent | 35.2 |

*Statistics for the highest-resolution shell are shown in parentheses. The coordinates and structure factors have been deposited to the Protein Data Bank under accession code 4P7U.

## B.4 Results

### B.4.1 Folding screen for TBRII-ECD

Our attempts to fold TBRII-ECD from inclusion bodies using previously reported conditions[282, 284, 293] afford highly variable yields (often less than 1 mg from 50 mg of solubilized inclusion bodies). Chicken Trx-TBRII-ECD was shown to be functional, but human Trx-TBRII-ECD cannot bind TGF-β1 despite being soluble[294, 295]. Human Trx-TBRII-ECD has been folded on Ni-NTA[283]; however, a high concentration of arginine at pH 8.0 interferes with protein binding to Ni-NTA[296]. Inspired by this on-resin folding strategy[283], we developed an on-plate folding screen to rapidly optimize TBRII-ECD folding conditions. In this assay, soluble Trx-TBRII-ECD was immobilized onto a Ni-NTA-coated 96-well plate (Figure B.2). Folding buffer was then applied to each well. In total, 81 conditions were screened. Each buffer contains Tris buffer, one additive, and one redox couple. The pH was maintained at 8.0, close to $pK_a$ of cysteine thiol, to facilitate disulfide bond formation. The additives examined were chosen based on literature precedence[297-300] (Table B.2). To quantify the amount of correctly folded TBRII-ECD, a functional assay that involved monitoring the amount of bound TGF-β1 using a TGF-β1-specific antibody.

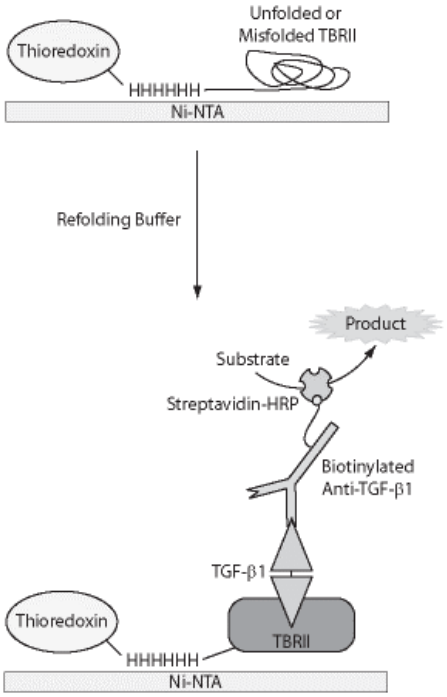



**Figure B.2**: ELISA scheme for the on-plate folding assay to detect correctly folded Trx-TBRII-ECD

**Table B.2**: Composition of buffers used in the on-plate folding screen.

| Buffer | Additives | Redox buffer |
|---|---|---|
| 75 mM Tris pH 8.0 | Sorbitol (0.5, 1, 2 M) | 2 mM GSH + 0.5 mM GSSG |
| | Urea (1, 2, 4 M) | 2 mM GSH + 0.2 mM GSSG |
| | NaCl (100, 250 mM) | 2 mM GSH + 0.05 mM GSSG |
| | PEG 3350 (0.05, 0.1, 0.5 %) | |
| | PEG 6000 (0.05, 0.1, 0.5 %) | |
| | $MgCl_2$+$CaCl_2$ (1 mM each) | |
| | α-Cyclodextrin (5, 10 mM) | |
| | β-Cyclodextrin (5, 10 mM) | |
| | NDSB-195 (0.5, 1 M) | |
| | NDSB-201 (0.5, 1 M) | |
| | NDSB-221 (0.5, 1 M) | |
| | NDSB-256 (0.5, 1 M) | |

A preliminary folding screen indicated that urea, NDSB-201, or NDSB-256 each serve as an effective additive (Figure B.3). We then held the redox couple constant (2 mM GSH + 0.5 mM GSSG) and repeated the folding screen in triplicate (Figure B.4). Using the level of folding with the redox buffer alone as a baseline, the effect of each additive was analyzed. When either urea, NDSB-201, or NDSB-256 was added, high levels of correctly folded Trx-TBRII-ECD was observed.

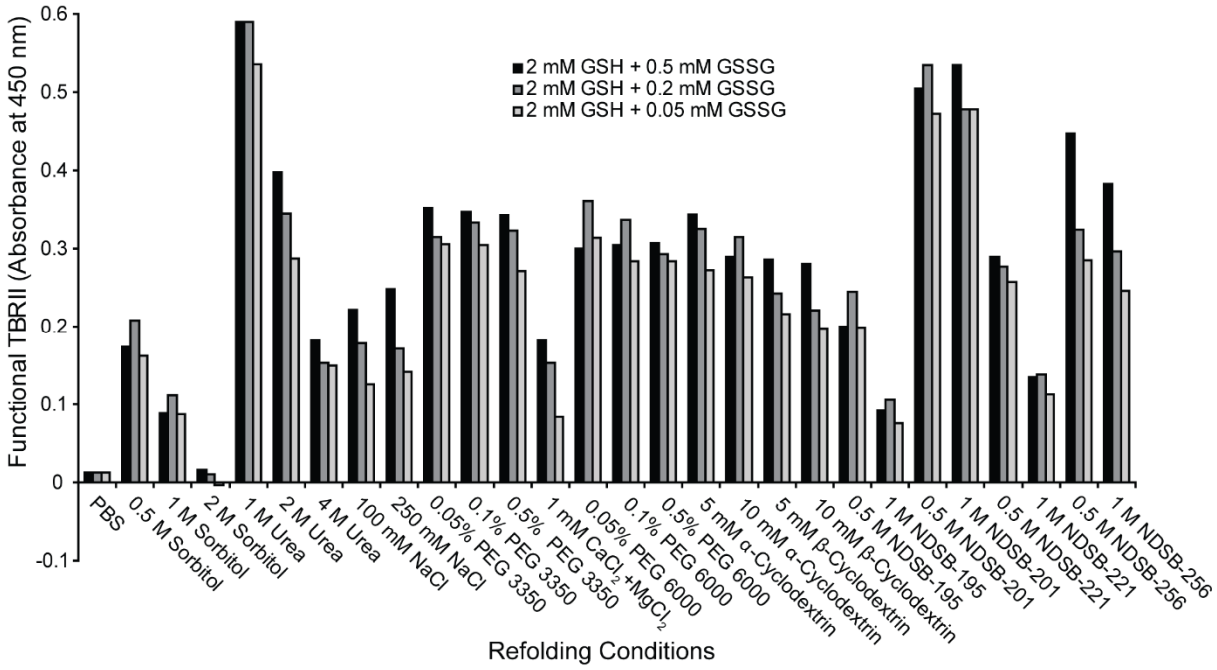



**Figure B.3**: Preliminary screen for folding conditions of Trx-TBRII-ECD. Relative amount of correctly folded protein was measured by absorbance reading of the ELISA product.

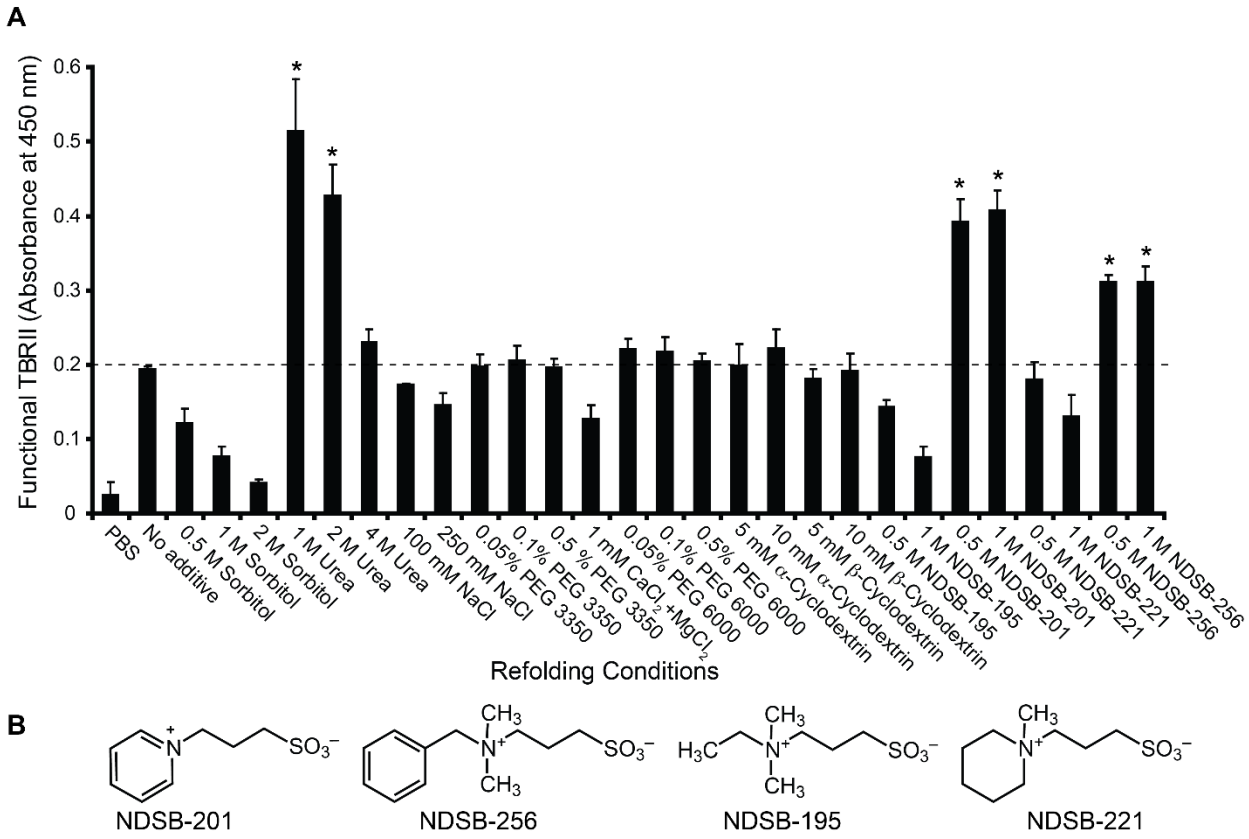



**Figure B.4**: (A) Trx-TBRII-ECD folding screen using a constant redox buffer (2 mM GSH + 0.5 mM GSSG). Error bars indicate standard deviation of triplicate absorbance measurements. Dashed line indicates the level of active Trx-TBRII-ECD when no additive is present, setting the baseline for effective additives. (B) Structures of NDSBs tested.

Although urea was an effective additive, we further investigated NDSBs because their mechanism of action is poorly understood. We postulated that there exists a specific binding site for NDSB-201 and NDSB-256 on TBRII-ECD because all NDSBs are zwitterionic, but not all NDSBs improved folding in our assay. Because NDSB-201 was more cost effective than NDSB-256, NDSB-201 was employed in larger scale solution phase folding experiments of untagged protease-resistant TBRII (TBRII-ECD-PR) previously used for structure determination[292]. The optimized conditions are 75 mM Tris (pH 8.0), 1 M NDSB-201, 2 mM GSH, and 0.5 mM GSSG for 40 h at 4 °C. This

protocol yielded 8-13 mg of purified TBRII-ECD-PR from 50 mg of urea-solubilized protein. This protocol was successful for both Trx-TBRII-ECD and TBRII-ECD-PR suggesting that the conditions found in the on-plate screen is applicable for folding in solution. In addition, we conclude that protease-resistant mutations in TBRII-ECD-PR do not affect folding significantly.

### B.4.2 Crystal structure of TBRII-ECD-PR in complex with NDSB-201

To examine whether TBRII-ECD contains a binding site for NDSB-201, we crystallized TBRII-ECD-PR using previously reported conditions[291]. When NDSB-201 was added, protein crystals grew and reached full size within 2-3 days as opposed to 1-2 weeks without NDSB-201. Crystallographic analysis revealed a molecule of NDSB-201 bound to TBRII-ECD-PR (Figure B.5). The binding site is a hydrophobic pocket formed by W65, F111, F126, and a disulfide bridge formed by C38 and C44. The cationic pyridinium ring of NDSB-201 stacks onto the phenyl ring π face of F126 in a staggered conformation. The pyridinium ring also interacts with W65 in an edge-on manner. The edge of the binding pocket also contains several hydrophilic residues (D39, N40, K67, E102, and T109). However, their side chains are more than 4 Å away from the sulfonate group of NDSB-201. Thus, these residues may only minimally contribute to NDSB-201 binding affinity. The total buried surface area of the complex is 375 $Å^2$ as determined by PISA[301]. These results suggest favorable and specific interactions between NDSB-201 and TBRII-ECD.

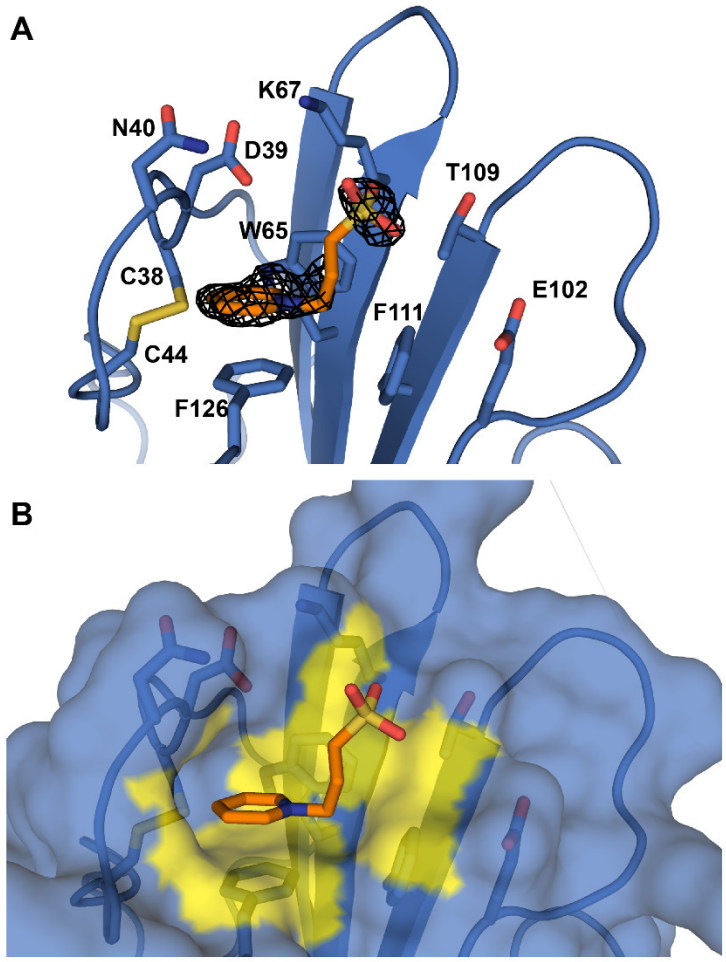



**Figure B.5**: (A) Structure of TBRII-ECD-PR (blue) with bound NDSB-201 (orange). The black mesh represents a $mF_o$-$DF_c$ map (3.0 σ) that was calculated after the ligand was removed and the structure re-refined. (B) Surface representation of the complex. TBRII-ECD-PR atoms that are within 4 Å of the ligand are highlighted in yellow.

**B.5 Discussion and Conclusions**

The on-plate assay allowed rapid identification of conditions to promote folding of Trx-TBRII-ECD. The conditions found were also effective for folding untagged TBRII (TBRII-ECD-PR) in solution. The assay provided information about the chemical sensitivity of TBRII-ECD during the folding process. For example, a low concentration of urea was highly effective at folding TBRII-ECD, but the efficiency decreased with increasing urea concentration. These results highlight the importance of controlling urea

concentration when applying a previously reported protocol[284]. Because the screen covers expansive chemical space of known additives, it could be adapted for other monomeric receptors or enzymes for which an activity assay for functional protein is available.

Because urea and NDSBs are aggregation suppressors[302, 303], their effectiveness as folding additives indicates that a major bottleneck of TBRII-ECD folding is aggregation. However, NDSB-201 and NDSB-256 outperformed other NDSBs, suggesting a specific binding site on the receptor. NDSB-201 and NDSB-256 contain pyridinium and benzyl groups, respectively, that can stack onto aromatic amino acids on the receptor[304]. NDSB-192 and NDSB-221, which lack an aromatic substituent, are less effective as a folding additive. Crystallographic analysis showed that NDSB-201 bind TBRII-ECD-PR in a shallow hydrophobic pocket. The pyridinium ring stacks onto aromatic phenyl alanine in the binding site. This binding mode may stabilize the correctly folded product in addition to preventing protein aggregation during the folding process.

To date, there are six protein structures in complex with NDSB-201 in the Protein Data Bank[305-310]. In each of them, NDSB-201 interacts with aromatic amino acids. However, NDSB-201 was used as a crystallization additive or a cryprotectant after the correctly folded protein has been obtained by other means without using NDSB-201 as a folding additive. Our results suggest that NDSB-201 could be used as both a folding agent and crystallization additive. We postulate that it will be advantageous to include NDSB-201 in crystallization trials if a protein has been folded or refolded with NDSB-201.

The presence of a small molecule binding pocket on TBRII-ECD suggests that this pocket could be targeted to generate probes for TGF-β signaling. Alignment of the NDSB-201-bound structure with the structure of the complex between TBRII-ECD, TBRI-ECD and TGF-β3 (PDB ID 2PJY, Figure B.6) reveals that the NDSB-201 binding site on TBRII-ECD does not overlap with the TBRI-ECD and TGF-β3 binding surface. This observation suggests that NDSB-201 may not interfere with the signaling complex formation. However, as demonstrated previously, molecules that bind in a remote site on the TGF-β signaling complex can still modulate TGF-β signaling[279, 280]. Our research not only suggests an alternative folding mechanism of TBRII-ECD by NDSB-201, but also reveals a potential pocket on TBRII-ECD that could be exploited for chemical modulator development.

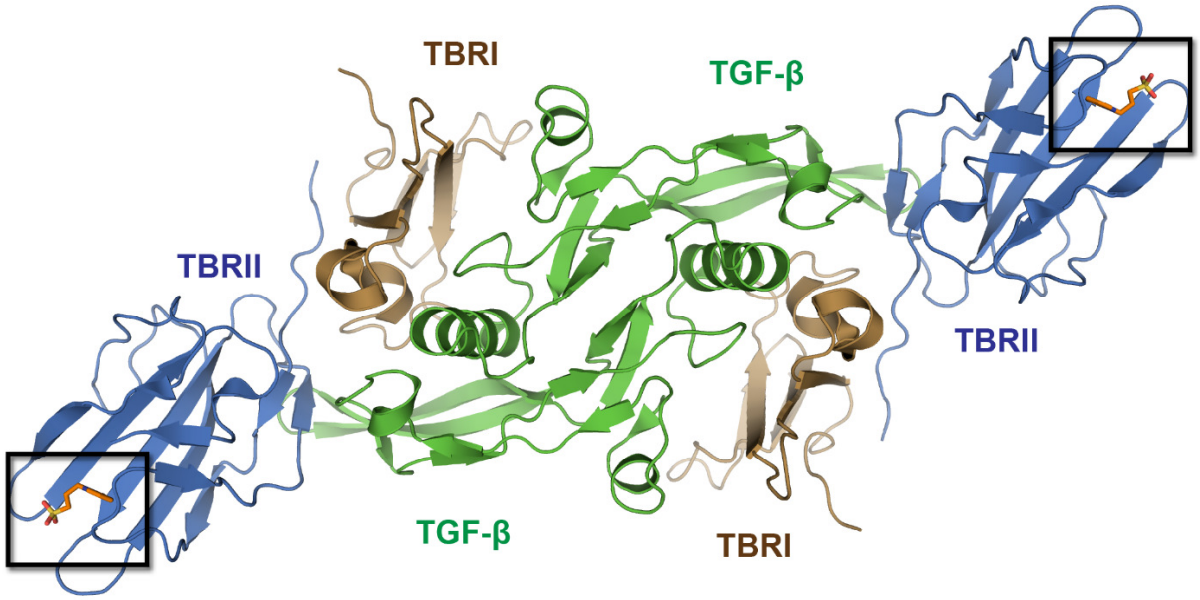



**Figure B.6**: Alignment of the TBRII-ECD-PR-NDSB-201 complex onto a structure of the TGF-β signaling complex. The location of NDSB-201 is indicated by the boxes.

## References


1. Adibekian, A., Stallforth, P., Hecht, M. L., Werz, D. B., Gagneux, P., and Seeberger, P. H. (2011) Comparative bioinformatics analysis of the mammalian and bacterial glycomes, *Chem. Sci. 2*, 337-344.

2. Herget, S., Toukach, P. V., Ranzinger, R., Hull, W. E., Knirel, Y. A., and von der Lieth, C. W. (2008) Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans, *BMC Struct. Biol. 8*, 35.

3. Varki A, C. R., Esko JD, et al. (2009) *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press.

4. Weinbaum, S., Tarbell, J. M., and Damiano, E. R. (2007) The structure and function of the endothelial glycocalyx layer, *Annu Rev Biomed Eng 9*, 121-167.

5. Yanagishita, M. (1993) Function of proteoglycans in the extracellular matrix, *Acta Pathol Jpn 43*, 283-293.

6. Schwartz, N. B., and Domowicz, M. (2002) Chondrodysplasias due to proteoglycan defects, *Glycobiology 12*, 57r-68r.

7. Stattin, E. L., Wiklund, F., Lindblom, K., Onnerfjord, P., Jonsson, B. A., Tegner, Y., Sasaki, T., Struglics, A., Lohmander, S., Dahl, N., Heinegard, D., and Aspberg, A. (2010) A missense mutation in the aggrecan C-type lectin domain disrupts extracellular matrix interactions and causes dominant familial osteochondritis dissecans, *Am J Hum Genet 86*, 126-137.

8. Weledji, E. P., and Assob, J. C. (2014) The ubiquitous neural cell adhesion molecule (N-CAM), *Ann Med Surg (Lond) 3*, 77-81.

9. Rapraeger, A. C. (1995) In the clutches of proteoglycans: how does heparan sulfate regulate FGF binding?, *Chem Biol 2*, 645-649.

10. Yan, D., and Lin, X. (2009) Shaping morphogen gradients by proteoglycans, *Cold Spring Harb Perspect Biol 1*, a002493.

11. Molinari, M. (2007) N-glycan structure dictates extension of protein folding or onset of disposal, *Nat Chem Biol 3*, 313-320.

12. Caramelo, J. J., and Parodi, A. J. (2015) A sweet code for glycoprotein folding, *FEBS Lett 589*, 3379-3387.

13. Cummings, R. D., and Smith, D. F. (1992) The selectin family of carbohydrate-binding proteins: structure and importance of carbohydrate ligands for cell adhesion, *Bioessays 14*, 849-856.

14. Clemetson, K. J., Lu, Q., and Clemetson, J. M. (2005) Snake C-type lectin-like proteins and platelet receptors, *Pathophysiol Haemost Thromb 34*, 150-155.

15. Figdor, C. G., van Kooyk, Y., and Adema, G. J. (2002) C-type lectin receptors on dendritic cells and Langerhans cells, *Nat Rev Immunol 2*, 77-84.

16. Gijzen, K., Cambi, A., Torensma, R., and Figdor, C. G. (2006) C-type lectins on dendritic cells and their interaction with pathogen-derived and endogenous glycoconjugates, *Curr Protein Pept Sci 7*, 283-294.

17. Curciarello, R., Steele, A., Cooper, D., MacDonald, T. T., Kruidenier, L., and Kudo, T. (2014) The role of Galectin-1 and Galectin-3 in the mucosal immune response to Citrobacter rodentium infection, *PLoS One 9*, e107933.

18. Fujita, T., Matsushita, M., and Endo, Y. (2004) The lectin-complement pathway--its role in innate immunity and evolution, *Immunol Rev 198*, 185-202.

19. Cooper, D. N., and Barondes, S. H. (1999) God must love galectins; he made so many of them, *Glycobiology 9*, 979-984.

20. Waisman, D. M. (1995) Annexin II tetramer: structure and function, *Mol Cell Biochem 149-150*, 301-322.

21. Gerke, V., Creutz, C. E., and Moss, S. E. (2005) Annexins: linking Ca2+ signalling to membrane dynamics, *Nat Rev Mol Cell Biol 6*, 449-461.

22. Kassam, G., Manro, A., Braat, C. E., Louie, P., Fitzpatrick, S. L., and Waisman, D. M. (1997) Characterization of the heparin binding properties of annexin II tetramer, *J Biol Chem 272*, 15093-15100.

23. Hajjar, K. A., and Acharya, S. S. (2000) Annexin II and regulation of cell surface fibrinolysis, *Ann N Y Acad Sci 902*, 265-271.

24. Shao, C., Zhang, F., Kemp, M. M., Linhardt, R. J., Waisman, D. M., Head, J. F., and Seaton, B. A. (2006) Crystallographic analysis of calcium-dependent heparin binding to annexin A2, *J Biol Chem 281*, 31689-31695.

25. Williams, D. B. (2006) Beyond lectins: the calnexin/calreticulin chaperone system of the endoplasmic reticulum, *J Cell Sci 119*, 615-623.

26. Parodi, A. J. (2000) Role of N-oligosaccharide endoplasmic reticulum processing reactions in glycoprotein folding and degradation, *Biochem J 348 Pt 1*, 1-13.

27. Ellgaard, L., and Helenius, A. (2003) Quality control in the endoplasmic reticulum, *Nat Rev Mol Cell Biol 4*, 181-191.

28. Schrag, J. D., Bergeron, J. J., Li, Y., Borisova, S., Hahn, M., Thomas, D. Y., and Cygler, M. (2001) The Structure of calnexin, an ER chaperone involved in quality control of protein folding, *Mol Cell 8*, 633-644.

29. Kozlov, G., Pocanschi, C. L., Rosenauer, A., Bastos-Aristizabal, S., Gorelik, A., Williams, D. B., and Gehring, K. (2010) Structural basis of carbohydrate recognition by calreticulin, *J Biol Chem 285*, 38612-38620.

30. Lee, C. G., Da Silva, C. A., Dela Cruz, C. S., Ahangari, F., Ma, B., Kang, M. J., He, C. H., Takyar, S., and Elias, J. A. (2011) Role of chitin and chitinase/chitinase-like proteins in inflammation, tissue remodeling, and injury, *Annu Rev Physiol 73*, 479-501.

31. Bigg, H. F., Wait, R., Rowan, A. D., and Cawston, T. E. (2006) The mammalian chitinase-like lectin, YKL-40, binds specifically to type I collagen and modulates the rate of type I collagen fibril formation, *J Biol Chem 281*, 21082-21095.

32. Chang, N. C., Hung, S. I., Hwa, K. Y., Kato, I., Chen, J. E., Liu, C. H., and Chang, A. C. (2001) A macrophage protein, Ym1, transiently expressed during inflammation is a novel mammalian lectin, *J Biol Chem 276*, 17497-17506.

33. Malette, B., Paquette, Y., Merlen, Y., and Bleau, G. (1995) Oviductins possess chitinase- and mucin-like domains: a lead in the search for the biological function of these oviduct-specific ZP-associating glycoproteins, *Mol Reprod Dev 41*, 384-397.

34. Sun, Y. J., Chang, N. C., Hung, S. I., Chang, A. C., Chou, C. C., and Hsiao, C. D. (2001) The crystal structure of a novel mammalian lectin, Ym1, suggests a saccharide binding site, *J Biol Chem 276*, 17507-17514.

35. Houston, D. R., Recklies, A. D., Krupa, J. C., and van Aalten, D. M. (2003) Structure and ligand-induced conformational change of the 39-kDa glycoprotein from human articular chondrocytes, *J Biol Chem 278*, 30206-30212.

36. Drickamer, K. (1999) C-type lectin-like domains, *Curr Opin Struct Biol 9*, 585-590.

37. Drickamer, K., and Taylor, M. E. (2015) Recent insights into structures and functions of C-type lectins in the immune system, *Curr Opin Struct Biol 34*, 26-34.

38. Zelensky, A. N., and Gready, J. E. (2005) The C-type lectin-like domain superfamily, *Febs j 272*, 6179-6217.

39. Weis, W. I., Taylor, M. E., and Drickamer, K. (1998) The C-type lectin superfamily in the immune system, *Immunol Rev 163*, 19-34.

40. Guo, Y., Feinberg, H., Conroy, E., Mitchell, D. A., Alvarez, R., Blixt, O., Taylor, M. E., Weis, W. I., and Drickamer, K. (2004) Structural basis for distinct ligand-

binding and targeting properties of the receptors DC-SIGN and DC-SIGNR, *Nat Struct Mol Biol 11*, 591-598.

41. Skaar, J. R., Pagan, J. K., and Pagano, M. (2013) Mechanisms and function of substrate recruitment by F-box proteins, *Nat Rev Mol Cell Biol 14*, 369-381.

42. Yoshida, Y., Adachi, E., Fukiya, K., Iwai, K., and Tanaka, K. (2005) Glycoprotein-specific ubiquitin ligases recognize N-glycans in unfolded substrates, *EMBO Rep 6*, 239-244.

43. Yoshida, Y., Tokunaga, F., Chiba, T., Iwai, K., Tanaka, K., and Tai, T. (2003) Fbs2 is a new member of the E3 ubiquitin ligase family that recognizes sugar chains, *J Biol Chem 278*, 43877-43884.

44. Mizushima, T., Hirao, T., Yoshida, Y., Lee, S. J., Chiba, T., Iwai, K., Yamaguchi, Y., Kato, K., Tsukihara, T., and Tanaka, K. (2004) Structural basis of sugar-recognizing ubiquitin ligase, *Nat Struct Mol Biol 11*, 365-370.

45. Matsushita, M., and Fujita, T. (2002) The role of ficolins in innate immunity, *Immunobiology 205*, 490-497.

46. Matsushita, M., and Fujita, T. (2001) Ficolins and the lectin complement pathway, *Immunol Rev 180*, 78-85.

47. Zhang, X. L., and Ali, M. A. (2008) Ficolins: structure, function and associated diseases, *Adv Exp Med Biol 632*, 105-115.

48. Krarup, A., Thiel, S., Hansen, A., Fujita, T., and Jensenius, J. C. (2004) L-ficolin is a pattern recognition molecule specific for acetyl groups, *J Biol Chem 279*, 47513-47519.

49. Zacho, R. M., Jensen, L., Terp, R., Jensenius, J. C., and Thiel, S. (2012) Studies of the pattern recognition molecule H-ficolin: specificity and purification, *J Biol Chem 287*, 8071-8081.

50. Teh, C., Le, Y., Lee, S. H., and Lu, J. (2000) M-ficolin is expressed on monocytes and is a lectin binding to N-acetyl-D-glucosamine and mediates monocyte adhesion and phagocytosis of Escherichia coli, *Immunology 101*, 225-232.

51. Liu, Y., Endo, Y., Iwaki, D., Nakata, M., Matsushita, M., Wada, I., Inoue, K., Munakata, M., and Fujita, T. (2005) Human M-ficolin is a secretory protein that activates the lectin complement pathway, *J Immunol 175*, 3150-3156.

52. Fujita, T. (2002) Evolution of the lectin-complement pathway and its role in innate immunity, *Nat Rev Immunol 2*, 346-353.

53. Lu, J., and Le, Y. (1998) Ficolins and the fibrinogen-like domain, *Immunobiology 199*, 190-199.

54. Garlatti, V., Belloy, N., Martin, L., Lacroix, M., Matsushita, M., Endo, Y., Fujita, T., Fontecilla-Camps, J. C., Arlaud, G. J., Thielens, N. M., and Gaboriaud, C. (2007) Structural insights into the innate immune recognition specificities of L- and H-ficolins, *Embo j 26*, 623-633.

55. Bishnoi, R., Khatri, I., Subramanian, S., and Ramya, T. N. (2015) Prevalence of the F-type lectin domain, *Glycobiology 25*, 888-901.

56. Vasta, G. R., Ahmed, H., Bianchet, M. A., Fernandez-Robledo, J. A., and Amzel, L. M. (2012) Diversity in recognition of glycans by F-type lectins and galectins: molecular, structural, and biophysical aspects, *Ann N Y Acad Sci 1253*, E14-26.

57. Chen, J., Xiao, S., and Yu, Z. (2011) F-type lectin involved in defense against bacterial infection in the pearl oyster (Pinctada martensii), *Fish Shellfish Immunol 30*, 750-754.

58. Odom, E. W., and Vasta, G. R. (2006) Characterization of a binary tandem domain F-type lectin from striped bass (Morone saxatilis), *J Biol Chem 281*, 1698-1713.

59. Bianchet, M. A., Odom, E. W., Vasta, G. R., and Amzel, L. M. (2010) Structure and specificity of a binary tandem domain F-lectin from striped bass (Morone saxatilis), *J Mol Biol 401*, 239-252.

60. Honda, S., Kashiwagi, M., Miyamoto, K., Takei, Y., and Hirose, S. (2000) Multiplicity, structures, and endocrine and exocrine natures of eel fucose-binding lectins, *J Biol Chem 275*, 33151-33157.

61. Baldus, S. E., Thiele, J., Park, Y. O., Hanisch, F. G., Bara, J., and Fischer, R. (1996) Characterization of the binding specificity of Anguilla anguilla agglutinin (AAA) in comparison to Ulex europaeus agglutinin I (UEA-I), *Glycoconj J 13*, 585-590.

62. Bianchet, M. A., Odom, E. W., Vasta, G. R., and Amzel, L. M. (2002) A novel fucose recognition fold involved in innate immunity, *Nat Struct Biol 9*, 628-634.

63. Elola, M. T., Wolfenstein-Todel, C., Troncoso, M. F., Vasta, G. R., and Rabinovich, G. A. (2007) Galectins: matricellular glycan-binding proteins linking cell adhesion, migration, and survival, *Cell Mol Life Sci 64*, 1679-1700.

64. Compagno, D., Jaworski, F. M., Gentilini, L., Contrufo, G., Gonzalez Perez, I., Elola, M. T., Pregi, N., Rabinovich, G. A., and Laderach, D. J. (2014) Galectins: major signaling modulators inside and outside the cell, *Curr Mol Med 14*, 630-651.

65. Liu, F. T., Patterson, R. J., and Wang, J. L. (2002) Intracellular functions of galectins, *Biochim Biophys Acta 1572*, 263-273.

66. Liu, F. T., Yang, R. Y., Saegusa, J., Chen, H. Y., and Hsu, D. K. (2011) Galectins in regulation of apoptosis, *Adv Exp Med Biol 705*, 431-442.

67. Rabinovich, G. A., and Toscano, M. A. (2009) Turning 'sweet' on immunity: galectin-glycan interactions in immune tolerance and inflammation, *Nat Rev Immunol 9*, 338-352.

68. Liu, F. T., and Rabinovich, G. A. (2005) Galectins as modulators of tumour progression, *Nat Rev Cancer 5*, 29-41.

69. Thijssen, V. L., Heusschen, R., Caers, J., and Griffioen, A. W. (2015) Galectin expression in cancer diagnosis and prognosis: A systematic review, *Biochim Biophys Acta 1855*, 235-247.

70. Saraboji, K., Hakansson, M., Genheden, S., Diehl, C., Qvist, J., Weininger, U., Nilsson, U. J., Leffler, H., Ryde, U., Akke, M., and Logan, D. T. (2012) The carbohydrate-binding site in galectin-3 is preorganized to recognize a sugarlike framework of oxygens: ultra-high-resolution structures and water dynamics, *Biochemistry 51*, 296-306.

71. Day, A. J., and Prestwich, G. D. (2002) Hyaluronan-binding proteins: tying up the giant, *J Biol Chem 277*, 4585-4588.

72. Knudson, C. B., and Knudson, W. (1993) Hyaluronan-binding proteins in development, tissue homeostasis, and disease, *Faseb j 7*, 1233-1241.

73. Milner, C. M., and Day, A. J. (2003) TSG-6: a multifunctional protein associated with inflammation, *J Cell Sci 116*, 1863-1873.

74. Blundell, C. D., Mahoney, D. J., Almond, A., DeAngelis, P. L., Kahmann, J. D., Teriete, P., Pickford, A. R., Campbell, I. D., and Day, A. J. (2003) The link module from ovulation- and inflammation-associated protein TSG-6 changes conformation on hyaluronan binding, *J Biol Chem 278*, 49261-49270.

75. Mahoney, D. J., Blundell, C. D., and Day, A. J. (2001) Mapping the hyaluronan-binding site on the link module from human tumor necrosis factor-stimulated gene-6 by site-directed mutagenesis, *J Biol Chem 276*, 22764-22771.

76. Crocker, P. R., Clark, E. A., Filbin, M., Gordon, S., Jones, Y., Kehrl, J. H., Kelm, S., Le Douarin, N., Powell, L., Roder, J., Schnaar, R. L., Sgroi, D. C., Stamenkovic, K., Schauer, R., Schachner, M., van den Berg, T. K., van der Merwe, P. A., Watt, S. M., and Varki, A. (1998) Siglecs: a family of sialic-acid binding lectins, *Glycobiology 8*, v.

77. Crocker, P. R., and Varki, A. (2001) Siglecs, sialic acids and innate immunity, *Trends Immunol 22*, 337-342.

78. Brinkman-Van der Linden, E. C., Angata, T., Reynolds, S. A., Powell, L. D., Hedrick, S. M., and Varki, A. (2003) CD33/Siglec-3 binding specificity, expression pattern, and consequences of gene deletion in mice, *Mol Cell Biol 23*, 4199-4206.

79. Nitschke, L. (2005) The role of CD22 and other inhibitory co-receptors in B-cell activation, *Curr Opin Immunol 17*, 290-297.

80. Yamaji, T., Teranishi, T., Alphey, M. S., Crocker, P. R., and Hashimoto, Y. (2002) A small region of the natural killer cell receptor, Siglec-7, is responsible for its preferred binding to alpha 2,8-disialyl and branched alpha 2,6-sialyl residues. A comparison with Siglec-9, *J Biol Chem 277*, 6324-6332.

81. Avril, T., Floyd, H., Lopez, F., Vivier, E., and Crocker, P. R. (2004) The membrane-proximal immunoreceptor tyrosine-based inhibitory motif is critical for the inhibitory signaling mediated by Siglecs-7 and -9, CD33-related Siglecs expressed on human monocytes and NK cells, *J Immunol 173*, 6841-6849.

82. Attrill, H., Imamura, A., Sharma, R. S., Kiso, M., Crocker, P. R., and van Aalten, D. M. (2006) Siglec-7 undergoes a major conformational change when complexed with the alpha(2,8)-disialylganglioside GT1b, *J Biol Chem 281*, 32774-32783.

83. Sharon, N., and Lis, H. (1990) Legume lectins--a large family of homologous proteins, *Faseb j 4*, 3198-3208.

84. Fiedler, K., and Simons, K. (1994) A putative novel class of animal lectins in the secretory pathway homologous to leguminous lectins, *Cell 77*, 625-626.

85. Yerushalmi, N., Keppler-Hafkemeyer, A., Vasmatzis, G., Liu, X. F., Olsson, P., Bera, T. K., Duray, P., Lee, B., and Pastan, I. (2001) ERGL, a novel gene related to ERGIC-53 that is highly expressed in normal and neoplastic prostate and several other tissues, *Gene 265*, 55-60.

86. Nufer, O., Mitrovic, S., and Hauri, H. P. (2003) Profile-based data base scanning for animal L-type lectins and characterization of VIPL, a novel VIP36-like endoplasmic reticulum protein, *J Biol Chem 278*, 15886-15896.

87. Kamiya, Y., Yamaguchi, Y., Takahashi, N., Arata, Y., Kasai, K., Ihara, Y., Matsuo, I., Ito, Y., Yamamoto, K., and Kato, K. (2005) Sugar-binding properties of VIP36, an intracellular animal lectin operating as a cargo receptor, *J Biol Chem 280*, 37178-37182.

88. Hauri, H. P., Kappeler, F., Andersson, H., and Appenzeller, C. (2000) ERGIC-53 and traffic in the secretory pathway, *J Cell Sci 113 ( Pt 4)*, 587-596.

89. Hosokawa, N., Wada, I., Hasegawa, K., Yorihuzi, T., Tremblay, L. O., Herscovics, A., and Nagata, K. (2001) A novel ER alpha-mannosidase-like protein accelerates ER-associated degradation, *EMBO Rep 2*, 415-422.

90. Braakman, I. (2001) A novel lectin in the secretory pathway. An elegant mechanism for glycoprotein elimination, *EMBO Rep 2*, 666-668.

91. Mast, S. W., Diekman, K., Karaveg, K., Davis, A., Sifers, R. N., and Moremen, K. W. (2005) Human EDEM2, a novel homolog of family 47 glycosidases, is involved in ER-associated degradation of glycoproteins, *Glycobiology 15*, 421-436.

92. Olivari, S., Galli, C., Alanen, H., Ruddock, L., and Molinari, M. (2005) A novel stress-induced EDEM variant regulating endoplasmic reticulum-associated glycoprotein degradation, *J Biol Chem 280*, 2424-2428.

93. Hirao, K., Natsuka, Y., Tamura, T., Wada, I., Morito, D., Natsuka, S., Romero, P., Sleno, B., Tremblay, L. O., Herscovics, A., Nagata, K., and Hosokawa, N. (2006) EDEM3, a soluble EDEM homolog, enhances glycoprotein endoplasmic reticulum-associated degradation and mannose trimming, *J Biol Chem 281*, 9650-9658.

94. Karaveg, K., Siriwardena, A., Tempel, W., Liu, Z. J., Glushka, J., Wang, B. C., and Moremen, K. W. (2005) Mechanism of class 1 (glycosylhydrolase family 47) {alpha}-mannosidases involved in N-glycan processing and endoplasmic reticulum quality control, *J Biol Chem 280*, 16197-16207.

95. Song, X., Lasanajak, Y., Olson, L. J., Boonen, M., Dahms, N. M., Kornfeld, S., Cummings, R. D., and Smith, D. F. (2009) Glycan microarray analysis of P-type lectins reveals distinct phosphomannose glycan recognition, *J Biol Chem 284*, 35201-35214.

96. Dahms, N. M., Lobel, P., and Kornfeld, S. (1989) Mannose 6-phosphate receptors and lysosomal enzyme targeting, *J Biol Chem 264*, 12115-12118.

97. Roberts, D. L., Weix, D. J., Dahms, N. M., and Kim, J. J. (1998) Molecular basis of lysosomal enzyme recognition: three-dimensional structure of the cation-dependent mannose 6-phosphate receptor, *Cell 93*, 639-648.

98. Olson, L. J., Dahms, N. M., and Kim, J. J. (2004) The N-terminal carbohydrate recognition site of the cation-independent mannose 6-phosphate receptor, *J Biol Chem 279*, 34000-34009.

99. Rutenber, E., and Robertus, J. D. (1991) Structure of ricin B-chain at 2.5 A resolution, *Proteins 10*, 260-269.

100. Raman, J., Fritz, T. A., Gerken, T. A., Jamison, O., Live, D., Liu, M., and Tabak, L. A. (2008) The catalytic and lectin domains of UDP-GalNAc:polypeptide alpha-

N-Acetylgalactosaminyltransferase function in concert to direct glycosylation site selection, *J Biol Chem 283*, 22942-22951.

101. Fiete, D. J., Beranek, M. C., and Baenziger, J. U. (1998) A cysteine-rich domain of the "mannose" receptor mediates GalNAc-4-SO4 binding, *Proc Natl Acad Sci U S A 95*, 2089-2093.

102. Liu, Y., Chirino, A. J., Misulovin, Z., Leteux, C., Feizi, T., Nussenzweig, M. C., and Bjorkman, P. J. (2000) Crystal structure of the cysteine-rich domain of mannose receptor complexed with a sulfated carbohydrate ligand, *J Exp Med 191*, 1105-1116.

103. Kawabata, S., and Tsuda, R. (2002) Molecular basis of non-self recognition by the horseshoe crab tachylectins, *Biochim Biophys Acta 1572*, 414-421.

104. Saito, T., Kawabata, S., Hirata, M., and Iwanaga, S. (1995) A novel type of limulus lectin-L6. Purification, primary structure, and antibacterial activity, *J Biol Chem 270*, 14493-14499.

105. Beisel, H. G., Kawabata, S., Iwanaga, S., Huber, R., and Bode, W. (1999) Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab Tachypleus tridentatus, *Embo j 18*, 2313-2322.

106. Saito, T., Hatada, M., Iwanaga, S., and Kawabata, S. (1997) A newly identified horseshoe crab lectin with binding specificity to O-antigen of bacterial lipopolysaccharides, *J Biol Chem 272*, 30703-30708.

107. Gokudan, S., Muta, T., Tsuda, R., Koori, K., Kawahara, T., Seki, N., Mizunoe, Y., Wai, S. N., Iwanaga, S., and Kawabata, S. (1999) Horseshoe crab acetyl group-recognizing lectins involved in innate immunity are structurally related to fibrinogen, *Proc Natl Acad Sci U S A 96*, 10086-10091.

108. Kairies, N., Beisel, H. G., Fuentes-Prior, P., Tsuda, R., Muta, T., Iwanaga, S., Bode, W., Huber, R., and Kawabata, S. (2001) The 2.0-A crystal structure of tachylectin 5A provides evidence for the common origin of the innate immunity and the blood coagulation systems, *Proc Natl Acad Sci U S A 98*, 13519-13524.

109. Roberson, M. M., and Barondes, S. H. (1982) Lectin from embryos and oocytes of Xenopus laevis. Purification and properties, *J Biol Chem 257*, 7520-7524.

110. Lee, J. K., Buckhaults, P., Wilkes, C., Teilhet, M., King, M. L., Moremen, K. W., and Pierce, M. (1997) Cloning and expression of a Xenopus laevis oocyte lectin and characterization of its mRNA levels during early development, *Glycobiology 7*, 367-372.

111. Lee, J. K., Baum, L. G., Moremen, K., and Pierce, M. (2004) The X-lectins: a new family with homology to the Xenopus laevis oocyte lectin XL-35, *Glycoconj J 21*, 443-450.

112. Yan, J., Xu, L., Zhang, Y., Zhang, C., Zhang, C., Zhao, F., and Feng, L. (2013) Comparative genomic and phylogenetic analyses of the intelectin gene family: implications for their origin and evolution, *Dev Comp Immunol 41*, 189-199.

113. Tsuji, S., Uehori, J., Matsumoto, M., Suzuki, Y., Matsuhisa, A., Toyoshima, K., and Seya, T. (2001) Human intelectin is a novel soluble lectin that recognizes galactofuranose in carbohydrate chains of bacterial cell wall, *J Biol Chem 276*, 23456-23463.

114. Pemberton, A. D., Knight, P. A., Gamble, J., Colledge, W. H., Lee, J. K., Pierce, M., and Miller, H. R. (2004) Innate BALB/c enteric epithelial responses to Trichinella spiralis: inducible expression of a novel goblet cell lectin, intelectin-2, and its natural deletion in C57BL/10 mice, *J Immunol 173*, 1894-1901.

115. Pemberton, A. D., Knight, P. A., Wright, S. H., and Miller, H. R. (2004) Proteomic analysis of mouse jejunal epithelium and its response to infection with the intestinal nematode, Trichinella spiralis, *Proteomics 4*, 1101-1108.

116. Nagata, S., Nishiyama, S., and Ikazaki, Y. (2013) Bacterial lipopolysaccharides stimulate production of XCL1, a calcium-dependent lipopolysaccharide-binding serum lectin, in Xenopus laevis, *Dev Comp Immunol 40*, 94-102.

117. Russell, S., Hayes, M. A., and Lumsden, J. S. (2009) Immunohistochemical localization of rainbow trout ladderlectin and intelectin in healthy and infected rainbow trout (Oncorhynchus mykiss), *Fish Shellfish Immunol 26*, 154-163.

118. Lin, B., Cao, Z., Su, P., Zhang, H., Li, M., Lin, Y., Zhao, D., Shen, Y., Jing, C., Chen, S., and Xu, A. (2009) Characterization and comparative analyses of zebrafish intelectins: highly conserved sequences, diversified structures and functions, *Fish Shellfish Immunol 26*, 396-405.

119. Lee, J. K., Schnee, J., Pang, M., Wolfert, M., Baum, L. G., Moremen, K. W., and Pierce, M. (2001) Human homologs of the Xenopus oocyte cortical granule lectin XL35, *Glycobiology 11*, 65-73.

120. Wesener, D. A., Wangkanont, K., McBride, R., Song, X., Kraft, M. B., Hodges, H. L., Zarling, L. C., R.A., S., Smith, D. F., Cummings, R. D., Paulson, J. C., Forest, K. T., and Kiessling, L. L. (2015) Recognition of Microbial Glycans by Human Intelectin-1, *Nat Struct Mol Biol 22*, 603-610.

121. Rudd, P. M., Elliott, T., Cresswell, P., Wilson, I. A., and Dwek, R. A. (2001) Glycosylation and the immune system, *Science 291*, 2370-2376.

122. Holmskov, U., Thiel, S., and Jensenius, J. C. (2003) Collectins and ficolins: Humoral lectins of the innate immune defense, *Annu. Rev. Immunol. 21*, 547-578.

123. Abe, Y., Tokuda, M., Ishimoto, R., Azumi, K., and Yokosawa, H. (1999) A unique primary structure, cDNA cloning and function of a galactose-specific lectin from ascidian plasma, *Eur J Biochem 261*, 33-39.

124. Yan, J., Wang, J., Zhao, Y., Zhang, J., Bai, C., Zhang, C., Zhang, C., Li, K., Zhang, H., Du, X., and Feng, L. (2012) Identification of an amphioxus intelectin homolog that preferably agglutinates gram-positive over gram-negative bacteria likely due to different binding capacity to LPS and PGN, *Fish Shellfish Immunol 33*, 11-20.

125. Yan, J., Zhang, C., Zhang, Y., Li, K., Xu, L., Guo, L., Kong, Y., and Feng, L. (2013) Characterization and comparative analyses of two amphioxus intelectins involved in the innate immune response, *Fish Shellfish Immunol 34*, 1139-1146.

126. Russell, S., Young, K. M., Smith, M., Hayes, M. A., and Lumsden, J. S. (2008) Identification, cloning and tissue localization of a rainbow trout (Oncorhynchus mykiss) intelectin-like protein that binds bacteria and chitin, *Fish Shellfish Immunol 25*, 91-105.

127. Datta, R., deSchoolmeester, M. L., Hedeler, C., Paton, N. W., Brass, A. M., and Else, K. J. (2005) Identification of novel genes in intestinal tissue that are regulated after infection with an intestinal nematode parasite, *Infect Immun 73*, 4025-4033.

128. French, A. T., Knight, P. A., Smith, W. D., Brown, J. K., Craig, N. M., Pate, J. A., Miller, H. R., and Pemberton, A. D. (2008) Up-regulation of intelectin in sheep after infection with Teladorsagia circumcincta, *Int J Parasitol 38*, 467-475.

129. Nishihara, T., Wyrick, R. E., Working, P. K., Chen, Y. H., and Hedrick, J. L. (1986) Isolation and characterization of a lectin from the cortical granules of Xenopus laevis eggs, *Biochemistry 25*, 6013-6020.

130. Shoji, H., Ikenaka, K., Nakakita, S., Hayama, K., Hirabayashi, J., Arata, Y., Kasai, K., Nishi, N., and Nakamura, T. (2005) Xenopus galectin-VIIa binds N-glycans of members of the cortical granule lectin family (xCGL and xCGL2), *Glycobiology 15*, 709-720.

131. Ishino, T., Kunieda, T., Natori, S., Sekimizu, K., and Kubo, T. (2007) Identification of novel members of the Xenopus Ca2+ -dependent lectin family and analysis of their gene expression during tail regeneration and development, *J Biochem 141*, 479-488.

132. Nagata, S. (2005) Isolation, characterization, and extra-embryonic secretion of the Xenopus laevis embryonic epidermal lectin, XEEL, *Glycobiology 15*, 281-290.

133. Nagata, S., Nakanishi, M., Nanba, R., and Fujita, N. (2003) Developmental expression of XEEL, a novel molecule of the Xenopus oocyte cortical granule lectin family, *Dev Genes Evol 213*, 368-370.

134. Dubaissi, E., Rousseau, K., Lea, R., Soto, X., Nardeosingh, S., Schweickert, A., Amaya, E., Thornton, D. J., and Papalopulu, N. (2014) A secretory cell type develops alongside multiciliated cells, ionocytes and goblet cells, and provides a protective, anti-infective function in the frog embryonic mucociliary epidermis, *Development 141*, 1514-1525.

135. Weis, W. I., Taylor, M. E., and Drickamer, K. (1998) The C-type lectin superfamily in the immune system, *Immunol. Rev. 163*, 19-34.

136. Gill, S. C., and von Hippel, P. H. (1989) Calculation of protein extinction coefficients from amino acid sequence data, *Anal. Biochem. 182*, 319-326.

137. Fullerton, S. W., Daff, S., Sanders, D. A., Ingledew, W. J., Whitfield, C., Chapman, S. K., and Naismith, J. H. (2003) Potentiometric analysis of UDP-galactopyranose mutase: stabilization of the flavosemiquinone by substrate, *Biochemistry 42*, 2104-2109.

138. Otwinowski, Z., and Minor, W. (1997) Processing of x-ray diffraction data collected in oscillation mode, In *Methods in enzymology: macromolecular crystallography, part A* (C.W. Carter, J., and Sweet, R. M., Eds.), pp 307-326, Academic Press New York.

139. Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H., and Hung, L. W. (2009) Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard, *Acta Crystallogr D Biol Crystallogr 65*, 582-601.

140. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., and Read, R. J. (2007) Phaser crystallographic software, *J Appl Crystallogr 40*, 658-674.

141. Afonine, P. V., Grosse-Kunstleve, R. W., and Adams, P. D. (2005) The Phenix refinement framework, *CCP4 Newsl. 42*, 8.

142. Emsley, P., Lohkamp, B., Scott, W. G., and Cowtan, K. (2010) Features and development of Coot, *Acta Crystallogr. D Biol. Crystallogr. 66*, 486-501.

143. Schuttelkopf, A. W., and van Aalten, D. M. (2004) PRODRG: a tool for high-throughput crystallography of protein-ligand complexes, *Acta Crystallogr D Biol Crystallogr 60*, 1355-1363.

144. Chen, V. B., III, W. B. A., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity:

all-atom structure validation for macromolecular crystallography, *Acta Crystallogr. D Biol. Crystallogr. 66*, 12-21.

145. The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC.

146. Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V., and Svergun, D. I. (2012) New developments in the ATSAS program package for small-angle scattering data analysis, *Journal of Applied Crystallography 45*, 342-350.

147. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J., and Svergun, D. I. (2003) PRIMUS: a Windows PC-based system for small-angle scattering data analysis, *Journal of Applied Crystallography 36*, 1277-1282.

148. Svergun, D. (1992) Determination of the regularization parameter in indirect-transform methods using perceptual criteria, *Journal of Applied Crystallography 25*, 495-503.

149. Svergun, D., Barberato, C., and Koch, M. H. J. (1995) CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates, *Journal of Applied Crystallography 28*, 768-773.

150. Franke, D., and Svergun, D. I. (2009) DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering, *Journal of Applied Crystallography 42*, 342-346.

151. Volkov, V. V., and Svergun, D. I. (2003) Uniqueness of ab initio shape determination in small-angle scattering, *Journal of Applied Crystallography 36*, 860-864.

152. Kozin, M. B., and Svergun, D. I. (2001) Automated matching of high- and low-resolution structural models, *Journal of Applied Crystallography 34*, 33-41.

153. M., L. T., D., S. B., M., R. T., and L., P. S. (1992) Computer-aided interpretation of analytical sedimentation data for proteins, In *Analytical ultracentrifugation in biochemistry and polymer science* (S.E., H., A.J., R., and J.C., H., Eds.), pp 90-125, Royal Society of Chemistry.

154. Laue, T. M. (1995) Sedimentation equilibrium as thermodynamic tool, *Methods Enzymol 259*, 427-452.

155. Kall, L., Krogh, A., and Sonnhammer, E. L. (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server, *Nucleic Acids Res 35*, W429-432.

156. Schmidt, T. G., and Skerra, A. (2007) The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins, *Nat Protoc 2*, 1528-1535.

157. Stivala, A., Wybrow, M., Wirth, A., Whisstock, J. C., and Stuckey, P. J. (2011) Automatic generation of protein structure cartoons with Pro-origami, *Bioinformatics 27*, 3315-3316.

158. Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids, *Nucleic Acids Res 38*, W529-533.

159. Mooij, W. T. M., Mitsiki, E., and Perrakis, A. (2009) ProteinCCD: enabling the design of protein truncation constructs for expression and crystallization experiments, *Nucleic Acids Research 37*, W402-W405.

160. Malashkevich, V. N., Chan, D. C., Chutkowski, C. T., and Kim, P. S. (1998) Crystal structure of the simian immunodeficiency virus (SIV) gp41 core: conserved helical interactions underlie the broad inhibitory activity of gp41 peptides, *Proc Natl Acad Sci U S A 95*, 9134-9139.

161. Bentley, S. D., Aanensen, D. M., Mavroidi, A., Saunders, D., Rabbinowitsch, E., Collins, M., Donohoe, K., Harris, D., Murphy, L., Quail, M. A., Samuel, G., Skovsted, I. C., Kaltoft, M. S., Barrell, B., Reeves, P. R., Parkhill, J., and Spratt, B. G. (2006) Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes, *PLoS Genet. 2*, 262-269.

162. Feinberg, H., Powlesland, A. S., Taylor, M. E., and Weis, W. I. (2010) Trimeric structure of langerin, *J Biol Chem 285*, 13285-13293.

163. Asensio, J. L., Arda, A., Canada, F. J., and Jimenez-Barbero, J. (2013) Carbohydrate-aromatic interactions, *Acc Chem Res 46*, 946-954.

164. Hudson, K. L., Bartlett, G. J., Diehl, R. C., Agirre, J., Gallagher, T., Kiessling, L. L., and Woolfson, D. N. (2015) Carbohydrate-aromatic interactions in proteins, *J Am Chem Soc*.

165. Tsuji, S., Yamashita, M., Hoffman, D. R., Nishiyama, A., Shinohara, T., Ohtsu, T., and Shibata, Y. (2009) Capture of heat-killed Mycobacterium bovis bacillus Calmette-Guerin by intelectin-1 deposited on cell surfaces, *Glycobiology 19*, 518-526.

166. Tsuji, S., Yamashita, M., Nishiyama, A., Shinohara, T., Li, Z., Myrvik, Q. N., Hoffman, D. R., Henriksen, R. A., and Shibata, Y. (2007) Differential structure and activity between human and mouse intelectin-1: human intelectin-1 is a disulfide-linked trimer, whereas mouse homologue is a monomer, *Glycobiology 17*, 1045-1051.

167. Rothlisberger, D., Honegger, A., and Pluckthun, A. (2005) Domain interactions in the Fab fragment: a comparative evaluation of the single-chain Fv and Fab format engineered with variable domains of different stability, *J Mol Biol 347*, 773-789.

168. Chang, S. G., Choi, K. D., Jang, S. H., and Shin, H. C. (2003) Role of disulfide bonds in the structure and activity of human insulin, *Mol Cells 16*, 323-330.

169. Zuniga, J. E., Groppe, J. C., Cui, Y., Hinck, C. S., Contreras-Shannon, V., Pakhomova, O. N., Yang, J., Tang, Y., Mendoza, V., Lopez-Casillas, F., Sun, L., and Hinck, A. P. (2005) Assembly of TbetaRI:TbetaRII:TGFbeta ternary complex in vitro with receptor extracellular domains is cooperative and isoform-dependent, *J Mol Biol 354*, 1052-1068.

170. Fass, D. (2012) Disulfide bonding in protein biophysics, *Annu Rev Biophys 41*, 63-79.

171. Wong, J. W., Ho, S. Y., and Hogg, P. J. (2011) Disulfide bond acquisition through eukaryotic protein evolution, *Mol Biol Evol 28*, 327-334.

172. Stowell, S. R., Arthur, C. M., Dias-Baruffi, M., Rodrigues, L. C., Gourdine, J. P., Heimburg-Molinaro, J., Ju, T., Molinaro, R. J., Rivera-Marrero, C., Xia, B., Smith, D. F., and Cummings, R. D. (2010) Innate immune lectins kill bacteria expressing blood group antigen, *Nat. Med. 16*, 295-301.

173. Vaishnava, S., Yamamoto, M., Severson, K. M., Ruhn, K. A., Yu, X., Koren, O., Ley, R., Wakeland, E. K., and Hooper, L. V. (2011) The antibacterial lectin RegIIIgamma promotes the spatial segregation of microbiota and host in the intestine, *Science 334*, 255-258.

174. Lis, H., and Sharon, N. (1998) Lectins: Carbohydrate-specific proteins that mediate cellular recognition, *Chem. Rev. 98*, 637-674.

175. Jack, D. L., Klein, N. J., and Turner, M. W. (2001) Mannose-binding lectin: targeting the microbial world for complement attack and opsonophagocytosis, *Immunol. Rev. 180*, 86-99.

176. Thomsen, T., Schlosser, A., Holmskov, U., and Sorensen, G. L. (2011) Ficolins and FIBCD1: soluble and membrane bound pattern recognition molecules with acetyl group selectivity, *Mol. Immunol. 48*, 369-381.

177. French, A. T., Bethune, J. A., Knight, P. A., McNeilly, T. N., Wattegedera, S., Rhind, S., Miller, H. R., and Pemberton, A. D. (2007) The expression of intelectin in sheep goblet cells and upregulation by interleukin-4, *Vet. Immunol. Immunopathol. 120*, 41-46.

178. Voehringer, D., Stanley, S. A., Cox, J. S., Completo, G. C., Lowary, T. L., and Locksley, R. M. (2007) Nippostrongylus brasiliensis: identification of intelectin-1 and -2 as Stat6-dependent genes expressed in lung and intestine during infection, *Exp. Parasitol. 116*, 458-466.

179. Kerr, S. C., Carrington, S. D., Oscarson, S., Gallagher, M. E., Solon, M., Yuan, S., Ahn, J. N., Dougherty, R. H., Finkbeiner, W. E., Peters, M. C., and Fahy, J. V.

(2014) Intelectin-1 is a prominent protein constituent of pathologic mucus associated with eosinophilic airway inflammation in asthma, *Am. J. Respir. Crit. Care Med. 189*, 1005-1007.

180. Kuperman, D. A., Lewis, C. C., Woodruff, P. G., Rodriguez, M. W., Yang, Y. H., Dolganov, G. M., Fahy, J. V., and Erle, D. J. (2005) Dissecting asthma using focused transgenic modeling and functional genomics, *J. Allergy Clin. Immunol. 116*, 305-311.

181. Suzuki, Y. A., Shin, K., and Lonnerdal, B. (2001) Molecular cloning and functional expression of a human intestinal lactoferrin receptor, *Biochemistry 40*, 15771-15779.

182. de Souza Batista, C. M., Yang, R. Z., Lee, M. J., Glynn, N. M., Yu, D. Z., Pray, J., Ndubuizu, K., Patil, S., Schwartz, A., Kligman, M., Fried, S. K., Gong, D. W., Shuldiner, A. R., Pollin, T. I., and McLenithan, J. C. (2007) Omentin plasma levels and gene expression are decreased in obesity, *Diabetes 56*, 1655-1661.

183. Tsuji, S., Yamashita, M., Nishiyama, A., Shinohara, T., Zhongwei, U., Myrvik, Q. N., Hoffman, D. R., Henriksen, R. A., and Shibata, Y. (2007) Differential structure and activity between human and mouse intelectin-1: Human intelectin-1 is a disulfide-linked trimer, whereas mouse homologue is a monomer, *Glycobiology 17*, 1045-1051.

184. Blixt, O., Head, S., Mondala, T., Scanlan, C., Huflejt, M. E., Alvarez, R., Bryan, M. C., Fazio, F., Calarese, D., Stevens, J., Razi, N., Stevens, D. J., Skehel, J. J., van Die, I., Burton, D. R., Wilson, I. A., Cummings, R., Bovin, N., Wong, C. H., and Paulson, J. C. (2004) Printed covalent glycan array for ligand profiling of diverse glycan binding proteins, *Proc. Natl. Acad. Sci. U.S.A. 101*, 17033-17038.

185. Stowell, S. R., Arthur, C. M., McBride, R., Berger, O., Razi, N., Heimburg-Molinaro, J., Rodrigues, L. C., Gourdine, J. P., Noll, A. J., von Gunten, S., Smith, D. F., Knirel, Y. A., Paulson, J. C., and Cummings, R. D. (2014) Microbial glycan microarrays define key features of host-microbial interactions, *Nat. Chem. Biol. 10*, 470-476.

186. Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart., P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution., *Acta Crystallogr. D Biol. Crystallogr. 66*, 213-221.

187. Matthews, B. W. (1968) Solvent content of protein crystals, *J Mol Biol 33*, 491-497.

188. Altona, C., and Sundaralingam, M. (1972) Conformational-Analysis of Sugar Ring in Nucleosides and Nucleotides - New Description Using Concept of Pseudorotation, *J. Am. Chem. Soc. 94*, 8205-&.

189. Taha, H. A., Richards, M. R., and Lowary, T. L. (2013) Conformational analysis of furanoside-containing mono- and oligosaccharides, *Chem. Rev. 113*, 1851-1876.

190. Richards, M. R., Bai, Y., and Lowary, T. L. (2013) Comparison between DFT- and NMR-based conformational analysis of methyl galactofuranosides, *Carbohyd. Res. 374*, 103-114.

191. Weis, W. I., and Drickamer, K. (1996) Structural basis of lectin-carbohydrate recognition, *Annu. Rev. Biochem. 65*, 441-473.

192. Angata, T., and Varki, A. (2002) Chemical diversity in the sialic acids and related alpha-keto acids: an evolutionary perspective, *Chem. Rev. 102*, 439-469.

193. Varghese, J. N., McKimm-Breschkin, J. L., Caldwell, J. B., Kortt, A. A., and Colman, P. M. (1992) The structure of the complex between influenza virus neuraminidase and sialic acid, the viral receptor, *Proteins 14*, 327-332.

194. Kraschnefski, M. J., Bugarcic, A., Fleming, F. E., Yu, X., von Itzstein, M., Coulson, B. S., and Blanchard, H. (2009) Effects on sialic acid recognition of amino acid mutations in the carbohydrate-binding cleft of the rotavirus spike protein, *Glycobiology 19*, 194-200.

195. Blanchard, H., Yu, X., Coulson, B. S., and von Itzstein, M. (2007) Insight into host cell carbohydrate-recognition by human and porcine rotavirus from crystal structures of the virion spike associated carbohydrate-binding domain (VP8*), *J Mol Biol 367*, 1215-1226.

196. Dormitzer, P. R., Sun, Z. Y., Wagner, G., and Harrison, S. C. (2002) The rhesus rotavirus VP4 sialic acid binding domain has a galectin fold with a novel carbohydrate binding site, *Embo j 21*, 885-897.

197. Pemberton, A. D., Rose-Zerilli, M. J., Holloway, J. W., Gray, R. D., and Holgate, S. T. (2008) A single-nucleotide polymorphism in intelectin 1 is associated with increased asthma risk, *J Allergy Clin Immunol 122*, 1033-1034.

198. Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., Schumm, L. P., Steinhart, A. H., Targan, S. R., Xavier, R. J., Consortium, N. I. G., Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J. P., de Vos, M., Vermeire, S., Louis, E., Belgian-French, I. B. D. C., Wellcome Trust Case

Control, C., Cardon, L. R., Anderson, C. A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N. J., Onnie, C. M., Fisher, S. A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C. G., Parkes, M., Georges, M., and Daly, M. J. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease, *Nat. Genet. 40*, 955-962.

199. Stranger, B. E., Stahl, E. A., and Raj, T. (2011) Progress and promise of genome-wide association studies for human complex trait genetics, *Genetics 187*, 367-383.

200. Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., Cozzi-Lepri, A., De Luca, A., Easterbrook, P., Francioli, P., Mallal, S., Martinez-Picado, J., Miro, J. M., Obel, N., Smith, J. P., Wyniger, J., Descombes, P., Antonarakis, S. E., Letvin, N. L., McMichael, A. J., Haynes, B. F., Telenti, A., and Goldstein, D. B. (2007) A whole-genome association study of major determinants for host control of HIV-1, *Science 317*, 944-947.

201. Kamatani, Y., Wattanapokayakit, S., Ochi, H., Kawaguchi, T., Takahashi, A., Hosono, N., Kubo, M., Tsunoda, T., Kamatani, N., Kumada, H., Puseenam, A., Sura, T., Daigo, Y., Chayama, K., Chantratita, W., Nakamura, Y., and Matsuda, K. (2009) A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians, *Nat Genet 41*, 591-595.

202. Ge, D., Fellay, J., Thompson, A. J., Simon, J. S., Shianna, K. V., Urban, T. J., Heinzen, E. L., Qiu, P., Bertelsen, A. H., Muir, A. J., Sulkowski, M., McHutchison, J. G., and Goldstein, D. B. (2009) Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance, *Nature 461*, 399-401.

203. Rauch, A., Kutalik, Z., Descombes, P., Cai, T., Di Iulio, J., Mueller, T., Bochud, M., Battegay, M., Bernasconi, E., Borovicka, J., Colombo, S., Cerny, A., Dufour, J. F., Furrer, H., Gunthard, H. F., Heim, M., Hirschel, B., Malinverni, R., Moradpour, D., Mullhaupt, B., Witteck, A., Beckmann, J. S., Berg, T., Bergmann, S., Negro, F., Telenti, A., and Bochud, P. Y. (2010) Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study, *Gastroenterology 138*, 1338-1345, 1345.e1331-1337.

204. Arnold, R. R., Cole, M. F., and McGhee, J. R. (1977) A bactericidal effect for human lactoferrin, *Science 197*, 263-265.

205. Arnold, R. R., Russell, J. E., Champion, W. J., Brewer, M., and Gauthier, J. J. (1982) Bactericidal activity of human lactoferrin: differentiation from the stasis of iron deprivation, *Infect Immun 35*, 792-799.

206. Arnold, R. R., Brewer, M., and Gauthier, J. J. (1980) Bactericidal activity of human lactoferrin: sensitivity of a variety of microorganisms, *Infect Immun 28*, 893-898.

207. Arnold, R. R., Russell, J. E., Champion, W. J., and Gauthier, J. J. (1981) Bactericidal activity of human lactoferrin: influence of physical conditions and metabolic state of the target microorganism, *Infect Immun 32*, 655-660.

208. Quill, T. A., and Hedrick, J. L. (1996) The fertilization layer mediated block to polyspermy in Xenopus laevis: isolation of the cortical granule lectin ligand, *Arch Biochem Biophys 333*, 326-332.

209. Levi, G., and Teichberg, V. I. (1981) Isolation and physicochemical characterization of electrolectin, a beta-D-galactoside binding lectin from the electric organ of Electrophorus electricus, *J Biol Chem 256*, 5735-5740.

210. Antcheva, N., Pintar, A., Patthy, A., Simoncsits, A., Barta, E., Tchorbanov, B., and Pongor, S. (2001) Proteins of circularly permuted sequence present within the same organism: the major serine proteinase inhibitor from Capsicum annuum seeds, *Protein Sci 10*, 2280-2290.

211. Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions, *Nat Methods 8*, 785-786.

212. Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2003) Genes required for mycobacterial growth defined by high density mutagenesis, *Mol Microbiol 48*, 77-84.

213. Pan, F., Jackson, M., Ma, Y., and McNeil, M. (2001) Cell wall core galactofuran synthesis is essential for growth of mycobacteria, *J Bacteriol 183*, 3991-3998.

214. Hsieh, P. F., Lin, T. L., Yang, F. L., Wu, M. C., Pan, Y. J., Wu, S. H., and Wang, J. T. (2012) Lipopolysaccharide O1 antigen contributes to the virulence in Klebsiella pneumoniae causing pyogenic liver abscess, *PLoS One 7*, e33155.

215. Tefsen, B., Ram, A. F., van Die, I., and Routier, F. H. (2012) Galactofuranose in eukaryotes: aspects of biosynthesis and functional impact, *Glycobiology 22*, 456-469.

216. Novelli, J. F., Chaudhary, K., Canovas, J., Benner, J. S., Madinger, C. L., Kelly, P., Hodgkin, J., and Carlow, C. K. (2009) Characterization of the Caenorhabditis elegans UDP-galactopyranose mutase homolog glf-1 reveals an essential role for galactofuranose metabolism in nematode surface coat synthesis, *Dev Biol 335*, 340-355.

217. Beverley, S. M., Owens, K. L., Showalter, M., Griffith, C. L., Doering, T. L., Jones, V. C., and McNeil, M. R. (2005) Eukaryotic UDP-galactopyranose mutase (GLF gene) in microbial and metazoal pathogens, *Eukaryot Cell 4*, 1147-1154.

218. Wesener, D. A., May, J. F., Huffman, E. M., and Kiessling, L. L. (2013) UDP-galactopyranose mutase in nematodes, *Biochemistry 52*, 4391-4398.

219. Soltero-Higgin, M., Carlson, E. E., Gruber, T. D., and Kiessling, L. L. (2004) A unique catalytic mechanism for UDP-galactopyranose mutase, *Nat Struct Mol Biol 11*, 539-543.

220. Nassau, P. M., Martin, S. L., Brown, R. E., Weston, A., Monsey, D., McNeil, M. R., and Duncan, K. (1996) Galactofuranose biosynthesis in Escherichia coli K-12: identification and cloning of UDP-galactopyranose mutase, *J Bacteriol 178*, 1047-1052.

221. Sanders, D. A., Staines, A. G., McMahon, S. A., McNeil, M. R., Whitfield, C., and Naismith, J. H. (2001) UDP-galactopyranose mutase has a novel structure and mechanism, *Nat Struct Biol 8*, 858-863.

222. Richards, M. R., and Lowary, T. L. (2009) Chemistry and biology of galactofuranose-containing polysaccharides, *Chembiochem 10*, 1920-1938.

223. Borrelli, S., Zandberg, W. F., Mohan, S., Ko, M., Martinez-Gutierrez, F., Partha, S. K., Sanders, D. A., Av-Gay, Y., and Pinto, B. M. (2010) Antimycobacterial activity of UDP-galactopyranose mutase inhibitors, *Int J Antimicrob Agents 36*, 364-368.

224. Dykhuizen, E. C., May, J. F., Tongpenyai, A., and Kiessling, L. L. (2008) Inhibitors of UDP-galactopyranose mutase thwart mycobacterial growth, *J Am Chem Soc 130*, 6706-6707.

225. Soltero-Higgin, M., Carlson, E. E., Phillips, J. H., and Kiessling, L. L. (2004) Identification of inhibitors for UDP-galactopyranose mutase, *J Am Chem Soc 126*, 10532-10533.

226. Carlson, E. E., May, J. F., and Kiessling, L. L. (2006) Chemical probes of UDP-galactopyranose mutase, *Chem Biol 13*, 825-837.

227. Partha, S. K., Sadeghi-Khomami, A., Slowski, K., Kotake, T., Thomas, N. R., Jakeman, D. L., and Sanders, D. A. (2010) Chemoenzymatic synthesis, inhibition studies, and X-ray crystallographic analysis of the phosphono analog of UDP-Galp as an inhibitor and mechanistic probe for UDP-galactopyranose mutase, *J Mol Biol 403*, 578-590.

228. Caravano, A., Dohi, H., Sinay, P., and Vincent, S. P. (2006) A new methodology for the synthesis of fluorinated exo-glycals and their time-dependent inhibition of UDP-galactopyranose mutase, *Chemistry 12*, 3114-3123.

229. Itoh, K., Huang, Z., and Liu, H. W. (2007) Synthesis and analysis of substrate analogues for UDP-galactopyranose mutase: implication for an oxocarbenium ion intermediate in the catalytic mechanism, *Org Lett 9*, 879-882.

230. Liautard, V., Desvergnes, V., Itoh, K., Liu, H. W., and Martin, O. R. (2008) Convergent and stereoselective synthesis of iminosugar-containing Galf and

UDP-Galf mimicks: evaluation as inhibitors of UDP-Gal mutase, *J Org Chem 73*, 3103-3115.

231. Ansiaux, C., N'Go, I., and Vincent, S. P. (2012) Reversible and efficient inhibition of UDP-galactopyranose mutase by electrophilic, constrained and unsaturated UDP-galactitol analogues, *Chemistry 18*, 14860-14866.

232. El Bkassiny, S., N'Go, I., Sevrain, C. M., Tikad, A., and Vincent, S. P. (2014) Synthesis of a novel UDP-carbasugar as UDP-galactopyranose mutase inhibitor, *Org Lett 16*, 2462-2465.

233. van Straaten, K. E., Kuttiyatveetil, J. R., Sevrain, C. M., Villaume, S. A., Jimenez-Barbero, J., Linclau, B., Vincent, S. P., and Sanders, D. A. (2015) Structural basis of ligand binding to UDP-galactopyranose mutase from Mycobacterium tuberculosis using substrate and tetrafluorinated substrate analogues, *J Am Chem Soc 137*, 1230-1244.

234. Kuppala, R., Borrelli, S., Slowski, K., Sanders, D. A., Ravindranathan Kartha, K. P., and Pinto, B. M. (2015) Synthesis and biological evaluation of nonionic substrate mimics of UDP-Galp as candidate inhibitors of UDP galactopyranose mutase (UGM), *Bioorg Med Chem Lett 25*, 1995-1997.

235. Baell, J., and Walters, M. A. (2014) Chemistry: Chemical con artists foil drug discovery, *Nature 513*, 481-483.

236. Lorber, D. M., and Shoichet, B. K. (2005) Hierarchical docking of databases of multiple ligand conformations, *Curr Top Med Chem 5*, 739-749.

237. Mysinger, M. M., and Shoichet, B. K. (2010) Rapid context-dependent ligand desolvation in molecular docking, *J Chem Inf Model 50*, 1561-1573.

238. Irwin, J. J., and Shoichet, B. K. (2005) ZINC--a free database of commercially available compounds for virtual screening, *J Chem Inf Model 45*, 177-182.

239. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012) ZINC: a free tool to discover chemistry for biology, *J Chem Inf model 52*, 1757-1768.

240. Chad, J. M., Sarathy, K. P., Gruber, T. D., Addala, E., Kiessling, L. L., and Sanders, D. A. (2007) Site-directed mutagenesis of UDP-galactopyranose mutase reveals a critical role for the active-site, conserved arginine residues, *Biochemistry 46*, 6723-6732.

241. Ponnusamy, R., Lebedev, A. A., Pahlow, S., and Lohkamp, B. (2014) Crystal structure of human CRMP-4: correction of intensities for lattice-translocation disorder, *Acta Crystallogr D Biol Crystallogr 70*, 1680-1694.

242. Sauter, N. K. (2011) Visualizing the raw diffraction pattern with LABELIT, In *Computational Crystallography Newsletter* pp 15-24.

243. Leslie, A. G. W., and Powell, H. R. (2007) Processing Diffraction Data with Mosflm, In *Evolving Methods for Macromolecular Crystallography*, pp 41-51.

244. Evans, P. R., and Murshudov, G. N. (2013) How good are my data and what is the resolution?, *Acta Crystallogr D Biol Crystallogr 69*, 1204-1214.

245. Evans, P. (2006) Scaling and assessment of data quality, *Acta Crystallogr D Biol Crystallogr 62*, 72-82.

246. Evans, P. R. (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics, *Acta Crystallogr D Biol Crystallogr 67*, 282-292.

247. Gruber, T. D., Westler, W. M., Kiessling, L. L., and Forest, K. T. (2009) X-ray crystallography reveals a reduced substrate complex of UDP-galactopyranose mutase poised for covalent catalysis by flavin, *Biochemistry 48*, 9171-9173.

248. Murshudov, G. N., Skubak, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., and Vagin, A. A. (2011) REFMAC5 for the refinement of macromolecular crystal structures, *Acta Crystallogr D Biol Crystallogr 67*, 355-367.

249. (1994) The CCP4 suite: programs for protein crystallography, *Acta Crystallogr D Biol Crystallogr 50*, 760-763.

250. Moriarty, N., Grosse-Kunstleve, R., and Adams, P. (2009) electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation, *Acta Crystallogr. D Biol. Crystallogr. 65*, 1074-1080.

251. Bragg, W. L., and Howells, E. R. (1954) X-ray diffraction by imidazole methaemoglobin, *Acta Crystallographica 7*, 409-411.

252. Wang, J., Kamtekar, S., Berman, A. J., and Steitz, T. A. (2005) Correction of X-ray intensities from single crystals containing lattice-translocation defects, *Acta Crystallogr D Biol Crystallogr 61*, 67-74.

253. Wang, J., Rho, S. H., Park, H. H., and Eom, S. H. (2005) Correction of X-ray intensities from an HslV-HslU co-crystal containing lattice-translocation defects, *Acta Crystallogr D Biol Crystallogr 61*, 932-941.

254. Scherman, M. S., Winans, K. A., Stern, R. J., Jones, V., Bertozzi, C. R., and McNeil, M. R. (2003) Drug targeting Mycobacterium tuberculosis cell wall synthesis: development of a microtiter plate-based screen for UDP-galactopyranose mutase and identification of an inhibitor from a uridine-based library, *Antimicrob Agents Chemother 47*, 378-382.

255. Tangallapally, R. P., Yendapally, R., Lee, R. E., Hevener, K., Jones, V. C., Lenaerts, A. J., McNeil, M. R., Wang, Y., Franzblau, S., and Lee, R. E. (2004) Synthesis and evaluation of nitrofuranylamides as novel antituberculosis agents, *J Med Chem 47*, 5276-5283.

256. Kincaid, V. A., London, N., Wangkanont, K., Wesener, D. A., Marcus, S. A., Heroux, A., Nedyalkova, L., Talaat, A. M., Forest, K. T., Shoichet, B. K., and Kiessling, L. L. (2015) Virtual Screening for UDP-Galactopyranose Mutase Ligands Identifies a New Class of Antimycobacterial Agents, *ACS Chem Biol 10*, 2209-2218.

257. Beis, K., Srikannathasan, V., Liu, H., Fullerton, S. W., Bamford, V. A., Sanders, D. A., Whitfield, C., McNeil, M. R., and Naismith, J. H. (2005) Crystal structures of Mycobacteria tuberculosis and Klebsiella pneumoniae UDP-galactopyranose mutase in the oxidised state and Klebsiella pneumoniae UDP-galactopyranose mutase in the (active) reduced state, *J Mol Biol 348*, 971-982.

258. Harding, M. M. (2002) Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium, *Acta Crystallogr D Biol Crystallogr 58*, 872-874.

259. Bennett, B. D., Kimball, E. H., Gao, M., Osterhout, R., Van Dien, S. J., and Rabinowitz, J. D. (2009) Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli, *Nat Chem Biol 5*, 593-599.

260. Boechi, L., de Oliveira, C. A., Da Fonseca, I., Kizjakina, K., Sobrado, P., Tanner, J. J., and McCammon, J. A. (2013) Substrate-dependent dynamics of UDP-galactopyranose mutase: Implications for drug design, *Protein Sci 22*, 1490-1501.

261. Tanner, J. J., Boechi, L., Andrew McCammon, J., and Sobrado, P. (2014) Structure, mechanism, and dynamics of UDP-galactopyranose mutase, *Arch Biochem Biophys 544*, 128-141.

262. Yuan, Y., Wen, X., Sanders, D. A., and Pinto, B. M. (2005) Exploring the mechanism of binding of UDP-galactopyranose to UDP-galactopyranose mutase by STD-NMR spectroscopy and molecular modeling, *Biochemistry 44*, 14080-14089.

263. Yuan, Y., Bleile, D. W., Wen, X., Sanders, D. A., Itoh, K., Liu, H. W., and Pinto, B. M. (2008) Investigation of binding of UDP-Galf and UDP-[3-F]Galf to UDP-galactopyranose mutase by STD-NMR spectroscopy, molecular dynamics, and CORCEMA-ST calculations, *J Am Chem Soc 130*, 3157-3168.

264. Yao, X., Bleile, D. W., Yuan, Y., Chao, J., Sarathy, K. P., Sanders, D. A., Pinto, B. M., and O'Neill, M. A. (2009) Substrate directs enzyme dynamics by bridging distal sites: UDP-galactopyranose mutase, *Proteins 74*, 972-979.

265. Partha, S. K., van Straaten, K. E., and Sanders, D. A. (2009) Structural basis of substrate binding to UDP-galactopyranose mutase: crystal structures in the reduced and oxidized state complexed with UDP-galactopyranose and UDP, *J Mol Biol 394*, 864-877.

266. Schomaker, V., and Trueblood, K. N. (1968) On the rigid-body motion of molecules in crystals, *Acta Cryst. B24*, 63-76.

267. Urzhumtsev, A., Afonine, P. V., and Adams, P. D. (2013) TLS from fundamentals to practice, *Crystallogr Rev 19*, 230-270.

268. Painter, J., and Merritt, E. A. (2006) Optimal description of a protein structure in terms of multiple groups undergoing TLS motion, *Acta Crystallogr D Biol Crystallogr 62*, 439-450.

269. Painter, J., and Merritt, E. A. (2006) TLSMD web server for the generation of multi-group TLS models, *J. Appl. Cryst. 39*.

270. Harding, M. M. (2001) Geometry of metal-ligand interactions in proteins, *Acta Crystallogr D Biol Crystallogr 57*, 401-411.

271. Wall, M. E. (2009) Methods and software for diffuse X-ray scattering from protein crystals, *Methods Mol Biol 544*, 269-279.

272. Van Benschoten, A. H., Afonine, P. V., Terwilliger, T. C., Wall, M. E., Jackson, C. J., Sauter, N. K., Adams, P. D., Urzhumtsev, A., and Fraser, J. S. (2015) Predicting X-ray diffuse scattering from translation-libration-screw structural ensembles, *Acta Crystallogr D Biol Crystallogr 71*, 1657-1667.

273. Massague, J. (1998) TGF-beta signal transduction, *Annual Review of Biochemistry 67*, 753-791.

274. Groppe, J., Hinck, C. S., Samavarchi-Tehrani, P., Zubieta, C., Schuermann, J. P., Taylor, A. B., Schwarz, P. M., Wrana, J. L., and Hinck, A. P. (2008) Cooperative assembly of TGF-β superfamily signaling complexes is mediated by two disparate mechanisms and distinct modes of receptor binding, *Mol. Cell 29*, 157–168.

275. Massague, J., and Wotton, D. (2000) Transcriptional control by the TGF-beta/Smad signaling system, *Embo j 19*, 1745-1754.

276. Massague, J. (2000) How cells read TGF-β signals, *Nat. Rev. Mol. Cell Bio. 1*, 169-178.

277. Blobe, G. C., Schiemann, W. P., and Lodish, H. F. (2000) Role of Transforming Growth Factor β in Human Disease, *N Engl J Med 342*, 1350-1358.

278. Halder, S. K., Beauchamp, R. D., and Datta, P. K. (2005) A specific inhibitor of TGF-beta receptor kinase, SB-431542, as a potent antitumor agent for human cancers, *Neoplasia 7*, 509-521.

279. Li, L., Orner, B. P., Huang, T., Hinck, A. P., and Kiessling, L. L. (2010) Peptide ligands that use a novel binding site to target both TGF-b receptors, *Mol. Biosyst. 6*, 2392–2402.

280. Li, L., Klim, J. R., Derda, R., Courtney, A. H., and Kiessling, L. L. (2011) Spatial control of cell fate using synthetic surfaces to potentiate TGF-β signaling, *Proc. Natl. Acad. Sci. U.S.A. 108*, 11745-11750.

281. Pedersen, R. O., Loboa, E. G., and LaBean, T. H. (2013) Sensitization of Transforming Growth Factor-beta Signaling by Multiple Peptides Patterned on DNA Nanostructures, *Biomacromolecules 14*, 4157-4160.

282. Boesen, C. C., Motyka, S. A., Patamawenu, A., and Sun, P. D. (2000) Development of a recombinant bacterial expression system for the active form of a human transforming growth factor β type II receptor ligand binding domain, *Protein Expr. Purif. 20*, 98-104.

283. Gasparian, M. E., Elistratov, P. A., Yakimov, S. A., Dolgikh, D. A., and Kirpichnikov, M. P. (2010) An efficient method for expression in Escherichia coli and purification of the extracellular ligand binding domain of the human TGFβ type II receptor, *J. Biotechnol. 148*, 113-118.

284. Hinck, A. P., Kerfoot P. Walker, I., Martina, N. R., Deepa, S., Hinck, C. S., and Freedberg, D. I. (2000) Sequential resonance assignments of the extracellular ligand binding domain of the human TGF-b type II receptor, *J. Biomol. NMR 18*, 369-370.

285. Janovick, J. A., Stewart, M. D., Jacob, D., Martin, L. D., Deng, J. M., Stewart, C. A., Wang, Y., Cornea, A., Chavali, L., Lopez, S., Mitalipov, S., Kang, E., Lee, H. S., Manna, P. R., Stocco, D. M., Behringer, R. R., and Conn, P. M. (2013) Restoration of testis function in hypogonadotropic hypogonadal mice harboring a misfolded GnRHR mutant by pharmacoperone drug therapy, *Proc Natl Acad Sci U S A 110*, 21030-21035.

286. Makley, L. N., and Gestwicki, J. E. (2013) Expanding the Number of 'Druggable' Targets: Non-Enzymes and Protein-Protein Interactions, *Chemical Biology & Drug Design 81*, 22-32.

287. Mecozzi, V. J., Berman, D. E., Simoes, S., Vetanovetz, C., Awal, M. R., Patel, V. M., Schneider, R. T., Petsko, G. A., Ringe, D., and Small, S. A. (2014) Pharmacological chaperones stabilize retromer to limit APP processing, *10*, 443-449.

288. Mu, T. W., Ong, D. S. T., Wang, Y. J., Balch, W. E., Yates, J. R., Segatori, L., and Kelly, J. W. (2008) Chemical and biological approaches synergize to ameliorate protein-folding diseases, *Cell 134*, 769-781.

289. Oh, M., Lee, J. H., Wang, W., Lee, H. S., Lee, W. S., Burlak, C., Im, W., Hoang, Q. Q., and Lim, H. S. (2014) Potential pharmacological chaperones targeting cancer-associated MCL-1 and Parkinson disease-associated alpha-synuclein, *Proc Natl Acad Sci U S A 111*, 11007-11012.

290. Huang, T., David, L., Mendoza, V., Yang, Y., Villarreal, M., De, K., Sun, L., Fang, X., Lopez-Casillas, F., Wrana, J. L., and Hinck, A. P. (2011) TGF-b signalling is mediated by two autonomously functioning TbRI:TbRII pairs, *EMBO J. 30*, 1263–1276.

291. Boesen, C. C., Motyka, S. A., Patamawenua, A., and Sun, P. D. (2002) Crystallization and preliminary crystallographic studies of human TGF-b type II receptor ligand-binding domain, *Acta Crystallogr. D Biol. Crystallogr. 58*, 1214-1216.

292. Boesen, C. C., Radaev, S., Motyka, S. A., Patamawenu, A., and Sun, P. D. (2002) The 1.1 Å crystal structure of human TGF-β type II receptor ligand binding domain, *Structure 10*, 913-919.

293. Glansbeek, H. L., Beuningen, H. M. v., Vitters, E. L., Kraan, P. M. v. d., and Berg, W. B. v. d. (1998) Expression of recombinant human soluble type II transforming growth factor-β receptor in *Pichia pastoris* and *Escherichia coli*: two powerful systems to express a potent inhibitor of transforming growth factor-β1, *Protein Expr. Purif. 12*, 201-207.

294. Marlow, M. S., Brown, C. B., Barnett, J. V., and Krezel, A. M. (2003) Solution structure of the chick TGFβ type II receptor ligand-binding domain, *J. Mol. Biol. 326*, 989-997.

295. Zuniga, J. E., Groppe, J. C., Cui, Y., Hinck, C. S., Contreras-Shannon, V., Pakhomova, O. N., Yang, J., Tang, Y., Mendoza, V., Lopez-Casillas, F., Sun, L., and Hinck, A. P. (2005) Assembly of TβRI:TβRII:TGFβ ternary complex in vitro with receptor extracellular domains is cooperative and isoform-dependent, *J. Mol. Biol. 354*, 1052-1068.

296. Abe, R., Kudou, M., Tanaka, Y., Arakawa, T., and Tsumoto, K. (2009) Immobilized metal affinity chromatography in the presence of arginine, *Biochem Biophys Res Commun 381*, 306-310.

297. Cleland, J. L., Hedgepeth, C., and Wang, D. I. C. (1992) Polyethylene glycol enhanced refolding of bovine carbonic anhydrase B, *J. Biol. Chem. 267*, 13327-13334.

298. Karuppiah, N., and Sharma, A. (1995) Cyclodextrins as protein folding aids, *Biochem Biophys Res Commun 211*, 60-66.

299. Mishra, R., Seckler, R., and Bhat, R. (2005) Efficient refolding of aggregation-prone citrate synthase by polyol osmolytes, *J. Biol. Chem. 280*, 15553-15560.

300. Vuillard, L., Rabilloud, T., and Goldberg, M. E. (1998) Interactions of non-detergent sulfobetaines with early folding intermediates facilitate in vitro protein renaturation, *Eur. J. Biochem. 256*, 128-135.

301. Krissinel, E., and Henrick, K. (2007) Inference of macromolecular assemblies from crystalline state, *J. Mol. Biol. 372*, 774-797.

302. Collins, T., D'Amico, S., Georlette, D., Marx, J. C., Huston, A. L., and Feller, G. (2006) A nondetergent sulfobetaine prevents protein aggregation in microcalorimetric studies, *Anal. Biochem. 352*, 299-301.

303. D'Amico, S., and Feller, G. (2009) A nondetergent sulfobetaine improves protein unfolding reversibility in microcalorimetric studies, *Anal. Biochem. 385*, 389-391.

304. Salonen, L. M., Ellermann, M., and Diederich, F. (2011) Aromatic Rings in Chemical and Biological Recognition: Energetics and Structures, *Angewandte Chemie-International Edition 50*, 4808-4842.

305. Bacik, J. P., Avvakumov, G., Walker, J. R., Xue, S., and Dhe-Paganon, S. (PDB ID 3JYU) Crystal structure of the N-terminal domains of the ubiquitin specific peptidase 4 (USP4), Structural Genomics Consortium.

306. Chang, C., Skarina, T., Kagan, O., Savchenko, A., Edwards, A. M., and Joachimiak, A. (PDB ID 2RFQ) Crystal structure of 3-HSA hydroxylase from Rhodococcus sp. RHA1, Midwest Center for Structural Genomics.

307. Fraser, M. E., Cherney, M. M., Marcato, P., Mulvey, G. L., Armstrong, G. D., and James, M. N. G. (2006) Binding of adenine to Stx2, the protein toxin from *Escherichia coli* O157:H7, *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 62*, 627-630.

308. Fraser, M. E., Fujinaga, M., Cherney, M. M., Melton-Celsa, A. R., Twiddy, E. M., O'Brien, A. D., and James, M. N. G. (2004) Structure of shiga toxin type 2 (Stx2) from *Escherichia coli* O157:H7, *J. Biol. Chem. 279*, 27511-27517.

309. Jung, T.-Y., Kim, Y.-S., Oh, B.-H., and Woo, E. (2013) Identification of a novel ligand binding site in phosphoserine phosphatase from the hyperthermophilic archaeon *Thermococcus onnurineus*, *Proteins 81*, 819-829.

310. Wang, H., Morita, M., Yang, X., Suzuki, T., Yang, W., Wang, J., Ito, K., Wang, Q., Zhao, C., Bartlam, M., Yamamoto, T., and Rao, Z. (2010) Crystal structure of the

human CNOT6L nuclease domain reveals strict poly(A) substrate specificity, *EMBO J. 29*, 2566-2576.