Mapping the structure of UDP-glucose:glycoprotein glucosyltransferase

Daniel E. Calles-Garcia
Department of Biochemistry
McGill University
Montreal, QC

This thesis was submitted to McGill University after completion of the requirements for the degree of Doctor of Philosophy.

June 2016

Table of Contents

Abbreviations

AoUGGT	*Aspergillus oryzae* UGGT
CNX/CRT	calnexin and calreticulin
DmUGGT	*Drosophila melanogaster* UGGT
EM	electron microscopy
ER	endoplasmic reticulum
ERAD	endoplasmic reticulum associated degradation
ERAF	endoplasmic reticulum associated folding
GST	glutathione S-transferase
HDX-MS	hydrogen-deuterium exchange and mass spectrometry
HOP	Hsp70-Hsp90 organizing protein
HSP	heat shock protein
MHC	major histocompatibility complex
SA	monovalent streptavidin
OST	oligosaccharyl transferase complex
PcUGGT	*Penicillium chrysogenum* UGGT
SAXS	small angle x-ray scattering
SEC-MALS	size exclusion chromatography & multi-angle light scattering
Sep15	selenoprotein of 15 kDa
Sep15cr	cysteine rich domain of Sep15
UGGT	uridine-5'-diphosphate-glucose:glycoprotein glucosyltransferase

List of Figures

List of Tables

Abstract

The enzyme UDP-glucose:glycoprotein glucosyltransferase (UGGT) is a key contributor to glycoprotein folding and to proper immune system function. The most characterized function of UGGT is that of a misfolded protein sensor in the endoplasmic reticulum (ER), where it selectively labels unfolded glycoproteins and sends them for chaperone-assisted folding. Prior to my work, the structure of UGGT and the mechanism for selection of misfolded proteins were unknown. First, I produced and characterized UGGT from various species, yielding pure and active protein adequate for structural analysis. Second, I obtained low-resolution models of UGGT through two different techniques: small angle x-ray scattering (SAXS) *ab initio* modeling and negative-stain electron microscopy (EM). Third, I defined the domain positions within UGGT by making streptavidin-labeled UGGT samples and carrying out negative-stain EM studies, identifying the catalytic and sensor domains. Fourth, I determined a medium-resolution cryo-EM structure of UGGT and fit homology models of UGGT individual domains within the structure. Lastly, in collaboration with Dr. Naoto Soya, we determined the binding site of Sep15 on UGGT using hydrogen/deuterium exchange and mass spectrometry experiments. Taking all the data into consideration, we propose a selection mechanism where the catalytic site of UGGT is located deep in a cavity lined by hydrophobic patches, which can only be reached by the glycan on or near flexible hydrophobic loops of the misfolded protein.

Résumé

L'enzyme UDP-glucose:glycoprotéine glucosyltransférase (UGGT) est un contributeur clé pour le repliement des glycoprotéines et pour le bon fonctionnement du système immunitaire. La fonction de UGGT la mieux étudiée est celle de détecteur de protéines mal-repliées dans le réticulum endoplasmique (ER), où cet enzyme ajoute une étiquette de manière sélective aux glycoprotéines dénaturées et en les dirigeant vers des cycles supplémentaires de repliement avec des chaperonnes. Avant cette étude, la structure de UGGT était inconnue et il n'existait pas de mécanisme de sélection pour les protéines mal-repliées par UGGT. En premier, j'ai produit et caractérisé UGGT de différentes espèces, en obtenant des échantillons de protéine active et de pureté adéquate pour mener des analyses structurales. Deuxièmement, j'ai obtenu des modèles de faible résolution de UGGT en utilisant deux différentes techniques: modélisation *ab initio* à partir de données de réfraction de rayons x à faible angle (SAXS) et par microscopie électronique (EM) à coloration négative. Troisièmement, j'ai défini la position des domaines de UGGT en utilisant des échantillons de UGGT étiquetés avec streptavidine et en faisant une analyse par EM à coloration négative, identifiant le domaine catalytique et le domaine senseur. Quatrièmement, j'ai déterminé une structure à résolution moyenne de UGGT par cryo-EM et j'ai positionné des modèles de homologie des domaines individuels de UGGT dans la structure. Cinquièmement, en collaboration avec Dr. Naoto Soya, on a déterminé le site d'interaction entre Sep15 et UGGT en utilisant des expériences d'échange d'hydrogène/deutérium couplé à la spectrométrie de masse. En considérant toutes les données acquises, on propose un mécanisme de sélection où le site catalytique de UGGT est localisé dans une profonde cavité hydrophobique, dans laquelle ont accès uniquement les chaînes glycosidiques portées sur ou proches de boucles hydrophobes flexibles de la protéine mal repliée.

Acknowledgements

Many people made this work possible. The most influential is Dr. Kalle Gehring, who welcomed me into his laboratory, gave me the opportunity to work on this project and encouraged me to try new approaches on this enzyme. Thanks to his corrections and reviews, my thesis was vastly improved. Throughout my work, I benefited from the guidance and insights of my research advisory committee members: Dr. Albert Berghuis, Dr. Justin Kollman and Dr. Huy Bui Khanh.

The expression of DmUGGT using insect cells relied on a construct and advice from Dr. David Y. Thomas, Dr. Daniel C. Tessier and Dr. Malcolm Whiteway. Sara Bastos and Dr. Katalin Kocsis introduced me to the insect cell protein expression and answered all my questions. Our Japanese collaborators, Dr. Yoichi Takeda and Dr. Yukishige Ito provided the constructs for Pc- and AoUGGT expression. They also provided the M9-MTX synthetic substrate, used for crystallization trials of UGGT.

Dr. Marie Ménade and Dr. Guennadi Kozlov were responsible for the initial constructs for Sep15 bacterial expression and for the PcUGGT mutants for biotinylation experiments and monovalent streptavidin labeling. Dr. Jean-François Trempe taught me the experimental conditions for UGGT activity assays using denatured glycoprotein substrates. Dr. Dmitry Rodionov taught me how to prepare samples, carry out and trouble shoot SAXS experiments, as well as the basics of data processing. Dr. Naoto Soya collected the HDX-MS data for DmUGGT and DmUGGT/Sep15, which allowed the identification of the Sep15-binding region of UGGT.

Dr. Justin Kollman taught me how to prepare samples for negative stain EM and cryo-EM, microscope handling, data collection methods and the fundamentals of EM data processing. Dr. Kaustuv Basu and Dr. Kelly Sears from the FEMR at McGill University and Dr. David Veesler at the University of Washington were of great help as well for microscope handling and troubleshooting.

Finally, my work was funded by the CIHR through the Chemical Biology Training Program Scholarship, and by the NSERC through the Bionano CREATE Training Program Scholarship. I was also the beneficiary of McGill University's Biochemistry Department Graduate Excellence Fellowship. I am immensely grateful to all my sources of funding, without which none of this work would have been possible.

Preface and Contribution of Authors

The UGGT enzyme is responsible for detecting misfolded glycoproteins in the endoplasmic reticulum of eukaryotes. It distinguishes between well-folded and defective glycoproteins, glucosylating and directing the latter for chaperone-assisted refolding. In this study, we used biochemical, biophysical and modeling techniques to better understand the structure of UGGT and to propose a structure-based mechanism for glycoprotein selection. This thesis provides structural data for UGGT, which had been extremely limited in the literature, and provides new research avenues to explore.

In the first chapter, I produced and purified Dm-, Pc- and AoUGGT and used the purified enzymes to collect SAXS data. Drs. David Y. Thomas, Daniel C. Tessier and Malcolm Whiteway provided the construct for DmUGGT. Drs. Yoichi Takeda and Yukishige Ito provided the constructs for Pc- and AoUGGT as well as the M9-MTX synthetic substrate. Drs. Marie Ménade and Guennadi Kozlov made the initial Sep15 construct, which I produced and purified and later modified to produce GST-Sep15cr. I carried out extensive crystallization trials with UGGT samples and complexes, but no crystals were obtained. The SAXS data I collected provided the first characterization of UGGT in solution, yielding new information on the size, fold and behavior of this enzyme in solution.

In the second chapter, I chose negative stain electron microscopy (EM) to reconstruct low-resolution structures of Pc- and DmUGGT. Dr. Justin Kollman provided me with invaluable training in data collection and analysis. As a cross-validation method, I used the SAXS data to calculate bead-models of Pc- and DmUGGT. The structures obtained showed conserved features across species and across techniques. The models show that UGGT has a claw shape, with a large central cavity between a large domain and a smaller hook-shaped domain. Though low-resolution, these models are the first of full length UGGT. Based on the structure, I proposed a simple mechanism for substrate selection, where the interior of the cavity contains both the glucosyltransferase active site and the misfolded protein sensor site.

In the third chapter, we used a combination of UGGT mutants, EM and homology modeling to pinpoint UGGT domains within the EM maps and to improve the proposed mechanism of action. Drs. Marie Ménade and Guennadi Kozlov made sixteen

PcUGGT constructs for biotinylation and purified five PcUGGT mutants with site-specific streptavidin labels. I collected negative stain EM data on the streptavidin-labeled PcUGGT and reconstructed maps for three of them. This way, we identified the positions of the Trx1, Trx3 and catalytic domains of UGGT. With the training and guidance of Dr. Justin Kollman, I collected cryo-EM data for DmUGGT and reconstructed a structure of UGGT to ~10 Å resolution. I calculated homology models for every domain of Dm- and PcUGGT, which I positioned within the EM maps. We found that the large lobe of UGGT is the sensor domain and is composed of the N-terminal domain and the Trx1-3 domains. Additionally, we found that the smaller lobe of UGGT is the catalytic domain, connected by the beta domain to the larger sensor domain. Altogether, we proposed a more detailed mechanism, where the top surface of the cavity contains a hydrophobic surface for misfolded protein sensing and the bottom of the cavity contains the catalytic glucosyltransferase site.

In the fourth and final chapter, we studied the DmUGGT/Sep15 complex using hydrogen-deuterium exchange coupled to mass spectrometry experiments (HDX-MS). I produced and purified the DmUGGT/Sep15 complex and characterized the complex using SEC-MALS and biochemistry. Dr. Naoto Soya, from the laboratory of Prof. Gergely Lukacs, collected mass spectrometry data under reducing and non-reducing conditions and performed HDX-MS experiments. The mass spectrometry results found various disulfide bonds, which were consistent with the homology models. More importantly, the HDX-MS data identified a 15-residue region of UGGT within the Trx1 that had reduced hydrogen-deuterium exchange in the presence of Sep15. All together, we concluded that this 15-residue region of UGGT Trx1 is the docking site for the cysteine-rich domain of Sep15, which was previously unknown.

I wrote this thesis and prepared two manuscripts covering the work of chapters I, II and III. This thesis, the manuscripts and this entire body of work have been carried out and reviewed under the guidance of Dr. Kalle Gehring, my supervisor, who has provided continuous support, ideas and encouragement throughout my PhD project.

I dedicate this work to…

Olivia Rose Saucier,
For her unwavering love and companionship,

Maria Valladares and Diego Benjamin Calles,
For looking after me for so many years and for being amazing role models,

Ana Marina Garcia and Francis Alberto Calles, my parents,
For giving me the strength and values to succeed in any endeavors.

1 Introduction

Cover figure:

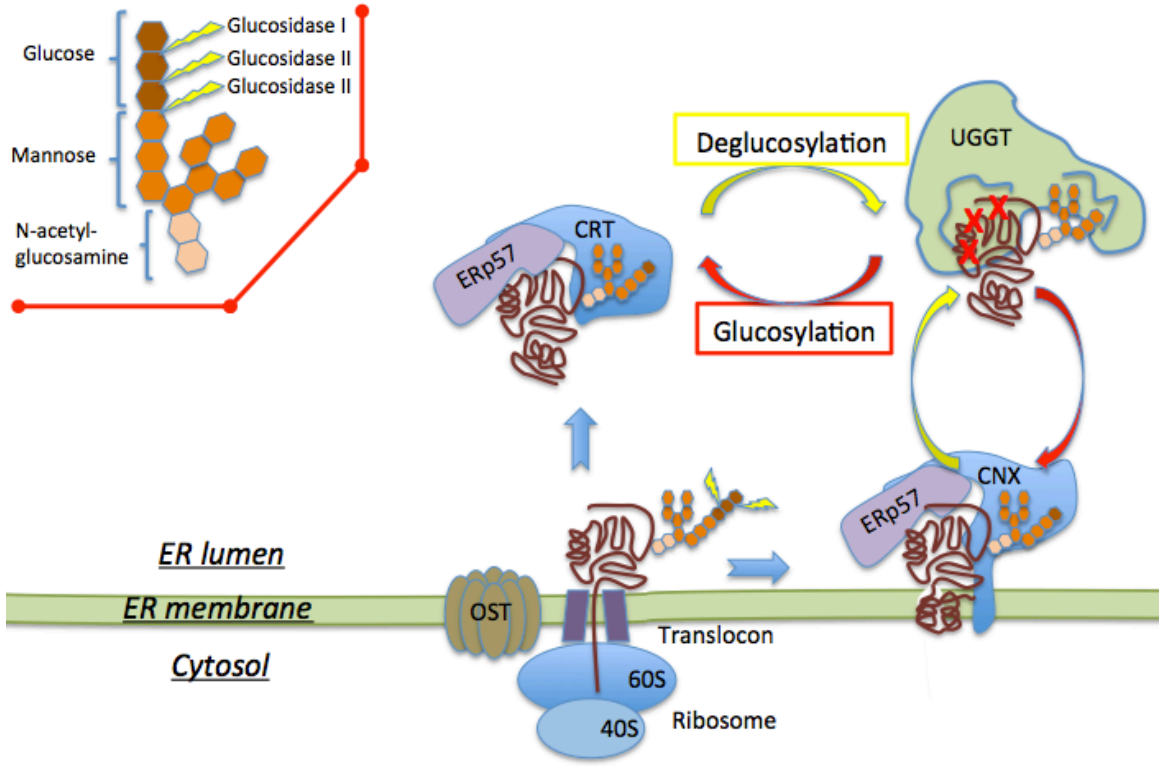


Figure 1: The Calnexin Cycle and UGGT in the endoplasmic reticulum

The oligosaccharyl transferase (OST) complex transfers the N-glycosylation (inset) onto the side chain of asparagine of **Asn**-X-Ser/Thr sequences. The glycan is trimmed by glucosidase I and II. The glycoprotein in its monoglucosylated form interacts with calnexin/calreticulin and with UGGT in its deglucosylated form.

1.1 Literature Review on Biological Context

In all cells, there is a dynamic equilibrium between protein synthesis and protein degradation. With the high number of proteins being synthesized, there are important folding pathways and quality control mechanisms in place to keep cells and organisms working efficiently. Some proteins will fold correctly by themselves, while others need assistance from other proteins to fold correctly. Proteins that don't acquire a native fold risk aggregating, causing damage inside of the cell. Proteins that present relatively minor

defects in their fold are described as misfolded, whereas those that present major defects or lack any folding are described as unfolded. To prevent defectively folded proteins from aggregating and causing cellular damage, a complex machinery in the cytosol[1, 2] and in the endoplasmic reticulum[3] (ER) detects misfolded proteins to correct their folding or to direct them for degradation by the proteasome. At the core of the folding machinery, there are chaperone proteins whose role is to bind misfolded or unfolded proteins, to keep them from aggregating and to help them acquire their native fold.

Equally important to the work of chaperones in the ER, there is a misfolded glycoprotein sensor ensuring the quality control of glycoprotein folding. This task is ensured by UDP-glucose:glycoprotein glucosyltransferase (UGGT), which prevents misfolded glycoproteins from leaving the ER and redirects them for additional chaperone assisted refolding. The objective of this thesis is to better understand how UGGT selectively recognizes misfolded substrates by collecting structural data using a variety of biochemical and biophysical techniques. The goal is to provide functional understanding of this enzyme through structural characterization, which is very limited to date. Though UGGT is not directly involved in any pathology, it is an essential housekeeping enzyme with a unique mode of action, which needs to be better understood.

1.1.1 Protein synthesis by the ribosome

Ribosomal protein synthesis starts in the cytosol with a ribosome binding to protein-coding messenger ribonucleotide acid (mRNA). Through scanning the mRNA sequence and sequential addition of amino acids, the ribosome will assemble a protein, which will start folding as it exits the ribosome. Cytosolic proteins will be entirely translated in the cytosol. Proteins targeted for the ER carry an N-terminal signal sequence, which upon recognition by the signal recognition particle (SRP) halts translation until the ribosome-SRP complex binds to the SRP receptor, on the ER membrane. Once the ribosome binds to the translocon Sec61 complex, translation of the protein resumes, and the polypeptide chain enters the ER lumen through the translocon complex[4-6].

1.1.2 Protein structure and folding

The folding of proteins is a complicated process, where the polypeptide chain needs to adopt a very specific three-dimensional arrangement where hydrophobic regions are buried within the core of the protein and hydrophilic regions and side chains are exposed to the solvent (see Figure 2). The polypeptide chain slowly adopts a combination of alpha helices and beta sheets connected by loops, referred to as secondary structure. Through a combination of hydrophobic and long-range inter-residue interactions, the secondary structure elements fold together in compact structures, which then fit tightly against each other in a specific conformation to yield the tertiary structure of the protein. Multiple proteins can then combine together to form a biologically active complex in a specific tridimensional arrangement, referred to as quaternary structure. Proteins will follow different folding pathways and different folds, depending on their roles and whether they are soluble or membrane proteins.

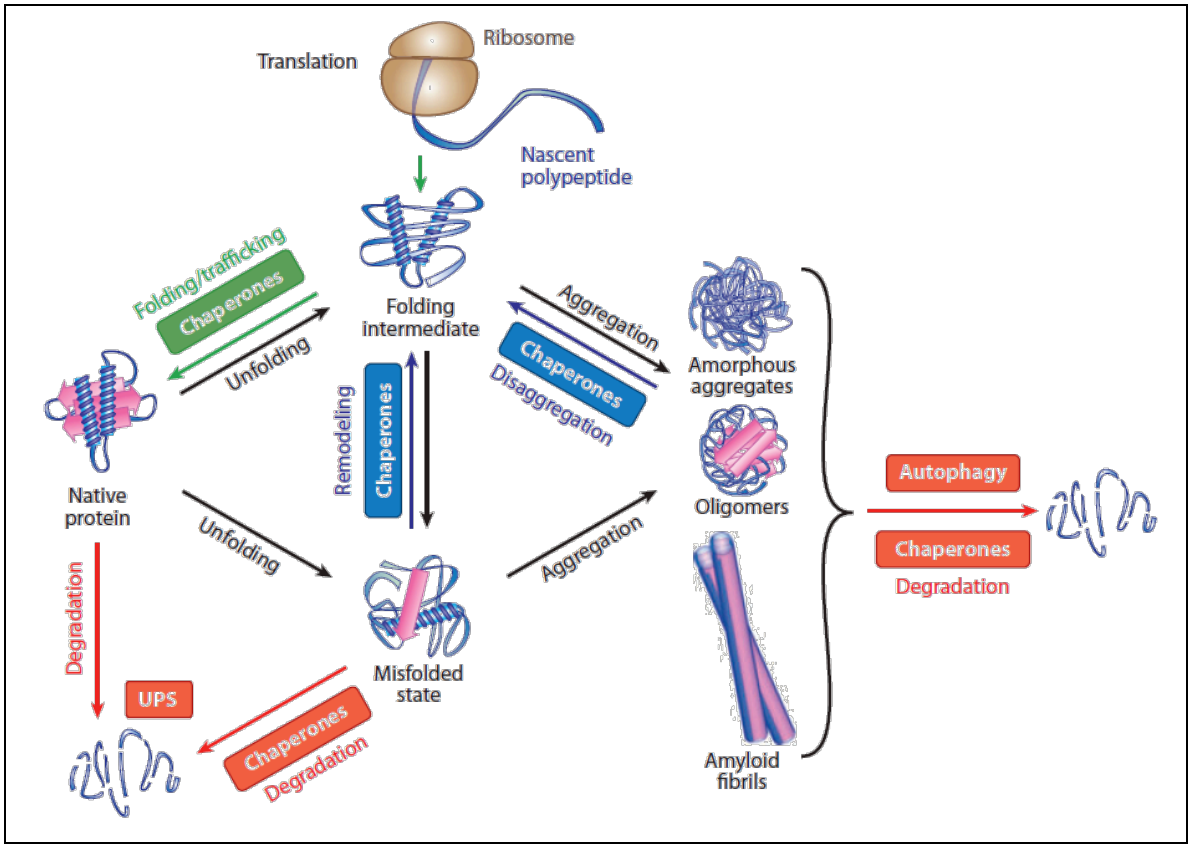


Figure 2: Protein folding pathways and chaperones

Protein folding can start the moment the polypeptide emerges from the ribosome. Once fully released, proteins can fold spontaneously into a native fold or can be helped by chaperones. The roles of chaperones are numerous, from preventing protein aggregation, remodeling protein fold, helping traffic protein to the right cellular compartment, or guiding terminally misfolded proteins towards degradation by the proteasome (figure source [2]).

1.1.3 Protein folding by chaperones in the cytosol

Some cytosolic proteins can fold spontaneously into their native structure, but others need help from other proteins, such as chaperones (see Figure 3). A first family of chaperones is HSP70, which bind to seven-residue long hydrophobic regions of unfolded proteins. These chaperones prevent unfolded proteins from aggregating, are regulated by ATP/ADP and by co-chaperones such as DNAJ proteins. A second family of chaperones is HSP90, which form homodimers. These chaperones bind to misfolded proteins in a clamp-like fashion, using ATP/ADP to induce rearrangements in the client protein. HSP70 and HSP90 can form complexes through adaptor proteins such as HOP, to collaborate in the folding of proteins[2, 7].

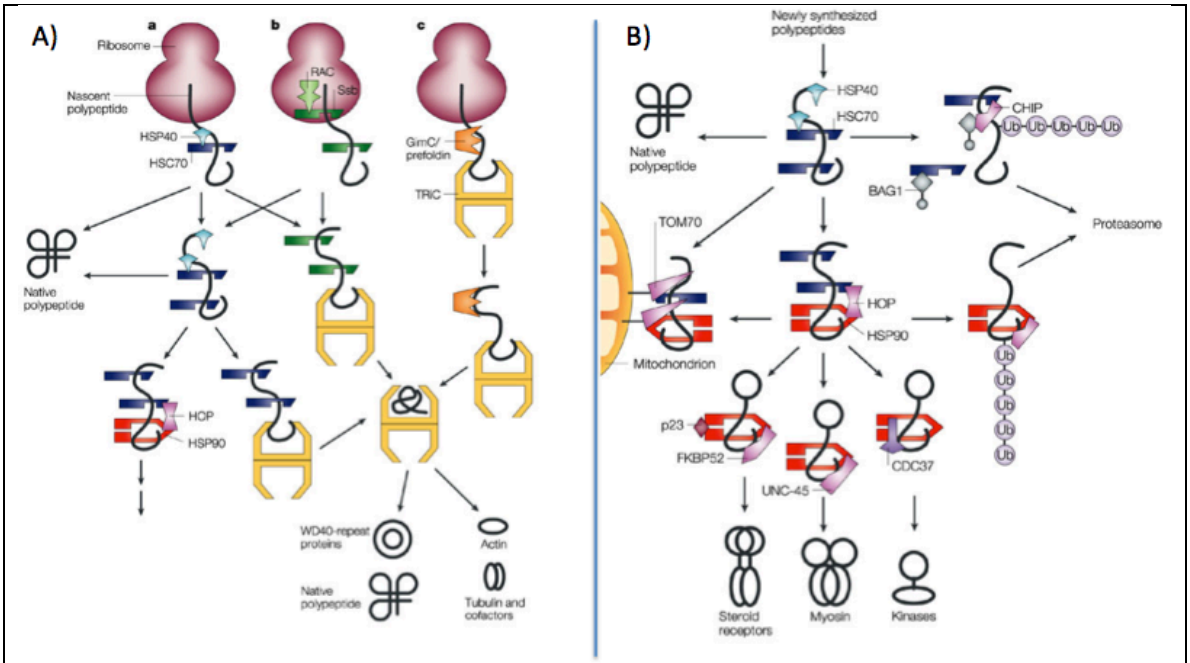


Figure 3: Cytosolic protein folding machinery: Hsp70, Hsp90 and Hsp60

HSP70 family members bind to unfolded proteins to keep them soluble and allow them to fold on their own or to bring them to other chaperones for assistance. HSP90 family members bind to partly folded proteins and can form complexes with HSP70 members through HOP. HSP60 forms a large double-sided "cage" where unfolded proteins are caught, isolated from other proteins, and protein folding is fostered through a progressive change in the hydrophobicity of the interior of the cage. HSP70 and HSP90 can guide ubiquitinated, misfolded proteins towards the proteasome for degradation and recycling (figure source [7]).

A third family of chaperones is the HSP60 chaperonins, best represented by the GroEL/ES large oligomeric complex (see Figure 3, left panel). The GroEL proteins assemble into two back-to-back large ring complexes, each of which has a large central cavity where a misfolded protein can be captured. Once the unfolded protein is present in the cavity of one of the rings, the subunits bind ATP and a conformational rearrangement takes place, followed by binding of GroES, which closes the barrel and traps the unfolded protein. The chamber initially provides a hydrophobic environment but upon ATP hydrolysis to ADP, the chamber surface becomes more hydrophilic and

forces the protein inside to fold. Once all ATP has been converted to ADP, the GroES cap dissociates and the folded protein is released.

1.1.4 Protein folding in the ER and the calnexin cycle

Proteins targeted for the secretory pathway and plasma membrane proteins go through the ER, where they are N-glycosylated before proceeding to the Golgi apparatus for further processing (see Figure 4). These proteins typically carry a N-terminal ER targeting sequence. N-glycosylation is an abundant post-translational modification[3, 4, 8, 9], in which the glycan Glc_3-Man_9-$GlcNAc_2$ is transferred onto the asparagine residue on Asn-X-Ser/Thr sequences by the oligosaccharyltransferase complex[10]. The N-glycan is then trimmed by glucosidase I and II, allowing the monoglucosylated glycoprotein to interact with ER-specific lectin chaperones (see Figure 5), calnexin and calreticulin, which will help the glycoprotein fold[11-14] and recruit additional chaperones to fold the nascent protein[15, 16]. In contrast to HSP chaperones that bind to a protein misfolded regions, calnexin and calreticulin are lectin chaperones that bind to client proteins by recognition of the monoglucosylated N-glycan[12, 13]. Upon release by these chaperones, the terminal glucose is cleaved by glucosidase II. Calmegin and calsperin are two additional lectin chaperone specific to the ER, though less well studied. Incompletely folded proteins are recognized by UDP-glucose:glycoprotein glucosyltransferase (UGGT), which selectively adds a glucose residue to the Man_9-$GlcNAc_2$ glycan to regenerate the monoglucosylated form for additional rounds of lectin chaperone assisted refolding[17-19] (see Figure 1).

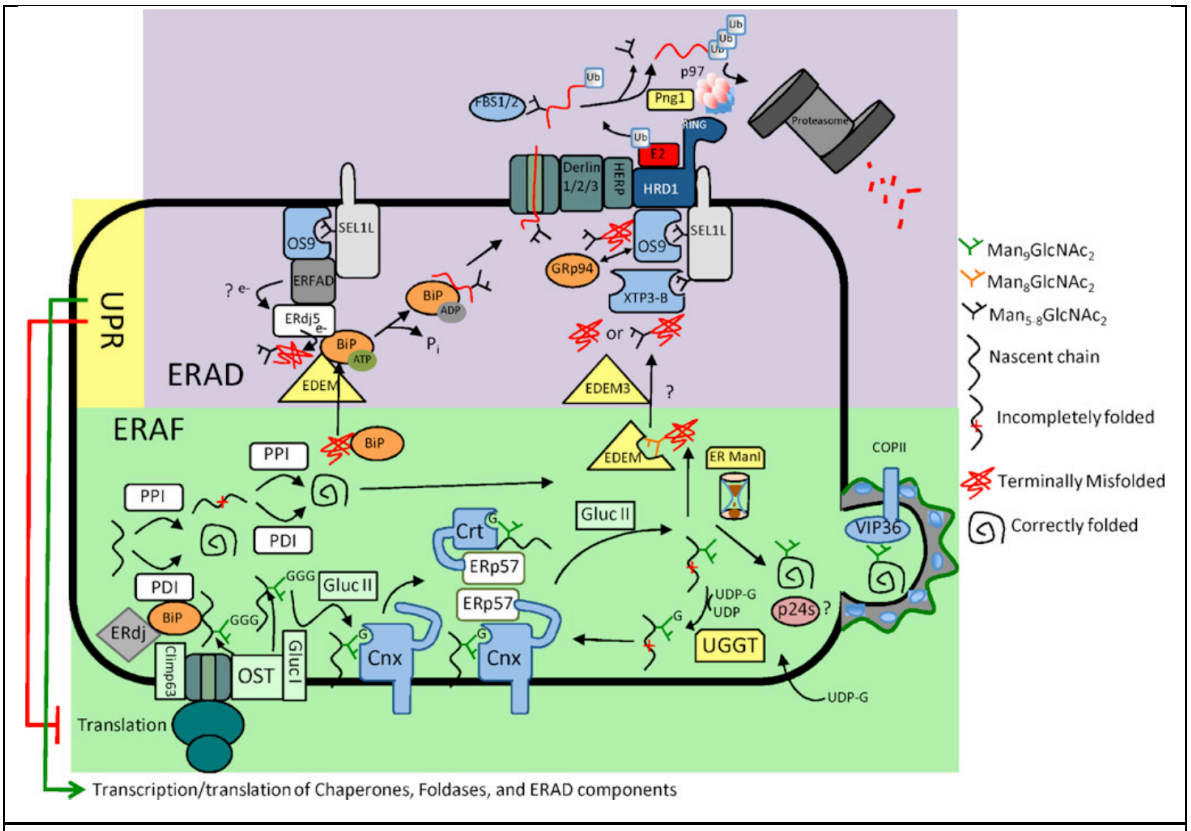


Figure 4: Folding (ERAF) and degradation (ERAD) in the ER

Folding is helped through members of the HSP family such as BiP and GRP94, through lectin chaperones such calreticulin and calnexin, and through various enzymes such as PDI, PPI, ERp57 and UGGT. Folded proteins leave the ER through COP-II vesicles towards the Golgi apparatus. For degradation, specific mannose residues on the glycan are trimmed, leading to escorting of the protein by EDEM or BiP towards a large membrane protein complex for export to the cytosol, ubiquitination and immediate degradation by the proteasome (image source [3]).

The calnexin/calreticulin binding and glucose-labeling by UGGT is known as the calnexin cycle (see Figure 5). The ER also contains other chaperones such as BiP, an HSP70 family protein, and ERdj proteins, members of the DNAJ co-chaperones; GRP94 proteins, part of the HSP90 family; and proteins capable of rearranging disulfide bonds, such as protein disulfide isomerase (PDI) or ERp57[3, 20]. Multichaperone complexes exist, just as in the cytosol, as ERdj proteins bring different chaperones together in a complex. Cyclophilin B is another ER component known to interact with PDIs, GRP94, BiP and

the lectin chaperones, likely participating in multichaperone complexes[12, 21]. Once the ER folding machinery has given a glycoprotein its native fold, the well folded glycoprotein can proceed to the Golgi in the secretion pathway[22]. Proteins that fail to acquire an adequate fold are eventually directed towards ER associated degradation [23].

1.1.5 The misfolded protein sensor UGGT

The role of UGGT is to detect misfolded glycoproteins and glucosylate the N-glycan, in order for the glycoprotein to interact with calnexin or calreticulin, which bind to monoglucosylated misfolded glycoproteins (see Figure 1 and Figure 5). This enzyme is conserved and essential across eukaryotes[24-26]. It is a large monomeric protein more than 1500 residues long, known to interact with two other ER proteins: Sep15, a 15 kDa selenoprotein that might contribute to disulfide bond breaking[27-31], and with ERdj3[21, 32], a DNAJ family protein that interacts with BiP and GRP94, which are HSP70 and HSP90 members respectively. In terms of clinical relevance, cancer sequencing has detected many multigene copy number variations as well as examples of single nucleotide polymorphism in the UGGT gene (see Figure 6). However, no missense mutations in UGGT have been associated with any pathology or clinical cases.

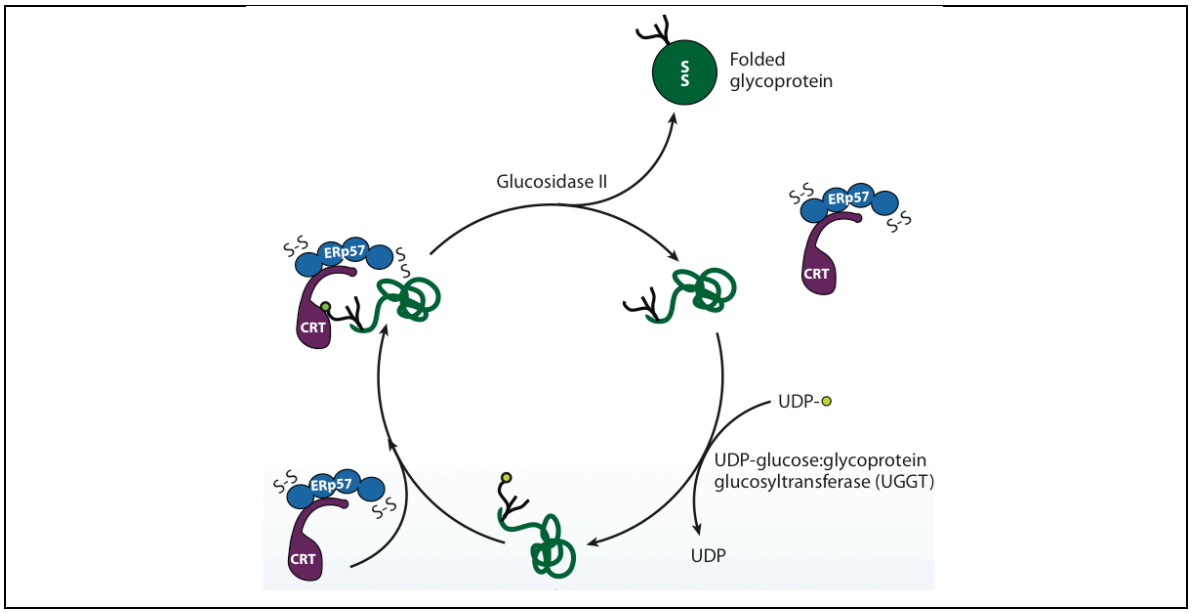


Figure 5: The Calnexin Cycle and UGGT activity on misfolded proteins

The monoglucosylated glycoprotein interacts with the lectin chaperone complex CRT/ERp57 or CRX/ERp57, capable of breaking incorrect disulfide bonds to help the protein acquire its native form. Upon release, glucosidase II removes the terminal glucose residue. At this point, the fold is verified by UGGT and glucosylation occurs on misfolded proteins, which can then interact with the lectin chaperones (source [33]).

The activity of UGGT has been extensively probed using glycoproteins in both native and misfolded states[18, 19, 34-38], glycopeptides[39-41] and small synthetic substrates[42-47]. Some general rules for misfolded substrate glucosylation have been found but no mechanism has been proposed for the mechanism of substrate selection and glucosylation. UGGT glucosylates misfolded substrates only and leaves native substrates unglucosylated. It binds to both the exposed hydrophobic loops and the 9-mannose-2-N-acetylglucosamine glycan on the substrate (see Figure 1). Both motifs need to be within 40 Å or less of each other and part of the same molecule to trigger glucosylation[38, 40, 48, 49]. In a simple labeling experiment where unfolded glycoproteins and non-glycoproteins were both present in the reaction mixture, the labeling activity of UGGT was remarkably reduced[34]. This competitive effect of unfolded proteins on UGGT activity points to one single binding pocket or cavity for substrate recognition and glucosylation by UGGT.

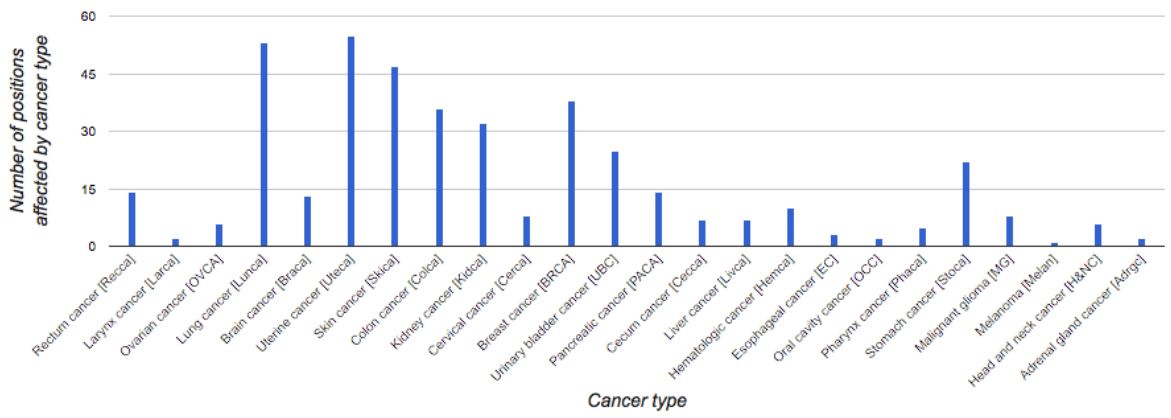


Figure 6: Single nucleotide polymorphism of UGGT in cancer

Sequencing of various cancer genomes has revealed the polymorphism of UGGT gene,

though no variants have been shown to have a clinical relevance. Lung cancer, uterine cancer and skin cancer present the largest number of mutations to the UGGT gene.

The study of UGGT has not been without controversy. The distance between the misfolded region that is recognized and between the glycan has been the subject of much debate. Though it has become obvious that UGGT can glucosylate misfolded substrates of any size, an initial study[39] used mass spectrometry on glycopeptides showing an ideal UGGT substrate to be at least 12 amino acids long, with hydrophobic residues within 14 residues from the glucosylation site. Later experiments have shown that the glycan can be as far as 40 Å away from the misfolded region of the protein[48] and that the glycan does not necessarily need to be within 14 residues of the glucosylation site. The 40-Å cutoff seems to have been validated, as a study with synthetic substrates used as molecular rulers[43, 50] showed that the reactivity of the substrate decreased as the length of the substrate is increased beyond ~40 Å.

Another point of contention is what the best UGGT substrate is. Initial studies regarded misfolded glycoproteins, in a molten globule form, as the ideal UGGT substrate[22, 34-36, 49]. However, a wave of studies using synthetic UGGT substrates[42-44] has shown clearly that the combination of a glycan and a hydrophobic molecule such as methotrexate or BODIPY are far more reactive to UGGT than any other molten globule glycoprotein. The latest point of discussion is whether or not UGGT2, a second homolog of UGGT in mammals, is active or not. The studies using molten globule glycoproteins concluded UGGT2 to be inactive[51, 52]. However, the recent use of synthetic UGGT substrates has shown UGGT2 to be active on those substrates[46]. Additionally, in contrast to UGGT1 activity, the UGGT2 activity is greatly increased in the presence of Sep15[46, 50]. It remains to be proven if UGGT2 is active *in vivo* with real misfolded glycoproteins.

1.1.6 Role of UGGT in the immune system maturation

The UGGT enzyme contributes to the maturation of the major histocompatibility complex of class 1 (MHC1), ensuring that a high-affinity peptide has been loaded onto MHC1 heavy chains[53, 54]. The class 1 MHC is composed of MHC-encoded heavy chains, which are glycosylated, and a β_2-microglobulin domain (see Figure 7). The mature MHC1 molecules are present at the cell surface and present a peptide of 8-10

residues to the antigen-specific CD8+ T cells. If pathogens are present inside the cell, MHC1 will present a peptide originating from a pathogen-related protein, which will trigger differentiation of CD8+ T cells into cytotoxic T cells capable of killing infected cells. The assembly of complete MHC1 involves many components of the ER folding pathways[33, 53, 55]. Among them, UGGT scans the peptide-MHC1 interaction and if the loaded-peptide is not well seated within the HC binding groove, UGGT glucosylates the HC and directs MHC1 for re-association with the calreticulin/ERp57/tapasin complex for a new peptide to be loaded onto the HC binding groove. This way, UGGT ensures that only high-affinity peptides are loaded onto the HC of MHC1. Without the activity of UGGT, weak-affinity peptides can dissociate from the HC binding groove, rendering MHC1 incapable of activating CD8+ T cells.

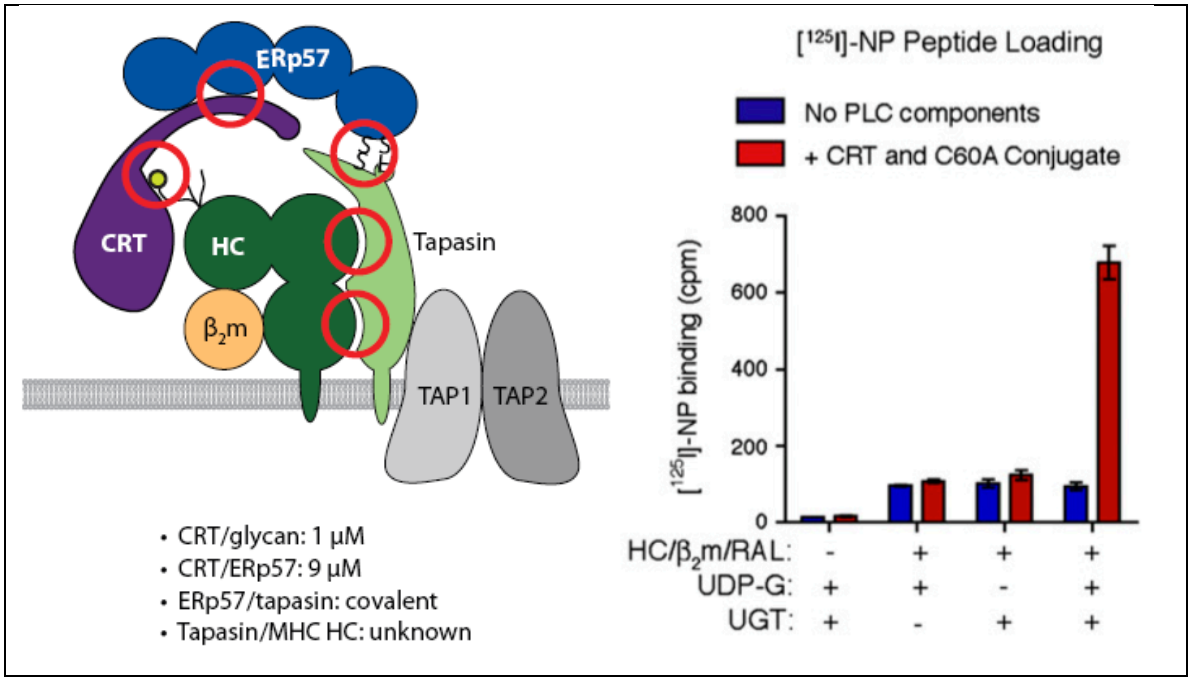


Figure 7: Assembly of MHC1 and role of UGGT in MHC1 optimization

The CRT/ERp57 complex and UGGT contribute to form a high-affinity MHC1-peptide complex, necessary for proper immune function. In a multi-protein complex formed by TAP1/2, Tapasin, CRT/ERp57 and HC/β_2m, a short peptide is loaded onto the HC binding groove (left panel). Following glucosidase II activity, UGGT verifies that the loaded peptide is tightly bound to the HC's groove. In the case of loosely bound peptides (experiment on right panel), the HC is reglucosylated and the peptide-loaded complex loads a different peptide in order to achieve a stronger and more stable complex (source [33, 53]).

1.1.7 Structural and bio-informatics data on UGGT

This enzyme has a N-terminal ER retention signal sequence, a sensor region accounting for 80% of the sequence and a C-terminal catalytic domain, accounting for 20% of the sequence (see Figure 8). The catalytic domain is highly conserved across species, belongs to the glycosyltransferase family 8 and is only active in the complete enzyme[25, 52]. The sensor region is less conserved among species, with ~20-40% sequence identity among species, and is responsible for misfolded protein selection. It is composed of at least four domains: three thioredoxin-like domains and a beta-sheet rich region, which connects to the catalytic domain[56].

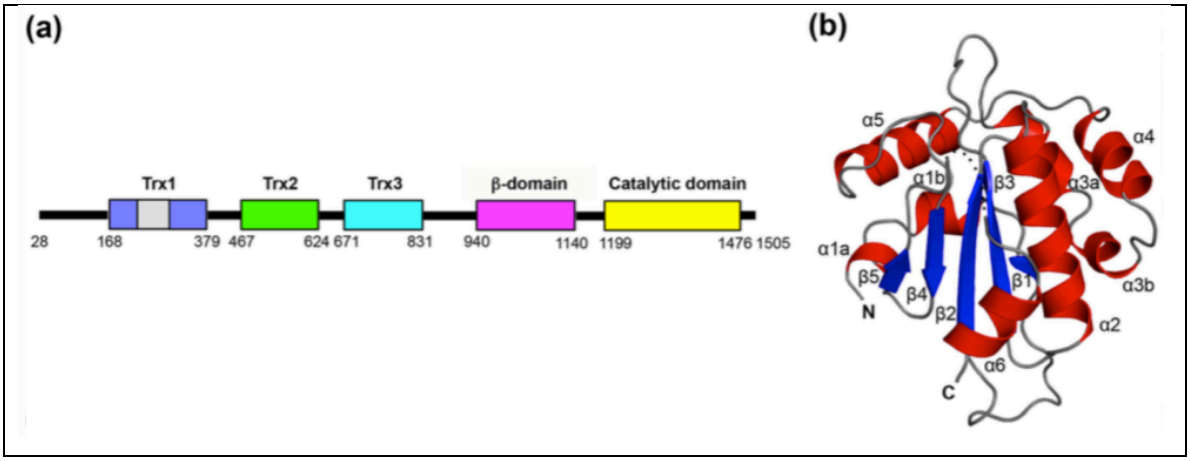


Figure 8: *C.thermophilum* UGGT domains and Trx3 crystal structure

The structure of UGGT is composed of three N-terminal thioredoxin-like domains, a beta-sheet rich region and a C-terminal catalytic domain. Long linker regions connect the domains. The structure of the third thioredoxin-like domain determined by x-ray crystallography consists of a beta-sheet core surrounded by alpha helices (source [56]).

The crystal structure of the third thioredoxin-like domain from *C. thermophillum* has been solved to 1.7 Å resolution (see Figure 8, right panel). Though it represents only ~11% of the sequence of UGGT, this structure offers some insight into the substrate recognition mechanism: the beta-sheets form an exposed hydrophobic surface, which was covered by detergents in the crystalized structure. If this feature is present in every thioredoxin-like domain in the full structure, these hydrophobic surfaces could work together as a docking surface for misfolded proteins. At a molecular weight of ~170 kDa, this enzyme has proven an extremely difficult crystallization target and is well beyond of NMR studies. The best techniques to study the complete structure of UGGT are small-angle X-ray scattering, which yields structural information of the enzyme in solution, and electron microscopy single particle reconstruction, which can achieve near-atomic resolution with state-of-the-art technologies and proteins of at least 100 kDa[57-60].

1.2 Methodologies for protein production and purification

All structural studies rely on the availability of milligrams of pure, active and homogenous protein samples as a foundation. During the first part of this thesis, the focus was placed on producing UGGT from various species in bacterial and eukaryotic expression hosts. The purification of UGGT was carried out through various chromatographic techniques, and the activity and folding status of UGGT were tested in solution. Our structural analysis was done in solution through SAXS, on negatively stained samples by EM and on vitreous-ice embedded native UGGT through cryo-EM. Various bio-informatics and modeling tools were used to supplement and interpret the experimental findings.

1.2.1 Recombinant protein production in *E. coli* bacterial cells

The easiest and fastest way to produce bacterial, yeast and some human proteins is to clone the gene of interest into expression vector optimized for bacterial expression. The bacteria *E. coli* is the most widely used bacterial host for expression of recombinant proteins; this bacterium has been widely studied, many strains are commercially available and it grows very rapidly, making it a low cost and versatile mean of protein production. However, *E. coli* is not very well suited for expression of large molecular weight proteins, glycoproteins or human membrane proteins. In this work, we produced various proteins in *E. coli*: a human selenoprotein Sep15 known to bind to UGGT, PcUGGT and PcUGGT mutants for biotinylation/EM experiments.

1.2.2 Recombinant protein production in *Sf*9 insect cells

The UGGT enzyme is an ER resident protein with several glycosylation sites, and of large molecular weight. All of these reasons make it a difficult enzyme to produce in *E. coli*. Even though PcUGGT was successfully produced and purified using *E. coli*, we chose to express the *Drosophila melanogaster* (Dm) UGGT expressed in *Spodoptera frugiperda* (*Sf*9) cells (see Figure 9) in an effort to obtain a purified enzyme in as close as possible to its native state, of paramount importance for structural studies. *Sf*9 cells are eukaryotic cells with a functional ER, capable of protein glycosylation and of producing proteins of large molecular weight in large quantities.

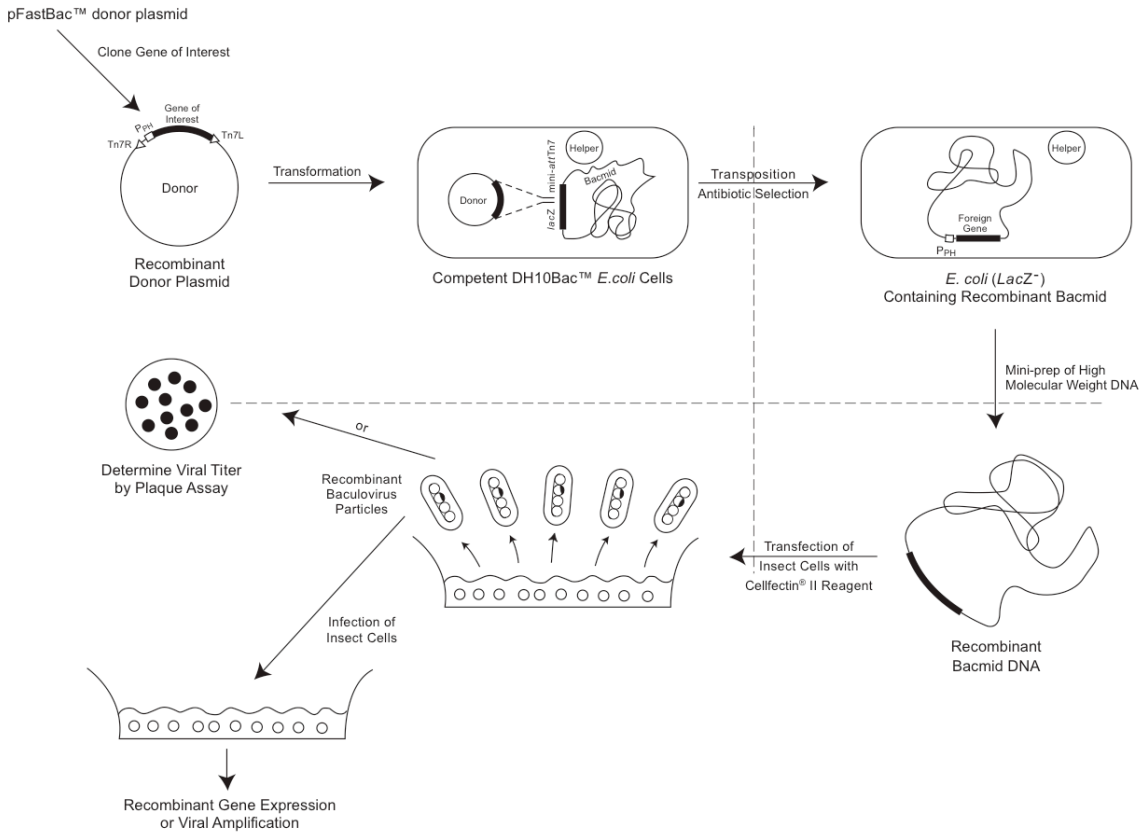


Figure 9: Protein expression using baculovirus/insect cells system

The gene of interest is cloned into a pFastBac donor plasmid, which is transformed into DH10Bac *E. coli* cells, which contain the bacmid. Following transposition, the bacmid with the gene of interest can be purified and used to transfect *Sf*9 cells. These cells will then begin producing the baculovirus and the protein of interest in small amount. Through amplification cycles, the virus titer is improved and concentrated, allowing for protein expression in high volumes and yields (source: BEVS Invitrogen manual).

The process is more complicated than in *E. coli* and the production is slower, as *Sf*9 double every ~24 hours. The gene of interest is first cloned into a bacterial plasmid, which is then introduced into a special strain of *E. coli* containing a baculovirus DNA (see Figure 9). Through recombination inside the bacteria, the gene is transferred from the initial plasmid into the baculovirus DNA, which can then be isolated and introduced into *Sf*9 cells. After transfection of *Sf*9 cells with the baculovirus DNA, the insect cells start to produce viral particles and the protein of interest in small quantities. The virus is purified

and amplified through several rounds of insect cell infection, until the viral load is high enough to infect enough culture for protein expression.

1.2.3 Purification of proteins and protein complexes

Once the protein of interest has been produced in bacterial cells or insect cells, it is necessary to purify the protein to a high degree of purity. Contaminants make any structural study of a given protein difficult to undertake. We used various techniques, starting by producing the protein with a hexa-histidine tag, which allows the use of nickel-nitriloacetic acid (Ni-NTA) resins in chromatographic columns to "fish-out" the protein, which can then be released using an imidazole gradient. The second technique we used separates proteins according to their global surface charge, using ion exchange chromatography columns. At low salt concentration, the protein adsorbs to the resin. As an increasing gradient of salt is applied, proteins are released from the column at various times. The third technique we used separates according to the size of proteins, using size exclusion chromatography columns. In this technique, the resin contains porous beads of a certain size. Small proteins can partially enter the beads and are thus slowed as they go through the column. Bigger proteins don't fit in the beads and thus pass through the column faster.

1.2.4 Unfolded RNase B labeling assay for UGGT

In order to verify that the purified samples used in our structural studies are well folded and in their native form, we tested its capacity to recognize and label a misfolded substrate. In our assays, we utilized a commercially available N-glycoprotein, RNase B, which we unfolded using urea. In the presence of radioactively labeled UDP-glucose and unfolded RNase B, UGGT can transfer radioactive glucose onto the glycan chain of RNase B (see Figure 10). After running the reaction mix in an SDS-PAGE gel, the labeled RNase B can be detected on photographic film by autoradiography. This offers a qualitative assessment of the quality of the purified samples.

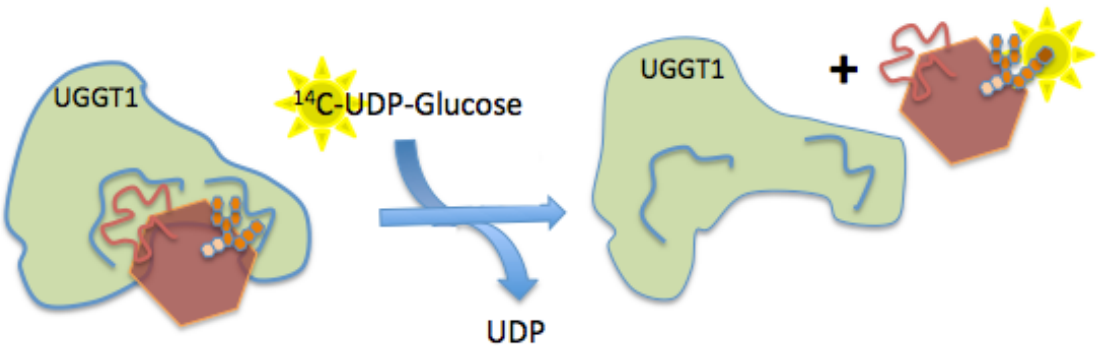


Figure 10: UGGT misfolded protein recognition and glucosylation assay

UGGT binds the exposed hydrophobic loop on the substrate through the N-terminal sensor region and the N-glycan through the C-terminal catalytic domain. Radioactive ^{14}C-glucose is transferred to the N-glycan of the substrate, releasing UDP. The labeled glycoprotein can be detected in photographic film.

1.3 Structural approach to understand UGGT

The function of proteins is tightly related to the structure they adopt in solution or in biological membranes. Determining the protein structure is key to understand their function. Various biophysical techniques are available to study protein structure at different levels of detail. The type and size of macromolecules studied can dictate the techniques that can be used to study its structure. The large molecular weight of UGGT has a big influence into the techniques that can be used.

1.3.1 Traditional techniques yielding high-resolution structure

Classically two techniques are capable of yielding atomic-resolution models of macromolecules. The first is protein crystallography and x-ray diffraction, which can be used on soluble and membrane proteins of all sizes but depends entirely on having crystals of the macromolecule in question[61]. Obtaining protein crystals that diffract to less than 4 Å is an absolute requirement for this technique. A protein crystal is formed under very specific conditions, which vary for every protein, and tens of thousands of conditions need to be screened when attempting crystallization. Once crystal-forming conditions have been found, the crystal needs to be optimized to yield high-resolution diffraction patterns, without which a structure can't be solved. Finding an initial hit and

crystal optimization can take anywhere from weeks to years. Crystallization of UGGT has been attempted extensively by our group and by many other groups, but only a small fragment of UGGT has been successfully crystallized (see Figure 8). The second technique yielding atomic resolution information is nuclear magnetic resonance (NMR), though in solution NMR is only applicable to proteins or complexes of less than 200 residues, and thus unusable to study the full-length UGGT.

1.3.2 Gaining structural data of UGGT in solution

Small angle x-ray scattering (SAXS) is used to study macromolecules in solution (see Figure 11). It is used to determine biophysical parameters, information on oligomeric states, and concentration-dependent behavior of macromolecules[62-64]. The sample is put in the path of an x-ray beam and the scattering at small angles measured; all molecules in the mixture will scatter the incoming beam, with larger molecules contributing more than smaller molecules to the scattering. The scattered x-rays are detected on a detector, and the measurements are performed for both buffer and protein-containing samples. By subtracting the buffer scattering from the protein solution scattering, the protein specific scattering is obtained, which can then be normalized to protein concentration.

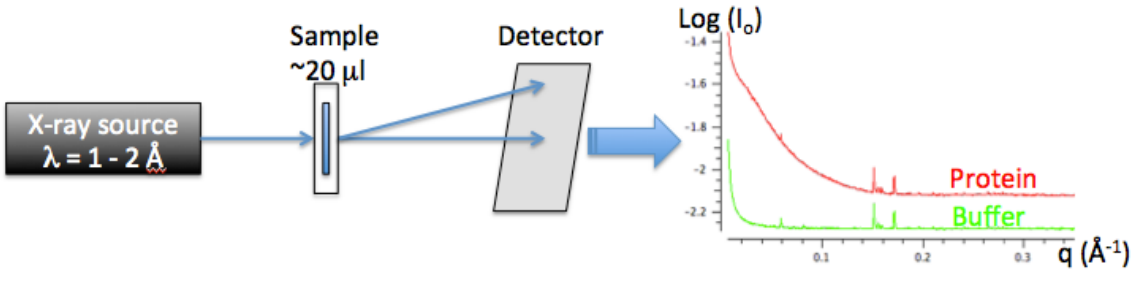


Figure 11: Small angle x-ray scattering principles

In SAXS, the x-ray scattering of molecules in solution is measured and the protein-specific scattering can be deduced by subtracting the solvent scattering. Once calculated, the protein specific scattering can be used to assess the protein shape in various buffers or at various protein concentrations. Biophysical parameters can be calculated and structural changes can be detected, by measuring data in the absence and presence of substrate or binding partners.

The analysis of protein-specific scattering (see Figure 12) will give the radius of gyration R_g, through the Guinier plot, and the maximal distance of the protein d_{max}, through the distance distribution plot. The Kratky plot is particularly useful to determine if a protein is insoluble or if there is any concentration-dependent aggregation. Using *ab initio* reconstruction techniques, the concentration-normalized SAXS data is used to generate low-resolution maps, giving an approximation of the shape of macromolecules[65, 66]. If high-resolution structures are available for the protein domains, rigid-body modeling can be used to determine the protein structure. If no structures are known for the protein being studied, it is difficult to assess whether or not an *ab initio* model is accurate, as various three-dimensional shapes can fit a single, one-dimensional scattering curve.

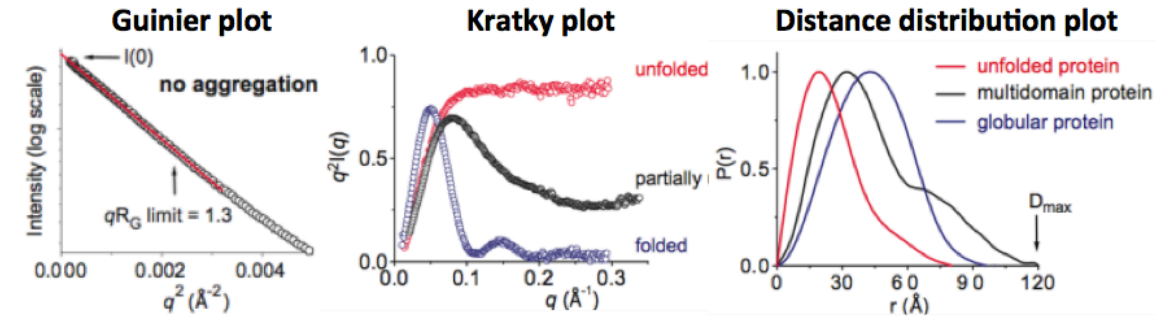


Figure 12: Analysis using Guinier, Kratky and distance distribution plots

The Guinier plot is used to detect the presence of high-molecular weight aggregates and to measure the radius of gyration of the protein. The shape of the Kratky plot reflects the folding of the protein: the curve for folded protein returns to the origin at larger angles, whereas that for unfolded proteins does not. The distance distribution curve or pair distribution plot gives an estimation of the maximal length, the compactness and shape of the protein. Globular proteins generally give a symmetric curve while multi-domain proteins can give two overlapping peaks (from Bioisis.net).

1.3.3 Low resolution structures by EM

Negative stain EM is capable of imaging proteins as small as 50 kDa and reconstructions of macromolecules can achieve resolutions of ~20 Å[67, 68]. An electron microscope utilizes a beam of electrons to image biological samples, relying on a series of magnets to focus the beam of electrons onto the sample. The resolution of transmission electron microscopes is far superior to that of light microscopes. For structural biology, micrographs were collected using photographic film but now they are routinely collected on electronic cameras in a high-throughput fashion. The biological sample is placed on a thin support, typically a copper grid with a thin layer of graphite. Because of the small difference in electron density between proteins and the supporting layer of carbon, the contrast is too small to visualize individual proteins. We increase the contrast in the micrographs by coating the proteins on the grid with a solution of electron-dense molecules such as uranyl formate or uranyl acetate. In negative stain EM, the electron-dense stain creates a dark background around proteins particles, which appear as light-gray objects on the micrographs. The micrographs can be analyzed in the computer, where stacks of proteins images are assembled, classified and eventually used to reconstruct three-dimensional models of the protein.

1.3.4 Towards high-resolution models by Cryo-EM

In recent years, the technology around electron microscopes has improved dramatically, most noticeably with the advent of direct electron detection cameras, making it possible to attain near atomic-resolution structures of high-molecular weight proteins and protein complexes by cryo-EM[57, 59, 69-71]. In contrast to negative stain EM, where the protein sample is dried in a layer of electron-dense stain, in cryo-EM the protein sample is frozen in ice captured on a copper or gold grid with very small holes. The cryo-grid is rapidly frozen in liquid ethane, leading to a very thin layer of vitreous ice on the cryo-grid, which is free of water crystals, preserving the macromolecules in their native conformation[72]. The samples are then kept at liquid nitrogen temperatures and the images are taken in the holes of the cryo-grid, where proteins are encased in the thin vitreous ice. Because the difference in electron density between water molecules and macromolecules is not very high, the contrast in cryo-EM is low. Contributing to the low contrast is the dose of electrons used in cryo-EM which must be minimized to avoid

radiation damage due to the high energy of the electron beam[73]. Direct detection cameras and advanced processing techniques[74-79] have alleviated these problems in several ways. These new cameras accurately count electrons for each pixel, leading to a high sensitivity and accuracy in detection. Furthermore, they can collect a large number of frames per second, allowing for compensation of movement so that individual frames can be aligned to improve the sharpness of the images[80, 81]. Advanced algorithms can account for beam-induced particle movement in the frames, which has led to near atomic resolution structures calculated by cryo-EM[58, 82-85].

1.3.5 Bio-informatics and homology modeling

In the absence of structural information for UGGT or its domains, we can use bio-informatics tools to build models of UGGT domains based on their similarity to proteins of known structure[86-88]. These models can then be incorporated within our electron density maps, useful in guiding mutagenesis studies or in postulating hypothesis of UGGT regulation mechanism[89]. As previously discussed, UGGT is a large enzyme with multiple domains. We utilized Phyre2[86], a server-based software, to build models of all domains, of both Pc- and DmUGGT species. In both cases, the domains fit well within our electron microscopy structure and helped improve our hypothesis for UGGT substrate recognition mechanism.

2 Production and characterization of UGGT in solution

Cover figure:

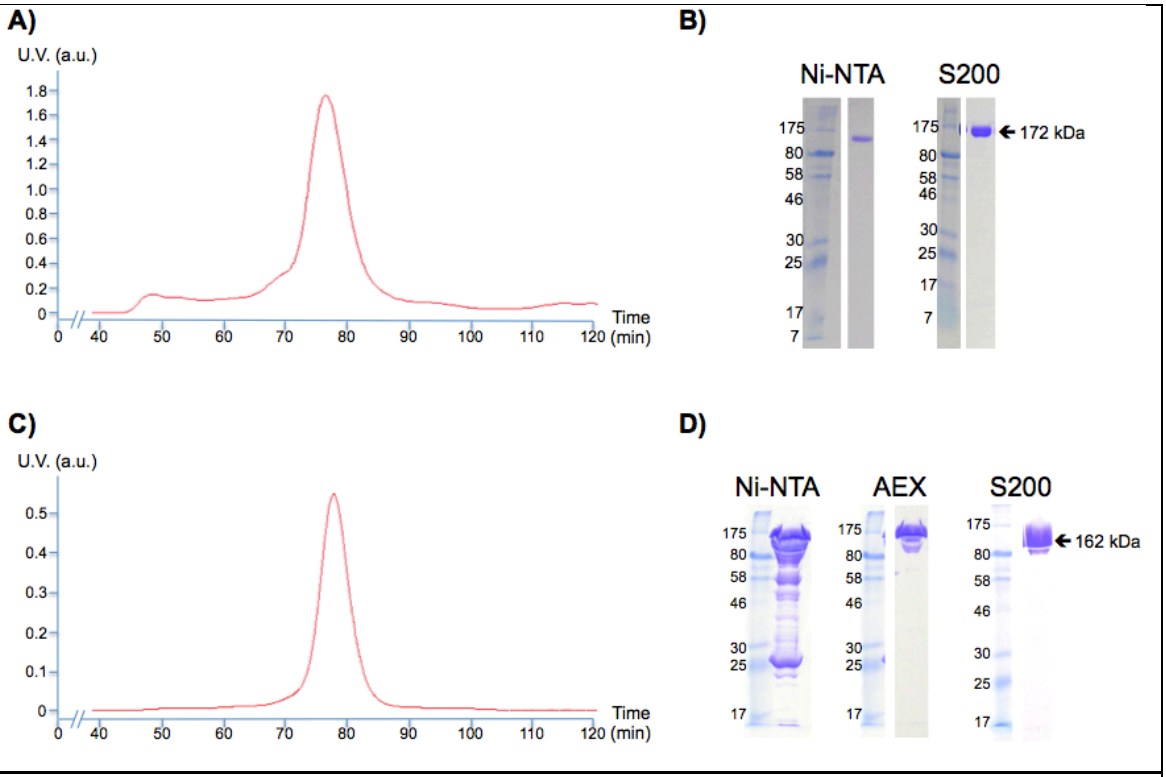


Figure 13: Production and purification of Dm- and PcUGGT

Purification of UGGT. Gel filtration chromatogram and SDS-PAGE of DmUGGT (A and B) and PcUGGT (C and D), respectively. All SDS-PAGE were 12% acrylamide.

2.1 Abstract

To date, there is only structural data for one small fragment of UGGT, where the third thioredoxin-like domain was crystallized. To gain information about the full-length enzyme, we established production and purification protocols for various UGGT species. The purified samples were not only soluble and stable, but they were capable of labeling unfolded RNase B and readily formed complexes with a 15 kDa selenoprotein, Sep15, which we produced and purified. Though our extensive crystallization screens did not succeed, we demonstrated through SEC-MALS and SAXS experiments that our samples were well folded and monodisperse even at high protein concentration.

2.2 Introduction

UGGT is an enzyme with a unique and essential activity. Its activity is both selective for misfolded or unfolded proteins containing the Man_9-$GlcNAc_2$ glycan and wide-scoped since it can glucosylate substrates of varying shapes and sizes, such as glycoproteins[34, 35, 39, 40, 49] and small synthetic substrates[42-44]. The common features to UGGT substrates are the Man_9-$GlcNAc_2$ glycan and a hydrophobic motif that are within 40 Å away[38, 39, 48, 49]. Similarly, UGGT contributes to the proper maturation of the major histocompatibility complex of class I by ensuring that only high-affinity MHC1-peptide complexes are formed[53, 54]. In this pathway, UGGT detects weakly bound peptides and labels the heavy chain of MHC1, sending it back to the peptide-loading complex to form a stable, high-affinity MHC1-peptide complex (see Figure 7).

When this thesis work began, no structures for UGGT or its domains had been determined. We thus set about to determine structures for full-length UGGT through crystallography, for which we setup robust production and purification protocols. Despite our extensive trials, we found no crystallization conditions. To date, only one domain of UGGT has been crystallized[56], though it is only a small fragment of UGGT (see Figure 8). We turned to biochemical assays and to SAXS experiments to verify UGGT was active, stable and able to bind to known UGGT effectors, such as Sep15. My results show that UGGT samples were capable of selectively labeling an unfolded substrate, interacting with Sep15 and remained monomeric even at high protein concentrations.

2.3 Results

2.3.1 Production and purification of PcUGGT and AoUGGT

I produced UGGT from the fungus *Penicillium chrysogenum* (PcUGGT) and *Aspergillus oryzae* (AoUGGT) using *E. coli* cells and achieved purification yields of 10 mg/L of protein (see Figure 14), with PcUGGT displaying the highest purity. We obtained the constructs from our collaborator Dr. Yukishige Ito, who did preliminary screens for various UGGT orthologs for bacterial expression. I purified the bacterially expressed enzymes through a combination of three chromatographic columns. The initial purification step was done using affinity chromatography, on a Ni-NTA column, which

removed the majority of contaminants. The second step of purification, using an anion exchange column, allowed the removal of additional contaminants. The final step of purification used a Superdex 200 gel filtration column, which removed the last contaminants from our samples (see Figure 13, panel C). The AoUGGT showed significantly more contaminants through all the purification steps, and as consequence, its final purity was lower than that of PcUGGT samples.

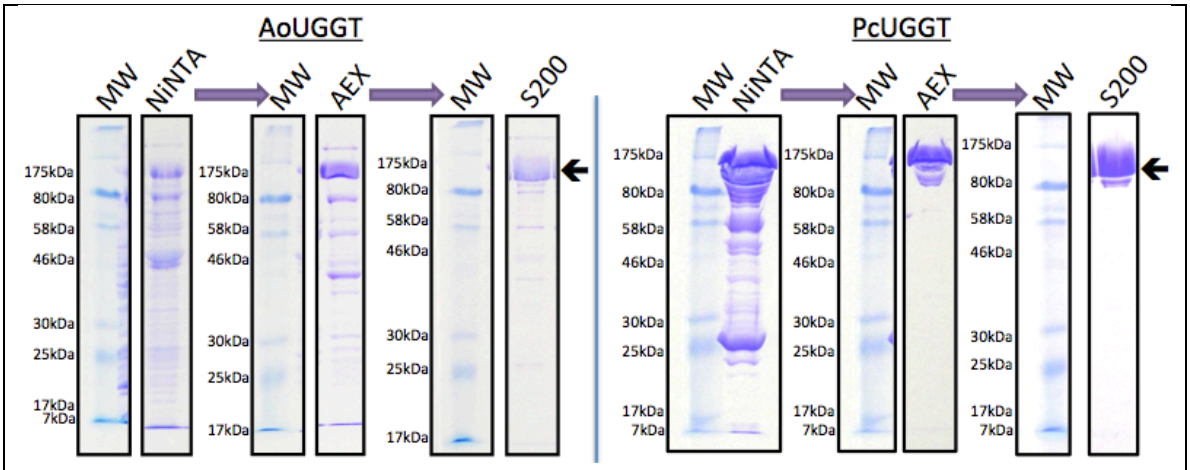


Figure 14: Purification of Ao- and PcUGGT

Analysis by SDS-PAGE gels for the Ni-NTA, AEX and gel filtration purification steps of AoUGGT (left panel) and of PcUGGT (right panel). All gels were 12% acrylamide. The black arrow indicates the band for UGGT.

2.3.2 Production and purification of DmUGGT

To improve the chances of succeeding in crystallization trials, I produced and purified *Drosophila melanogaster* UGGT (DmUGGT) expressed in *Sf*9 insect cells, using as starting point previously published protocols[11]. The DmUGGT enzyme was secreted into the cell culture media during production, which was collected and used as starting material for purification with Ni-NTA affinity columns (see Figure 15). The purity of DmUGGT after only one step of purification was very high. Similarly to the purification of Pc- and AoUGGT, the initial protocol used an intermediary anion exchange column, which I later abandoned as the longer purification protocol led to protein degradation products. The final gel filtration purification step on a Superdex 200 column yielded a

very homogeneous and pure sample, free of degradation products and contaminants (see Figure 13, panel A).

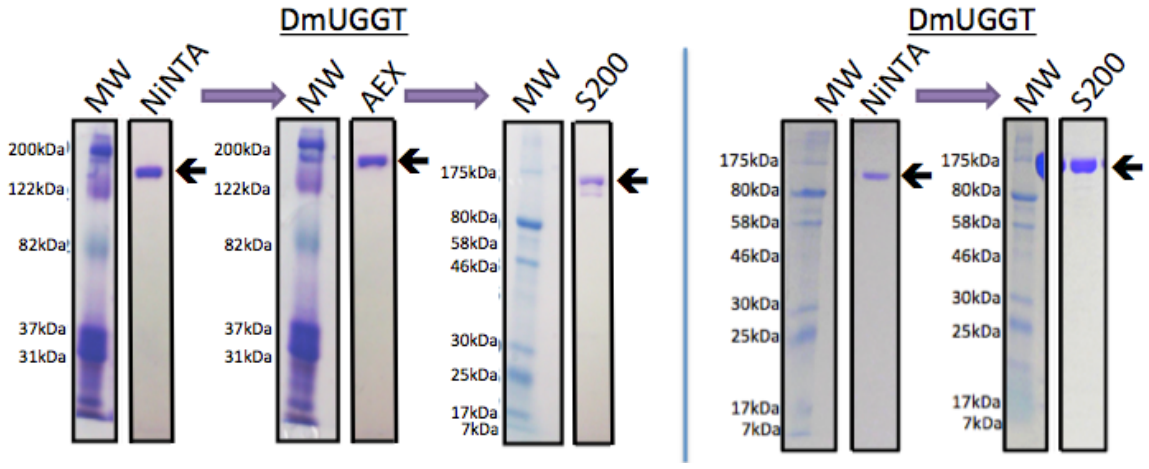


Figure 15: Purification of DmUGGT

Analysis by SDS-PAGE gels for the Ni-NTA, AEX and S200 purification of DmUGGT (left panel), displaying some degradation products, and for the Ni-NTA and S200 purification (right panel), free of degradation products. All gels were 12% acrylamide. The black arrow indicates the UGGT band.

2.3.3 Production and purification of Sep15 and GST-Sep15cr

I produced and purified the UGGT interacting partner, Sep15, as the full-length protein (see Figure 16, left panel) and as a fusion protein between GST and the cysteine-rich domain of Sep15 (Sep15cr), which is responsible for binding to UGGT (see Figure 16, left panel). Both constructs were expressed in bacterial cells. The purification of full-length Sep15 was done using Ni-NTA affinity chromatography, followed by gel filtration using a Superdex 75 column. The purification of GST-Sep15cr was carried out using a glutathione Sepharose column, followed by gel filtration using a Superdex 200 column. The purified GST-Sep15cr protein was unstable and proteolysis was observed, yielding a mixture of GST-Sep15 and free GST in the purified sample. Both Sep15 and GST-Sep15cr were nonetheless capable of binding UGGT and the complex could be successfully purified.

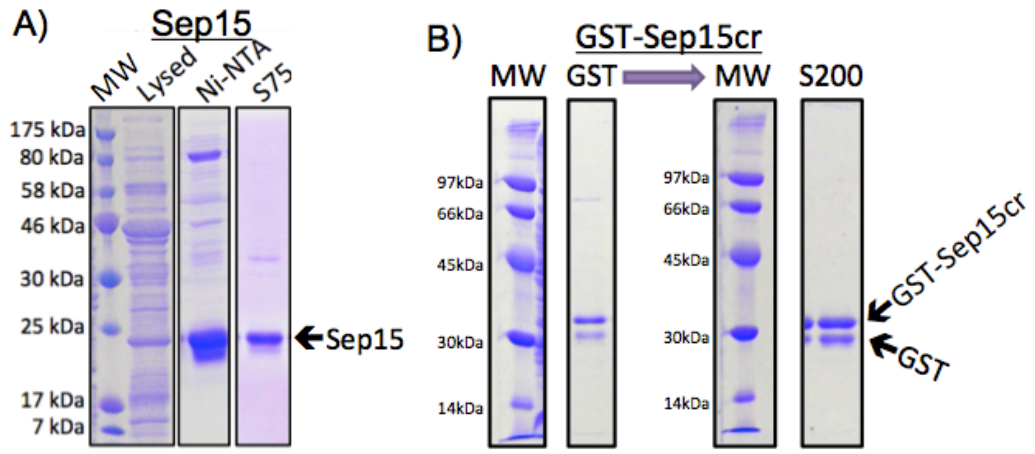


Figure 16: Purification of Sep15 and GST-Sep15cr

Analysis by SDS-PAGE gels for the Ni-NTA and gel filtration purification of Sep15 (left panel) and of GST-Sep15cr with glutathione Sepharose and gel filtration columns (right panel). All gels were 15% acrylamide.

2.3.4 Co-purification of Dm- and PcUGGT with Sep15

The purified UGGT samples from both species were able to bind Sep15 and GST-Sep15cr; both protein complexes were purified for structural studies. After incubation of GST-Sep15cr with DmUGGT, I separated the DmUGGT/GST-Sep15cr complex from free GST and GST-Sep15cr through size exclusion chromatography. The chromatogram gave two peaks (A and B of Figure 17), peak A with UGGT/GST-Sep15cr and peak B with free GST and GST-Sep15cr. Consistent with the literature[29], the cysteine-rich domain of Sep15 mediated interaction with UGGT.

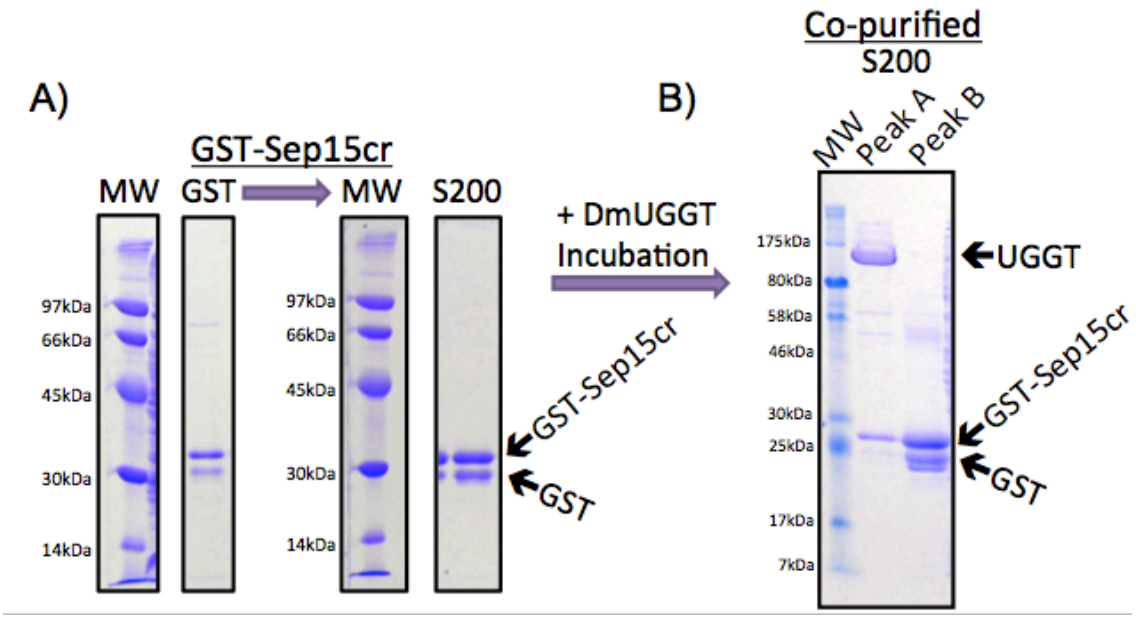


Figure 17: Co-purification of DmUGGT and GST-Sep15cr

Analysis by SDS-PAGE gels for co-purification of DmUGGT/GST-Sep15cr using gel filtration (gel on the right). The peak for DmUGGT/GST-Sep15cr (lane A) and for GST-Sep15cr/GST (lane B) were concentrated separately and analyzed by SDS-PAGE.

Similarly to the purification of UGGT/GST-Sep15cr complex, after incubating purified Sep15 and UGGT together, I separated the UGGT/Sep15 complex from free Sep15 using a Superdex 200 column. From the relative intensity of the bands in the Coomassie stained gel, the UGGT/Sep15 binding was agreed with a 1:1 ratio. Though consistent with the literature, we validated this using SEC-MALS. Though the results were stronger for DmUGGT, both species associated with Sep15 in solution (see Figure 18), and both were purified in high-enough concentrations for crystallographic screens.

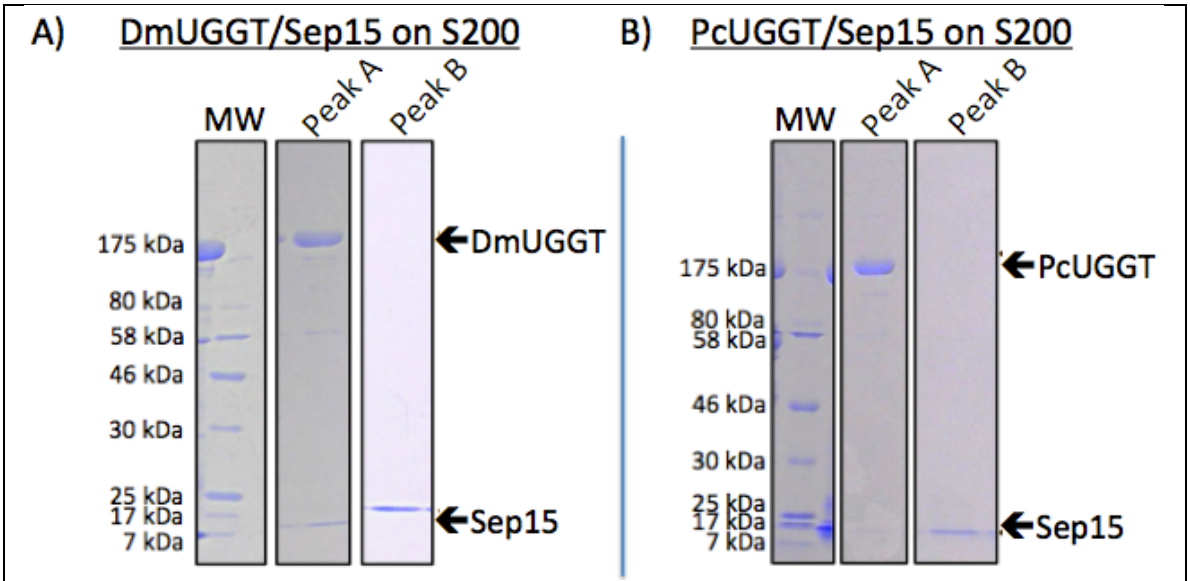


Figure 18: Co-purification of Dm- and PcUGGT with Sep15

Co-purification of DmUGGT/Sep15 (gel on the left) and of PcUGGT/Sep15 (gel on the right) using gel filtration, analyzed by SDS-PAGE on 12% acrylamide gels.

2.3.5 UGGT activity assays

I showed through qualitative misfolded glycoprotein labeling assays that all three purified UGGT samples were capable of glucosylating unfolded RNase B (see Figure 19). The labeling activity of DmUGGT did not show a discernible increase in the presence of Sep15. This results show that the purified samples are active and suitable for structural studies through protein crystallography and other biophysical techniques.

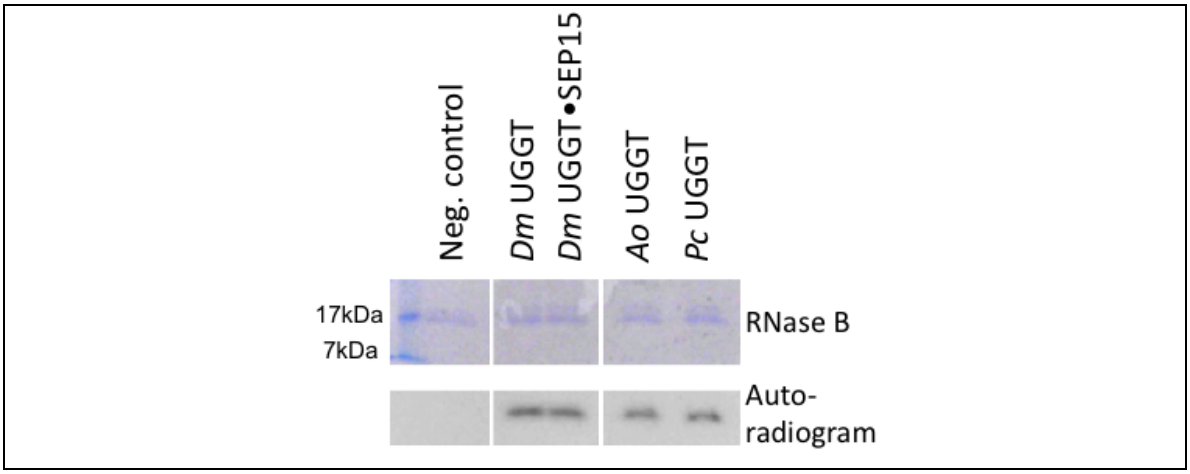


Figure 19: Unfolded RNase B labeling assay using Dm-, Ao- and PcUGGT

SDS-PAGE and autoradiogram for the glucosylation assay of unfolded RNase B using radioactively labeled ^{14}C-UDP-Glucose and purified Ao-, Dm- and PcUGGT, and the DmUGGT/Sep15 complex. Negative control, reaction mixture without UGGT added.

2.3.6 SEC-MALS of DmUGGT and DmUGGT-Sep15

To characterize the molecular weight and aggregation state of my samples, I used multi-angle light scattering coupled to size exclusion chromatography (SEC-MALS) on a Superdex 200 column (see Figure 20). The experiments confirmed that UGGT is monomeric in solution, with an apparent molecular weight of 175 kDa (see Figure 20, top panel), and that UGGT/Sep15 is a 1:1 protein complex with a molecular weight of 196 kDa (see Figure 20, bottom panel). As can be seen in the SEC-MALS data, the UGGT/Sep15 peak eluted a couple of minutes earlier from the column than the UGGT peak. The light scattering data for these two peaks confirmed that the molecular weight associated with these two samples matched the theoretical molecular of monomeric UGGT (172 kDa) and the molecular weight of a one-to-one UGGT/Sep15 complex (172 kDa UGGT plus 15 kDa Sep15).

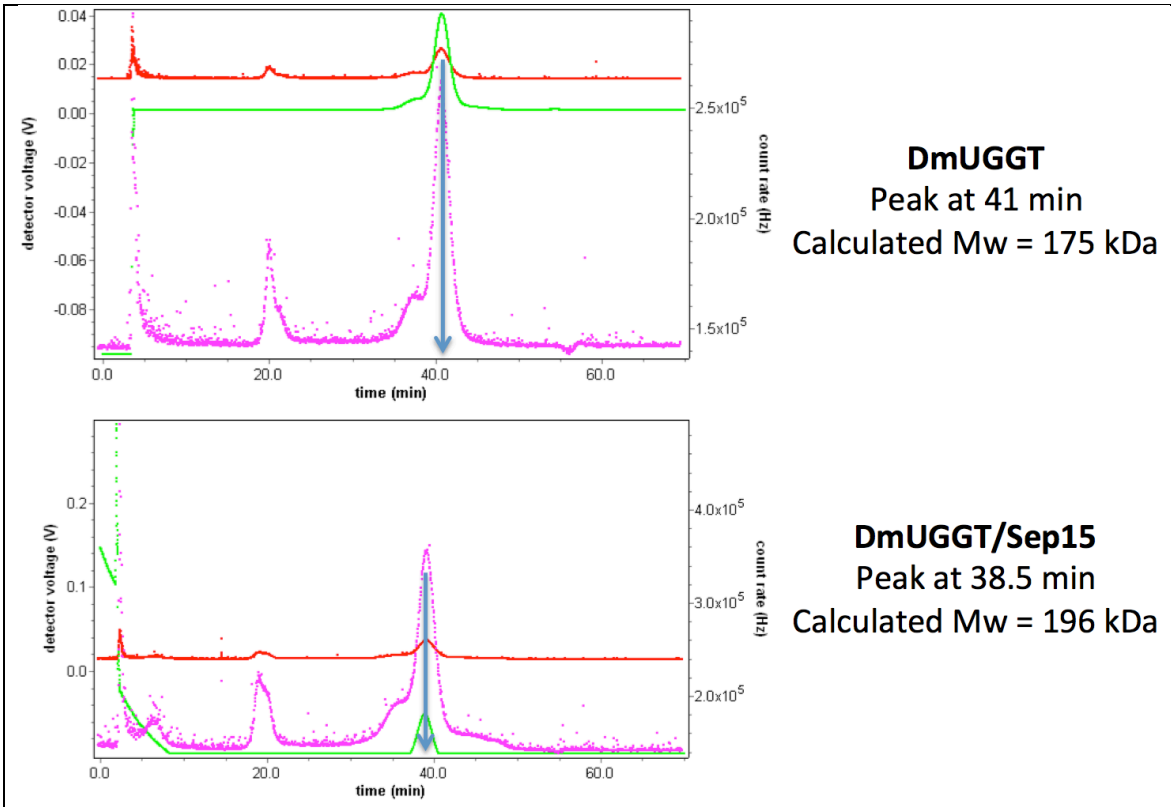


Figure 20: SEC-MALS of DmUGGT and the DmUGGT/Sep15 complex

Analysis of purified DmUGGT (top panel) and of DmUGGT/Sep15 complex (bottom panel) by size exclusion and multi-angle light scattering. The DmUGGT peak eluted at minute 41 and corresponded to a molecular weight of 175 kDa. The DmUGGT/Sep15 peak eluted earlier, at minute 38.5 and corresponded to a molecular weight of 196 kDa.

2.3.7 Crystallization trials of UGGT and UGGT/Sep15

Despite having purified UGGT from various species to a high degree of purity, extensive crystallization trials (see Table 1) failed to yield any encouraging hits to optimize for x-ray diffraction experiments. I tested a wide range of conditions using five different crystallization screen sets, each of which contains 96 different solutions. Using those screens, the first variable I tuned was the initial protein concentrations, testing ranges from 1 to 20 mg/ml. UGGT was stable with salt concentrations between 200 and 450 mM, I therefore used salt concentration as a second variable to tune. Initial screens were performed at room temperature and later screens carried out at 18 °C and 4 °C. The

lower temperature helped minimized protein precipitation in the crystallization drops. In this manner, I screened using DmUGGT, DmUGGT/Sep15, PcUGGT and PcUGGT/Sep15 purified samples at various protein and salt concentrations. In a similar way, I screened for crystals in the presence of the synthetic substrate M9-MTX provided to us by our Japanese collaborators, using variations of molar ratios between UGGT and M9-MTX from 1:1 to 1:10. I also tried crystallization screens using UDP-glucose and other UDP-glucose analogs such as UDP-galactose, UDP-glucuronic acid and tunicamycin. Finally, in an attempt to aid crystallization by removing flexible loops from UGGT, I carried out screens in the presence of varying concentrations of two proteases in the drop – chymotrypsin and trypsin.

	DmUGGT	PcUGGT	DmUGGT/ Sep15	PcUGGT/ Sep15
Crystallization suites	Classics I & II; PEGs; JCSG+; PACS; Ammo.Sulfate			
[Protein]	1 - 20 mg/ml	1 - 20 mg/ml	2 - 15 mg/ml	2 - 15 mg/ml
[NaCl]	200 - 400mM	200 - 400mM	200 - 400mM	200 - 400mM
Temperature	22, 18, 4°C	22, 18, 4°C	22, 18, 4°C	22, 18, 4°C
M9-MTX	1:1, 1:2, 1:10	1:1, 1:2, 1:10		
Analogs	UDP-glucose, UDP-galactose, UDP-glucuronic acid, tunicamycin			
Protease	Chymotrypsin and Trypsin at two concentrations in drops			
	Upwards of 30,000 conditions screened			

Table 1: Crystallization trials using UGGT and UGGT/Sep15

The conditions include variations of protein concentration for five different crystallization solution kits. Drops were set at different temperatures, varying salt concentration in the purification buffer and in the presence of various UDP-glucose analogs. Trials also included a synthetic substrate, M9-MTX, and treatment with trypsin and chymotrypsin.

2.3.8 Characterization in solution by SAXS

The lack of crystallization hits suggested the use of SAXS to study UGGT in solution. For this, I used the purified Dm- and PcUGGT at protein concentrations comparable to those used in my crystallization screens. The Guinier plot was the first tool I used to assess UGGT behavior (see Figure 21). The data curves remain parallel at increasing concentrations, which is a sign of a soluble and stable protein sample. The Guinier curves were linear and did not show curving in the scattering signal at low angles, consistent with an absence of aggregation for both Dm- and PcUGGT.

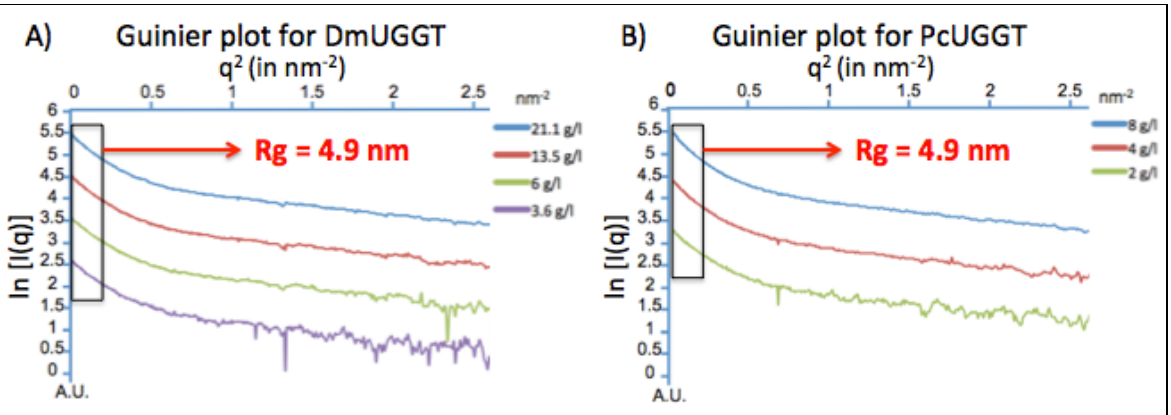


Figure 21: Guinier plots for Dm- and PcUGGT

Protein specific scattering for four concentrations of DmUGGT (plot on the left) and for three concentrations of PcUGGT (plot on the right) were charted using the Guinier plot. For easier comparison, the curves were staggered.

Using the low-angle region of the Guinier curves (see Figure 22), I was able to estimate the radius of gyration (R_g) of UGGT from both species to ~4.9 nm. The R_g is estimated from the slope of the Guinier plot in the low angle region. This region of the curve is highly sensitive to the presence of high-molecular weight aggregates. In a case where proteins aggregate, this translates into a steeper curve in the low-angle region. With both UGGT species and at all concentrations tested, the slope and calculated R_g values at low-angle remained constant. This confirms that the protein remains soluble and monodisperse as function of the protein concentration.

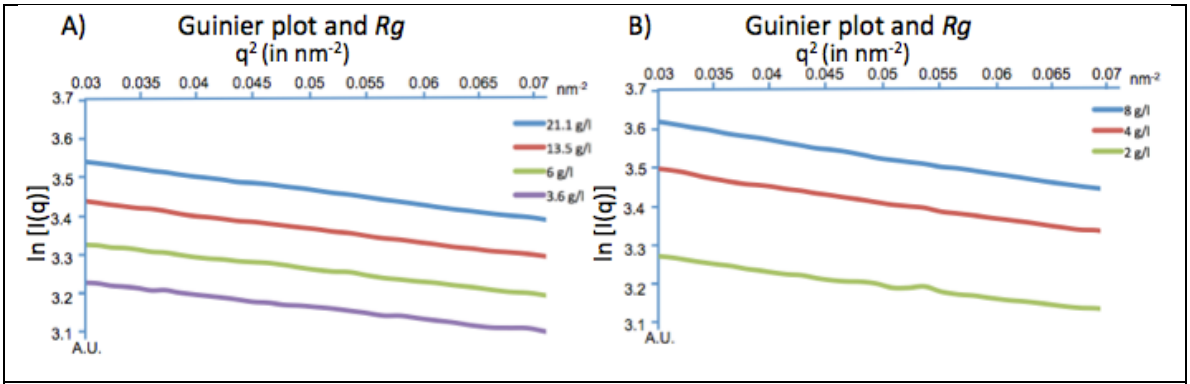


Figure 22: Zoom over Guinier plot region for R_g determination

The region of the Guinier plot near the origin, where the scattering angle is the smallest, is highly sensitive to aggregation. Focusing on this region for DmUGGT (plot on the left) and for PcUGGT (plot on the right), a 4.9 nm radius of gyration was calculated for samples at all concentrations. Curves were staggered to ease comparisons across samples.

I evaluated the folded state of UGGT using the Kratky plot (see Figure 23). The data showed that UGGT samples from both species are well folded at all concentrations tested. The shape of Kratky curves is indicative of how well a protein is folded: curves that do not come back to zero indicate unfolded or partially unfolded proteins, whereas curves that come back to zero indicate properly folded proteins. All of the samples I tested showed a tail that comes back to zero.

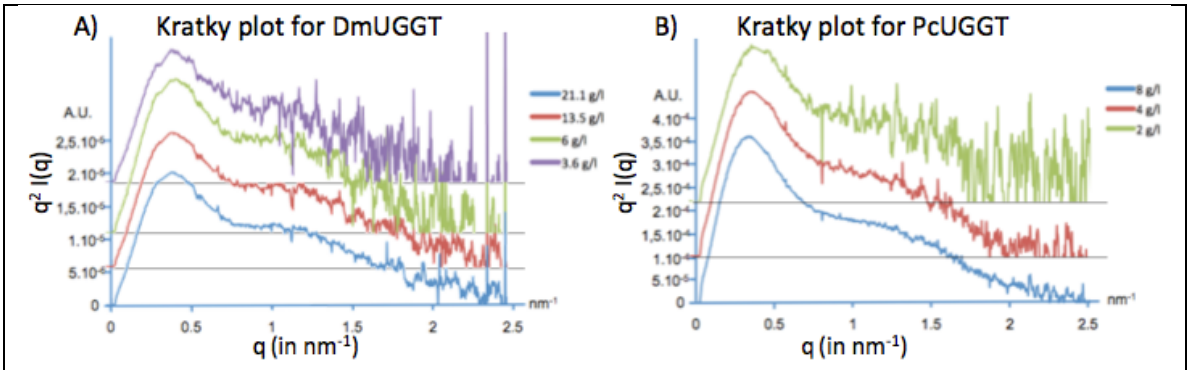


Figure 23: Kratky plots for Dm- and PcUGGT

The scattering data for DmUGGT (left panel) and for PcUGGT (right panel) was represented using the Kratky plot. As previously, the curves were staggered to facilitate

the comparison across the different concentration. Lower concentrations curves displayed significant noise on the high-angle region, though all curves returned to the origin.

As a final tool for SAXS data analysis, I calculated the distance distribution curve or pair distribution curve (P_R) to estimate the size of UGGT in solution (see Figure 24). The analysis measured a maximum diameter of roughly 140 Å in length. The P_R curve gives a distribution of the interatomic distance within the protein. In the case of conformational changes or oligomer formation, the size of a protein would increase and this can be quantified using the P_R plot. In the case of UGGT, the pair distribution curve for both species yielded a maximal length of 140 Å at all concentrations and was consistent with all of the previous analysis. The shape of the curve also correlates with the shape of the protein: globular and compact proteins have a Gaussian-shaped curve, whereas multi-domain proteins have a more complex shape with multiple peaks or shoulders. The latter was observed for UGGT, pointing to a non-compact organization of UGGT domains.

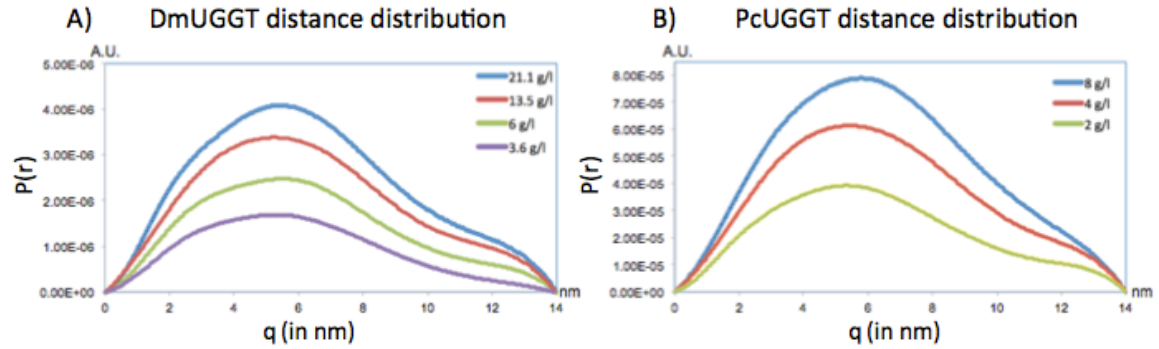


Figure 24: Distance distribution curves for Dm- and PcUGGT

The distance distribution curves were calculated for DmUGGT (left panel) and for PcUGGT (right panel), for every concentration tested. The d_{max} calculated in the case of both species, at all concentrations, was consistently estimated to 140 Å. The shape of the curves is consistent with a multi-domain protein. The curves were scaled down relative to the highest protein concentration to facilitate comparisons.

2.4 Discussion

For over two decades[18], UGGT has been at the subject of many studies, many of which have tried to obtained high-resolution structures through crystallography and X-ray diffraction[56]. The activity and the sequence of UGGT are unique; as this protein shares no significant similarity to other know cellular proteins. Though cultured cells can survive without UGGT, this enzyme is essential for embryonic development and cellular viability[90] and is well conserved across eukaryotes, particularly in its catalytic C-terminal domain[22, 36]. Following the footsteps of previous groups, I initially aimed my efforts at solving crystallographic structure of UGGT (see Table 1).

I was able to produce and purify UGGT from various species (see Figure 13), but despite extensive trials, I was unable to find any crystallization conditions. The yeast species and the fruit fly UGGT share 21-30%, 41% and 61% sequence identity between their thioredoxin-like domains, the beta-rich region and the catalytic domain respectively. The lower conservancy in the thioredoxin-like domains is similar across other species. The higher conservancy of the catalytic domains seems to be a requirement for the precise catalytic activity of UGGT, conserved across species[22, 36, 51, 91]. I verified through activity assays, SEC-MALS (see Figure 20) and SAXS experiments that the protein samples we prepared were well folded, active and stable even at high protein concentrations (see Figure 21, Figure 23 and Figure 24). My SAXS results did not point towards flexibility as being the culprit for the lack of crystallization hits. Those experiments indicate that the protein is folded correctly and consistent with a monomeric enzyme. The calculated radius of gyration and the profile of the distance distribution curve indicate UGGT has a complex multi-domain organization, with a non-compact overall shape[25, 52].

Through bio-informatics, five domains have been identified for UGGT, with long linker regions between each domain[56] (see Figure 8). With the exception of the third thioredoxin-like domain, the domains themselves are not stable when produced on their own and the enzyme is catalytically inactive if any domain is removed suggesting the domains work in a concerted fashion during substrate recognition and glucosylation[25, 52]. Given the selective activity of UGGT and its multiple domains, a non-compact multi-domain organization is the most likely arrangement allowing UGGT to fulfill its cellular

role. Though we can't completely rule out large domain movements yet within UGGT, the SAXS data is more in line with a relatively compact domain organization with a small degree of flexibility for recognition and labeling of various unfolded glycoproteins, which may be the reason why crystallization conditions have failed to materialize.

To improve on these results, a first experiment would be to accurately quantify the activities of DmUGGT, of PcUGGT and of AoUGGT using RNase B as a substrate, in order to better compare their enzymatic activities and to better assess their quality after purification. The purification of AoUGGT could also be improved and SAXS data could be collected to compare to the Dm- and PcUGGT data presented here.

2.5 Material and Methods

2.5.1 Production and purification of PcUGGT and AoUGGT

Penicillium chrysogenum and *Aspergillus oryzae* UGGT were codon-optimized and cloned for bacterial expression using *E. coli* with the pCold I vector and a hexa-histidine C-terminal tag. We used *E. coli* Rosetta Gami 2 for expression in LB-ampicillin media. We produced Pc- and AoUGGT at 37°C at 200 rpm, inducing with 0.5 mM IPTG at an OD_{600nm} of 0.6 and harvesting cells 4 hours post-induction by centrifugation at 8,000 *g* at 4 °C for 30 minutes. The bacterial pellets were suspended in ~40 ml of lysis buffer (Buffer A, with PMSF at 1 mM) and lysed by sonication. The lysates were centrifuged at 18,000 *g* at 4 °C for 30 minutes and the supernatant was added to ~5 ml of pre-equilibrated Ni-NTA resin in gravity columns. Pc- and AoUGGT were eluted using 150 mM Imidazole buffer A, concentrated to ~5 ml and diluted to a 50 mM NaCl (final). The diluted sample was loaded onto anion exchange columns and eluted using a linear gradient of 50 mM to 1 M NaCl with UGGT eluting at ~350 mM NaCl. UGGT was concentrated to ~2 ml volume in a 50 kDa MWCO concentrator by centrifuging at 4000 *g* at 4 °C and injected into a Superdex 200 size exclusion column (Buffer A, without anti-protease). Fractions were analyzed by SDS-PAGE, concentrated to less than 5 mg/ml and stored at -80 °C.

2.5.2 Production and purification of DmUGGT

Drosophila melanogaster UGGT was previously cloned in the pFastBac plasmid with a N-terminal melittin secretion sequence and with a C-terminal hexa-histidine tag. The

plasmid was used to transform DH10-Bac cells and the baculovirus DNA was produced according to the *Sf*9 baculovirus expression system manual, using serum free media. For production, we amplified *Sf*9 insect cells to 2 x 10^6 cells per ml, infected with a P3 virus at 0.2% (v/v, final) and let the cells grow for 72 hours at 27 °C at 120 rpm in 2.8 L flasks, with 0.5% (v/v, final) heat-inactivated fetal bovine serum to minimize UGGT proteolysis. Cells were centrifuged at 1000 G and the supernatant containing soluble DmUGGT incubated with 5 ml of pre-equilibrated Ni-NTA resin per 500 ml of supernatant for ~3 hours at 4 °C. After loading the resin onto a gravity column, DmUGGT was eluted in a 150 mM Imidazole buffer (Buffer A: 30 mM Tris-HCl, pH 7.5, 300 mM NaCl, 3% glycerol, Roche Complete EDTA-free anti-protease). UGGT was concentrated to ~2 ml volume in a 50 kDa MWCO concentrator by centrifuging at 4000 *g* at 4 °C and injected into a Superdex 200 size exclusion column (Buffer A, without anti-protease). The fractions were analyzed by SDS-PAGE and stored at 5 mg/ml and -80 °C.

2.5.3 Production and purification of Sep15

Human Sep15 was cloned in the pET29a plasmid with a C-terminal hexa-histidine tag. A shorter domain of Sep15, responsible for binding to UGGT, was cloned in pGEX vector to produce it as a fusion protein with glutathione S-transferase (GST). The plasmids were used to transform *E. coli* Rosetta Gami 2 cells in LB media, supplemented with ampicillin and kanamycin for pGEX and pET29a vectors respectively. We produced Sep15 and Sep15cr-GST at 37°C at 200 rpm, inducing with 0.1 mM IPTG at an OD_{600nm} of 0.6 and harvesting cells 4 hours post-induction by centrifugation at 8,000 *g* at 4 °C for 30 minutes. The bacterial pellets were suspended in ~40 ml of lysis buffer (Buffer A, with PMSF at 1 mM) and lysed by sonication. The lysed cells were centrifuged at 18,000 *g* at 4 °C for 30 minutes and the supernatant was added to ~5 ml of pre-equilibrated Ni-NTA resin in gravity columns for His-tagged Sep15 and with GST-resin for Sep15cr-GST. Sep15 was eluted using 150 mM Imidazole buffer A, concentrated to ~3 ml using a 10 kDa MWCO concentrator at 4000 *g* at 4 °C. Sep15cr-GST was eluted using a 20 mM glutathione buffer A from the GST-column and concentrated to ~3 ml using a 30 kDa MWCO concentrator at 4000 *g* at 4 °C. Samples were further purified using Superdex 75 size exclusion column (Buffer A, without anti-protease). After analysis by SDS-PAGE, the protein was stored at 3 mg/ml and -80 °C.

2.5.4 Co-purification of UGGT with Sep15

Aliquots of purified Dm- or PcUGGT were mixed with an excess of purified Sep15 or Sep15cr-GST and concentrated to less than 4 ml using 10 kDa MWCO concentrators at 4000 G 4°C. The mixture was injected into a size exclusion Superdex 200 column equilibrated with 30 mM Tris-HCl, pH 7.5, 300 mM NaCl, 3% glycerol (w/v). The fractions from the purification were analyzed using SDS-PAGE and the fractions containing UGGT-Sep15 were separated from the fractions containing Sep15, which eluted later from the column. The excess of free Sep15 or Sep15cr-GST could be recovered and recycled for other experiments. The purified UGGT-Sep15 was concentrated to 2 mg/ml and stored at -80°C.

2.5.5 Qualitative RNase B labeling assays

RNase B is commercially available and a small percentage of the RNase B carries the 2-N-acetylglucosamine-9-mannose N-glycan that UGGT requires for effective glucosylation. The purchased RNase B was dissolved in a 50 mM potassium phosphate buffer at pH 7.5 with 8 M urea and 10 mM DTT. The unfolded substrate was dialyzed overnight at 4 °C against 50 mM sodium phosphate buffer at pH 5.8 with 10 mM DTT. The prepared substrate was flash-frozen in liquid nitrogen and lyophilized overnight. For glucosylation assays, UGGT was incubated in a buffer containing 50 mM Tris-HCl pH 7.5, 0.2 mM 1-deoxynojirimycin, 5 mM $CaCl_2$, 0.42g/l RNase B, 20 μM ^{14}C-UDP-glucose and the mixture was incubated at 37°C for 2 hours. The reaction was stopped by addition of 5x Laemmli loading buffer and the samples were loaded into a 12% SDS-PAGE gel. After electrophoretic migration, the gel was dried and exposed to photographic film for up to 24 hours at -80 °C.

2.5.6 SEC-MALS of DmUGGT and DmUGGT-Sep15

DmUGGT and DmUGGT-Sep15 were injected into an analytical size exclusion Superdex 200 column coupled to UV and small angle light scattering detectors. The buffer used was identical to that used for purification purposes (Buffer A) and the experiment was carried out at ~22 °C. After calibration using BSA at 4 mg/ml as a standard for molecular weight, we applied DmUGGT and DmUGGT-Sep15 at 5 mg/ml

and 3 mg/ml respectively. The chromatographic and scattering profiles were analyzed using the proprietary software, yielding an estimation of the molecular weight.

2.5.7 Crystallization trials of UGGT and UGGT-Sep15

Utilizing purified UGGT samples, crystallization trials were undertaken using the sub-microlimiter handling Phoenix robot. We systematically screened using the solutions contained in the following kits: Classics 1, Classics 2, PEGs, PACS, JCSG+, and Ammonium Sulfate Suites. The initial round of crystallization trials focused on using purified Dm-, Pc- and AoUGGT at concentrations ranging from 1 mg/ml up to 20 mg/ml, at 22 °C, 18 °C and 4 °C for all, and varying the NaCl concentrations from 400 mM to 200 mM. The second round of trials focused on Dm- and PcUGGT-Sep15 complex crystallization, sampling protein concentrations ranging from 2 to 15 mg/ml, varying the NaCl concentration between 200 mM and 400 mM, and at temperature of 22 °C and 4 °C. The third round of trials focused on Dm- and PcUGGT crystallization in the presence of 1:1, 1:2 and 1:10 molar ratios of UGGT:M9-MTX, a small synthetic substrate provided by Drs. Yoichi Takeda and Yukishige Ito, carried out at proteins concentrations between 2 and 10 mg/ml and at 4 °C. A fourth set of trials tested crystallization drops containing Dm- or PcUGGT and three glucose-donor analogs: UDP-glucose, tunicamycin, UDP-galactose and UDP-glucuronic acid. In a final round of crystallization trials, Dm- and PcUGGT drops were prepared in the presence of trypsin or chymotrypsin to attempt to crystallize proteolysis-resistant fragments of UGGT.

2.5.8 Characterization in solution by SAXS

Freshly purified UGGT samples were taken during concentration, measuring the protein concentration using a NanoDrop 2000c from ThermoScientific. Concentrations between 0.5 mg/ml and 20 mg/ml were tested. Data collection was done using an in-house SAXSess Anton Par system, with exposures of six hours per sample on a CCD chip for both protein and buffer samples. Processing was done using the Primus software package. Scattering curves were normalized to protein concentration and baseline level. The buffer scattering data was subtracted from the protein scattering curves. Data was then analyzed using Guinier plots to determine radius of gyration values and find signs of aggregation. Distance distribution plots were then used to find optimal d_{max} values.

Preface to low-resolution UGGT structural studies

As we discussed in the previous chapter, the samples I prepared were highly amenable for structural studies: the protein samples were of high purity, were enzymatically active and remained monomeric and stable in solution, even at high protein concentrations. Despite all of this, my extensive crystallization trials did not give any promising leads. To gain structural insights in the absence of crystals, I benefitted from the supervision and direction of Dr. Kalle Gehring and from the EM training and expertise of Dr. Justin Kollman, an expert in the field of protein studies through EM.

For this chapter, I chose negative stain electron microscopy and single particle analysis reconstructions to determine low-resolution models for Dm- and PcUGGT. I worked in collaboration with Dr. Justin Kollman, who provided me with valuable training in terms of sample preparation, microscope operation and data analysis. I used the SAXS data I collected to carry out *ab initio* models of both UGGT species to cross-validate my EM findings. My results provide the first structural models for full-length UGGT and the structures across species and across techniques display a conserved topology, attesting to the validity of my findings. Though the structures are low-resolution and don't provide domain orientation details, I proposed a simple model for misfolded protein selection that is consistent with all the biochemistry literature available.

3 SAXS *ab initio* modeling and negative stain EM

Cover figure:

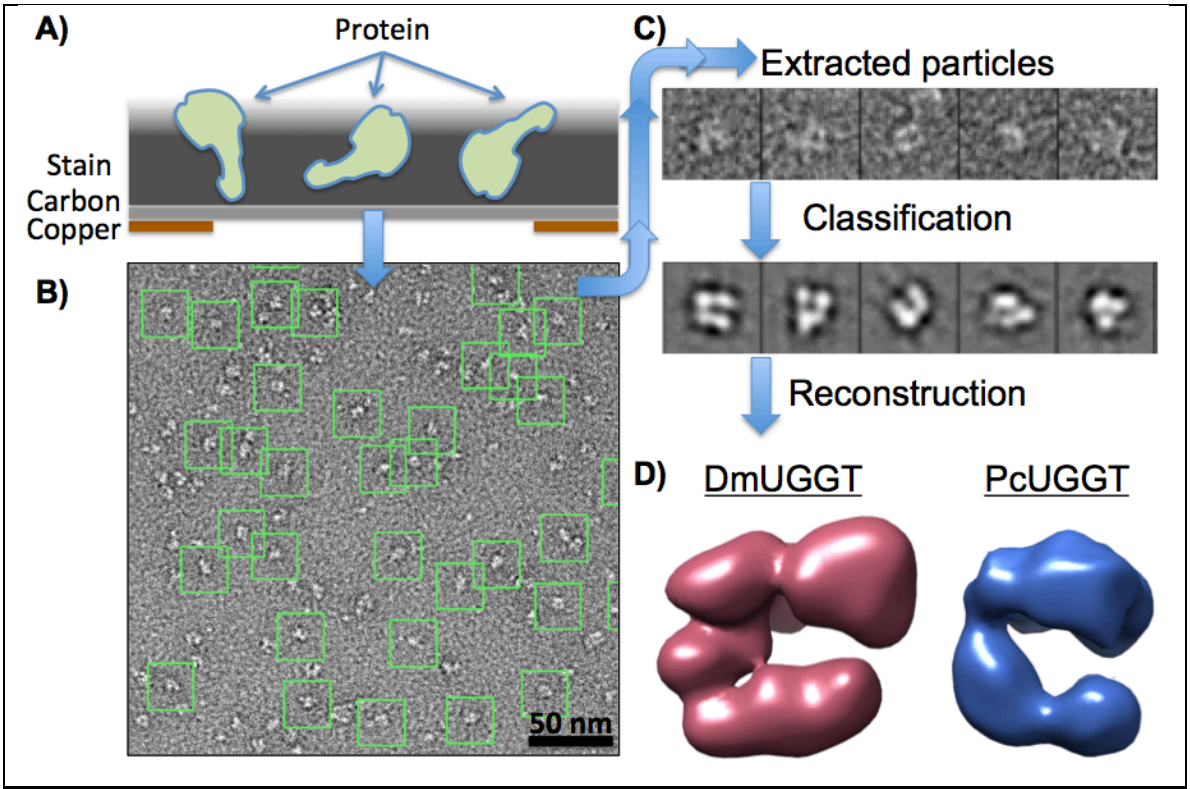


Figure 25: Structures of Dm- and PcUGGT by negative stain EM

Protein particles embedded in negative stain (A) are imaged and particles are picked from the micrographs (B). The extracted particles are normalized and classified (C) to yield low-resolution structures of Dm- and PcUGGT (D).

3.1 Abstract

Many groups have tried crystallizing UGGT without success, including ours, despite having very pure and active samples available. To determine full-length structures of UGGT, we utilized negative stain EM and single particle analysis to reconstruct three-dimensional models for Pc- and DmUGGT to resolutions of ~20 Å. These maps display a two-lobed "claw-shaped" structure. To cross-validate our EM maps, we used our SAXS data to reconstruct *ab initio* maps of UGGT, thus providing an independent approximation of the UGGT structure in solution. Both techniques revealed a two-lobed structure with an internal cavity. Although the position of the individual domains remains

to be identified, the low-resolution structures provide insight into the mechanism of substrate recognition by UGGT.

3.2 Introduction

As previously discussed, UGGT distinguishes between well-folded and defectively folded proteins (see Figure 1) and for glucosylation to be triggered, the glycan and the unfolded motif need to be in close proximity and part of the same molecule[38, 39, 48, 50]. *In vitro*, the labeling of unfolded glycoproteins is inhibited in the presence of unfolded non-glycosylated proteins[34], while the same reaction will be unaffected by the addition of folded proteins. This indicates glycosylated and non-glycosylated proteins compete for a common binding site on UGGT. This suggests that the misfolded protein sensor surface and the glucosyltransferase pocket of UGGT are close to each other and could be part of one large catalytic pocket. The sensor region of UGGT binds to solvent-exposed hydrophobic residues on the substrate, likely carried on flexible loops. The binding might be mediated through hydrophobic patches on UGGT, as the crystal structure of the third thioredoxin-like domain displayed a hydrophobic region masked by detergent molecules[56]. Through biochemical assays, it has been shown that the catalytic domain is inactive without the sensor region of UGGT[25, 52]. In mammals, there is a second isoform that has been reported to be inactive on misfolded glycoproteins, though seems active on small synthetic substrates[46]. A chimeric protein with the sensor region of UGGT1 and the catalytic domain of UGGT2 displayed 46.5% glycoprotein labeling activity relative to the wild-type UGGT1[52]. Thus, the recognition and labeling of unfolded glycoproteins require a close communication and interplay between the sensor region and the glucosyltransferase region of UGGT.

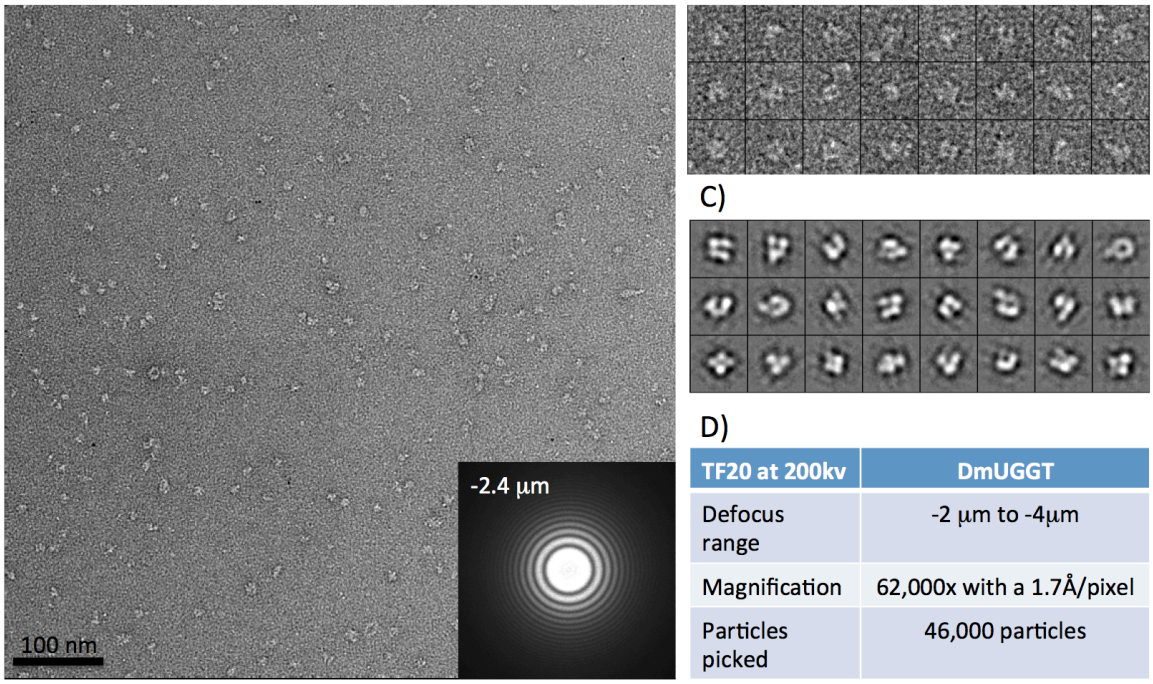
The large molecular weight of UGGT puts this enzyme out of reach for nuclear magnetic resonance and leaves only a few usable techniques to determine tridimensional structures. Crystallography of UGGT has eluded the efforts of many groups, which might be in part due to UGGT's large size and in part to some intrinsic flexibility of the domains of this enzyme. Sequence analysis and biochemical data of UGGT have shown that the N-terminal domain, involved in the misfolded protein detection[52], has three thioredoxin-like domains, one of which has been crystallized[56], and a beta-rich domain (see Figure 8). The C-terminal catalytic domain of UGGT belongs to the

glycosyltransferase 8 family, is highly conserved among species, and is only active in the context of the intact protein[25]. To gain insights into the structure of UGGT we utilized a two-pronged approach using negative stain electron microscopy[67, 76] and SAXS[62, 92]. The analysis of UGGT from two distantly related organisms, fruit flies and an ascomycetous fungus, and through two distinct techniques shows strong conservation in the overall shape: a large lobe and a smaller hook-like appendage, forming a large central cavity (see Figure 25). The open structure suggests a mechanism for how the catalytic activity of UGGT is restricted to unfolded or misfolded glycoproteins.

3.3 Results

3.3.1 Single particle analysis for Dm- and PcUGGT

The single particle analysis of DmUGGT using uranyl formate stain shows that the protein particles have no visible preferred orientation on the grid and the size distribution of the particles is very uniform, key characteristics to facilitate three-dimensional reconstructions (see Figure 26). After collecting negative stain data for DmUGGT with defocus values ranging from -2 to -4 μm, I boxed roughly 46,000 particles, which I classified using the reference-free method built into the e2refine2d.py routine of Eman2.0. Among the class averages I calculated, the number of particles within each class was very even, with no classes being over or under populated. The individual particles and the two-dimensional classes are between 10 and 12 nm in size. The class averages show diverse views of DmUGGT, with features that could be formed by the domains of UGGT.

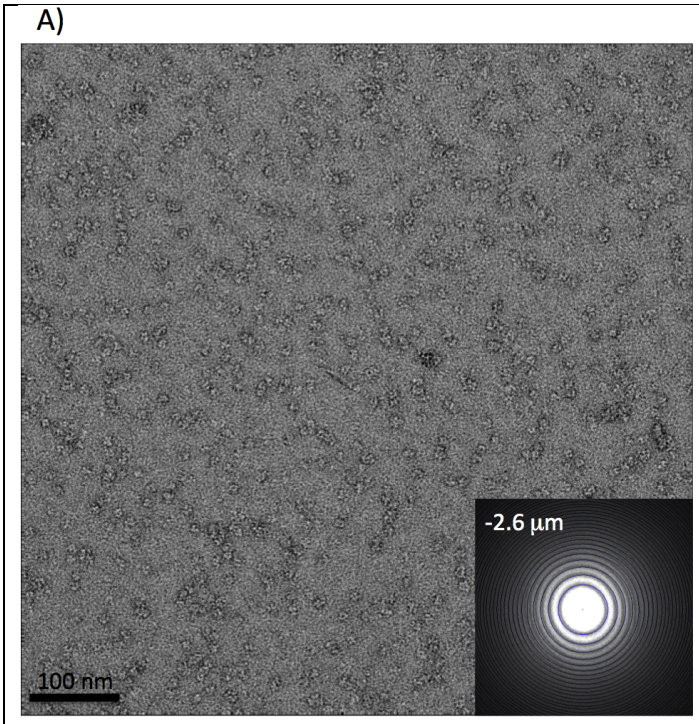


TF20 at 200kv	DmUGGT
Defocus range	-2 µm to -4µm
Magnification	62,000x with a 1.7Å/pixel
Particles picked	46,000 particles

Figure 26: Negative stain EM on DmUGGT

Micrograph (A) for DmUGGT at -2.4 µm defocus. Comparison between raw particles (B) and reference-free class averages (C). The table (D) provides data collection details.

Similar to the results observed for DmUGGT, the negative stain EM data for PcUGGT shows that the particles display random orientations on the grid and there are no signs of higher molecular weight aggregates, consistent with a well folded and stable protein sample (see Figure 27). The size of the particles is also in agreement with the particles from DmUGGT, with the size of particles ranging from 10 to 12 nm. The total number of particles boxed was ~48,000 particles, with defocus values ranging from -2 to -4 µm. The reference free classification of the boxed particles revealed very similar classes to those observed for DmUGGT. Some classes display a C-shape, while others display a ring-shape with three or four lobes. As was the case before, the 2D class averages showed an even distribution of particles across classes. There were no classes that over represented or over populated with particles. The combined results of the negative stain Dm- and PcUGGT show that both enzymes share a similar topology and are amenable for three-dimensional reconstructions.



D)

TF20 at 200kv	PcUGGT
Defocus range	-2 µm to -4µm
Magnification	62,000x with a 1.7Å/pixel
Particles picked	48,000 particles

Figure 27: Negative stain EM on PcUGGT

Micrograph (A) for PcUGGT at -2.6 µm defocus. Comparison between raw particles (B) and reference-free class averages (C). The table (D) provides data collection details.

3.3.2 EM map of Dm- and PcUGGT

The structure of DmUGGT by negative stain EM shows a claw-like architecture, with a large ring-shaped lobe and with a smaller globular lobe (see Figure 25 and Figure 28). I reconstructed initial models for DmUGGT using different methods. The first used the common-lines methodology built into Eman to reconstruct an initial three-dimensional model based on the Fourier space correlations of two-dimensional class averaged images. The second used the SIMPLE software, a software package specially developed to reconstruct initial models of asymmetric macromolecules using particle stacks or class averaged images as initial input and relying on a more elaborate algorithm using the common-lines method. The initial models using both methods looked very similar, with an asymmetric C-shape and a large central cavity. The final refined structure of DmUGGT showed a bipartite architecture, with a large ring-shaped lobe at

the top and a smaller hook-shaped globular lobe. The structure is 110 Å in its longest dimension and the calculated volume of the map matches that of a 170 kDa monomer.

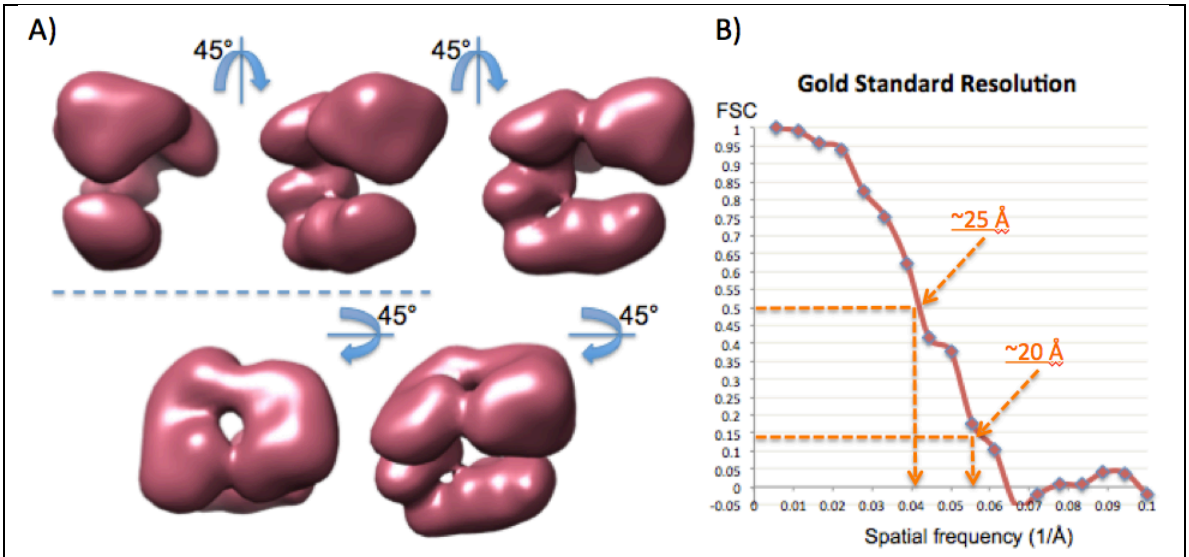


Figure 28: Structure of DmUGGT by negative stain EM

Refined structure of DmUGGT (A), displaying an asymmetric "claw"-shape with a large central cavity, and Fourier shell correlation curve (B).

The structure of PcUGGT by negative stain EM displays the same architecture as that of DmUGGT, with a claw-like shape formed by a large ring-shaped lobe and a small hook-shaped lobe (see Figure 25 and Figure 29). Just as with the DmUGGT, I calculated initial three-dimensional models using the common-lines algorithm with comparable results. Using the refinement method where the particles are split in two random halves and refined separately, the structure of PcUGGT displayed a large ringed lobe at the top and the smaller globular domain, forming a large central cavity. The structural features of the map are comparable to those of the class averages. The volume of the EM map for PcUGGT is smaller than that of DmUGGT, more in line with a 150 kDa protein rather than the 163 kDa theoretical molecular weight. The refined model is more compact than that of DmUGGT, being 100 Å in its longest dimension. In terms of volume and dimensions, the structure agrees with the parameters derived through SEC-MALS (see Figure 20) and SAXS for both species (see Figure 22 and Figure 24).

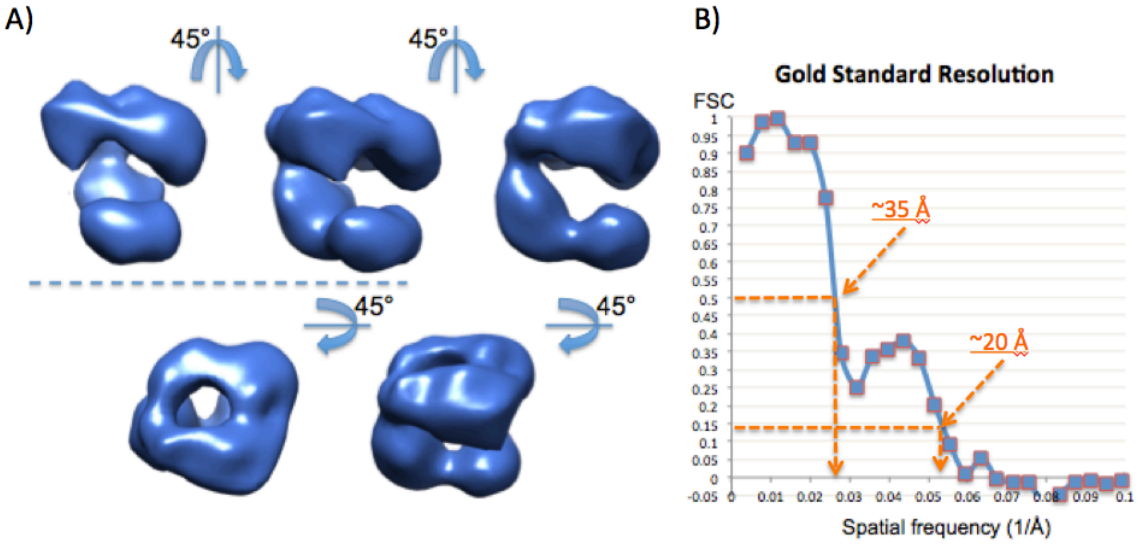


Figure 29: Structure of PcUGGT by negative stain EM

Refined structure of PcUGGT (A), displaying an asymmetric "claw"-shape with a large central cavity, and Fourier shell correlation curve (B).

3.3.3 SAXS *ab initio* modeling of Dm- and PcUGGT

SAXS data can be used to calculate structural models of proteins in solution at very low resolution, typically at 30 to 40 Å resolution. The models include the protein hydration shell and are sufficient to fit crystal structure of the protein within the structure. The method is useful in the absence of other data, although the accuracy of the structural models requires scrutiny, especially for proteins that have not been previously crystallized. The *ab initio* modeling takes the data distribution curves as a starting point (see Figure 30). Based on the maximal length of the protein, a sphere is created filled with beads. The algorithm will randomly remove beads from the sphere, and calculate the theoretical scattering of the beads left in the sphere. The theoretical scattering is then compared to the experimental scattering and the process is continued until there is good agreement between the theoretical scattering and the experimental SAXS data. The process generates variation in the resulting arrangement of the beads; many structures can be calculated that agree with the experimental SAXS data. To compensate for this, the structures are averaged to create a more representative structure of the protein.

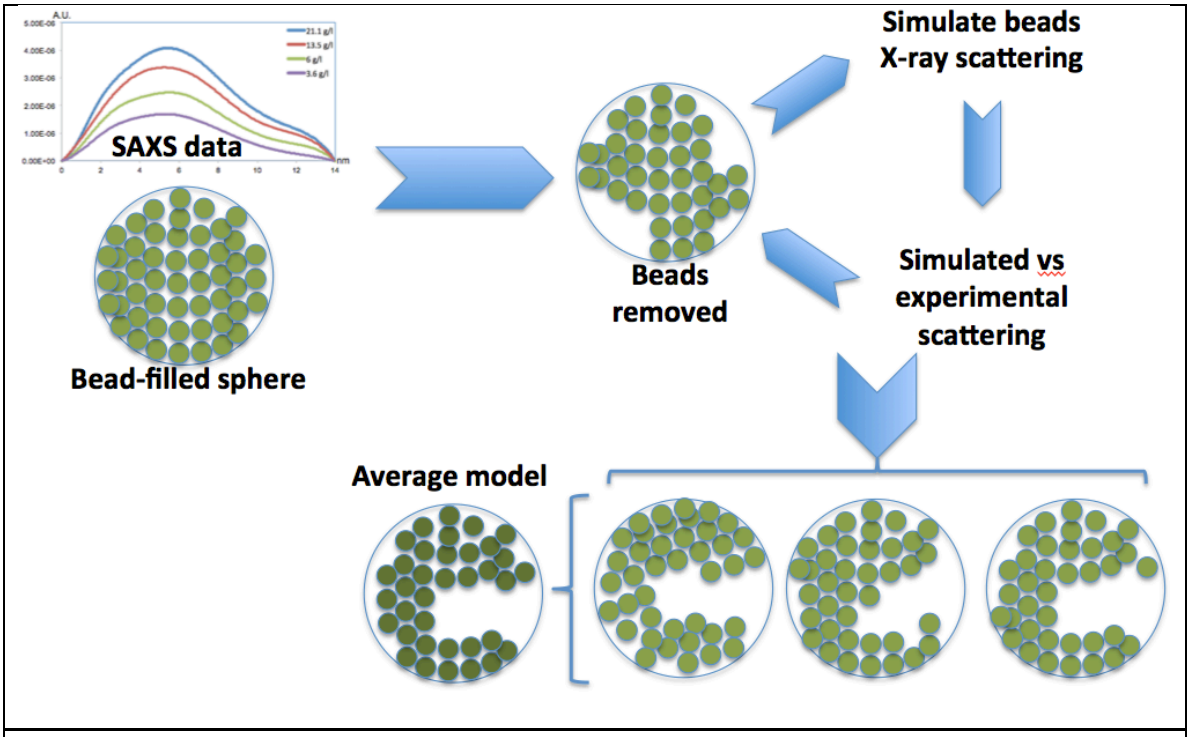


Figure 30: Basics for *ab initio* reconstructions using SAXS data

Workflow for SAXS *ab initio* modeling, relying on the processed scattering data and the generation of bead models. The models are hydration shell approximations that have theoretical scattering profiles consistent with the experimental protein scattering. Due to the variability across models, a final averaged model is calculated.

Given that the P(r) analysis suggested the maximal size of DmUGGT is less than 140 Å, I calculated *ab initio* models using spheres of 140 Å, 120 Å and 100 Å in diameter (see Figure 31). All of the resulting models display a bipartite topology, with a large lobe with a depression and a small lobe. The arrangement between the two lobes varied, with the two lobes coming closer to each other as the initial sphere diameter was reduced. To facilitate comparison between the SAXS *ab initio* models and the EM data, I calculated back projections of the SAXS bead models. The back-projections of SAXS model calculated with a 100 Å sphere was remarkably reminiscent of the two-dimensional classes of DmUGGT in negative stain (see Figure 26).

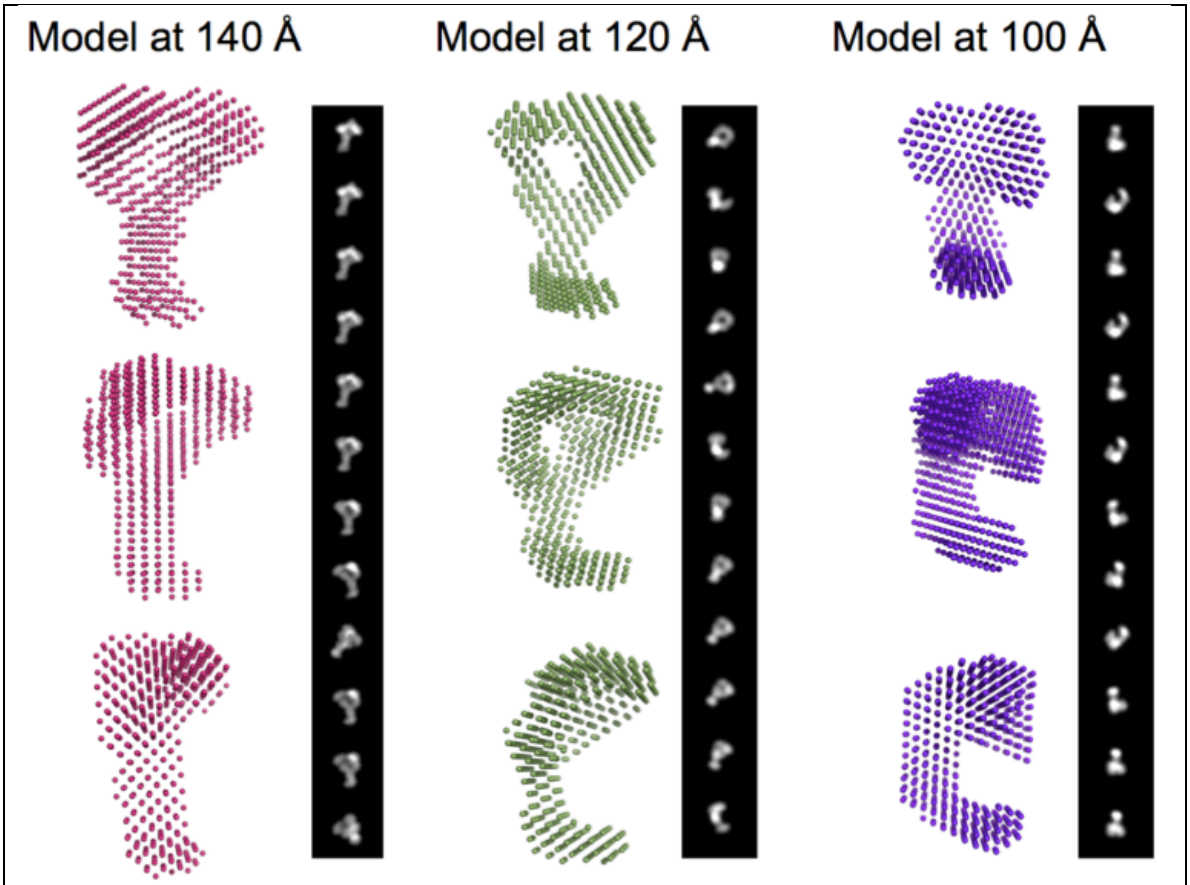


Figure 31: DmUGGT models using initial spheres of decreasing diameter

DmUGGT *ab initio* models using initial volume bead-filled spheres with decreasing diameter, of 140 Å (left panel), 120 Å (middle panel) and of 100 Å (right panel). For each model, three views are presented. Back-projections of every model were calculated to facilitate comparison to the negative stain EM structures.

I took the same approach for the PcUGGT SAXS data, calculating *ab initio* models using spheres of 140 Å, 120 Å and 100 Å in diameter (see Figure 32). As before, the resulting models display a two-lobe topology, with a large lobe with a depression and a small lobe, with some variation in the arrangement between the two lobes. As the sphere's diameter used for modeling was reduced, the space between the two lobes became smaller. The back-projections of 100-Å-diameter model were the closest to the class averages observed in the negative stain EM data of PcUGGT (see Figure 27).

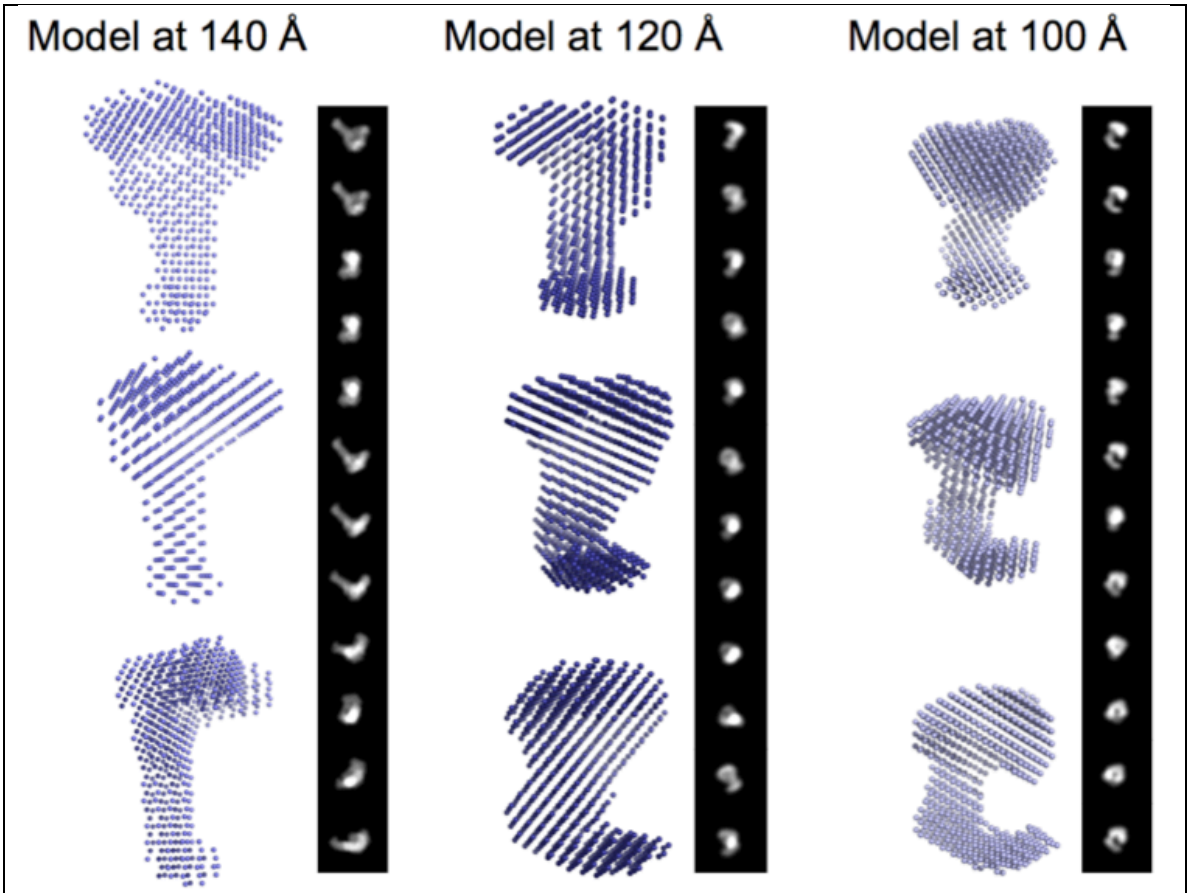


Figure 32: PcUGGT structures using initial spheres of decreasing diameter

PcUGGT *ab initio* models using initial volume bead-filled spheres with decreasing diameter, of 140 Å (left panel), 120 Å (middle panel) and of 100 Å (right panel). For each model, three views are presented. Back-projections of every model were calculated to facilitate comparison to the negative stain EM structures.

3.3.4 Comparison between EM and *ab initio* SAXS models

The side-by-side interspecies comparison between the structures for UGGT determined by negative stain EM and by SAXS *ab initio* modeling shows remarkable agreement, with a conserved two-lobed arrangement and similar dimensions (see Figure 33). The front and side views of the UGGT models share the characteristic C-shape, which can also be observed in the back-projections of the SAXS and EM models. The space between the two regions of UGGT is larger in the SAXS models, which can be explained due to the way SAXS modeling has been optimized. The major assumption in

bead modeling is that proteins don't have big cavities inside them, which makes modeling UGGT a complicated target to tackle through this technique. The use of two different techniques bolsters the confidence in these structures, especially considering the structural conservation of two distant UGGT species.

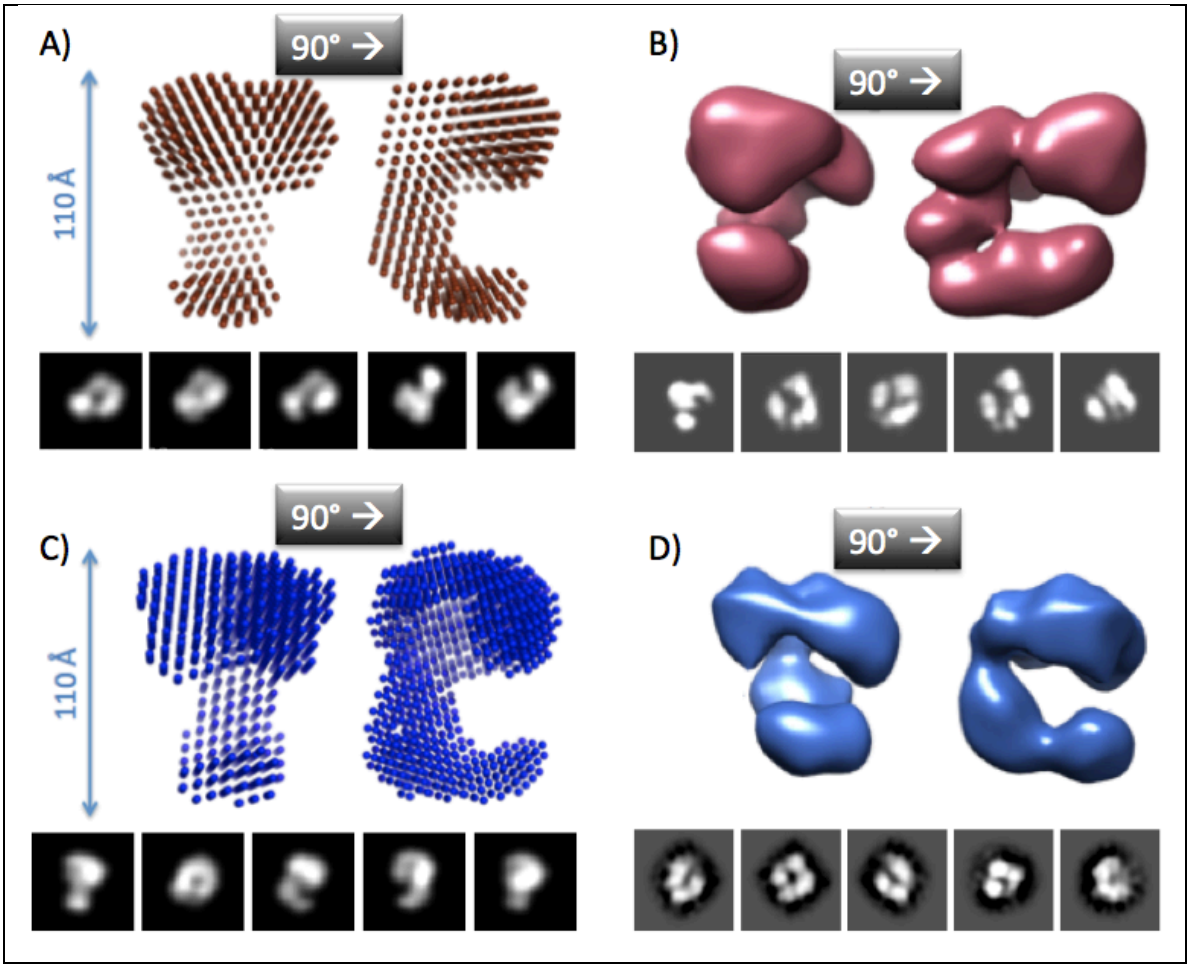


Figure 33: Comparison between SAXS and negative stain EM models

Panel A and C, structures of DmUGGT and PcUGGT, respectively, by SAXS *ab initio* modeling. Panel B and D, structures of DmUGGT and PcUGGT, respectively, by single particle negative stain EM. Back-projections are provided for every structure.

3.3.5 Putative substrate recognition by UGGT

Based on the conserved global features of our UGGT models and on the biochemical data available, we have put forward a basic mechanism for selective

misfolded glycoprotein labeling by UGGT (see Figure 34) that excludes folded proteins from glucosylation and that stands apart from other misfolded protein binding proteins, such as members of the Heat Shock Protein family and ER lectin chaperones. Many studies have shown that UGGT binds to a hydrophobic region and to the glycan on the substrate, with glucosylation occurring for substrates that carry these two motifs within less than 40 Å of each other. This suggests that the misfolded protein sensor region and the glucosyltransferase catalytic pocket are relatively close together in space. This is reinforced by the experimental observation that glucosylation of unfolded glycoproteins is competitively inhibited by unfolded non-glycosylated proteins.

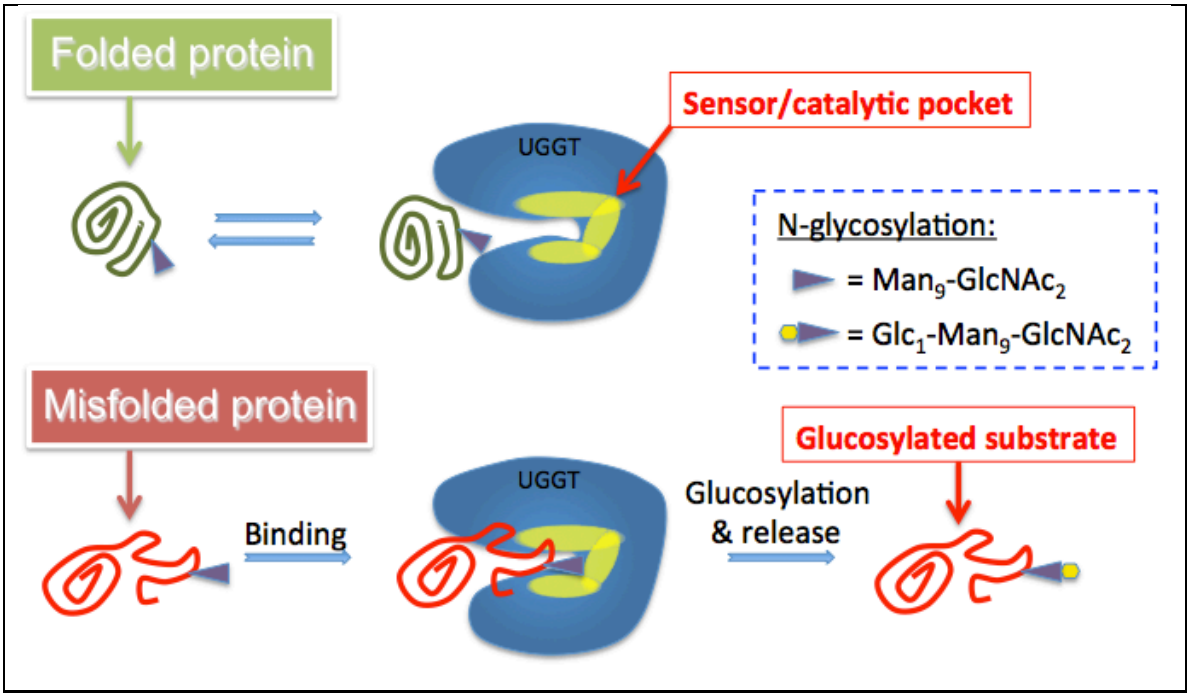


Figure 34: Substrate selection model by UGGT

In the top half, the glycan on a well-folded protein is incapable of reaching into the deep central cavity of UGGT, and can't be glucosylated. In the bottom half, the N-glycan on a flexible hydrophobic loop is pulled deep into the central cavity for glucosylation, with the help of hydrophobic interactions between the sensor domain and the flexible loop.

Our EM and SAXS structure feature a large central cavity formed by two asymmetric lobes (see Figure 33). It is this cavity we hypothesize houses the catalytic site

for glucosylation and the misfolded protein recognition patch. In our model, if a glycoprotein with exposed hydrophobic loops comes near UGGT, the hydrophobic loops with a nearby N-glycan could enter the central cavity of UGGT (see Figure 34, bottom half), which would stabilize the complex and trigger glucosylation of the N- glycan on the client protein. On the other hand, if a protein does not present any exposed hydrophobic loops, the glycan would not be able to reach into the transferase catalytic pocket and glucosylation would not occur (see Figure 34, top half).

3.4 Discussion

As stated previously, the structure and mechanism of action of UGGT has been the subject of many studies, although the structural insights have been very limited[56]. Our results have provided the first full-length structures for UGGT from two different species, using two different low-resolution techniques (see Figure 33). Our structures are not only novel, they also provide a putative substrate recognition mechanism (see Figure 34) that is consistent with the literature and provides a unique mechanism of action, different than that of other cytosolic or ER chaperones.

The UGGT structures showed remarkably conserved structural features across techniques and species: a large ring-shape lobe connected to a hook-shaped lobe, separated by a ~25 Å wide central cavity. At this stage, the domain composition of these two lobes is unknown. Though the structure for both species agreed with the observed class averages and back-projections, the EM structure of DmUGGT (see Figure 28) was of better quality than that of PcUGGT (see Figure 29), as the latter seems to have a smaller volume than expected and its FSC curve was not a smooth curve (see Figure 25). Given the known requirement for substrate recognition, we hypothesize that the central cavity between the two lobes contains the misfolded protein recognition surface as well as the catalytic pocket for the glucosyltransferase domain.

Though SAXS *ab initio* modeling is not optimized to model protein structures with large cavities or pockets, the SAXS models display the two-lobe topology observed in the EM maps (see Figure 33), particularly the most compact *ab initio* models. Our models are consistent with our previous findings of the SAXS scattering data, which showed UGGT had a non-compact multi-domain organization. Though small-scale flexibility between

the domains may exist *in vivo*, the low-resolution structure analysis by SAXS and EM did not reveal any conformational changes in UGGT.

As discussed previously, UGGT substrates must present exposed hydrophobic residues near the N-glycosylation sites, and these substrates can be competed out of the binding site by unfolded non-glycosylated proteins[34], showing that the substrate-binding surface and the glycan binding surface are present in one large cavity or very close to each other. Our models provide exactly that, a large central cavity that could house the catalytic pocket and the substrate-binding surface. Furthermore, the restricted space in which UGGT functional sites are present would prevent folded proteins from reaching into these two sites, effectively excluding them from glucosylation.

This mechanism stands apart from that of Heat Shock Proteins, chaperonins like GroEL/ES or ER lectin chaperones (see Figure 3). HSP proteins bind to misfolded proteins through hydrophobic regions, are heavily regulated by ATP/ADP binding, and have other effectors regulating their action[2, 7]. They follow a global bind-and-forced-release mechanism, using ATP/ADP as a switch between conformations. Chaperonins use ATP hydrolysis cycles to bind and release proteins, but they provide a large cage-like environment where a misfolded protein is caught and inside of which folding is promoted. ER lectin chaperones such as calnexin and calreticulin bind to their substrate through the monoglucosylated N-glycan[3, 12, 13] (see Figure 5). These proteins follow a no-questions-asked binding depending on that single glucosylation site and a timed release. UGGT mode of action is unique in that binding occurs through both exposed hydrophobic residues and Man_9-$GlcNAc_2$ glycan, which triggers mono-glucosylation and release.

To improve on these results, it would be of interest to probe UGGT flexibility in the presence of Sep15 or of UGGT substrates of various sizes. SAXS experiments could easily compare the biophysical parameters of the UGGT/Sep15 complex or of UGGT in the presence of synthetic substrates. This could reveal conformational changes upon substrate binding by UGGT and bead models could be calculated as well. It could also be possible to collect negative stain EM data on UGGT/Sep15 or of UGGT in complex with a synthetic substrate such as M9-MTX. Single particle analysis and structure reconstruction could illustrate changes in the structure of UGGT.

3.5 Material and Methods

3.5.1 SAXS *ab initio* modeling of UGGT

After doing the Guinier, Krakty and distance distribution analysis for Dm- and PcUGGT SAXS data at all concentrations, we utilized Dammif[66] to generate 20 *ab initio* models using bead-filled spheres of decreasing diameter, sampling spheres of 100 Å to 180 Å in diameter in 20 Å increments. The generated bead models for each diameter were aligned and averaged using Damaver. To facilitate comparison to the 2D dimensional negative stain EM data, bead models were converted into density maps, low-pass filtered to 25 Å resolution and two-dimensional back-projections were calculated.

3.5.2 Sample preparation for negative stain EM and data collection

A solution of uranyl formate at 0.75% (w/v) and neutral pH was used as negative stain. Protein concentrations raging from 8 to 20 ng/µL were tested, with an optimal protein concentration of ~12 ng/µL. A volume of 5 µL of protein solution was applied to carbon-coated copper grids for 60 seconds and then blotted, immediately applying 5 µL of uranyl formate stain, which was blotted after 60 seconds. The grids were allowed at least an hour to dry before inserting into the microscope. Imaging was done using a FEI Tecnai G^2 TF20 at 200kV equipped with a Gatan Ultrascan 4000 CCD camera (model 895), at 62,000x magnification at a 1.8 Å/pixel and micrographs were recorded at varying defocus values.

3.5.3 Reconstructing EM map of UGGT at 20 Å resolution

Micrographs and particle processing was carried out with Eman2.0 software[74, 76]. Particles were picked using e2boxer with a 140x140 pixel box size, yielding ~46,000 particles for DmUGGT and ~48,000 particles for PcUGGT, boxed from about 250 micrographs for each species dataset. The particles were classified using Eman2.0 reference-free algorithm K-means, and only particles from well-defined class-averages were kept for further processing. The high-quality particles were used to generate initial models by the common-lines algorithm within Eman2.0 and by the SIMPLE PRIME[77, 93] method. We refined these models within Eman2.0 using the Fourier-shell correlation method, which randomly splits the data in two halves and refines each half separately. This allows a Fourier-shell correlation to be calculated between the two maps and an

accurate estimation of the resolution. UGGT models were compared using Chimera, at thresholds levels giving equivalent voxel-volumes. Back-projections of the models were calculated using Eman2.0's e2project3d.py for comparison to the class averages.

Preface to domain identification and cryo-EM

The structures presented in the previous chapter provide a unique view into UGGT structure, which allowed us to provide a simple substrate recognition mechanism. To improve upon our novel UGGT structures, we needed to identify UGGT domains within our maps and improve the amount of details available in our structures. To address this, we first labeled PcUGGT with bulky position-specific labels and carried out negative stain EM experiments. Secondly, we took advantage of the recent developments in the cryo-EM field to attempt data collection of UGGT embedded in vitreous ice utilizing state-of-the-art direct electron detection cameras.

For this chapter, I spent three months at the University of Washington collaborating with Dr. Justin Kollman in order to learn cryo-EM sample preparation techniques and microscope handling for cryo-EM data collection. I collected various cryo-EM datasets, albeit only the DmUGGT dataset was of high-enough quality to achieve medium resolution maps. With additional advice from Dr. Huy Bui Khanh, I treated the cryo-EM DmUGGT data to achieve a more detailed and complete map of UGGT. With its relatively small molecular weight, UGGT is at the limit of what has been done using cryo-EM[57, 60] and our results are proof that cryo-EM is ready to tackle sub-200 kDa protein samples. To better orient our maps, Dr. Marie Ménade and Dr. Guennadi Kozlov produced PcUGGT mutants with biotinylation sites, allowing site specific labeling with monovalent streptavidin. I collected negative stain EM data for the successfully SA-labeled PcUGGT mutants and reconstructed maps for three PcUGGT-biotin/SA samples. Furthermore, I calculated homology models for the domains of Dm- and PcUGGT, which I was able to fit within the electron density maps. Together, these results offer a clear picture of UGGT architecture and provide a stronger model for misfolded substrate selection by this enzyme.

4 UGGT domain identification and cryo-EM

Cover figure:

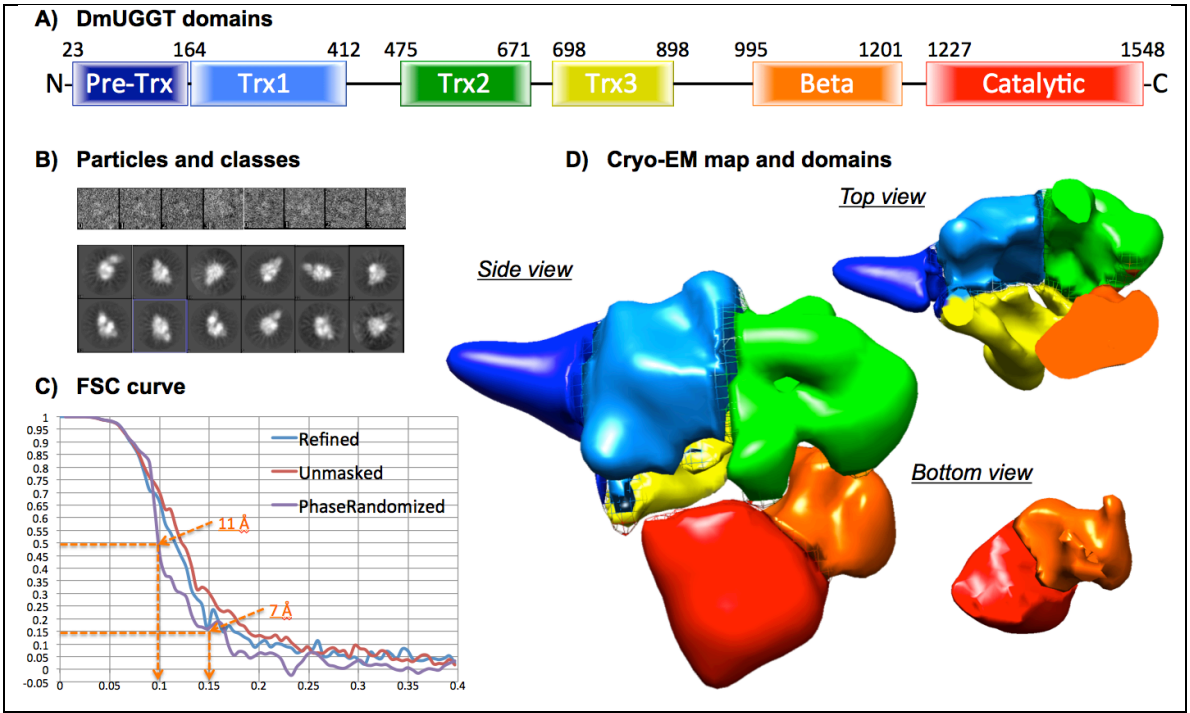


Figure 35: Structure of DmUGGT by cryo-EM

Schematic representation of DmUGGT domains (A). Comparison between cryo-EM particles and class averages (B). Refined structure (D) by cryo-EM for full-length DmUGGT at 11 Å resolution and corresponding Fourier shell correlation (C).

4.1 Abstract

The full-length structure of UGGT has eluded many groups, principally because the only technique previously suited to take UGGT was crystallography. Despite numerous attempts, UGGT has remained an elusive target for crystallization. Our work focused on using electron microscopy to determine structures of this unique enzyme, through which we previously showed that UGGT has a claw-like two-lobed structure. To identify UGGT domains, we made PcUGGT mutants with biotinylation sites in different domains. We used negative stain EM to reconstruct maps of the PcUGGT-biotin/SA (monovalent streptavidin) through which we unequivocally identified UGGT domains within our EM maps. Furthermore, we successfully imaged DmUGGT through cryo-EM

and reconstructed a 10 Å resolution map using the latest image processing tools. Our structure provides a solid fit for homology models of UGGT individual domains and reveals the mechanism of unfolded substrate selectivity by UGGT.

4.2 Introduction

The field of cryo-electron microscopy has rapidly expanded in the past five years to tackle increasingly smaller targets and achieving structures at resolutions rivaling those achieved by crystallography and x-ray diffraction experiments[57-60, 71, 94]. Studying asymmetric proteins of less than 200 kDa was not possible by cryo-EM because the conventional detectors were not capable of acquiring detailed micrographs of small proteins, and the lack of details and symmetry of the particles prevented accurate alignment to build high-resolution models. These issues have been addressed through a series of technological advances and advanced image-processing software. The advent of direct electron detection cameras has revolutionized the cryo-EM field, allowing the collection of movie micrographs where random particle drift can be corrected. This is useful to account for beam induced damage and to produce sharper images with improved signal to noise ratios[80, 81]. Together with the development of automated data collection software, it is possible to acquire very large datasets in a matter of days, from which hundreds of thousands of particles can be collected. Technological improvements in computer power and advanced data processing software have allowed the treatment of these large datasets, while remaining time and cost effective.

The resolution barriers for sub-megadalton proteins continue to decrease through automated data collection using high-voltage top-of-the-line electron microscopes equipped with the K2 Summit direct detection camera and a contrast-boosting energy filter. Until very recently, the highest resolution achieved so far for a sub-megadalton protein was for the structure of β-galactosidase, a protein complex of 465 kDa, resolved at 2.2 Å resolution[59]. One of the smallest proteins tackled so far using cryo-EM is γ-secretase, a heavily glycosylated (30-70 kDa worth) membrane protein complex of 170 kDa, whose structure was determined to a resolution of 4.5 Å using cryo-EM[57]. In the most recent study, the structures of glutamate dehydrogenase (334 kDa), of lactate

dehydrogenase (145 kDa) and of isocitrate dehydrogenase (93 kDa) were resolved to 1.8 Å, 2.8 Å and 3.8 Å, proof of how fast the field is expanding[60].

My previous results showed that UGGT has a claw-like architecture (see Figure 25), which we hypothesize houses the glucosyltransferase catalytic pocket and the misfolded protein sensor surface within the interior of the claw (see Figure 33 and Figure 34). Though the orientation and the position of UGGT domains could not be identified at the resolution levels afforded by negative stain EM or SAXS modeling, our proposed structure-based substrate recognition model is consistent with the vast biochemical data on UGGT. To address the limitations of my previous work, I used a combination of negative stain EM with monovalent streptavidin site-specific labeled PcUGGT samples[89] and cryo-EM of DmUGGT using a state-of-the-art direct detection camera to identify the position of UGGT domains within our electron density maps and to improve the resolution of our structure (see Figure 35). Furthermore, I used homology modeling[86] to build structural models for the individual domains of UGGT, which I was able to fit within the EM maps of both Dm- and PcUGGT. Using these three approaches, I was able to improve the model we proposed for UGGT substrate selection. Our revised model gives a simple mechanism for unfolded glycoprotein recognition, which is unique and different from the recognition mechanism of other cellular chaperones.

4.3 Results

4.3.1 Purification of PcUGGT-biotin/SA mutants

After carrying out sequence analysis and introducing sixteen different biotinylation sites, we were able to successfully produce, purify and label with SA a total of five PcUGGT-biotinylated mutants (see Figure 36). Using secondary structure prediction software, we targeted sixteen positions where a biotinylation sequence could be added. Those sites were generally after an alpha helix, in an area predicted to be an unfolded loop, and the mutation sites spread throughout the sequence of PcUGGT. Two mutants contained a biotinylation site in the N-terminal pre-thioredoxin-domain region; two mutants, in the first thioredoxin-like domain (Trx1); one mutant, in the linker region between the first and second thioredoxin-like domain; one mutant, in the second

thioredoxin-like domain (Trx2); two mutants, in the third thioredoxin-like domain (Trx3); three mutants, in the beta-sheet rich region (Beta); one mutant, in the linker region between the beta-rich region and the catalytic domain; and four mutants, in the C-terminal glucosyltransferase domain. Eleven mutants were not amenable for structural work, showing a combination of poor expression yields and protein degradation and precipitation during purification trials. Five mutants were stable, labeled with SA and purified through gel filtration (see Figure 36, panel B). The biotinylation sites in SA-labeled mutants are in the Trx1 domain (mutant 3 and 4), in the Trx3 domain (mutant 7), in the linker region between the beta-rich domain and the catalytic domain (mutant 12), and in the catalytic domain (mutant 15).

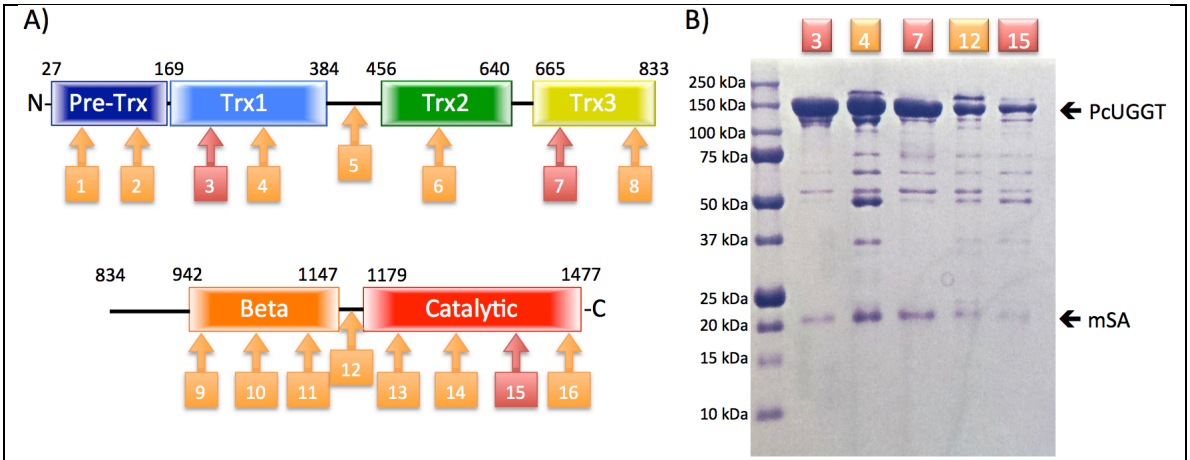


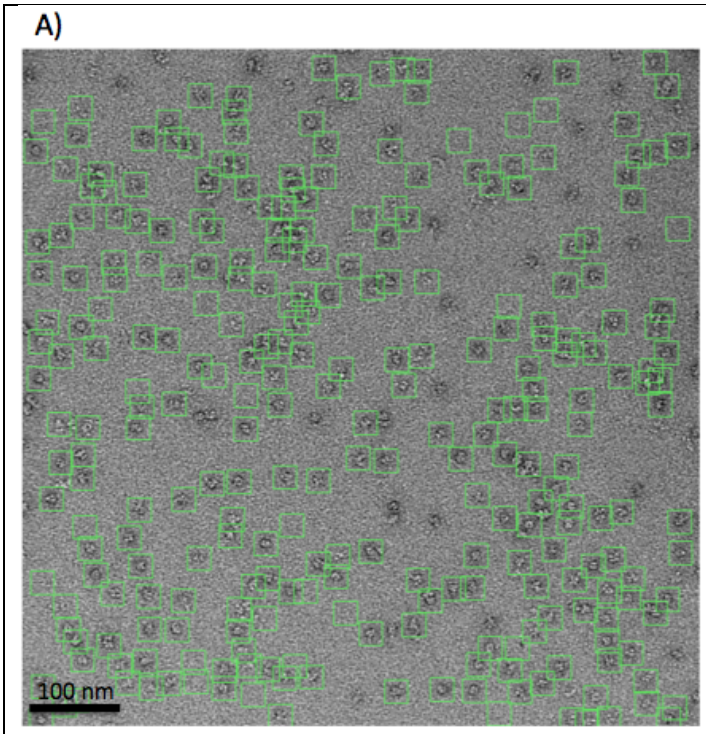
Figure 36: Domains, biotinylation sites and purified SA labeled PcUGGT

Schematic representation of PcUGGT domains (A). SDS-PAGE analysis of purified PcUGGT mutants after SA-labeling (B), on a 12% acrylamide gel.

4.3.2 Single particle analysis for PcUGGT-biotin/SA mutants

Similar to the EM results observed in the previous chapter, the negative stain EM data for PcUGGTm3-biotin/SA shows that the particles display random orientations on the grid and there are no signs of higher molecular weight aggregates (see Figure 37). The size of the particles agrees with the wild-type PcUGGT EM results, with particles ranging from 10 to 12 nm in length. The total number of particles boxed was ~30,000 particles, with defocus values ranging from -1.5 to -3.5 μm. The reference free classification of the

boxed particles revealed very similar classes to those observed previously, with some classes showing a smaller additional lobe, consistent with a 50 kDa SA label. The negative stain single particle analysis for PcUGGTm4-biotin/SA, which also carried the SA in the Trx1 domain, revealed that protein sample to present significant higher molecular weight aggregates, which made further analysis of that sample unnecessary.



TF20 at 200kv	PcUGGTm3-biotin/SA
Defocus	-1.5 µm to -3.5µm
Magnification	62,000x with a 1.7Å/pixel
Particles picked	~30,000 particles

Figure 37: Negative stain EM on PcUGGTm3-biotin/SA

Representative micrograph of negative stain (A), particles extracted (B), reference-free class averages (C) and summary table (D) for PcUGGTm3-biotin/SA EM data.

As with wild type PcUGGT and PcUGGTm3-biotin/SA, the negative stain EM data for PcUGGTm7-biotin/SA shows particles with random orientations on the grid and no signs of higher molecular weight aggregates (see Figure 38). As before, the length of particles ranges from 10 to 12 nm. The total number of particles boxed was ~30,000 particles, with defocus values ranging from -1.5 to -3.5 µm. The reference-free class averages were consistent with the rest of the negative stain work, particularly that of PcUGGTm3-biotin/SA (see Figure 37), with class averages reminiscent of the wild-type

protein data but with some classes displaying an additional small lobe. The size of the lobe is consistent with streptavidin label and seems positioned differently than that observed in PcUGGTm3-biotin/SA.

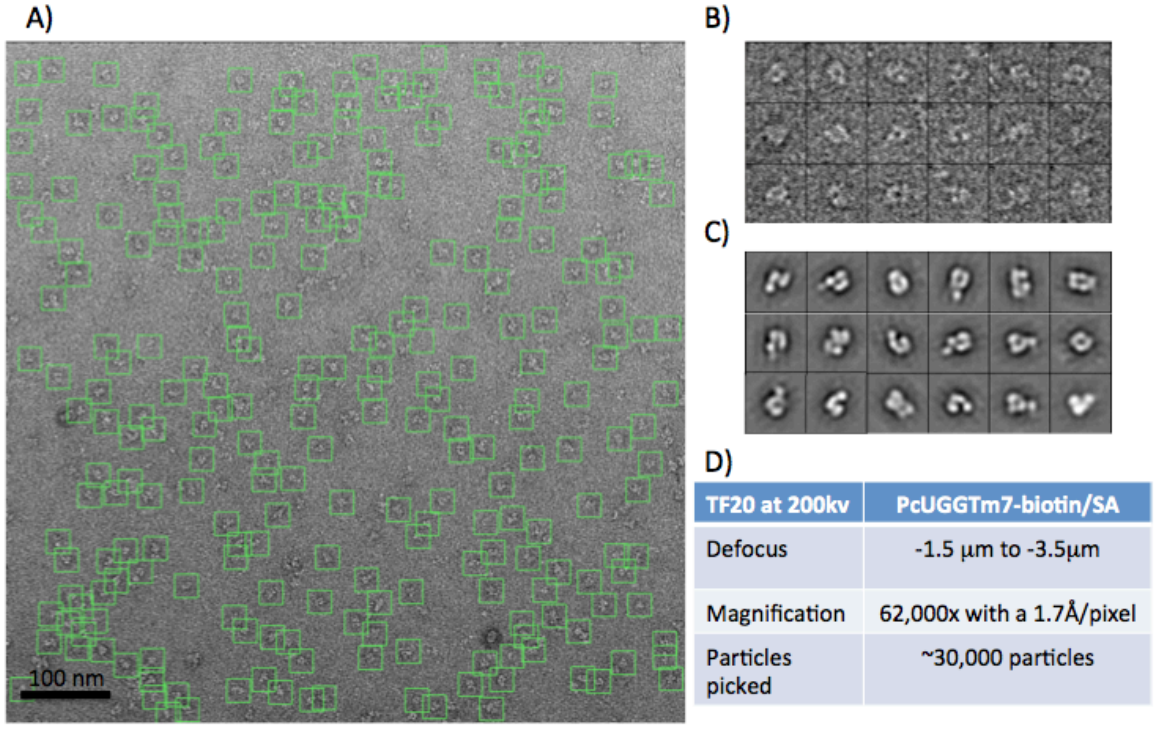
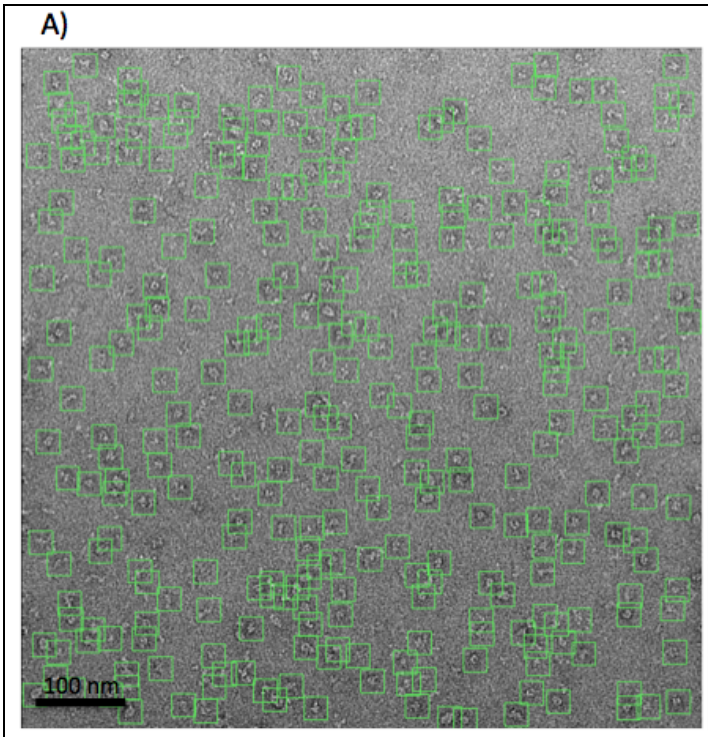


Figure 38: Negative stain EM on PcUGGTm7-biotin/SA

Representative micrograph of negative stain (A), particles extracted (B), reference-free class averages (C) and summary table (D) for PcUGGTm7-biotin/SA EM data.

As with PcUGGTm3-biotin/SA and PcUGGTm7-biotin/SA, the negative stain EM data for PcUGGTm15-biotin/SA presents particles in random orientations and devoid of higher molecular weight aggregates (see Figure 39). The particles and class averages are slightly longer than those observed previously, with lengths ranging from 10 to 14 nm. I boxed ~30,000 particles, with defocus values ranging from -1.5 to -3.5 µm. The reference free class averages offer similar classes to the ones observed previously, with some newer classes unique to this mutant. In this sample, the additional lobe corresponding to the 50 kDa SA label visible in the C-shape class averages, which now look more like a larger ring. The negative stain single particle analysis for PcUGGTm12-

biotin/SA, which carried the SA in the linker region between the beta-rich and catalytic domain, revealed the presence of higher molecular weight aggregates, and thus further processing was abandoned.



TF20 at 200kv	PcUGGTm15-biotin/SA
Defocus	-1.5 µm to -3.5µm
Magnification	62,000x with a 1.7Å/pixel
Particles picked	~30,000 particles

Figure 39: Negative stain EM on PcUGGTm15-biotin/SA

Representative micrograph of negative stain (A), particles extracted (B), reference-free class averages (C) and summary table (D) for PcUGGTm15-biotin/SA EM data.

4.3.3 EM map of PcUGGT-biotin/SA mutant 3

The structure of PcUGGTm3-biotin/SA, determined through refinement using PcUGGT structure as initial model, shows an additional density between the ring lobe and the smaller hook-shaped lobe (see Figure 40). The extra density we observed is attributed to the SA label. Combining the information provided by the three PcUGGT mutant structures determined, I identified the ring-shaped lobe to the sensor region of UGGT, formed by the three thioredoxin-like domains. In the structure of PcUGGTm3-biotin/SA, the bulk of the streptavidin has pushed the sensor region away from the catalytic domain in an asymmetric way.

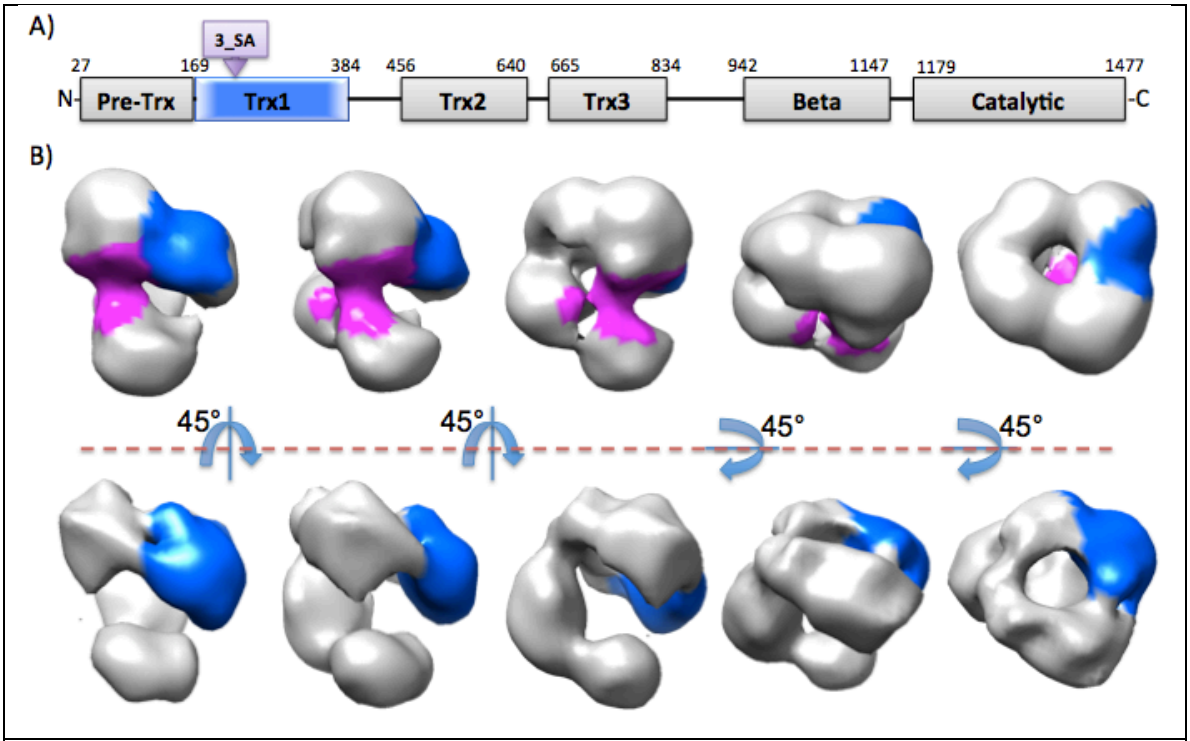


Figure 40: Structures of PcUGGTm3-biotin/SA and PcUGGT

Schematic representation of PcUGGT domains (A) and comparison between the refined structure of PcUGGTm3-biotin/SA (B, top panel) and wild-type PcUGGT (B, bottom panel). The structures are color coded by domains: Trx1 domain shown in light blue and SA label in pink.

4.3.4 EM map of PcUGGT-biotin/SA mutant 7

The structure of PcUGGTm7-biotin/SA, determined through refinement using PcUGGT structure as initial model, shows an additional density at the top of the ring-shaped region of UGGT (see Figure 41). The added density is compatible with the molecular weight and volume of SA. Between the structural information provided by this mutant and the PcUGGTm3 mutant, I identified the relative position of the Trx1 and Trx3 domains. The only possible position for the Trx2 domain is at the back of the ring-shaped region and between the Trx1 and Trx3 domains.

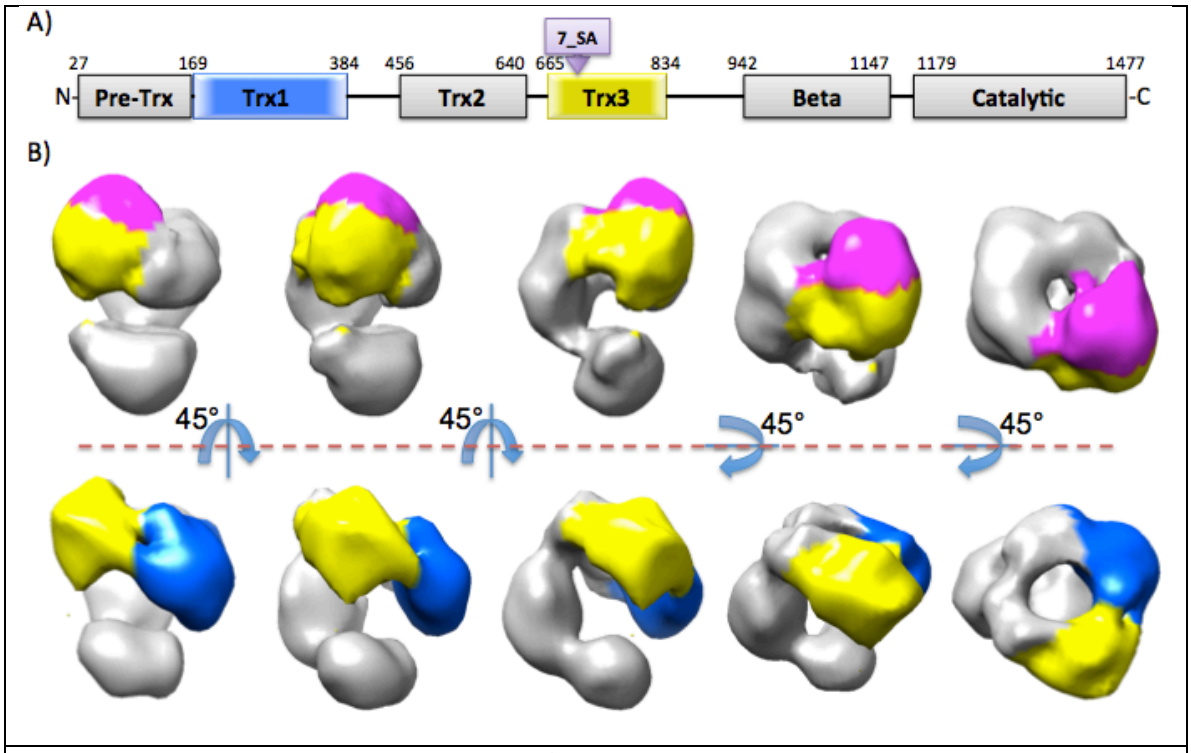


Figure 41: Structures of PcUGGTm7-biotin/SA and PcUGGT

Schematic representation of PcUGGT domains (A) and comparison between the refined structure of PcUGGTm7-biotin/SA (B, top panel) and wild-type PcUGGT (B, bottom panel). The structures are color coded by domains: Trx1 domain shown in light blue, Trx3 domain in yellow and SA label in pink.

4.3.5 EM map of PcUGGT-biotin/SA mutant 15

The structure of PcUGGTm15-biotin/SA, determined through refinement using PcUGGT structure as initial model, shows a very distinct additional density at the end of the hook-shaped domain of UGGT (see Figure 42). As before, the added density is compatible with the volume and weight of SA. This mutant structure clearly identifies the position of the catalytic domain of PcUGGT and confirms our identification of the ring-shaped region as the sensor region of UGGT. The only possible position for the beta-sheet rich domain is immediately next to the catalytic domain and below the ring-shaped lobe. The position of the SA in our structure explains why in the class averages the C-shaped class was replaced by a larger ring-shaped class average, which corresponds to a

side view of PcUGGTm15-biotin/SA where the SA is almost in contact with the sensor region formed by the thioredoxin-like domains.

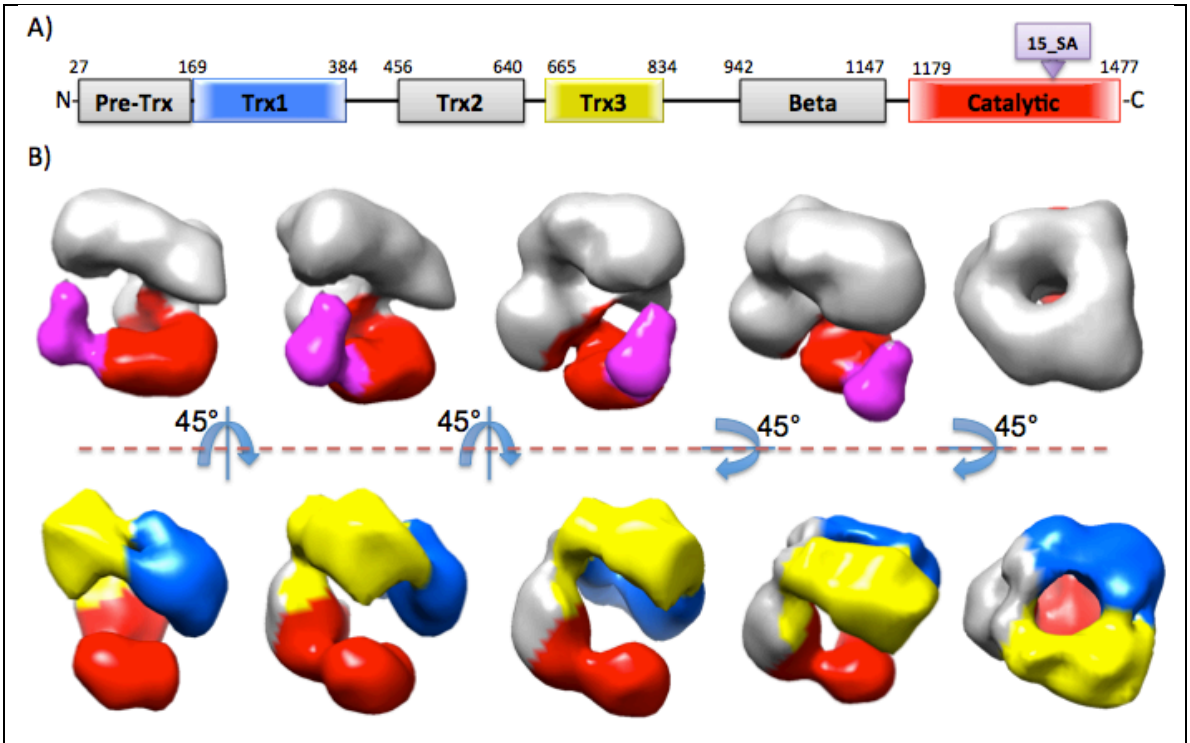


Figure 42: Structures of PcUGGTm15-biotin/SA and PcUGGT

Schematic representation of PcUGGT domains (A) and comparison between the refined structure of PcUGGTm15-biotin/SA (B, top panel) and wild-type PcUGGT (B, bottom panel). The structures are color coded by domains: Trx1 domain shown in light blue, Trx3 domain in yellow, catalytic domain in red and SA label in pink.

4.3.6 Homology models of PcUGGT within the EM map

Combining the negative stain EM structure of the wild-type, SA-labeled and homology models of the individual domains of PcUGGT, I found the relative position of the homology models within the EM structure (see Figure 43, panel B), enriching our structural understanding of UGGT substrate recognition. Based on secondary structure prediction results (see appendices, from Figure 62 to Figure 66) and previous domains identification from other groups (see Figure 8), I calculated homology models for the three Trx domains, the beta-sheet rich and the catalytic domains of PcUGGT using the

Phyre2 webserver modeler (see Figure 43, panel A). The structure of the pre-Trx region could not be modeled for PcUGGT, and the EM map in negative stain did not show density for this region. The models were calculated using protein structures with less than 26% sequence identity to UGGT. The resulting models for the three Trx domains and for the catalytic domain were calculated with a +98% confidence score and with +95% sequence coverage. Only 50% of the beta -sheet rich region could be modeled, albeit with a 95% confidence score. The EM map provides enough space for the five domains.

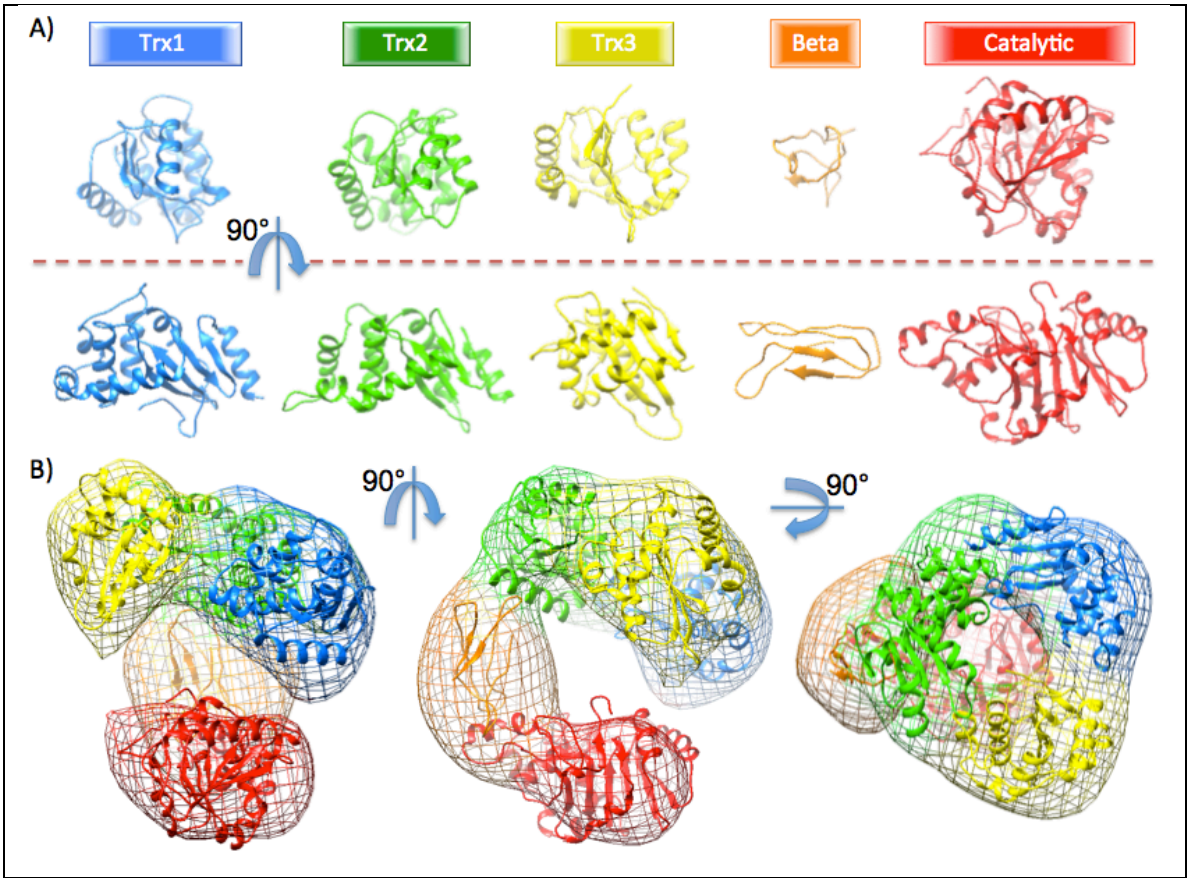


Figure 43: Homology models within the PcUGGT map

Homology models of PcUGGT domains (A) and fitting of the models within the negative stain structure of wild-type PcUGGT (B). The models are color coded as before: Trx1 domain in light blue, Trx2 domain in green, Trx3 domain in yellow, beta-sheet rich domain in orange and catalytic domain in red.

4.3.7 Single particle analysis for DmUGGT by cryo-EM

With the goal of achieving higher resolution maps of UGGT, I collected a cryo-EM data set for DmUGGT using the Leginon data collection software and the K2 Summit direct electron detection camera, allowing for frame drift correction and particle movement tracking and correction (see Figure 44). At 172 kDa, DmUGGT is among the smallest proteins to date for which cryo-EM data has been collected and analyzed, and very thin ice was crucial to high-quality data collection. I collected a total of 1500 movie micrographs, each of which consists of 36 frames, at a magnification of ~29,000x and a total dose of ~40 e/Å^2. I corrected for the global micrograph drift using driftcorr running on a GPU computer, and generated drift-corrected merged images to use for the single particle analysis (see Figure 44, panel B and C).

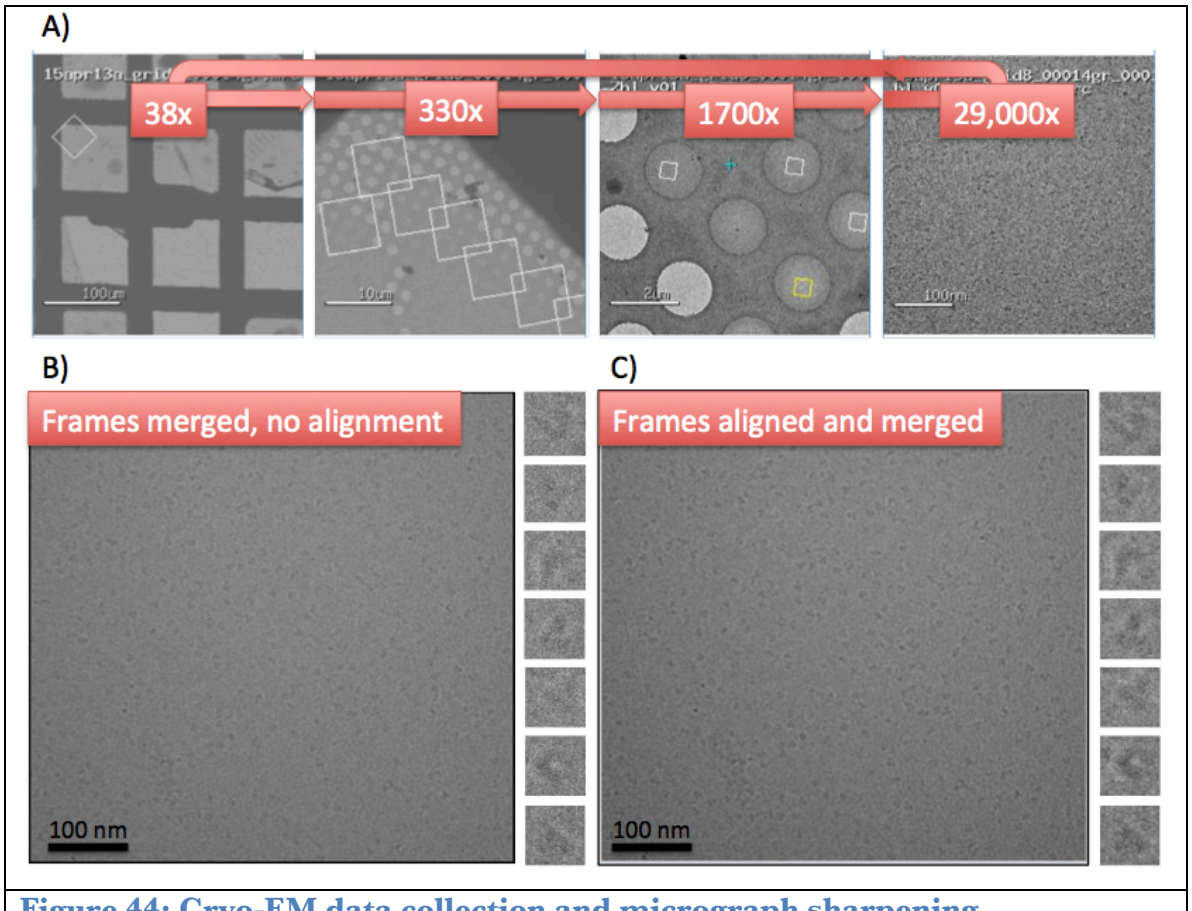


Figure 44: Cryo-EM data collection and micrograph sharpening

Workflow for automated cryo-EM data collection using Leginon (A), sequentially picking

imaging targets at increasing magnification and collecting the final micrograph at ~29,000 magnification. When merged, raw frames (B) result in a blurry micrograph. Correcting for global drift across frames (C) yields sharper micrographs and clearer particles, for more accurate particle alignment and reconstructions.

Using the drift-corrected micrographs, the single particle analysis showed a homogeneous sample, free of high molecular aggregate and offering random views of DmUGGT embedded in vitreous ice (see Figure 45). I utilized Eman to box particles from the micrographs, yielding ~191,000 particles in total, with picking done in a supervised semi-automated manner. The micrographs and the particle coordinates were used in Relion to extract and normalize the protein particles (see Figure 45, panel D). The defocus estimation was calculated using CTFfind3. The maximum-likelihood two-dimensional classification, carried out with Relion and CTF corrected particles, revealed detailed class averages with finer details (see Figure 45, panel E) than those previously seen through negative stain EM, which ultimately allowed a more complete map of UGGT to be calculated. A circular mask of 180 Å was used during the 2D classification.

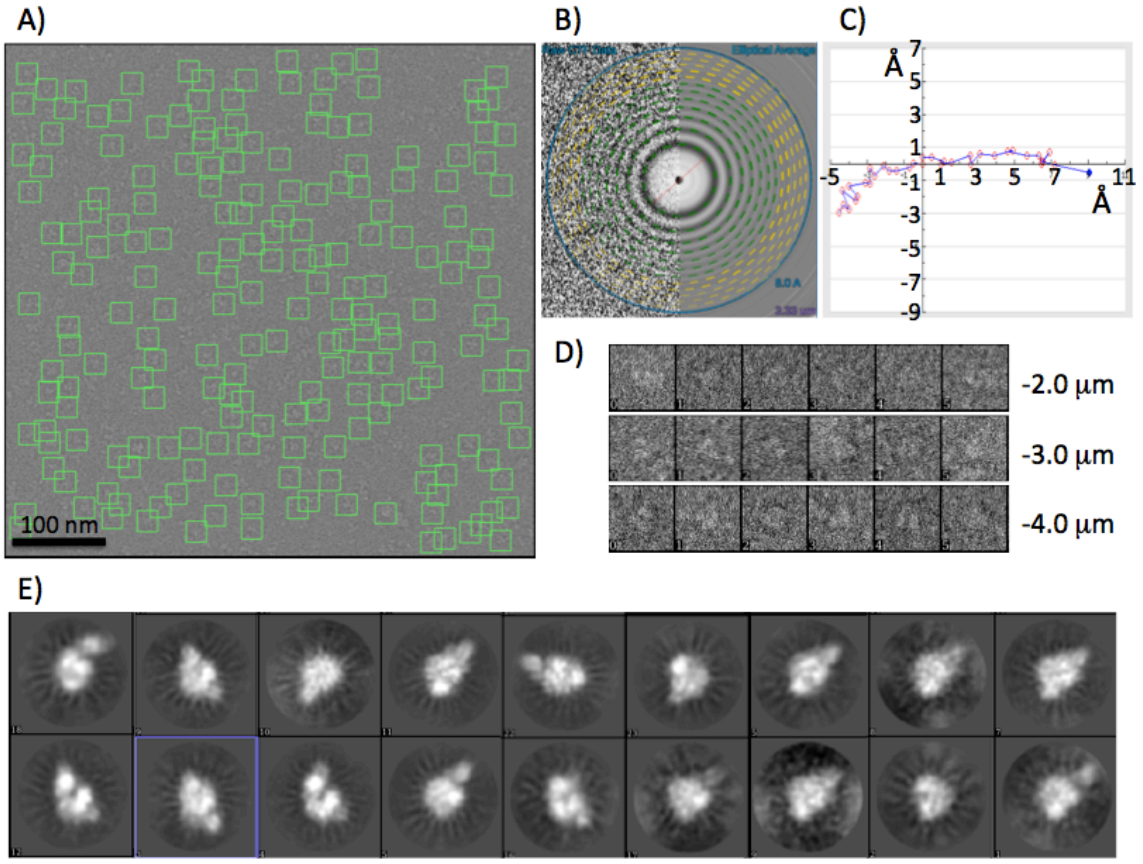


Figure 45: Single particle analysis on cryo-EM data from DmUGGT

Drift-corrected cryo-EM micrograph (A), with contrast inverted, used for semi-automated particle picking. Defocus estimation (B) and drift estimation (C) for the micrograph presented. Representative sampling of cryo-EM particles (D) at various defocus values and reference-free class averages (E) of DmUGGT.

4.3.8 EM map of DmUGGT by cryo-EM

Through three-dimensional maximum-likelihood classification and refinement of the most homogeneous particles sets, I determined a map of DmUGGT at a resolution of 10 Å that provides the most complete and detailed map of UGGT to date (see Figure 47). After cleaning the particle dataset through two-dimensional classification runs, I performed the initial three-dimensional classification with a single class using as initial reference the structure of DmUGGT by negative stain, heavily low-pass filtered to 60 Å.

Through subsequent classification runs with multiple classes, I isolated a subset of ~46,000 particles to be used for refinement within Relion.

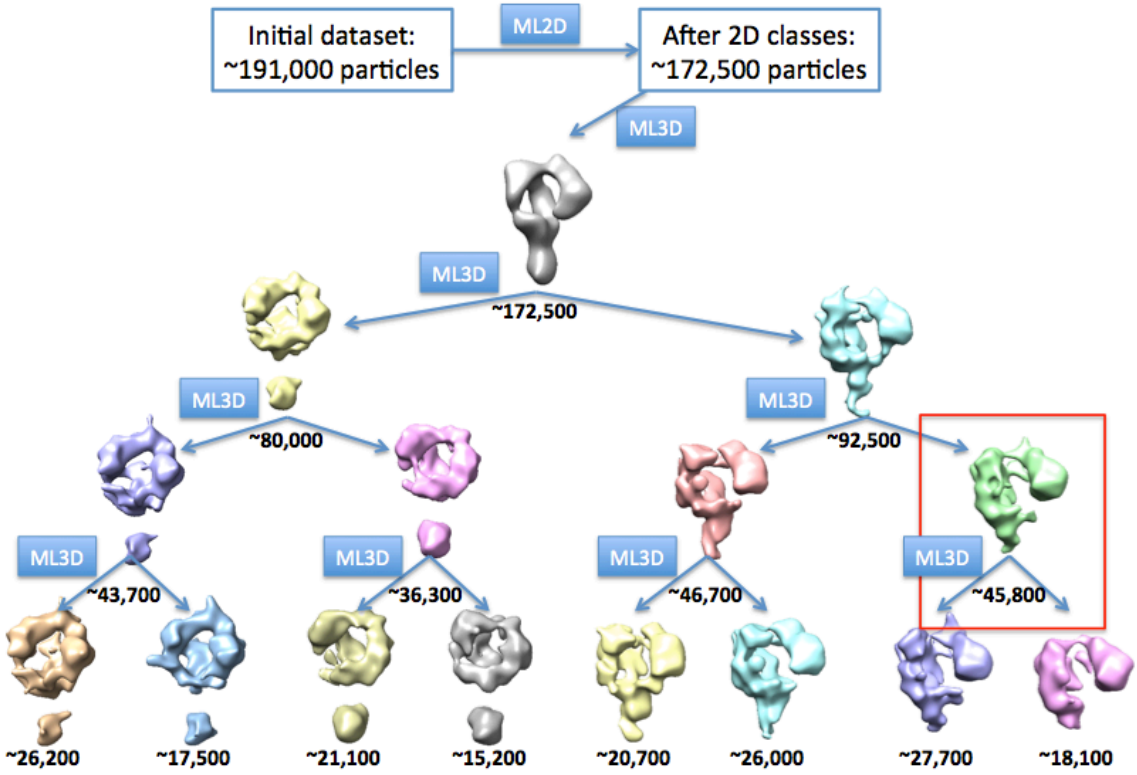


Figure 46: Particle sorting through 2D and 3D classification

A total of ~191,000 particles were picked. After 2D classification, ~172,500 were kept for further processing and used for 3D classification. After each 3D classification, particles were separated into two classes, each of which was used for a new 3D classification run. The green structure marked in a red box was used for refinement, with ~45,800 particles.

Using the best particle set, I refined a structure for DmUGGT at a resolution of 10 Å (see Figure 47, panel C). The structure is about 115 Å long, 105 Å tall and 90 Å wide, with a large central cavity (see Figure 47, panel A). The slices through the model show features inside of the EM map, possibly secondary structure elements such as alpha helices and beta sheets (see Figure 47, panel B). The structure features additional density on the sensor region, attributed to the pre-thioredoxin N-terminal region of UGGT.

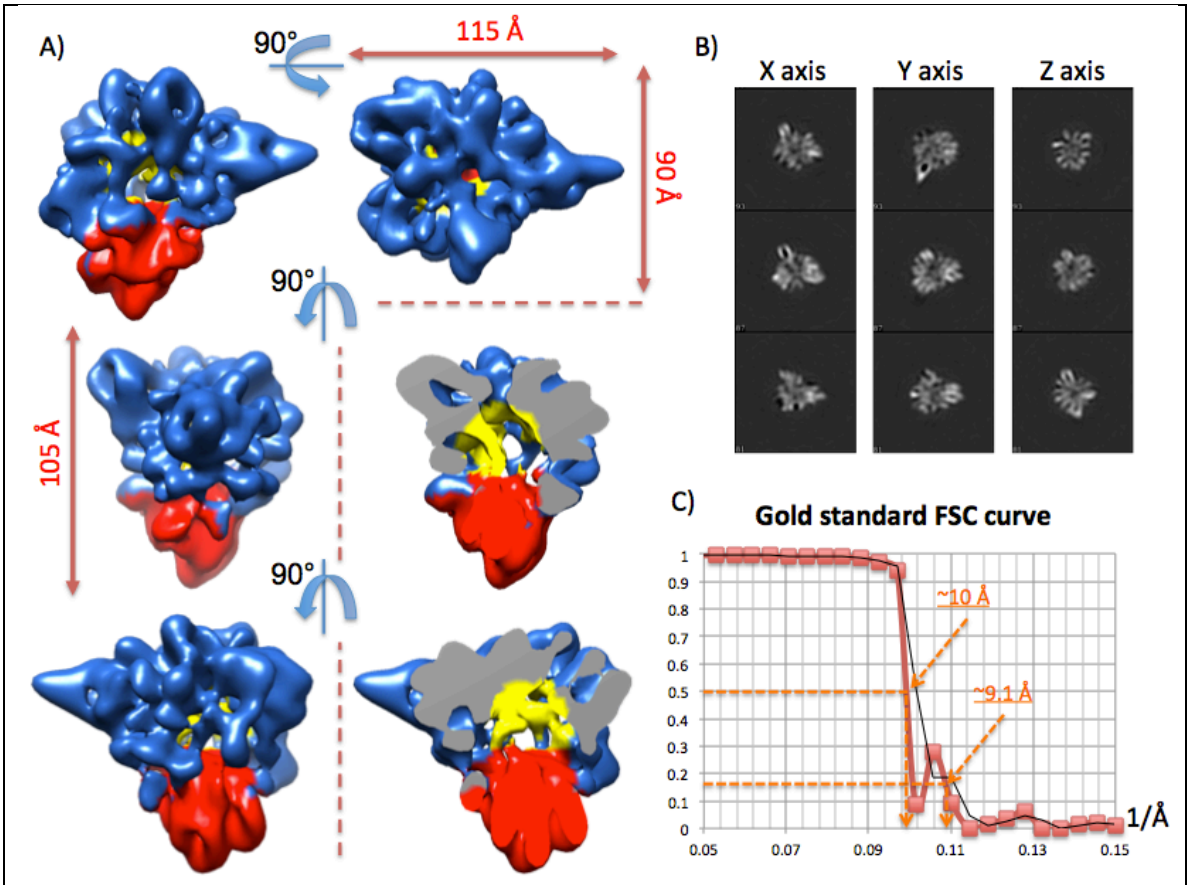


Figure 47: Cryo-EM structure of DmUGGT

Cryo-EM structure refined at 10 Å resolution with ~45,800 particles (A), slices of the EM map through different axis (B) and the Fourier shell correlation for DmUGGT refined structure (C). The catalytic and sensor surface are colored in red and yellow respectively.

In an effort to decrease calculation time and to make the sub-frame particle tracking possible, I downscaled the cryo-EM data by a factor of 2. The particles had a 2.52 A/pixel size and the memory burden was reduced by a factor of 4. I then reclassified the data and refined new DmUGGT structures. A structure with all ~172,500 particles was refined (Figure 48, A), followed by 3D classification to tease apart different populations. Two classes were identified and separately refined to ~10 Å. The most populated class (Figure 48, B) showed a better-defined pre-thioredoxin region and a partially closed entrance to the cavity. The opposite was observed with the second class (Figure 48, C), with a less-defined pre-thioredoxin region and an unobstructed entrance

to the cavity. Though the resolution was not increased, these models displayed fewer artifacts from the reconstruction process than the structure refined using the full-sized particles. Further, these structures suggest at least two conformations of UGGT are present in our cryo-EM dataset, likely contributing to decreased resolution. However, even with the reduced dataset size we could not complete sub-frame particle tracking.

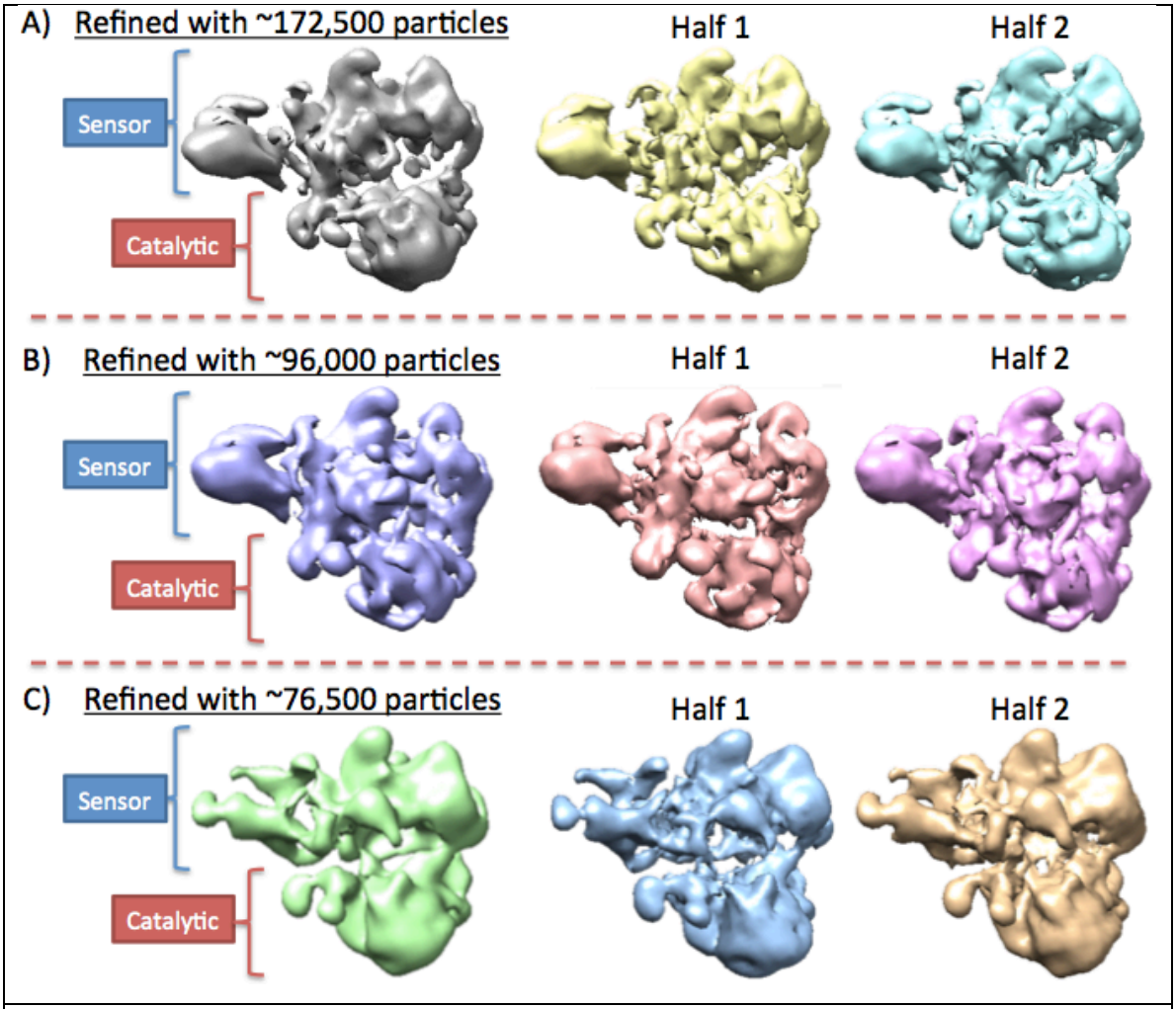


Figure 48: DmUGGT maps with downscaled particles

Structures of DmUGGT: A) refined map using ~172,500 particles, which were then separated into two classes; B) map refined using ~96,000 particles from class 1; and C) map refined using ~76,500 particles from class 2. For each map, the sensor and catalytic domains are indicated and views of the half-maps are shown.

4.3.9 Homology model of DmUGGT domains

Combining the domain identification results using the SA-labeled PcUGGT mutants and the homology models for the domains of DmUGGT, I found the relative position of the domains within the cryo-EM map (see Figure 49, panel B). I used secondary structure prediction (see appendices, from Figure 67 to Figure 72) and previously identified domains (see Figure 8) to calculate homology models for all domains of DmUGGT, including the pre-thioredoxin N-terminal domain, using the Phyre2 modeling server (see Figure 49, panel A). The templates used for modeling had less than 23% sequence identity to UGGT. The homology model for the pre-Trx domain for DmUGGT covered 95% of the sequence, albeit with a 41% confidence score. The models for the three Trx domains and for the catalytic domain were calculated with a +97% confidence score and with +95% sequence coverage. Only 60% of the beta -sheet rich region could be modeled, albeit with a 95% confidence score. A low-pass filtered EM map provided a good fit for all the domains modeled, similar to that of PcUGGT.

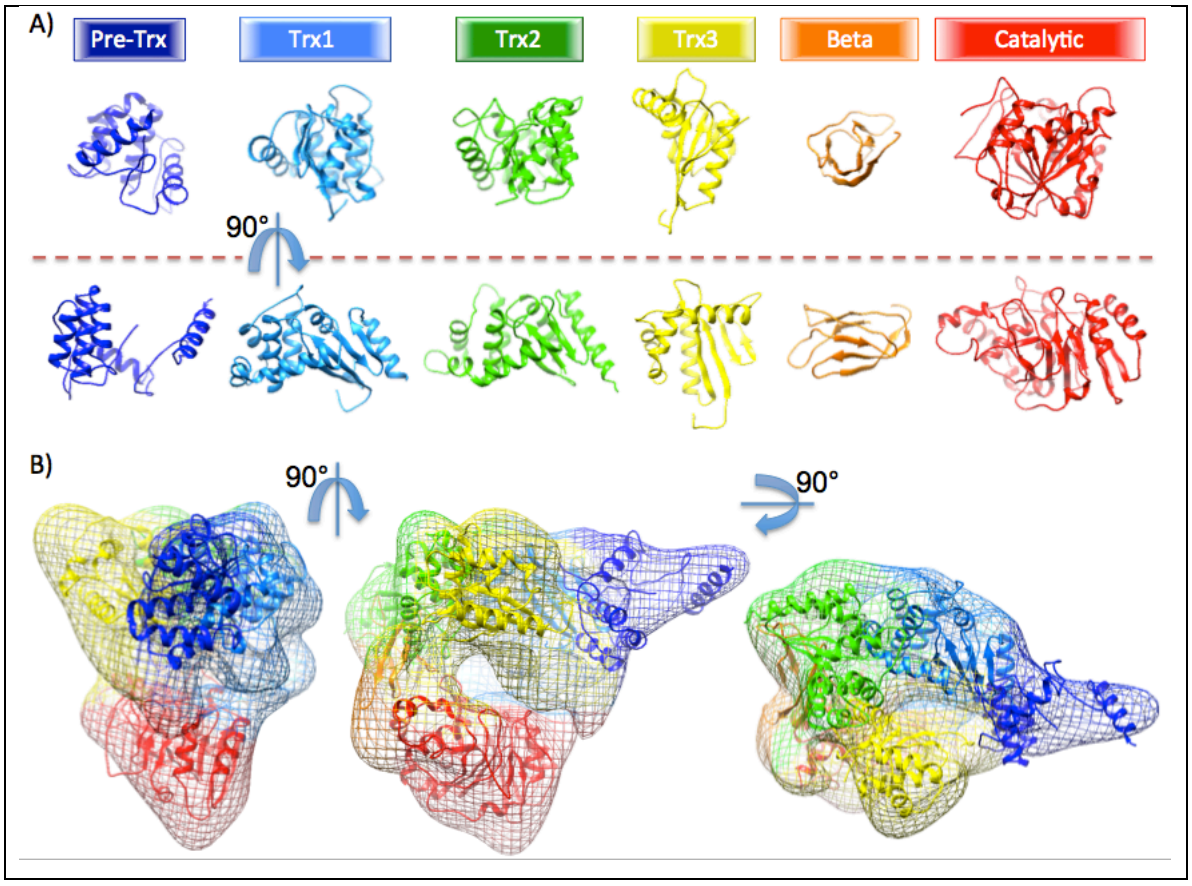
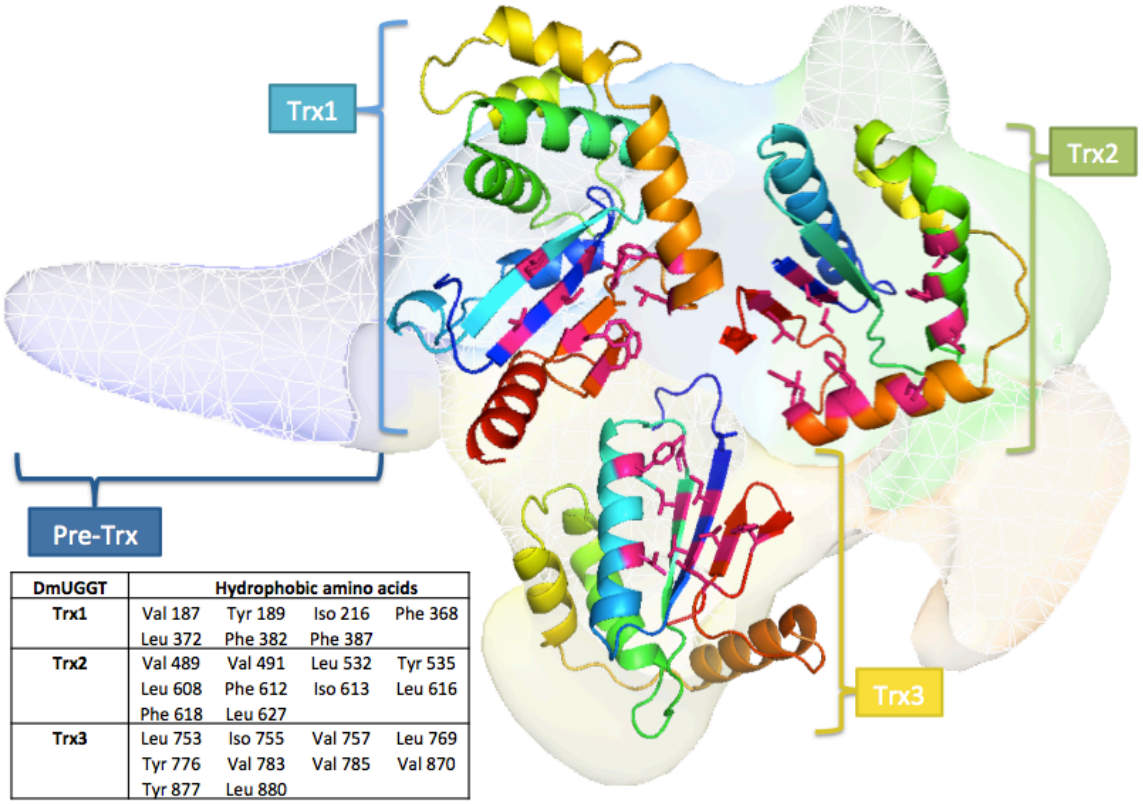


Figure 49: Homology models within DmUGGT cryo-EM map

Homology models of DmUGGT domains (A) and fitting of the models within the cryo-EM structure of DmUGGT (B). The models are color coded as before: Pre-thioredoxin domain in dark blue, Trx1 domain in light blue, Trx2 domain in green, Trx3 domain in yellow, beta-sheet rich domain in orange and catalytic domain in red.

The Trx1-3 homology models of DmUGGT contain numerous hydrophobic residues between the core beta sheets and the surrounding alpha helices. The models within the EM map form a triad of hydrophobic surfaces on the inside of the sensor domain. This arrangement was represented in Figure 50, where the clusters of hydrophobic residues can be seen in magenta. In the schematic figure, the domains were twisted outwards from their original position to better visualize the beta sheets forming the core of the sensor domain. It is important to note that the detailed validation of the Trx1-3 organization remains to be carried out.



DmUGGT	Hydrophobic amino acids			
Trx1	Val 187	Tyr 189	Iso 216	Phe 368
	Leu 372	Phe 382	Phe 387	
Trx2	Val 489	Val 491	Leu 532	Tyr 535
	Leu 608	Phe 612	Iso 613	Leu 616
	Phe 618	Leu 627		
Trx3	Leu 753	Iso 755	Val 757	Leu 769
	Tyr 776	Val 783	Val 785	Val 870
	Tyr 877	Leu 880		

Figure 50: Putative hydrophobic surface formed by Trx1-3 domains

The homology models of the Trx1-3 domains of DmUGGT are presented as ribbon diagrams, with hydrophobic residues highlighted in magenta with visible side chains. The beta sheets of Trx1-3 form a triangular hydrophobic surface for misfolded protein binding. For clarity, the models were spread and tilted outwards from their original position The inside of the cryo-EM map is shown in transparency for context.

4.3.10 Revised substrate selection and glucosylation by UGGT

Based on our structural findings, we revised our UGGT substrate recognition mechanism (see Figure 51), finding the position of the sensor and catalytic domains of UGGT within our EM maps. In our model, the large lobe of UGGT is formed by the thioredoxin-like domains, which are arranged in a ring (see Figure 43 and Figure 49). Each thioredoxin-like domain has a core of beta-sheets, which are oriented towards the large central cavity and could provide the hydrophobic-motif recognition surface (see Figure 50). The beta-sheet rich and catalytic domains form the smaller hook-shaped domain. The catalytic pocket of the glucosyltransferase domain is facing the central cavity and thioredoxin-like domains. This arrangement of the domains, with a large but narrow cavity formed between the sensor and catalytic domains, explains why substrates need to present both N-glycan and hydrophobic moiety within the same molecule and within 40 Å or less. This is also consistent with unfolded glycoproteins being competed out of the central cavity by unfolded non-glycosylated proteins.

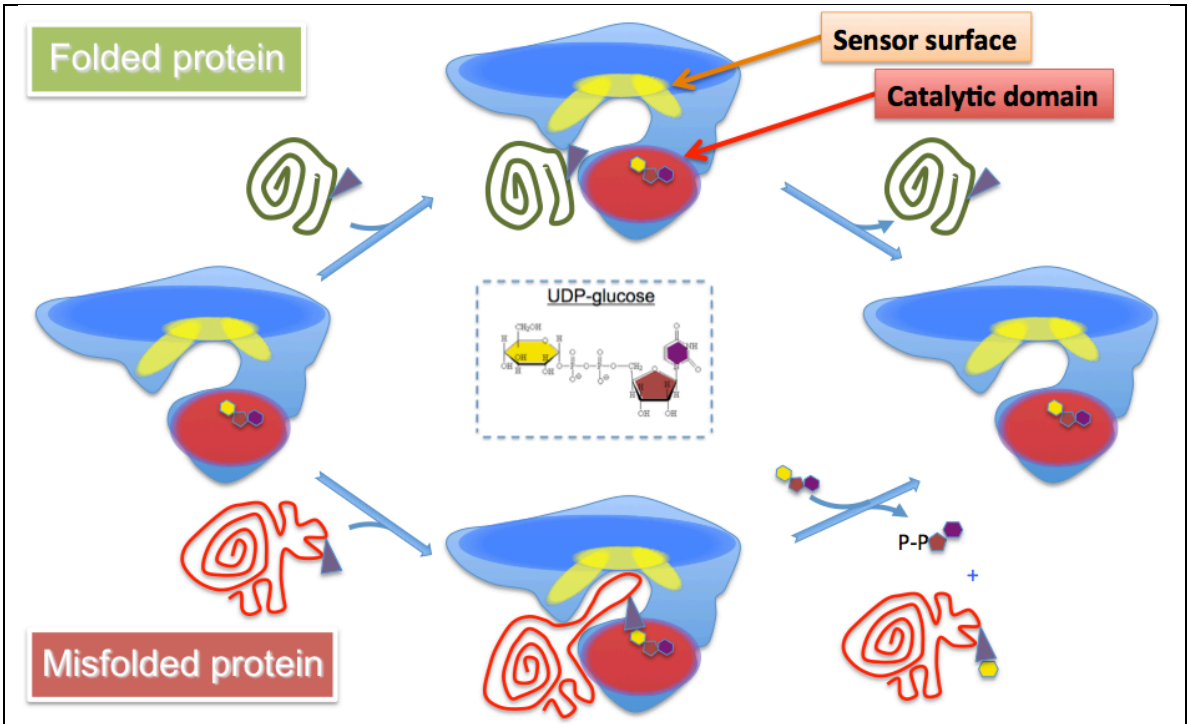


Figure 51: Improved model for substrate selection and glucosylation

The asymmetric structure has a large lobe, formed by the thioredoxin-like domains, and a small lobe, made up of the catalytic domain. The top of the cavity provides the hydrophobic-group biding surface, while the bottom contains the catalytic pocket. The glycan on a native protein can't reach into the catalytic pocket (top). Hydrophobic interactions between the sensor surface of UGGT and hydrophobic loop on the misfolded substrate pull the N-glycan into the cavity, triggering glucosylation (bottom).

4.4 Discussion

Combining negative stain EM of SA-labeled PcUGGT and cryo-EM of DmUGGT, we have clearly identified the position of the domains of UGGT, confirming and enriching our understanding of this peculiar enzyme (see Figure 43 and Figure 49). The maps presented in the previous chapter provided novel structures with strong implications for substrate recognition, but did not provide evidence for the precise domain identification within them. The work in this chapter clearly identified the position of three important domains, which allowed us to infer the location of the remaining domains (see Figure 40 to Figure 42). Furthermore, we modeled the PcUGGT domains

and found these models fit well within the electron density map and agreed with our assignment of their positions.

Our cryo-EM structure provided a complete UGGT structure, in which the homology for every domain of UGGT was positioned (see Figure 49). The resolution we achieved was ~10 Å, at which density for secondary structure elements starts to be visible (see Figure 47). The resolution limitation is a combination of factors, mainly particles suffering from blurriness caused by particle-drift within frames and the difficult task of aligning such small particles accurately. UGGT is one of the smallest asymmetrical proteins studied through cryo-EM single particle analysis and we could not correct for the random drift of particles within frames. Furthermore, we did not have an energy filter nor did we collect data using super resolution. The usage of an energy filter at the time of data collection, or usage of a phase plate, would have helped with the alignment process and might help future studies achieve a higher resolution.

We showed that the thioredoxin-like domains form the large sensor lobe and that the catalytic domain forms the smaller hook-shaped lobe. The secondary structure elements in the cross sections of the cryo-EM map point towards the center of the cavity (see Figure 47, panel B), which we used when fitting the domains within the map. The thioredoxin-like domains have a core of beta-sheets surrounded by near parallel alpha helices. In the crystal structure of the third thioredoxin-like domain (see Figure 8), the lower region of the beta sheets has a hydrophobic surface[56]. Our homology models for the Trx1-3 domain and for the catalytic domain were based on thioredoxin-like proteins and on a galactosyl transferase enzyme, respectively, and we trust they are good approximations of their true structure (see Figure 43 and Figure 49, panels A). For the Trx1-3 domains, we can see from our models that each of them has a cluster of hydrophobic residues on the surface of the core beta-sheet (see Figure 50). We hypothesize that these hydrophobic clusters form a platform where hydrophobic flexible loops can bind (see Figure 50), while remaining shielded from the solvent and other proteins within the large central cavity. The detailed orientation and position of each domain of UGGT still needs experimental validation, and this could be addressed through high-resolution cryo-EM structures and possibly through cross-linking experiments coupled with mass spectrometry.

The structure of DmUGGT positioned the model of the N-terminal pre-thioredoxin-like domain next to the Trx1 domain and within the plane formed by the Trx1-3 domains (see Figure 49). Though this domain has no known functional role attributed, we think it may be necessary in binding to other ER components, such as ERDJ3[21, 32], a J-domain protein that binds HSP family chaperones in the ER. The interaction between UGGT and ERDJ3 has been shown in the literature, showing that complex folding pathways could take place[21]. Thanks to these networks, UGGT could hand-off an unfolded substrate directly to lectin chaperones, such as calnexin or calreticulin, or through the help of ERDJ3 to HSP machinery, such as BiP or GRP94. These interesting networks offer a rich research pathway to follow and explore.

To expand the results of this chapter, several possibilities exist. First, given more microscope time, it would be ideal to collect a cryo-data set of DmUGGT using a higher voltage, combined with a K2 Summit camera and an energy filter, in order to maximize the quality of the data. The improved experimental setup, along with data collection under super-resolution mode, could allow reconstruction of sub-4 Å density maps for UGGT. This would make possible pseudo-atomic resolution model building, as recent publications have shown[60]. To improve crystallization trials, and given our homology models, it is conceivable to produce mutants of the isolated Trx domains with decreased hydrophobicity. As was seen in the Trx3 crystal structure[56], the hydrophobic surface on the Trx domains could be the source of the instability of the domains when isolated from the full-length UGGT. Mutating these surfaces could render the Trx domains more stable and thus improve the chance of obtaining crystal structures.

4.5 Material and Methods

4.5.1 Production and purification of PcUGGT-biotin/SA mutants

In order to produce streptavidin-labeled PcUGGT, we followed the procedure outlined by Lau et al. 2012[89]. We carried out a secondary structure prediction of PcUGGT to identify as many mutation sites as possible to introduce the biotinylation sequence (LNDILEAQ**K**IEWHEQ) across the various domains. A total of 16 sites were identified, distributed across the entire sequence of PcUGGT. The modified AviTag

sequence was introduced to all 16 sites, presumably loops or unstructured linkers. We performed expression trials using *E. coli* using the same conditions as for wild-type PcUGGT. Purification was done through Ni-NTA purification as previously, followed by *in vitro* biotinylation using the biotin ligase enzyme BirA. The successfully biotinylated mutants were then incubated with an excess of monovalent streptavidin and further purified by gel filtration on a Superdex 200 column.

4.5.2 Negative stain EM sample and data collection

Preparation of samples for negative stain EM and data collection were carried out using the same procedures outlined in chapter 3, section 5.2. Briefly, protein concentrations of ~12 ng/µL were used, stained with a solution of uranyl formate at 0.75% (w/v) and imaged at 62,000x magnification on the FEI Tecnai G^2 TF20 operated at 200 kV.

4.5.3 Reconstructing EM maps of PcUGGT-biotin/SA mutants

Similarly to Dm- and PcUGGT processing, we treated the data using Eman2.0[74, 76]. We collected roughly 100 micrographs for each of the five PcUGGT-biotin/SA mutants, from which we picked upwards of 30,000 particles for each one. We used reference-free classification to sort and select the high-quality particles. We utilized the negative stain PcUGGT structure as initial model for refinement of each of the SA-labeled mutants. We attributed the extra density observed in the mutant maps to the 50 kDa SA label. We color coated our maps using Chimera, based on the analysis of the position of the streptavidin labels among the various mutants.

4.5.4 Sample preparation and data collection for cryo-EM

We tested concentrations for Dm- and PcUGGT between 50 to 200 ng/ml and blotting times from 4 to 7 seconds. We lowered the glycerol concentration in the purified UGGT sample to less than 0.1 % and the best sample was obtained at 100 ng/ml and a 6 second blot time, with Quantifoil 1.2/1.3 200 mesh holey-carbon grids. Given UGGT's monomeric nature and 172 kDa molecular weight, it was difficult to find thin ice conditions to visualize individual protein particles. We collected 1500 movie micrographs for DmUGGT using a FEI Tecnai G^2 TF20 at 200kV equipped with K2 Summit direct detection camera, operated through Leginon[95-97] for automated data collection. Each

movie was collected over 7.2 seconds, with 0.2 seconds per frame, with a dose of ~8.8 e/Å^2/second and a 1.26 Å/pixel size.

4.5.5 Single particle analysis and advanced image processing

The global frame alignment for the micrograph movies was done with motioncorr and CTFfind3[79, 98] was used to determine defocus values of the merged frames. We utilized Eman2's e2boxer to pick particles from the 1500 micrographs in a semi-automated fashion, cleaning out the data set from ice contaminants. The particle coordinates for about 191,000 particles were used in Relion[78, 99] to extract and normalize the particles from both the merged micrographs and the movies frames. The 2D classification, 3D classification and refinement were done using Relion (1.3) on the CalculQuebec Guillimin computer cluster. After extensive classification, the cleanest dataset so far contains ~46,800 particles. The particle tracking within the movie-frames could not be completed with the dataset with had available.

4.5.6 Homology modeling of Dm- and PcUGGT domains

The signal sequence for both UGGT species was removed for the modeling process. UniProt tools were used to align the sequences of Dm- and PcUGGT. The sequences for each of the UGGT domains were separated and submitted for homology modeling using the Phyre2 web server[86, 88]. The Pc- and DmUGGT share 21-30%, 41% and 61% sequence identity between their thioredoxin-like domains, the beta-rich region and the catalytic domain, respectively. The domains modeled were as follow: pre-thioredoxin-like domain, thioredoxin-like domain 1, thioredoxin-like domain 2, thioredoxin-like domain 3, beta-sheet rich domain and glucosyltransferase domain. For PcUGGT, the Trx1 domain was modeled based on the structure of a thioredoxin-like oxidoreductase from *Sicilibacter pomeroyi* (3GYK); the Trx2, on the structure of a thiol:disulfide interchange protein from *Acinetobacter baumanii* (4P3Y); the Trx3, on the structure of UGGT Trx3 from *Chaetomium thermophilum* (3WZS); the beta-domain, on part of the structure of carboxypeptidase gp180 from *Lophoretta specularioides* (1H8L); and the catalytic domain, on the structure of a galactosyl transferase from *Neisseria meningitides* (1GA8). For DmUGGT, the pre-thioredoxin domain was modeled based on the structure of a periplasmic thioredoxin-like protein from *Salmonella enterica* (4GXZ); the Trx1, on the

structure of a thioredoxin-like protein from *Corynbacterium diphtheria* (4PWO); the Trx2, on the structure of an oxydoreductase from *Actinomyces oris* (4Z7X); the Trx3, on the structure of UGGT Trx3 from *Chaetomium thermophilum* (3WZS); the beta-domain, on the structure of a human carboxypeptidase (2NSM); and the catalytic domain, on the structure of a galactosyl transferase from *Neisseria meningitides* (1GA8). The modeled structures were fitted into the EM maps using Chimera, following the domain localization from the PcUGGT-biotin/SA mutants.

Preface to HDX-MS on UGGT/Sep15

The best-characterized UGGT binding partner is Sep15. The small selenoprotein Sep15 binds to UGGT very tightly, but the precise role and position of binding to UGGT has not yet been identified. To identify the Sep15-binding region of UGGT, we used hydrogen-deuterium exchange experiments coupled to mass spectrometry on DmUGGT and DmUGGT/Sep15. I put these findings in the context of the full length UGGT using the structural models I determined using electron microscopy and homology modeling.

For the fourth chapter, I produced and purified DmUGGT and DmUGGT/Sep15 samples for characterization through HDX-MS experiments. Dr. Naoto Soya, from Prof. Gergely Lukacs laboratory, prepared peptide digests for MS and HDX-MS experiments using both DmUGGT and the Sep15-bound DmUGGT samples. The HDX-MS data was crucial in identifying where Sep15 binds onto UGGT. In combination with the results from chapters III and IV, I was able to put the HDX-MS results in the context of the full-length UGGT structure. Together, these results place Sep15 on the sensor domain of UGGT, where it would have an ideal position to potentially reduce disulfide bonds on UGGT substrates, particularly during oxidative stress conditions in the ER.

5 Sep15 binding region of UGGT by HDX-MS

Cover figure:

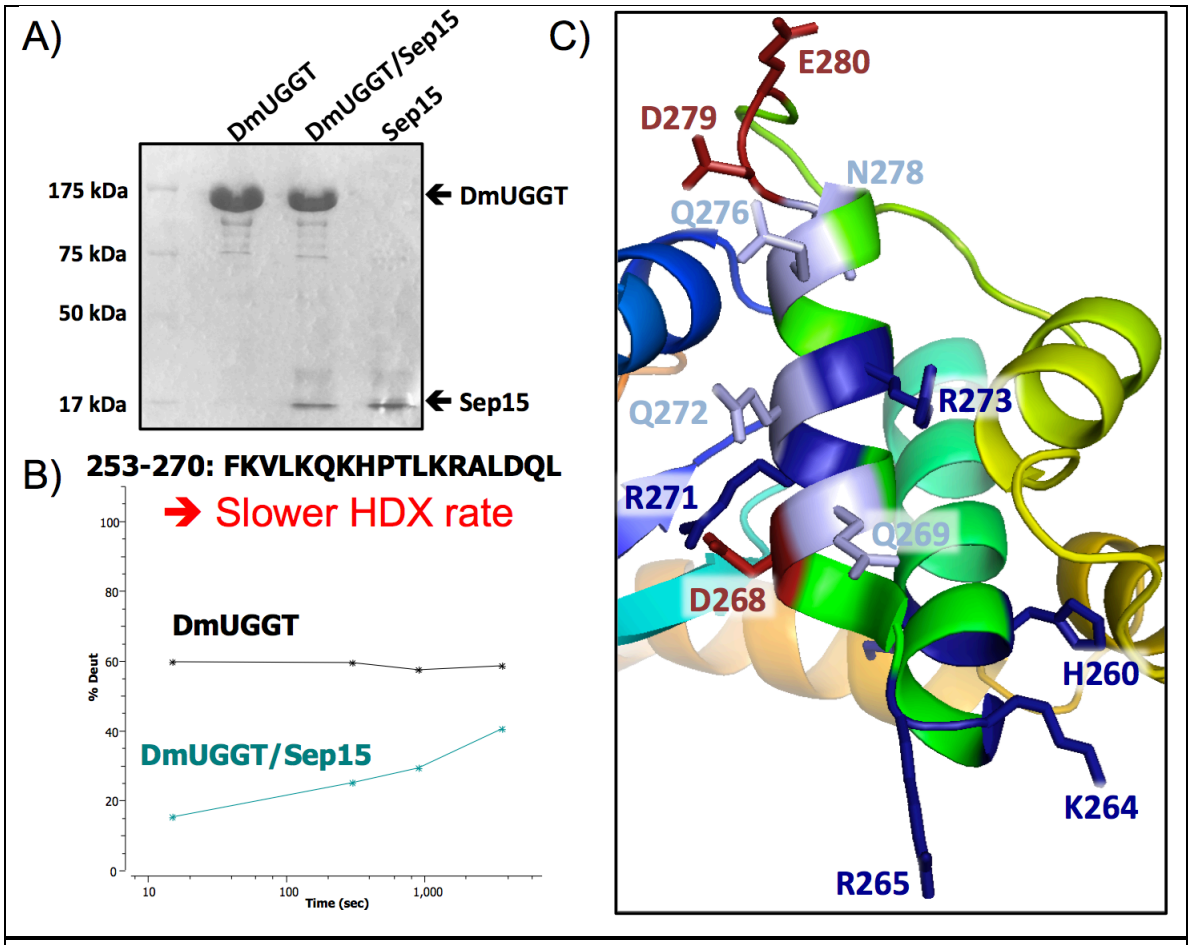


Figure 52: Identification of Sep15-binding region of DmUGGT

SDS-PAGE comparison between purified DmUGGT, DmUGGT/Sep15 and Sep15 samples (A). Hydrogen-deuterium exchange rate curves for the Sep15-binding region of DmUGGT, in the presence and absence of Sep15 (B). Focus on the Sep15-binding helix of the thioredoxin-like domain 1 homology model of DmUGGT (C).

5.1 Abstract

The role of Sep15, a selenoprotein found in complex with UGGT, has not yet been clearly identified though it is required to resolve oxidative stress in the ER. To better characterize the UGGT/Sep15 complex, we carried out hydrogen-deuterium exchange experiments coupled to mass spectrometry (HDX-MS) to map the Sep15 binding site on

UGGT. We identified a short amino acid sequence on the first thioredoxin-like domain of UGGT where Sep15 binds. This small sequence forms a solvent exposed helix, has significantly reduced solvent accessibility in the presence of Sep15 and presents a series of charged amino acids that compliment charged amino acids on the cysteine-rich domain of Sep15, known to mediate the binding to UGGT. Based on our structural models of UGGT, Sep15 binds on the sensor domain at the lip of the central cavity where substrate binding and glucosylation occurs. This could allow Sep15 to reduce disulfide bonds on misfolded UGGT substrates to help cope with heavy oxidative stress.

5.2 Introduction

The 15 kDa selenoprotein, Sep15, is found in the ER of higher eukaryotes and was first identified in humans[31]. The protein lacks an ER retention signal but is always found associated to UGGT, which maintains it in this cellular compartment[27]. This small protein seems to not be essential for cellular viability, though it seems to play a role during oxidative stress in the ER, as it is up-regulated during these events[30, 100]. Sep15 has been shown to have redox activity and to contribute to the redox homeostasis of the ER[28, 100, 101]. The influence of Sep15 on UGGT and the Sep15-binding region of UGGT are not yet clear. It has no major effect on the activity of UGGT1, though it seems to activate and enhance the activity of UGGT2 *in vitro* with synthetic substrates[46, 50]. The interaction between these two proteins is very strong, with a 20 nM affinity, and is mediated through the N-terminal cysteine-rich domain of Sep15[29]. To date, only the structure of the C-terminal half of Sep15 has been determined through NMR, which displays a thioredoxin-like fold[28].

To better characterize the DmUGGT/Sep15 complex we previously purified and characterized, we utilized a combination of HDX and MS. We carried out careful deuterium exchange experiments for DmUGGT alone and for the Sep15-complex, through which we identified the Sep15-binding region of DmUGGT (see Figure 52). With our structural data from EM and with the homology models of DmUGGT domains, we identified the general position of Sep15 in the context of the full structure of UGGT. Docked onto the sensor domain and next to the central cavity of UGGT, Sep15 would be ideally positioned to reduce disulfide bonds on UGGT substrates.

5.3 Results

5.3.1 Purification and characterization of DmUGGT/Sep15

As previously discussed in chapter I, we produced and purified human Sep15 using the *E. coli* expression system and the *Drosophila melanogaster* UGGT using the *Sf9* insect expression system (see Figure 53). We then mixed a sample of purified DmUGGT with a molar excess of Sep15 to form the complex, which was then purified using a gel filtration Superdex 200 column. The calculated molecular weight determined through SEC-MALS was 175 kDa and 196 kDa for DmUGGT and DmUGGT/Sep15, respectively, consistent with a 1:1 binding ratio (see Figure 20). The purified DmUGGT and DmUGGT/Sep15 complex were used for the HDX-MS experiments.

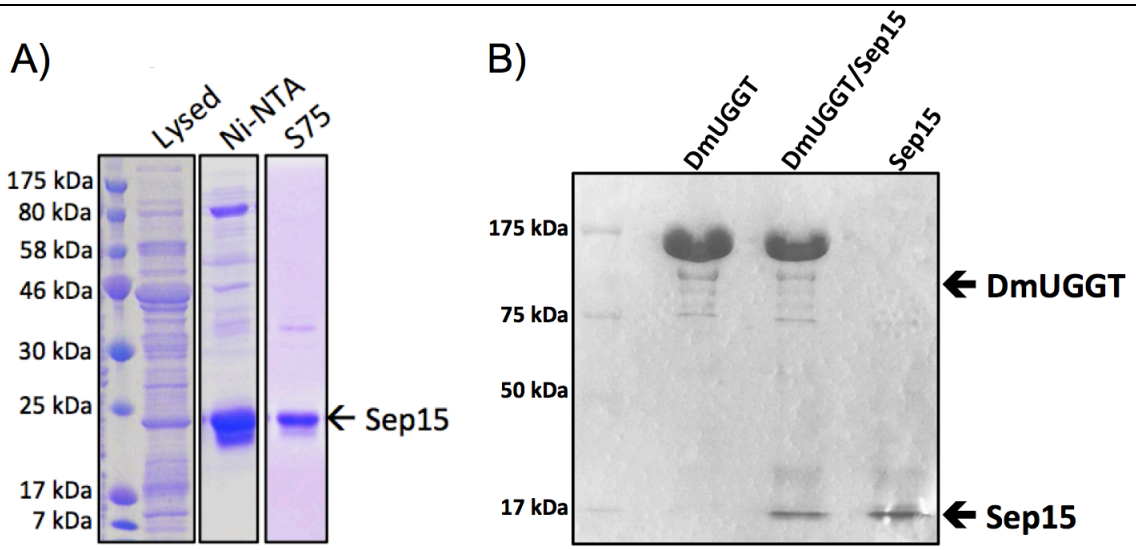


Figure 53: Purification of Sep15 and co-purification of DmUGGT/Sep15

SDS-PAGE analysis of the purification of Sep15 (A) and of the gel-filtration co-purification of the DmUGGT/Sep15 complex (B).

5.3.2 Mass spectrometry and disulfide bonds in DmUGGT

The peptide digestion and MS analysis gave good coverage of the entire sequence of DmUGGT (see appendices, Figure 74), particularly under reducing conditions (see

appendices, Figure 74), which suggests the presence of disulfide bonds. Under non-reducing conditions, there were missing peptides in the pre-thioredoxin domain, between the residues 104 to 166. Carrying out the MS analysis under reducing conditions made it possible to find peptides from the missing region. This indicates the presence of a disulfide bond between C109 and C123 (see Figure 54).

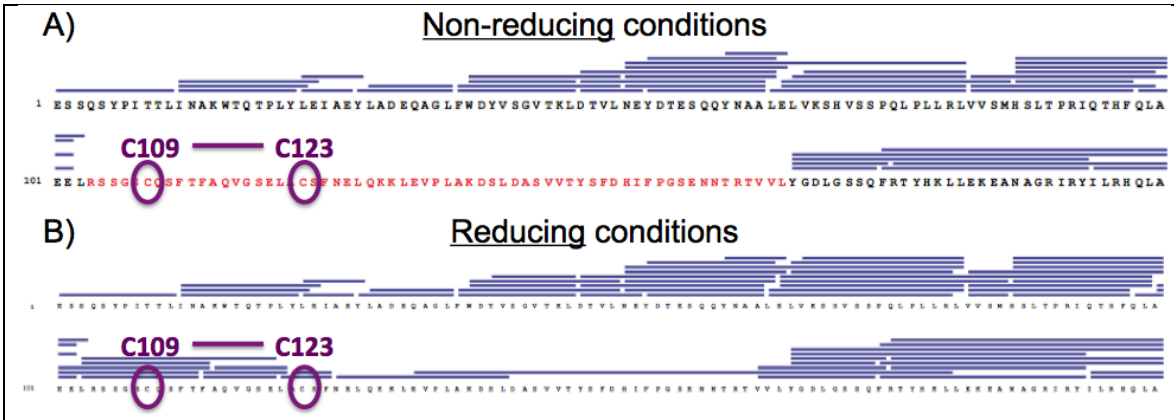


Figure 54: Missing peptides and disulfide bonds in DmUGGT N-terminus

Comparison between the peptide coverage by mass spectrometry for DmUGGT under non-reducing (A) and under reducing conditions (B). For clarity, only the first 200 residues are displayed and the region where no peptides were identified is shown in red.

The homology model for the pre-thioredoxin region of DmUGGT is consistent with the disulfide bond between the C109 and C123. In the model, two cysteine residues are separated by ~10 Å (see Figure 55). Although the distance is long, there are no structural features or residues separating the two residues. This region did not have a strong similarity to other proteins, explaining the difficulty of modeling it. Despite that, the model does support the presence of a disulfide bond between C109 and C123.

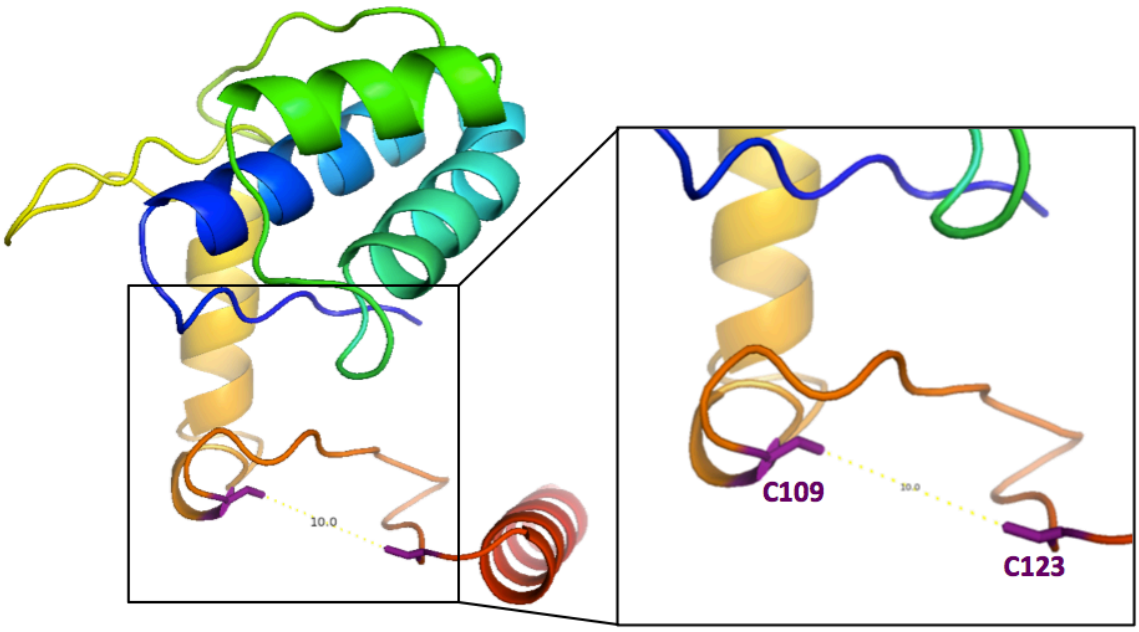


Figure 55: Homology model of DmUGGT N-terminal domain

Homology model of the most N-terminal domain of DmUGGT. The cysteine residues forming the disulfide bonds, C109-C123, are displayed in purple and stick format. The distance between the residues is 10 Å but the space is unobstructed by any residues or structural elements.

The catalytic domain of DmUGGT also displayed a difference in the coverage of peptides detected under non-reducing and reducing conditions (see Figure 56). Under non-reducing conditions, peptides were missing from the residues 1354-1377 and 1449-1488, which contained one and three different cysteine residues respectively. The coverage for both of these regions improved under reducing conditions. As before, this indicates the presence of disulfide bonds between the four cysteines in these two regions.

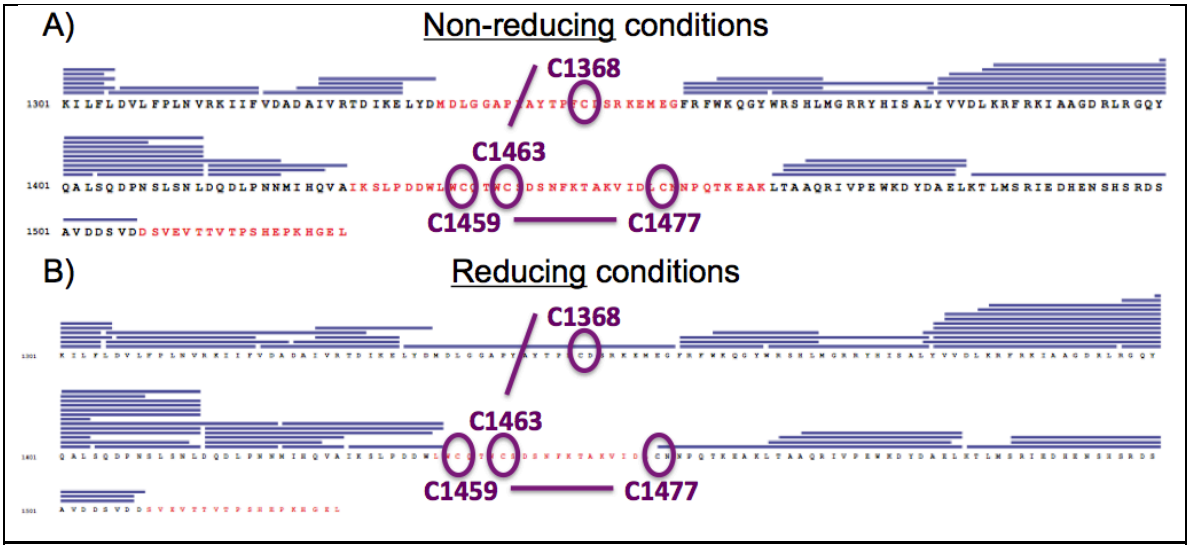


Figure 56: Missing peptides and disulfide bonds in the catalytic domain

Comparison between the peptide coverage by mass spectrometry for DmUGGT under near-native (A) and under reducing conditions (B). For clarity, only the last ~250 residues are displayed and the regions where no peptides were identified are shown in red.

The homology model for the catalytic domain of DmUGGT offered information on which disulfide bonds were forming in this domain (see Figure 57). Analysis of the homology model, which was based on a highly homologous galactosyl transferase enzyme (1GA8), shows that the four cysteines are relatively close to each other. The residues C1368 and C1463 are within 9 Å, while the residues C1459 and C1477 are within 4.6 Å. There are no obstructions between the residues we highlighted. Though some small rearrangement would be required to bring the cysteine residues closer together, disulfide bonds between the C1368-C1463 and C1459-C1477 are in agreement with the observed peptide coverage under reducing and non-reducing conditions.

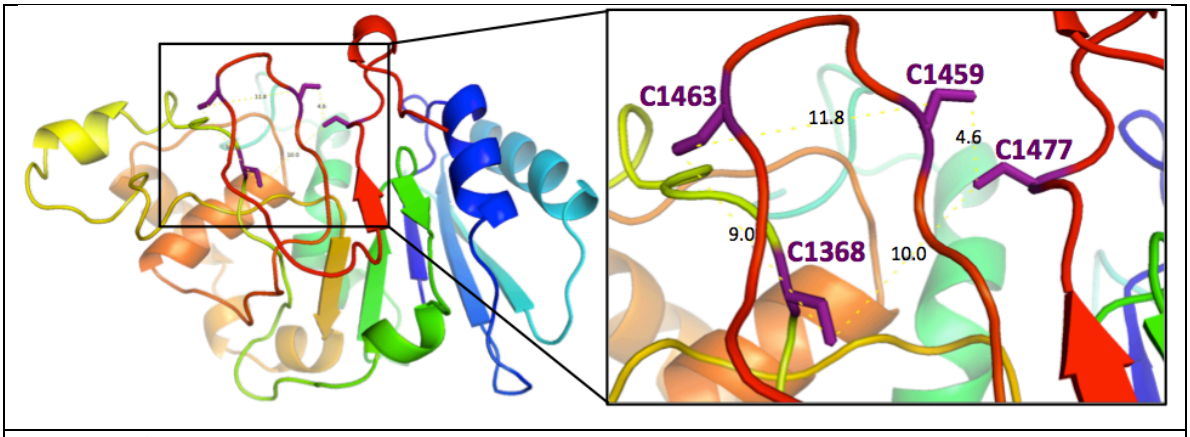


Figure 57: Homology model of DmUGGT catalytic domain

Homology model of the C-terminal catalytic domain of DmUGGT. The cysteine residues forming the disulfide bonds, C1368-C1463 and C1459-C1477, are displayed in purple and stick format. The distance between C1459-C1477 is ~4.6 Å and between C1368-C1463 is ~9 Å, with no structural elements obstructing the residues.

5.3.3 Sep15-binding region of DmUGGT by HDX-MS

The HDX-MS data for DmUGGT and DmUGGT/Sep15 only diverged in one region of DmUGGT, corresponding to a 15 residues sequence with a number of basic amino acids (see Figure 58). The peptide coverage in both cases was ~95%, covering the entire sequence of DmUGGT (see appendices, Figure 75 and Figure 76). Remarkably for a protein of 1500 amino acids, only a region of 15 residues in the first thioredoxin-like domain displayed a significantly reduced HDX rate in the presence of Sep15 (see Figure 58 and appendices, Figure 77). In the absence of Sep15, this region readily exchanged with the solvent. The reduction in HDX rate was of about 40% with the DmUGGT/Sep15 complex, strongly pointing to this region as the Sep15 binding region.

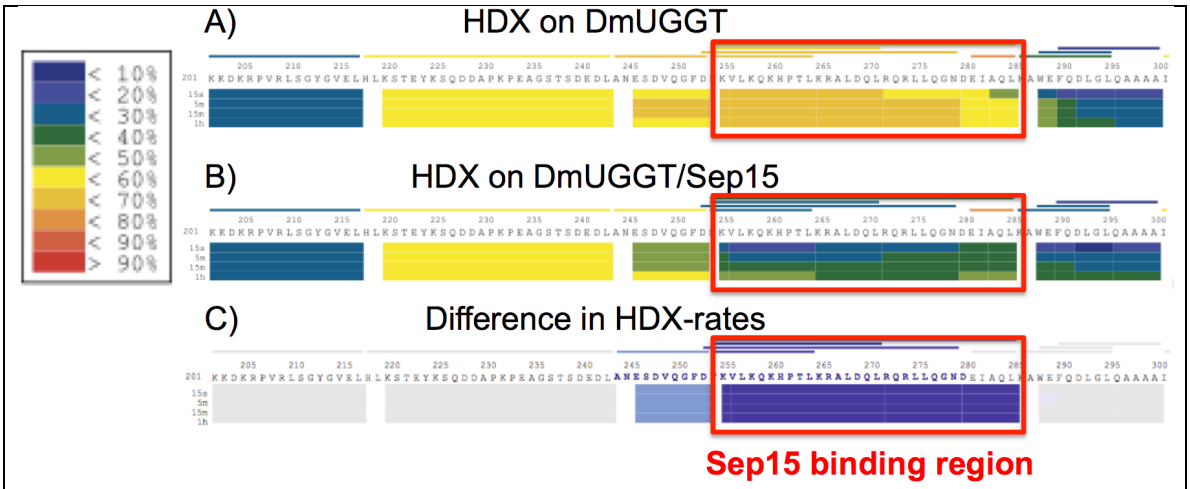


Figure 58: Differences in the HDX-MS results between DmUGGT samples

HDX-MS results for DmUGGT (A) and for the DmUGGT/Sep15 complex (B). The difference in exchange rates (C) between the two samples is displayed. For clarity, only the DmUGGT residues 200-300 are displayed and a red box is shown over the residues where Sep15 binds. Cool colors represent areas with low HDX rates, whereas brighter colors display regions with high HDX rates.

The peptides between residues 252 and 270 systemically displayed a reduced deuterium content as well as significantly slow exchange speed with the solvent when in the presence of Sep15 (see Figure 59). The deuterium content for these peptides from DmUGGT remained constant through time at around 60%, whereas those same peptides displayed only a 20% deuterium content for DmUGGT/Sep15. The deuterium content for these residues of the complex increased very slowly from 20% to a maximum of 50%, over the course of a 60-minute incubation. These results illustrate the lowered availability of these residues to the solvent, caused by a shielding effect upon Sep15 binding through its C-terminal cysteine-rich domain to DmUGGT.

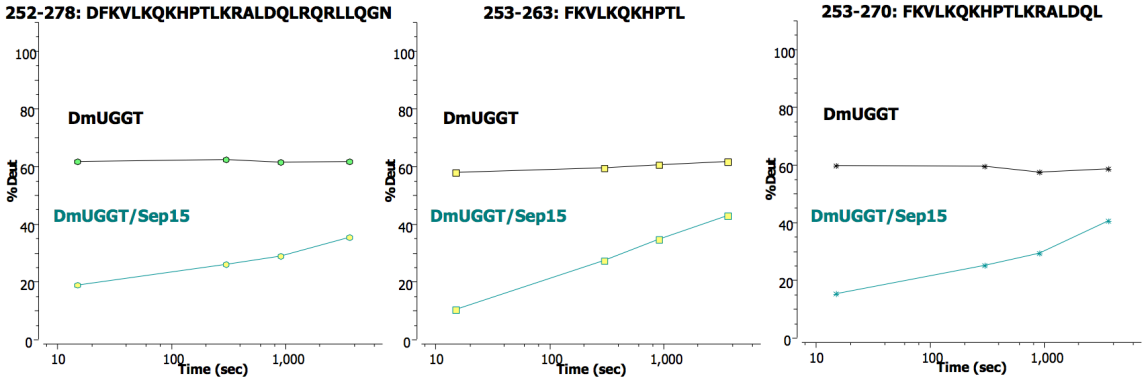


Figure 59: Reduced HDX-rate for the Sep15 binding peptides on DmUGGT

Comparison between HDX rates for the peptides involved in Sep15-binding. Black curves show the HDX-MS data for DmUGGT alone, whereas the green curves display the data for the DmUGGT/Sep15 complex. Peptides in this region only displayed a significant 40% reduction in HDX rates in the presence of Sep15.

The Sep15-binding region of DmUGGT forms a heavily charged helix, that in the homology model of the first thioredoxin-like domain is solvent exposed (see Figure 60). The core sequence contains lysine (K264), arginines (R265, R271, R273) and glutamines (Q269, Q272, Q276, Q278) residues, which are positively charged or polar, as well as an acidic aspartate (D268) residue in the middle of the sequence. Though there is no structure for the cysteine-rich domain of Sep15, its cysteine-rich domain contains an acidic sequence, D(-)PD(-)CR(+)GCCQE(-)E(-), with complimentary opposing charges to those observed in the core Sep15-binding sequence of DmUGGT, K(+)R(+)ALD(-)QLR(+)QR(+). The strong charge complementarity, in conjunction with interlocking tridimensional structures of these domains, could explain the 20 nm affinity between these two proteins[29]. Based on the position of Sep15 binding on DmUGGT and on the position of this domain within the EM map, Sep15 would be positioned near the lip of the central cavity of DmUGGT.

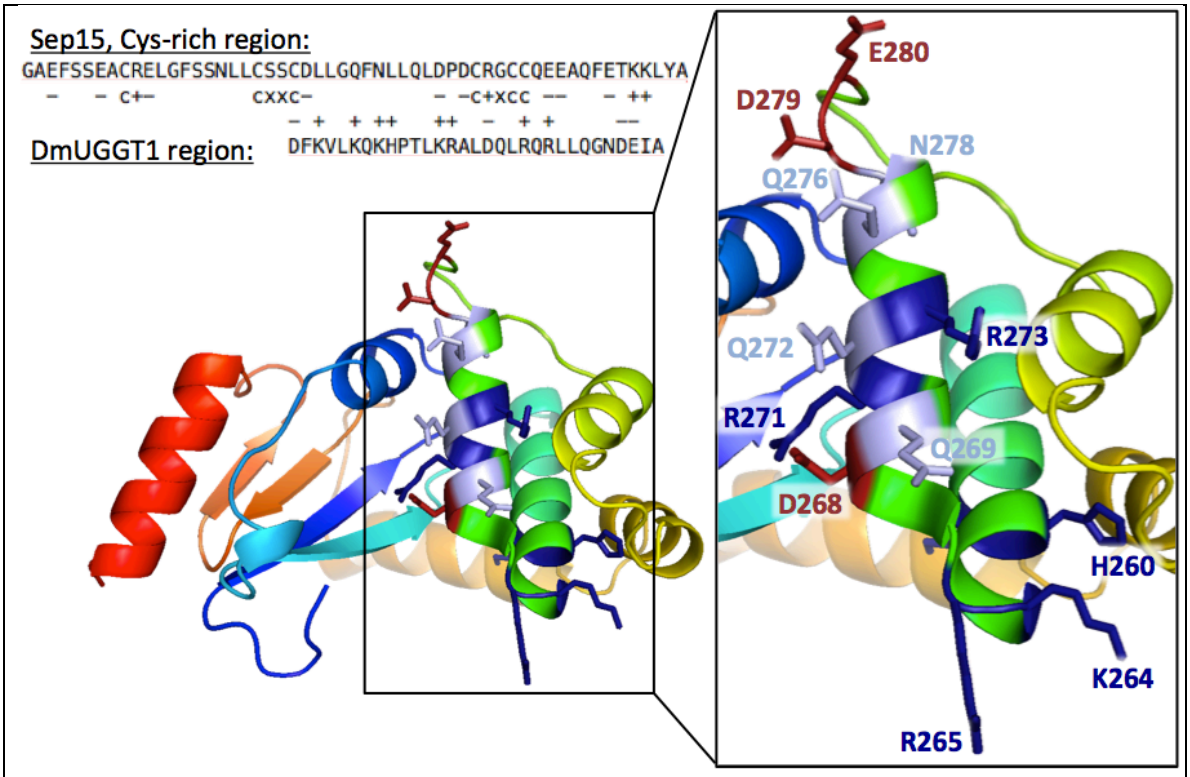


Figure 60: Details of the Sep15-binding helix from DmUGGT Trx1 domain

Homology model of the C-terminal catalytic domain of DmUGGT. The sequence of the cysteine-rich domain of Sep15 is shown and compared to the Sep15-binding region of DmUGGT, with positive (+) and negative (-) charges. The Sep15-binding residues form a heavily charged helix, with charged or polar side-chains displayed as sticks. Side-chains are color coded according to their charge or polarity: negatively charged in red, positively charged in blue and grey for positive polarity.

5.4 Discussion

We first utilized MS under reducing and non-reducing conditions DmUGGT (see Figure 54 and Figure 56), finding disulfide bonds in the N-terminal pre-thioredoxin domain (see Figure 55) and in the catalytic domain (see Figure 57). No other modifications were found, consistent with our construct for DmUGGT in which the N-glycosylation sites were removed. Though inter-domain disulfides could have stabilized the structure of UGGT, only intra-domain disulfide bonds were identified. The quality of our homology models was high enough to agree with the disulfide bonds we identified.

More importantly, we identified the core sequence of DmUGGT that binds to Sep15 cysteine-rich domain (see Figure 52). The binding is isolated to a ~15 amino sequence, which is basic on DmUGGT and acidic on Sep15, with a complementary distribution of charges on either side of the binding surface (see Figure 60). Though there is no current crystal structure for the UGGT/Sep15 or for the full-length Sep15, we think the 20 nM affinity interaction[27, 28, 31] between these two proteins is a product of both shape and charge complementary. The charges on DmUGGT are conserved across higher eukaryotes[31], but are less conserved in PcUGGT, which lacks the Sep15 gene. This agrees with the observation that Sep15 is only necessary to help cope with high levels of oxidative stress[30, 100, 101].

With our structural models of DmUGGT and the position of the domains, Sep15 would tightly bind to UGGT sensor domain, next to the entrance of the large central cavity, where the misfolded glycoprotein binds and is glucosylated (see Figure 35 and Figure 51). Under normal circumstances, PDI and ERp57 are enough to break or correct improperly folded disulfide bonds in folding glycoproteins. We hypothesize that under oxidative stress, where PDI and ERp57 are no longer enough to deal with too many proteins with abnormal disulfide bonds, Sep15 in conjunction with UGGT might help by reducing disulfide bonds on UGGT substrates. This makes Sep15 a non-essential component under normal conditions[101], consistent with its absence in lower eukaryotes, but an additional safety mechanism evolved in multi-cellular eukaryotes to help resolve intense oxidative stress conditions.

For future direction, it would be important to biochemically verify that the interaction is indeed taking part between the Trx1 domain of UGGT and the cysteine-rich domain of Sep15. For this, the amino acids in the helix of the Trx1 domain could be mutated to non-polar residues for example, to see if the binding to Sep15 is lost, after disrupting the predicted binding surface on UGGT. Once mutated, the interaction could be tested by pull-down assays using a Sep15-GST fusion protein and glutathione Sepharose beads. SDS-PAGE analysis of the pull-down assays could easily reveal whether or not the binding is lost after Trx1 is mutated. This method would require low quantities of protein and could possibly be done with cellular lysates.

5.5 Material and Methods

5.5.1 Production of Sep15 and DmUGGT

As described in chapter I, DmUGGT was produced using the *Sf9* expression system. Briefly, the *Drosophila melanogaster* UGGT was cloned in the pFastBac plasmid and a baculovirus containing the DmUGGT gene was optimized using *Sf9* insect cells. The protein was purified through a combination of Ni-NTA affinity and Superdex 200 gel filtration chromatography. The gene for human Sep15 was cloned into the pET29a vector for production using the *E. coli* expression system, using the Rosetta Gami 2 strain. After production, the protein was purified using a combination of Ni-NTA affinity, anion exchange and Superdex 75 gel filtration chromatography.

5.5.2 Co-purification of UGGT with Sep15

The DmUGGT/Sep15 copurification was detailed in chapter I. Briefly, purified samples of DmUGGT were mixed with a molar excess of purified Sep15 and incubated briefly at 4°C. Care was taken to have mixture volumes of less than 4 ml, to inject directly after incubation in a Superdex 200 gel filtration column, equilibrated with a 30 mM Tris-HCl, pH 7.5, 300 mM NaCl, 3% glycerol buffer. After analysis of the peak fractions, the DmUGGT/Sep15 peak fractions were pooled and concentrated to 2 mg/ml and stored at -80°C.

5.5.3 Hydrogen-Deuterium Exchange and Mass Spectrometry

HDX experiments were carried out similar to those previously described[102]. Briefly, HDX was initiated by diluting stock DmUGGT solution 1.5:8.5 into D_2O-based buffer (30 mM Tris, 300 mM NaCl, 3% Glycerol, pD 7.5). HDX incubation periods were 15 sec, 5 min, 15 min and 1 hour, and temperature was set at 25 °C. HDX was quenched with chilled quenching buffer (300 mM glycine, 6 M Gdn-HCl and 400 mM TCEP in H_2O, pH 2.5) using a 1:1 dilution ration. Quenched samples were flash frozen in methanol containing dry ice, and frozen solutions were stored at -80 °C until used.

Prior to UHPLC-MS analysis, deuterated DmUGGT was digested in an on-line immobilized pepsin column, prepared in-house. Resulting peptides were loaded onto a C_{18} analytical column (1 mm i.d. × 50 mm, Thermo Fisher Scientific) equipped to Agilent 1290 UHPLC system. Peptides were separated using a 5-40% liner gradient of

acetonitrile containing 0.1% formic acid for 10 min at 65 μL/min flow rate. To minimize back-exchange, the column, solvent delivery lines, injector and other accessories were placed in an ice bath. The C_{18} column was directly connected to the electrospray ionization source of LTQ Orbitrap XL (Thermo Fisher Scientific), and mass spectra of peptides were acquired in positive-ion mode for m/z 200-2000. The deuteration (%) as a function of incubation time was determined using HDExaminer 2.1 (Sierra Analytics, Modesto, CA).

6 Conclusions about UGGT

The UGGT enzyme has been the focus of many studies, most of them through biochemical assays using various glycoprotein substrates or elaborate synthetic substrates carrying different hydrophobic groups and glycan chains[39, 50]. Through the biochemical work, it has been established UGGT recognizes exposed hydrophobic loops and a $Man_9GlcNAc_2$ glycan on the substrate (see Figure 1), which need to be within 40 Å of each other to trigger glucosylation[48]. The glucosylation is selective for misfolded glycoproteins (see Figure 51), not affecting well-folded glycoproteins[22, 34, 49]. The affinity between UGGT and its substrate is low and once glucosylation occurs, the substrate is released. This intricate mechanism is unique to this enzyme, whose sequence shows small similarity to other proteins. Structural data is key to understand the mechanism of action of UGGT, though the success on this front has been extremely limited.

6.1 UGGT characterization in solution

Though our crystallization efforts were fruitless, our biophysical analysis of UGGT in solution established that the purified enzyme was highly stable, behaved consistently as a monomer and was catalytically active (see Figure 13). Purified UGGT from both species was able to bind to Sep15, the best-characterized UGGT binding partner[27-29, 31], in a one-to-one binding ratio (see Figure 20). Most importantly, our SAXS experiments point to a complex three-dimensional arrangement of UGGT domains (see Figure 24). The analysis of the scattering data shows UGGT has a non-compact multi-domain organization[25, 52]. Though we can't completely rule out the possibility of large flexibility between domains, our results are more consistent with subtle small-scale domain movement, which might be necessary for substrate recognition and glucosylation, explaining the difficulties encountered in crystallization trials of UGGT.

6.2 Novel EM and SAXS structural models for UGGT

To characterize the UGGT structure, I turned to negative stain EM and single particle reconstruction techniques to successfully reconstruct maps for two UGGT species (see Figure 25). Though low-resolution, the EM maps revealed a conserved claw-shaped

architecture, with a large lobe and a smaller lobe forming a wide central cavity. I propose that this central cavity houses the misfolded protein sensor surface as well as the catalytic pocket for glucosylation. I used SAXS to confirm and validate our EM findings, successfully finding *ab initio* models that present the same topology observed in our EM data (see Figure 33). The claw-shape with the asymmetric lobes and a large central cavity is thus conserved across species and across techniques. Our structure-based activity model explains the vast biochemical data available for UGGT and provides a unique mode of action for unfolded protein binding, labeling and release (see Figure 34).

6.3 Domain identification and medium-resolution cryo-EM

To improve upon our previous findings, I used a combination of EM techniques and homology modeling to find the position of every UGGT domain in our maps, and made steps toward a higher-resolution map for UGGT through cryo-EM (see Figure 35). Using domain-specific SA-labeling of PcUGGT mutants and negative stain EM, I determined three maps that allowed for domain identification (see Figure 43). I was able to create homology models for the assigned domains, and to fit them within the electron density maps. The cryo-EM structure calculated appears to be a complete structure for DmUGGT, with an additional density for the N-terminal pre-thioredoxin-like domain that was missing in the negative stain map. The cryo-EM map was refined to a resolution of 10 Å and slices of the maps suggest the presence of secondary structure elements (see Figure 47). My current efforts focus on further sharpening the images in order to boost the resolution of the structure. The cryo-EM map of DmUGGT allowed the homology models of its domains to be fitted into the electron density (see Figure 49). Our results validated and enriched our model for selective recognition and glucosylation of misfolded glycoproteins (see Figure 51).

6.4 Sep15 binding to UGGT, localization and implications

To pinpoint the binding location of Sep15 on UGGT, Dr. Naoto Soya carefully carried out hydrogen-deuterium exchange experiments with samples I provided of UGGT alone and of UGGT in complex with Sep15. These experiments revealed that a

small putative helix on the thioredoxin-like domain 1 of UGGT displays significantly reduced solvent accessibility in the presence of Sep15 (see Figure 52). In the context of our cryo-EM map and homology models, these results suggest that Sep15 binds to the sensor region on the lip of the large central cavity of UGGT (see Figure 35 and Figure 51). At this location, Sep15 could act by reducing disulfide bonds on misfolded substrates[28, 29], particularly under ER oxidative stress conditions[30, 100]. This requires validation through biochemical assays, potentially through using misfolded synthetic substrates with a disulfide bonds.

6.5 Structural implications for the regulation of UGGT regulation

The asymmetric claw of our models reconciles twenty-five years of research into UGGT activity. The proposed mechanism (see Figure 51) is capable of distinguishing between a folded and an unfolded glycoprotein. The mechanism is unique to UGGT and different from the mechanism of HSP family chaperones[2, 7] or lectin chaperones[3, 12, 13] (see Figure 3 and Figure 5). The structure and the position of the domains within the EM maps provide a large central cavity, about 40 Å wide inside and with a narrow 15 Å wide entrance, enough to allow N-glycans on flexible hydrophobic loops to enter but narrow and deep enough to exclude N-glycans on folded proteins (see Figure 43 and Figure 49). On one side, the inside of the central cavity is formed by the thioredoxin-like domains, which provide a hydrophobic surface to bind exposed hydrophobic loops from the substrate (see Figure 50). The glucosyltransferase domain forms the opposite side of the central cavity, with the catalytic domain facing the thioredoxin-like domains (see Figure 51). This arrangement allows flexible loops with exposed hydrophobic residues on the substrate to be pulled into the central cavity, dragging along the N-glycan, which is thus positioned next to the catalytic pocket and primed for glucosylation. Once glucose transfer occurs, the glucosylated misfolded protein is released to interact with the lectin chaperones for proper folding.

6.6 Future directions

The structure of UGGT has evaded numerous attempts to study it at high resolution using protein crystallography and x-ray diffraction, but in this study we succeeded in determining various structures at low and medium resolution using SAXS and EM, as well as homology modeling. Though we made great progress in understanding the mechanism of substrate selection by UGGT, many improvements can be made and various directions for future work have opened up.

The relative position of the thioredoxin-like domains could be improved through usage of cross-linking agents of 5-10 Å with the full length UGGT. Doing short-length cross-links between the UGGT domains, followed by unfolding and limited proteolysis and analysis by mass spectrometry could help identify peptides that are close in space. A list of neighboring peptides could be established and used to better model the assembly of the individual domains within the medium resolution EM maps. This approach would require careful planning and would benefit from collaboration with a group with experience using this technique.

Determining a near atomic resolution structure of UGGT by cryo-EM is possible and the feasibility is much higher than trying to crystallize UGGT and carry out x-ray diffraction experiments. The approach would follow our general protocols, relying on freshly purified UGGT samples and collecting data on direct detection cameras. The improvement could be made using a more stable microscope with a higher operating voltage, as a Titan Krios at 300 kV, combined with a K2 Summit camera for detection as it shows the best performance for sub-megadalton proteins and using an energy filter during data collection, as this improves the overall quality of the collected movie-micrographs. Finally, sub-frame particle tracking on the highest quality dataset would unlock near atomic resolution structures, to build pseudo-atomic resolution models.

Combining high-resolution cryo-EM with purified UGGT in complex with a stable molten globule glycoprotein, it could be possible to confirm the substrate binding mechanism, including the functions of the sensor domain and that of the glucosyltransferase domain, which we have discussed in this thesis.

7 Appendices

7.1 Secondary structure of PcUGGT domains

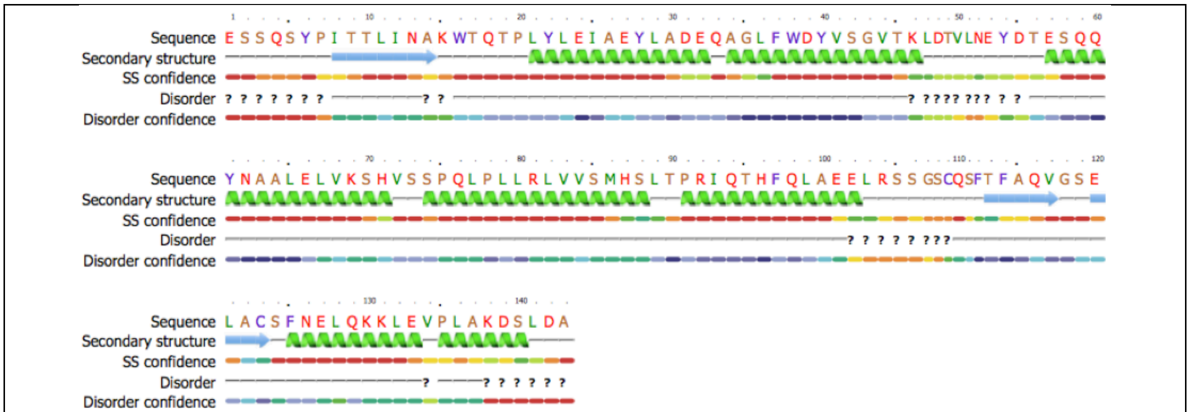


Figure 61: Secondary structure prediction of PcUGGT Pre-Trx domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

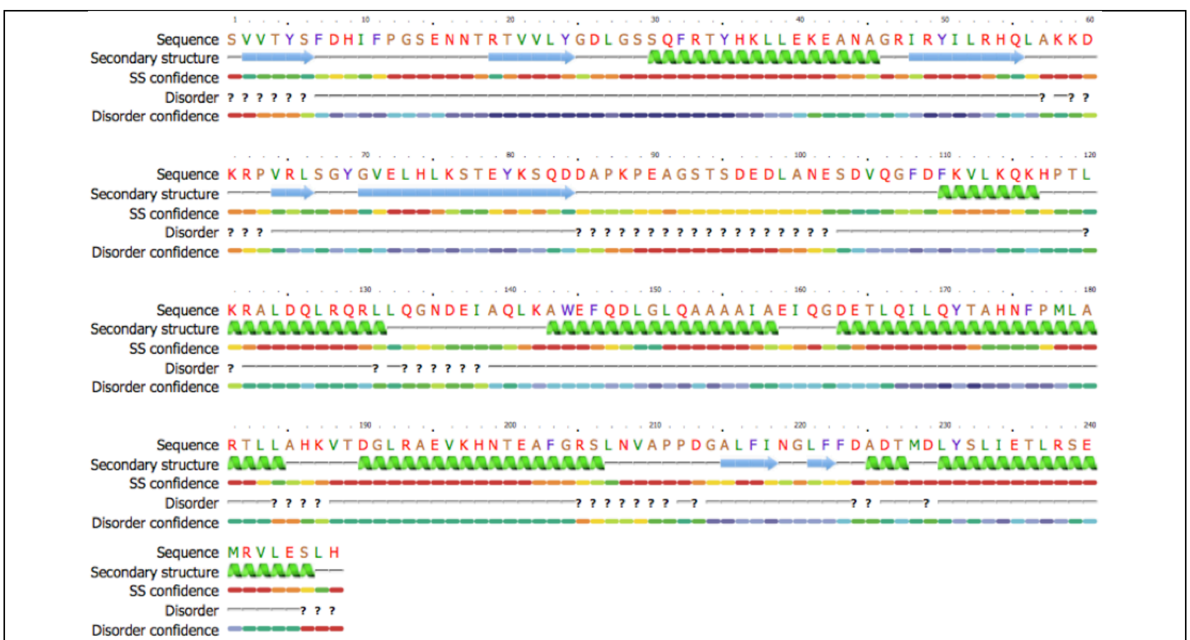


Figure 62: Secondary structure prediction of PcUGGT Trx1 domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

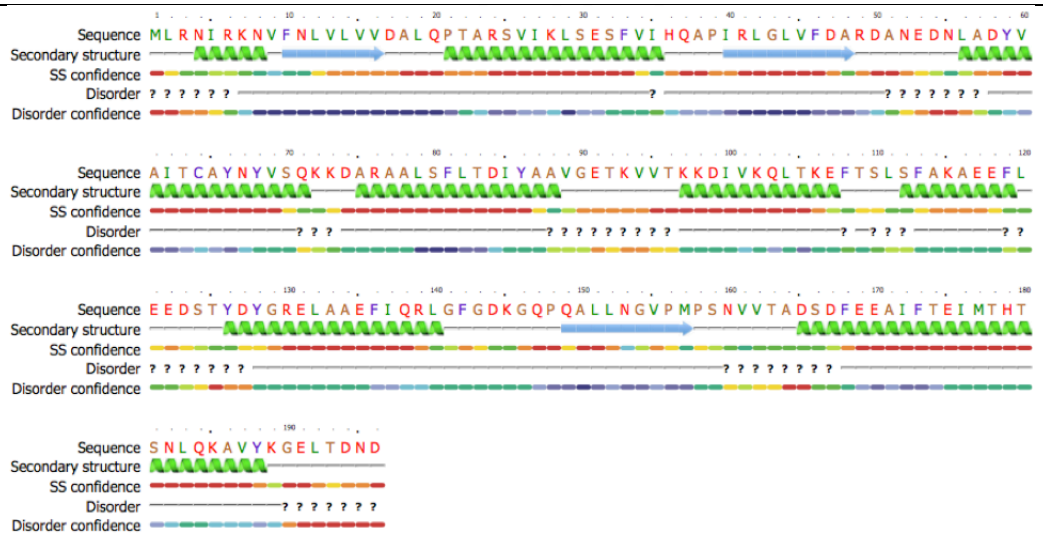


Figure 63: Secondary structure prediction of PcUGGT Trx2 domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

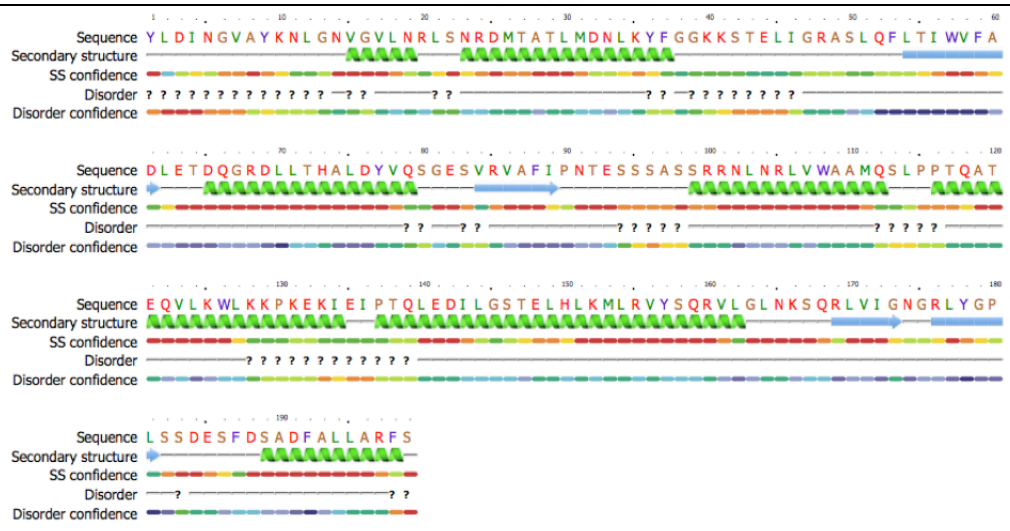


Figure 64: Secondary structure prediction of PcUGGT Trx3 domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

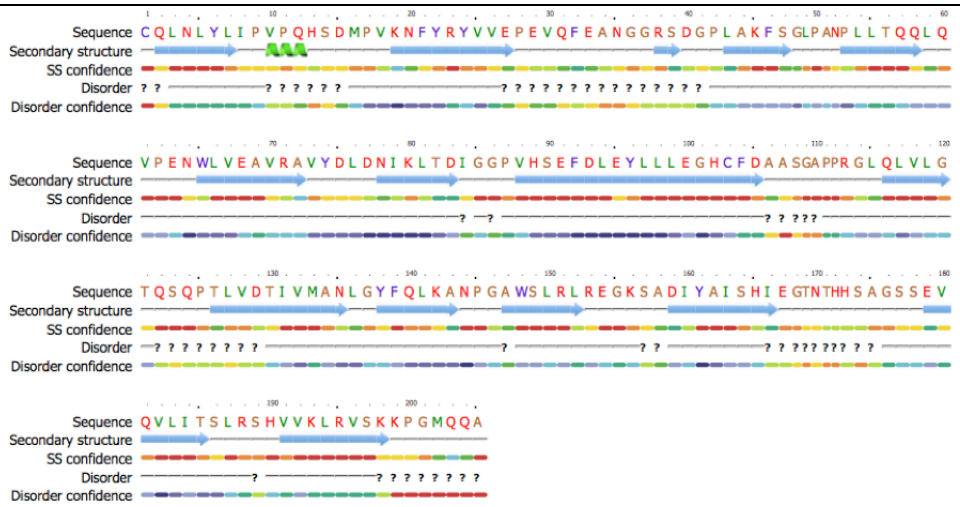


Figure 65: Secondary structure prediction of PcUGGT beta-rich domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

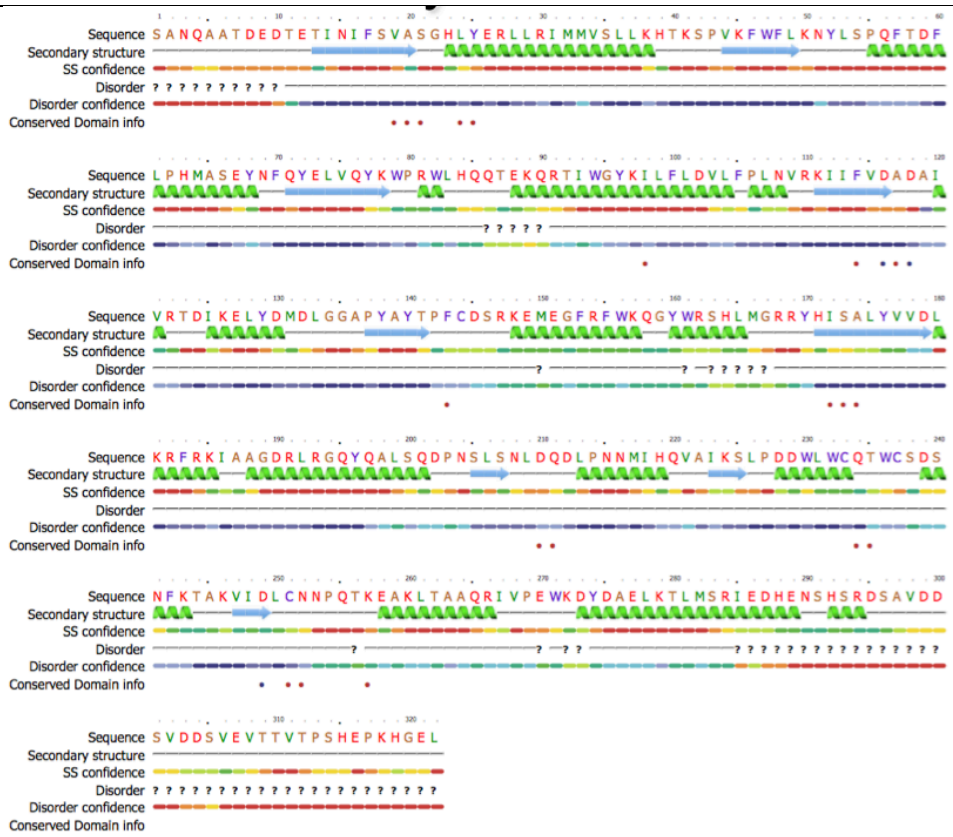


Figure 66: Secondary structure prediction of PcUGGT catalytic domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

7.2 Secondary structure of DmUGGT domains

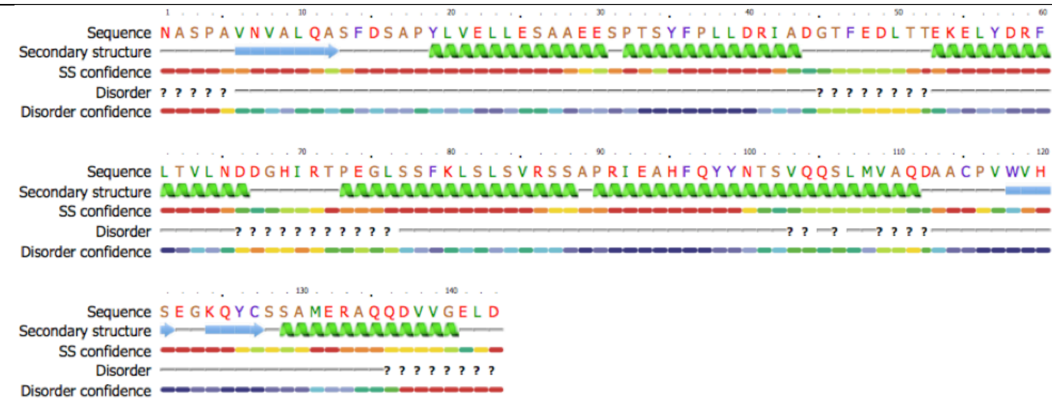


Figure 67: Secondary structure prediction of DmUGGT Pre-Trx domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

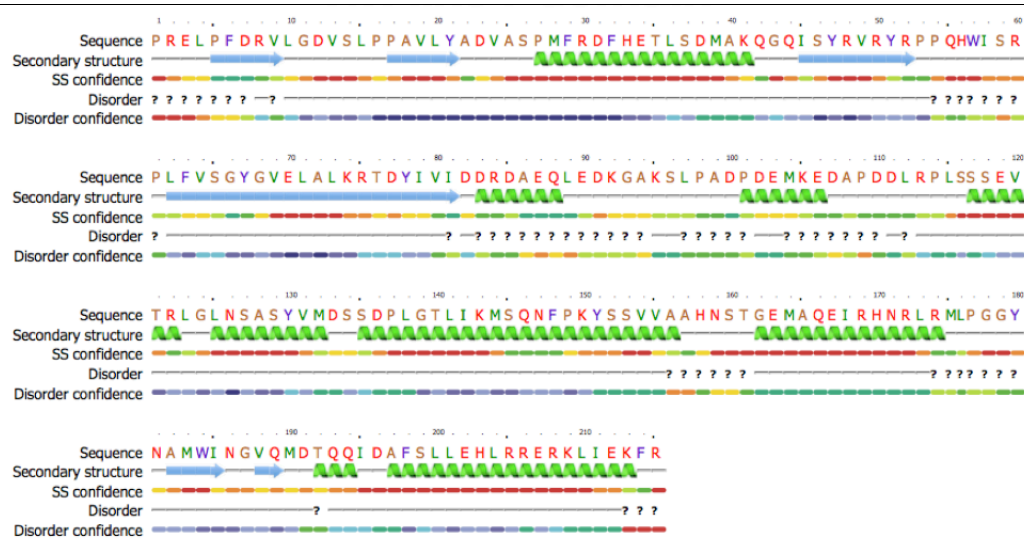


Figure 68: Secondary structure prediction of DmUGGT Trx1 domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

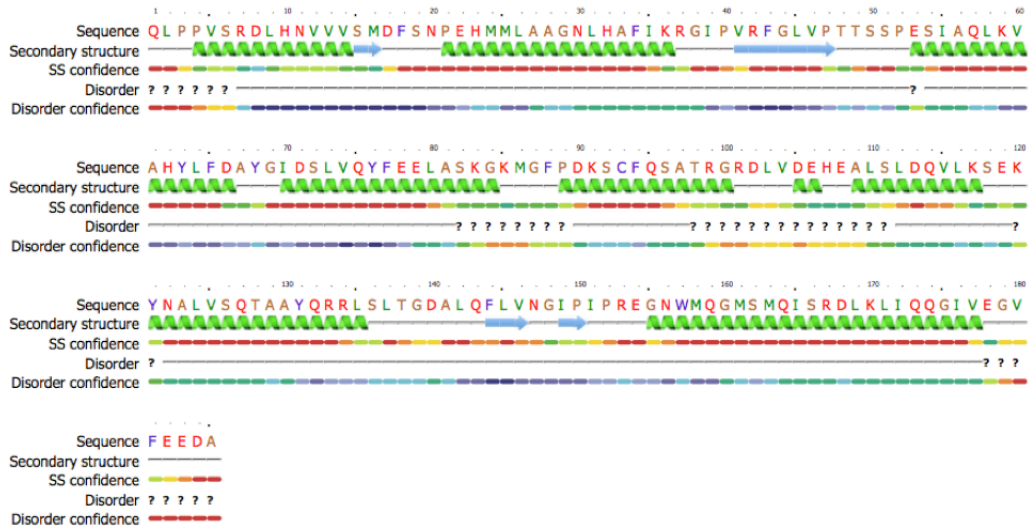


Figure 69: Secondary structure prediction of DmUGGT Trx2 domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

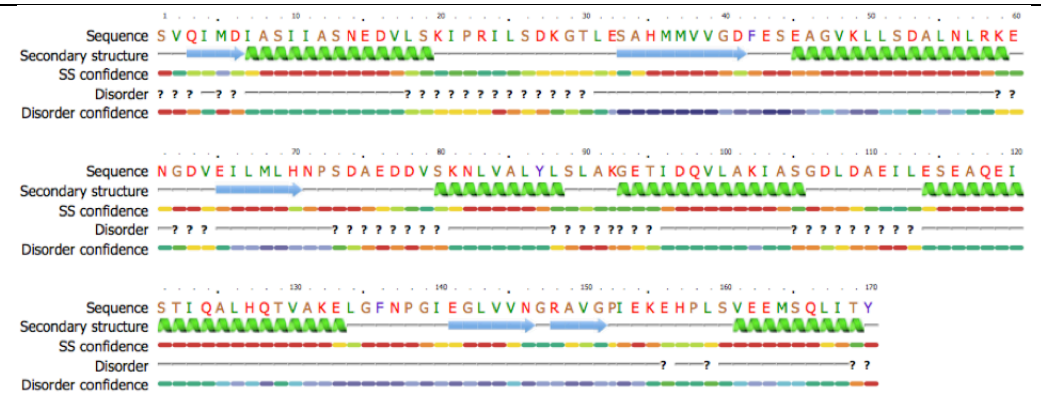


Figure 70: Secondary structure prediction of DmUGGT Trx3 domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

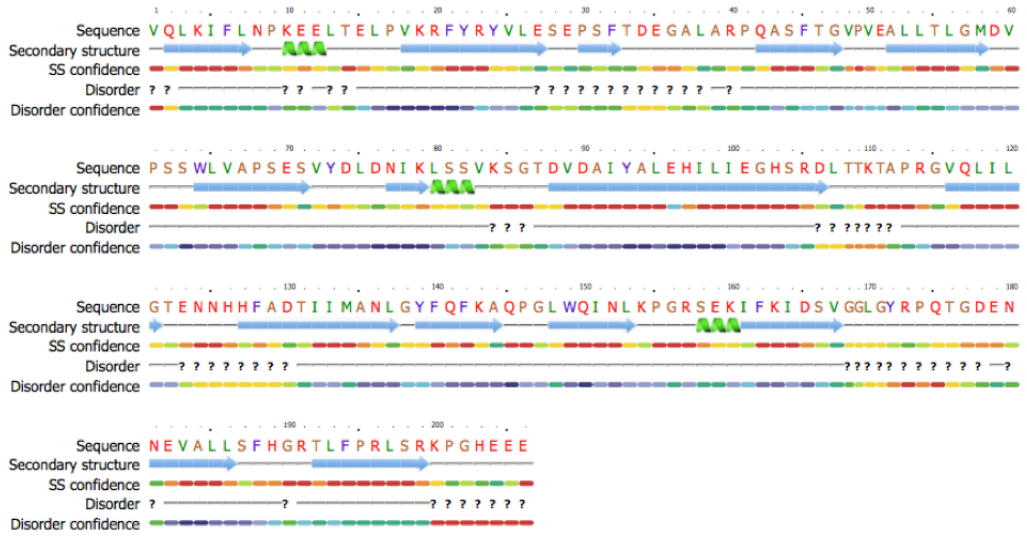


Figure 71: Secondary structure prediction of DmUGGT beta-rich domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

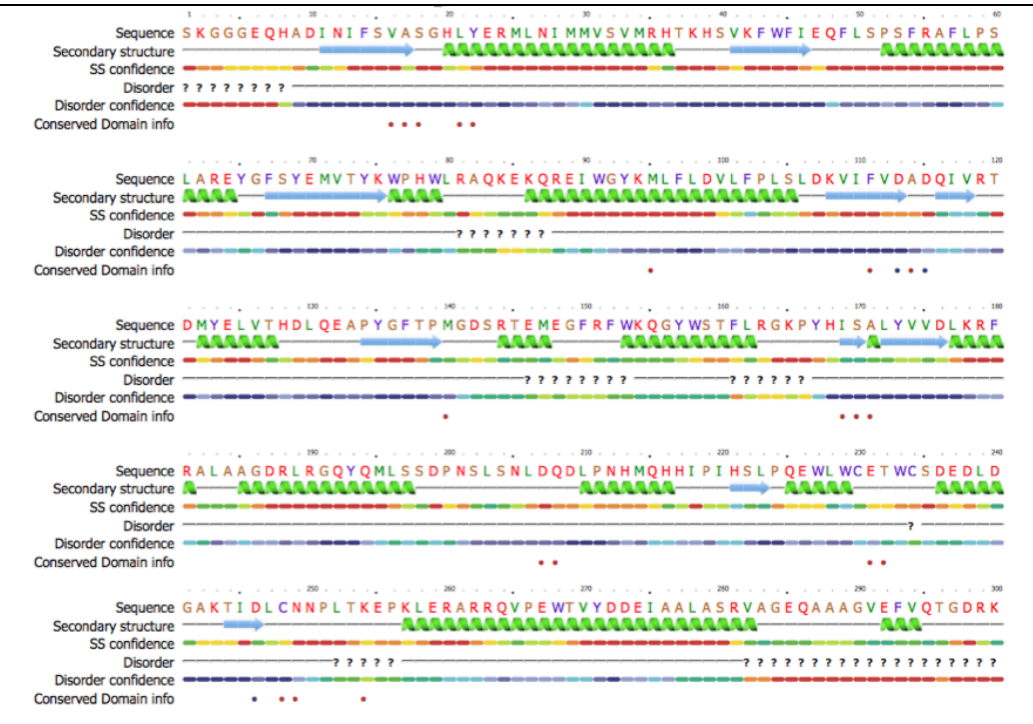


Figure 72: Secondary structure prediction of DmUGGT catalytic domain

Prediction of secondary structure, generated during the homology modeling process using the Phyre2 web server. Beta sheets, alpha helices and disordered regions are represented by blue arrows, green helices and question marks, respectively. Confidence values are color coded, with reds for high confidence and blues for low confidence values.

7.3 MS and HDX-MS data on DmUGGT and DmUGGT/Sep15

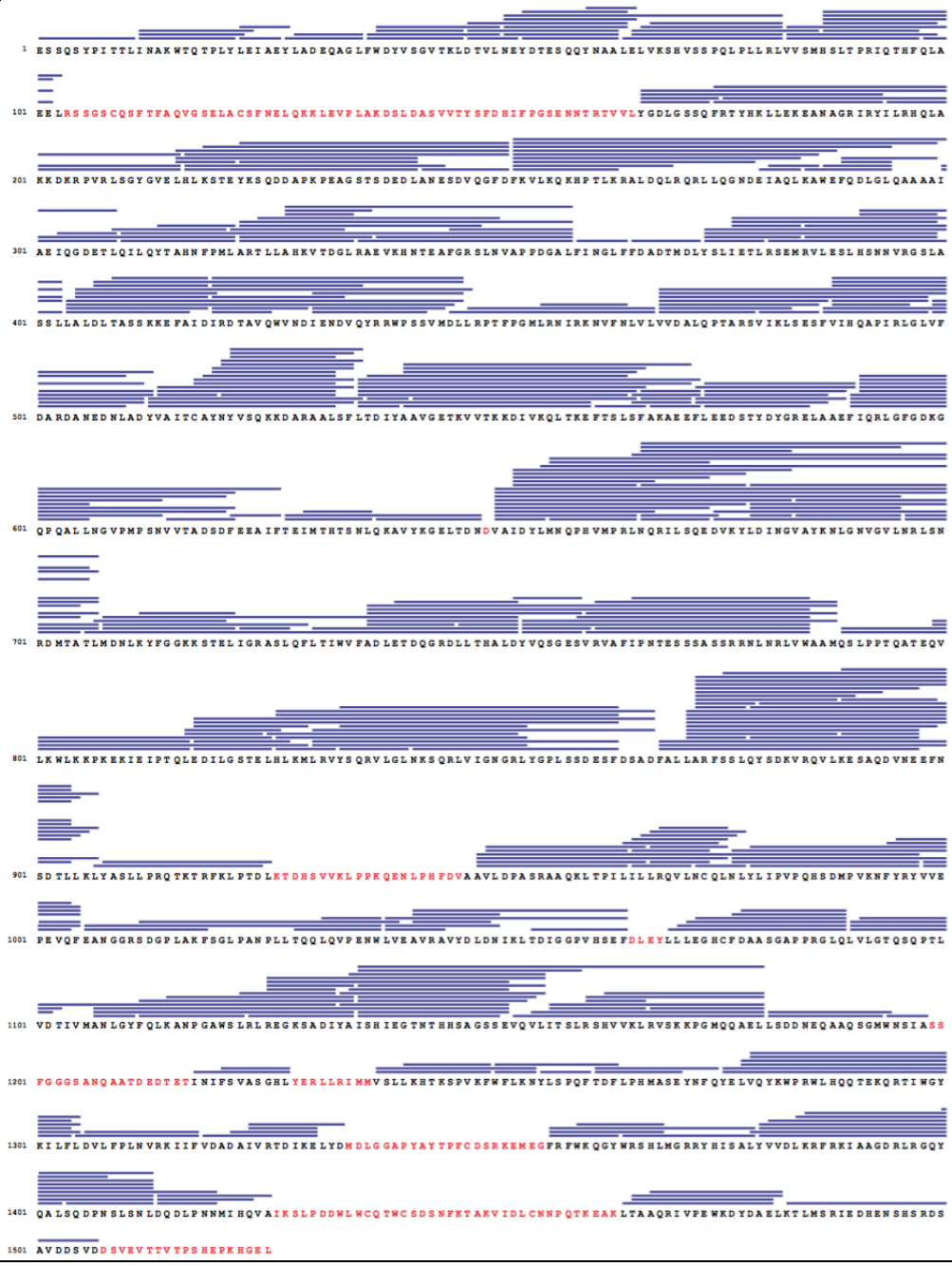


Figure 73: MS and peptide coverage under non-reducing conditions.

Peptides detected by mass spectrometry are represented as blue bars above the sequence to which they were assigned. No peptides could be detected in red colored regions.

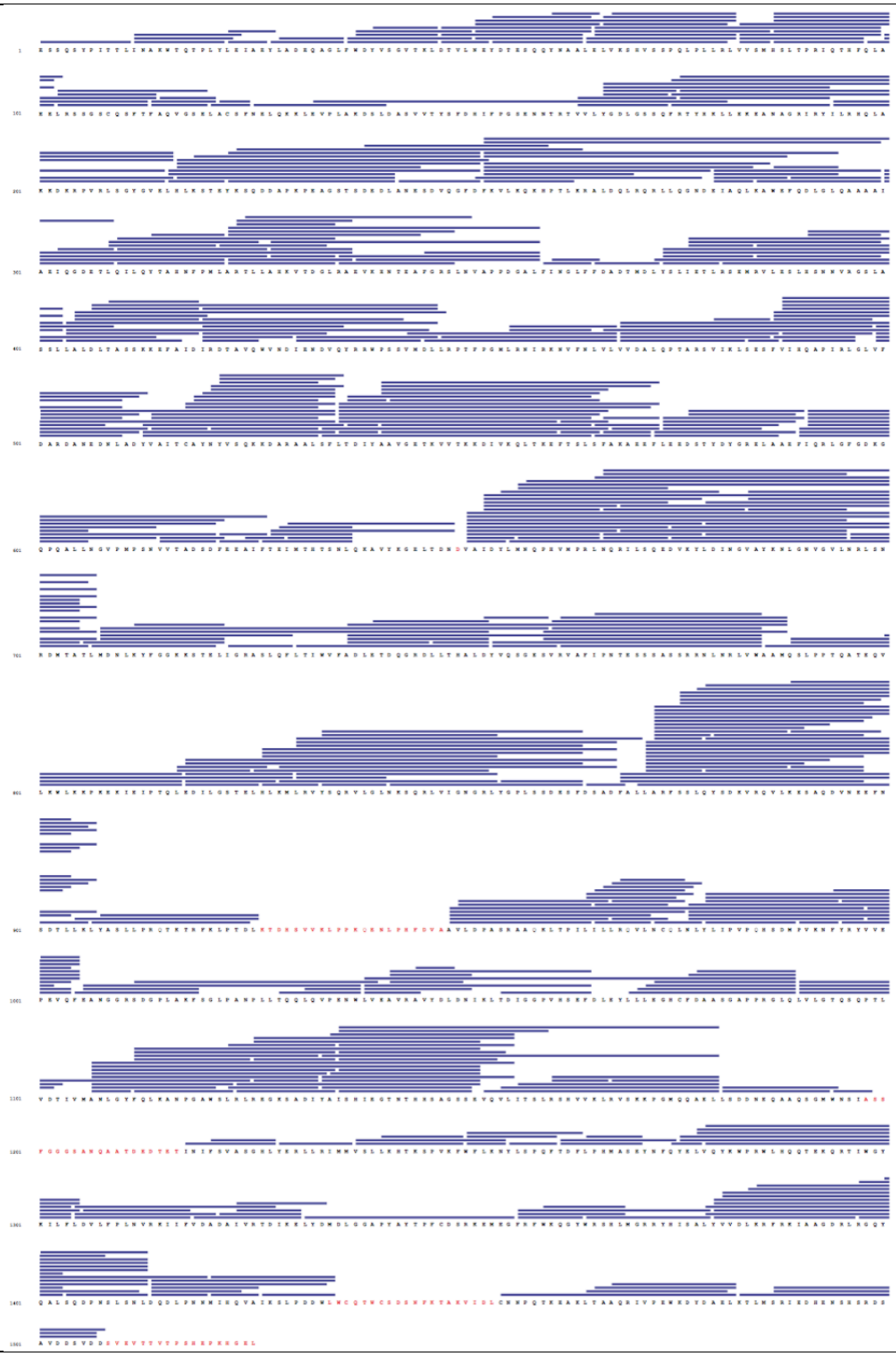


Figure 74: MS and higher peptide coverage under reducing conditions.

Peptides detected by mass spectrometry are represented as blue bars above the sequence to which they were assigned. No peptides could be detected in red colored regions.

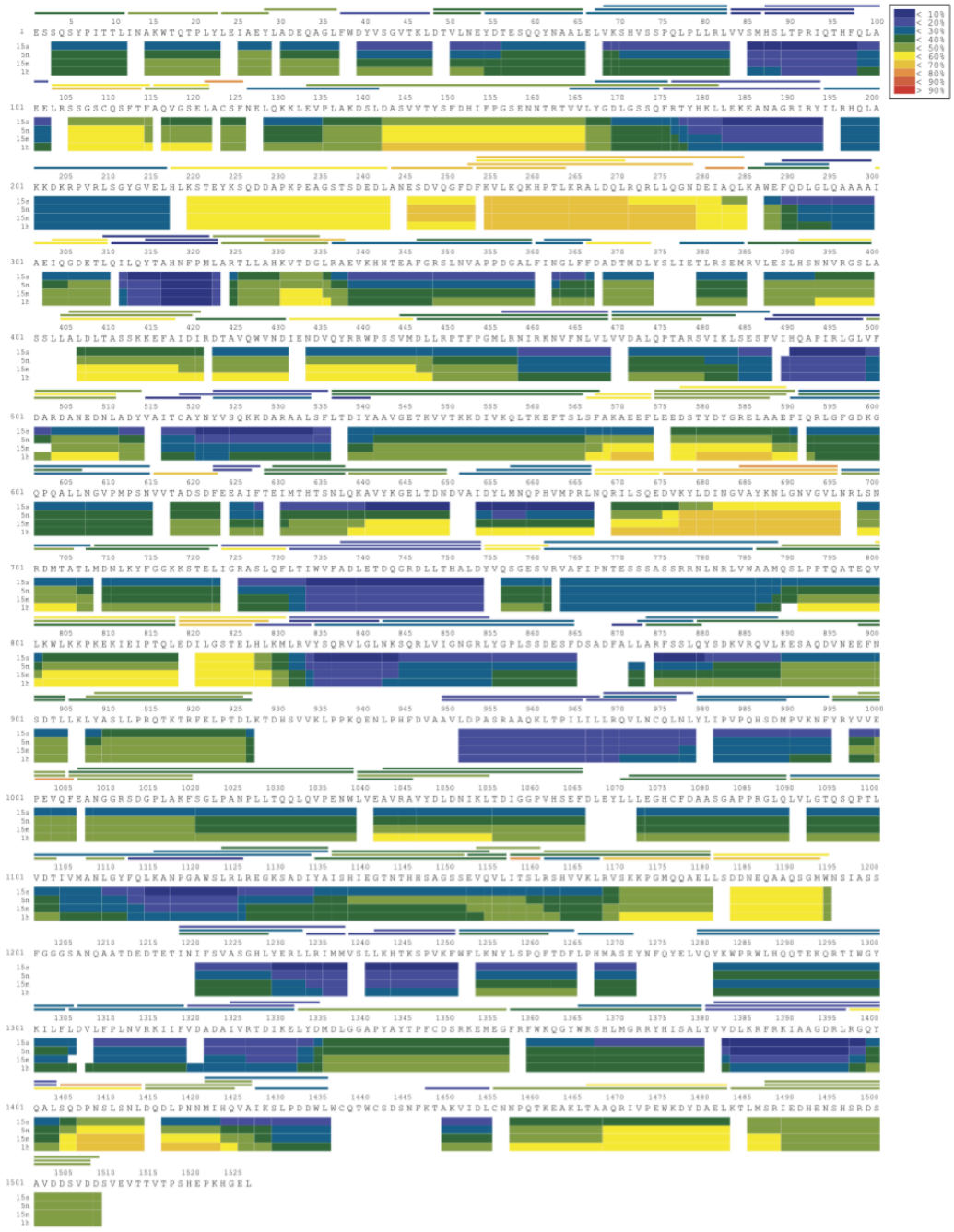


Figure 75: HDX-MS data on DmUGGT.

Peptides detected by mass spectrometry are represented as thin bars above the sequence to which they were assigned. No peptides could be detected in blank regions. The thicker bars below the sequence correspond to HDX rates for four difference incubation periods, from 15 seconds to 1 hour. The bars are color coded from blue to red corresponding to

low to high HDX rates, respectively.

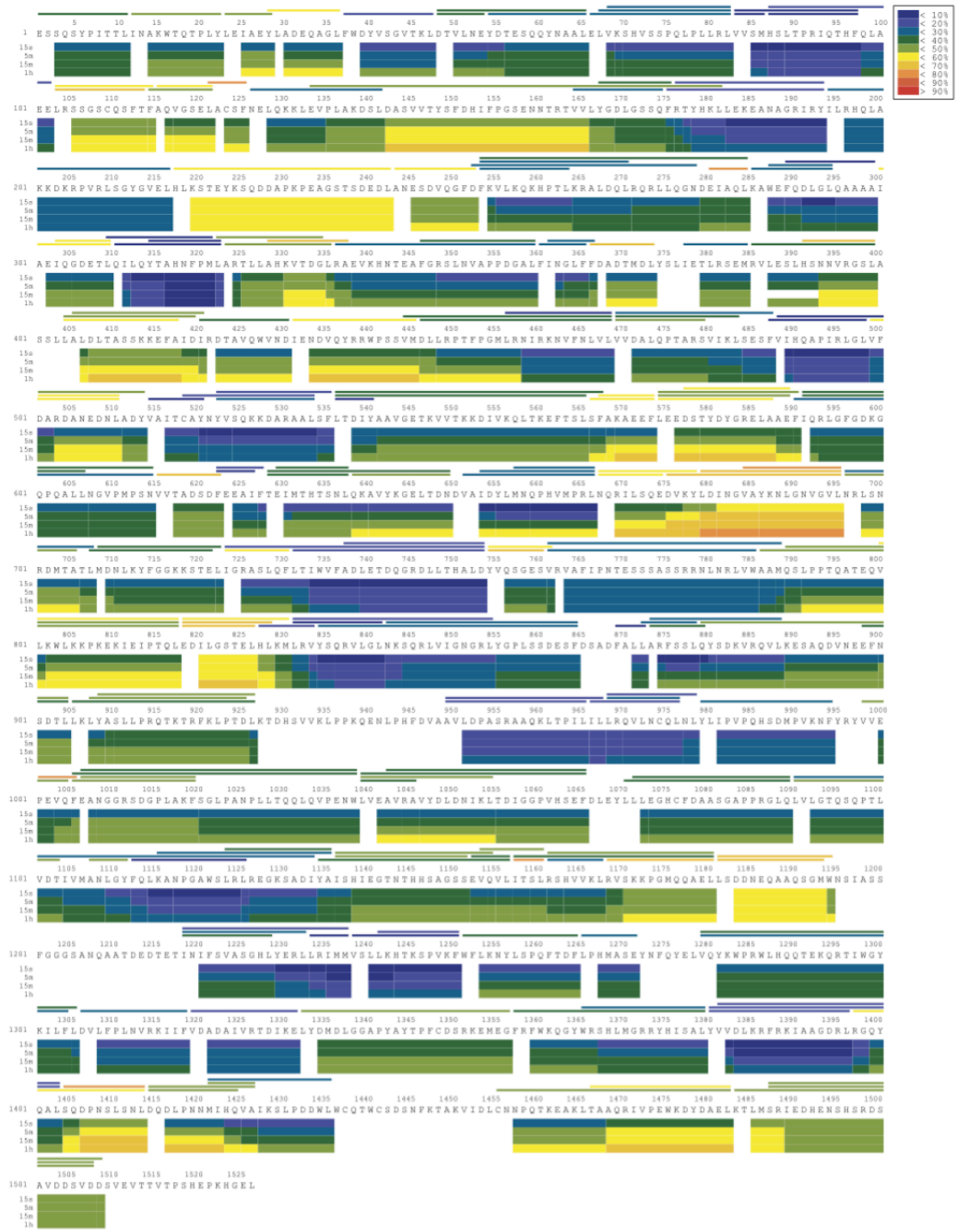


Figure 76: HDX-MS data on the DmUGGT/Sep15 complex

Peptides detected by mass spectrometry are represented as thin bars above the sequence to which they were assigned. No peptides could be detected in blank regions. The thicker bars below the sequence correspond to HDX rates for four difference incubation periods,

from 15 seconds to 1 hour. The bars are color coded from blue to red corresponding to low to high HDX rates, respectively.

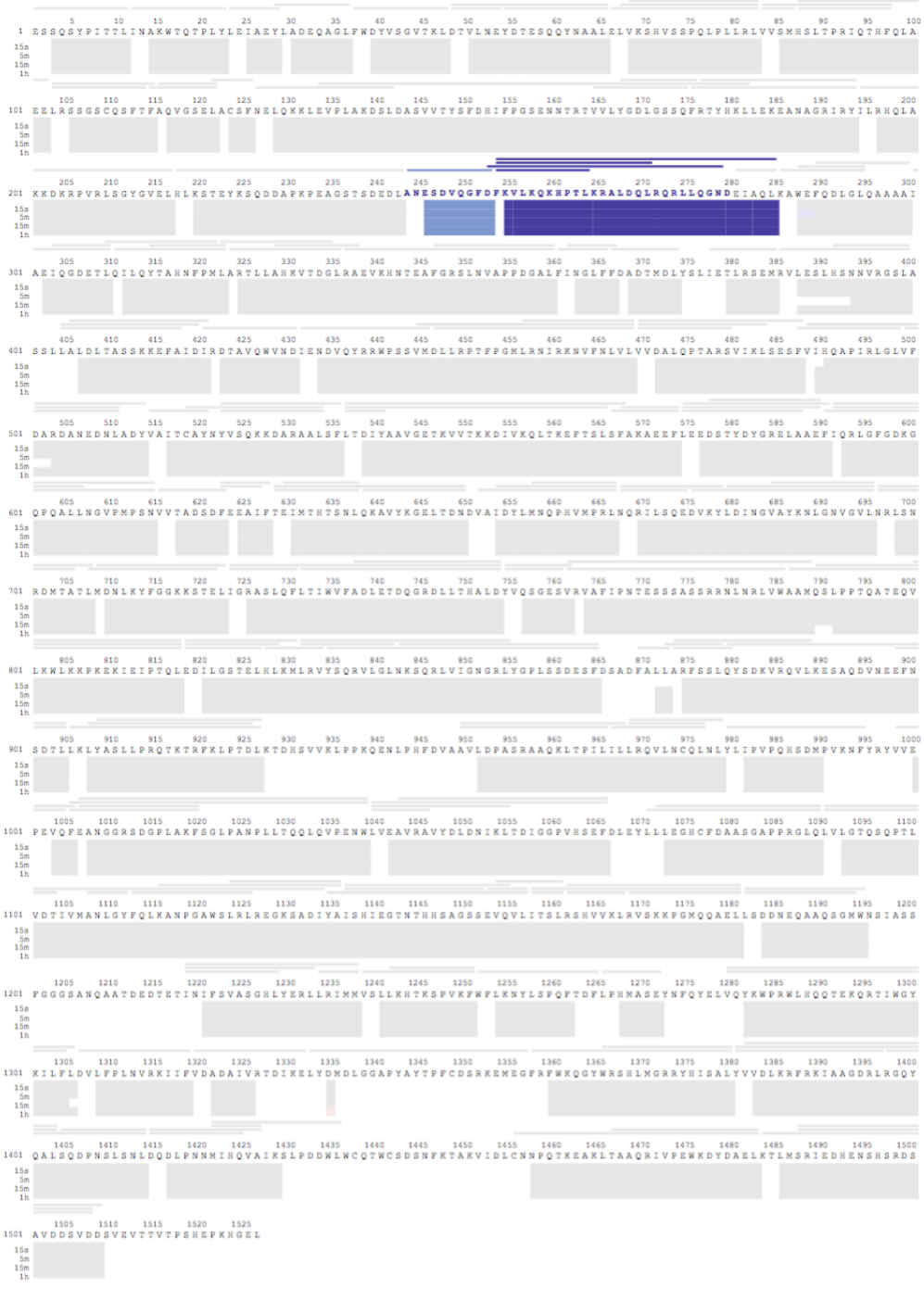


Figure 77: Difference in HDX rates with and without Sep15

Peptides detected by mass spectrometry are represented as thin bars above the sequence

to which they were assigned. No peptides could be detected in blank regions. The thicker bars below the sequence correspond to HDX rates for four difference incubation periods, from 15 seconds to 1 hour. Regions colored in grey showed no difference in HDX rates in the presence or absence of Sep15. Regions colored in light blue and dark blue showed moderate and high decreases in HDX rates in the presence of Sep15.

8 Bibliography

1. Saibil, H.R., *Chaperone machines in action.* Curr Opin Struct Biol, 2008. **18**(1): p. 35-42.
2. Kim, Y.E., et al., *Molecular chaperone functions in protein folding and proteostasis.* Annu Rev Biochem, 2013. **82**: p. 323-355.
3. Maattanen, P., et al., *Protein quality control in the ER: the recognition of misfolded proteins.* Semin Cell Dev Biol, 2010. **21**(5): p. 500-511.
4. Hebert, D.N., S.C. Garman, and M. Molinari, *The glycan code of the endoplasmic reticulum: asparagine-linked carbohydrates as protein maturation and quality-control tags.* Trends Cell Biol, 2005. **15**(7): p. 364-370.
5. Chavan, M. and W. Lennarz, *The molecular basis of coupling of translocation and N-glycosylation.* Trends Biochem Sci, 2006. **31**(1): p. 17-20.
6. Nyathi, Y., B.M. Wilkinson, and M.R. Pool, *Co-translational targeting and translocation of proteins to the endoplasmic reticulum.* Biochim Biophys Acta, 2013. **1833**(11): p. 2392-2402.
7. Young, J.C., et al., *Pathways of chaperone-mediated protein folding in the cytosol.* Nat Rev Mol Cell Biol, 2004. **5**(10): p. 781-791.
8. Roth, J., et al., *Protein N-glycosylation, protein folding, and protein quality control.* Mol Cells, 2010. **30**(6): p. 497-506.
9. Apweiler, R., H. Hermjakob, and N. Sharon, *On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database.* Biochimica et Biophysica Acta, 1999. **1473**: p. 4-8.
10. Dempski, R.E., Jr. and B. Imperiali, *Oligosaccharyl transferase: gatekeeper to the secretory pathway.* Curr Opin Chem Biol, 2002. **6**(6): p. 844-50.
11. Zapun, A., et al., *Enhanced catalysis of ribonuclease B folding by the interaction of calnexin or calreticulin with ERp57.* J Biol Chem, 1998. **273**(11): p. 6009-6012.
12. Kozlov, G., et al., *Structural basis of cyclophilin B binding by the calnexin/calreticulin P-domain.* J Biol Chem, 2010. **285**(46): p. 35551-35557.
13. Kozlov, G., et al., *Structural basis of carbohydrate recognition by calreticulin.* J Biol Chem, 2010. **285**(49): p. 38612-38620.
14. Kozlov, G., et al., *Crystal structure of the bb' domains of the protein disulfide isomerase ERp57.* Structure, 2006. **14**(8): p. 1331-1339.
15. Parodi, A.J., *Role of N-oligosaccharide endoplasmic reticulum processing reactions in glycoprotein folding and degradation.* Biochem J, 2000. **348**(1): p. 1-13.
16. Dejgaard, S., et al., *The ER glycoprotein quality control system.* Curr Issues Mol Biol, 2004. **6**(1): p. 29-42.
17. D'Alessio, C., J.J. Caramelo, and A.J. Parodi, *UDP-GlC:glycoprotein glucosyltransferase-glucosidase II, the ying-yang of the ER quality control.* Semin Cell Dev Biol, 2010. **21**(5): p. 491-499.
18. Trombetta, S.E., S.A. Gañan, and A.J. Parodi, *The UDP-Glc:glycoprotein glucosyltransferase is a soluble protein of the endoplasmic reticulum.* Glycobiology, 1991. **1**(2): p. 155-161.
19. Trombetta, S.E. and A.J. Parodi, *Purification to Apparent Homogeneity andPartial Characterization of Rat Liver UDP-G1ucose:Glycoprotein Glucosyltransferase.* J Biol Chem, 1992. **267**(13): p. 9236-9240.
20. Kozlov, G., et al., *Structure of the noncatalytic domains and global fold of the protein disulfide isomerase ERp72.* Structure, 2009. **17**(5): p. 651-659.

21. Jansen, G., et al., *An interaction map of endoplasmic reticulum chaperones and foldases.* Mol Cell Proteomics, 2012. **11**(9): p. 710-723.
22. Parker, C.G., et al., *Drosophila UDP-glucose:glycoprotein glucosyltransferase: sequence and characterization of an enzyme that distinguishes between denatured and native proteins.* The EMBO Journal, 1995. **14**(7): p. 1294-1303.
23. Vembar, S.S. and J.L. Brodsky, *One step at a time: endoplasmic reticulum-associated degradation.* Nat Rev Mol Cell Biol, 2008. **9**(12): p. 944-957.
24. Fanchiotti, S., et al., *The UDP-Glc:Glycoprotein Glucosyltransferase Is Essential for Schizosaccharomyces pombe Viability under Conditions of Extreme Endoplasmic Reticulum Stress.* The Journal of Cell Biology, 1998. **143**(3): p. 625–635.
25. Guerin, M. and A.J. Parodi, *The UDP-glucose:glycoprotein glucosyltransferase is organized in at least two tightly bound domains from yeast to mammals.* J Biol Chem, 2003. **278**(23): p. 20540-20546.
26. Herrero, A.B., et al., *KRE5 gene null mutant strains of Candida albicans are avirulent and have altered cell wall composition and hypha formation properties.* Eukaryot Cell, 2004. **3**(6): p. 1423-1432.
27. Korotkov, K.V., et al., *Association between the 15-kDa selenoprotein and UDP-glucose:glycoprotein glucosyltransferase in the endoplasmic reticulum of mammalian cells.* J Biol Chem, 2001. **276**(18): p. 15330-15336.
28. Ferguson, A.D., et al., *NMR structures of the selenoproteins Sep15 and SelM reveal redox activity of a new thioredoxin-like family.* J Biol Chem, 2006. **281**(6): p. 3536-3543.
29. Labunskyy, V.M., et al., *A novel cysteine-rich domain of Sep15 mediates the interaction with UDP-glucose:glycoprotein glucosyltransferase.* J Biol Chem, 2005. **280**(45): p. 37839-37845.
30. Labunskyy, V.M., et al., *Sep15, a thioredoxin-like selenoprotein, is involved in the unfolded protein response and differentially regulated by adaptive and acute ER stresses.* Biochemistry, 2009. **48**(35): p. 8458-8465.
31. Gladyshev, V.N., et al., *A new human selenium-containing protein. Purification, characterization, and cDNA sequence.* J Biol Chem, 1998. **273**(15): p. 8910-8915.
32. Shen, Y. and L.M. Hendershot, *ERdj3, a Stress-inducible Endoplasmic Reticulum DnaJ Homologue, Serves as a CoFactor for BiP's Interactions with Unfolded Substrates.* Molecular Biology of the Cell, 2005. **16**: p. 40–50.
33. Peaper, D.R. and P. Cresswell, *Regulation of MHC class I assembly and peptide binding.* Annu Rev Cell Dev Biol, 2008. **24**: p. 343-368.
34. Sousa, M.C., M.A. Ferrero-Garcia, and A.J. Parodi, *Recognition of the Oligosaccharide and Protein Moieties of Glycoproteins by the UDP-G1c:Glycoprotein Glucosyltransferase.* Biochemistry, 1992. **31**: p. 97-105.
35. Choudhury, P., et al., *Intracellular Association between UDP-glucose:Glycoprotein Glucosyltransferase and an Incompletely Folded Variant of a1-Antitrypsin.* Journal of Biological Chemistry, 1997. **272**(20): p. 13446–13451.
36. Tessier, D.C., et al., *Cloning and characterization of mammalian UDP-glucose glycoprotein: glucosyltransferase and the development of a specific substrate for this enzyme.* Glycobiology, 2000. **10**(4): p. 403–412.
37. Tannous, A., et al., *Reglucosylation by UDP-glucose:glycoprotein glucosyltransferase 1 delays glycoprotein secretion but not degradation.* Mol Biol Cell, 2015. **26**(3): p. 390-405.
38. Ritter, C., et al., *Minor folding defects trigger local modification of glycoproteins by the ER folding sensor GT.* EMBO J, 2005. **24**(9): p. 1730-1738.

39. Taylor, S.C., et al., *Glycopeptide specificity of the secretory protein folding sensor UDP-glucose glycoprotein:glucosyltransferase.* EMBO Rep, 2003. **4**(4): p. 405-411.
40. Izumi, M., et al., *Chemical synthesis of intentionally misfolded homogeneous glycoprotein: a unique approach for the study of glycoprotein quality control.* J Am Chem Soc, 2012. **134**(17): p. 7238-7241.
41. Dedola, S., et al., *Folding of synthetic homogeneous glycoproteins in the presence of a glycoprotein folding sensor enzyme.* Angew Chem Int Ed Engl, 2014. **53**(11): p. 2883-2887.
42. Totani, K., et al., *Synthetic substrates for an endoplasmic reticulum protein-folding sensor, UDP-glucose: glycoprotein glucosyltransferase.* Angew Chem Int Ed Engl, 2005. **44**(48): p. 7950-7954.
43. Takeda, Y., et al., *Chemical approaches toward understanding glycan-mediated protein quality control.* Curr Opin Chem Biol, 2009. **13**(5-6): p. 582-591.
44. Totani, K., et al., *The recognition motif of the glycoprotein-folding sensor enzyme UDP-Glc:glycoprotein glucosyltransferase.* Biochemistry, 2009. **48**(13): p. 2933-2940.
45. Sakono, M., et al., *Biophysical properties of UDP-glucose:glycoprotein glucosyltransferase, a folding sensor enzyme in the ER, delineated by synthetic probes.* Biochem Biophys Res Commun, 2012. **426**(4): p. 504-510.
46. Takeda, Y., et al., *Both isoforms of human UDP-glucose:glycoprotein glucosyltransferase are enzymatically active.* Glycobiology, 2014. **24**(4): p. 344-350.
47. Ohara, K., et al., *Profiling Aglycon-Recognizing Sites of UDP-glucose:glycoprotein Glucosyltransferase by Means of Squarate-Mediated Labeling.* Biochemistry, 2015. **54**(31): p. 4909-4917.
48. Taylor, S.C., et al., *The ER protein folding sensor UDP-glucose glycoprotein-glucosyltransferase modifies substrates distant to local changes in glycoprotein conformation.* Nat Struct Mol Biol, 2004. **11**(2): p. 128-134.
49. Ritter, C. and A. Helenius, *Recognition of local glycoprotein misfolding by the ER folding sensor UDP-glucose:glycoprotein glucosyltransferase.* Nat Struct Biol, 2000. **7**(4): p. 278-280.
50. Ito, Y., et al., *Functional analysis of endoplasmic reticulum glucosyltransferase (UGGT): Synthetic chemistry's initiative in glycobiology.* Semin Cell Dev Biol, 2015. **41**: p. 1-9.
51. Arnold, S.M., et al., *Two homologues encoding human UDP-glucose:glycoprotein glucosyltransferase differ in mRNA expression and enzymatic activity.* Biochemistry, 2000. **39**(9): p. 2149-2163.
52. Arnold, S.M. and R.J. Kaufman, *The noncatalytic portion of human UDP-glucose: glycoprotein glucosyltransferase I confers UDP-glucose binding and transferase function to the catalytic domain.* J Biol Chem, 2003. **278**(44): p. 43320-43328.
53. Wearsch, P.A., D.R. Peaper, and P. Cresswell, *Essential glycan-dependent interactions optimize MHC class I peptide loading.* Proc Natl Acad Sci U S A, 2011. **108**(12): p. 4950-4955.
54. Zhang, W., et al., *A role for UDP-glucose glycoprotein glucosyltransferase in expression and quality control of MHC class I molecules.* Proc Natl Acad Sci U S A, 2011. **108**(12): p. 4956-4961.
55. Wearsch, P.A. and P. Cresswell, *The quality control of MHC class I peptide loading.* Curr Opin Cell Biol, 2008. **20**(6): p. 624-631.
56. Zhu, T., T. Satoh, and K. Kato, *Structural insight into substrate recognition by the endoplasmic reticulum folding-sensor enzyme: crystal structure of third thioredoxin-like domain of UDP-glucose:glycoprotein glucosyltransferase.* Sci Rep, 2014. **4**(7322): p. 1-6.

57. Lu, P., et al., *Three-dimensional structure of human gamma-secretase.* Nature, 2014. **512**(7513): p. 166-170.
58. Scheres, S.H., *Beam-induced motion correction for sub-megadalton cryo-EM particles.* Elife, 2014. **3**: p. e03665.
59. Bartesaghi, A., et al., *2.2 A resolution cryo-EM structure of beta-galactosidase in complex with a cell-permeant inhibitor.* Science, 2015. **348**(6239): p. 1147-1151.
60. Merk, A., et al., *Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery.* Cell, 2016. **165**: p. 1-10.
61. Krauss, R.I., et al., *An overview of biological macromolecule crystallization.* Int J Mol Sci, 2013. **14**(6): p. 11643-11691.
62. Rambo, R.P. and J.A. Tainer, *Accurate assessment of mass, models and resolution by small-angle scattering.* Nature, 2013. **496**(7446): p. 477-481.
63. Petoukhov, M.V. and D.I. Svergun, *Applications of small-angle X-ray scattering to biomacromolecular solutions.* Int J Biochem Cell Biol, 2013. **45**(2): p. 429-437.
64. Mertens, H.D. and D.I. Svergun, *Structural characterization of proteins and complexes using small-angle X-ray solution scattering.* J Struct Biol, 2010. **172**(1): p. 128-141.
65. Blanchet, C.E. and D.I. Svergun, *Small-angle X-ray scattering on biological macromolecules and nanocomposites in solution.* Annu Rev Phys Chem, 2013. **64**: p. 37-54.
66. Konarev, P.V., et al., *PRIMUS: a Windows PC-based system for small-angle scattering data analysis.* Journal of Applied Crystallography, 2003. **36**: p. 1277–1282.
67. Ohi, M., et al., *Negative Staining and Image Classification - Powerful Tools in Modern Electron Microscopy.* Biol Proced Online, 2004. **6**(1): p. 23-34.
68. De Carlo, S. and J.R. Harris, *Negative staining and cryo-negative staining of macromolecules and viruses for TEM.* Micron, 2011. **42**(2): p. 117-131.
69. Henderson, R., *The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules.* Q Rev Biophys, 1995. **28**: p. 171-193.
70. Frank, J., *Single-particle reconstruction of biological macromolecules in electron microscopy--30 years.* Q Rev Biophys, 2009. **42**(3): p. 139-158.
71. Kuhlbrandt, W., *Cryo-EM enters a new era.* Elife, 2014. **3**: p. e03678.
72. Dobro, M.J., et al., *Plunge freezing for electron cryomicroscopy.* Methods Enzymol, 2010. **481**: p. 63-82.
73. Baker, L.A. and J.L. Rubinstein, *Radiation damage in electron cryomicroscopy.* Methods Enzymol, 2010. **481**: p. 371-388.
74. Baldwin, P.R. and P.A. Penczek, *The Transform Class in SPARX and EMAN2.* J Struct Biol, 2007. **157**(1): p. 250-261.
75. Hohn, M., et al., *SPARX, a new environment for Cryo-EM image processing.* J Struct Biol, 2007. **157**(1): p. 47-55.
76. Tang, G., et al., *EMAN2: an extensible image processing suite for electron microscopy.* J Struct Biol, 2007. **157**(1): p. 38-46.
77. Elmlund, D. and H. Elmlund, *SIMPLE: Software for ab initio reconstruction of heterogeneous single-particles.* J Struct Biol, 2012. **180**(3): p. 420-427.
78. Scheres, S.H., *RELION: implementation of a Bayesian approach to cryo-EM structure determination.* J Struct Biol, 2012. **180**(3): p. 519-530.
79. Rohou, A. and N. Grigorieff, *CTFFIND4: Fast and accurate defocus estimation from electron micrographs.* J Struct Biol, 2015. **192**(2): p. 216-221.

80. Li, X., et al., *Influence of electron dose rate on electron counting images recorded with the K2 camera.* J Struct Biol, 2013. **184**(2): p. 251-260.
81. Li, X., et al., *Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM.* Nat Methods, 2013. **10**(6): p. 584-590.
82. Bai, X.C., et al., *Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles.* Elife, 2013. **2**: p. e00461.
83. Banterle, N., et al., *Fourier ring correlation as a resolution criterion for super-resolution microscopy.* J Struct Biol, 2013. **183**(3): p. 363-367.
84. Cardone, G., J.B. Heymann, and A.C. Steven, *One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions.* J Struct Biol, 2013. **184**(2): p. 226-236.
85. Chen, S., et al., *High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy.* Ultramicroscopy, 2013. **135**: p. 24-35.
86. Kelley, L.A., et al., *The Phyre2 web portal for protein modeling, prediction and analysis.* Nat Protoc, 2015. **10**(6): p. 845-858.
87. Bennett-Lovsey, R.M., et al., *Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre.* Proteins, 2008. **70**(3): p. 611-625.
88. Kelley, L.A. and M.J. Sternberg, *Protein structure prediction on the Web: a case study using the Phyre server.* Nat Protoc, 2009. **4**(3): p. 363-371.
89. Lau, P.W., et al., *DOLORS: versatile strategy for internal labeling and domain localization in electron microscopy.* Structure, 2012. **20**(12): p. 1995-2002.
90. Molinari, M., et al., *Persistent glycoprotein misfolding activates the glucosidase II/UGT1-driven calnexin cycle to delay aggregation and loss of folding competence.* Mol Cell, 2005. **20**(4): p. 503-12.
91. Buzzi, L.I., et al., *The two Caenorhabditis elegans UDP-glucose:glycoprotein glucosyltransferase homologues have distinct biological functions.* PLoS One, 2011. **6**(11): p. e27025.
92. Kikhney, A.G. and D.I. Svergun, *A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins.* FEBS Lett, 2015. **589**(19): p. 2570-2577.
93. Elmlund, H., D. Elmlund, and S. Bengio, *PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy.* Structure, 2013. **21**(8): p. 1299-1306.
94. Bartesaghi, A., et al., *Structure of beta-galactosidase at 3.2-A resolution obtained by cryo-electron microscopy.* Proc Natl Acad Sci U S A, 2014. **111**(32): p. 11709-11714.
95. Carragher, B., et al., *Leginon: an automated system for acquisition of images from vitreous ice specimens.* J Struct Biol, 2000. **132**(1): p. 33-45.
96. Pulokas, J., et al., *Automated EM Data Acquisition using Leginon II.* Microscopy and Microanalysis, 2004. **10**(S02): p. 1510-1511.
97. Suloway, C., et al., *Automated molecular microscopy: the new Leginon system.* J Struct Biol, 2005. **151**(1): p. 41-60.
98. Mindell, J.A. and N. Grigorieff, *Accurate determination of local defocus and specimen tilt in electron microscopy.* J Struct Biol, 2003. **142**(3): p. 334-47.
99. Scheres, S.H., *A Bayesian view on cryo-EM structure determination.* J Mol Biol, 2012. **415**(2): p. 406-418.
100. Kasaikina, M.V., et al., *Roles of the 15-kDa selenoprotein (Sep15) in redox homeostasis and cataract development revealed by the analysis of Sep 15 knockout mice.* J Biol Chem, 2011. **286**(38): p. 33203-33212.
101. Yin, N., et al., *Knockdown of 15-kDa selenoprotein (Sep15) increases hLE cells' susceptibility to tunicamycin-induced apoptosis.* J Biol Inorg Chem, 2015. **20**(8): p. 1307-1317.

102. Okiyoneda, T., et al., *Mechanism-based corrector combination restores DeltaF508-CFTR folding and function.* Nat Chem Biol, 2013. **9**(7): p. 444-454.