INVESTIGATION OF THE STRUCTURAL REQUIREMENTS FOR OLIGOSACCHARYL TRANSFERASE FUNCTION

by

Marcie B. Jaffee

B.A. Biochemistry
Columbia University, 2007

ARCHIVES
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SEP 04 2013
LIBRARIES

SUBMITTED TO THE DEPARTMENT OF BIOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2013
[SEPTEMBER]

2013 Massachusetts Institute of Technology. All rights reserved.

Signature of Author: ______________________
Department of Biology
August 22, 2013

Certified by: ______________________
Barbara Imperiali
Class of 1922 Professor of Chemistry and Professor of Biology
Thesis Supervisor

Accepted by: ______________________
Steve P. Bell
Professor of Biology
Chair, Departmental Committee on Graduate Students

INVESTIGATION OF THE STRUCTURAL REQUIREMENTS FOR OLIGOSACCHARYL TRANSFERASE FUNCTION

by

Marcie B. Jaffee

Submitted to the Department of Biology
on August 23, 2013 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

ABSTRACT

N-linked glycosylation is a ubiquitous protein modification involved in a wide range of cellular functions and diseases (Varki, 1993). The oligosaccharyl transferase (OTase), the principle enzyme responsible for catalysis of N-linked glycosylation, facilitates the transfer of a pre-assembled oligosaccharide from a polyprenyl-lipid donor to the side-chain nitrogen of asparagine acceptors. In a majority of eukaryotes the OTase is a multimeric, membrane-bound complex, which has precluded detailed structural and mechanistic studies. The discovery of a homologous bacterial N-linked glycosylation system in 1999 provided new opportunities to study the fundamental biosynthetic and mechanistic components of the process. Specifically, the bacterial OTase, PglB, is monomeric and homologous to the catalytic subunit of the eukaryotic OTase. The following chapters describe studies designed to learn about the OTase transfer reaction using PglB as a model. Optimization of expression and purification of PglB is first described, as initial expression of PglB in *E. coli* showed poor yields and instability upon purification. Bioinformatic analysis was performed to define sequence regions of functional importance and to ascertain the extent of homology between PglB and other OTases, which verified the value of PglB studies for understanding OTases generally. The bioinformatics analysis was based on topology predictions for PglB and various homologs. The general conservation in OTase membrane configuration was used to structurally guide and simplify sequence analyses for a large number of diverse OTases. The method revealed highly conserved motifs within the soluble loops appearing between transmembrane domains and biochemical analysis was carried out to establish the functional importance of these domains. These results were published concurrently with an independently determined X-ray crystal structure of PglB, which validated and complemented the bioinformatic and biochemical results. Subsequently, a sensitive luminescent assay was designed and developed as a way to investigate the dynamics of substrate binding and conformational changes in the OTase reaction. Initial results demonstrate the informative value of the system. Measurements using this system will be continued by Imperiali lab members.

Thesis Supervisor: Barbara Imperiali
Class of 1922 Professor of Chemistry and Professor of Biology

ACKNOWLEDGMENTS

For my mother, Amelia Marie Marcinko,

to whom I owe my love for science,

my strength, my faith,

and my solidarity with the gypsies of the world.

TABLE OF CONTENTS

ABSTRACT **2**

ACKNOWLEDGMENTS **3**

TABLE OF CONTENTS **5**

LIST OF FIGURES **9**

LIST OF TABLES **14**

LIST OF ABBREVIATIONS **15**

CHAPTER 1: Introduction **20**

Asparagine-linked glycosylation: significance, applications, and progress **21**

The oligosaccharyl transferase **29**

The peptidyl substrate of the oligosaccharyl transferase **37**

Conclusion **45**

CHAPTER 2: An optimized protocol for expression and purification of PglB **49**

Acknowledgments: **53**

Summary **54**

Introduction **55**

Results **58**
- Expression optimization of PglB in E. coli: general strategy 58
- Protein Tags and Gene Truncations 58
- Expression and induction conditions 61
- Optimization of PglB purification 67
- Isolation and processing of cell-membrane fraction 70
- Solubilization of membrane proteins 73
- Affinity chromatography 75
- Buffer Exchange 77
- PglB Purification protocol 80

Discussion and Conclusion **82**

Materials and Methods **86**
Vectors and cloning 86
Protein Expression 86
Activity Assay 88
Protein Quantification 88
Cell lysis 89
Isolation of Cell-Envelope Fraction (CEF) 89
Extraction of membrane proteins from cell membrane 90
Ni-NTA purification 90

References **92**

CHAPTER 3: Sequence-driven bioinformatic analysis of the evolutionary diversity of oligosaccharyl transferases **95**

Summary **96**

Introduction **97**

Results **102**
Predicted topologies are similar for PglB and Stt3 homologs 102
Identification of conserved motifs 106
Assessing structural conservation in loops 109

Conclusions **123**

Materials and Methods **124**
Bioinformatic analysis 124
Summary of topology sequence-extraction programs 125

References **128**

CHAPTER 4: Biochemical validation and characterization of sequence motifs conserved in oligosaccharyl transferases 131

Acknowledgments 131

Summary 132

Introduction 133

Results 137
  Mutational and kinetic characterization of residues in conserved motifs 137
  Mutants maintain tertiary structure 142
  Correlation of mutation effects with concentration of Und-PP-disaccharide substrate 144

Discussion 148
  Interpretation of PglB mutants in context of kinetic and structural data. 148
  Mutants maintain tertiary structure 152

Conclusion 156

Materials and Methods 158
  Mutant production, expression, and purification 158
  PglB Activity Assays 160
  Synthesis of lipid-linked sugar substrate at three specific activities 160
  Limited Proteolysis 163

References 164

CHAPTER 5: Design and development of LRET-based system to investigate the dynamics of PglB catalysis 167

Acknowledgments 167

Summary 168

Introduction 169

Results 176
Expression of LBT-PglB construct 176
Optimization of LBT-PglB luminescence 184
Insertion of cysteine mutations to allow labeling of PglB with thiol-reactive acceptor dye... 197
Synthesis and labeling of cysteine-containing peptide substrates of PglB 202
Initial LRET measurements 207

Conclusion 210

Materials and Methods 213
Synthesis of substrate and negative-control peptides 213
Purification and labeling of substrate-peptide and negative-control peptide 213
Emission spectra 214
Lifetime measurements 215
Calculation of distances using lifetimes 215

References 217

LIST OF FIGURES

CHAPTER 1

Figure 1: The N-linked glycosylation biosynthetic pathway in S. cerevisiae. 22

Figure 2: The N-linked glycosylation biosynthetic pathway in C. jejuni. 23

Figure 3: The cotranslational process of eukaryotic N-linked glycosylation. 24

Figure 4: Structure of the polyprenyl-diphosphate-glycan substrate common in eukaryotes. 25

Figure 5: Structures of biologically relevant monosaccharides in eukaryotes. 26

Figure 6: Structural complexity of monosaccharides and oligosaccharides. 27

Figure 7: Comparison of the OTases in *C. jejuni* (bacteria) and *S. cerevisiae* (eukaryotes).30

Figure 8: N-linked glycosylation transfer reaction in *C. jejuni*. 32

Figure 9: Topology model of PglB agrees with X-ray structure data 33

Figure 10: Alignment of PglB sequences from *C. jejuni* and *C. lari*. 36

Figure 11: Comparison of the Asx-type and β-type turn peptide Ac-Asn-Xaa-Thr-NH_2. ...37

Figure 12: Proposed OTase mechanisms. A) Marshall. B) Bause. C) Imperiali. 39

Figure 13: Structures of peptides constrained to bias Asx-type or β-type turn formation ..41

CHAPTER 2

Figure 1: Comparison of the OTases in *C. jejuni* (bacteria) and *S. cerevisiae* (eukaryotes).57

Figure 2: Investigating a natural degradation product of PglB. 59

Figure 3: Coomassie-stained SDS-PAGE of fractions of multiple PglB constructs. 61

Figure 4: Expression improvement with IPTG induction at high O.D and auto-induction.64

Figure 5: Procedure for optimizing purification protocol for PglB. 67

Figure 6: Activity-rate data for fractions used to calculate Table 3 purification values. 69

Figure 7: Optimization of Ni-NTA purification. 76

Figure 8: Activity of PglB over multiple days post-purification. 79

CHAPTER 3

Figure 1: Translation of topological probabilities to a picture of protein architecture........ 97

Figure 2: Pymol representation of crystal structure of *P. furiosus* Stt3 soluble domain...... 99

Figure 3: Representation summarizing the topology-based sequence analysis performed on Stt3 homologs. 101

Figure 4. Sequence-based topology predictions for several Stt3 homologs......................... 104

Figure 5: Alignments for highly conserved regions of OTase sequences. 107

Figure 6: Topology model of PglB showing locations of conserved motifs. 108

Figure 7: Secondary structure predictions for first major conserved loop in several Stt3 homologs. 111

Figure 8: Example of one source of ambiguity in topology predictions: when a helix is both 'soluble' and integral-membrane'. 113

Figure 9: Secondary-structure predictions for regions of Stt3 sequences predicted to correspond to the first soluble loop containing the conserved L/I-xxx-D-x-Y/F/W motif. . 115
  Figure 9A: Eukaryotic Stt3 homologs 116
  Figure 9B: Bacterial Stt3 homologs 117
  Figure 9C: Archaeal Stt3 homologs 118

Figure 10: Secondary-structure predictions for regions of Stt3 sequences predicted to correspond to the soluble loop containing the conserved S/T-I/V-x-E motif. 119
  Figure 10A: Eukaryotic Stt3 homologs 120
  Figure 10B: Bacterial Stt3 homologs 121
  Figure 10C: Archaeal Stt3 homologs 122

Figure 11: Python code for programs developed to parse and group sequence data according to predicted topological constraints. 127

CHAPTER 4

Figure 1: Topology model of PglB showing locations of mutants. 134

Figure 2: PglB-catalyzed N-glycosylation reaction in *C. jejuni*. 135

Figure 3: Western blots of PglB wild-type and mutant CEFs show similar expression levels 138

Figure 4: PglB activity assays performed on CEFs of loop mutants. 139

Figure 5: Activity data for soluble-domain mutants and extended time points for loop mutants. 140

Figure 6: Probing effects of mutations on PglB tertiary structure using limited proteolysis. 143

Figure 7: Partial rescue of mutant activity with increasing sugar-substrate concentrations. 145

Figure 8: Quantitative western blotting of mutants with partial activity loss. 146

Figure 9: Relative conservation of residues within and surrounding conserved motifs. 150

Figure 10: N-terminal sequencing of degradation products reveals sites of protease susceptibility 153

CHAPTER 5

Figure 1: Jablonski diagram representing fluorescent and phosphorescent emission. 170

Figure 2: Sensitization of Tb^{3+} by the LBT. 173

Figure 3: Simple illustration of the planned LRET experiments. 175

Figure 4: Poor initial expression of LBT-PglB and improvements with TEV insertion 178

Figure 5: LBT before and after optimization of nucleotide content. 182

Figure 6: Purification of stocks of T7-PglB and LBT-PglB. 183

Figure 7: Expression and purification of an LBT-ubiquitin standard. 185

Figure 8: Measurements of LBT-Ubiquitin intensities and lifetime. 186

Figure 9: Effects of buffer components on LBT-Ubiquitin luminescence intensities. 187

Figure 10: Tb^{3+} can satisfy cation requirement for T7-PglB activity. 188

Figure 11: LBT-PglB shows relatively low luminescence intensities. 189

Figure 12: Effects of varying phosphorimeter parameters on luminescence intensity........190

Figure 13: Improvement in luminescent intensities measured for LBT-Ubiquitin and LBT-PglB when the number of flashes is increased. 191

Figure 14: Comparison of luminescent intensities of three proteins fused to LBT........193

Figure 15: Additional data from luminescent titration experiments shown in Figure 14..194

Figure 16: Comparison of LBT-PglB luminescence intensities and lifetime-curve fits in the presence of three buffers........195

Figure 17: Effect of detergent identity and concentration on luminescence intensity of LBT-PglB in presence of Tb^{3+}........196

Figure 18: Effect of cation identity and concentration on LBT-PglB luminescence. 196

Figure 19: Mutating native cysteines in PglB to serine........197

Figure 20: Efficiency (E) of LRET as a function of Tb^{3+} (donor) and Bodipy-TMR (acceptor) distance (r). 198

Figure 21: Structure and photophysical properties of LRET pair. 200

Figure 22: Sites of cysteines mutated into PglB to act as handles for LRET acceptor. 201
Synthesis and labeling of cysteine-containing peptide substrates of PglB 202

Figure 23: Two peptides synthesized for intermolecular LRET studies. 203

Figure 24: Distance measurements determine the preferred location of cysteines in substrate peptide and site-specific mutagenesis. 204

Figure 25: MALDI-MS validation of masses of peptides for LRET studies. 205

Figure 26: HPLC purification of Bodipy-TMR labeled peptides. 206

Figure 27: Bodipy-TMR is a favorable LRET acceptor to LBT-Tb^{3+} donor luminescence. 206

Figure 28: PglB activity assays in presence of substrate and negative-control peptides....207

Figure 29: Increase and saturation acceptor emission........208

Figure 30: LRET measurements with labeled peptide substrate..209

LIST OF TABLES

CHAPTER 2

Table 1: Fusion constructs used in PglB expression screening. 60

Table 2: Induction conditions varied to screen for improved expression of PglB. 62

Table 3: Values from final optimized purification protocol of PglB. 69

Table 4: Minor change in centrifugal spin significantly affects yield of purification step ... 71

Table 5: Comparison of several salt solutions in CEF salt wash efficiency. 72

Table 6: Comparison of several detergent solutions in CEF-solubilization efficiency. 74

CHAPTER 3

Table 1: Identities of sequences used in all bioinformatic analysis. 105

CHAPTER 4

Table 1: Relative activity and locations of PglB mutants. 141

Table 2: Activity of mutants relative to wild-type at various sugar-substrate concentrations. 147

CHAPTER 5

Table 1: Summary of major vectors discussed in this chapter. 177

Table 2: Comparison of nucleotide content of the four codons following the start codon for several constructs discussed. 179

LIST OF ABBREVIATIONS

Standard 1-letter and 3-letter codes are used for the 20 natural amino acids.

Amino Acid	Single-Letter	Three-Letter	DNA codons
Isoleucine	I	Ile	ATT, ATC, ATA
Leucine	L	Leu	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	Val	GTT, GTC, GTA, GTG
Phenylalanine	F	Phe	TTT, TTC
Methionine	M	Met	ATG
Cysteine	C	Cys	TGT, TGC
Alanine	A	Ala	GCT, GCC, GCA, GCG
Glycine	G	Gly	GGT, GGC, GGA, GGG
Proline	P	Pro	CCT, CCC, CCA, CCG
Threonine	T	Thr	ACT, ACC, ACA, ACG
Serine	S	Ser	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	Tyr	TAT, TAC
Tryptophan	W	Trp	TGG
Glutamine	Q	Gln	CAA, CAG
Asparagine	N	Asn	AAT, AAC
Histidine	H	His	CAT, CAC
Glutamic acid	E	Glu	GAA, GAG
Aspartic acid	D	Asp	GAT, GAC
Lysine	K	Lys	AAA, AAG
Arginine	R	Arg	CGT, CGC, CGA, CGG, AGA, AGG
STOP	-	-	TAA, TAG, TGA

Abbreviation	Meaning
Ac	acetyl
ALG	Asparagine-linked glycosylation
ATP	adenosine 5'-triphosphate
Bac	Bacillosamine, (2,4-diacetamido-2,4,6-trideoxyglucose)
Bn	benzyl
Bodipy	6-((4,4-Difluoro-1,3-Dimethyl-5-(4-Methoxyphenyl)-4-Bora-3a,4a-Diaza-s-Indacene-2-Propionyl)amino)hexanoic Acid
Bz	benzoyl
Ci	Curies
C. lari	*Campylobacter lari*
C. jejuni	*Campylobacter jejuni*
DDM	n-dodecyl β-D-maltoside
DIEA	diisopropylethylamine
DMF	dimethylformamide
DMSO	dimethylsulfoxide
Dol	dolichyl
Dol-P	dolichyl phosphate
Dol-PP	dolichylpyrophosphate
DPM	disintegrations per minute
DTT	dithiothreitol
ER	endoplasmic reticulum
ESI	electrospray ionization mass spectrometry
Fmoc	fluoren-9-ylmethyloxycarbonyl
E. coli	*Escherichia coli*
Eu	europium
GDP-Man	guanosine 5’-diphospho-α-D-mannose
GalNAc	N-acetyl-D-galactosamine
GlcNAc	N-acetyl-D-glucosamine
Gal	D-galactose

gi	Gene identifier
Glc	D-glucose
His-X, His_x	Protein tag consisting of 'X'sequential encoded histidines
HPLC	high-performance liquid chromatography
IgG	Immunoglobulin G
K_i	inhibition constant
K_m	Michaelis constant
LB	Lysogeny Broth
LBT	Lanthanide-binding Tag
MALDI-MS	Matrix assisted laser desorption ionization mass spectrometry
Man	D-mannose
MBP	Maltose-binding Protein
MHC	Major histocompatibility complex
N-glycan	Asparagine-linked glycan
N-linked	Asparagine-linked
Ngl	N-linked glycosylation
NMR	nuclear magnetic resonance
O.D.	optical density (absorbance at 600 nm)
OG	n-octyl β-D-pyranoglucoside
OT	oligosaccharyl transferase
P	phosphate
*p*N-Phe	*para*-nitro phenylalanine
PP	pyrophosphate
PCR	polymerase chain reaction
PDB	Protein Database
Pfam	Protein families (database)
PglX	Protein glycosyl transferase X
PSUP	pure solvent upper phase
PyBOP	benzotriazole- l –yl-oxy-tris-pyrrolidino-phosphoniumhexafluorophosphate
SDS	sodium dodecyl sulfate
SPPS	Solid-phase peptide synthesis

SUMO	Small ubiquitin-like modifier
Tb	terbium
TFA	trifluoroacetic acid
TIS	triisopropylsilane
TMHMM	TransMembrane prediction using Hidden Markov Models
TMR	tetramethylrhodamine
TRX	Thioredoxin
TUP	theoretical upper phase
UDP-GlcNAc	uridine 5'diphospho-N-acetyl-β-D-glucosamine
UDP-Glc	uridine 5'diphospho-β-D-glucose
UMP	uridine 5'-monophosphate
Und	undecaprenyl
Xaa	used to denote any amino acid

CHAPTER 1:

INTRODUCTION

Asparagine-linked glycosylation: significance, applications, and progress

Asparagine-linked glycosylation (Ngl) is a ubiquitous, complex protein modification found in all domains of life [1,2]. While details vary with regard to function, structure, and identity of N-glycans between organisms, a general biosynthetic infrastructure is conserved (**Figures 1, 2**) [3]. A series of glycosyl transferases sequentially add specific monosaccharides to a polyprenyl-phosphate or polyprenyl-diphosphate lipid carrier. The result is a specific core oligosaccharide anchored to the membrane by a phosphate- or diphosphate-polyprenyl carrier, which serves as the donor substrate in the N-glycosylation of proteins.

The N-glycosylation reaction is catalyzed by the oligosaccharyl transferase (OTase), which transfers the core oligosaccharide to select asparagines appearing within the consensus sequence Asn-Xaa-Ser/Thr, in which Xaa represents any amino acid other than proline [4]. In the vast majority of eukaryotes, the OTase is a multi-subunit, integral-membrane complex that interacts with the translocon and transfers the core oligosaccharide to asparagines within the nascent polypeptide chain (**Figure 3**) [5-7]. In the species of Gram-negative bacteria known to possess a pathway for protein N-glycosylation, such as *Campylobacter jejuni*, the transfer reaction occurs on the periplasmic face of the inner membrane by a single-subunit, integral-membrane OTase. There is some evidence that the process occurs post-translationally and post-translocationally in bacteria, although there are still very limited studies addressing this issue with diverse proteins [8]. Nearly all eukaryotes transfer the conserved tetradecasaccharide $Glc_3Man_9GlcNAc_2$ (**Figure 4**). After transfer, the tetradecasaccharide is modified, abridged or extended in the ER and Golgi, resulting in proteins with a diversity of glycan structures. In contrast, the structure of the core oligosaccharide varies for different bacterial species, though no further modification occurs after transfer to a protein [9,10].

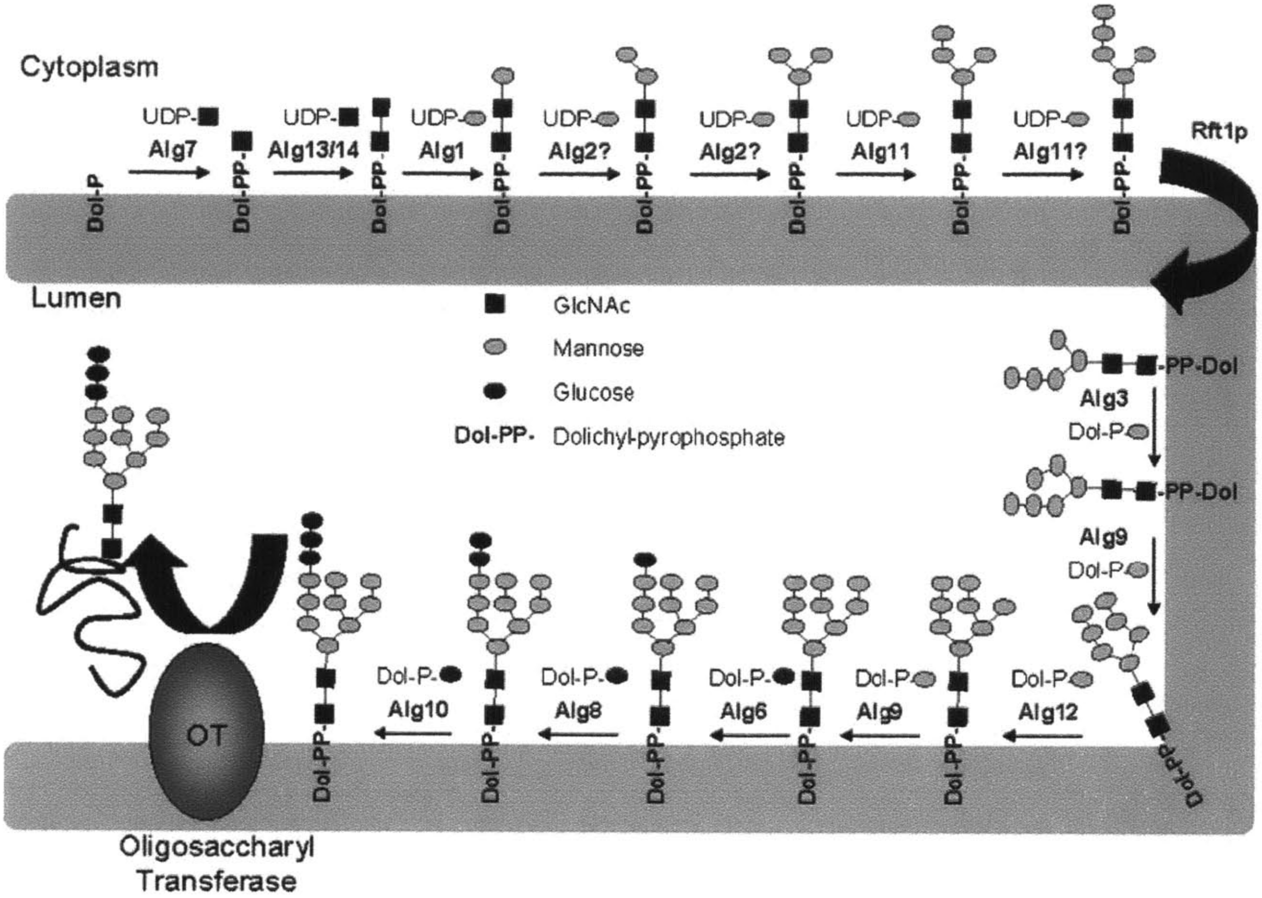

Cytoplasm
UDP-■
Alg7
UDP-■
Alg13/14
UDP-
Alg1
UDP-
Alg2?
UDP-
Alg2?
UDP-
Alg11
UDP-
Alg11?
Rft1p
Dol-P
Dol-PP-
Lumen
■ GlcNAc
Mannose
● Glucose
Dol-PP- Dolichyl-pyrophosphate
-PP-Dol
Alg3
Dol-P-
-PP-Dol
Alg9
Dol-P-
Dol-P-
Alg12
Dol-P-
Alg9
Dol-P-●
Alg6
Dol-P-●
Alg8
Dol-P-●
Alg10
Dol-PP-
OT
Oligosaccharyl Transferase

Figure 1: The N-linked glycosylation biosynthetic pathway in S. cerevisiae. Monosaccharides are sequentially added by a series of glycosyl transferases to form dolichyl-diphosphate-$Glc_3Man_9GalNAc_2$. Synthesis begins in the cytoplasm and is completed in the lumen after action of a flippase. The completed glycan is transferred by the OTase to asparagines within the consensus Asn-Xaa-Ser/Thr (Xaa ≠ Pro).

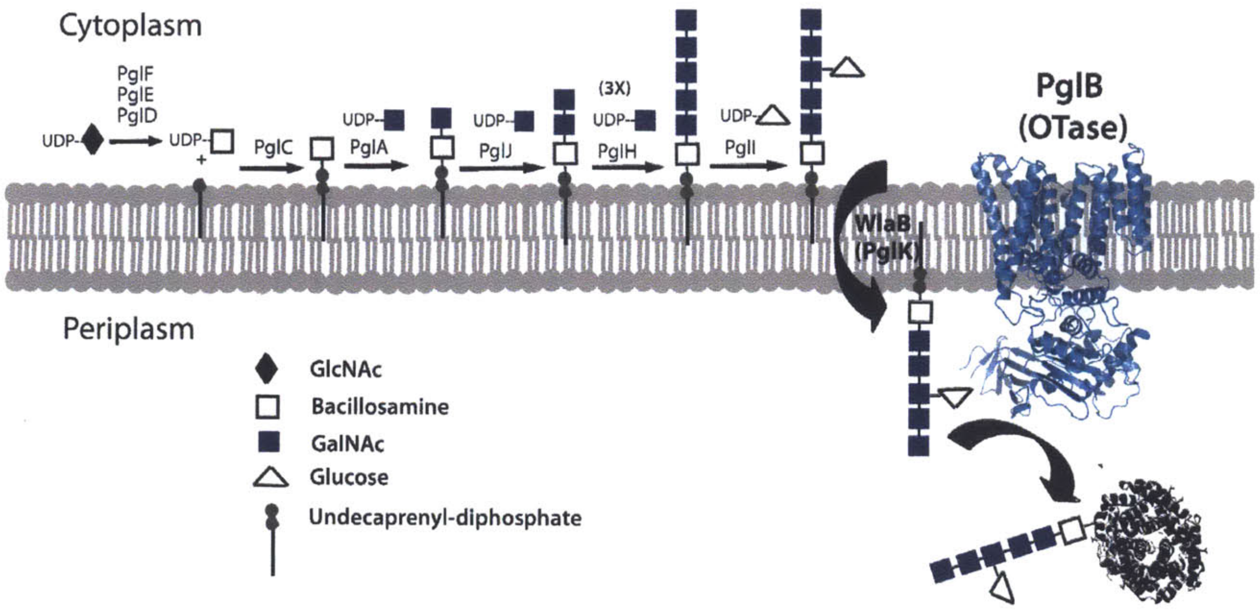

Cytoplasm
PglF
PglE
PglD
UDP-
UDP-
+
PglC
UDP-
PglA
UDP-
PglJ
(3X)
UDP-
PglH
UDP-
PglI
PglB
(OTase)
WlaB
(PglK)
Periplasm
GlcNAc
Bacillosamine
GalNAc
Glucose
Undecaprenyl-diphosphate

Figure 2: The N-linked glycosylation biosynthetic pathway in C. jejuni.
The core oligosaccharide is $GlcGalNAc_5Bac$, where Bac is Bacillosamine, or 2,4-diacetamido-2,4,6-trideoxyglucose. The polyprenyl carrier is undecaprenol.

The structural diversity of eukaryotic N-glycans accounts for the involvement of Ngl in diverse cellular functions [11]. The addition of the core oligosaccharide is a cotranslational process; thus, the addition of the chemical group affects the energy landscape of sampled polypeptide configurations and accordingly plays a major role in protein folding [9,12-14]. The presence of the glycan goes on to affect critical structural aspects of the protein, including stability, solubility, and myriad forms of interactions [12,15-19]. Not only does the glycan affect protein-protein interactions, but it is frequently a determining factor in protein trafficking and localization, signaling, and the functionality of the protein [20-26]. Ngl, which is exclusive to proteins synthesized on the ER membrane, largely affects cell-surface and secretory proteins. Consequently, Ngl is extensively associated with broad, multicellular processes, specifically the immune response, signal transduction, cell migration and cell-cell interactions [23,27-31].

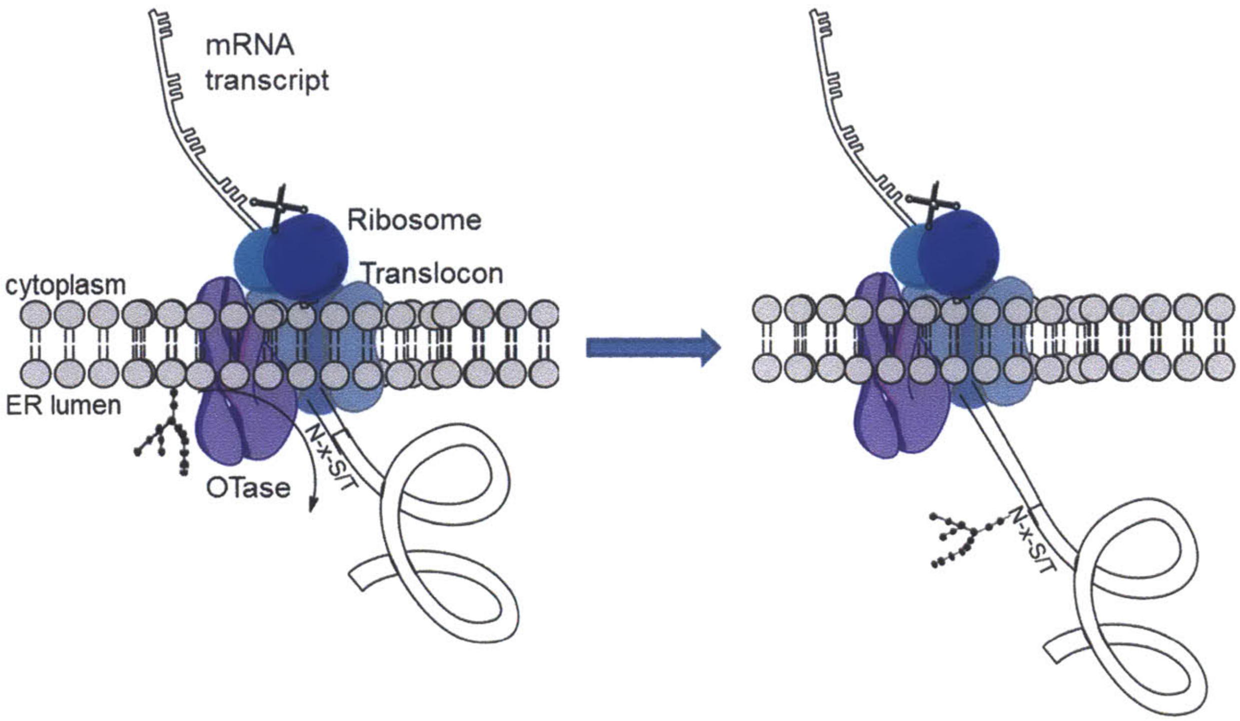


Figure 3: The cotranslational process of eukaryotic N-linked glycosylation.
The OTase is a multi-subunit membrane-bound protein complex that interacts with the translocon machinery, scans the nascent peptide chain, and transfers the glycan to select asparagines in the consensus Asn-Xaa-Ser/Thr sequence (Xaa ≠ Pro).

Ngl is of major interest to researchers due to the varied and critical role that it plays in human health and disease. The host Ngl pathway is frequently involved in the maturation and secretion of proteins made by intracellular pathogens; for example, glycosylation of viral envelope proteins allows many viruses to evade immune detection and invade new cells, with key examples including Influenza A and Human Immunodeficiency Virus (HIV) [32-36]. The ability of influenza to rapidly evolve new strains that continually challenge the immune system is often due to the alteration of the position or presence of Ngl at specific sites within viral coat proteins [35,37-41]. Similarly, Ngl of envelope proteins accounts for antibody neutralization by HIV [13,34,42,43]. The bacterial pathogen *Campylobacter jejuni*, a major cause of gastroenteritis, requires its innate Ngl pathway for effective colonization and invasion of host cells [44-46]. Proper functioning of the immune response in mammals also depends greatly on Ngl of key proteins such as IgG antibodies and the major histocompatibility complex (MHC); hence, autoimmune in addition to numerous congenital disorders have been linked to dysfunction of the Ngl pathway [47-51]. Regrettably, variability in disease phenotypes complicates diagnoses and hinders informative genetic screens, leaving many of the disorders ill-defined [52-55].

Conversely, the variable nature of Ngl makes it a useful indicator of cell state: cellular profiles and serum markers of N-glycosylation represent prospective methods for diagnosing disease states or stages of cancer progression [56-59]. Many promising methods are being evaluated for utility in better understanding correlations of Ngl with different diseases [22,60-63]. Additionally, researchers have positively used Ngl in developing and improving drug efficacy and in modifying protein properties for research purposes [15,27,64,65]. Correspondingly, considerable research is directed toward development of a system to synthesize glycosylated proteins robustly and with a high degree of control and specificity [66-69].

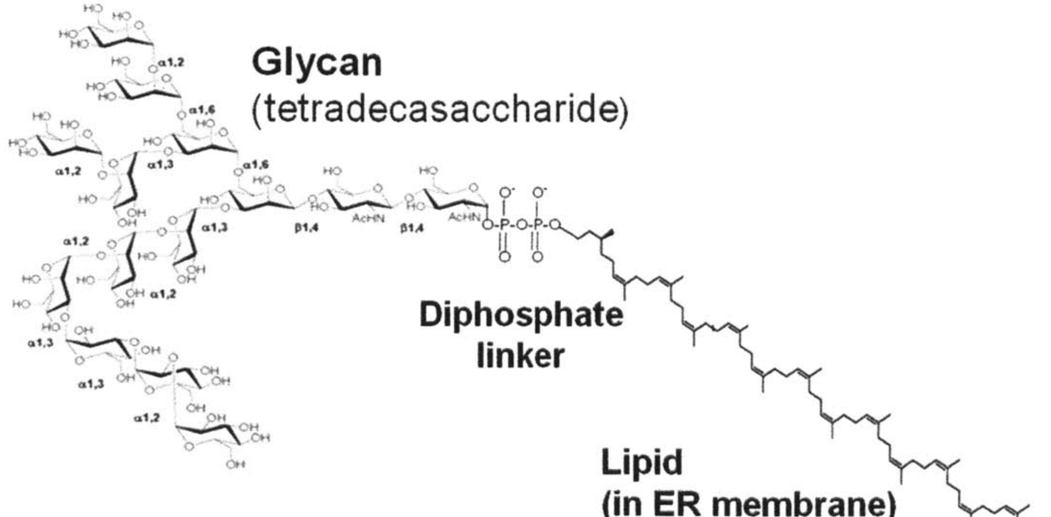
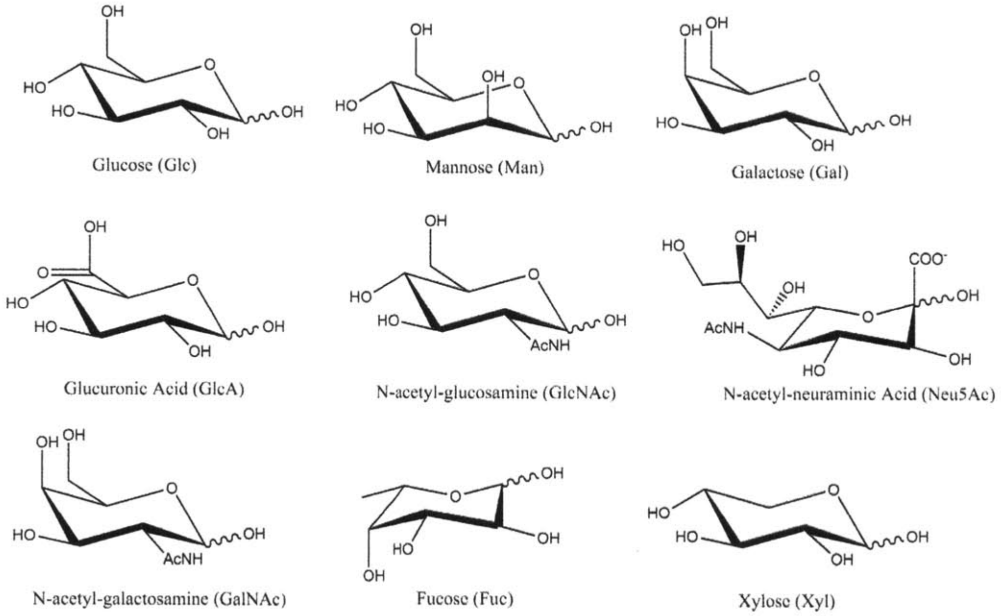


Figure 4: Structure of the polyprenyl-diphosphate-glycan substrate common in eukaryotes.

Despite the importance of Ngl in so many cellular processes, little is known about the details of the process relative to other common protein modifications such as phosphorylation, O-linked glycosylation, and ubiquitination. The factors that determine which appearances of the consensus sequon will be successfully glycosylated are not well understood, preventing the prediction of site occupancy from protein sequence. It is unknown which factors influence the nature of modifications for specific glycosylated sites on proteins. The effect of specific glycan structures on the ultimate function of the modified proteins is also unpredictable.



Glucose (Glc) Mannose (Man) Galactose (Gal)

Glucuronic Acid (GlcA) N-acetyl-glucosamine (GlcNAc) N-acetyl-neuraminic Acid (Neu5Ac)

N-acetyl-galactosamine (GalNAc) Fucose (Fuc) Xylose (Xyl)

Figure 5: Structures of biologically relevant monosaccharides in eukaryotes.

The minimal understanding of N-linked glycosylation can be explained by a number of factors. Carbohydrates are structurally challenging and diverse, and characterization of a specific oligosaccharide requires knowledge of the length, sugar composition, order, and the position and stereochemistry of each glycosidic linkage. In eukaryotes, there is a known group of monosaccharides that most commonly is associated with biological function (**Figure 5**). However, bacteria and particularly archaea are increasingly expanding this paradigm, which increases the number of possible structures that must be considered when characterizing a novel glycan [70,71]. Characterization of glycan presents many challenges due to the role of stereochemistry in identifying each sugar. Multiple monosaccharides may have the same mass and molecular formula and differ only in the stereochemistry of a single C-OH bond (such as in the case of glucose and galactose, **Figure 6**). Yet this single difference has enormous effects on determining the function and interactions of the sugar molecule. The multiple sites on a monosaccharide available for linkage to additional sugars result in an extremely high number of possible structures. The stereochemistry of each sugar-sugar linkage, along with branching locations and sugar identities, are similarly critical in determining oligosaccharide structure and function (**Figure 6**). Characterization of an unknown glycan can quickly become overwhelming upon consideration of the diversity of possible structures and the stereochemical details required to fully describe an oligosaccharide [72]. Further complexity results from the fact that a particular protein can be heterogeneously glycosylated at one or more sites.

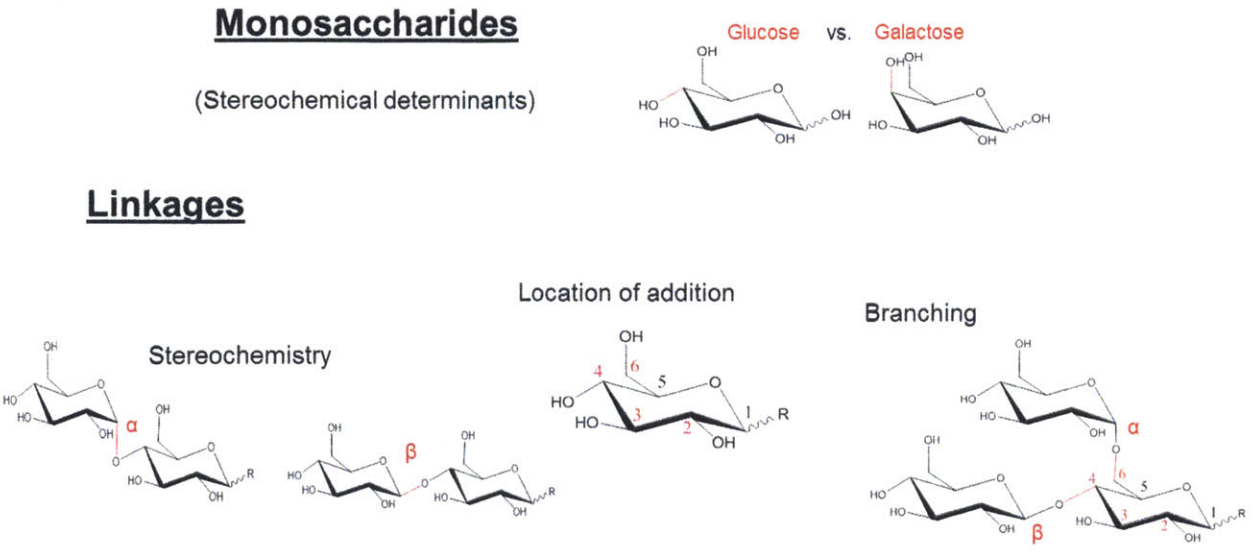


Figure 6: Structural complexity of monosaccharides and oligosaccharides. Stereochemical and isomeric properties cannot be evaluated using mass spectrometry.

The stereochemistry of a molecule cannot be ascertained using mass spectrometry, which reduces the number of methods available for identifying carbohydrates. Methods used to identify a particular oligosaccharide include NMR and enzymatic and chemical separation steps in addition to mass spectrometry. Determining the identity of an individual oligosaccharide requires direct analysis because carbohydrates are not template-driven products (as opposed to proteins and nucleic acids); therefore, direct purification and examination of material from biological sources is often required. When chemical and enzymatic steps are used, the material available for late-stage examination is reduced, which can be particularly problematic provided the quantity of material necessary for conclusive NMR analysis. Thus, low sample quantities and a shortage of sufficiently sensitive instruments have significantly stymied characterization of N-linked glycosylation.

The oligosaccharyl transferase

Despite challenges, the methods used to analyze oligosaccharides have advanced remarkably in sensitivity and reliability over the last decades. In addition, progress has been made in classifying the common Ngl structures and linking structural classes to particular functions [71,72]. Continued progress would be facilitated and accelerated greatly by successful development of Ngl inhibitors and other small molecules allowing control of enzyme activation for *in vivo* studies. A realm of prospective strategies using protein engineering has been suggested, such as modifying enzyme specificity for investigation of glycotherapeutics and enhancement of *in vitro* activity for rapid, homogenous glycosylation of proteins, which could be applied toward improving therapeutic effects of current protein therapies and optimizing those of novel drugs [73-75]. The biophysical impacts of glycosylation on proteins, such as enhancement of solubility and effects on stability and folding, also present opportunities for pursuing general strategies to control, study, and optimize proteins for a range of applications [74].

However, these multifold uses require the ability to produce homogenous and efficient protein glycosylation, which necessitates an understanding of the enzymes that accomplish the tasks in nature. In addition to the intrinsic complexity of oligosaccharides described above, additional impediments to Ngl research result from the nature of the involved enzymes. Both biosynthetic and modifying Ngl enzymes are diverse, numerous, and frequently membrane-associated, making isolation and characterization challenging. Consequently, structural and mechanistic characterization of key enzymes remains elusive.

In 1999, the first bacterial system of N-linked glycosylation was discovered in the enteropathogen *Campylobacter jejuni*. Prior to this discovery, the process was thought to exist exclusively in eukaryotes and archaea [76]. The presence of an N-linked glycosylation pathway represented a breakthrough in the field of Ngl research; the evolutionary conservation further validates the importance of the modification throughout all domains of life, and the existence of a bacterial system has potential to function as a simplified model to study. Since then, the N-linked glycosylation pathway in *C. jejuni* has successfully been functionally transferred to *E. coli*, characterized and established as a relatively accessible model system of Ngl [77-84]. A comparison of the pathways depicted in **Figures 1** and **Figure 2** reveals fundamental similarities, though also apparent is the increased complexity of the Ngl system in *S. cerevisiae* relative to that in *C. jejuni*.

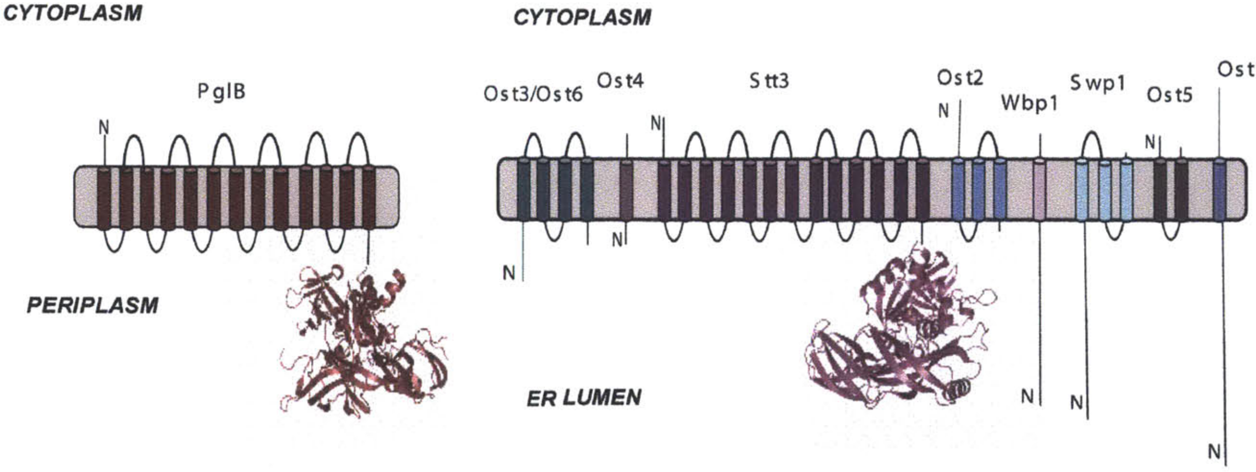


Figure 7: Comparison of the OTases in *C. jejuni* (bacteria) and *S. cerevisiae* (eukaryotes). Images highlight the similarity between the *S. cerevisiae* catalytic subunit 'STT3' and PglB.

Figure 7 compares the eukaryotic and the prokaryotic OTases. In *C. jejuni,* the OTase, PglB, is a single-subunit enzyme, while the eukaryotic OTase is composed of at least eight subunits, all of which are membrane-bound and five of which are required for function. Importantly, the catalytic subunit of the eukaryotic OTase, Stt3, is homologous to PglB. As described in the following chapters, homology between all Stt3 subunits and monomeric OTases is far more extensive than originally imagined [85,86]. Additionally, the fundamental chemical nature of the transfer reaction is conserved (**Figure 8**). However, the substrates differ; the heptasaccharide in *C. jejuni* is simpler than the eukaryotic tetradecasaccharide (**Figures 4, 8**). Thus, PglB presents an exceptional opportunity for learning about the fundamental biochemistry involved in asparagine glycosylation, as well as studying the effects of Ngl in bacteria and as a tool for protein engineering and high-level N-glycoprotein synthesis [67,87].

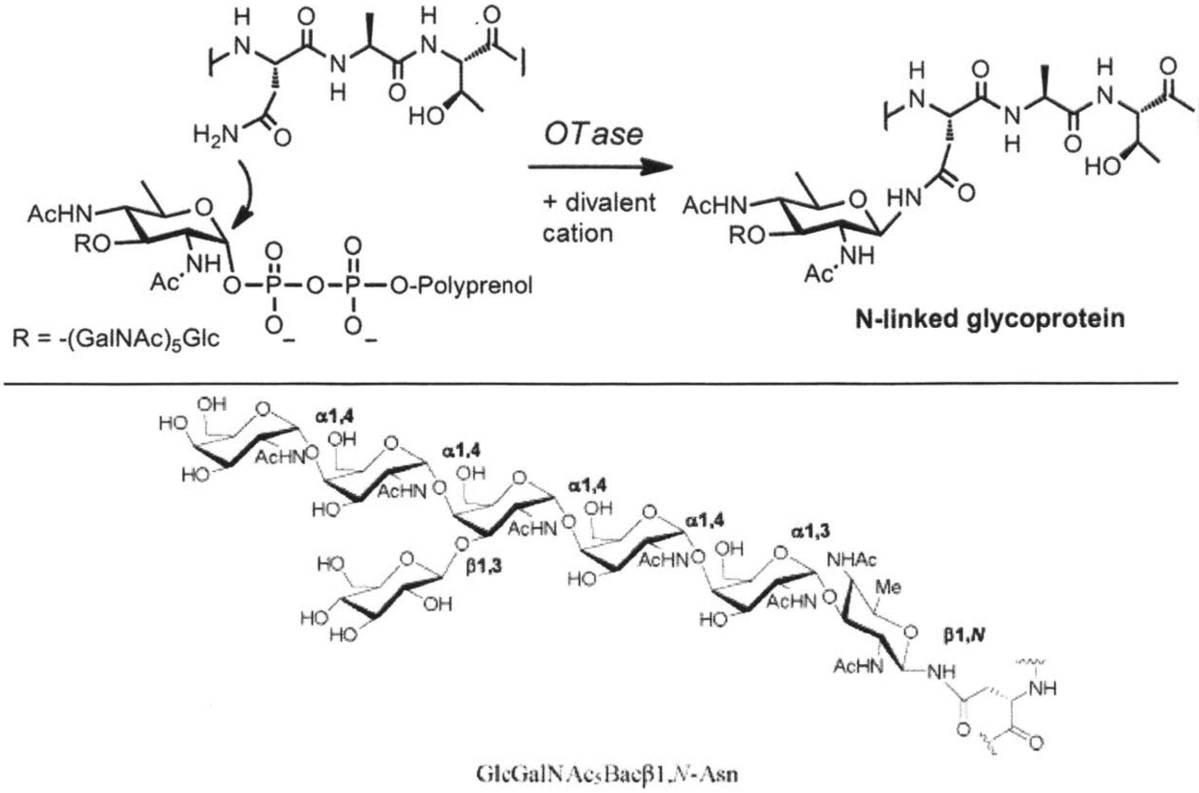


Figure 8: N-linked glycosylation transfer reaction in *C. jejuni.*
Top: In *C. jejuni*, PglB transfers the heptasaccharide $GlcGalNAc_5Bac$ (Bac is di-N-acetyl-bacillosamine or 2,4-diacetamido-2,4,6-trideoxyglucose) to the polyprenyl lipid. The eukaryotic transfer reaction is similar except the Bac sugar is GlcNAc and the R-group is –$GlcNAcMan_9Glc_3$ (**Figure 4**). *Bottom*: Structure of the full heptasaccharide from *C. jejuni*, shown linked to the nitrogen on the asparagine side chain.

The recent structural and biochemical data published on PglB show the motifs responsible for catalysis are conserved throughout all domains of life, solidifying its role as an important and general mechanistic model for N-linked glycosylation (**Figure 9**) [85,86]. The publication of the crystal structure of a close homolog of PglB from *Campylobacter lari* represents a reliable model for PglB in *C. jejuni*, as the protein sequences are highly similar (**Figure 10**) [86]. The crystal structure is resolved to 3.5 Å and shows the placement of the major conserved loop motifs in the vicinity of one another, forming a presumed active site (**Figure 9**) [86].

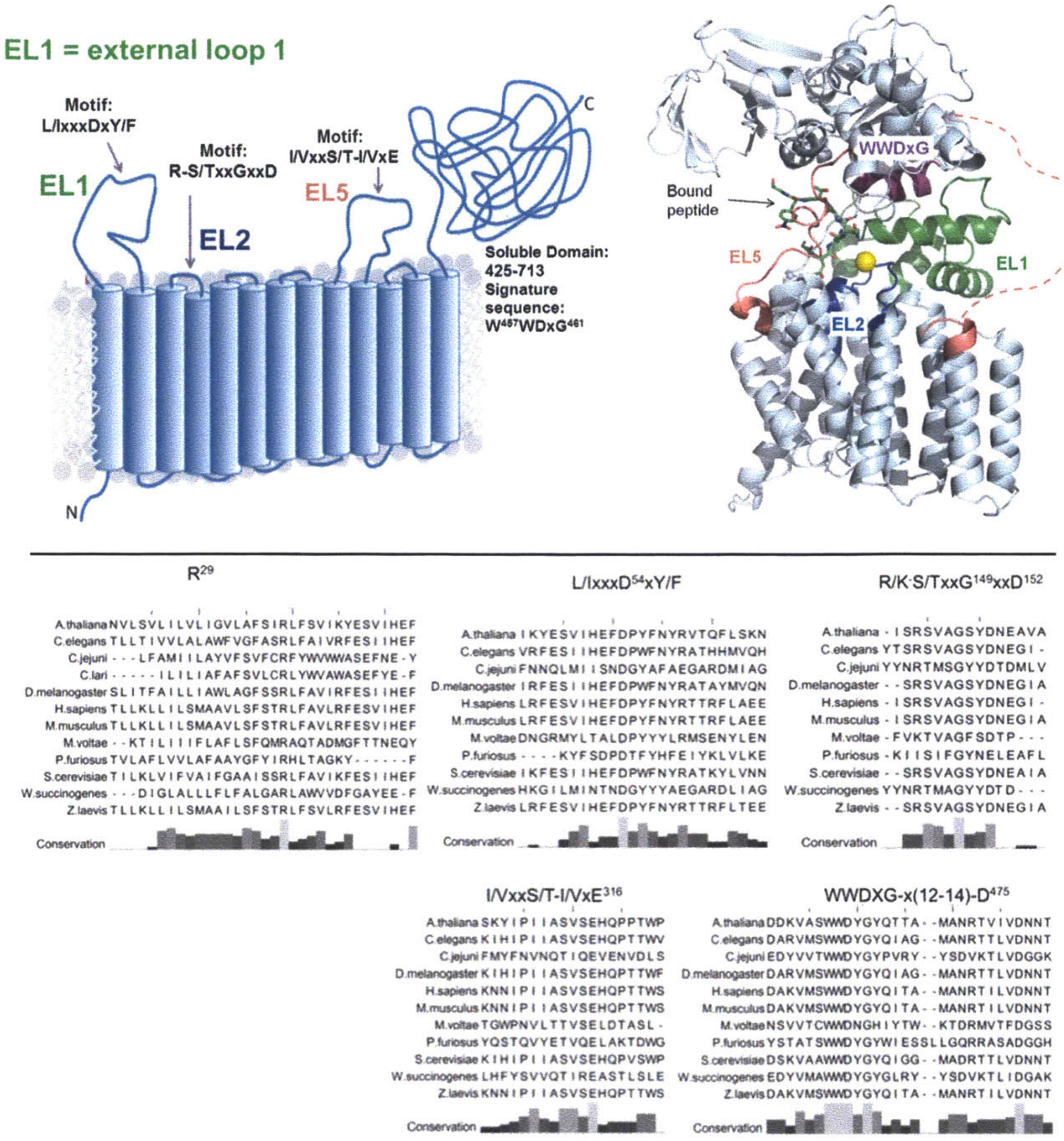

EL1 = external loop 1
Motif: L/IxxxDxY/F
Motif: R-S/TxxGxxD
Motif: I/VxxS/T-I/VxE
EL1
EL2
EL5
C
N
Soluble Domain: 425-713
Signature sequence: W457WDxG461
WWDxG
Bound peptide
EL5
EL1
EL2
R29
A.thaliana NVLSVLILVLIGVLAFSIRLFSVIKYESVIHEF
C.elegans TLLTIVVLALAWFVGFASRLFAIVRFESIIHEF
C.jejuni ---LFAMIILAYVFSVFCRFYWVWWASEFNE-Y
C.lari -----ILILIAFAFSVLCRLYWVAWASEFYE-F
D.melanogaster SLITFAILLIAWLAGFSSRLFAVIRFESIIHEF
H.sapiens TLLKLLILSMAAVLSFSTRLFAVLRFESVIHEF
M.musculus TLLKLLILSMAAVLSFSTRLFAVLRFESVIHEF
M.voltae --KTILIIIFLAFLSFQMRAQTADMGFTTNEQY
P.furiosus TVLAFLVVLAFAAYGFYIRHLTAGKY------F
S.cerevisiae TILKLVIFVAIFGAAISSRLFAVIKFESIIHEF
W.succinogenes ---DIGLALLLFLFALGARLAWVVDFGAYEE-F
Z.laevis TLLKLLILSMAAILSFSTRLFSVLRFESVIHEF
Conservation
L/IxxxD54xY/F
A.thaliana IKYESVIHEFDPYFNYRVTQFLSKN
C.elegans VRFESIIHEFDPWFNYRATHHMVQH
C.jejuni FNNQLMIISNDGYAFAEGARDMIAG
D.melanogaster IRFESIIHEFDPWFNYRATAYMVQN
H.sapiens LRFESVIHEFDPYFNYRTTRFLAEE
M.musculus LRFESVIHEFDPYFNYRTTRFLAEE
M.voltae DNGRMYLTALDPYYYLRMSENYLEN
P.furiosus ----KYFSDPDTFYHFEIYKLVLKE
S.cerevisiae IKFESIIHEFDPWFNYRATKYLVNN
W.succinogenes HKGILMINTNDGYYYAEGARDLIAG
Z.laevis LRFESVIHEFDPYFNYRTTRFLTEE
Conservation
R/K-S/TxxG149xxD152
A.thaliana -ISRSVAGSYDNEAVA
C.elegans YTSRSVAGSYDNEGI-
C.jejuni YYNRTMSGYYDTDMLV
D.melanogaster --SRSVAGSYDNEGIA
H.sapiens -ISRSVAGSYDNEGI-
M.musculus -ISRSVAGSYDNEGIA
M.voltae -FVKTVAGFSDTP---
P.furiosus -KIISIFGYNELEAFL
S.cerevisiae --SRSVAGSYDNEAIA
W.succinogenes YYNRTMAGYYDTD---
Z.laevis --SRSVAGSYDNEGIA
Conservation
I/VxxS/T-I/VxE316
A.thaliana SKYIPIIASVSEHQPPTWP
C.elegans KIHIPIIASVSEHQPTTWV
C.jejuni FMYFNVNQTIQEVENVDLS
D.melanogaster KIHIPIIASVSEHQPTTWF
H.sapiens KNNIPIIASVSEHQPTTWS
M.musculus KNNIPIIASVSEHQPTTWS
M.voltae TGWPNVLTTVSELDTASL-
P.furiosus YQSTQVYETVQELAKTDWG
S.cerevisiae KIHIPIIASVSEHQPVSWP
W.succinogenes LHFYSVVQTIREASTLSLE
Z.laevis KNNIPIIASVSEHQPTTWS
Conservation
WWDXG-x(12-14)-D475
A.thaliana DDKVASWWDYGYQTTA--MANRTVIVDNNT
C.elegans DARVMSWWDYGYQIAG--MANRTTLVDNNT
C.jejuni EDYVVTWWDYGYPVRY--YSDVKTLVDGGK
D.melanogaster DARVMSWWDYGYQIAG--MANRTTLVDNNT
H.sapiens DAKVMSWWDYGYQITA--MANRTILVDNNT
M.musculus DAKVMSWWDYGYQITA--MANRTILVDNNT
M.voltae NSVVTCWWDNGHIYTW--KTDRMVTFDGSS
P.furiosus YSTATSWWDYGYWIESSLLGQRRASADGGH
S.cerevisiae DSKVAAWWDYGYQIGG--MADRTTLVDNNT
W.succinogenes EDYVMAWWDYGYGLRY--YSDVKTLIDGAK
Z.laevis DAKVMSWWDYGYQITA--MANRTILVDNNT
Conservation

Figure 9: Topology model of PglB agrees with X-ray structure data.

Top. Topology model of PglB showing locations of conserved motifs. Model is based on a combination of topology prediction and conservation of loop regions and agrees with the *C. lari* crystal structure (shown right, PDB:3RCE). The N-terminus is located in the cytoplasm *Bottom*. Alignments of conserved residues and motifs that informed the topology model above. Alignments include a selection of representative sequences. The corresponding residue number in PglB (*C. jejuni*) is designated in superscript. Alignments and conservation histograms shown were made using Jalview multiple alignment editor. (doi: 10.1093/bioinformatics/btp033). Histogram shading: light to dark designates most to least conserved.

However, the structure provides very little conclusive information on the mechanism of the OTase. Unfortunately, *in vitro* activity is not shown for the purified enzyme and it is unclear whether the structure represents an active conformation. Additionally, the 3.5- Å resolution leaves uncertainty in the positioning of the side chains and potential hydrogen bonding networks. At this resolution, it is impossible to differentiate between the nitrogen and the oxygen on the side chain of the nucleophilic asparagine. The structural data also lacks one of the major loops expected to be involved in catalysis (designated EL5, containing the I/V-xx-S/T-I/V-E motif). The loop is unresolved due to flexibility and mobility of the region; though the conservation of its length and amino-acid content implies this very mobility may implicate the loop in a conformational change of the enzyme upon binding or release of the substrate.

The conditions of the crystallization provide further reasons to interpret the structural data with skepticism; the pH of crystallization is 9.4, which is not representative of the state of the active enzyme *in vivo* and increases uncertainty with respect to the charge of key amino acids and thus the ability of the enzyme to adopt an active conformation or participate in necessary acid-base-type chemistry. Like many membrane-protein crystal structures, the structure of PglB was determined in the absence of its native lipid environment. The detergent micelles surrounding the PglB monomers bear little resemblance to a membrane bilayer, casting further doubt on whether the structure represents the active enzyme.

The peptide shown bound in the PglB crystal structure has been assumed to represent a picture of the native binding behavior of PglB to its substrate, and mechanistic arguments were made based on this assumption [86]. However, there are multiple reasons to call this assumption into question. First, the structure is solved for only a single conformation in the presence of substrate peptide, yet it remains unknown in which order the peptide and oligosaccharide substrates bind, and thus it is uncertain whether the conformation and interactions are representative of the active state. As noted above, the activity of the crystallized enzyme fraction has not been shown, calling into question the biological relevance of the enzyme-substrate interaction [86].

It is also important to note that N-glycosylation in *C. jejuni* has not been determined as occurring co- or post-translocationally. The system in *S. cerevisiae* is co-translational and co-translocational; the OTase closely interacts with the translocation machinery, which feeds to the OTase the nascent peptide chain. N-glycosylation by PglB has shown capability *in vitro* of glycosylating exposed, flexible loops on fully folded proteins, which has led to a hypothesis that PglB glycosylates proteins post-translocationally and after periplasmic folding [8]. Nonetheless, evidence beyond *in vitro* capability is lacking; thus, it is unclear whether PglB is associated with translocational or other cellular machinery when glycosylating proteins and whether that interaction affects binding of the substrate. Therefore, the nature of the peptide-substrate bound in the crystal structure of PglB is further drawn into question.

```
C. jejuni       -MLKKEYLKNP-YLVLFAMIILAYVFSVFCRFYWVWWASEFNEYFFNNQLMIISNDGYAF 58
C. lari         MKLQQNFTDNNSIKYTCILILIAFAFSVLCRLYWVAWASEFYEFFFNDQLMITTNDGYAF 60
                  *:::: .*        :*::*:.***:**:*** ***** *:***:**** :******

C. jejuni       AEGARDMIAGFHQPNDLSYYGSSLSALTYWLYKITPFSFESIILYMSTFLSSLVVIPTIL 118
C. lari         AEGARDMIAGFHQPNDLSYFGSSLSTLTYWLYSILPFSFESIILYMSAFFASLIVVPIIL 120
                *******************:*****:******.* ************:*::**:*:* **

C. jejuni       LANEYKRPLMGFVAALLASIANSYYNRTMSGYYDTDMLVIVLPMFILFFMVRMILKKDFF 178
C. lari         IAREYKLTTYGFIAALLGSIANSYYNRTMSGYYDTDMLVLVLPMLILLTFIRLTINKDIF 180
                :*.***    **:****.*********************:****:**: ::*: ::**:*

C. jejuni       SLIALPLFIGIYLWWYPSSYTLNVALIGLFLIYTLIFHRKEKIFYIAVILSSLTLSNIAW 238
C. lari         TLLLSPVFIMIYLWWYPSSYSLNFAMIGLFGLYTLVFHRKEKIFYLTIALMIIALSMLAW 240
                :*:  *:** **********:**.*:**** :***:*********::: *  ::** :**

C. jejuni       FYQSAIIVILFALFALEQKRLNFMIIGILGSATLIFLILSGGVDPILYQLKFYIFRSDES 298
C. lari         QYKLALIVLLFAIFAFKEEKINFYMIWALIFISILILHLSGGLDPVLYQLKFYVFKASDV 300
                 *: *:**:***:**::::::** :*  *   ::::* ****:**:*******:*::.:

C. jejuni       ANLT-QGFMYFNVNQTIQEVENVDLSEFMRRISGSEIVFLFSLFGFVWLLRKHKSMIMAL 357
C. lari         QNLKDAAFMYFNVNETIMEVNTIDPEVFMQRISSSVLVFILSFIGFILLCKDHKSMLLAL 360
                 **.  .*******:** **:.:* . **:***.* :**::*::**: * :.****::**

C. jejuni       PILVLGFLALKGGLRFTIYSVPVMALGFGFLLSEFKAIMVKKYSQL---TSNVCIVFATI 414
C. lari         PMLALGFMALRAGLRFTIYAVPVMALGFGYFLYAFFNFLEKKQIKLSLRNKNILLILIAF 420
                *:*.***:**:.*******:*********::*  *  :: **  :*   ..*: ::: ::

C. jejuni       LTLAPVFIHIYNYKAPTVFSQNEASLLNQLKNIANREDYVVTWWDYGYPVRYYSDVKTLV 474
C. lari         FSISPALMHIYYYKSSTVFTSYEASILNDLKNKAQREDYVVAWWDYGYPIRYYSDVKTLI 480
                ::::*.::*** **: ***:. ***:**:*** *:******:*******:*********:

C. jejuni       DGGKHLGKDNFFPSFALSKDEQAAANMARLSVEYTEKSFYAPQNDILKTDILQAMMKDYN 534
C. lari         DGGKHLGKDNFFSSFVLSKEQIPAANMARLSVEYTEKSFKENYP-----DVLKAMVKDYN 535
                ************ **.***::  ****************          *:*:**:****

C. jejuni       QSNVDLFLASLSKPDFKIDTPKTRDIYLYMPARMSLIFSTVASFSFINLDTGVLDKPFTF 594
C. lari         KTSAKDFLESLNDKDFKFDTNKTRDVYIYMPYRMLRIMPVVAQFANTNPDNGEQEKSLFF 595
                ::... ** **.. ***:** ****:*:*** **  *: .**.*:  * *.*  :* : *

C. jejuni       STAY--PLDVKNGEIYLSNGVVLSDDFRSFKIGDNVVSVNSIVEINSIKQGEYKITPIDD 652
C. lari         SQANAIAQDKTTGSVMLDNGVEIINDFRALKVEGASIPLKAFVDIESITNGKFYYNEIDS 655
                * *     * ..*.: *.*** : :***::*:    : ::::*:*:**.:*::  . **.

C. jejuni       KAQFYIFYLKDSAIPYAQFILMDKTMFNSAYVQMFFLGNYDKNLFDLVINSRDAKVFKLK 712
C. lari         KAQIYLLFLRE----YKSFVILDESLYNSSYIQMFLLNQYDQDLFEQITNDTRAKIYRLK 711
                ***:*:::*::    * .*:::*:::::**:*:***:* :**::**: : *.  **:::**
```

Figure 10: Alignment of PglB sequences from *C. jejuni* and *C. lari*.
Alignment constructed using ClustalW. Gene identifiers of sequences: *C. lari* gi:7410986; *C. jejuni* gi:218562740

The peptidyl substrate of the oligosaccharyl transferase

As discussed above, the X-ray crystal structure of the OTase in the presence of bound substrate peptide (Ac-Gly-Asp-Gln-Asn-Ala-Thr-Gly-*p*N-Phe, where *p*N-Phe is *para*-nitro phenylalanine, [83]) represents a remarkable achievement, though also leaves significant uncertainty with regard to speculative mechanistic conclusions [86]. Particularly, an ongoing question has revolved around the function of the consensus sequence of the peptide substrate, Asn-Xaa-Ser/Thr, in which Xaa can be any amino acid other than proline.

Substantial biochemical evidence exists to support the formation of a specific conformation of the consensus sequence. **Figure 11** shows the possible local conformations that a peptide containing the minimal consensus sequence may adopt [88]. The Asx-turn involves hydrogen bonding between the carbonyl oxygen of the side-chain of asparagine and the proton donors on the hydroxyl side chain and backbone amide of the amino acid at the +2 site. The β-turn, conversely, does not include interactions between the asparagine side chain and other chemical groups within the consensus sequon, and intramolecular interactions consist largely of backbone hydrogen bonds.

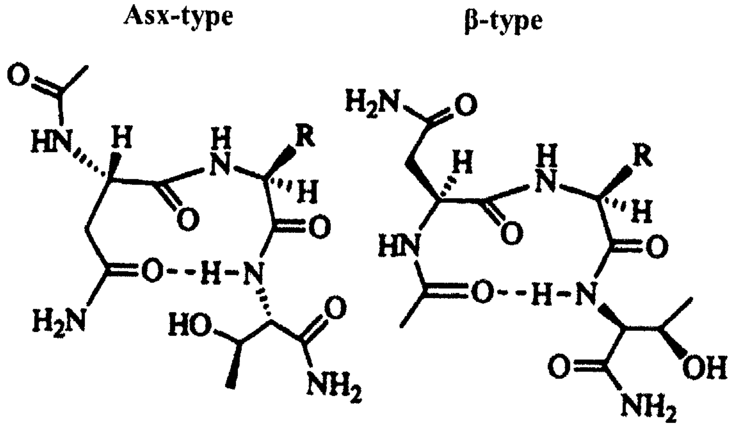


Figure 11: Comparison of the Asx-type and β-type turn peptide Ac-Asn-Xaa-Thr-NH_2. Figure adapted from [89].

Several mechanisms have been proposed that implicate the Asn-Xaa-Ser/Thr consensus sequon in the activation of the asparagine side-chain nitrogen, which is usually a very poor nucleophile. A proposal by Marshall suggests that the hydroxyl group of the Ser/Thr residue engages in a hydrogen bond with the carbonyl oxygen of the asparagine side chain, which promotes deprotonation of the amide nitrogen and subsequent nucleophilic attack (**Figure 12-A**) [90]. Bause proposes that the amide nitrogen of asparagine engages in a hydrogen bond with the hydroxyl group of the Ser/Thr residue, facilitating deprotonation of the amide nitrogen upon attack of the hydroxyl proton by a general base (**Figure 12-B**) [91]. A third mechanism has been formally proposed by Imperiali, which implicates the formation of the Asx-type conformation in the activation of the amide nitrogen (**Figure 12-C**) [92,93]. In this scheme, the carbonyl on the Asn side chain is engaged in hydrogen bonding with protons on the hydroxyl side chain in the +2 position and a backbone nitrogen amide proton, which shifts the electron density away from the nitrogen toward the carbonyl group. The pKa of the nitrogen protons is therefore lowered, facilitating removal of a proton by a basic group and freeing a nitrogen electron pair for subsequent attack of the electron-poor phosphate-linked carbon on the first sugar of the oligosaccharide [92,93].

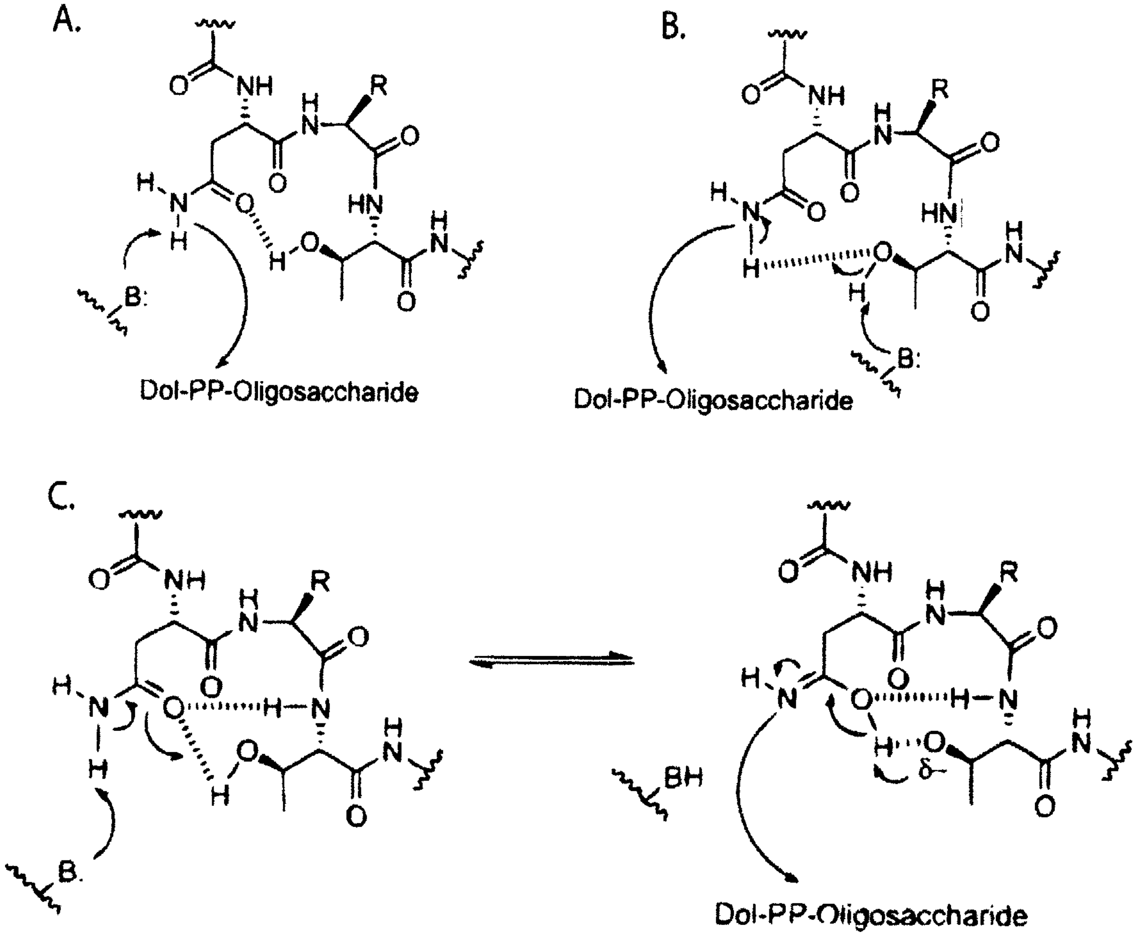
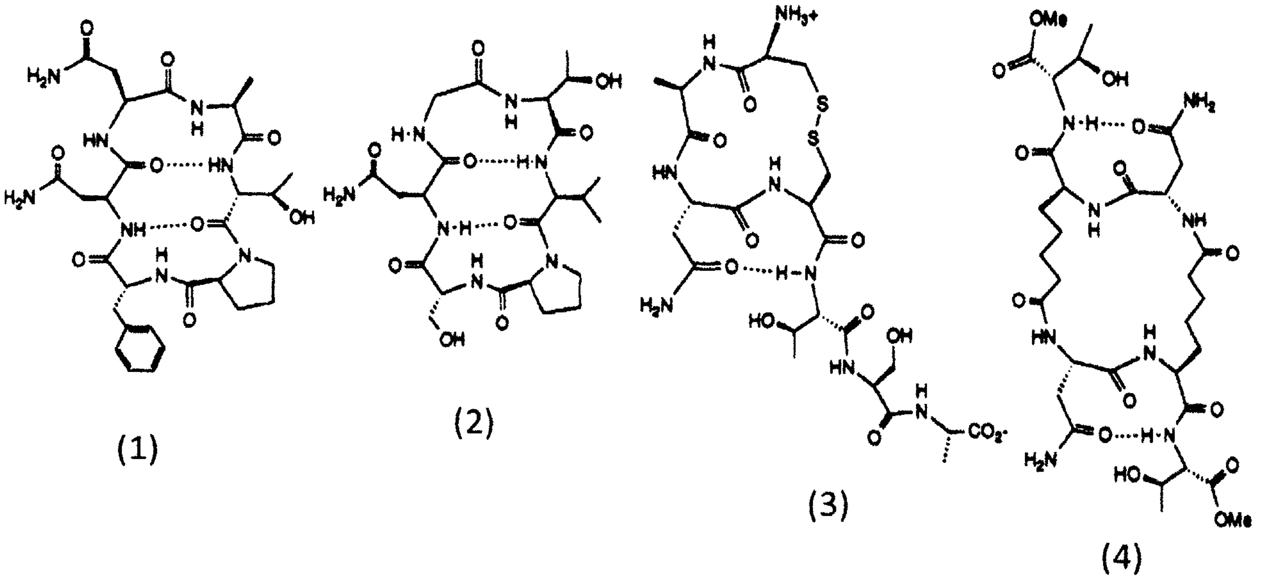


Figure 12: Proposed OTase mechanisms. A) Marshall. B) Bause. C) Imperiali.

Reasons to favor the mechanism proposed by Imperiali include the substantial evidence that the peptidyl substrate adopts the Asx-type conformation when binding the OTase. Importantly, *in vitro* characterization of tetrapeptides has shown that the β-type turn is favored and the Asx-type turn conformation is structurally prohibited when a proline is present at the *i*+2 position [94]. This structural requirement accounts for the fact that the OTase can turn over the consensus sequence peptide with any amino acid at this position except proline [4]. The interactions between the side chains of the two consensus residues in the sequon also provide an explanation of the strict consensus requirement. Notably, the involvement in catalysis explains absolute requirement for asparagine over glutamine. Given the chemical similarity of the amino acids, it is striking that glutamine is not accepted by the OTase even as a weak substrate [83]. The central role of the Asx-turn in catalysis would require the precise alignment of chemical groups to achieve the activation of the poorly nucleophilic nitrogen, thus explaining this rigid specificity. The catalytic role is also highly plausible given the fact that the consensus sequence is conserved in all systems of N-linked glycosylation, as it is far more likely that an enzyme evolves the specificity of binding interactions over time than the chemical mechanism.

In addition to these observations, rigorous biochemical studies have been performed that support the suggested mechanism (**Figure 12-C**) [93]. Particularly, the use of synthetic peptide chemistry in the creation of constrained peptides provides a convincing method for testing the Asx-turn mechanistic proposal [88,89,92,94]. **Figure 13** shows four peptides synthesized in a fashion that structurally bias the consensus sequon into formation of a β-type or an Asx-type turn conformation. Two β-type peptides were synthesized using the prolyl-D-amino acid dipeptide to circularize and constrain the structure (**Figure 13**, structures **1** and **2**). NMR and CD studies were performed to support the presence of the predicted conformations [88,89,92,94]. Two additional peptides were synthesized and constrained via oxidation of thiol groups (**Figure 13**, structures **3** and **4**) [92]. Structural studies confirmed the predicted Asx–type conformation for **4**. Peptide **3** is not highly constrained to form the Asx-turn; however, it is incapable of forming the β-type turn and due to its flexible nature the peptide is expected to be capable of adopting the Asx-turn [92].



(1) (2) (3) (4)

Figure 13: Structures of peptides constrained to bias formation of Asx-type or β-type turn. Peptides **1** and **2** are shown in the β-type conformation. Peptides 3 and 4 are shown in the Asx-conformation, which is stabilized by the constrained structure of the peptide. The predicted conformations of peptides **1**, **2**, and **4** were validated using NMR and CD. Peptide 3 is not fixed as an Asx-type turn; however, the peptide is restricted from formation of the alternate β-type conformation and the flexibility of the peptide suggests that it is capable of adopting the Asx-type conformation. Kinetic results show that **1** and **2** are not substrates for the OTase, while **3** and **4** are substrates and show similar kinetic parameters to control substrate peptides containing the consensus motif [92].

The constrained peptides shown in **Figure 13** were subjected to kinetic analysis. Activity studies showed that **1** and **2** are not accepted as substrates by the OTase, while **3** and **4** have kinetic parameters similar to standard substrate peptides [92]. The results described above are supplemented by kinetic studies showing potent inhibition by non-substrate peptides adopting the Asx-turn type conformation [95]. Additional substrate-specificity studies suggest the absence of a formal negative charge occurring during the catalytic reaction, which is supported by the mechanism proposed in **Figure 12-C** [93].

The experiments discussed above provide convincing evidence of the role of the Asx-turn in the OTase catalytic mechanism. Nonetheless, publication of the *C. lari* crystal structure shows the bound peptide in a conformation dissimilar to the Asx-turn, and the authors conclude that the mechanism does not involve adoption of the Asx-turn. Instead, the consensus Asn-Xaa-Ser/Thr consensus sequence is explained by the binding interactions of the peptide with the conserved Trp-Trp-Asp-Xaa-Gly (WWDxG) motif [86].

An argument posed against the Asx-turn requirement for OTase catalysis include the existence of exceptions to the consensus rule, in which non-canonical consensus sites are successfully glycosylated [2]. There have been several anecdotal examples of non-canonical sequons in specific proteins that, when purified from cells, are shown to have been glycosylated [96-98]. Other than anecdotal evidence, very few large-scale studies have been performed to show the frequency of non-canonical glycosylation sites. One notable exception involved a large-scale precision mapping of N-glycosylated sites present in the *in vivo* glycoproteome in mouse tissues, which included thorough analysis of the false discovery rate and confidence levels for identified sites [99]. This study showed that the vast majority (> 96%) of glycosylated asparagines are followed by a serine or threonine in the +2 position. Strikingly, not one of these sites contains a proline in the +1 position. The 3.5% of non-canonical sites were dominated by Asn-Xaa-Cys (Xaa ≠ Pro) , which account for 1.3% of the total glycosylated sites and agrees with anecdotal studies showing asparagines within Asn-Xaa-Cys can be glycosylated [100]. In this site, it is easy to see that the cysteine resembles a serine or threonine in the length of the non-carbon atom from the backbone and harboring an accessible proton. Of the remaining 2.2% glycosylated asparagines, two motifs could be identified with statistical significance, with Asn-Gly accounting for 0.5% and Asn-Xaa-Val/Leu accounting for 0.4% of the total glycosylated asparagine residues. The Asn-Xaa-Val motif also agrees with an anecdotal study showing the ability of a valine or leucine to substitute for the serine of threonine in the canonical Asn-Xaa-Ser/Thr sequon [98].

Yet, the presence of non-canonical glycosylation sites is consistently an exception to the rule, and organisms from bacteria through archaeal and up through mammals maintain the nearly exclusive preference for the canonical Asn-Xaa-Ser/Thr. The studies showing glycosylation of non-canonical sequons lack the crucial evidential component of performing the reaction *in vitro* using purified enzymes and substrates. Showing the reaction occurring *in vitro* will be necessary to determine which specific sites can be glycosylated by the unaccompanied enzyme. However, the evidence for some non-canonical sites, such as Asn-Gly and Asn-Xaa-Val/Leu, appears compelling [99]. It is not clear that the infrequent nature of these occurrences contradicts the role of the Asx-type turn in OTase catalysis. It can be imagined that atypical sequons can acquire Asx-type turns with hydrogen bonding supplied by the peptide backbone or neighboring residues. Most importantly, no examples have been seen of glycosylated canonical sequons when a proline is present at the +1 site. Evidence of OTase glycosylation of a peptide constrained in a way that disallows formation of the Asx-type turn would provide the strongest challenge to the proposed role of the Asx-type conformation in the OTase mechanism.

The recent discovery of an unusual N-linked glycosylation system may provide the most convincing evidence yet for the fundamental role of the Asn-Xaa-Ser/Thr consensus in catalysis of N-linked glycosylation [101]. HMW1C is a soluble N-glycosyl transferase from *Haemophilus influenzae.* The enzyme glycosylates two adhesin proteins, which mediate adherence to epithelial cells and thereby facilitate pathogenesis of *H. influenzae* [102]. HMW1C functions in the cytoplasm, adding glucose monosaccharides from nucleotide-activated precursors to asparagine residues [101]. The localization in the cytoplasm, the nucleotide-activated monosaccharide substrate, and the soluble structure defy a whole set of qualities that were thought to be common to all oligosaccharyl transferases. The sequence motifs and structural features that are conserved in other OTases are all lacking in HMW1C. Structurally, HMW1C is considered a part of the GT41 family of glycosyl transferases according to CAZy (Carbohydrate Active Enzymes database at www.cazy.org). The GT41 family is otherwise composed entirely of *O*-GlcNAc transferases (OGTs). Like other OGTs, HMW1C contains an N-terminal all α-domain fold and a C-terminal GT-B fold with two Rossmann-like domains. However, HMW1C lacks the tetratricopeptide repeat fold at the N-terminus that is characteristic of the rest of the GT41 family [103].

The single conserved feature in the HMW1C N-glycosylation system is the presence of the Asn-Xaa-Ser/Thr consensus sequence determining the sites of glycosylation [101]. The evolutionary background shared with GT41 OGTs clearly indicates N-glycosylation function of HMW1C is a result of convergent evolution. The fact that the independent evolution of asparagine glycosylation was accompanied by the specificity for the Asn-Xaa-Ser/Thr sequon is a striking indication of the fundamental mechanistic role that the sequon is playing in catalysis—specifically, the precise combination of residues with correct order and spacing and thus conformation that allows vastly different proteins to activate the poorly nucleophilic asparagine nitrogen for attack.

Conclusion

The significance of N-linked glycosylation to human health has been well-established, with N-glycosylation demonstrating deterministic role in signal transduction, protein-protein and cell-cell interactions, protein folding and targeting, and microbial immune evasion. The chemical complexity of the modification and the structural complexity of pathway enzymes have stymied efforts to understand the process in greater detail. The discovery of a bacterial pathway of N-glycosylation has provided a convenient model system that is relatively simple to characterize. Importantly, the conservation of certain sequence motifs suggests the monomeric bacterial OTase, PglB, functions in a catalytically similar manner to the eukaryotic OTase.

The following chapters describe steps taken toward characterization of the bacterial OTase PglB. Chapter 2 describes extensive efforts devoted to optimizing heterologous overexpression of PglB in *E. coli* and the purification procedure. These efforts resulted in significant increases in PglB expression and attainment of improved levels of pure PglB that remains stable and active for several months. The availability of stable PglB at mid-micromolar concentration in solution allows for precise kinetic, biochemical, and biophysical investigations into the enzyme function. Chapters 4 and 5 demonstrate two examples of successful studies that make use of the well-expressed, pure, and stable PglB.

Chapter 3 describes bioinformatic analysis of the sequences of PglB and homologs. A method was developed to parse the sequences based on topological predictions in order to discover pockets of conservation within otherwise divergent regions of the protein sequences. Chapter 4 uses the predictions derived from the bioinformatics analysis to guide site-directed mutagenesis studies of PglB. Results validated the functional relevance of previously unrecognized, highly conserved motifs.

The publication of a crystal structure of a homolog of PglB from a closely related bacterial species provided striking validation of the predictions and biochemical data from Chapters 3 and 4. The structural data represents an impressive step toward visualizing and understanding the OTase mechanism. However, the crystal structure provides visualization of only a single conformation of PglB resolved to 3.5 Å with a bound substrate peptide. There is inevitable uncertainty about the relevance of data obtained for any protein in the non-native environment of a crystal. However, additional factors indicate the PglB structure should be interpreted with caution: activity data is not provided for the PglB fraction used to obtain crystals, the membrane protein is in a detergent micelle and thus is in a highly non-native environment even prior to crystallization, and the crystallization conditions include an alkaline pH at which PglB shows dramatically reduced activity in solution. Nonetheless, the peptide is bound in a conformation that does not resemble the Asx-type turn that was previously proposed function in OTase catalysis, and bold conclusions were drawn that dismissed the significant biochemical data previously presented in support of the Asx-turn proposal.

Described in Chapter 5 is the design and development of a method to measure distances between PglB and the peptide substrate and between locations within the enzyme using resonant energy transfer. Requisite components were assembled and distances measurements were conducted between PglB and the peptide substrate, demonstrating the success of the approach. This system provides the groundwork for precise measurements of conformational changes within the enzyme and substrate-binding location and order. The measurements will be made on the pure solutions of active PglB, providing a degree of confidence that crystallographic data cannot provide. Future use of this system for studies of PglB as well as an archaeal monomeric OTase, AglB, are currently being continued and advanced by Imperiali lab members..

References

1. Larkin A, Imperiali B (2011) The expanding horizons of asparagine-linked glycosylation. Biochemistry 50: 4411-4426.
2. Schwarz F, Aebi M (2011) Mechanisms and principles of N-linked protein glycosylation. Curr Opin Struct Biol 21: 576-582.
3. Zielinska DF, Gnad F, Schropp K, Wisniewski JR, Mann M (2012) Mapping N-glycosylation sites across seven evolutionarily distant species reveals a divergent substrate proteome despite a common core machinery. Mol Cell 46: 542-548.
4. Bause E (1983) Structural requirements of N-glycosylation of proteins. Studies with proline peptides as conformational probes. Biochem J 209: 331-336.
5. Rothman JE, Lodish HF (1977) Synchronised transmembrane insertion and glycosylation of a nascent membrane protein. Nature 269: 775-780.
6. Chavan M, Yan A, Lennarz WJ (2005) Subunits of the translocon interact with components of the oligosaccharyl transferase complex. J Biol Chem 280: 22917-22924.
7. Nilsson IM, von Heijne G (1993) Determination of the distance between the oligosaccharyltransferase active site and the endoplasmic reticulum membrane. J Biol Chem 268: 5798-5801.
8. Kowarik M, Numao S, Feldman MF, Schulz BL, Callewaert N, et al. (2006) N-linked glycosylation of folded proteins by the bacterial oligosaccharyltransferase. Science 314: 1148-1150.
9. Aebi M (2013) N-linked protein glycosylation in the ER. Biochim Biophys Acta.
10. Moremen KW, Tiemeyer M, Nairn AV (2012) Vertebrate protein glycosylation: diversity, synthesis and function. Nat Rev Mol Cell Biol 13: 448-462.
11. Varki A (1993) Biological roles of oligosaccharides: all of the theories are correct. Glycobiology 3: 97-130.
12. Mitra N, Sinha S, Ramya TN, Surolia A (2006) N-linked oligosaccharides as outfitters for glycoprotein folding, form and function. Trends Biochem Sci 31: 156-163.
13. Ashkenazi A, Viard M, Wexler-Cohen Y, Blumenthal R, Shai Y (2011) Viral envelope protein folding and membrane hemifusion are enhanced by the conserved loop region of HIV-1 gp41. FASEB J 25: 2156-2166.
14. Zai J, Mei L, Wang C, Cao S, Fu ZF, et al. (2013) N-glycosylation of the premembrane protein of Japanese encephalitis virus is critical for folding of the envelope protein and assembly of virus-like particles. Acta Virol 57: 27-33.
15. Price JL, Culyba EK, Chen W, Murray AN, Hanson SR, et al. (2012) N-glycosylation of enhanced aromatic sequons to increase glycoprotein stability. Biopolymers 98: 195-211.
16. Wang XY, Ji CG, Zhang JZ (2012) Exploring the molecular mechanism of stabilization of the adhesion domains of human CD2 by N-glycosylation. J Phys Chem B 116: 11570-11577.
17. Arshad N, Ballal S, Visweswariah SS (2013) Site-specific N-linked glycosylation of receptor guanylyl cyclase C regulates ligand binding, ligand-mediated activation and interaction with vesicular integral membrane protein 36, VIP36. J Biol Chem 288: 3907-3917.
18. Walker AK, Soo KY, Levina V, Talbo GH, Atkin JD (2013) N-linked glycosylation modulates dimerization of protein disulfide isomerase family A member 2 (PDIA2). FEBS J 280: 233-243.

19. Zou S, Huang S, Kaleem I, Li C (2013) N-Glycosylation enhances functional and structural stability of recombinant beta-glucuronidase expressed in Pichia pastoris. J Biotechnol 164: 75-81.
20. Wang J, White AL (1999) Role of N-linked glycans, chaperone interactions and proteasomes in the intracellular targeting of apolipoprotein(a). Biochem Soc Trans 27: 453-458.
21. Hurt JK, Fitzpatrick BJ, Norris-Drouin J, Zylka MJ (2012) Secretion and N-linked glycosylation are required for prostatic acid phosphatase catalytic and antinociceptive activity. PLoS One 7: e32741.
22. Mazumder R, Morampudi KS, Motwani M, Vasudevan S, Goldman R (2012) Proteome-wide analysis of single-nucleotide variations in the N-glycosylation sequon of human genes. PLoS One 7: e36212.
23. Tang X, Wu Y, Belenkaya TY, Huang Q, Ray L, et al. (2012) Roles of N-glycosylation and lipidation in Wg secretion and signaling. Dev Biol 364: 32-41.
24. Zou S, Xie L, Liu Y, Kaleem I, Zhang G, et al. (2012) N-linked glycosylation influences on the catalytic and biochemical properties of Penicillium purpurogenum beta-d-glucuronidase. J Biotechnol 157: 399-404.
25. Mant A, Williams S, Roncoroni L, Lowry E, Johnson D, et al. (2013) N-glycosylation-dependent control of functional expression of background potassium channels K2P3.1 and K2P9.1. J Biol Chem 288: 3251-3264.
26. Veldhuis NA, Lew MJ, Abogadie FC, Poole DP, Jennings EA, et al. (2012) N-glycosylation determines ionic permeability and desensitization of the TRPV1 capsaicin receptor. J Biol Chem 287: 21765-21772.
27. Fares F (2006) The role of O-linked and N-linked oligosaccharides on the structure-function of glycoprotein hormones: development of agonists and antagonists. Biochim Biophys Acta 1760: 560-567.
28. Marth JD, Grewal PK (2008) Mammalian glycosylation in immunity. Nat Rev Immunol 8: 874-887.
29. Janik ME, Litynska A, Vereecken P (2010) Cell migration-the role of integrin glycosylation. Biochim Biophys Acta 1800: 545-555.
30. Gu J, Isaji T, Xu Q, Kariya Y, Gu W, et al. (2012) Potential roles of N-glycosylation in cell adhesion. Glycoconj J 29: 599-607.
31. Sekine SU, Haraguchi S, Chao K, Kato T, Luo L, et al. (2013) Meigo governs dendrite targeting specificity by modulating Ephrin level and N-glycosylation. Nat Neurosci.
32. Sayce AC, Miller JL, Zitzmann N (2010) Targeting a host process as an antiviral approach against dengue virus. Trends Microbiol 18: 323-330.
33. Vigerust DJ, Shepherd VL (2007) Virus glycosylation: role in virulence and immune interactions. Trends Microbiol 15: 211-218.
34. Wei X, Decker JM, Wang S, Hui H, Kappes JC, et al. (2003) Antibody neutralization and escape by HIV-1. Nature 422: 307-312.
35. Skehel JJ, Stevens DJ, Daniels RS, Douglas AR, Knossow M, et al. (1984) A carbohydrate side chain on hemagglutinins of Hong Kong influenza viruses inhibits recognition by a monoclonal antibody. Proc Natl Acad Sci U S A 81: 1779-1783.
36. Yan Q, Li XP, Tumer NE (2012) N-glycosylation does not affect the catalytic activity of ricin a chain but stimulates cytotoxicity by promoting its transport out of the endoplasmic reticulum. Traffic 13: 1508-1521.

37. Hause BM, Stine DL, Sheng Z, Wang Z, Chakravarty S, et al. (2012) Migration of the swine influenza virus delta-cluster hemagglutinin N-linked glycosylation site from N142 to N144 results in loss of antibody cross-reactivity. Clin Vaccine Immunol 19: 1457-1464.
38. Hillaire ML, van Eijk M, Nieuwkoop NJ, Vogelzang-van Trierum SE, Fouchier RA, et al. (2012) The number and position of N-linked glycosylation sites in the hemagglutinin determine differential recognition of seasonal and 2009 pandemic H1N1 influenza virus by porcine surfactant protein D. Virus Res 169: 301-305.
39. Kim JI, Park MS (2012) N-linked glycosylation in the hemagglutinin of influenza A viruses. Yonsei Med J 53: 886-893.
40. Hutter J, Rodig JV, Hoper D, Seeberger PH, Reichl U, et al. (2013) Toward animal cell culture-based influenza vaccine design: viral hemagglutinin N-glycosylation markedly impacts immunogenicity. J Immunol 190: 220-230.
41. Vigerust DJ, Ulett KB, Boyd KL, Madsen J, Hawgood S, et al. (2007) N-linked glycosylation attenuates H3N2 influenza viruses. J Virol 81: 8593-8600.
42. Pabst M, Chang M, Stadlmann J, Altmann F (2012) Glycan profiles of the 27 N-glycosylation sites of the HIV envelope protein CN54gp140. Biol Chem 393: 719-730.
43. Wang W, Nie J, Prochnow C, Truong C, Jia Z, et al. (2013) A systematic study of the N-glycosylation sites of HIV-1 envelope protein on infectivity and antibody-mediated neutralization. Retrovirology 10: 14.
44. Bacon DJ, Szymanski CM, Burr DH, Silver RP, Alm RA, et al. (2001) A phase-variable capsule is involved in virulence of Campylobacter jejuni 81-176. Mol Microbiol 40: 769-777.
45. Guerry P, Szymanski CM, Prendergast MM, Hickey TE, Ewing CP, et al. (2002) Phase variation of Campylobacter jejuni 81-176 lipooligosaccharide affects ganglioside mimicry and invasiveness in vitro. Infect Immun 70: 787-793.
46. Szymanski CM, Burr DH, Guerry P (2002) Campylobacter protein glycosylation affects host cell interactions. Infect Immun 70: 2242-2244.
47. Demetriou M, Granovsky M, Quaggin S, Dennis JW (2001) Negative regulation of T-cell activation and autoimmunity by Mgat5 N-glycosylation. Nature 409: 733-739.
48. Grigorian A, Mkhikian H, Demetriou M (2012) Interleukin-2, Interleukin-7, T cell-mediated autoimmunity, and N-glycosylation. Ann N Y Acad Sci 1253: 49-57.
49. Grigorian A, Mkhikian H, Li CF, Newton BL, Zhou RW, et al. (2012) Pathogenesis of multiple sclerosis via environmental and genetic dysregulation of N-glycosylation. Semin Immunopathol 34: 415-424.
50. Gu J, Cao R, Zhang Y, Lian X, Ishag H, et al. (2012) Deletion of the single putative N-glycosylation site of the porcine circovirus type 2 Cap protein enhances specific immune responses by DNA immunisation in mice. Vet J 192: 385-389.
51. Lauc G, Huffman JE, Pucic M, Zgaga L, Adamczyk B, et al. (2013) Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. PLoS Genet 9: e1003225.
52. Freeze HH (2001) Update and perspectives on congenital disorders of glycosylation. Glycobiology 11: 129R-143R.
53. Freeze HH (2002) Human disorders in N-glycosylation and animal models. Biochim Biophys Acta 1573: 388-393.
54. Schachter H, Freeze HH (2009) Glycosylation diseases: quo vadis? Biochim Biophys Acta 1792: 925-930.

55. (2013) Neuromuscular disease: Defects in asparagine-linked protein glycosylation implicated in congenital mysathenic syndromes. Nat Rev Neurol 9: 183.
56. Arnold JN, Saldova R, Hamid UM, Rudd PM (2008) Evaluation of the serum N-linked glycome for the diagnosis of cancer and chronic inflammation. Proteomics 8: 3284-3293.
57. Potapenko IO, Haakensen VD, Luders T, Helland A, Bukholm I, et al. (2010) Glycan gene expression signatures in normal and malignant breast tissue; possible role in diagnosis and progression. Mol Oncol 4: 98-118.
58. Gilgunn S, Conroy PJ, Saldova R, Rudd PM, O'Kennedy RJ (2013) Aberrant PSA glycosylation--a sweet predictor of prostate cancer. Nat Rev Urol 10: 99-107.
59. Tousi F, Bones J, Iliopoulos O, Hancock WS, Hincapie M (2012) Multidimensional liquid chromatography platform for profiling alterations of clusterin N-glycosylation in the plasma of patients with renal cell carcinoma. J Chromatogr A 1256: 121-128.
60. Ozohanics O, Turiak L, Puerta A, Vekey K, Drahos L (2012) High-performance liquid chromatography coupled to mass spectrometry methodology for analyzing site-specific N-glycosylation patterns. J Chromatogr A 1259: 200-212.
61. Kuo CW, Wu IL, Hsiao HH, Khoo KH (2012) Rapid glycopeptide enrichment and N-glycosylation site mapping strategies based on amine-functionalized magnetic nanoparticles. Anal Bioanal Chem 402: 2765-2776.
62. Han H, Stapels M, Ying W, Yu Y, Tang L, et al. (2012) Comprehensive characterization of the N-glycosylation status of CD44s by use of multiple mass spectrometry-based techniques. Anal Bioanal Chem 404: 373-388.
63. Chuang GY, Boyington JC, Joyce MG, Zhu J, Nabel GJ, et al. (2012) Computational prediction of N-linked glycosylation incorporating structural properties and patterns. Bioinformatics 28: 2249-2255.
64. Wojtowicz K, Szaflarski W, Januchowski R, Zawierucha P, Nowicki M, et al. (2012) Inhibitors of N-glycosylation as a potential tool for analysis of the mechanism of action and cellular localisation of glycoprotein P. Acta Biochim Pol 59: 445-450.
65. Galonic DP, Gin DY (2007) Chemical glycosylation in the synthesis of glycoconjugate antitumour vaccines. Nature 446: 1000-1007.
66. Igura M, Kohda D (2011) Selective control of oligosaccharide transfer efficiency for the N-glycosylation sequon by a point mutation in oligosaccharyltransferase. J Biol Chem 286: 13255-13260.
67. Guarino C, DeLisa MP (2012) A prokaryote-based cell-free translation system that efficiently synthesizes glycoproteins. Glycobiology 22: 596-601.
68. Glover KJ, Weerapana E, Numao S, Imperiali B (2005) Chemoenzymatic synthesis of glycopeptides with PglB, a bacterial oligosaccharyl transferase from Campylobacter jejuni. Chem Biol 12: 1311-1315.
69. Baker JL, Celik E, Delisa MP (2013) Expanding the glycoengineering toolbox: the rise of bacterial N-linked protein glycosylation. Trends Biotechnol 31: 313-323.
70. Kandiba L, Aitio O, Helin J, Guan Z, Permi P, et al. (2012) Diversity in prokaryotic glycosylation: an archaeal-derived N-linked glycan contains legionaminic acid. Mol Microbiol 84: 578-593.
71. Schmaltz RM, Hanson SR, Wong CH (2011) Enzymes in the synthesis of glycoconjugates. Chem Rev 111: 4259-4307.
72. Jones CJ, Larive CK (2011) Carbohydrates: Cracking the glycan sequence code. Nat Chem Biol 7: 758-759.

73. Sola RJ, Griebenow K (2010) Glycosylation of therapeutic proteins: an effective strategy to optimize efficacy. BioDrugs 24: 9-21.
74. Sola RJ, Rodriguez-Martinez JA, Griebenow K (2007) Modulation of protein biophysical properties by chemical glycosylation: biochemical insights and biomedical implications. Cell Mol Life Sci 64: 2133-2152.
75. Walsh G, Jefferis R (2006) Post-translational modifications in the context of therapeutic proteins. Nat Biotechnol 24: 1241-1252.
76. Szymanski CM, Yao R, Ewing CP, Trust TJ, Guerry P (1999) Evidence for a system of general protein glycosylation in Campylobacter jejuni. Mol Microbiol 32: 1022-1030.
77. Wacker M, Linton D, Hitchen PG, Nita-Lazar M, Haslam SM, et al. (2002) N-linked glycosylation in Campylobacter jejuni and its functional transfer into E. coli. Science 298: 1790-1793.
78. al. Ue (2003) Bacterial glycoproteins: Functions, biosynthesis and applications. Proteomics.
79. Karlyshev AV, Ketley JM, Wren BW (2005) The Campylobacter jejuni glycome. FEMS Microbiol Rev 29: 377-390.
80. Linton D, Dorrell N, Hitchen PG, Amber S, Karlyshev AV, et al. (2005) Functional analysis of the Campylobacter jejuni N-linked protein glycosylation pathway. Mol Microbiol 55: 1695-1703.
81. Weerapana E, Glover KJ, Chen MM, Imperiali B (2005) Investigating bacterial N-linked glycosylation: synthesis and glycosyl acceptor activity of the undecaprenyl pyrophosphate-linked bacillosamine. J Am Chem Soc 127: 13766-13767.
82. Glover KJ, Weerapana E, Chen MM, Imperiali B (2006) Direct biochemical evidence for the utilization of UDP-bacillosamine by PglC, an essential glycosyl-1-phosphate transferase in the Campylobacter jejuni N-linked glycosylation pathway. Biochemistry 45: 5343-5350.
83. Chen MM, Glover KJ, Imperiali B (2007) From peptide to protein: comparative analysis of the substrate specificity of N-linked glycosylation in C. jejuni. Biochemistry 46: 5579-5585.
84. Chen MM, Weerapana E, Ciepichal E, Stupak J, Reid CW, et al. (2007) Polyisoprenol specificity in the Campylobacter jejuni N-linked glycosylation pathway. Biochemistry 46: 14342-14348.
85. Jaffee MB, Imperiali B (2011) Exploiting topological constraints to reveal buried sequence motifs in the membrane-bound N-linked oligosaccharyl transferases. Biochemistry 50: 7557-7567.
86. Lizak C, Gerber S, Numao S, Aebi M, Locher KP (2011) X-ray structure of a bacterial oligosaccharyltransferase. Nature 474: 350-355.
87. Chen MM, Bartlett AI, Nerenberg PS, Friel CT, Hackenberger CP, et al. (2010) Perturbing the folding energy landscape of the bacterial immunity protein Im7 by site-specific N-linked glycosylation. Proc Natl Acad Sci U S A 107: 22528-22533.
88. Imperiali B, Shannon KL (1991) Differences between Asn-Xaa-Thr-containing peptides: a comparison of solution conformation and substrate behavior with oligosaccharyltransferase. Biochemistry 30: 4374-4380.
89. Imperiali B, Hendrickson, T.L. (1995) Asparagine-linked glycosylation: specificity and function of oligosaccharyl transferase. Bioorg & Med Chem 3: 1565-1578.
90. Marshall RD (1972) Glycoproteins. Annu Rev Biochem 41: 673-702.

91. Bause E, Legler G (1981) The role of the hydroxy amino acid in the triplet sequence Asn-Xaa-Thr(Ser) for the N-glycosylation step during glycoprotein biosynthesis. Biochem J 195: 639-644.
92. B. Imperiali KLS, K. W. Rickert (1992) Role of peptide conformation in asparagine-linked glycosylation. J Am Chem Soc 114: 7942-7944.
93. B. Imperiali KLS, M. Unno, K. W. Rickert (1992) A Mechanistic Proposal for Asparagine-Linked Glycosylation. J Am Chem Soc 114.
94. B. Imperiali SLF, R.A. Moats, T. J. Prins (1992) A conformational study of peptides with the general structure Ac-L-Xaa-Pro-D-Xaa-L-Xaa-NH2: spectroscopic evidence for a peptide with significant .beta.-turn character in water and in dimethyl sulfoxide. J Am Chem Soc 114: 3182-3188.
95. Hendrickson TLS, J. R.; Kato, M.; Imperiali, B., (1996) Design and evaluation of potent inhibitors of asparagine-linked protein glycosylation. J Am Chem Soc 118: 7636-7637.
96. Valliere-Douglass JF, Eakin CM, Wallace A, Ketchem RR, Wang W, et al. (2010) Glutamine-linked and non-consensus asparagine-linked oligosaccharides present in human recombinant antibodies define novel protein glycosylation motifs. J Biol Chem 285: 16012-16022.
97. Valliere-Douglass JF, Kodama P, Mujacic M, Brady LJ, Wang W, et al. (2009) Asparagine-linked oligosaccharides present on a non-consensus amino acid sequence in the CH1 domain of human antibodies. J Biol Chem 284: 32493-32506.
98. Zeitler R, Hochmuth E, Deutzmann R, Sumper M (1998) Exchange of Ser-4 for Val, Leu or Asn in the sequon Asn-Ala-Ser does not prevent N-glycosylation of the cell surface glycoprotein from Halobacterium halobium. Glycobiology 8: 1157-1164.
99. Zielinska DF, Gnad F, Wisniewski JR, Mann M (2010) Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. Cell 141: 897-907.
100. Vance BA, Wu W, Ribaudo RK, Segal DM, Kearse KP (1997) Multiple dimeric forms of human CD69 result from differential addition of N-glycans to typical (Asn-X-Ser/Thr) and atypical (Asn-X-cys) glycosylation motifs. J Biol Chem 272: 23117-23122.
101. Flavio Schwarz Y-YF, Mario Schubert, Markus Aebi (2011) Cytoplasmic N-Glycosyltransferase of Actinobacillus pleuropneumoniae Is an Inverting Enzyme and Recognizes the NX(S/T) Consensus Sequence. J Biol Chem 286: 35267-35274.
102. Grass S, Buscher AZ, Swords WE, Apicella MA, Barenkamp SJ, et al. (2003) The Haemophilus influenzae HMW1 adhesin is glycosylated in a process that requires HMW1C and phosphoglucomutase, an enzyme involved in lipooligosaccharide biosynthesis. Mol Microbiol 48: 737-751.
103. Kawai F, Grass S, Kim Y, Choi KJ, St Geme JW, 3rd, et al. (2011) Structural insights into the glycosyltransferase activity of the Actinobacillus pleuropneumoniae HMW1C-like protein. J Biol Chem 286: 38546-38557.

CHAPTER 2:

AN OPTIMIZED PROTOCOL FOR EXPRESSION AND PURIFICATION OF PGLB

Acknowledgments:

Insertion of *pglB* gene into vectors with 5 different fusion tags, expression of fusion-tag-PglB constructs, purification, and SDS-PAGE analysis was performed by Dr. Mark Chen, a former Imperiali lab member.

The majority of the data in this chapter has been published in:

Jaffee, M.B. and Imperiali, B. (2013) Optimized protocol for expression and purification of membrane-bound PglB, a bacterial oligosaccharyltransferase. *Prot. Exp. Purif.* 89:241–250

Summary

Characterization of the oligosaccharyl transferase (OTase) is critical for developing ways to inhibit, engineer, and otherwise manipulate the enzyme for research and therapeutic purposes. The minimal understanding of this enzyme can be attributed to its complex, transmembrane structure, and the resulting instability and resistance to overexpression and purification. The following chapter describes an optimized procedure for recombinant expression and purification of PglB in a stably active form. The conditions screened at each step, the order of screening, and the method of comparing conditions are described. Ultimately, the following approach increased expression levels from tens of micrograms to several milligrams of active protein per liter of *E. coli* culture, and increased stability from several hours to greater than six months post-purification. This represents the first detailed procedure for attaining a pure, active, and stable OTase in milligram quantities. These quantities of pure protein makes detailed kinetic and biophysical analysis of an active OTase a realistic possibility. In addition to presenting an optimized protocol for expression and purification of PglB, these results present a general guide for the systematic optimization of the expression, purification, and stability of a large, transmembrane protein.

Introduction

Biological researchers are invariably familiar with the importance of membrane proteins in therapeutic development and throughout cell biology. Several well-known statistics demonstrate this import; most notably, membrane proteins account for approximately one-third of the human proteome and comprise a majority of current drug targets [1-3]. However, standard overexpression and purification techniques are often unsuitable for this class of proteins, which presents a major obstacle to research progress. There exist promising developments for acquiring large quantities of membrane proteins, including cell-free translation systems, directed evolution of well-expressing bacteria, and the ever-increasing advances in the efficiency of current approaches [4,5]. However, expressing and purifying membrane proteins currently remains a largely empirical, time-consuming, and high-risk endeavor, leaving many important membrane-bound enzymes uncharacterized and presenting significant gaps in the understanding of cellular pathways.

The proteins in the eukaryotic N-linked glycosylation (NLG) biosynthetic pathway provide an example of this phenomenon, as virtually all of the enzymes are membrane-bound. NLG plays a major role in many cell processes, including immune-system response, protein signaling and trafficking, and pathogenic invasion strategies [6-9]. In addition, NLG introduces several prospective tools in therapeutics; N-linked sugars are capable of functioning as indicators of cell state and type [10-12] and represent a novel chemical platform for developing new therapeutics and enhancing efficacy of current drugs [10,13]. Most current studies involving NLG, however, focus on determining the specific effects of the glycan modification on a target of interest, often aimed primarily at establishing the glycosylated sites within a protein and the effects of the modification on function [14-17]. Far less is known about the enzymes comprising the biosynthetic pathway of NLG due to the difficulty of expressing and purifying the involved membrane proteins in a stable and active state. Closing this knowledge gap would allow for enhanced ability to efficiently manipulate, study, and thus control and utilize NLG.

In yeast, the OTase is a complex composed of eight subunits, all of which have one or more transmembrane domains. In comparison, PglB is comprised of a single subunit, which is homologous to the catalytic subunit of the eukaryotic OTase (**Figure 1**). Although PglB is ostensibly a tractable target relative to the eukaryotic OTase, the enzyme represents a challenge in its own right. PglB has thirteen transmembrane domains and is fairly large (82 kDa), which accounts for poor recombinant expression and instability in *E. coli*. Thus, characterization of PglB has lagged relative to other *C. jejuni* NLG enzymes, despite the potential of this OTase to reveal fundamental principles about the mechanism of OTases across the evolutionary spectrum. This chapter describes the systematic approach used to optimize the expression, purification, and stability of active PglB. The conditions screened at each step, the order of screening, and the method of comparison for each condition are described.

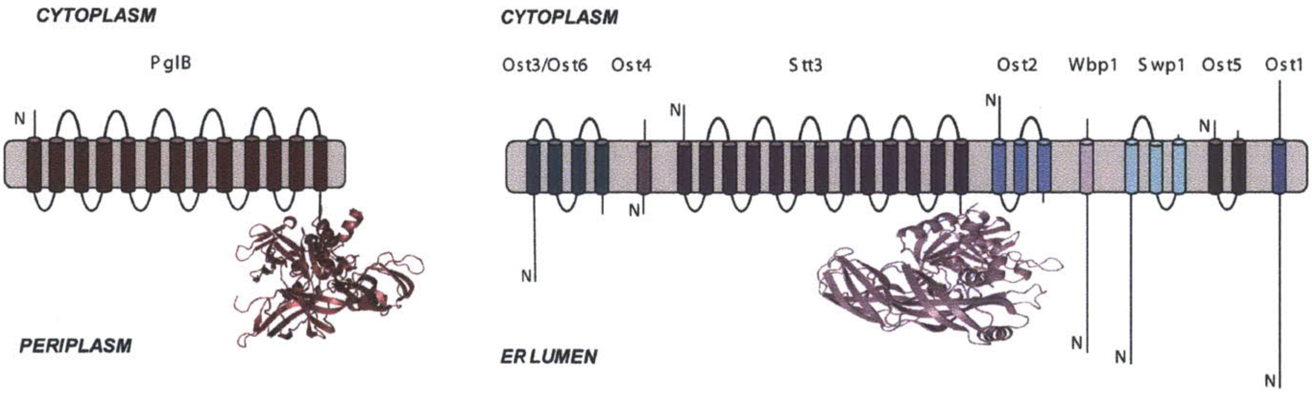


Figure 1: Comparison of the OTases in *C. jejuni* (bacteria) and *S. cerevisiae* (eukaryotes). Images highlight the similarity between the *S. cerevisiae* catalytic subunit 'STT3' and PglB. The relative simplicity of the bacterial OTase when compared to the eukaryotic enzyme complex is also demonstrated.

Specific activity values are used to determine the optimal conditions for balancing protein recovery with activity recovery. The activity-guided approach can be contrasted with expression screening approaches undertaken with the primary interest in X-ray crystallographic studies, which generally focus on protein quantity and monodispersity and do not account for the proportion of active to inactive protein purified. Thus, the following method is intended to aid researchers interested in *C. jejuni* N-linked glycosylation, and also to illustrate an activity-guided approach to optimizing expression, purification, and stability of a specific membrane protein of interest. This following data represents the first available protocol for expressing and purifying milligram quantities of a stable and active OTase.

Results

Expression optimization of PglB in E. coli: general strategy

Optimization of PglB expression in *E. coli* involved screening many expression constructs and conditions. To facilitate testing the multitude of expression conditions, 5-mL cultures are grown for each condition specified (direct comparisons are always grown simultaneously). Unless specified otherwise, the cultures are grown at 37°C until they reach an O.D. of 0.6-0.8, at which point the temperature is shifted to 16°C for overnight growth. These mini-cultures are then centrifuged to concentrate the cells. The cell pellets are weighed and resuspended in a volume of lysis buffer for 0.1 g/mL and are then lysed using sonication. Activity rates are compared for series of lysates derived from cultures with varied expression conditions. The lysates with highest activity are used to determine which expression conditions yield the most functional enzyme. Expression and activity comparisons are carried out on at least two separate occasions to ensure conclusions drawn are accurate.

Protein Tags and Gene Truncations

Membrane proteins are often expressed as truncations of the native protein to impart higher expression or stability to the protein [18]. It has been shown that the C-terminal, soluble domain of PglB is not functional when expressed on its own (roughly from residue 420 to C-terminus) [19]. However, an apparent proteolytic degradation product of PglB is consistently observed via western blot when full-length PglB is over-expressed in *E. coli* (**Figure 2**). The roughly 50 kDa band is expected to contain the C-terminal domain of PglB because it is visible on western blots when an anti-His (C-terminal) antibody is used but not when an anti-T7 (N-terminal) antibody is used. However, at 50 kDa, the degradation product is significantly larger than the soluble domain alone and thus likely contains several transmembrane domains, as well.

It was considered that this degradation product might, in contrast to the soluble domain alone, represent an active version of PglB that is more stable or easily expressed than the full-length protein. To test this hypothesis, several truncated PglB constructs were made (**Figure 2**). Constructs were expressed and normalized for expression via western blot. However, it was found that all of the constructs lacked catalytic activity, including those that lacked only approximately 50 residues from the C- or N-terminus (**Figure 2**).

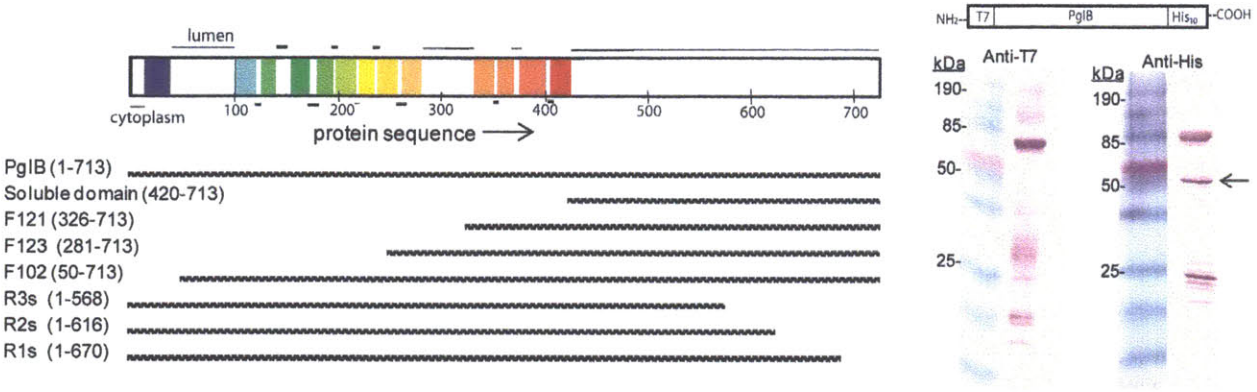

lumen
cytoplasm
100
200
300
400
500
600
700
protein sequence ⟶
PglB (1-713)
Soluble domain (420-713)
F121 (326-713)
F123 (281-713)
F102 (50-713)
R3s (1-568)
R2s (1-616)
R1s (1-670)
NH₂-
T7
PglB
His₁₀
-COOH
kDa
Anti-T7
190-
85-
50-
25-
kDa
Anti-His
190-
85-
50-
25-

Figure 2: Investigating a natural degradation product of PglB.
Top left: topology diagram for PglB based on available structural data (PDB 3RCE). The horizontal bar represents PglB with the protein sequence numbered along the bottom starting with the N-terminus. The location of each transmembrane helix is represented by a vertical bar and the thin black horizontal lines represent the location of the soluble regions (lumen on top, cytoplasm on bottom). *Bottom left:* Bars representing the various truncated constructs of PglB, with length aligned to X-axis of topology diagram. The name assigned to each construct is listed on the left along with the residues encompassed by that construct. *Right:* Western blots showing the degradation product of interest (indicated by the arrow) when stained with the anti-His antibody (C-terminal tag) but not when stained with the anti-T7 antibody (N-terminal tag).

The observations that even short N-terminal truncations cause complete activity loss can now be explained by the recent structural and biochemical data, which reveal that a required catalytic motif appears in transmembrane-loop regions close to the N-terminus [20,21]. However, the available data – structural, biochemical, and bioinformatic- has not located major conservation or apparent involvement of the C-terminal portion of the soluble domain in catalysis, so it remains to be seen why the presence of last 50 or so residues of this enzyme is required for activity.

Table 1: Fusion constructs used in PglB expression screening.

Vector	N-terminal fusion tag	Fusion tag MW (kDa, approx.)	References
pET24a(+)	**T7**	**1**	[22]
pGBH	**Protein G, B1 domain (GB1)**	**6**	[23]
pET SUMO	**Small Ubiquitin-Like Protein (SUMO)**	**11**	[24]
pET Trx	**Thioredoxin (Trx)**	**12**	[25]
pGEX	**Glutathione-S-Transferase (GST)**	**26**	[26]
pMAL-c2x	**Maltose-Binding Protein (MBP)**	**42.5**	[25,27]

Screening of expression tags represents another common approach used to alter expression, solubility, localization, and other functional aspects of proteins [28]. The *pglB* gene was cloned into several vectors encoding N-terminal fusion tags that have been shown to positively affect expression (**Table 1**). These tags include: T7, Glutathione-S-transferase (GST), Maltose Binding Protein (MBP), SUMO, Thioredoxin, and GB1 [23-27]. Within those tested, it was estimated that the MBP (pMAL-c2X), the T7 (pET-24a(+)) and the GB1 (pGBH, [23]) tags are associated with the highest expression under standard expression conditions, which can easily be discerned from a visual inspection of SDS-PAGE analyses (**Figure 3**). These three constructs were thereafter used for systematic optimization.

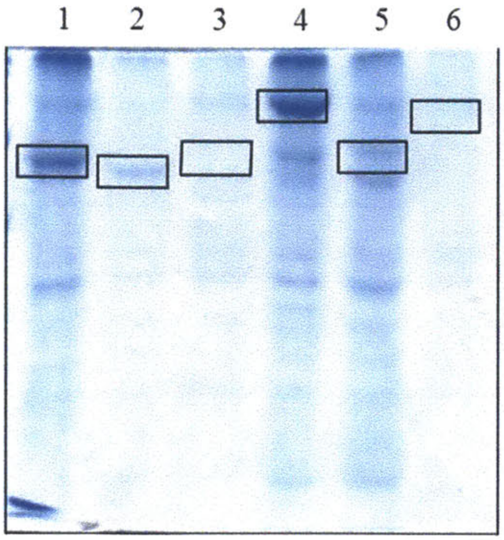


Figure 3: Coomassie-stained SDS-PAGE of fractions of multiple PglB constructs. Each lane contains a version of PglB with the expression tag indicated: 1. GB1, 2.T7, 3.Thioredoxin, 4. MBP, 5. SUMO, 6. GST. (Figure courtesy of Dr. Mark Chen).

Expression and induction conditions

A variety of *E. coli* competent cells were screened to determine the highest-yielding strain for PglB expression. Seven types of *E. coli* expression strains were transformed with the MBP-PglB and T7-PglB vector constructs; these included C41(DE3), Rosetta 2, Rosetta 2 gami, and the following BL21(DE3) strains: RIL, Gold, RP, pLys. It was observed that BL21(DE3) RIL cells show the highest level of expression of both MBP-PglB and T7-PglB as measured by activity levels and western blotting. This result suggests that this strain is optimal for PglB generally and that the outcome is not tag-specific.

A range of induction parameters were investigated, including IPTG concentration used to induce expression, optical density of cultures at the time of induction, incubation temperature post-induction, and length of time cultures were grown following induction. The effects of each of these variables are interdependent; thus, each combination should ideally be tested. Table II displays the organization scheme for this screening process. Ultimately it was found that cultures induced at a very high O.D. (>1.2) rather than the typical O.D. of 0.6-0.8 results in significantly improved expression levels (Figure 4A).

Table 2: Induction conditions varied to screen for improved expression of PglB.

O.D. of induction[a]	Temperature shift upon induction	[IPTG] (mM) to induce
0.6	37°C → 16°C	0.2
0.6	37°C → 16°C	0.7
0.6	37°C → 16°C	1
1	37°C → 16°C	0.2
1	37°C → 16°C	0.7
1	37°C → 16°C	1
0.6	37°C ——|	0.2
0.6	37°C ——|	0.7
0.6	37°C ——|	1
1	37°C ——|	0.2
1	37°C ——|	0.7
1	37°C ——|	1

a. For each condition, time points were taken at 3, 6, and 20 hours post induction.
b. ——| indicates no shift.

The improvements in PglB expression due to induction at high O.D. suggested that auto-induction, or 'high-density cell growth', may be an advantageous way to express PglB. Auto-induction involves growing the expression culture (expressing a protein under the control of the inducible T7 system) in complex media containing a small amount of lactose. Amino acids and other carbon energy sources in the media prevent induction by lactose during log phase of cell growth. Depletion of these components and approach to a saturating density automatically prompts the cells to uptake lactose, causing induction of the target protein [29]. The procedure is similar to IPTG-induced expression, except auto-induction requires a media specifically developed and optimized to couple induction of protein expression by lactose with the approach of the culture to saturating density (see *Materials and Methods*). This method not only eliminates the need to frequently monitor O.D. of the culture, it also ensures that the expression cultures have the highest beneficial O.D. when induction occurs, and prevents any significant expression of the target protein prior to induction.

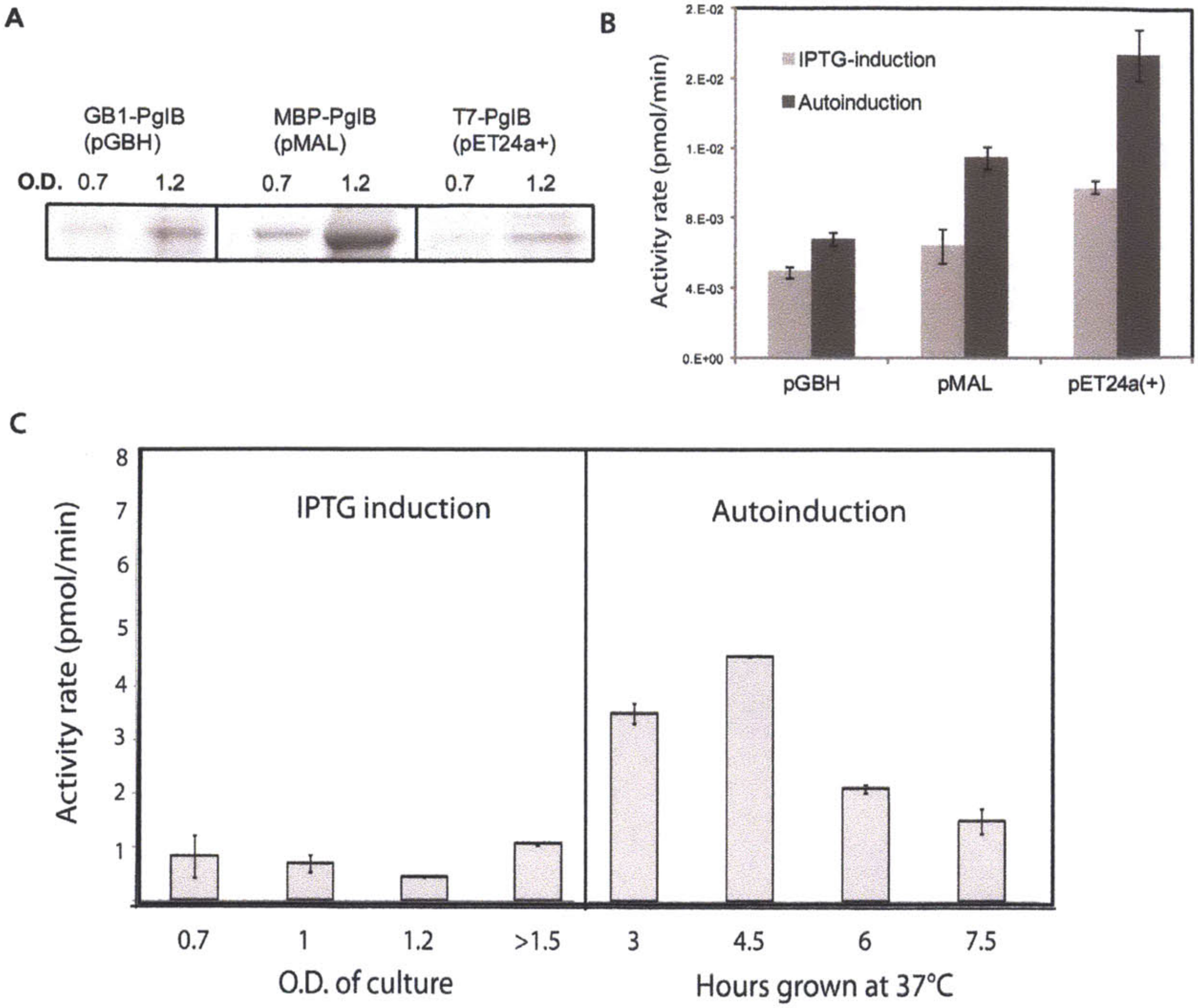


Figure 4: Expression improvement with IPTG induction at high O.D and auto-induction. *A.* Cell cultures were grown at 37°C to an O.D. of 0.7 and 1.2, at which point the cultures were induced with 1 mM IPTG and the temperature was shifted from 37 to 16°C. Fractions from equal cell-pellet weights were purified over Ni-NTA resin and elutions were compared via SDS-PAGE. *B.* Cell cultures of each construct were grown in LB using IPTG (induced at O.D. = 1) as well as using auto-induction media. Equal weights of cell pellet were lysed and spun at 10000 x g to remove debris. The cleared lysates were used to measure initial rates of activity for each fraction. *C.* Graph indicates activity levels for PglB-pET24a (+) expression cultures that were grown under the specified conditions. For the auto-induced cultures, it was of interest to determine the optimal time to grow the cultures at 37 °C before the temperature shift to 16 °C; hence, the hours post-inoculation at 37 °C were varied and compared. Error bars indicate the standard error of activity measurements. Activity assays were performed as described above using cleared lysate fractions from equal weights of cell pellet from each condition.

Due to the fact that auto-induction involves growing cell cultures to saturation, the cell weight obtained per liter of culture is dramatically increased relative to other expression and induction conditions. While one liter of IPTG-induced culture (induced at an O.D. of 0.6-0.8 with 1 mM IPTG) gives roughly 2 grams of cells, one liter of auto-induction culture generally yields at least 25 grams of cell weight (with expression parameters otherwise held constant).

Therefore, when comparing auto-induction expression to IPTG-induced expression, it was of interest to determine whether improvements in protein yield are due to the greater number of cells produced per liter of culture or to improved expression of PglB per cell. Activity per gram of cell weight was measured for PglB in pMAL, pET24a(+), and pGBH vectors grown with both auto-induction and IPTG induction (**Figure 4B**). In order to determine the optimal time of induction (or temperature shift) to use in the comparison, PglB-pET24a(+) was expressed using both IPTG-induction at various O.D.s and auto-induction with varying time spent growing at 37°C before shifting the temperature to 16°C (**Figure 4C**).

Results show that PglB expressed using auto-induction is optimal with a temperature shift occurring at 4.5 hours post-inoculation and that PglB expressed using IPTG induction is optimal when induced at an O.D. ~ 1.6. Furthermore, auto-induction yields significant increase per gram of cell weight over expression using IPTG induction in addition to yielding more cell weight per liter of culture (**Figure 4B**). It was additionally shown that with auto-induction, PglB expressed in the pET24a(+) vector yields higher levels of active protein than PglB expressed in the pMAL or the pGBH vector (**Figure 4B**).

For this reason, the PglB-pET24a(+) construct was used for subsequent optimization and experiments. Use of this construct was also advantageous because the T7- tag is significantly smaller than both MBP and GB1 (see **Table I**), and thus most closely resembles the native (untagged) PglB.

Optimization of PglB purification

Once expression levels of PglB had been improved using auto-induction, the purification procedure was optimized so that the higher expression would translate into higher levels of pure, active protein. A flow chart of the general procedure for PglB purification is shown in **Figure 5A**. Optimization begins at the top of the flow chart, and the following steps were not optimized until optimized conditions were established for all previous steps. Optimizing a purification step involved dividing the 'crude' fraction into equal parts. Each part was subjected to one of several conditions. The total activity and protein concentration in the fractions before and after the purification step were measured for each condition.

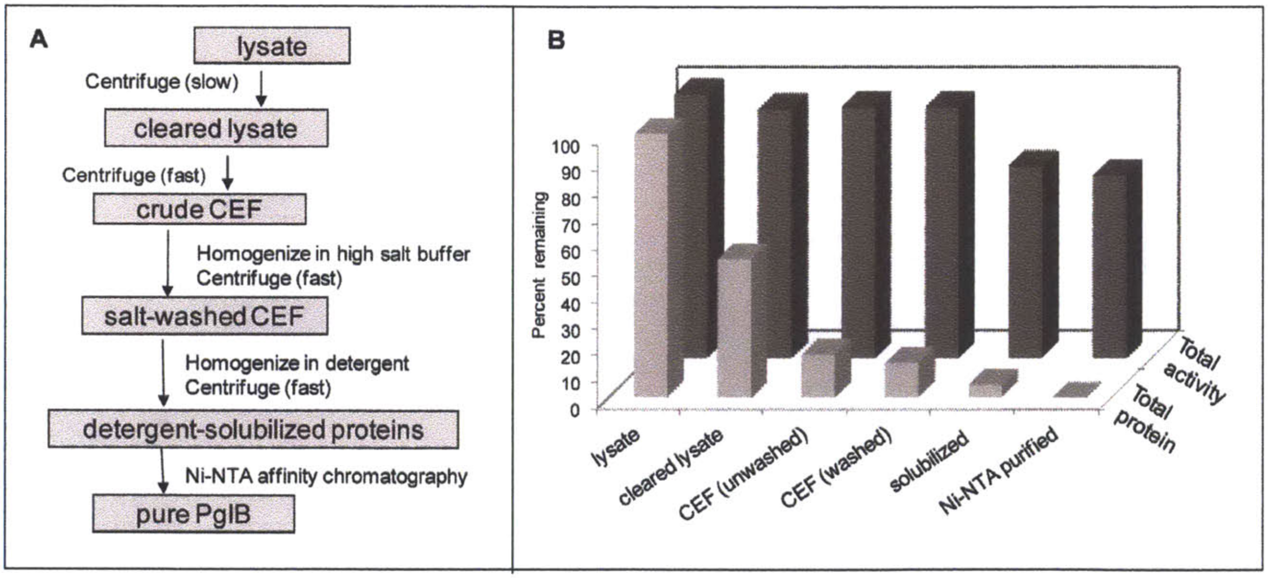


Figure 5: Procedure for optimizing purification protocol for PglB.
A. Flow chart of the general procedure for purifying PglB. The boxes contain the fraction acquired at each step, and adjacent to the arrows is the physical procedure performed to get from one fraction to the next. *B*. The values of total protein and total activity in percent of initial (where 'initial values' refer to those in lysate) obtained using optimized procedure. Values are shown for each purification step diagrammed in the flow chart in part *A* and correspond to the values in Table 3. Note the steep decrease in total protein concentration associated with each step and the comparatively mild loss of total activity.

The activity and protein-concentration measurements allowed calculation of specific activity, fold purification, and percent yield (see *Materials and Methods*). These values provide a definitive measure by which to judge the optimal condition for each purification step, where optimal is defined as maintaining maximal activity and minimal levels of total protein (**Figure 5B**). **Table 3** provides the values obtained using the final optimized purification protocol, while **Figure 6** shows the activity data which was used (along with the quantities of total protein present in each fraction) to calculate the values in **Table 3**. A stepwise procedure for the purification steps can be found at the conclusion of this chapter's results.

Table 3: Values from final optimized purification protocol of PglB.
A 3.6-gram pellet was used, representing approximately 1/7 the weight from 1L culture.

Fraction	Concentration (mg/mL)[a,b]	Total protein (mg)[c]	Rate (nmol/min)[b,d]	Total activity (nmol/min)[c]	Specific activity (nmol/min*mg)	Percent Yield	Fold Purification
lysate	12.85	515	1.0E-05	8.4E-02	1.6E-04	100	1.0
cleared lysate	6.74	270	9.5E-06	7.6E-02	2.83E-04	91	1.7
CEF (unwashed)	2.25	83	1.1E-05	7.9E-02	9.48E-04	93	5.8
CEF (washed)	1.6	68	8.6E-06	7.3E-02	1.08E-03	87	6.6
solubilized	0.59	25	7.0E-06	5.9E-02	2.36E-03	70	15
Ni-NTA purified	0.2	0.6	8.5E-05	5.4E-02	9.06E-02	64	550

a. Concentration of total protein in fraction.
b. Measured value (see Methods).
c. Volume-corrected.
d. Rate of transfer of radioactive sugar substrate to peptide substrate by active PglB in fraction.

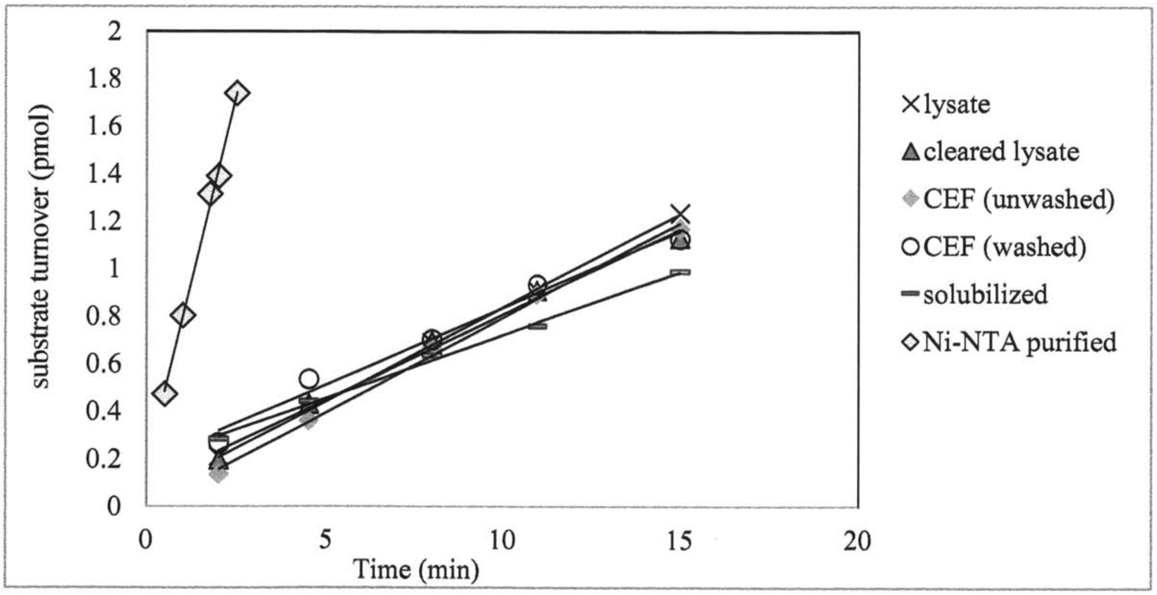


Figure 6: Activity-rate data for fractions used to calculate Table 3 purification values. Best fit lines were used to estimate the activity rates of PglB contained in 5 μL of each purification fraction. The rates from 5 μL were volume corrected to find the total activity present in the entire fraction.

Isolation and processing of cell-membrane fraction

Cell lysis was performed with sonication in three sets of two minutes each. Cells were incubated on ice before, during, and post-sonication to prevent the temperature of the cell suspension from rising and potentially inducing cellular stress response and increased protein degradation and denaturation. Sonication is considered a relatively harsh method of lysis, so as a comparison two PglB cell pellets of equal weight were lysed either by French Press or by sonication. Purification and activity results were unaffected by the method of lysis, indicating sonication is not causing significant impairment to protein yields.

After lysis of cells, a 10000 x g centrifugal spin is performed to remove unbroken cells and other extraneous cell debris, yielding the 'cleared lysate' supernatant. While there are general recommendations for the speed at which to achieve this desired separation, at times it is beneficial to vary the exact speed of the spin to improve retention of the protein of interest. In this case, the first spin was initially performed at the standard speed of 10000 x g. Calculation of the fold purification and the percent yield showed that the step was increasing the specific activity by only a small amount, but 40% of the activity was being lost (**Table 4**). The speed was lowered to 8000 x g, which improved the yield and fold purification (**Table 4**). Particularly, as this is one of the first purification steps, the effects of changes in percent yield on the final yield of protein are compounded. Thus, a relatively small difference in initial the centrifugal separation speed had considerable effects on the ultimate yield of pure protein.

Table 4: Minor change in centrifugal spin significantly affects yield of purification step

	Fraction	Concentration (mg/mL)	Total protein (mg)	Rate (nmol/min)	Total activity (nmol/min)	Specific activity (nmol/min*mg)	Percent Yield	Fold Purification
Prep 1[a]	lysate	9.5	380	1.2E-04	2.3E-01	6.12E-04	100	1.0
	cleared lysate (10000 x g)	8.0	304	7.5E-05	1.4E-01	4.70E-04	61	0.8
Prep 2	lysate	12.9	514	351	4.7E+06	9.1E+03	100	1
	cleared lysate (8000 x g)	6.7	270	318	4.2E+06	1.6E+04	91	2

a. Numbers are given for two PglB preparations representative of numbers seen in additional preps.

The supernatant ('cleared lysate') is then subjected to a second centrifugal spin at 150000 x g, which pellets the membrane fraction, and soluble proteins remain in the supernatant. The membrane fraction, or the 'cell-envelope fraction' (CEF), is then resuspended as a semi-pure fraction containing only the cell-membrane content (cell-membrane lipids and all membrane-associated proteins). A good percent yield (90-95%) and high fold purification (3-6 fold) were observed with standard parameters; thus, it was not necessary to optimize this step further.

A subsequent salt wash of the membrane fraction involves homogenizing the pellet from the first 150000 x g spin (which pellets the 'unwashed CEF') in a buffer containing a high concentration of salt. The high salt presumably disrupts electrostatic interactions between proteins associated with membrane lipids or integral-membrane proteins, such that any non-integral membrane proteins are released into the aqueous surrounding. The CEF is then repelleted in a second 150000 x g spin, which is expected to contain only integral-membrane proteins ('washed CEF'). It was of interest to determine whether the identity of the salt would play a role in the effectiveness of the wash.

Table 5: Comparison of several salt solutions in CEF salt wash efficiency.

Fraction	Total Protein (mg)	Rate (nmol/min)	Total activity	Specific activity	Percent Yield	Fold Purification
NaCl CEF (pre-wash)	143	5.6E-06	7.2E-02	5.0E-04	100	1.0
NaCl CEF (post-wash)[a]	75	2.5E-06	3.6E-02	4.7E-04	49	0.9
KCl CEF (pre-wash)	152	5.2E-06	6.8E-02	4.5E-04	100	1.0
KCl CEF (post-wash)	75	1.6E-06	2.5E-02	3.3E-04	36	0.7
NaCl + KCl CEF (pre-wash)	189	6.2E-06	8.0E-02	4.2E-04	100	1.0
NaCl + KCl CEF (post-wash)	75	3.6E-06	5.4E-02	7.2E-04	68	1.7

a. The percent yield and fold-purification values for 'post-wash' fractions were determined relative to the pre-washed state only.

To test this possibility, a comparison was made between three solutions with the following salt contents: 500 mM NaCl, 500 mM KCl, 250 mM NaCl + 250 mM KCl (**Table 5**). The results show the fold purification and percent yield are highest for the CEF wash performed using the combination of KCl and NaCl. While the salt wash using the combined salt solution was advantageous, the use of solutions of only 500 mM KCl or 500 mM NaCl surprisingly resulted in both unfavorable yields and fold purification. The comparison was repeated to show reproducibility of this result, and the combination salt wash was used routinely thereafter[1]. At this point, the CEF is resuspended in a small volume (less than 10 mL lysis buffer) so that membrane structure remains intact and will provide stability. An additional 10-20% glycerol can be added as a cryoprotectant. This concentrated solution can be dispensed into aliquots and stored at -80°C, where it remains stable for a year or more.

---

[1] It is worth noting that the specific activity measurements were indispensable for determining that this customary purification step is actually obstructive for PglB purification unless it is carried out in specific salt conditions.

Solubilization of membrane proteins

Solubilization of the CEF involves addition of a high concentration of detergent to the membrane fraction, such that the native lipid structure surrounding the membrane proteins is disrupted and the membrane proteins may be released into solution. The lost lipid periphery on these proteins is presumably replaced with stabilizing detergent micelles, or a mixture of lipids and detergent. The identity and concentration of the detergent used imparts highly variable yields of active protein. However, many detergents are very costly and large detergent screens are often time-consuming. Rather than investigating a large range of detergents, four detergent conditions were initially screened to resolve the amount of optimization that needed to be performed. The detergents initially screened were DDM (*n*-dodecyl-β-D-maltoside), OG (*n*-octyl-β-D-glucoside), Triton X-100, and an equal combination of the former three. A final concentration of 0.5% (w/v) was used for the four detergent conditions (see *Materials and Methods*). These detergents were chosen for the following reasons: DDM has traditionally been successful in the solubilization of membrane proteins [30,31], OG has a high CMC (critical micelle concentration) which facilitates easy removal of detergent at later points, and Triton X-100 is economical and oft-used in protein biochemistry. The concentration of detergent (0.5% w/v) was used as an initial trial concentration based on previous reports of OT solubilization efforts [32]. While it was not necessary in the presently described case, the critical micelle concentration (CMC) – the concentration of detergent at which detergent micelles form from the individual detergent molecules—is an important factor that can guide detergent choice and concentration. Often a range of detergents are screened concomitantly at 40X CMC of each detergent, which leads to very different percentages of detergent in each solution tested. However, it also important to note that reported CMC values are estimates and the true CMC is highly dependent on specific environment, including buffer (identity and concentration of buffer used), salt concentration, temperature, and any other buffer additives.

Table 6: Comparison of several detergent solutions in CEF-solubilization efficiency.

Fraction	Total protein (mg)	Rate (nmol/min)	Total activity (nmol/min)	Specific activity (nmol/min*mg)	Percent yield	Fold purification
pre-solubilized[a]	**27.9**	**1.4E-05**	**8.3E-03**	**3.0E-04**	**100**	**1.0**
post-DDM[b]	**9**	**1.2E-05**	**7.1E-03**	**7.9E-04**	**86**	**2.7**
post-OG	**2.7**	**1.1E-06**	**6.6E-04**	**2.4E-04**	**8**	**0.8**
post-Triton	**6.3**	**9.8E-07**	**5.9E-04**	**9.3E-05**	**7**	**0.3**
post-combination	**7.2**	**7.0E-06**	**4.2E-03**	**5.8E-04**	**51**	**2.0**

a. Percent yield and fold-purification values for 'post-solubilized' fractions were determined relative to pre-solubilized fraction.

b. Fraction after solubilization with DDM. All conditions had a final detergent concentration of 0.5%.

A 5% solution of detergent is diluted 10-fold into the resuspended CEF, which has been diluted to 10 mg/mL total protein. This mixture is homogenized and vortexed vigorously, followed by dilution to lower the detergent concentration (see *Materials and Methods*). Results (**Table 6**) indicate clearly that DDM is most efficient at solubilizing active PglB relative to the other detergent solutions tested. The use of 0.5% DDM gives a very favorable percent yield, eliminating the need to further optimize this step. In the common case that initial detergent conditions fail to provide satisfactory results, one can expand the comparison to include additional detergents with a range of properties and vary the concentration of the detergents.

Affinity chromatography

The specificity of PglB affinity purification using Ni-NTA affinity was inefficient under batch binding conditions initially used. Initial concentrations for batch-binding were 50 mM HEPES, pH 7.5, 0.17% DDM, 10 mM imidazole. A high relative concentration of contaminating proteins co-eluted with PglB and recovery of PglB from the column was incomplete. There are several additives which are known to decrease non-specific binding to Ni-NTA resin and between proteins [33]. **Figure 7A** shows a gel of PglB elutions after batch binding with several different conditions. Addition of glycerol to all buffers and increasing imidazole from 10 to 20 mM successfully decreased the amount of the low-molecular weight impurities relative to the full-length PglB. Diluting the solution two-fold during batch binding significantly decreased impurities as well; however, only when batch binding proceeded for a longer duration did the dilution result in increased purity without loss of significant protein in flow through. This is not surprising, because in general, longer batch binding time is required for maximal binding of a more dilute solution of His-tagged proteins. For PglB, batch binding diluted, solubilized CEF to fresh Ni-NTA resin overnight, in the presence of glycerol and increased imidazole, resulted in the highest purity with the least loss of protein in the flow through and washes.

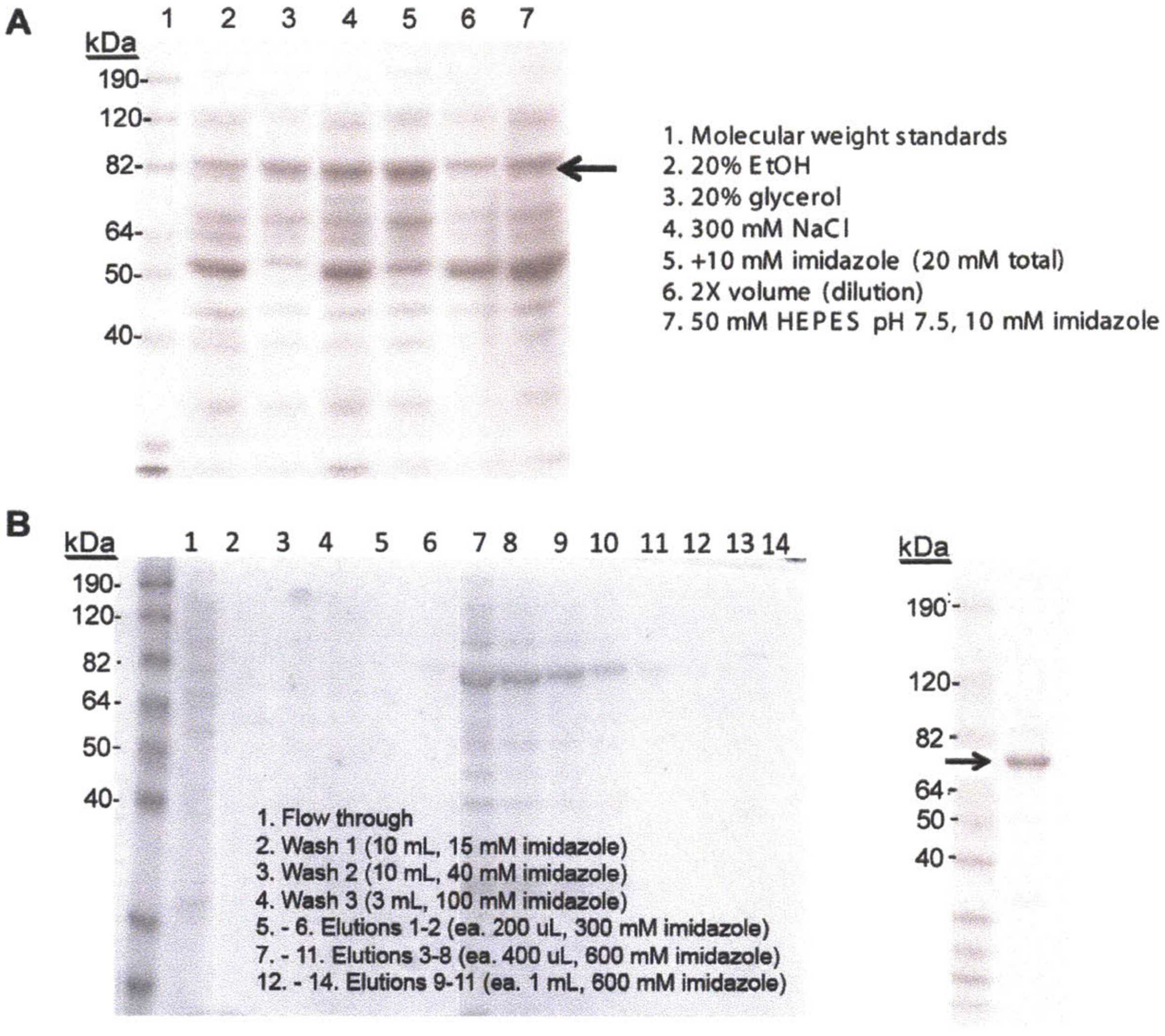

A
kDa
1 2 3 4 5 6 7
190-
120-
82-
64-
50-
40-
1. Molecular weight standards
2. 20% EtOH
3. 20% glycerol
4. 300 mM NaCl
5. +10 mM imidazole (20 mM total)
6. 2X volume (dilution)
7. 50 mM HEPES pH 7.5, 10 mM imidazole
B
kDa
1 2 3 4 5 6 7 8 9 10 11 12 13 14
190-
120-
82-
64-
50-
40-
1. Flow through
2. Wash 1 (10 mL, 15 mM imidazole)
3. Wash 2 (10 mL, 40 mM imidazole)
4. Wash 3 (3 mL, 100 mM imidazole)
5. - 6. Elutions 1-2 (ea. 200 uL, 300 mM imidazole)
7. - 11. Elutions 3-8 (ea. 400 uL, 600 mM imidazole)
12. - 14. Elutions 9-11 (ea. 1 mL, 600 mM imidazole)
kDa
190-
120-
82-
64-
50-
40-

Figure 7: Optimization of Ni-NTA purification.

A. Batch binding conditions screened for effect on PglB purity after Ni-NTA column. Lane 1 contains the Benchmark pre-stained standard (Invitrogen), Lanes 2-6 contain standard buffer of 50 mM HEPES, pH 7.5, 10 mM imidazole, 0.17% DDM, plus the additive noted to the right of the gel. Buffer 7 contained only standard buffer. *B.* Improved purification of PglB over Ni-NTA column. Coomassie-stained SDS-PAGE of flow through, washes, and elutions are shown (*left*). Elution fractions in lanes 7-11 were combined and concentrated, shown *right*. Apparent improvement in purity in the concentrated solution can be attributed to the concentration method using a 100-kDa MWCO filter, which allows smaller impurities to flow through. In addition, PglB dispersity becomes more homogenous during concentration, removing the higher order oligomers and thus decreasing intensities of bands above 82 kDa.

Another alteration that significantly affected the purification was replacement of the His6 tag (native to the pET24a(+) vector) with a His10 tag. This change was made during expression optimization, and thus all purification optimization heretofore described was performed with the His10 tag. However, it is mentioned because lengthening the His-tag did cause a clear increase in the affinity of PglB for the Ni-NTA resin. For this reason, much higher imidazole concentrations are used in column washes than are considered typical, and similarly very high imidazole is required to elute all the bound PglB from the resin (up to 600 mM from 300 mM used with the His6 tag). In fact, using the typical concentration of 300 mM imidazole for elution recovers only a fraction of the bound PglB. Accordingly, very little protein remained in the flow through and washes based on gel analysis (**Figure 7B**) and activity assays. The batchbinding changes resulted in a final value fold-purification of approximately 500 (**Table 3**), relative to the ~50-fold purification values achieved before batchbinding optimization and increasing imidazole in elution buffer.

Buffer Exchange

It has been observed that PglB activity is rapidly lost after elution from the Ni-NTA column due to prolonged exposure to very high imidazole concentration. Dialysis, which was previously used to exchange the buffer and remove the imidazole after elution, proceeded too slowly for PglB to maintain activity. Therefore, an alternative method was used to exchange buffer: a HiTrap desalting column (G.E. Healthcare) was used immediately after elution to replace the elution buffer with one that lacked imidazole. However, use of the desalting column diluted the protein and caused significant loss of activity. Next, an Amicon Ultra 100-kDa MWCO cellulose filter was used, which concentrates the fractions containing PglB, and also serves as a method to exchange the buffer and remove the imidazole (**Figure 7B**). While the molecular weight of PglB (82 kDa) is below the molecular-weight cut-off of 100 kDa, the mass is high enough that there is virtually no PglB lost in the filtrate. This may in part be due to the presence of the DDM detergent micelle surrounding PglB. In contrast, the DDM detergent micelle alone (without PglB bound) flows through the 100 kDa-MWCO filter despite the reasonably large size of the DDM micelle (~ 50 kDa) [34]. This selectivity is very advantageous because it allows the protein to be concentrated without simultaneously concentrating the detergent to a degree that is deleterious to the protein stability. In addition, the loss of lower molecular-weight impurities is consistently observed when using these high molecular-weight cut-off filters presumably due to the fact that these contaminants can pass through the filter.

In summary, this method of buffer exchange results in negligible loss of activity whereas dialysis of a PglB solution overnight at 4°C results in complete loss of enzyme activity. In fact, it was found that PglB solutions lose activity when dialyzed overnight, even in the absence of imidazole. Thus, it is likely that the dialysis procedure was a cause of activity loss in PglB in addition to the high imidazole. Using the 100-kDa MWCO Amicon filter, Ni-NTA-purified PglB is concentrated and buffer-exchanged to roughly 10 μM in 50 mM HEPES pH 7.5, 100 mM NaCl, 30% glycerol, 0.01% DDM. This solution is distributed into aliquots and stored at -80°C and, under these conditions, remains stable for at least two months (**Figure 8**).

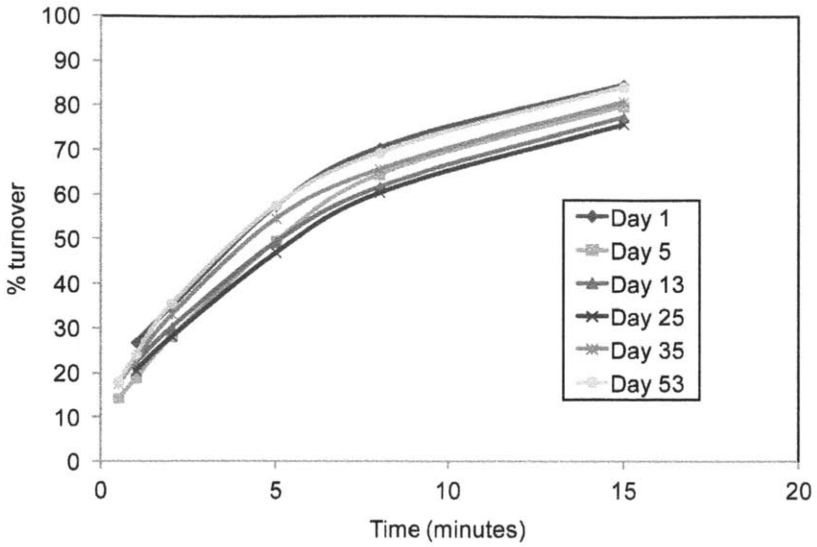


Figure 8: Activity of PglB over multiple days post-purification.
Pure PglB (shown in **Figure 7B**) was distributed into aliquots and stored at -80^{0}C. An aliquot from this batch was used for each activity measurement shown.

PglB Purification protocol

1L culture yields ~ 25 grams of cells

All steps performed on ice or at 4°C

Purification of ~3.5 gram pellet:

1. Resuspension: Remove pellet from -80°C storage. Add Buffer 1 at a volume for ~ 0.1 grams of pellet weight per mL. Agitate to resuspend. Add protease inhibitor cocktail solution, DNase and lysozyme at product manual recommended concentrations. Agitate solution gently for ~ 1 hour.

2. Lysis: Sonication of lysis solution on ice:
    a. Between rounds of sonication lysis solution should be cooled on ice and periodically inverted to mix.
    b. 3 rounds of sonication at the following settings: Time: 1 min, Amplitude: 50%, Pulse: 1 second on/1 second off.

3. Differential centrifugation, Round I: 8000 x g, 4°C, 35'

4. Decant supernatant. Save supernatant. Pellet can be discarded (or sample saved for blot analysis).

5. Differential centrifugation, Round 2: 150000 x g 4°C, 60'

6. Decant supernatant. Save pellet (contains crude cell envelope fraction, or "CEF"). Supernatant may be discarded (or sample saved for blot analysis).

7. CEF salt wash: Quantitatively transfer cell pellet to homogenizer. Use spatula to transfer as much of the pellet as possible. Then add 2 mL Buffer 2 to centrifuge tube. Scrape bottom of centrifuge tube to resuspend remaining CEF. Add this solution to homogenizer containing rest of the CEF. Homogenize cell pellet in a final volume of 35 mL of Buffer 2 (33 mL added, after the 2 mL wash).

8. Differential centrifugation, Round 3: return homogenized CEF to centrifuge tube. Spin, at 150000 x g 4°C, 60'

9. Decant supernatant. Save pellet (contains washed cell envelope fraction, or "CEF"). Supernatant may be discarded (or sample saved for blot analysis).

10. Transfer washed CEF quantitatively to homogenizer. Use 2 mL of Buffer 3 to wash CEF remaining in centrifuge tube. Add this solution to homogenizer containing dry CEF pellet.

11. [Sample may be homogenized, aliquotted, and frozen at -80° at this point and purification continued at a later date. To continue, thaw stored CEF on ice, pour into homogenizer, wash container with 200 μL of buffer and combine with neat CEF solution.]

12. Solubilization of washed CEF:
    a. Homogenize CEF at for a concentration of 10 mg/mL protein
    b. Add 1/10 volume of 5% DDM solution, → protein: detergent ratio of 2
    c. Re-homogenize
    d. Vortex 2 minutes
    e. Dilute 1:3 in Buffer 3 for a final concentration of 30% glycerol, 0.17% DDM, 20 mM imidazole, and 3.3 mg/mL protein)

13. Differential centrifugation, Round 4: spin at 100000 x g, 4°C, 60'.

14. Decant supernatant. Save supernatant. Pellet may be discarded.

15. Batchbinding to Ni-NTA: 1-2 mL Ni-NTA resin equilibrated in 50 mM HEPES pH 7.5, 30% glycerol, 20 mM imidazole, 0.01% DDM. Add supernatant to equilibrated resin and gently agitate overnight.

16. Ni-NTA purification: let each wash or elution fraction completely flow through before adding the following fraction.
    a. Flow through of solution over gravity flow filter.
    b. Wash 1: 1 x ~ 25 mL Buffer 4 + 20 mM imidazole.
    c. Wash 2: 3 x 1 mL Buffer 4 + 80 mM imidazole.
    d. Elution A: 6 x 0.5 mL fractions with Buffer 5 + 300 mM imidazole
    e. Elution B: 6 x 0.5 mL fractions with Buffer 5 + 600 mM imidazole

17. Measure activity of elution fractions (or use SDS-PAGE) to determine fractions containing protein.

18. Concentrate and buffer exchange using 100 -kDa concentrator, exchange into Buffer 5

19. Storage of solution in aliquots at -80°C.

Buffer 1: 50 mM HEPES, pH 7.5; 10% glycerol, 100 mM NaCl.
Buffer 2: 50 mM HEPES, pH 7.5; 10% glycerol, 250 mM NaCl, 250 mM KCl.
Buffer 3: 50 mM HEPES pH 7.5, 30% glycerol, 20 mM imidazole.
Buffer 4: 50 mM HEPES pH 7.5, 20% glycerol, 300 mM NaCl, 0.01% DDM
Buffer 5: 50 mM HEPES pH 7.5, 30% glycerol, 100 mM NaCl, 0.01% DDM

Discussion and Conclusion

The purification values shown in **Table 3** were recorded for a 3.6 g cell pellet, which represents roughly one seventh of the cell weight obtained on average from one liter of culture using auto-induction. The 0.6 mg protein remaining at this point is > 95% purified PglB as judged by SDS-PAGE (**Figure 7B**). Thus, we estimate that per liter of culture, using auto-induction expression and the optimized purification conditions, 1-5 mg pure PglB is obtained. This represents a 100-fold increase in levels of pure, stable PglB relative to pre-optimized levels, which were estimated to be ~50 μg per liter of culture. Importantly, multiple PglB preparations have been completed using the current optimized protocol, demonstrating the reproducibility of yields obtained, and indicating that the values shown in **Table 3** are representative of an average purification.

Nonetheless, the optimized procedure for purifying PglB is far more time-consuming than purification of most soluble proteins. The obligatory steps of isolating and processing the *E. coli* membrane fraction add considerable time and effort to the purification process. Additionally, a greater emphasis is placed on maintaining the protein at a cold temperature, working rapidly, and following protocols strictly. Furthermore, optimized quantities of membrane proteins often seem scant relative to typical yields for soluble proteins. These difficulties, however, are well-recognized by those intending to express and purify an uncharacterized membrane protein, and there are many thorough reviews that provide an overview of these challenges and potential solutions [35-38]. While optimizing PglB expression and purification, several additional obstacles and solutions were encountered, and it is hoped that these may be beneficial for non-specialists to consider when commencing such an effort.

Ironically, it proved challenging to navigate the abundant options available that have potential to improve the expression or stability of a membrane protein. Various cell lines, fusion tags, expression-culture additives, detergent alternatives, protein stabilizers, and countless other products and methods have reported instances of success in handling of certain membrane proteins. Yet, there are very few guarantees, and optimizing most steps remains an empirical process. In the case of PglB, it proved to be more time-efficient and economical to begin optimization of each step with a limited range of commonly used conditions and expand based on those results, when necessary. Deciding on a standard for sufficiency is useful, and when a step is particularly problematic, it may be practical to allot a specific amount of time to the task. If progress is not made by the end of the time period, perhaps it is wiser to try another route or target entirely.

It was also found helpful to develop a clear strategy from the onset, and to begin with the first step of the procedure. Once the first step of the procedure has been optimized, the next step of the procedure should only then be addressed, because conditions used at each step in expression and purification can have unpredictable results on those that follow.

Along the same lines, one should carefully choose the specific construct to optimize. Ideally it should be the same construct that will be used for anticipated future experiments. An optimal procedure for expression and purification for a protein in one vector can be significantly different in another vector, or when adding or exchanging the expression tag. The importance of designing a systematic procedure for optimization and an objective measure of comparison cannot be underestimated. The inevitably qualitative nature of many aspects of protein biochemistry—resulting from variability in individual transformation colonies, expression cultures, loading of gels and western-blot transfers—multiplies the uncertainty of qualitative judgments.

The presented optimization of PglB expression and purification was performed with the ultimate purpose of kinetic and mechanistic characterization, rather than crystallographic studies. Implications of this goal include a primary interest in obtaining high quantities of pure, active, and stable protein. In contrast, crystallography-driven optimizations often prioritize obtaining monodisperse, concentrated protein, which are absolute requirements for obtaining a crystal structure.

Because the goal of the presented procedure is biochemical, SDS-PAGE and western blots alone represent poor methods for comparing the optimal conditions for expression and purification of the enzyme. While a condition used in a purification step may yield a bright band on a western blot, it may be unfavorable for activity yields. If constructs with tags of various sizes are screened, the differences in molecular weight can be a perplexing factor to account for in gel densitometry measurements. Additionally, expression tags can have effects on structure, activity and monodispersity of the enzyme, which cannot be accounted for using SDS-PAGE. Purifying a protein based on specific activity and yields allows a more rigorous and precise measure of the most efficient purification condition to use for each step.

However, the requirement for a quantitative activity assay can be problematic, as activity assays are enzyme-specific, and reagents and equipment can make the assays laborious or expensive to perform frequently. Undeniably, in the developing field of membrane-protein biology, most information acquired is important and useful even if lacking in certain methodological precision. Though, while it may be exciting to speculate on results from experiments and techniques available, it is equally important to be conscientious of the questions that these experiments leave unanswered.

Materials and Methods

Vectors and cloning

The PglB gene was amplified by PCR from the *C. jejuni* genome NCTC 11168 [39,40]. Primers used in the PCR encoded a BamHI site at the N-terminus prior to the start codon and His10-UGA-XhoI on the C-terminus prior to the native stop codon. The PCR product was purified and digested with BamHI and XhoI (New England Biolabs, NEB) and ligated into the corresponding sites in the pET24a(+) vector (Invitrogen) using T4 DNA ligase (NEB) and standard molecular biology procedures. The resulting vector was sequenced and then transformed using manufacturer-supplied protocols into BL21 (DE3) RIL *E. coli* competent cells (Agilent) for expression. Additional vectors screened for expression of PglB with alternate fusion tags include pGEX with Glutathione-S-transferase (GST) (GE Healthcare), pMAL-c2X with MBP (NEB), pET SUMO with SUMO (Invitrogen), pET Trx with Thioredoxin (EMD Millipore), and pGBH with GB1 [23] (**Table I**).

Protein Expression

Pre-optimized expression was carried out according to the following procedure except when specifically noted: 5-mL solutions of LB at 25g/L were at 37°C, until reaching an O.D.= 0.6-0.8. At this point the temperature was turned to 16°C and cultures were induced by adding IPTG to a final concentration of 1 mM. Cultures were left to shake overnight. The following day, cells were harvested and lysed according to the conditions described below. Competent cells screened included BL21 (DE3) RIL, BL21 (DE3) Gold, BL21 (DE3) RP, BL21 (DE3) pLys (all BL21 strains from Agilent), Rosetta 2 (DE3) (Novagen), Rosetta gami-2 (DE3) (Novagen), and C41 (Lucigen). In contrast, the optimized expression procedure involves using auto-induction media

ZYM-5052, a high-density growth media [29]. For auto-induction expression, one liter of auto-induction media was made up in a 6-liter flask to allow adequate aeration of the cultures. 10 g tryptone and 5 g yeast extract were combined with 960 mL of deionized water and autoclaved. Once the media had cooled, Kanamycin and Chloramphenicol were added for final concentrations of 100 μg/mL and 170 μg/mL, respectively. Just before inoculation, the following media components were added:

- 25 mL of 40X Salt Solution: 1 M Na_2HPO_4, 1 M KH_2PO_4, 2 M NH_4Cl, 0.2 M Na_2SO_4
- 20 mL of 50X 5052: 25% glycerol (v/v), 2.5% glucose (w/v), 10% α-lactose monohydrate (w/v)
- 0.2 mL of 1000X trace metals
- 2 mL of 1 M $MgSO_4$

This gives the final composition of ZYM-5052: 1% N-Z-amine (tryptone), 0.5% yeast extract, 25 mM Na_2HPO_4, 25 mM KH_2PO_4, 50 mM NH_4Cl, 5 mM Na_2SO_4, 2 mM $MgSO_4$, 0.2X trace metals, 0.5% glycerol, 0.05% glucose, 0.2% lactose. Optimized expression is carried out by inoculating the 1L auto-induction media with a 5-mL culture grown from a recent transformation of the PglB vector into BL21 (DE3) RIL. The inoculated media is grown at 37°C for 4-5 hours shaking at 200 RPM. The temperature is then turned down to 16°C and cultures are left shaking at 200 RPM overnight. The following day, cultures are harvested and cell pellets are weighed and stored at -80°C until purification.

Activity Assay

The PglB activity assay has been described in detail elsewhere [41]. Briefly, to a tube of radiolabeled substrate, Und-PP-Bac-[^{3}H]GalNAc, DMSO (10 µL), 2X PglB Assay Buffer containing 100 mM HEPES, pH 7.5 / 280 mM sucrose / 2.4% (v/v) Triton X-100 (100 µL), 1 M $MnCl_2$ (2 µL), H_20 (73 µL), and 5 µL PglB fraction are combined. The assay is initiated by the addition of 10 µL of a 2 mM stock of the peptide Acetyl-DQNAT-*p*NF in DMSO [42]. Time points of the reaction are taken by quenching aliquots of the reaction in biphasic solutions of 3:2:1 $CHCl_3$: MeOH: 4 mM $MgCl_2$ (1.2 mL). The aqueous layer is then extracted and the organic layer washed twice with theoretical upper phase (TUP) with salt (2 x 600 µL). The aqueous layer and washes are combined and mixed with 5 mL of scintillation fluid (Ecolite, MP Biomedicals). The organic layer was mixed with 5 mL of scintillation fluid (OptiFluor, Perkin Elmer) and all fractions are subjected to scintillation counting.

Protein Quantification

Total protein in each of the purification fractions is estimated using the Bio-Rad protein assay (Bio-Rad, #500-0006) according to the manufacturer-supplied protocol. Bovine Serum Albumin (BSA) standard (Thermo Scientific Pierce) is used to create a standard curve. Fractions are taken during each step of the purification from a given cell pellet and stored at -80°C. Upon purifying to the desired end-point, the protein quantities for all samples are measured at the same time to minimize error. Pure PglB protein is quantified by measuring the absorbance at 280 nm and using an extinction coefficient of 117,300 M^{-1} cm^{-1} (for T7-PglB-His_{10}).

Cell lysis

All steps of purification are performed on ice or at 4°C. A cell pellet of 3-5 g in weight is thawed on ice and resuspended in 50 mM HEPES pH 7.5, 10% glycerol, 100 mM NaCl for 0.1 grams of cell pellet per mL. Protease inhibitor cocktail solution, EDTA-free (Calbiochem) and hen-egg lysozyme powder (Amresco) is added for 1 μL/mL and 1 mg/mL, respectively. Mixture is agitated gently for ~1 hour. Cell lysis is performed by sonication (Sonics Vibracell, VC 505 (500 watts) & VC 750 (750 watts)). Sonication is performed on ice for 3 x 1 min at 50% amplitude, pulsing at 1 s on/1 s off, and with breaks between cycles to prevent warming of the mixture. Lysates are centrifuged at 8000 x g, 4°C for 35 minutes to remove unlysed cells and insoluble cell debris. The supernatant (cleared lysate) is decanted and pellet is discarded.

Isolation of Cell-Envelope Fraction (CEF)

The cleared lysate is centrifuged at 150000 x g, 4°C for 60 minutes. After the spin, the supernatant is discarded. The pelleted fraction (the CEF) is transferred as quantitatively as possible to a Pyrex homogenizer using a small spatula. Two mL of high-salt buffer (50 mM HEPES pH 7.5, 10% glycerol, 250 mM NaCl, 250 mM KCl) is then added to the centrifuge tube and the remaining CEF is resuspended by gently scraping the bottom of the tube with the spatula. This wash solution is added to the homogenizer containing the rest of the CEF. The CEF is homogenized in a final volume of 35 mL of the high-salt buffer (33 mL added, after the 2 mL wash). The homogenized CEF is returned to the centrifuge tube and this solution is centrifuged at 150000 x g, 4°C for 60 minutes. Again, the supernatant is discarded. The pellet, containing the washed CEF, is again transferred and homogenized in 5-10 mL of 50 mM HEPES pH 7.5, 30% glycerol, 20 mM imidazole. This washed CEF fraction is either stored at -80°C or purification is continued.

Extraction of membrane proteins from cell membrane

The optimization procedure for PglB solubilization from the membrane is modeled after one used previously for the yeast oligosaccharyl transferase [32]. Washed CEF is solubilized by homogenizing in 50 mM HEPES pH 7.5, 30% glycerol, 20 mM imidazole in a volume that roughly yields a final concentration of 10 mg total protein per mL. The 5% DDM detergent solution (in 50 mM HEPES pH 7.5) is diluted 1:10 in the homogenized CEF volume for a final concentration of 0.5% detergent for solubilization. This 0.5% (w/v) concentration corresponds to roughly 200X CMC of DDM (100X CMC of Triton X-100, and 9X CMC of OG, also screened at 0.5%) [34]. This solution is re-homogenized thoroughly, and the solution is vortexed at the maximum setting for two minutes. The solution is then diluted 1:3 in 50 mM HEPES pH 7.5, 30% glycerol, 20 mM imidazole, giving a final concentration of 50 mM HEPES pH 7.5, 30% glycerol, 0.17% DDM, 20 mM imidazole, and roughly 3.3 mg/mL protein. The solubilized CEF is centrifuged at 100000 x g, 4°C for 60 minutes. The supernatant is decanted into a clean, pre-chilled 50-mL conical tube for further purification, and the pellet is discarded.

Ni-NTA purification

The supernatant, containing the solubilized membrane-proteins, is added to Ni-NTA agarose resin (Qiagen) that has been pre-equilibrated in 50 mM HEPES pH 7.5, 30% glycerol, 20 mM imidazole, 0.01% DDM. The protein solution is gently agitated overnight with the resin. Ni-NTA purification is carried out with Buffer M (50 mM HEPES pH 7.5, 20% glycerol, 300 mM NaCl, 0.01% DDM) plus the specified amount of imidazole used for washes and Buffer E (50 mM HEPES pH 7.5, 30% glycerol, 100 mM NaCl, 0.01% DDM) plus specified imidazole used for elutions. Batchbinding solution is allowed to flow through a disposable gravity-filtration column (BioRad). The following fractions are collected: Flow through, Wash A: 1 x 25 mL Buffer M + 40 mM imidazole. Wash B: 3 x 1 mL Buffer M + 100 mM imidazole, Elution A: 6 x 0.5 mL fractions of Buffer E + 300 mM imidazole, Elution B: 6 x 0.5 mL Buffer E + 600 mM imidazole. Five microliters of every other elution is removed and used to determine the location of the active PglB using the activity assay. Fractions containing most activity are combined. The combined solution is concentrated and buffer is exchanged using an Amicon Ultra-100K centrifugal filter (Millipore) into Buffer E (no imidazole). SDS-PAGE is used to verify purity and then the concentration of the pure PglB solution is quantified using UV-absorbance at λ = 280 nm. The solution is then divided into aliquots and stored at -80°C.

References

1. Ahram M, Litou ZI, Fang R, Al-Tawallbeh G (2006) Estimation of membrane proteins in the human proteome. In Silico Biol 6: 379-386.
2. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5: 993-996.
3. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M (2007) Drug-target network. Nat Biotechnol 25: 1119-1126.
4. Katzen F, Peterson TC, Kudlicki W (2009) Membrane protein expression: no cells required. Trends Biotechnol 27: 455-460.
5. Sarkar CA, Dodevski I, Kenig M, Dudli S, Mohr A, et al. (2008) Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. Proc Natl Acad Sci U S A 105: 14808-14813.
6. Chen Q, Miller LJ, Dong M (2010) Role of N-linked glycosylation in biosynthesis, trafficking, and function of the human glucagon-like peptide 1 receptor. Am J Physiol Endocrinol Metab 299: E62-68.
7. Luthi F, Leibundgut K, Niggli FK, Nadal D, Aebi C, et al. (2012) Serious medical complications in children with cancer and fever in chemotherapy-induced neutropenia: results of the prospective multicenter SPOG 2003 FN study. Pediatr Blood Cancer 59: 90-95.
8. Ubelhart R, Bach MP, Eschbach C, Wossning T, Reth M, et al. (2010) N-linked glycosylation selectively regulates autonomous precursor BCR function. Nat Immunol 11: 759-765.
9. Londrigan SL, Turville SG, Tate MD, Deng YM, Brooks AG, et al. (2011) N-linked glycosylation facilitates sialic acid-independent attachment and entry of influenza A viruses into cells expressing DC-SIGN or L-SIGN. J Virol 85: 2990-3000.
10. van Berkel PH, Gerritsen J, Perdok G, Valbjorn J, Vink T, et al. (2009) N-linked glycosylation is an important parameter for optimal selection of cell lines producing biopharmaceutical human IgG. Biotechnol Prog 25: 244-251.
11. Bartusik D, Aebisher D, Lyons AM, Greer A (2012) Bacterial inactivation by a singlet oxygen bubbler: identifying factors controlling the toxicity of (1)O2 bubbles. Environ Sci Technol 46: 12098-12104.
12. Deniaud A, Bernaudat F, Frelet-Barrand A, Juillan-Binard C, Vernet T, et al. (2011) Expression of a chloroplast ATP/ADP transporter in E. coli membranes: behind the Mistic strategy. Biochim Biophys Acta 1808: 2059-2066.
13. Lizak C, Fan YY, Weber TC, Aebi M (2011) N-Linked glycosylation of antibody fragments in Escherichia coli. Bioconjug Chem 22: 488-496.
14. Wei Z, Lin T, Sun L, Li Y, Wang X, et al. (2012) N-linked glycosylation of GP5 of porcine reproductive and respiratory syndrome virus is critically important for virus replication in vivo. J Virol 86: 9941-9951.
15. Hurt JK, Fitzpatrick BJ, Norris-Drouin J, Zylka MJ (2012) Secretion and N-linked glycosylation are required for prostatic acid phosphatase catalytic and antinociceptive activity. PLoS One 7: e32741.
16. Hillaire ML, van Eijk M, Nieuwkoop NJ, Vogelzang-van Trierum SE, Fouchier RA, et al. (2012) The number and position of N-linked glycosylation sites in the hemagglutinin

determine differential recognition of seasonal and 2009 pandemic H1N1 influenza virus by porcine surfactant protein D. Virus Res 169: 301-305.

17. Somnuke P, Hauhart RE, Atkinson JP, Diamond MS, Avirutnan P (2011) N-linked glycosylation of dengue virus NS1 protein modulates secretion, cell-surface expression, hexamer stability, and interactions with human complement. Virology 413: 253-264.
18. Graslund S, Sagemark J, Berglund H, Dahlgren LG, Flores A, et al. (2008) The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. Protein Expr Purif 58: 210-221.
19. Maita N, Nyirenda J, Igura M, Kamishikiryo J, Kohda D (2010) Comparative structural biology of eubacterial and archaeal oligosaccharyltransferases. J Biol Chem 285: 4941-4950.
20. Lizak C, Gerber S, Numao S, Aebi M, Locher KP (2011) X-ray structure of a bacterial oligosaccharyltransferase. Nature 474: 350-355.
21. Jaffee MB, Imperiali B (2011) Exploiting topological constraints to reveal buried sequence motifs in the membrane-bound N-linked oligosaccharyl transferases. Biochemistry 50: 7557-7567.
22. Nilsson J, Larsson M, Stahl S, Nygren PA, Uhlen M (1996) Multiple affinity domains for the detection, purification and immobilization of recombinant proteins. J Mol Recognit 9: 585-594.
23. Bao WJ, Gao YG, Chang YG, Zhang TY, Lin XJ, et al. (2006) Highly efficient expression and purification system of small-size protein domains in Escherichia coli for biochemical characterization. Protein Expr Purif 47: 599-606.
24. Panavas T, Sanders C, Butt TR (2009) SUMO fusion technology for enhanced protein production in prokaryotic and eukaryotic expression systems. Methods Mol Biol 497: 303-317.
25. Sachdev D, Chirgwin JM (1998) Solubility of proteins isolated from inclusion bodies is enhanced by fusion to maltose-binding protein or thioredoxin. Protein Expr Purif 12: 122-132.
26. Davies AH, Jowett JB, Jones IM (1993) Recombinant baculovirus vectors expressing glutathione-S-transferase fusion proteins. Biotechnology (N Y) 11: 933-936.
27. Pryor KD, Leiting B (1997) High-level expression of soluble protein in Escherichia coli using a His6-tag and maltose-binding-protein double-affinity fusion system. Protein Expr Purif 10: 309-319.
28. Xie H, Guo XM, Chen H (2009) Making the most of fusion tags technology in structural characterization of membrane proteins. Mol Biotechnol 42: 135-145.
29. Studier FW (2005) Protein production by auto-induction in high-density shaking cultures. Protein Expression and Purification 41: 207-234.
30. VanAken T, Foxall-VanAken S, Castleman S, Ferguson-Miller S (1986) Alkyl glycoside detergents: synthesis and applications to the study of membrane proteins. Methods Enzymol 125: 27-35.
31. Rosevear P, VanAken T, Baxter J, Ferguson-Miller S (1980) Alkyl glycoside detergents: a simpler synthesis and their effects on kinetic and physical properties of cytochrome c oxidase. Biochemistry 19: 4108-4115.
32. Sharma CB, Lehle L, Tanner W (1981) N-Glycosylation of yeast proteins. Characterization of the solubilized oligosaccharyl transferase. Eur J Biochem 116: 101-108.
33. QIAGEN (2008) N-NTA reagent compatibility table. 1-4.

34. Bhairi SM, Mohan, C., EMD Biosciences (2007) Detergents Booklet. San Diego, CA.
35. Rosenbusch JP (2001) Stability of membrane proteins: relevance for the selection of appropriate methods for high-resolution structure determinations. J Struct Biol 136: 144-157.
36. Grisshammer R (2006) Understanding recombinant expression of membrane proteins. Curr Opin Biotechnol 17: 337-340.
37. Aebi A, Neumann P (2011) Endosymbionts and honey bee colony losses? Trends Ecol Evol 26: 494.
38. Lluis MW, Godfroy JI, 3rd, Yin H (2012) Protein engineering methods applied to membrane protein targets. Protein Eng Des Sel.
39. Taylor DE, Eaton M, Yan W, Chang N (1992) Genome maps of Campylobacter jejuni and Campylobacter coli. J Bacteriol 174: 2332-2337.
40. Karlyshev AV, Henderson J, Ketley JM, Wren BW (1998) An improved physical and genetic map of Campylobacter jejuni NCTC 11168 (UA580). Microbiology 144 (Pt 2): 503-508.
41. Glover KJ, Weerapana E, Numao S, Imperiali B (2005) Chemoenzymatic synthesis of glycopeptides with PglB, a bacterial oligosaccharyl transferase from Campylobacter jejuni. Chem Biol 12: 1311-1315.
42. Chen MM, Glover KJ, Imperiali B (2007) From peptide to protein: comparative analysis of the substrate specificity of N-linked glycosylation in C. jejuni. Biochemistry 46: 5579-5585.

CHAPTER 3:

SEQUENCE-DRIVEN BIOINFORMATIC ANALYSIS OF THE EVOLUTIONARY DIVERSITY OF OLIGOSACCHARYL TRANSFERASES

The majority of the data in this chapter has been published in:

Jaffee, M.B. and Imperiali, B. (2011) Exploiting topological constraints to reveal buried sequence motifs in the membrane-bound N-linked oligosaccharyl transferases. *Biochemistry.* 50(35):7557-67.

Summary

In an effort to learn more about the functionally relevant regions of PglB/STT3—specifically, to investigate the importance of the transmembrane domain for function—the sequences for several divergent STT3 homologs were analyzed with a range of bioinformatic software programs. It was noted that the topology predicted for STT3 homologs is conserved. However, the sequences of the transmembrane helices are highly divergent, preventing recognition of potential conserved motifs in full-length sequence alignments. To circumvent this issue, a computer program was developed to parse each STT3 sequence according to predicted topology and group the sequences of topologically related regions from each homolog. Alignments were then made from each group of sequence segments individually. For example, the sequence of the first predicted cytoplasmic loop from each STT3 homolog would comprise one group. This analysis was performed using 28 homologs from evolutionarily distant organisms ranging from humans to yeast to bacteria and archaea. The results show that several soluble loops between transmembrane helices contain strictly conserved motifs, suggesting a critical role of the transmembrane region in PglB/STT3 function. Sequence and topological analysis was supplemented by comparison of the secondary structure of the loops containing the conserved motifs. Secondary-structure predictions indicate the motifs share a conserved local structure, further demonstrating the importance of these loops. On a general methodological level, these studies establish the utility of combining topology prediction and sequence analysis for exposing buried pockets of homology in large membrane proteins.

Introduction

The *C. jejuni* OTase PglB comprises a single 82-kDa protein that shows homology to the catalytic subunit of the eukaryotic OTase, called Stt3 [2,3]. Both have a similar predicted architecture, with 11-13 transmembrane domains in the N-terminal region followed by a C-terminal soluble domain (**Figure 1**). The soluble domain projects into the periplasmic space (or ER lumen in eukaryotes) and contains the distinctive Stt3 signature motif: WWDxG. However, sequence identity between PglB and Stt3 is low (17.9%), and apart from the aforementioned motif, little conservation had been clearly demonstrated between PglB and other Stt3s at the time of these studies.

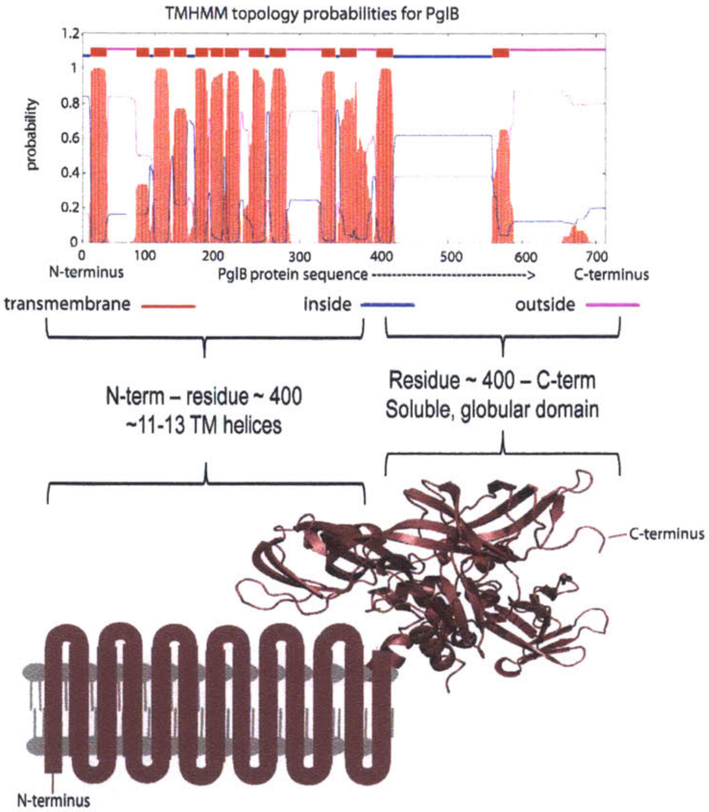


Figure 1: Translation of topological probabilities to a picture of protein architecture. *Top:* Graphical representation of the topology prediction for PglB. TMHMM algorithm uses sequence data to predict location of each residue as transmembrane, inside (cytoplasmic), or outside (luminal or extracellular). Cartoon (*bottom*) depicts the architecture predicted by the graph above, with multiple transmembrane domains followed by a C-terminal globular domain.

At the time this research was performed, little was understood of the structure-activity relationship of STT3/OTase. Crystal structures existed only for the C-terminal soluble domains of PglB (*C. jejuni*) and *Pyrococcus furiosus* Stt3. It was expected that the soluble domain contained the catalytic site, since the only known conserved and essential motif at the time (WWDxG) appears in this region. Also, the soluble domain alone is relatively simple to express and crystallize, accounting for the available structural data for this region alone [4-6]. In contrast, the many N-terminal transmembrane domains make the full-length enzyme very intractable for *in vitro* and structural studies (**Figure 1**). Yet the soluble domain by itself is neither functional nor capable of binding substrate, limiting the interpretability of the structural data [5,6].

Nonetheless, defining the residues required for OTase function presented a potential starting point for investigation of the structural prerequisites for catalysis. Therefore, although lacking activity, additional residues were suggested to function in OTase catalysis based on the structural data from the soluble domains of *P. furiosus* STT3 and *C. jejuni* PglB. The structure of the *P. furiosus* soluble domain shows an aspartic acid and a lysine (separated by two residues, DxxK) that appear to interact with the WWDxG motif (**Figure 2**). Researchers showed that mutating these residues led to decreased OTase activity, indicating their potential importance for enzyme function [5]. However, the general frequency of the DxxK pattern makes it difficult to define homologous residues in other Stt3 congeners using sequence analysis alone; thus, the level of conservation throughout evolutionarily diverse Stt3s could not be clearly established. Some additional loosely conserved residues were shown to display decreased glycosylation when mutated [4-6]. These studies generated interesting hypotheses about the importance of additional regions of the OTase, though the lack of strict conservation and quantitative biochemical studies complicated efforts to define the role of these regions in the catalysis of glycan transfer.

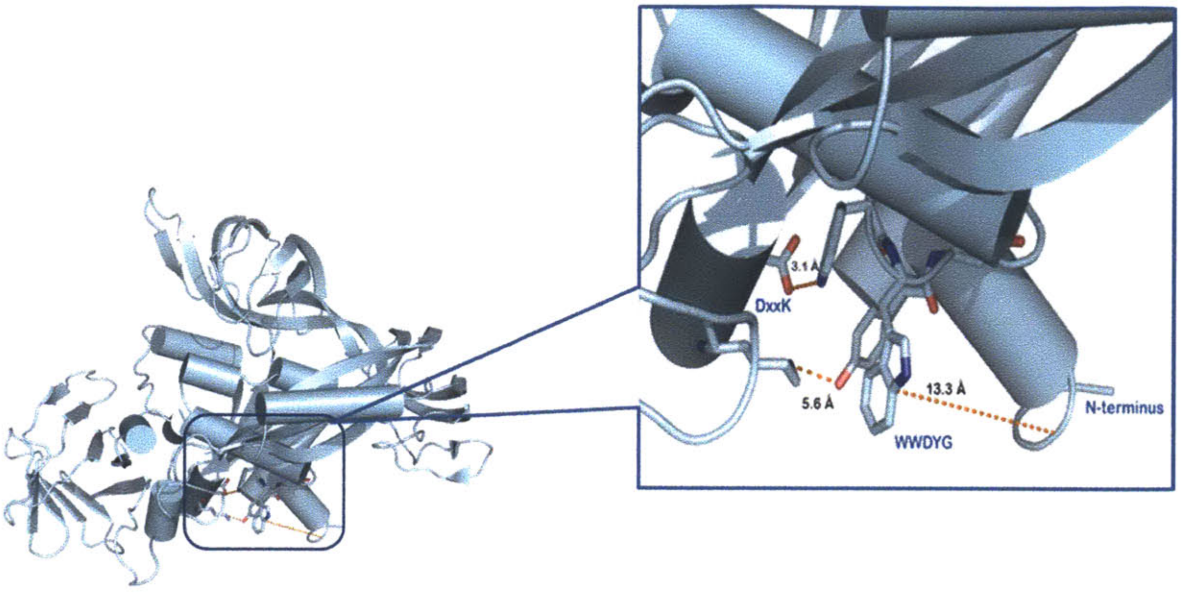


Figure 2: Pymol representation of crystal structure of *P. furiosus* Stt3 soluble domain. (PDB: 2zai, 2.7 Å resolution, [5]). The left image shows the full soluble domain and the right image shows a region of interest enlarged for clarity. Shown as sticks are the residues comprising the WWDxG motif and the aspartate and lysine from the proposed DxxK motif. Orange dashed lines show approximate distances for the following: N-terminus and the second tryptophan in the WWDxG motif (13.3 Å), aspartate from DxxK and the first tryptophan in WWDxG (3.1 Å), and lysine from DxxK and the tyrosine situated in the 'x' position of WWDxG (5.6 Å).

In contrast, there existed no direct investigation of the significance of the transmembrane region of PglB and other Stt3 proteins. This region is predicted to be significant in catalysis due to the highly hydrophobic, and membrane-associated, undecaprenyldiphosphate-linked substrates that the OTases act upon. In addition, the crystal structure of the *P. furiosus* Stt3 soluble domain shows the WWDxG residues in close proximity to the N-terminus (the N-terminus being the native site connecting the soluble domain to the transmembrane region) (**Figure 2**). The location of the WWDxG in the structure suggests these residues are interacting with atoms located on or near the surface of the membrane bilayer, implying that residues within the transmembrane domain interact with the expected site of catalysis [5].

In order to investigate the transmembrane domain of the enzyme directly, sequence alignments of diverse Stt3 homologs were created. However, as mentioned above, alignment of full-length Stt3 sequences failed to expose any homology in the transmembrane region. It was hypothesized that some regions of local conservation may exist within the predicted transmembrane domain, but in general, these regions are not well defined due to the large size of the protein and the low overall sequence conservation. In other words, divergence of surrounding transmembrane helices may create a background noise that prevents detection of intervening conserved motifs. To test this hypothesis, computational topology predictions, which predict whether a residue in a protein sequence appears inside or outside of the membrane, were assembled for several Stt3 proteins (**Figure 1**). The Stt3 topology predictions were used to narrow regions of focus, for example, to the amino-acid sequence of a specific loop between two transmembrane domains. Sequence analysis was then performed on these localized segments (**Figure 3**).

In order to expand this analysis to a set of 28 non-redundant Stt3 sequences, which included 13 eukaryotes, 7 archaea, and 8 bacteria, a program was developed which accepts a list of homologous protein sequences and the related topology predictions. The program returns the sequence from each homolog that corresponds to a specific topological feature (e.g. the first TM sequence, or the third loop of each sequence). This method allowed topology-driven sequence analysis of a large number of Stt3 sequences, which would have been prohibitively onerous to evaluate manually. Using this program, substantial homology was identified between PglB and other Stt3s in the N-terminal transmembrane region of the protein, with several motifs that appear between the transmembrane helices in predicted periplasmic soluble loops. Remarkably, these motifs exhibit strict conservation from bacteria to humans. The secondary structure of the sequence regions surrounding the motifs was analyzed as well; predictions showed that local secondary-structure surrounding the motifs are conserved, further supporting the expected importance of these motifs. In addition to identifying conserved motifs in the OTase enzyme prior to the availability of supporting structural data [7], the following analysis clearly demonstrates the utility of combining topology prediction and sequence analysis to identify conserved, functional motifs in large membrane proteins.

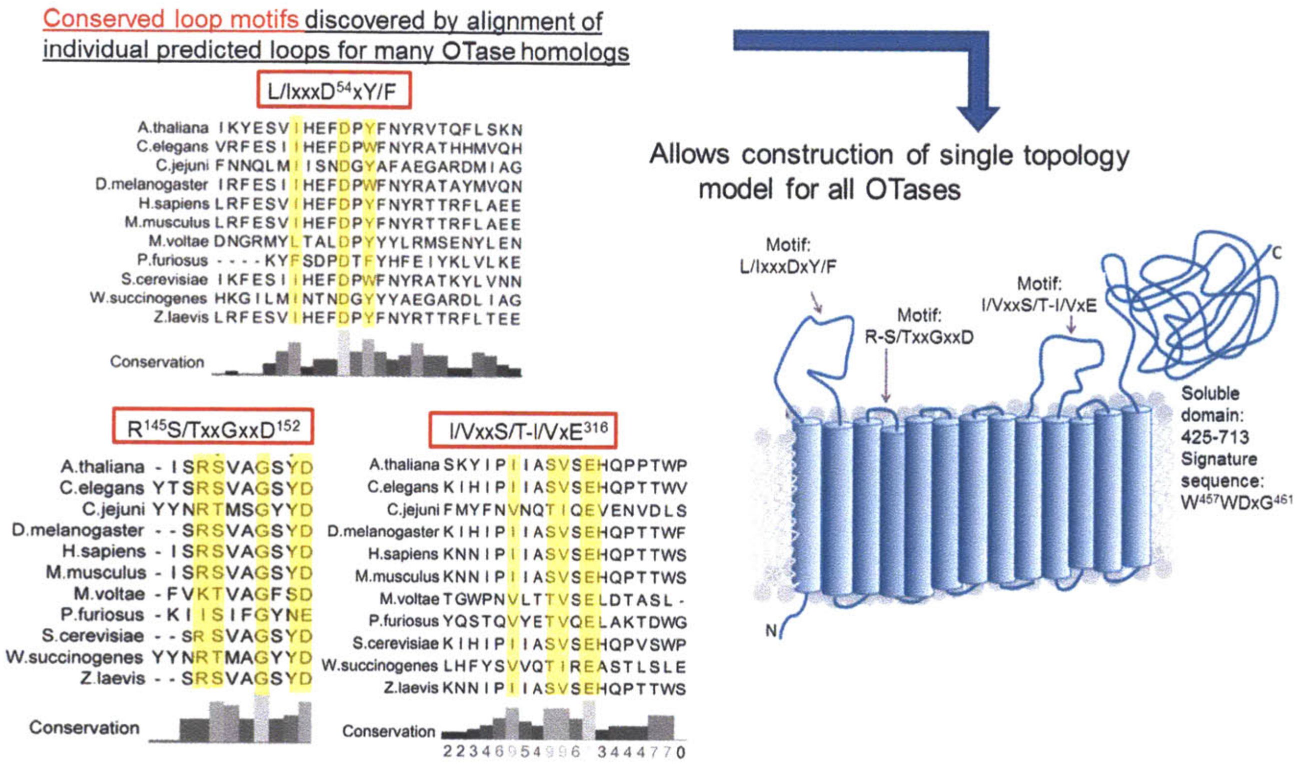

Conserved loop motifs discovered by alignment of individual predicted loops for many OTase homologs
L/IxxxD^{54}xY/F
A.thaliana IKYESVIHEFDPYFNYRVTQFLSKN
C.elegans VRFESIIHEFDPWFNYRATHHMVQH
C.jejuni FNNQLMIISNDGYAFAEGARDMIAG
D.melanogaster IRFESIIHEFDPWFNYRATAYMVQN
H.sapiens LRFESVIHEFDPYFNYRTTRFLAEE
M.musculus LRFESVIHEFDPYFNYRTTRFLAEE
M.voltae DNGRMYLTALDPYYYLRMSENYLEN
P.furiosus ----KYFSDPDTFYHFEIYKLVLKE
S.cerevisiae IKFESIIHEFDPWFNYRATKYLVNN
W.succinogenes HKGILMINTNDGYYYAEGARDLIAG
Z.laevis LRFESVIHEFDPYFNYRTTRFLTEE
Conservation
R^{145}S/TxxGxxD^{152}
A.thaliana -ISRSVAGSYD
C.elegans YTSRSVAGSYD
C.jejuni YYNRTMSGYYD
D.melanogaster --SRSVAGSYD
H.sapiens -ISRSVAGSYD
M.musculus -ISRSVAGSYD
M.voltae -FVKTVAGFSD
P.furiosus -KIISIFGYNE
S.cerevisiae --SRSVAGSYD
W.succinogenes YYNRTMAGYYD
Z.laevis --SRSVAGSYD
Conservation
I/VxxS/T-I/VxE^{316}
A.thaliana SKYIPIIASVSEHQPPTWP
C.elegans KIHIPIIASVSEHQPTTWV
C.jejuni FMYFNVNQTIQEVENVDLS
D.melanogaster KIHIPIIASVSEHQPTTWF
H.sapiens KNNIPIIASVSEHQPTTWS
M.musculus KNNIPIIASVSEHQPTTWS
M.voltae TGWPNVLTTVSELDTASL-
P.furiosus YQSTQVYETVQELAKTDWG
S.cerevisiae KIHIPIIASVSEHQPVSWP
W.succinogenes LHFYSVVQTIREASTLSLE
Z.laevis KNNIPIIASVSEHQPTTWS
Conservation
22346954996 3444770
Allows construction of single topology model for all OTases
Motif: L/IxxxDxY/F
Motif: R-S/TxxGxxD
Motif: I/VxxS/T-I/VxE
C
N
Soluble domain: 425-713
Signature sequence: W^{457}WDxG^{461}

Figure 3: Representation summarizing the topology-based sequence analysis performed on Stt3 homologs.

The sequence segments corresponding to individual predicted loops (1st, 2nd, etc.) were isolated from several Stt3-homologs. These segments were then aligned, at times revealing conserved residues. Pockets of conservation were then used to construct a revised topology map common to all OTases.

Results

Predicted topologies are similar for PglB and Stt3 homologs

Much of the following analysis was motivated by observed similarities between topology predictions for STT3 homologs. Topology predictions for Stt3 homologs, including PglB, indicate a conserved overall structural arrangement. Stt3 homologs share a large N-terminal multi-transmembrane domain, followed by a C-terminal soluble domain (**Figure 1**). Graphical depictions of predicted topologies are displayed in **Figure 4** for select Stt3 homologs, chosen to represent the evolutionary diversity of the protein.

It should be noted that the predicted number of transmembrane helices varies from 11-13 amongst sequences from various organisms. However, topology predictions vary subtly according the specific algorithm used in analysis, which leaves uncertainty about whether the number of transmembrane domains in STT3 homologs varies or whether some predictions are incorrect. Some sequences are also predicted to contain a transmembrane domain near the middle of the C-terminal soluble domain, although the majority of these are predicted to simply contain a hydrophobic region at this position. The latter scenario is almost certainly the correct one, as indicated by many studies of the soluble domain [5,8-10]. However, this discrepancy demonstrates the tentative nature of topology predictions, which often differ depending on the specific algorithm used, and thus are used only as a guide for investigating protein topology. The topology prediction server TMHMM was used throughout the present analysis [11].

Closer inspection of the topology diagrams reveals the consistent presence of two sizeable loops (> 40 residues) located in similar regions of each of the Stt3 topology predictions. One loop of roughly 50 residues appears consistently after the first transmembrane domain; a second loop of roughly 40 residues appears 2-3 transmembrane domains before the start of the soluble domain (**Figure 4**). Investigation of the sequences corresponding to the first loop reveals a prominently conserved aspartic acid (residue D54 in PglB). While other groups have suggested the existence of conserved residues, a systematic investigation of potential regions of conservation or their functional significance has not been previously carried out [4,9]. Indeed, such investigations are hindered by the difficulty of detecting small motifs within the context of the large and variable full-length sequence. Utilizing the conserved topology of Stt3 homologs, sequences corresponding to a specific topological feature of each Stt3 protein were extracted, and then sequence analysis was confined to these sections. To systematize the approach, a computer program was developed to automatically extract the sequence corresponding to a given topological feature for a series of 28 divergent Stt3 homologs (**Table 1**).

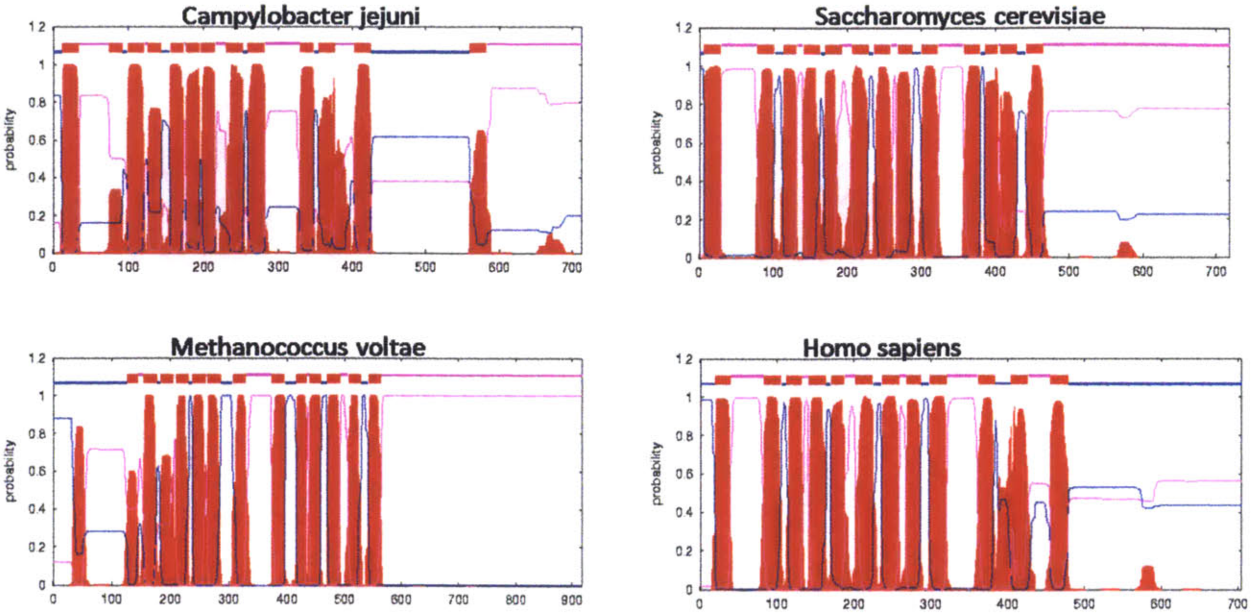


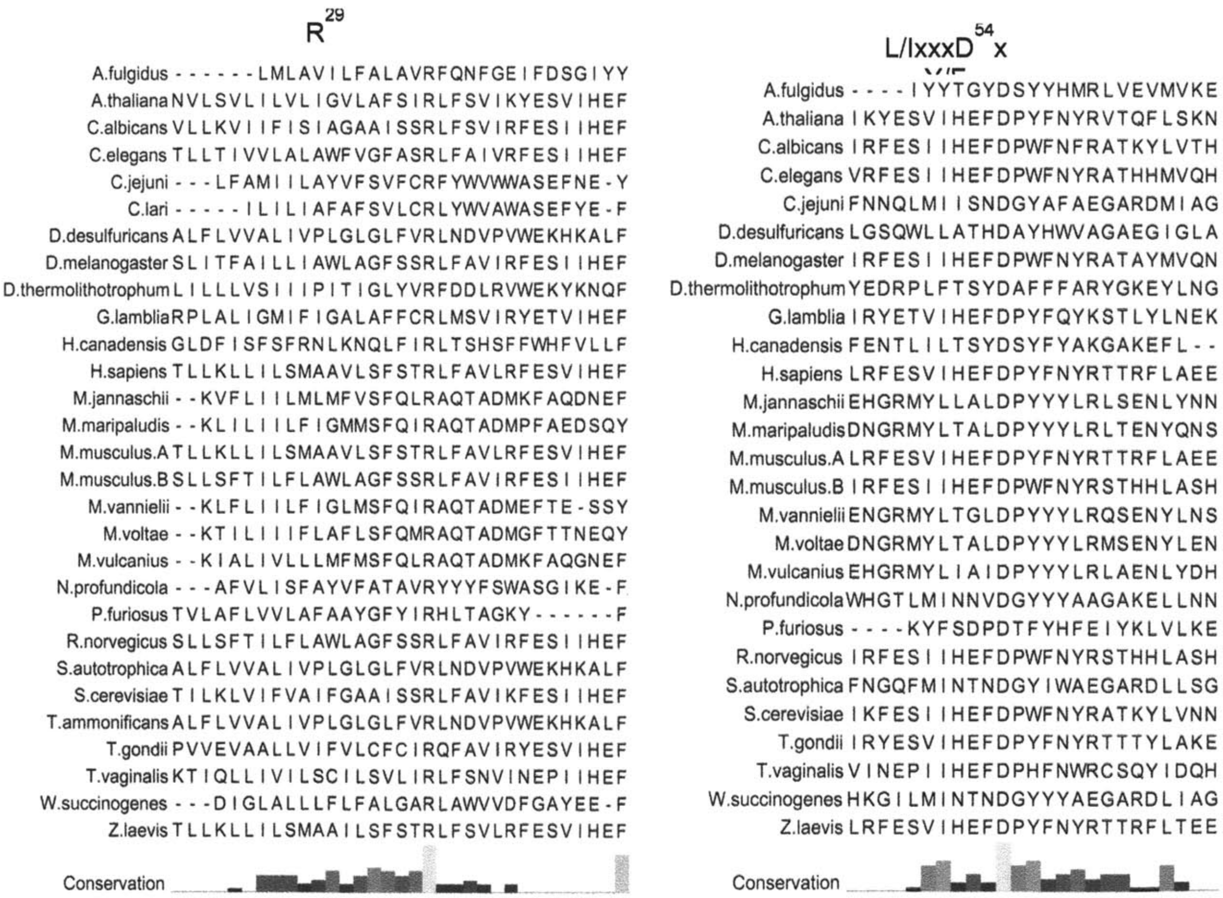
Figure 4. Sequence-based topology predictions for several Stt3 homologs.
The predictions shown are for archaea (*M. voltae*), bacteria (*C. jejuni,* PglB), fungi (*S. cerevisiae*) and a mammal (*H. sapiens*). The protein sequence from N- to C-terminus is represented along the x-axis, and the y-axis shows the probability that a given residue appears within a transmembrane domain or outside the membrane. The tall blocks represent predicted transmembrane domains. Predictions and illustrations were generated by the Transmembrane Hidden Markov Model (TMHMM) topology prediction program [11]. *Note:* The trace running horizontally along the top of the graph is included in the program output to designate upon which side of the membrane a soluble section is predicted to appear (i.e. cytoplasmic is blue or luminal, pink). For the figure shown, this trace can be disregarded, as these predictions tend to be highly inaccurate when processing sequences with many consecutive predicted transmembrane helices.

Table 1: Identities of sequences used in all bioinformatic analysis.

Domain	Gene Identifier	Organism
eukaryota	gi|1322489|emb|CAA96722.1|	Saccharomyces cerevisiae
archaea	gi|11497941|ref|NP_069165.1|	Archaeoglobus fulgidus DSM 4304
archaea	gi|15669720|ref|NP_248533.1|	Methanocaldococcus jannaschii DSM 2661
archaea	gi|87045854|gb|ABD17750.1|	Methanococcus voltae
archaea	gi|134045200|ref|YP_001096686.1|	Methanococcus maripaludis C5
archaea	gi|150012189|gb|ABR54641.1|	Methanococcus vannielii SB
archaea	gi|261402368|ref|YP_003246592.1|	Methanocaldococcus vulcanius M7
eukaryota	gi|281485606|ref|NP_001164010.1|	Rattus norvegicus homolog B
eukaryota	gi|148226196|ref|NP_001083986.1|	Xenopus laevis homolog A
eukaryota	gi|211965698|gb|EEB00894.1|	Toxoplasma gondii ME49
eukaryota	gi|121904152|gb|EAY09105.1|	Trichomonas vaginalis G3
bacteria	gi|86153042|ref|ZP_01071247.1|	Campylobacter jejuni HB93-13
eukaryota	gi|308159114|gb|EFO61662.1|	Giardia lamblia P15
eukaryota	gi|127799903|gb|AAH48348.2|	Homo sapiens homolog A
eukaryota	gi|55154464|gb|AAH85313.1|	Mus musculus homolog A
eukaryota	gi|30851502|gb|AAH52433.1|	Mus musculus homolog B
eukaryota	gi|238881972|gb|EEQ45610.1|	Candida albicans WO-1
archaea	gi|18976528|ref|NP_577885.1|	Pyrococcus furiosus DSM 3638
eukaryota	gi|18419993|ref|NP_568380.1|	Arabidopsis thaliana
eukaryota	gi|529357|gb|AAC24442.1|	Caenorhabditis elegans
bacteria	gi|253510784|gb|EES89443.1|	Helicobacter canadensis MIT 98-5491
eukaryota	gi|17738187|ref|NP_524494.1|	Drosophila melanogaster
bacteria	gi|34556499|ref|NP_906314.1|	Wolinella succinogenes DSM 1740
bacteria	gi|307721432|ref|YP_003892572.1|	Sulfurimonas autotrophica DSM 16294
bacteria	gi|224373660|ref|YP_002608032.1|	Nautilia profundicola AmH
bacteria	gi|219868319|gb|ACL48654.1|	Desulfovibrio desulfuricans ATCC 27774
bacteria	gi|325066002|gb|ADY74009.1|	Desulfurobacterium thermolithotrophum DSM 11699
bacteria	gi|317114826|gb|ADU97316.1|	Thermovibrio ammonificans HB-1

Identification of conserved motifs

The sequence analysis revealed strong conservation in several Stt3 inter-TM loops, which is not easily detected when the full-length Stt3 sequences are aligned. Alignments of these motifs are shown for PglB (*C. jejuni*) and other key Stt3 homologs in **Figure 5**. The first transmembrane helix contains a conserved arginine residue (residue R29 in PglB). The first loop (which appears in the lumen in eukaryotic Stt3s and the periplasm of the *C. jejuni* PglB) contains the [L/I]xx[D^{54}]x[Y/F] motif. The third loop in the ER lumen or the periplasm contains the [R/K][S/T]xx[G^{149}]xx[D^{152}] motif, and the seventh loop contains the [I/V]xxx[S/T][I/V]x[E^{316}] motif. A topology model, shown in **Figure 6**, was constructed based on a combination of topology predictions for Stt3 homologs and the conserved motifs, and is used to indicate the locations of the identified motifs.


R29
A.fulgidus ------LMLAVILFALAVRFQNFGEIFDSGIYY
A.thaliana NVLSVLILVLIGVLAFSIRLFSVIKYESVIHEF
C.albicans VLLKVIIFISIAGAAISSRLFSVIRFESIIHEF
C.elegans TLLTIVVLALAWFVGFASRLFAIVRFESIIHEF
C.jejuni ---LFAMIILAYVFSVFCRFYWVWWASEFNE-Y
C.lari -----ILILIAFAFSVLCRLYWVAWASEFYE-F
D.desulfuricans ALFLVVALIVPLGLGLFVRLNDVPVWEKHKALF
D.melanogaster SLITFAILLIAWLAGFSSRLFAVIRFESIIHEF
D.thermolithotrophum LILLLVSIIIPITIGLYVRFDDLRVWEKYKNQF
G.lamblia RPLALIGMIFIGALAFFCRLMSVIRYETVIHEF
H.canadensis GLDFISFSFRNLKNQLFIRLTSHSFFWHFVLLF
H.sapiens TLLKLLILSMAAVLSFSTRLFAVLRFESVIHEF
M.jannaschii --KVFLIILMLMFVSFQLRAQTADMKFAQDNEF
M.maripaludis --KLILIILFIGMMSFQIRAQTADMPFAEDSQY
M.musculus.A TLLKLLILSMAAVLSFSTRLFAVLRFESVIHEF
M.musculus.B SLLSFTILFLAWLAGFSSRLFAVIRFESIIHEF
M.vannielii --KLFLIILFIGLMSFQIRAQTADMEFTE-SSY
M.voltae --KTILIIIFLAFLSFQMRAQTADMGFTTNEQY
M.vulcanius --KIALIVLLLMFMSFQLRAQTADMKFAQGNEF
N.profundicola ---AFVLISFAYVFATAVRYYYFSWASGIKE-F
P.furiosus TVLAFLVVLAFAAYGFYIRHLTAGKY------F
R.norvegicus SLLSFTILFLAWLAGFSSRLFAVIRFESIIHEF
S.autotrophica ALFLVVALIVPLGLGLFVRLNDVPVWEKHKALF
S.cerevisiae TILKLVIFVAIFGAAISSRLFAVIKFESIIHEF
T.ammonificans ALFLVVALIVPLGLGLFVRLNDVPVWEKHKALF
T.gondii PVVEVAALLVIFVLCFCIRQFAVIRYESVIHEF
T.vaginalis KTIQLLIVILSCILSVLIRLFSNVINEPIIHEF
W.succinogenes ---DIGLALLLFLFALGARLAWVVDFGAYEE-F
Z.laevis TLLKLLILSMAAILSFSTRLFSVLRFESVIHEF
Conservation
L/IxxxD54x
A.fulgidus ----IYYTGYDSYYHMRLVEVMVKE
A.thaliana IKYESVIHEFDPYFNYRVTQFLSKN
C.albicans IRFESIIHEFDPWFNFRATKYLVTH
C.elegans VRFESIIHEFDPWFNYRATHHMVQH
C.jejuni FNNQLMIISNDGYAFAEGARDMIAG
D.desulfuricans LGSQWLLATHDAYHWVAGAEGIGLA
D.melanogaster IRFESIIHEFDPWFNYRATAYMVQN
D.thermolithotrophum YEDRPLFTSYDAFFFARYGKEYLNG
G.lamblia IRYETVIHEFDPYFQYKSTLYLNEK
H.canadensis FENTLILTSYDSYFYAKGAKEFL--
H.sapiens LRFESVIHEFDPYFNYRTTRFLAEE
M.jannaschii EHGRMYLLALDPYYYLRLSENLYNN
M.maripaludis DNGRMYLTALDPYYYLRLTENYQNS
M.musculus.A LRFESVIHEFDPYFNYRTTRFLAEE
M.musculus.B IRFESIIHEFDPWFNYRSTHHLASH
M.vannielii ENGRMYLTGLDPYYYLRQSENYLNS
M.voltae DNGRMYLTALDPYYYLRMSENYLEN
M.vulcanius EHGRMYLIAIDPYYYLRLAENLYDH
N.profundicola WHGTLMINNVDGYYYAAGAKELLNN
P.furiosus ----KYFSDPDTFYHFEIYKLVLKE
R.norvegicus IRFESIIHEFDPWFNYRSTHHLASH
S.autotrophica FNGQFMINTNDGYIWAEGARDLLSG
S.cerevisiae IKFESIIHEFDPWFNYRATKYLVNN
T.gondii IRYESVIHEFDPYFNYRTTTYLAKE
T.vaginalis VINEPIIHEFDPHFNWRCSQYIDQH
W.succinogenes HKGILMINTNDGYYYAEGARDLIAG
Z.laevis LRFESVIHEFDPYFNYRTTRFLTEE
Conservation

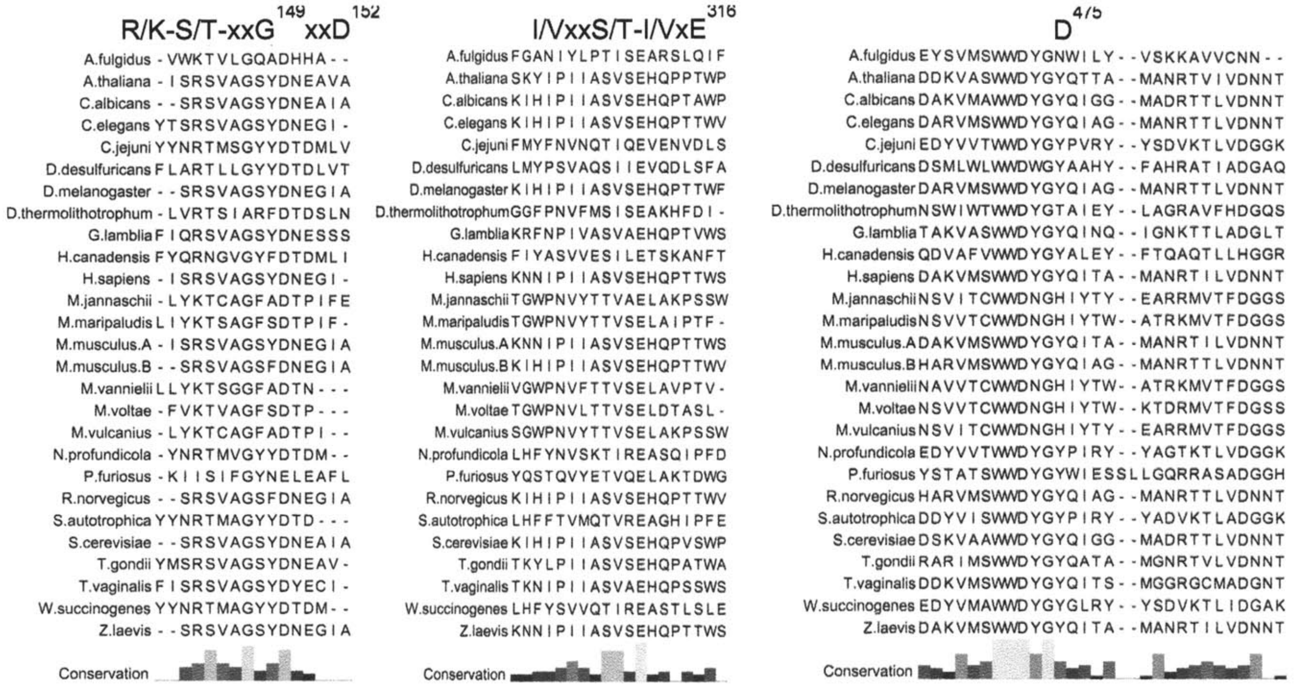

R/K-S/T-xxG 149 xxD 152
A.fulgidus -VWKTVLGQADHHA--
A.thaliana -ISRSVAGSYDNEAVA
C.albicans --SRSVAGSYDNEAIA
C.elegans YTSRSVAGSYDNEGI-
C.jejuni YYNRTMSGYYDTDMLV
D.desulfuricans FLARTLLGYYDTDLVT
D.melanogaster --SRSVAGSYDNEGIA
D.thermolithotrophum -LVRTSIARFDTDSLN
G.lamblia FIQRSVAGSYDNESSS
H.canadensis FYQRNGVGYFDTDMLI
H.sapiens -ISRSVAGSYDNEGI-
M.jannaschii -LYKTCAGFADTPIFE
M.maripaludis LIYKTSAGFSDTPIF-
M.musculus.A -ISRSVAGSYDNEGIA
M.musculus.B --SRSVAGSFDNEGIA
M.vannielii LLYKTSGGFADTN---
M.voltae -FVKTVAGFSDTP---
M.vulcanius -LYKTCAGFADTPI--
N.profundicola -YNRTMVGYYDTDM--
P.furiosus -KIISIFGYNELEAFL
R.norvegicus --SRSVAGSFDNEGIA
S.autotrophica YYNRTMAGYYDTD---
S.cerevisiae --SRSVAGSYDNEAIA
T.gondii YMSRSVAGSYDNEAV-
T.vaginalis FISRSVAGSYDYECI-
W.succinogenes YYNRTMAGYYDTDM--
Z.laevis --SRSVAGSYDNEGIA
Conservation
I/VxxS/T-I/VxE 316
A.fulgidus FGANIYLPTISEARSLQIF
A.thaliana SKYIPIIASVSEHQPPTWP
C.albicans KIHIPIIASVSEHQPTAWP
C.elegans KIHIPIIASVSEHQPTTWV
C.jejuni FMYFNVNQTIQEVENVDLS
D.desulfuricans LMYPSVAQSIIEVQDLSFA
D.melanogaster KIHIPIIASVSEHQPTTWF
D.thermolithotrophum GGFPNVFMSISEAKHFDI-
G.lamblia KRFNPIVASVAEHQPTVWS
H.canadensis FIYASVVESILETSKANFT
H.sapiens KNNIPIIASVSEHQPTTWS
M.jannaschii TGWPNVYTTVAELAKPSSW
M.maripaludis TGWPNVYTTVSELAIPTF-
M.musculus.A KNNIPIIASVSEHQPTTWS
M.musculus.B KIHIPIIASVSEHQPTTWV
M.vannielii VGWPNVFTTVSELAVPTV-
M.voltae TGWPNVLTTVSELDTASL-
M.vulcanius SGWPNVYTTVSELAKPSSW
N.profundicola LHFYNVSKTIREASQIPFD
P.furiosus YQSTQVYETVQELAKTDWG
R.norvegicus KIHIPIIASVSEHQPTTWV
S.autotrophica LHFFTVMQTVREAGHIPFE
S.cerevisiae KIHIPIIASVSEHQPVSWP
T.gondii TKYLPIIASVSEHQPATWA
T.vaginalis TKNIPIIASVAEHQPSSWS
W.succinogenes LHFYSVVQTIREASTLSLE
Z.laevis KNNIPIIASVSEHQPTTWS
Conservation
D 475
A.fulgidus EYSVMSWWDYGNWILY--VSKKAVVCNN--
A.thaliana DDKVASWWDYGYQTTA--MANRTVIVDNNT
C.albicans DAKVMAWWDYGYQIGG--MADRTTLVDNNT
C.elegans DARVMSWWDYGYQIAG--MANRTTLVDNNT
C.jejuni EDYVVTWWDYGYPVRY--YSDVKTLVDGGK
D.desulfuricans DSMLWLWWDWGYAAHY--FAHRATIADGAQ
D.melanogaster DARVMSWWDYGYQIAG--MANRTTLVDNNT
D.thermolithotrophum NSWIWTWWDYGTAIEY--LAGRAVFHDGQS
G.lamblia TAKVASWWDYGYQINQ--IGNKTTLADGLT
H.canadensis QDVAFVWWDYGYALEY--FTQAQTLLHGGR
H.sapiens DAKVMSWWDYGYQITA--MANRTILVDNNT
M.jannaschii NSVITCWWDNGHIYTY--EARRMVTFDGGS
M.maripaludis NSVVTCWWDNGHIYTW--ATRKMVTFDGGS
M.musculus.A DAKVMSWWDYGYQITA--MANRTILVDNNT
M.musculus.B HARVMSWWDYGYQIAG--MANRTTLVDNNT
M.vannielii NAVVTCWWDNGHIYTW--ATRKMVTFDGGS
M.voltae NSVVTCWWDNGHIYTW--KTDRMVTFDGSS
M.vulcanius NSVITCWWDNGHIYTY--EARRMVTFDGGS
N.profundicola EDYVVTWWDYGYPIRY--YAGTKTLVDGGK
P.furiosus YSTATSWWDYGYWIESSLLGQRRASADGGH
R.norvegicus HARVMSWWDYGYQIAG--MANRTTLVDNNT
S.autotrophica DDYVISWWDYGYPIRY--YADVKTLADGGK
S.cerevisiae DSKVAAWWDYGYQIGG--MADRTTLVDNNT
T.gondii RARIMSWWDYGYQATA--MGNRTVLVDNNT
T.vaginalis DDKVMSWWDYGYQITS--MGGRGCMADGNT
W.succinogenes EDYVMAWWDYGYGLRY--YSDVKTLIDGAK
Z.laevis DAKVMSWWDYGYQITA--MANRTILVDNNT
Conservation

Figure 5: Alignments for highly conserved regions of OTase sequences.

All 28 homologs listed are listed in **Table 1**. *Top left* alignment shows a conserved arginine within the first transmembrane domain. The corresponding residue number in PglB (*C. jejuni*) is designated in superscript. Alignments and conservation histograms shown were made using Jalview multiple alignment editor. (doi: 10.1093/bioinformatics/btp033). Histogram shading: light to dark designates most to least conserved. Alignment conservation calculation uses the MAAFT method of multiple sequence alignment (Livingstone, C. D., Barton, G. J. (1993) *CAMBIOS* 9: 745-756).

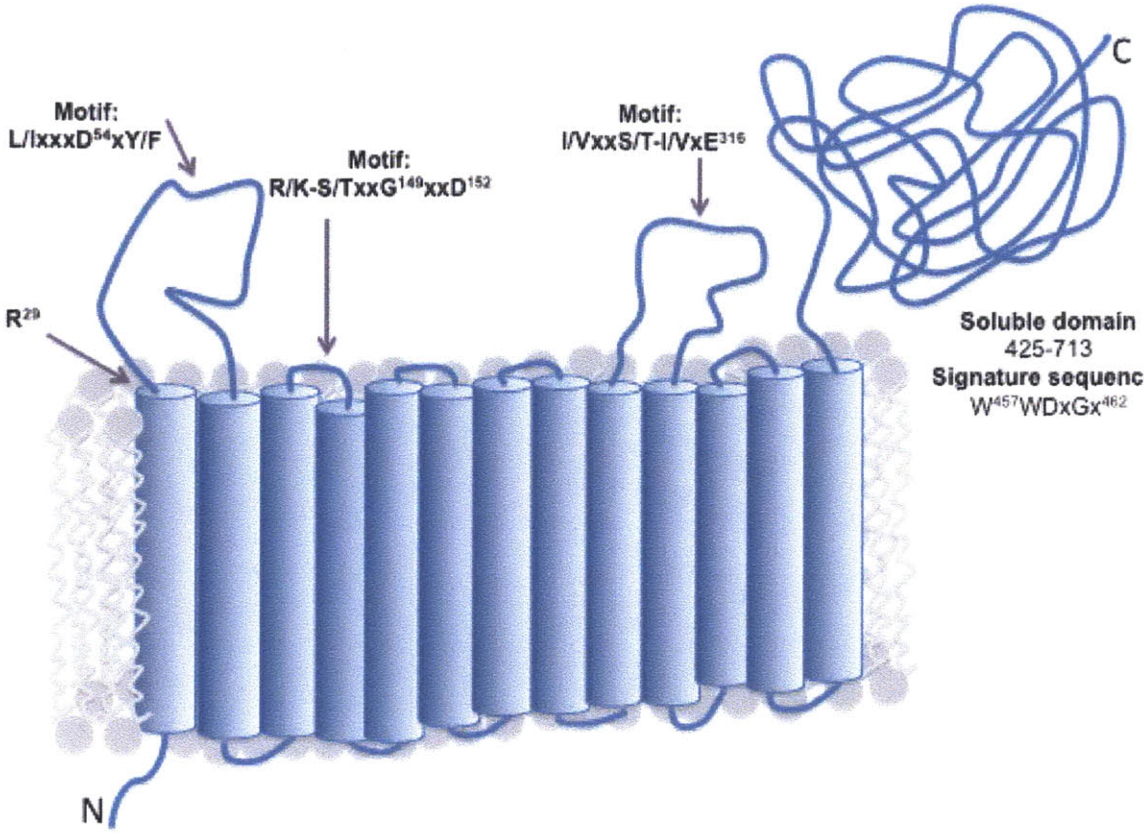

Motif:
L/IxxxD^{54}xY/F
Motif:
R/K-S/TxxG^{149}xxD^{152}
Motif:
I/VxxS/T-I/VxE^{316}
C
R^{29}
Soluble domain
425-713
Signature sequence:
W^{457}WDxGx^{462}
N

Figure 6: Topology model of PglB showing locations of conserved motifs.
Model is based on a combination of topology prediction programs (see Methods) and conservation of loop regions, and agrees with topology of the *C. lari* crystal structure. Thirteen transmembrane helices are followed by a globular domain located in the periplasm. Predicted locations of identified motifs are indicated by arrows. The N-terminus is located in the cytoplasm.

Assessing structural conservation in loops

Sequence analysis of the various Stt3 proteins provided strong predictive data that these regions are fundamentally involved in PglB function. However, the motifs consisted of one to three highly conserved residues interspersed with stretches of residues with little or no conservation. Additionally, these motifs are short and potentially difficult to identify within a loop. For example, the first major loop contains a D-x-Y/F/W motif, in which the aspartate is invariably conserved. But it is not improbable that several Stt3 homologs will show several appearances of this motif pattern appearing in the first luminal loop.

Therefore, to supplement the sequence alignments, the series of Stt3 homolog sequences for the two largest loops were submitted to the free online software program NetSurfP, which predicts a protein's secondary structure and level of accessibility from the sequence [12]. While NetSurfP is less well-known than several alternative programs available for this purpose, the program both allows input of a list of sequences at a single time and the output data can easily be parsed to provide visually informative graphs, which makes for simpler comparison and detection of patterns.

The conserved Stt3 motifs invariably appear in the same loops relative to the remainder of the protein. Therefore, it is expected that the motif-containing loops exhibit secondary-structure conservation in this region in order to structurally situate the motif into a functional position. In other words, is likely that the location of the motif within the loop will fall in the same location with respect to the secondary structure of the loop (e.g. the motif may appear continually in a coil between two alpha helices, etc.). As a result, the secondary structure pattern presumably provides an alternate way to quickly identify the motif of interest with high confidence. Additionally, the conservation of secondary structure can also be informative of the flexibility, and thus potential mobility of the loop, in the OTase reaction. Finally, a conserved secondary structural pattern can function as an additional way to refine the predicted border residues of the loops and help identify additional regions in the loops that are critical for function.

The sequences for two of the predicted loop regions were analyzed for all of the Stt3 homologs used to make the alignments. The two loops analyzed included the first soluble loop containing the highly conserved aspartate (in the motif D-x-F/Y/W, D54 in *C. jejuni*) and the large soluble loop toward the C-terminus of the transmembrane region (containing the motif S/T-I/V-x-E, E316 in *C. jejuni*). These two loops, both ranging from 30 to 100 residues, are both long enough to get a meaningful prediction of structure (in contrast to the roughly10-residue R-S/T-xx-G-xx-D –containing loop) and yet these two soluble loops are short enough to identify subtle patterns and variations. Graphs of the secondary structure predictions for the first loop are shown in **Figure 7** for several representative organisms across the evolutionary spectrum. The graphs for all Stt3 homologs involved in the sequence analysis are shown at the end of this section **(Figures 9 and 10)**.

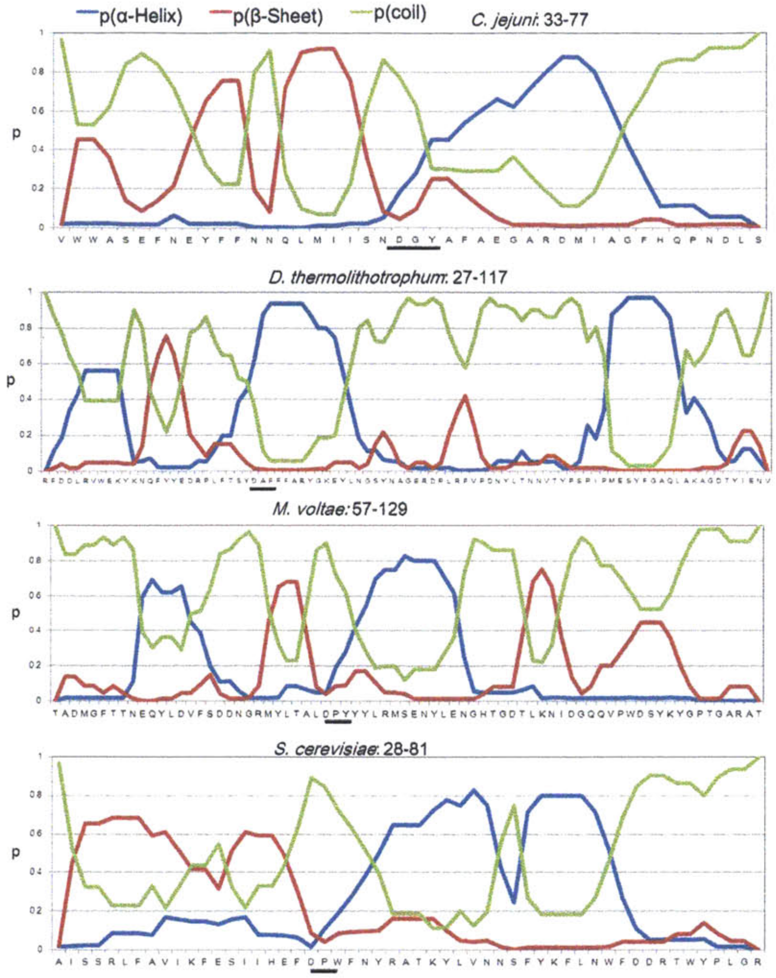


Figure 7: Secondary structure predictions for first major conserved loop in several Stt3 homologs.
The x-axis shows the protein sequence for the first conserved loop of the homologs lsited. The y-axis gives the calculated probability (p) for each residue that it is located within an alpha helix (blue line), beta strand (red line), or coil (green line). The conserved aspartate in the D-x-W/F/Y motif is underlined for reference.

The known motifs are useful in comparing the secondary structure predictions, because they create anchor points around which to align the graphs. The graphs are based both on the topology prediction of the length of the loop as well as the prediction of secondary structure. The combination of predictions from several programs increases the associated error. This ambiguity makes it difficult to determine whether a given graph may be representing, for example, the first or latter half of the loop sequence. It is important to recognize that the presence of a given section of secondary structure in some graphs and not others does not necessarily imply a lack of structural conservation. As noted above, the accuracy of these graphs is not only dependent on the accuracy of the secondary-structure predictions but also on the topology predictions of the length of the loop region. Since nature rarely adheres to convenient categories such as membrane, inside, or outside in location, the situation is often complicated when the borders of the protein loop are not well defined. A comparison of two highly similar bacterial OTase sequences from *C. jejuni* and *W. succinogenes* highlights how similar sequences can produce highly variable results. **Figure 8** demonstrates this principle making use of the *C. lari* structural model (PDB 3RCE, [7]).

The structure in **Figure 8** shows the region encompassed by the first conserved loop for *C. jejuni* and for *W. succinogenes*. The *W. succinogenes* predicted 'loop region' **(Figure 8D)** includes an additional section of an alpha helix that appears to lie atop the membrane and then merge with the following TM helix. Thus, although these sequences are virtually identical, it is clear that the topology prediction may differ for each in defining the precise borders of 'loop' versus 'TM helix', as the structural border between these two defined categories is inherently unclear.

The conserved residues, however, can be used to clarify this potential complication. For example, in the graphs shown in **Figure 7**, there is considerable variation at first glance. However, the appearance of the D-x-W/Y/F motif is underlined in each sequence. Using this anchor point, a clear pattern becomes visible. The motif consistently appears at the start of a clearly defined alpha helical segment, which is invariably bordered on both sides by a short region of coil structure, which is then interspersed with short occurrences of backbone structure in the form of beta strand or alpha helix.

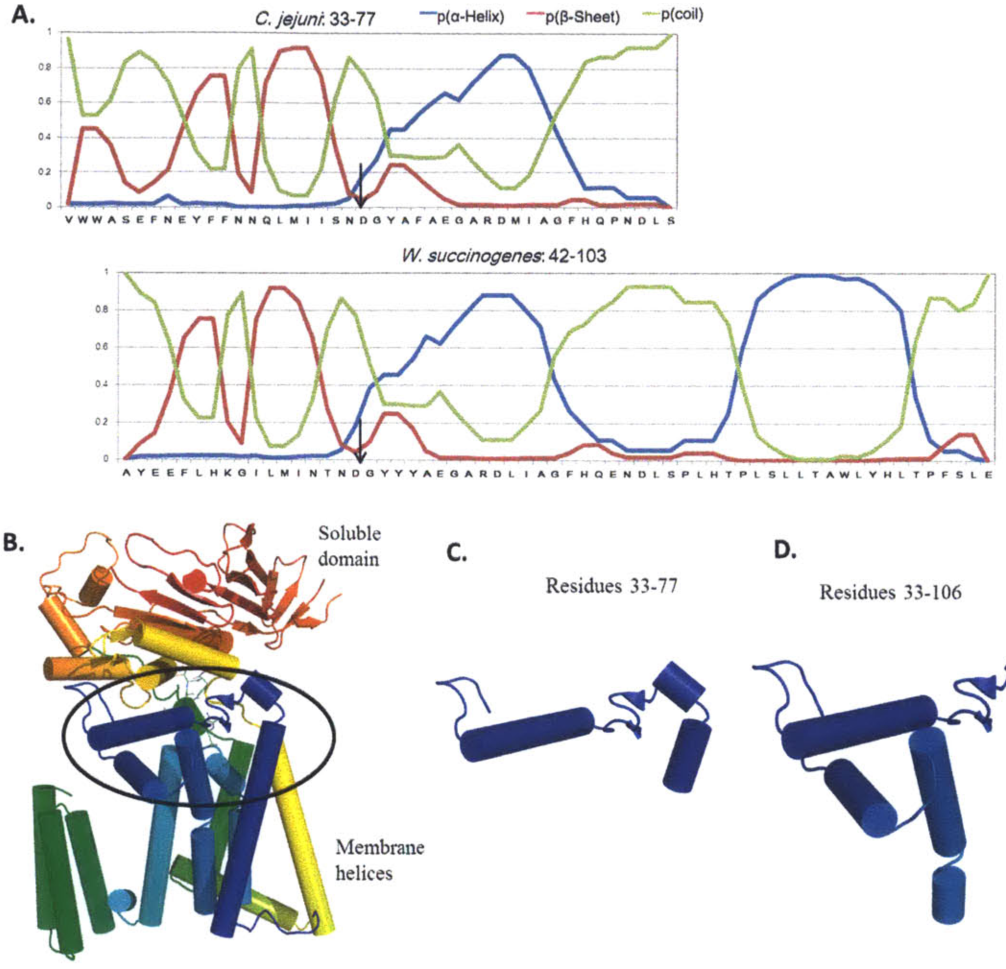


Figure 8: Example of one source of ambiguity in topology predictions: when a helix is both 'soluble' and integral-membrane'.

(*A*) Secondary-structure predictions for first transmembrane loop in *C. jejuni* and *W. succinogenes*, are highly similar, showing the formation of 1-2 strong helical regions. Though the sequences are highly similar, the prediction for where the soluble loop transitions into a transmembrane helix differs for each protein (ranges of graphs, part *A*). The large, well-defined helices in this loop account for the inconsistency. Both predicted ranges are mapped onto the *C. lari* X-ray crystal structure shown in (*B*). The range predicted to be 'soluble' for *C. jejuni* is shown in (*C*); the region predicted to be 'soluble' for *W. succinogenes* is shown in (*D*). The fusing of the second helix in the loop with the following TM helix is shown and accounts for variability in predictions.

The graphs can also be interpreted generally as a prediction of structural rigidity versus flexibility (alpha helix or beta strand represent formation of rigidity through peptide backbone bonds, while coils are generally more disordered, or flexible, and less confined by energetic nadirs formed by secondary-structure bonding). Consequently, the invariable presence of some secondary-structure order indicates that the region of the loop containing the motif is constrained, and any movement involving this unit is expected to be tightly controlled.

The first loop, as noted above, is predicted to contain a high content of alpha-helix and beta-sheet structural order, which implies a general ordered nature. In contrast, the second large loop in the transmembrane region, which contains the motif S/T-I/V-x-E, consistently is predicted to contain a high degree of disorder. The predictions for the sequences comprising this loop consistently have high coil content interrupted by 1-2 precipitous drops in coil toward the center of the loop. The conserved motif unfailingly appears in one of these sudden, brief appearances of secondary-structural order, and at times is predicted to be an alpha-helix and other times predicted to be a beta-sheet. This prediction is in agreement with the available structural data, in which a large portion of this loop is unresolved in the crystal structure [7]. The conservation of the general size and secondary structure composition of this loop, in addition to the presence of the conserved motif, implicates this region in a critical motion for catalysis, perhaps in substrate binding or release. The secondary-structure predictions for these two large loops are shown for all 28 Stt3 homologs in the following pages. **Figure 9** shows predictions for Loop 1 containing the conserved motif L/I-xxx-D-x-Y/F/W and **Figure 10** shows predictions luminal/periplasmic loop 5 containing the conserved motif S/T-V/I-x-E.

For **Figure 9**, sections **9A**, **9B**, and **9C** respectively appear on the three pages immediately following the current one.

Figure 9: Secondary-structure predictions for regions of Stt3 sequences predicted to correspond to the first soluble loop containing the conserved L/I-xxx-D-x-Y/F/W motif. Along the x-axis runs the protein sequence predicted for this region (residue numbers indicated next to species name) and the y-axis gives the probability that each residue is part of a coil (green), helix (blue), or β-sheet (red). Loop 1 contains the highly conserved motif L/I-xxx-D-x-Y/F/W in Stt3 homologs (see **Figure 6**). The location of the aspartate in the motif is highlighted with a purple star. The graphs are group according to kingdom: (*A*) eukaryote, (*B*) bacteria, (*C*) archaeal.

Figure 9A: Eukaryotic Stt3 homologs

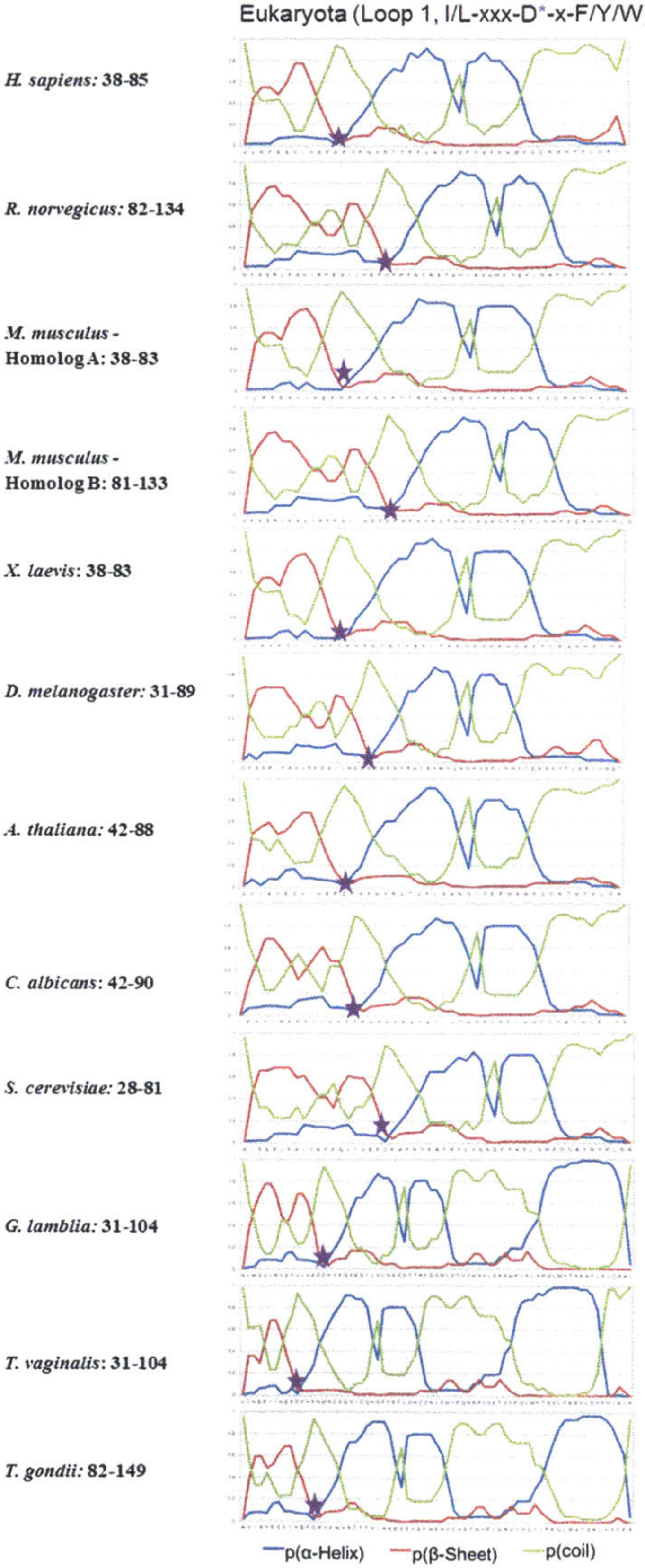


Figure 9B: Bacterial Stt3 homologs

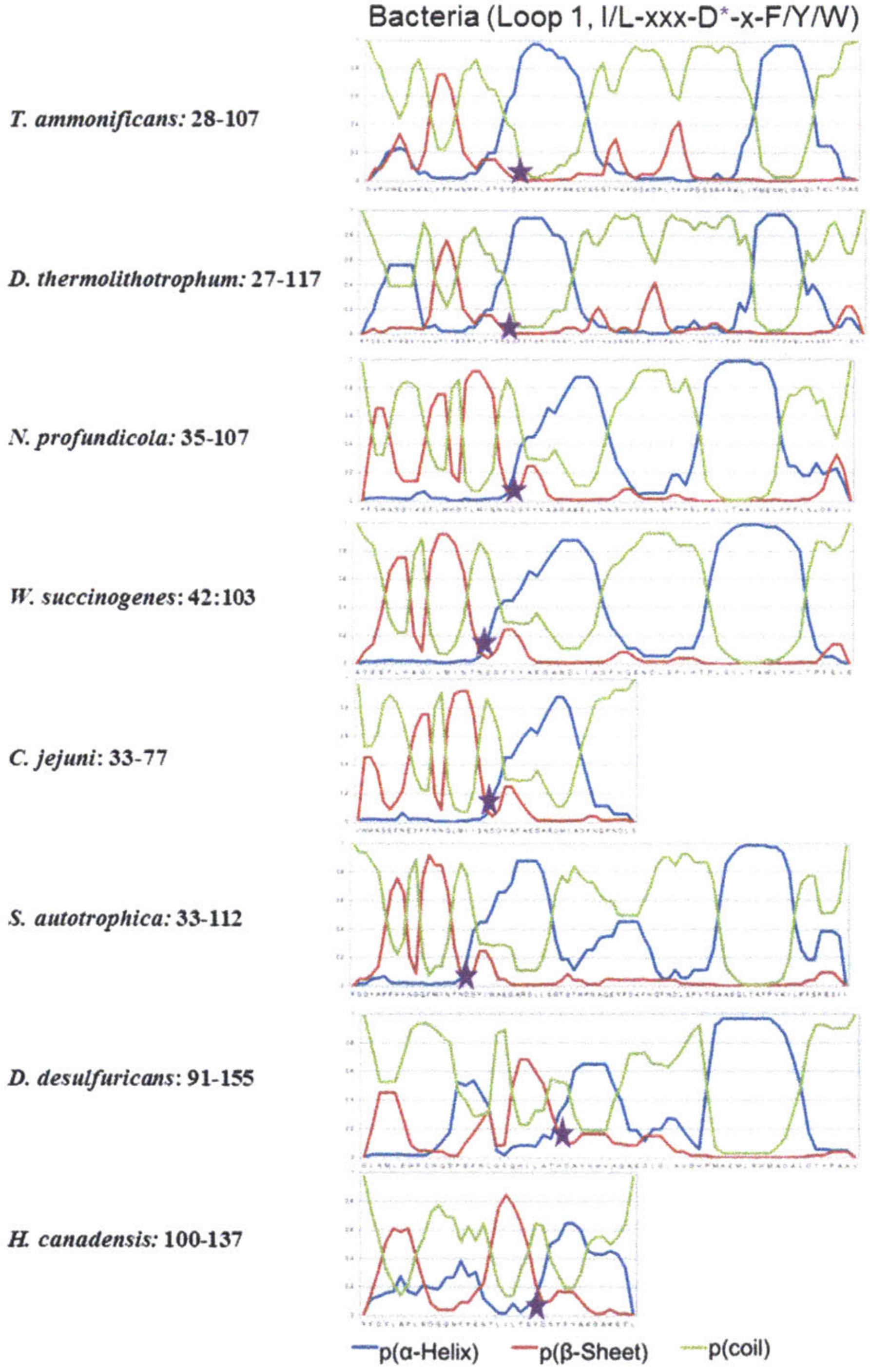
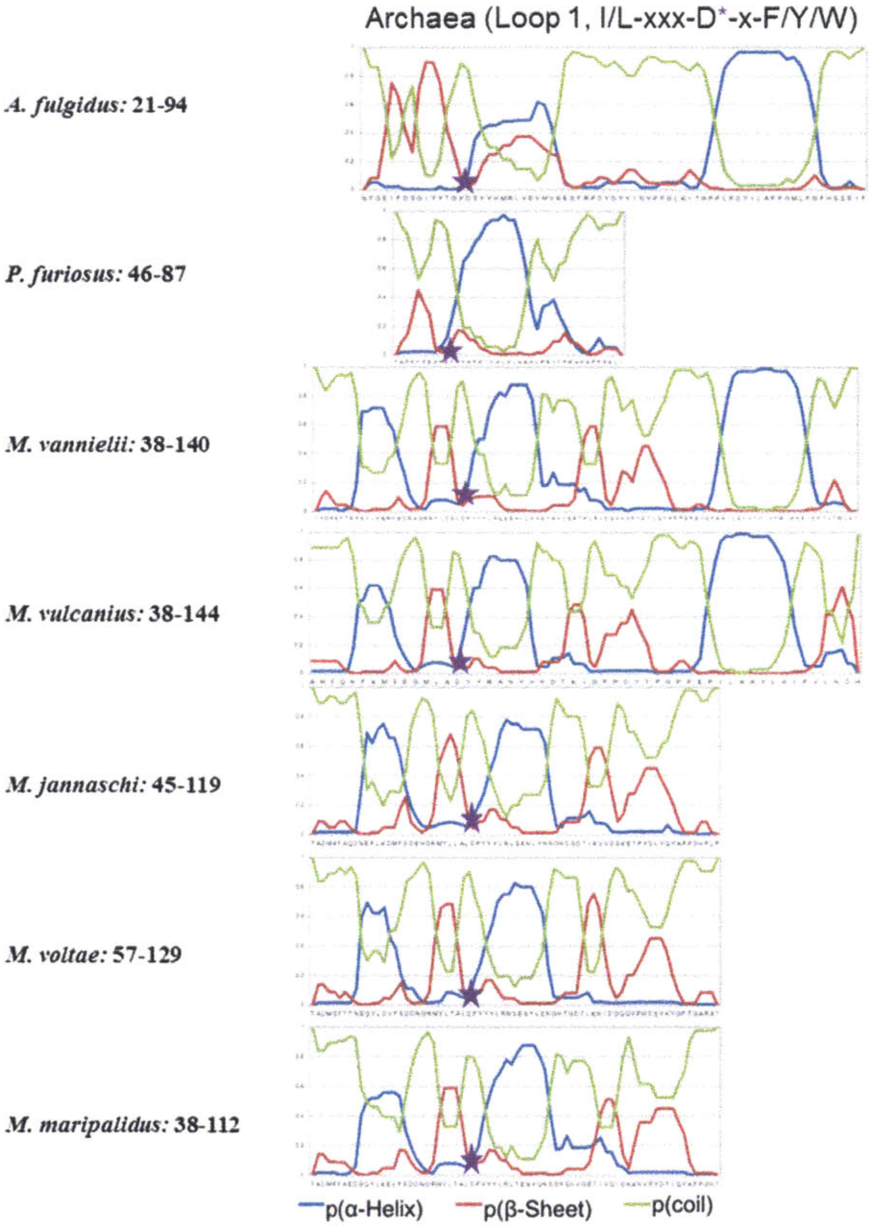


Figure 9C: Archaeal Stt3 homologs



For **Figure 10**, sections **10A**, **10B**, and **10C** respectively appear on the three pages immediately following the current one.

Figure 10: Secondary-structure predictions for regions of Stt3 sequences predicted to correspond to the soluble loop containing the conserved S/T-I/V-x-E motif.
Along the x-axis runs the protein sequence predicted for this region (residue numbers indicated next to species name) and the y-axis gives the probability that each residue is part of a coil (green), helix (blue), or β-sheet (red). Loop 5 contains the highly conserved motif S/T-I/V-x-E in Stt3 homologs (see **Figure 6**). The location of the glutamate in the motif is highlighted with a purple star. The graphs are group according to kingdom: (*A*) eukaryote, (*B*) bacteria, (*C*) archaeal.

Figure 10A: Eukaryotic Stt3 homologs

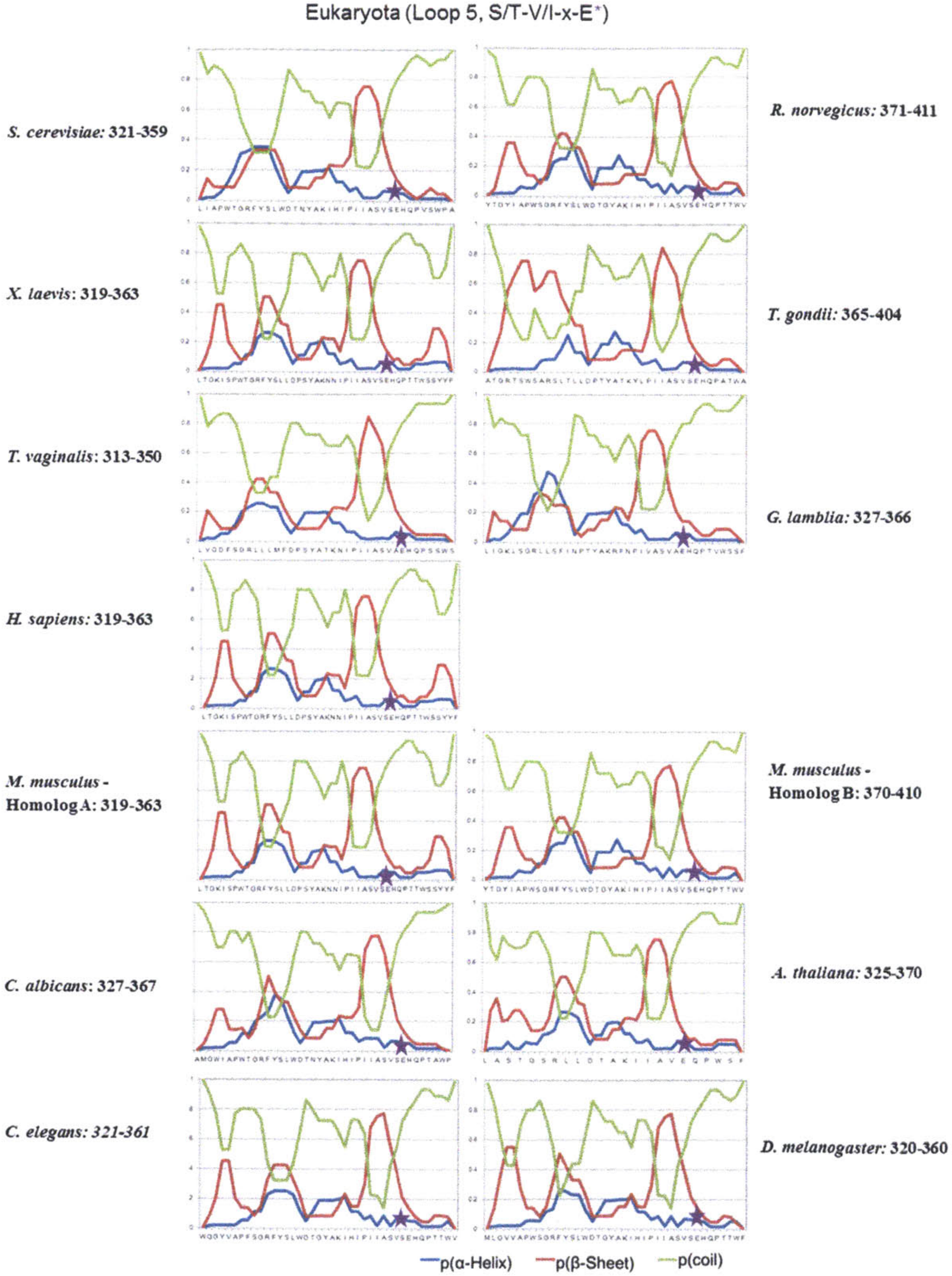


Figure 10B: Bacterial Stt3 homologs

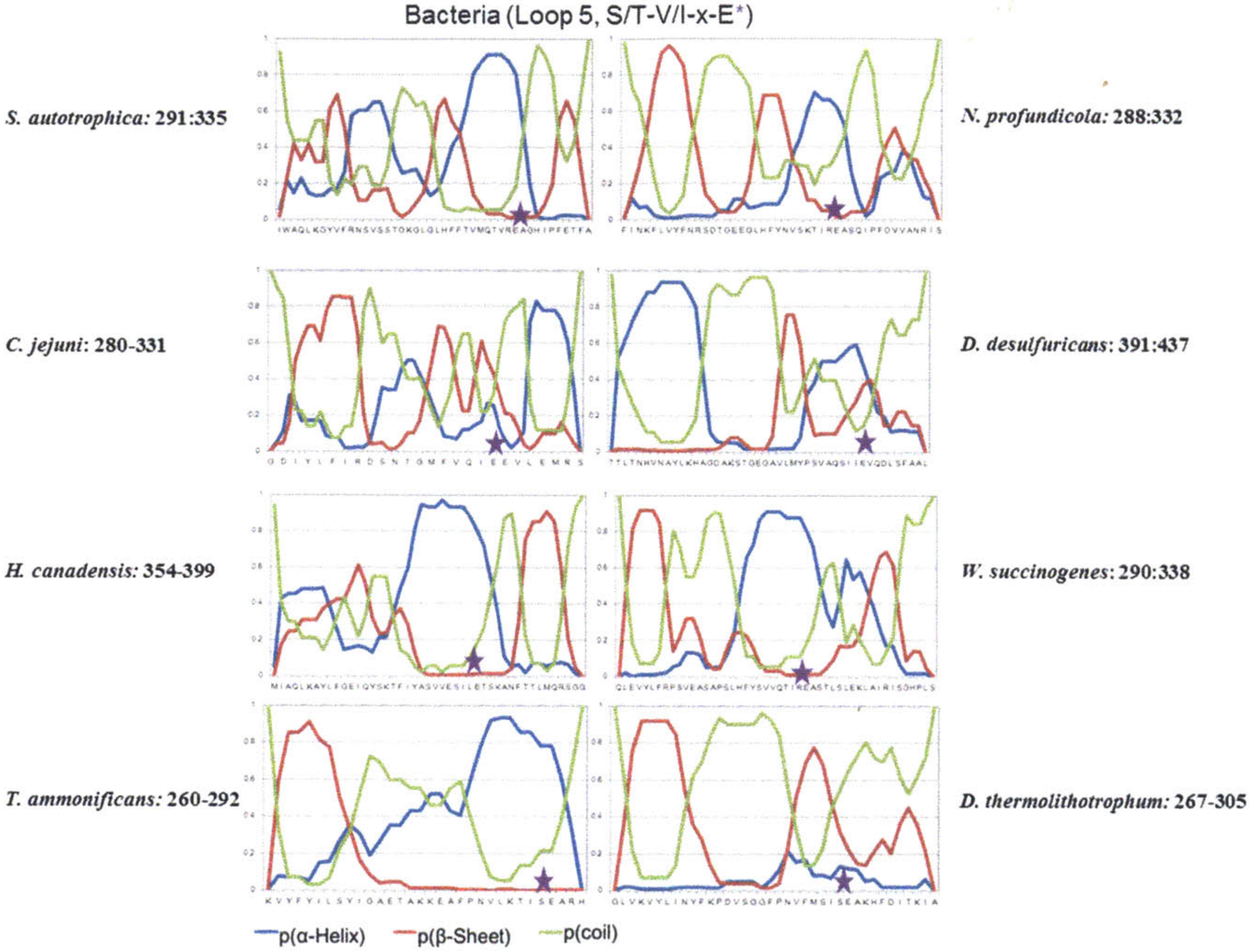
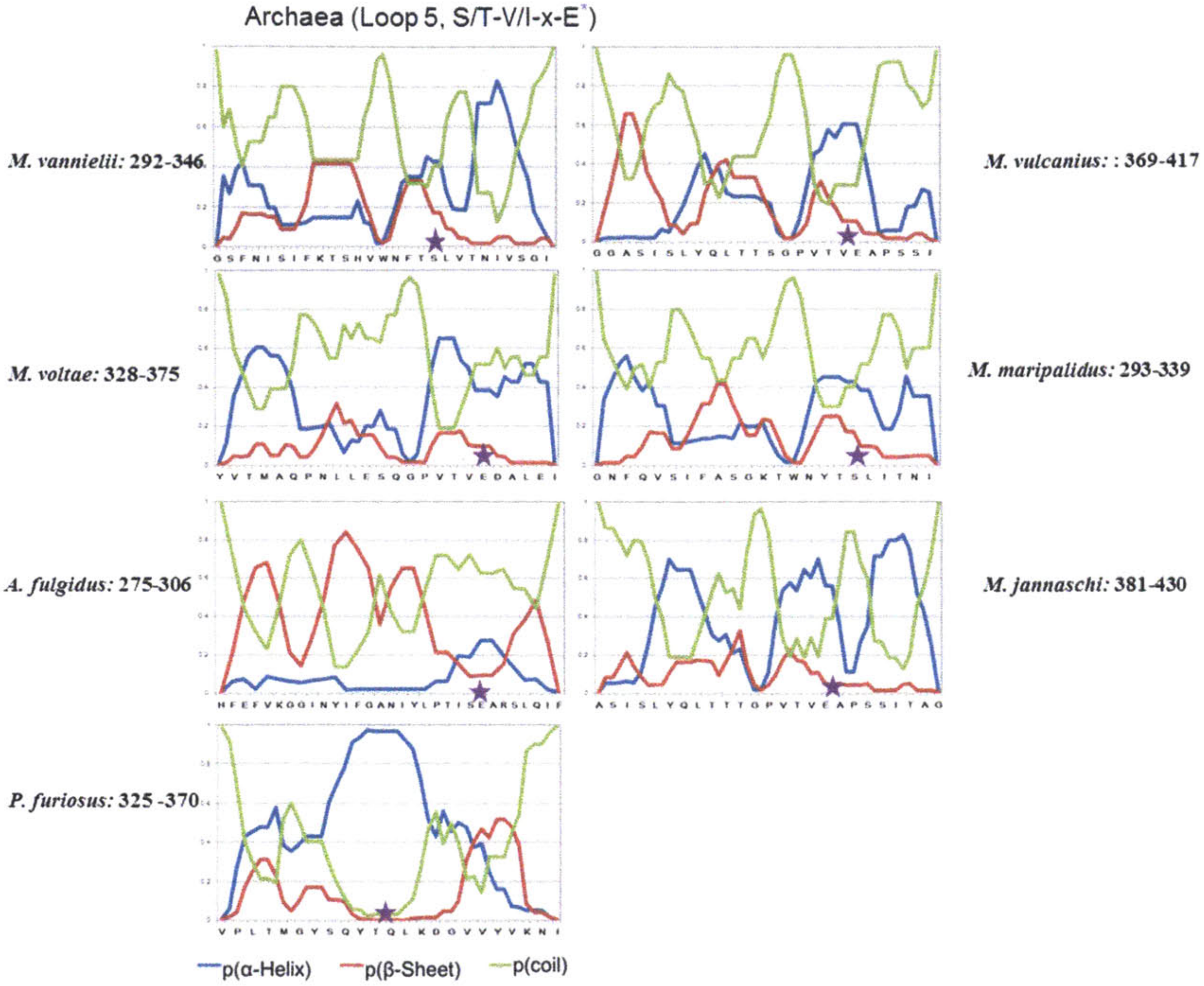


Figure 10C: Archaeal Stt3 homologs



Conclusions

In recent years, it has become abundantly clear that the conserved loop segments between transmembrane domains often play a fundamental role in substrate recognition and/or catalysis in integral membrane proteins [13-16]. However, it is particularly difficult to observe sequence conservation in these regions because motifs may be embedded within transmembrane regions that have diverged considerably through evolution, thereby complicating sequence alignments. Demonstrated here is a straightforward method for defining these buried regions of conservation applied to the complex integral-membrane OTase PglB. Topology predictions are generated for a list of homologous sequences using freely available software; a designed algorithm then parses each sequence by topological feature. These sequence segments are then more fruitful for detecting regions of conservation through sequence alignment. The systematic nature of the method allows for position-specific sequence analysis of a large number of divergent sequences of a given protein, which is crucial for determining the extent and significance of proposed motifs. These results expose the extraordinary level of conservation that exists in Stt3 homologs from bacteria through humans. Conservation over this evolutionary span implies that these regions play an essential role in the OTase activity.

With the breadth of computational prediction programs currently available, information on a protein can be gleamed without entering a laboratory. While the predictions generated will generally require experimental validation in order to be accepted, the predictions can also make evident the experiments that will likely be successful in providing results.

Materials and Methods

Bioinformatic analysis

Sequences of Stt3 homologs were chosen by searching NCBI Entrez Protein Database for 'Stt3' and selecting Stt3 homologs from evolutionarily diverse organisms. The collated set of 28 non-redundant sequences (Table 1) were subjected to topology prediction analysis via the TMHMM 2.0 server (http://www.cbs.dtu.dk/services/TMHMM/) [11]. Batch topology output was run through our 'TMH.py' program, which was written in Python code (all codes provided in Figure 5). The 'TMH' program parsed the topology prediction data and returned output files with each Stt3 GenBank Identifier (gi) number and the predicted sequence ranges for the appropriate topological location (i.e. inside (cytoplasmic), outside, or transmembrane). Each output file and the Stt3-sequence file were directed to our 'Extract.py' program, which extracted the region of the sequence corresponding to the specified range indicated for a topological characteristic. The final output file for a specific topological feature showed, for each Stt3 sequence, the gi number followed by the sequence range and its corresponding sequence (e.g. >1322489 [1:9] MGSDRSCV \n >1322489 [100:116] ALRNWLGLPIDIRNVC, etc.). Sequence analysis was carried out by selecting a feature of each Stt3 (e.g. the first 'inside loop' prediction) and aligning the corresponding sequences using MAFFT [17]. The alignments were used to identify conserved motifs within the given feature. The alignments were then manually adjusted for those Stt3 sequences in which the conserved motif did not appear exactly within the residue range predicted by TMHMM, but in a proximal region.

For secondary structure predictions, the same extracted fragments for the two major TM loops were input as a list in Fasta format to NetSurfP (http://www.cbs.dtu.dk/services/NetSurfP/)

[12]. The output contained a probability for each residue assessing that whether it is contained in an alpha-helix, beta-sheet, or coil. These probabilities were graphed in Excel.

Summary of topology sequence-extraction programs

1. Download list of desired FASTA sequences from Entrez, call this file 'fastas.txt'

2. Input fastas.txt to TMHMM topology prediction and select 'extensive output without graphics', name output file predicts.txt

3. Run TMHMM.py with command as follows:

```
python TMHMM_inside.py predicts.txt inside_ranges.txt
```

the output file 'input_ranges.txt' will have a list that looks like the following:

```
1322489 1 8
1322489 100 115
1322489 162 171
```

etc... where the first number is the gi,
the second number is the beginning residue of the inside range,
and the third is the last residue of the range

When the ranges for all of the inside-predicted regions for gi 1322489 are listed, the inside regions for the next gi (sequence) are listed, etc

4. Run extract2.py with command as follows:

```
python extract2.py fastas.txt inside_ranges.txt inside_seqs.txt
```

The output file 'inside_seqs.txt will have a list that looks like the following:

```
>1322489 [1:9]
MGSDRSCV
>1322489 [100:116]
ALRNWLGLPIDIRNVC
```

etc... giving sequences for all the ranges listed in 'inside_ranges.txt'

5. Repeat from step 3 replacing 'inside' at all points with 'outside' or 'TMhelix'
e.g. TMHMM_outside.py, outside_ranges.txt, etc.

```
#Program 1: TMH.py
#extracts topology prediction data from TMHMM output
#takes in topology results and extracts the gi number and ranges in the sequences predicted to be TMhelix (or inside, or #outside,
by changing 'TMhelix' to appropriate
S
import sys

if len(sys.argv) == 1:
    print "Error: Must input a file name"; sys.exit()
else:
    fn = sys.argv[1]
fh = open(fn)
lines = fh.readlines()
output = sys.argv[2]
sys.stdout = open(output, 'a')
seqs = []
r = ''
for line in lines:
        string = line
        if len(string) > 1:
                if string[2] == '_':
                        m = 1
                        splat = string.split()
                        j = splat[0]
                        j = j.split('_')
                        gi = j[1]
                        if splat[2] == 'outside':
                                a = splat[3]
                                a = int(a)
                                a = a - 2
                                if a < 1:
                                        a = 1
                                b = splat[4]
                                b = int(b)
                                b = b + 2
                                print gi, a, b
```

```
#Program 2: Extract.py
#first input- takes in list of fasta sequences and removes newline
#second input- file output from TMHMM program with gi and its range for a specific topological section
#output- prints the gi, the range, and the sequence corresponding to that range

import sys

if len(sys.argv) == 1:
    print "Error: Must input a file name"; sys.exit()
else:
    fn = sys.argv[1]
         seq = sys.argv[2]
fh = open(fn)
lines = fh.readlines()

output = sys.argv[3]
sys.stdout = open(output, 'a')
seqs = []
gi = ''
for line in lines:
        if line[0] == '>':
                seqs.append(gi)
                r = ''
                line = line.split('|')
                gi = line[1]
                gi = gi + '_'
        else:
                line = line.rstrip("\n")
                line = str(line)
                gi+=line
seqs = seqs[1:]
fh = open(seq)
linez = fh.readlines()
for i in linez:
        line = i.split( )
        gi = line[0]
        range1 = line[1]
        if range1 < 0:
                range1 = 1
        range2 = line[2]
        range1, range2 = int(range1), int(range2)
        range1 = range1 - 1
        for j in seqs:
                k = j.split('_')
                if gi == k[0]:
                        sec = k[1]
                        #range2 = len(sec)
                        sec1 = sec[range1:range2]
                        range1, range2 = range1 +1, range2 +1
                        range1, range2 = str(range1), str(range2)
                        print '>' + gi, '[' + range1 + ':' + range2 + ']'
                        print sec1
```

Figure 11: Python code for programs developed to parse and group sequence data according to predicted topological constraints.

References

1. Chen MM, Weerapana E, Ciepichal E, Stupak J, Reid CW, et al. (2007) Polyisoprenol specificity in the Campylobacter jejuni N-linked glycosylation pathway. Biochemistry 46: 14342-14348.
2. Yan Q, Lennarz WJ (2002) Studies on the function of oligosaccharyl transferase subunits. Stt3p is directly involved in the glycosylation process. J Biol Chem 277: 47692-47700.
3. Young NM, Brisson JR, Kelly J, Watson DC, Tessier L, et al. (2002) Structure of the N-linked glycan present on multiple glycoproteins in the Gram-negative bacterium, Campylobacter jejuni. J Biol Chem 277: 42530-42539.
4. Igura M, Kohda D (2011) Selective control of oligosaccharide transfer efficiency for the N-glycosylation sequon by a point mutation in oligosaccharyltransferase. J Biol Chem 286: 13255-13260.
5. Igura M, Maita N, Kamishikiryo J, Yamada M, Obita T, et al. (2008) Structure-guided identification of a new catalytic motif of oligosaccharyltransferase. EMBO J 27: 234-243.
6. Maita N, Nyirenda J, Igura M, Kamishikiryo J, Kohda D (2010) Comparative structural biology of eubacterial and archaeal oligosaccharyltransferases. J Biol Chem 285: 4941-4950.
7. Lizak C, Gerber S, Numao S, Aebi M, Locher KP (2011) X-ray structure of a bacterial oligosaccharyltransferase. Nature 474: 350-355.
8. Karamyshev AL, Kelleher DJ, Gilmore R, Johnson AE, von Heijne G, et al. (2005) Mapping the interaction of the STT3 subunit of the oligosaccharyl transferase complex with nascent polypeptide chains. J Biol Chem 280: 40489-40493.
9. Kim H, von Heijne G, Nilsson I (2005) Membrane topology of the STT3 subunit of the oligosaccharyl transferase complex. J Biol Chem 280: 20261-20267.
10. Nilsson I, Kelleher DJ, Miao Y, Shao Y, Kreibich G, et al. (2003) Photocross-linking of nascent chains to the STT3 subunit of the oligosaccharyltransferase complex. J Cell Biol 161: 715-725.
11. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580.
12. Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 9: 51.
13. Ashkenazi A, Viard M, Wexler-Cohen Y, Blumenthal R, Shai Y (2011) Viral envelope protein folding and membrane hemifusion are enhanced by the conserved loop region of HIV-1 gp41. FASEB J 25: 2156-2166.
14. Estrada-Mondragon A, Reyes-Ruiz JM, Martinez-Torres A, Miledi R (2010) Structure-function study of the fourth transmembrane segment of the GABArho1 receptor. Proc Natl Acad Sci U S A 107: 17780-17784.
15. McCracken LM, McCracken ML, Gong DH, Trudell JR, Harris RA (2010) Linking of Glycine Receptor Transmembrane Segments Three and Four Allows Assignment of Intrasubunit-Facing Residues. ACS Chem Neurosci 1: 482.
16. Ponsaerts R, De Vuyst E, Retamal M, D'Hondt C, Vermeire D, et al. (2010) Intramolecular loop/tail interactions are essential for connexin 43-hemichannel activity. FASEB J 24: 4378-4395.

17. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059-3066.

CHAPTER 4:

BIOCHEMICAL VALIDATION AND CHARACTERIZATION OF SEQUENCE MOTIFS CONSERVED IN OLIGOSACCHARYL TRANSFERASES

Acknowledgments

Dr. Angelyn Larkin provided the *pglB* mutant genes D519A and K522A in pET24a(+) vector.

The majority of the data in this chapter has been published in:

Jaffee, M.B. and Imperiali, B. (2011) Exploiting topological constraints to reveal buried sequence motifs in the membrane-bound N-linked oligosaccharyl transferases. *Biochemistry.* 50(35):7557-67.

Summary

The topology-driven sequence analysis discussed in Chapter 3 identified three highly conserved motifs present in the inter-transmembrane loops of oligosaccharyl transferases. This chapter discusses the mutagenesis, expression, activity, and limited proteolysis studies performed on PglB to determine whether key residues in the loop regions of PglB/Stt3s are required for function. The level of specificity required at these sites was further investigated by mutating the key residues to amino acids with similar properties to the native residue. Finally, partially active mutants were analyzed using a substrate rescue assay to determine potential involvement in substrate binding. Recent suggestions of a conserved DxxK motif based on the *P. furiosus* soluble-domain crystal structure were investigated in a similar manner by mutating and assaying appearances of this DxxK pattern in the PglB soluble domain. Coincident to the publication of this date, a structure of the full-length PglB was published [Lizak, C. et al (2011) Nature 474, 350-355]. Both the following biochemical studies and structural data have strongly verified the functional importance of the predicted motifs in catalysis.

Introduction

The previous chapter discusses a strategy used to simplify detection of conserved motifs in transmembrane proteins. When applied to the sequences for PglB and a range of Stt3 homologs, several strongly conserved motifs are revealed. Prior to detection of these motifs, very little conservation had been definitively demonstrated in Stt3 homologs, excepting the WWDxG motif present in the soluble domain of the enzyme. The publication of the structure of the soluble domain from the *P. furiosus* Stt3 had led to speculation that an aspartic acid and lysine (separated by two non-conserved residues, called DxxK) are conserved and may be involved in catalysis. However, the general frequency of the DxxK pattern makes it difficult to define homologous residues in other Stt3 congeners using sequence analysis alone.

To investigate the hypothesis of an essential DxxK motif in PglB, the three appearances of this sequence pattern in the soluble domain of PglB were mutated and each mutant was assessed in a similar way to the loop mutants. A small number of additional conserved regions had also been suggested that showed decreased glycosylation when mutated; however, the conservation demonstrated was relatively low [1-3]. All of these studies generated interesting hypotheses about the importance of additional regions of the OTase; although the lack of quantitative biochemical studies complicated efforts to define the role of these regions in the catalysis of glycan transfer.

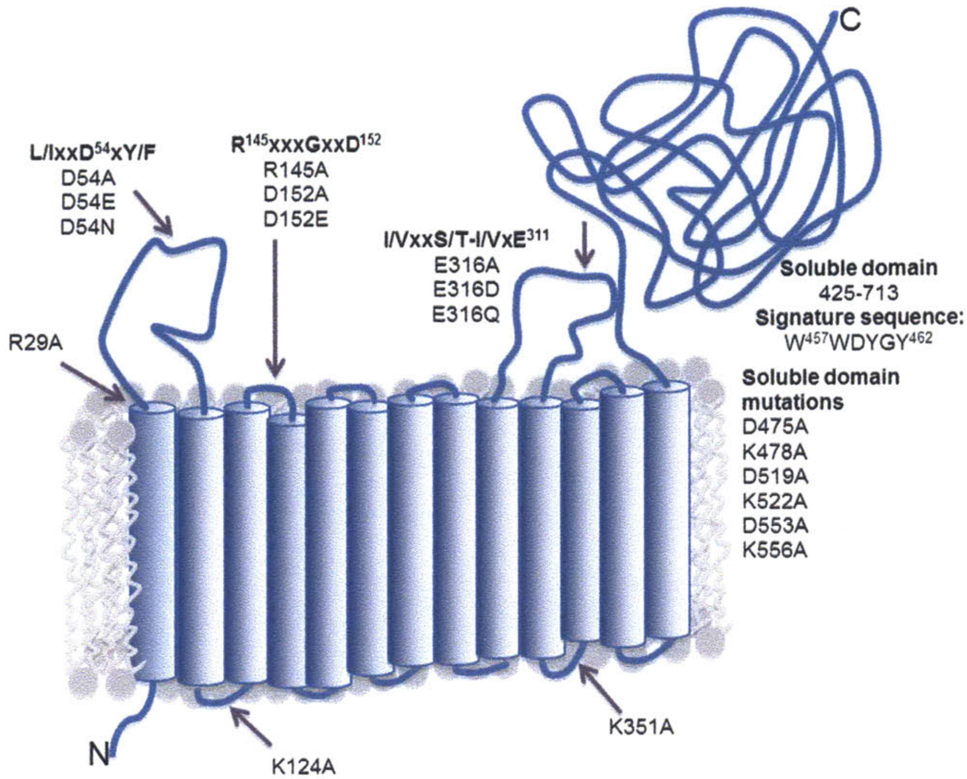

C
L/IxxD54xY/F
D54A
D54E
D54N
R145xxxGxxD152
R145A
D152A
D152E
I/VxxS/T-I/VxE311
E316A
E316D
E316Q
Soluble domain
425-713
Signature sequence:
W457WDYGY462
Soluble domain
mutations
D475A
K478A
D519A
K522A
D553A
K556A
R29A
N
K124A
K351A

Figure 1: Topology model of PglB showing locations of mutants.
Model is based on a combination of topology prediction programs and conservation of loop regions (Chapter 3) and agrees with topology of the *C. lari* crystal structure [4]. The N-terminus is located in the cytoplasm, along with the two loops containing non-conserved residues that were mutated to serve as positive controls.

Figure 1 shows a topology model of PglB with the approximate locations of the mutants described in this chapter. To experimentally examine the functional relevance of the loop motifs and the presence of an essential DxxK motif, key residues in the PglB gene were mutated and the resulting mutant constructs were transformed into and expressed in *E. coli*. In addition, several residues were mutated which were predicted to be in cytoplasmic loops and which showed no detectable conservation (**Figure 1**). These mutants (K124A and K351A) served as controls for loop-residue essentiality. The recombinant mutant proteins were purified as CEFs. Western blots were run on all CEF mutant fractions to show that each mutant was expressed at a level qualitatively similar to wild-type PglB. Fractions of the CEFs were then subjected to kinetic analysis using a highly specific assay for PglB activity. Most mutants were assayed with time points ranging from one minute to three hours; those which failed to show turnover after three hours were subjected to a time-course that extended overnight. Several of the mutants showed a complete loss of activity, while others showed only a partial loss. Loop mutant controls and several of the DxxK-candidate mutants showed wild-type activity levels. The structural integrity of the mutants relative to the wild type was probed using limited proteolysis in order to establish that the observed loss of activity for a given mutant was not the result of misfolding or large structural changes.

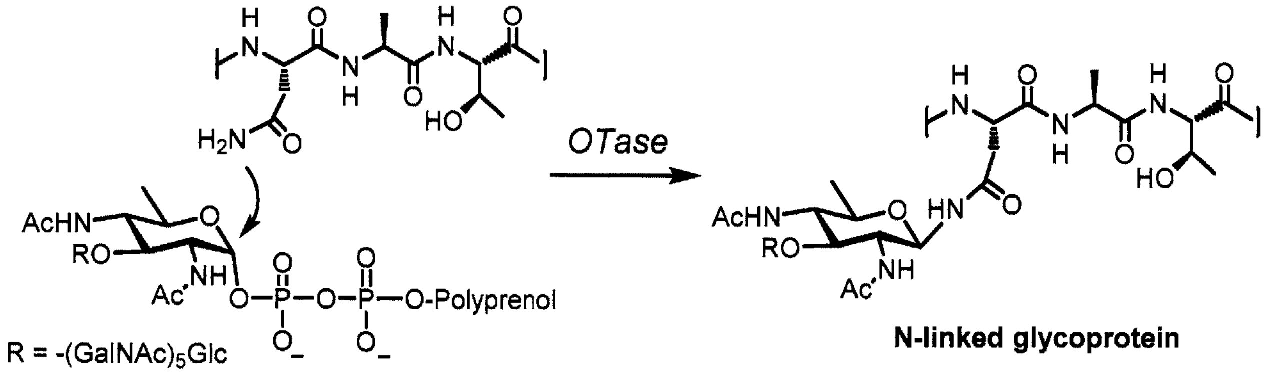


Figure 2: PglB-catalyzed N-glycosylation reaction in *C. jejuni*.
PglB transfers the heptasaccharide GlcGalNac$_5$Bac (where Bac is Bacillosamine, or 2, 4-diacetamido-2,4,6-trideoxylglucose) from undecaprenyl pyrophosphate to asparagine side chains that appear in the sequon D/E-X$_1$-N-X$_2$-S/T, in which X$_1$ and X$_2$ can be any residue other than proline.

To characterize the role of some of these mutations in PglB catalysis (**Figure 2**), an assay was designed to investigate differential effects of substrate concentration on the reaction rates. It was reasoned that the mutational effects would be most sensitive to the concentration of the undecaprenyldiphosphate-linked glycan substrate in cases where the altered residues directly impacted interactions with this substrate. Therefore, in order to establish whether the mutations influenced the binding of the glycan substrate, the effect of increasing glycosyl donor substrate concentration on catalysis was assessed. Results show that effects of certain mutations are directly correlated with substrate concentration; the effect of the mutation becomes distinctly less severe as the concentration of sugar substrate is increased. This substrate correlation provides further biochemical insight into the potential importance of these residues in the binding of the polyprenyldiphosphate-linked glycan.

Coincident with completion of these studies, a 3.4-Å resolution crystal structure was published for the PglB homolog from *Campylobacter lari* [4]. This structure represents a remarkable leap forward in our understanding of the structure and topology of the OTases. The crystal structure provides strong evidence that the aforementioned loop motifs are critical for catalysis and enzyme activity in vivo. The complementary biochemical evidence described in this chapter demonstrates the importance of the transmembrane region of PglB and indeed all Stt3 proteins for OTase function.

Results

Mutational and kinetic characterization of residues in conserved motifs

In order to investigate the importance of the three motifs in catalysis, site-directed mutagenesis was performed on several conserved residues in PglB. As negative controls, residues (K124 and K351) that would be predicted to appear in non-conserved loop regions were also mutated (**Figure 1**). All candidate residues were mutated to alanine, and each acidic residue was additionally mutated to its amide-containing complement (Asn for Asp, Gln for Glu), and/or its acidic complement residue (Asp for Glu, or vice versa).

In addition, the proposal that a DxxK motif is required for OT catalysis was investigated [2]. The crystal structure of the soluble domain of *P. furiosus* Stt3 suggests a role for an aspartic acid and a lysine, separated by two residues, fifty residues C-terminal to the signature WWDxG motif; when mutated, a decrease in OTase activity was previously observed [2]. The natural frequency of the aspartic acid and lysine residues and thus the DxxK pattern within the PglB sequence makes it difficult to surmise functionally equivalent residues in other Stt3 sequences. PglB has three appearances of DxxK in the C-terminal soluble domain: $D^{475}xxK^{478}$, $D^{519}xxK^{522}$, and $D^{553}xxK^{556}$. To establish whether a functionally conserved DxxK motif exists in PglB, each aspartic acid and lysine within these DxxKs were mutated to alanine and analyzed along with the loop mutations.

All mutants were overexpressed in *E. coli*, as described in Materials and Methods. A wild type (WT) PglB strain and a strain containing the empty expression vector (referred to as the 'blank') served as the positive and negative controls, respectively. All mutants and WT were expressed with an N-terminal T7-tag and a C-terminal His_{10}-tag. Mutant cultures were purified to the point of 'washed CEF' (see Chapter 2), so each mutant solution was expected to contain the integral-membrane proteins and associated lipids only. Ideally, each mutant enzyme would be purified to homogeneity for quantification; however, the unstable nature of membrane proteins upon solubilization and purification can result in a source of error with unpredictable effects on activity measurements. Thus, it was considered preferable to use the mutant and wild-type enzymes in the more stable cell envelope fractions and to establish similar enzyme concentrations by using western blot analysis. A fraction of each semi-pure mutant CEF was run on a western blot alongside the WT and blank. Immunostaining with the anti-His and anti-T7 antibodies confirmed comparable expression levels of PglB in the fractions assayed and that the full-length gene was being expressed. **(Figure 3)**.

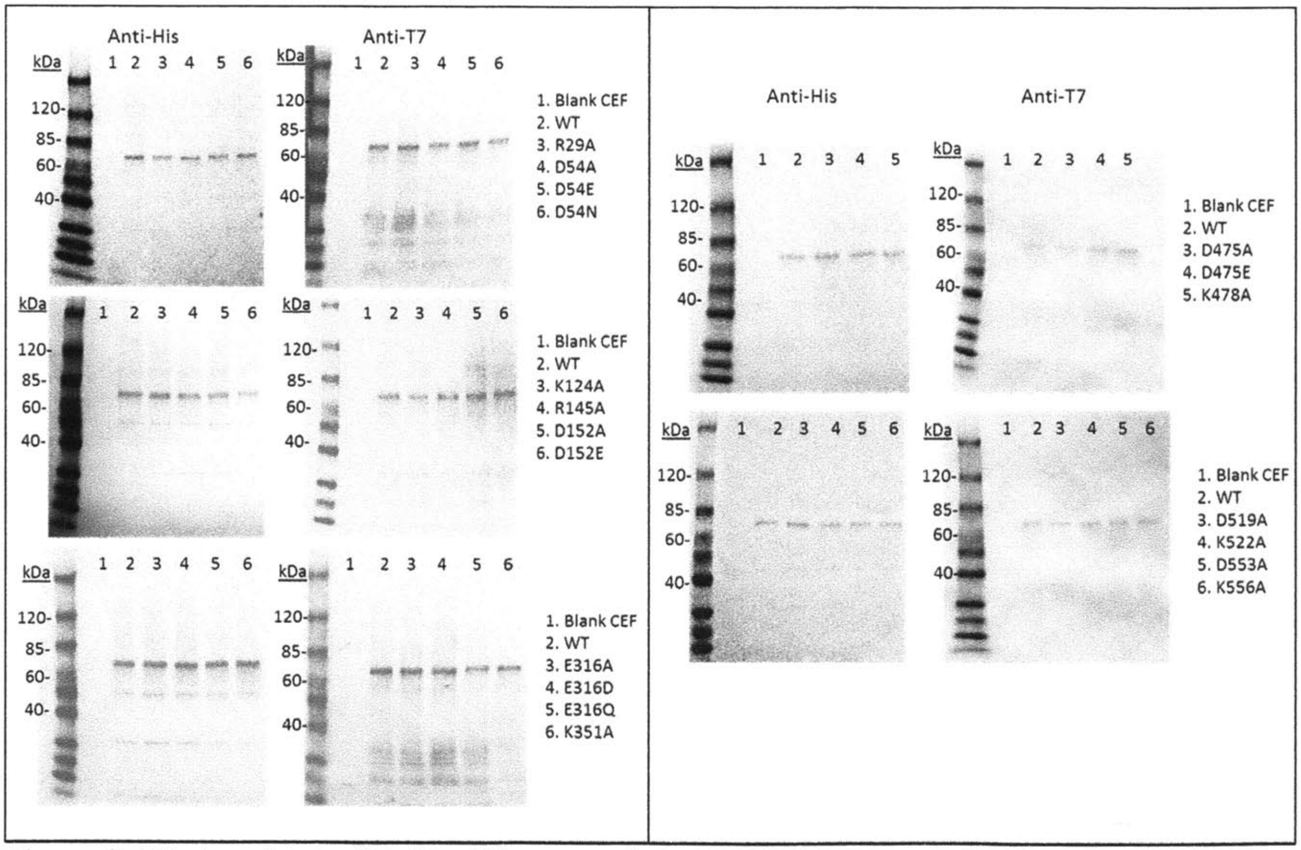

Anti-His
Anti-T7
kDa
1 2 3 4 5 6
120-
85-
60-
40-
1. Blank CEF
2. WT
3. R29A
4. D54A
5. D54E
6. D54N
1. Blank CEF
2. WT
3. K124A
4. R145A
5. D152A
6. D152E
1. Blank CEF
2. WT
3. E316A
4. E316D
5. E316Q
6. K351A
Anti-His
Anti-T7
1 2 3 4 5
1. Blank CEF
2. WT
3. D475A
4. D475E
5. K478A
1. Blank CEF
2. WT
3. D519A
4. K522A
5. D553A
6. K556A

Figure 3: Western blots of PglB wild-type and mutant CEFs show similar expression levels. All blots include a negative control ('Blank CEF') and a positive wild-type PglB control. Western blots were visualized using the anti-tetra-His antibody (which also recognizes the C-terminal His_{10} tag) as well as the T7-tag antibody (which recognizes the N-terminal T7 tag of each protein).

The activities of PglB and each of the mutants were measured using a radioactivity-based assay, which measures transfer of a radiolabeled glycan from the undecaprenyldiphosphate carrier to a peptide bearing the PglB consensus sequence (see Materials and Methods). Results from activity assays for PglB enzymes with mutations in conserved loop motifs and the first DxxK motif are shown in **Figure 4**. Activity assays for mutants showing wild-type levels of activity along with extended time points for loop mutants are shown in **Figure 5**. Mutant activity results are summarized in **Table 1**. Mutants were divided into broad categories based on relative activity level, as follows: WT-level activity (+++), decreased activity (++), activity detectable only with overnight incubation (+), or no detected activity (-).

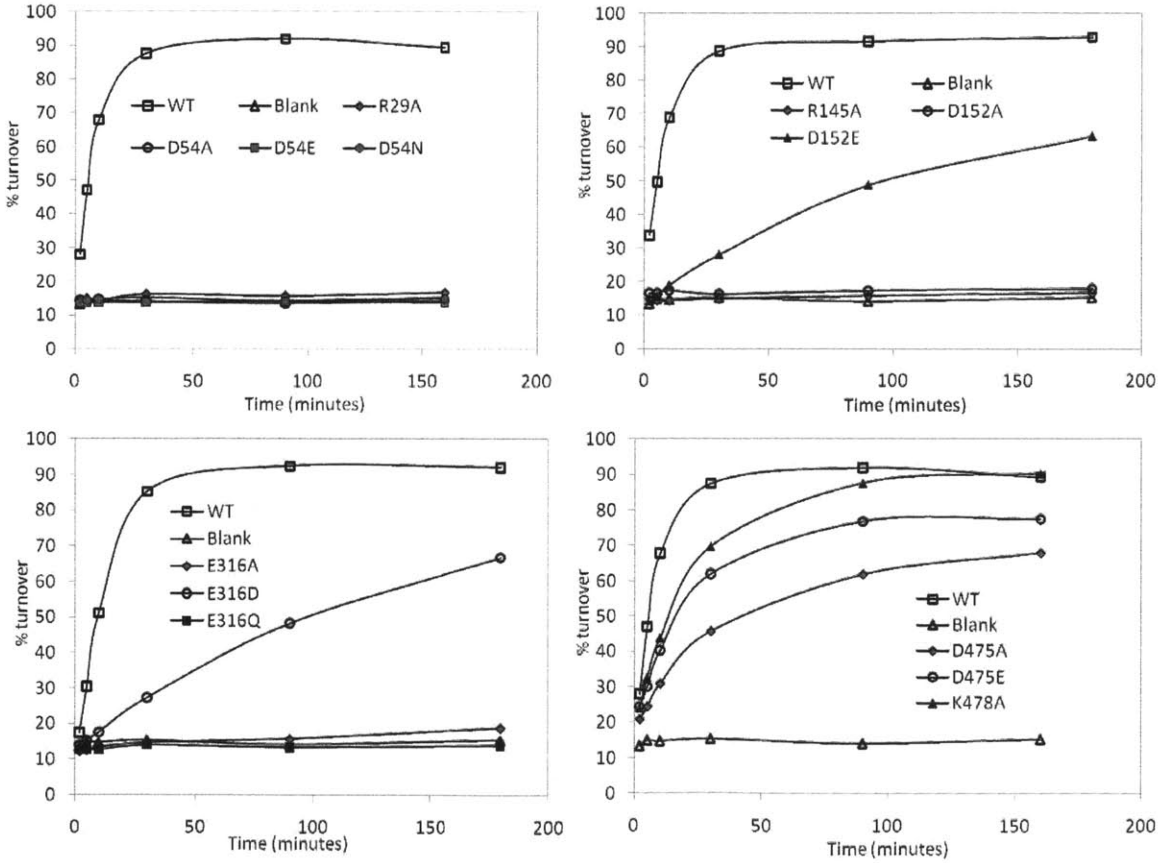


Figure 4: PglB activity assays performed on CEFs of loop mutants.
Each assay set included a negative control ('Blank CEF') and a positive PglB control (WT). Each plot indicates the mutants assayed in the legend. Shown here are those with mutations in the conserved loop motifs and the D^{475}xxK motif.

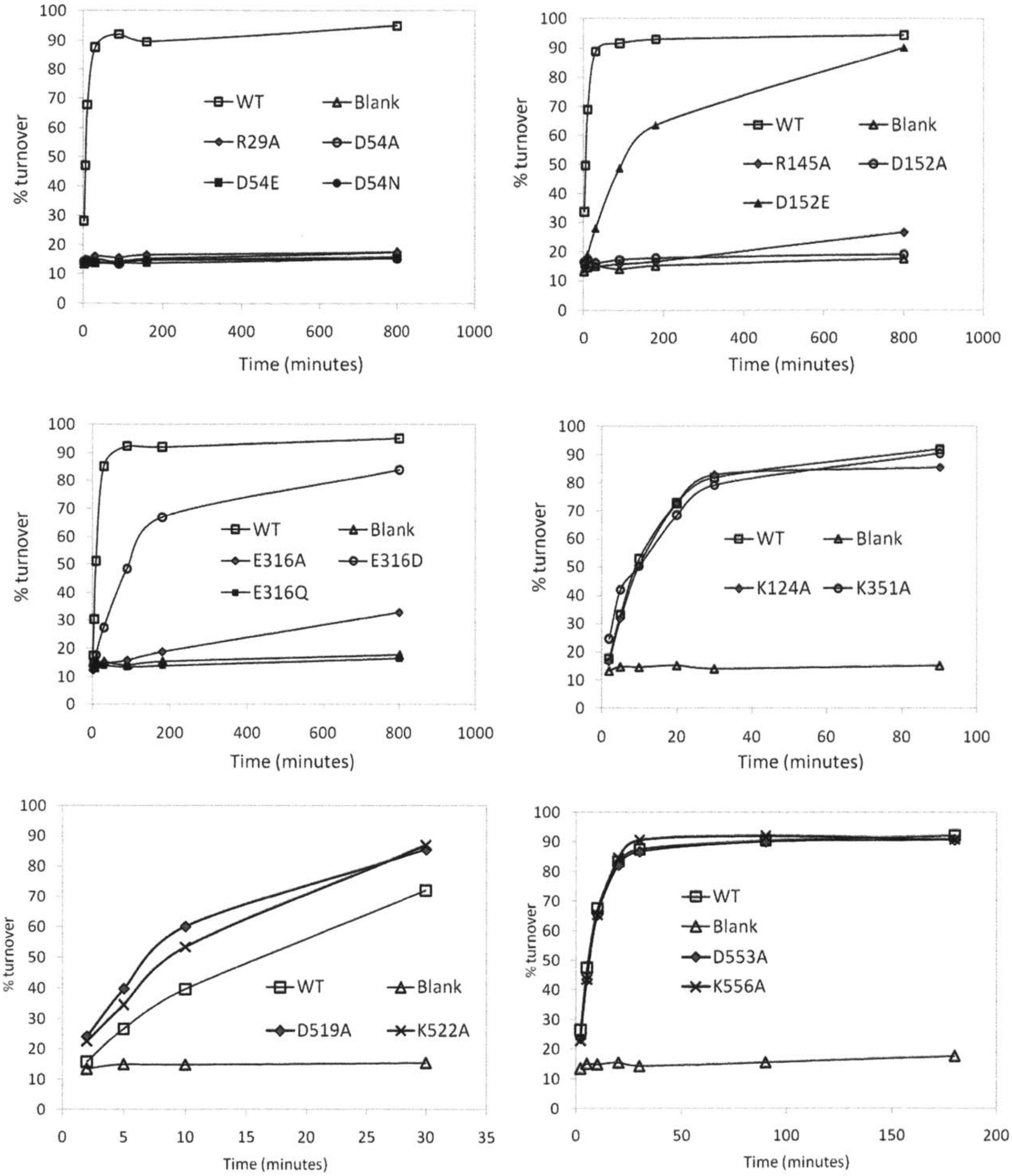


Figure 5: Activity data for soluble-domain mutants and extended time points for loop mutants.

Assays for negative-control mutants K124A and K351A are also shown along with those for the D^{519}xxK and D^{553}xxK mutants.

Table 1: Relative activity and locations of PglB mutants.
Activity was considered null (-) if no activity is seen over the blank CEF background, +++ if levels are comparable to wild type (WT), ++ if levels are distinctly lower than WT, and + if levels are detected only with overnight incubation.

Mutant	**Location**	**Activity**
Blank CEF	---	-
WT	---	+++
R29A	between TM domains 1 & 2	-
D54A	between TM domains 1 & 2	-
D54E	between TM domains 1 & 2	-
D54N	between TM domains 1 & 2	-
K124A	between TM domains 2 & 3	+++
R145A	between TM domains 3 & 4	+
D152A	between TM domains 3 & 4	-
D152E	between TM domains 3 & 4	++
E316A	between TM domains 7 & 8	+
E316D	between TM domains 7 & 8	++
E316Q	between TM domains 7 & 8	-
K351A	between TM domains 10 & 11	+++
D475A	soluble domain	++
D475E	soluble domain	++
K478A	soluble domain	++
D519A	soluble domain	+++
K522A	soluble domain	+++
D553A	soluble domain	+++
K556A	soluble domain	+++

Mutants maintain tertiary structure

Mutants that showed a decrease or loss in activity were analyzed by limited proteolysis, which examines the possibility that the mutation causes a loss in activity by disrupting folding of the mutant. **Figure 6** shows western-blot degradation profiles for wild-type PglB and all mutants. Before protease is added (time 0) the bands of the major digestion products are already visible at very low levels, perhaps due to endogenous *E. coli* protease activity. Over time, individual bands appear or change in intensity to give a characteristic profile, which was used to establish structural integrity in mutants with loss of activity. As can be seen in **Figure 6**, all mutants display similar degradation profiles to the wild-type protein. Limited proteolysis cannot provide conclusive evidence that the mutants are as stable as the WT enzyme or that the mutants are capable of adopting an active conformation. However, the presence of a clear time-dependent degradation profile is a strong indicator that on a global, tertiary level the mutant enzymes adopt a WT-like fold.

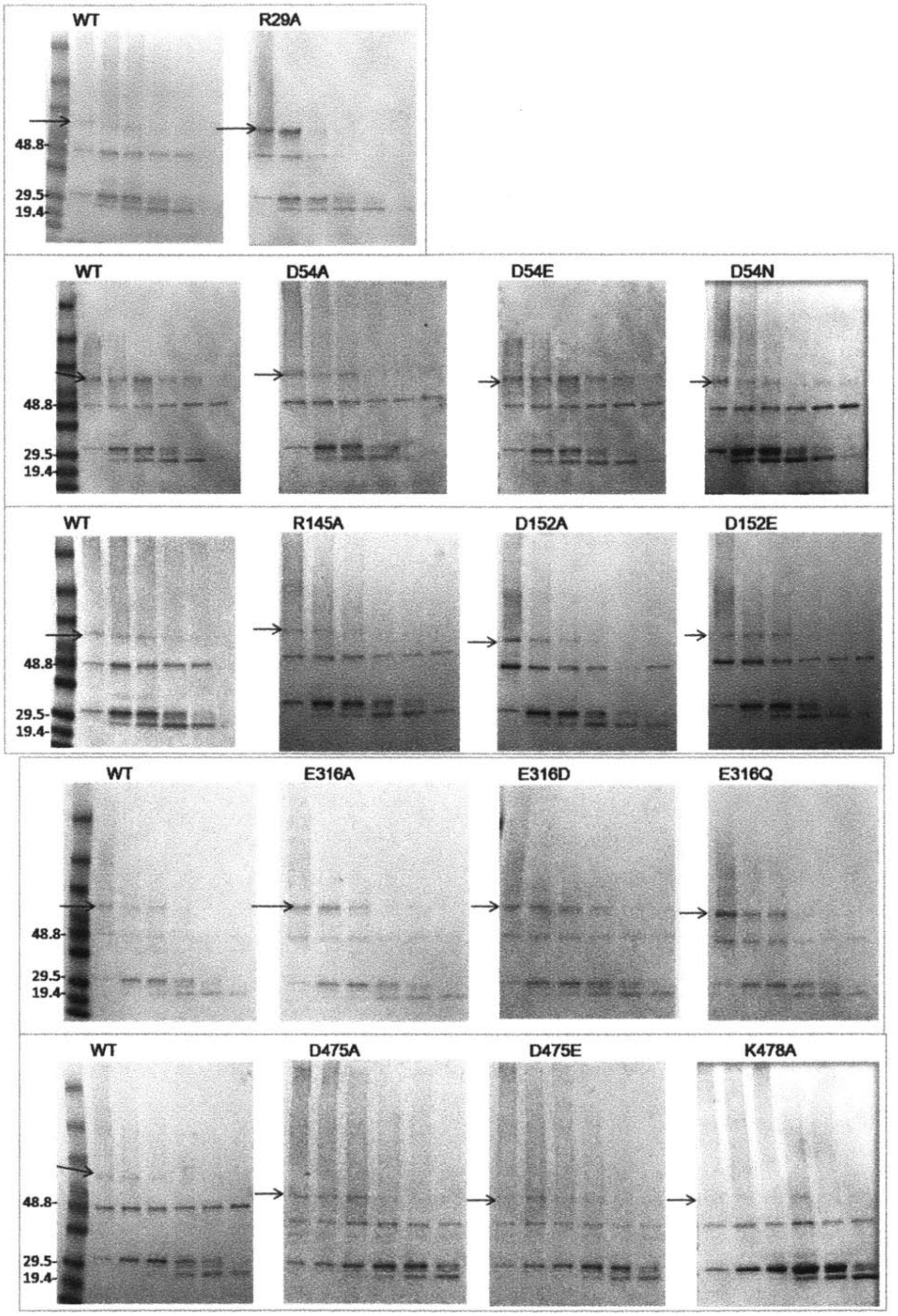

WT
R29A
48.8-
29.5-
19.4-
WT
D54A
D54E
D54N
48.8-
29.5-
19.4-
WT
R145A
D152A
D152E
48.8-
29.5-
19.4-
WT
E316A
E316D
E316Q
48.8-
29.5-
19.4-
WT
D475A
D475E
K478A
48.8-
29.5-
19.4-

Figure 6: Probing effects of mutations on PglB tertiary structure using limited proteolysis. Mutants were assayed in sets, each accompanied by wild-type PglB for comparison. Anti-His western blot analysis was performed on fractions of the reaction quenched at 0, 5, 10, 30, 60, and 180 minutes. Arrows indicate the location of the full-length PglB band at the start of the reaction. Numbers to the left of wild-type blots indicate molecular weights (kDa) of standards.

Correlation of mutation effects with concentration of Und-PP-disaccharide substrate

It was determined that the effect of mutants D152E, E316D, and D475A and D475E are inversely correlated with concentration of Und-PP-glycan substrate: at increasing substrate concentrations, the mutant rates approached the wild-type rate (**Figure 7**). Initial mutant enzyme assays (**Figures 4, 5**) were performed in the presence of saturating peptide substrate (50 x apparent K_M) and a relatively low concentration of radiolabeled disaccharide substrate (0.01x apparent K_M), which causes the measured rates to be highly sensitive to changes in binding efficiency of undecaprenyldiphosphate-disaccharide and insensitive to changes in peptide binding. The initial assay was constructed in this manner to test the hypothesis that the conserved loop motifs, given the acidic nature of many key residues, might be involved in metal-ion mediated binding of the diphosphate in the glycosyl donor substrate since PglB is known to require divalent cations (generally Mn(II) or Mg(II)) for activity. Then, in order to establish whether the mutations influenced the binding of the undecaprenyldiphosphate-linked glycan, we assessed the effect of increasing glycosyl donor substrate concentration on catalysis. The initial rates for these mutants were measured at three separate concentrations of this substrate: 0.01, 0.1, and 1.0 μM (**Figure 7**).

Initial western blot analyses ensured that a similar level of each PglB construct was expressed, providing confidence such that mutants could be reliably grouped into broad categories based on activity levels (**Table 1**). However, to further characterize the partially active mutants by directly comparing initial turnover rates, a more precise western blotting method was used to quantify relative levels of protein (see Materials and Methods). It was determined that levels of all mutants were very similar, with a maximum concentration difference of +1.7-fold relative to WT (**Figure 8, Table 2**). The relative level of each mutant in CEF measured by western blot analyses was reproducible, allowing determination and correction of any effects on rate comparisons.

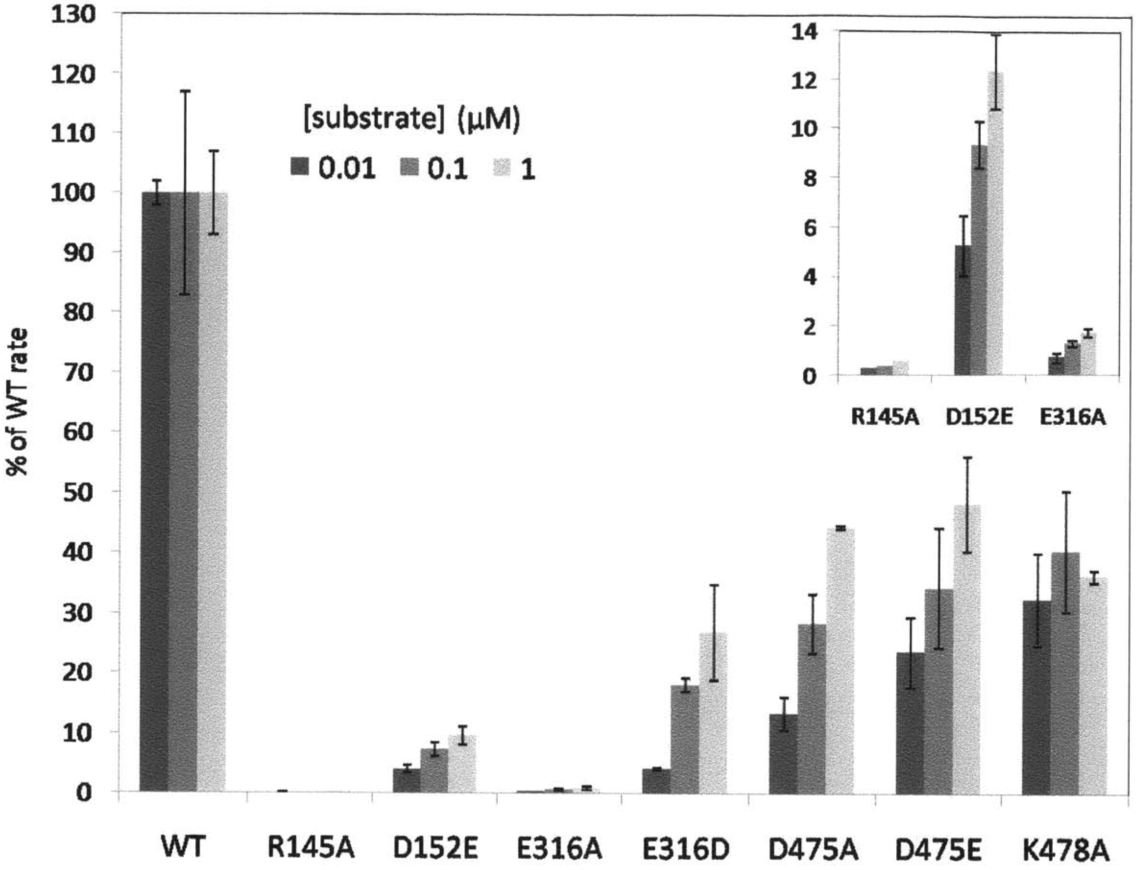


Figure 7: Partial rescue of mutant activity with increasing sugar-substrate concentrations. Average initial rates for low-activity PglB mutants assayed at three concentrations of polyprenyldiphosphate disaccharide substrate (0.01, 0.1, and 1.0 μM). Percent of WT rate for each mutant is determined at each concentration, with the average rate of WT set to 100%. Inset: magnification of R145A, D152E, and E316A values for clarity. Rate values shown are corrected for small variations in concentration, as determined by quantitative western blot analysis. Raw rate data can be found in the **Table 2**.

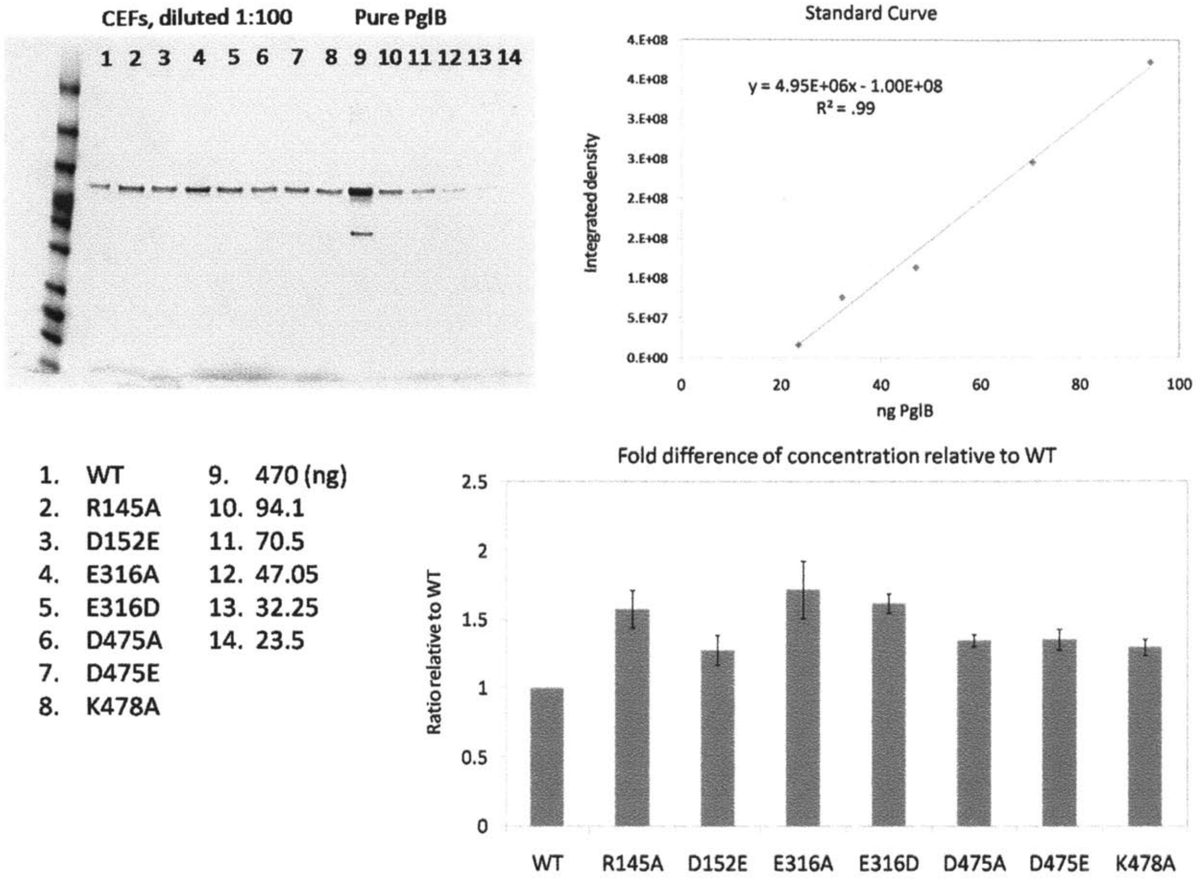


Figure 8: Quantitative western blotting of mutants with partial activity loss. Quantitative blotting was performed to ensure observed activity losses are not significantly influenced by differences in enzyme concentration. Average fold differences in activity were used to correct for variations when measuring relative rate data at varying substrate concentrations. As indicated in the legend below the blot, lanes 1-8 show mutant CEFs and lanes 9-14 show various levels of pure, quantified PglB. The intensities of the varied amounts of pure protein are used to derive a standard curve (upper right). This curve was used to estimate the quantities of protein in the CEFs shown in the blot. Fold differences in enzyme levels are shown in the bar graph on the bottom right. Error bars indicate the standard error from averages of multiple western blots.

Table 2: Activity of mutants relative to wild-type at various sugar-substrate concentrations. Percent of wild-type rate at varied concentrations of polyprenyldiphosphate-disaccharide substrate are shown for mutants tested. Left table contains numbers derived directly from measured rates. Right table contains numbers that have been corrected for concentration variation as measured by quantitative western blot.

% of WT rate,
directly from measured rates

	0.01 μM	**0.1 μM**	**1 μM**
WT	100	100	100
R145A	0.15	0.24	0.37
D152E	4.2	7.4	9.7
E316A	0.3	0.8	1.0
E316D	4.6	18.2	27.0
D475A	18.1	28.3	44.5
D475E	33.5	34.4	48.4
K478A	45.4	40.5	36.4

% of WT rate,
corrected for concentration differences

	0.01 μM	**0.1 μM**	**1.0 μM**
WT	100	100	100
R145A	0.13	0.15	0.23
D152E	3.2	5.8	7.6
E316A	0.2	0.4	0.6
E316D	3	11	17
D475A	10	21	33
D475E	17	25	36
K478A	25	31	28

Discussion

Interpretation of PglB mutants in context of kinetic and structural data.

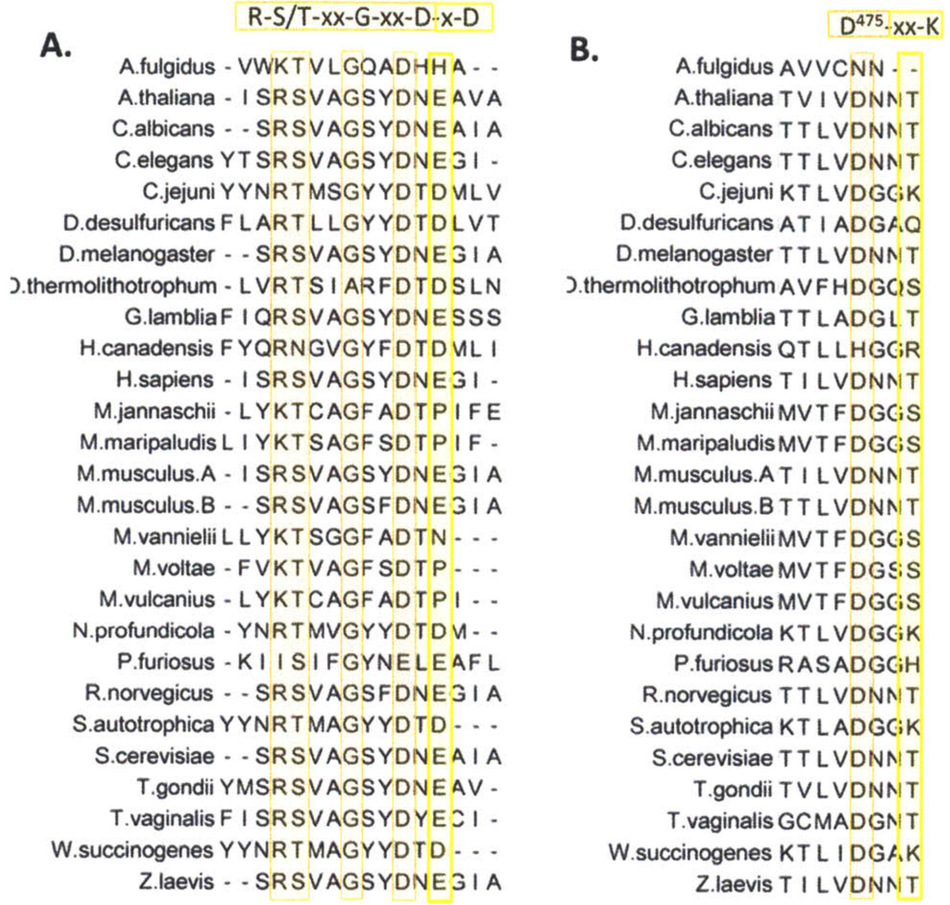
The bioinformatic and biochemical data that we report can now be framed in the context of the recent structure analysis of the Stt3 homolog from *C. lari*, also designated as PglB [4]. In this structure, the conserved acidic residues in each loop motif (D56, D154, and E319 of the *C. lari* PglB) are proposed to form a pocket that accommodates a divalent magnesium cation and the nucleophilic asparagine. In these structural studies, the essentiality of these residues is supported by use of an in vivo glycosylation assay, which exploits the gel shift observed upon glycosylation.

The biochemical results that we present provide a more detailed characterization of the residues implicated by the structure, and additionally expose critical residues that are not made evident by the structural data. The aspartic acid in the [L/I]xx[D^{54}]x[Y/F] motif (D54) can be set apart from the other two implicated acidic residues (D152 and E316) by the in vitro activity measurements. In particular, D152 and E316 can both be mutated to their acidic counterparts (D152E and E316D) and still retain activity, albeit at a notably decreased level. This indicates the primary role that the negative charge plays at these sites, as the alanine mutations at these sites show minimal activity, and E316Q showed no activity. In contrast, with D54, we observe that when mutated to alanine, asparagine, or glutamic acid, activity is completely eliminated. This establishes D54 as highly specific for its role in function with respect to both charge and size and suggests a pivotal role in catalysis.

It is also observed that the effects of certain mutations can be directly correlated with substrate concentration; the effect of the mutation becomes less pronounced as the concentration of the glycosyl donor substrate is increased. For example, this is observed with D152E and E316D

(Figure 7). For other mutations, including D54 and K478A the increased Und-PP-Bac-GalNAc substrate concentration had little effect. This substrate correlation analysis suggests that these residues are implicated in the binding of the polyprenyldiphosphate-linked glycan. This trend is consistent with the reported structural analysis, which suggests that these residues coordinate to the divalent cation, and the hypothesis that the divalent cation is further involved in coordinating the disphosphate of the glycan substrate. Further biochemical analyses using more complex glycosyl donor substrates may help determine the relative effect of the mutations on substrate binding and catalysis.

With respect to the [R/K][S/T]xx[G^{149}]xx[D^{152}] motif, we find that R145 and D152, when individually mutated to alanine, lead to an almost complete loss of PglB activity. Mutation of D152 to a glutamate restores partial activity. In this context, a fundamental role for the aspartic acid at the +2 position to the D152 (D154 in *C. jejuni* PglB, D156 in *C. lari* PglB) has been proposed [4]. Specifically, this pair of aspartic acids is proposed to represent a 'DxD' motif that is characteristic in the glycosyltransferase superfamily GT-C [4-6]. However, the alignments derived from our topology-guided motif analysis suggest that the sequence characteristics in this region of the OTase are not definitively conserved across evolution. Indeed, it is evident from our alignment of the motif **(Figure 9A)** that the residue +2 to the defined motif is a glutamic acid in almost fifty percent of homologs examined, and this residue is a proline in the seven archaeal homologs that were analyzed in this study. Site-directed mutagenesis data are not yet available to evaluate the kinetic significance of the second aspartic acid in the DxD motif in PglB and the natural frequency of the generic and variable motif leaves significant room for further exploration. Therefore the association with the characteristic GT-C superfamily motif may not be clear.



A.

R-S/T-xx-G-xx-D-x-D

A.fulgidus	- V W K T V L G Q A D H H A - -
A.thaliana	- I S R S V A G S Y D N E A V A
C.albicans	- - S R S V A G S Y D N E A I A
C.elegans	Y T S R S V A G S Y D N E G I -
C.jejuni	Y Y N R T M S G Y Y D T D M L V
D.desulfuricans	F L A R T L L G Y Y D T D L V T
D.melanogaster	- - S R S V A G S Y D N E G I A
D.thermolithotrophum	- L V R T S I A R F D T D S L N
G.lamblia	F I Q R S V A G S Y D N E S S S
H.canadensis	F Y Q R N G V G Y F D T D M L I
H.sapiens	- I S R S V A G S Y D N E G I -
M.jannaschii	- L Y K T C A G F A D T P I F E
M.maripaludis	L I Y K T S A G F S D T P I F -
M.musculus.A	- I S R S V A G S Y D N E G I A
M.musculus.B	- - S R S V A G S F D N E G I A
M.vannielii	L L Y K T S G G F A D T N - - -
M.voltae	- F V K T V A G F S D T P - - -
M.vulcanius	- L Y K T C A G F A D T P I - -
N.profundicola	- Y N R T M V G Y Y D T D M - -
P.furiosus	- K I I S I F G Y N E L E A F L
R.norvegicus	- - S R S V A G S F D N E G I A
S.autotrophica	Y Y N R T M A G Y Y D T D - - -
S.cerevisiae	- - S R S V A G S Y D N E A I A
T.gondii	Y M S R S V A G S Y D N E A V -
T.vaginalis	F I S R S V A G S Y D Y E C I -
W.succinogenes	Y Y N R T M A G Y Y D T D - - -
Z.laevis	- - S R S V A G S Y D N E G I A

B.

D^{475}-xx-K

A.fulgidus	A V V C N N - -
A.thaliana	T V I V D N N T
C.albicans	T T L V D N N T
C.elegans	T T L V D N N T
C.jejuni	K T L V D G G K
D.desulfuricans	A T I A D G A Q
D.melanogaster	T T L V D N N T
D.thermolithotrophum	A V F H D G Q S
G.lamblia	T T L A D G L T
H.canadensis	Q T L L H G G R
H.sapiens	T I L V D N N T
M.jannaschii	M V T F D G G S
M.maripaludis	M V T F D G G S
M.musculus.A	T I L V D N N T
M.musculus.B	T T L V D N N T
M.vannielii	M V T F D G G S
M.voltae	M V T F D G S S
M.vulcanius	M V T F D G G S
N.profundicola	K T L V D G G K
P.furiosus	R A S A D G G H
R.norvegicus	T T L V D N N T
S.autotrophica	K T L A D G G K
S.cerevisiae	T T L V D N N T
T.gondii	T V L V D N N T
T.vaginalis	G C M A D G N T
W.succinogenes	K T L I D G A K
Z.laevis	T I L V D N N T

Figure 9: Relative conservation of residues within and surrounding conserved motifs. *A.* Alignment of sequences containing OTase motif in second transmembrane loop R-S/T-xx-G-xx-D^{152} (orange outline). The proposed overlapping motif D^{152}-x-D is included in alignment (aspartate at position +2 is highlighted yellow). *B.* Alignment of first appearance of D-xx-K pattern (D^{475}, K^{478}) in soluble domain of PglB and analogous positions in other OTases. Aspartate 475 is highlighted in orange and lysine 478 is highlighted yellow.

Interestingly, the arginine (R29) in *C. jejuni* PglB that appears at the C-terminus of the first transmembrane helix is also conserved and essential: when mutated to alanine, activity is eliminated. In view of this data alone, the role that this arginine may play in PglB function is not obvious. However, when the equivalent residue (R31) is examined in the *C. lari* structure one observes that this arginine side chain appears to be in direct contact with threonine 148 (or T146 in *C. jejuni* PglB). In this context, we note that this position is highly conserved as a hydroxyamino acid (either a serine or threonine) throughout our 28 examined Stt3 homologs and appears in the

[R/K][S/T]xx[G^{149}]xx[D^{152}] motif in the third transmembrane loop (**Figure 1**). This apparent interaction again highlights the unique ability of this topology-guided sequence analysis to reveal residues key to enzyme function. The high conservation of this hydroxyamino acid and its apparent interaction with the essential arginine in the first transmembrane helix allows us to propose that these residues may be involved in mediating in a conformational shift of the enzyme, as the interaction represents a clear link between the catalytic site and an integral membrane helix. In the future, it will be very interesting to gain a structural analysis of other substrate-bound forms of the enzyme to determine the likelihood of this scenario.

Interpretation of DxxK mutants in context of kinetic and structural data

Of the three appearances of DxxK sequon in the PglB soluble domain, mutations in D475 and K478 negatively impacted PglB activity, while [D^{519}xxK^{522}]and [D^{553}xxK^{556}] alanine mutants showed levels comparable to wild type (**Figure 4**). The effect on D475A was more significant than that on K478A (**Figure 4**), which calls the proposed DxxK motif into question. Considering the enhanced impact of D475A relative to K478A, and the additional fact that this aspartic acid is highly conserved throughout all OTases while the lysine is not (**Figure 9B**), it appears that a 'DxxK motif' is not playing an essential role in OTase activity. However, the high level of conservation of D475 and the clear impact of the mutation on activity indicates that this residue is likely involved the OTase function. Interestingly, an increase in the concentration of the polyprenyldisphosphate-linked glycan in the assay steadily attenuated the effect of D475A and D475E on activity, while no such trend is seen for K478A (**Figure 7**). As in the case of D152 and E316, this trend suggests that mutation at these sites affects the ability of PglB to bind the polyprenyldiphosphate glycan substrate.

Mutants maintain tertiary structure

Limited proteolysis represents powerful tool for identifying flexible, exposed regions of proteins and for studying how tertiary structure (and thus susceptibility to proteolysis) is affected by mutations, substrate presence and many other factors [7-11]. Mutants that showed a decrease or loss in activity were analyzed by limited proteolysis to establish that activity loss is not due to misfolding (see Materials and Methods), and all mutants assayed display similar degradation profiles to the wild-type protein (**Figure 6**). In addition to reinforcing that the mutants that are analyzed are properly folded, it was of interest to determine the precise location of proteolysis. The three major C-terminal degradation products appear at roughly 50, 30, and 23 kDa, which provides a rough approximation of the cut site. To further narrow the location of the proteolysis sites, N-terminal Edman degradation sequencing of the digestion bands was performed.

Using the N-terminal sequencing data and the estimated molecular weights of the fragments, it was determined that the enzyme was being proteolyzed at G331/S332 ('1st cut site') and Y467/S468 ('2nd cut site'), yielding C-terminal fragments of 45.5 and 29.3 kDa, respectively. **Figure 10** shows the location of the analogous residues in a model of the *C.lari* PglB X-ray structure [4]. The first cut site appears at the C-terminal end of a large loop between two transmembrane domains (**Figure 10B, C**). In the *C. lari* structure, this loop appears as a highly disordered coil, with the N-terminal half of the loop unresolved [4]. In comparison, the second cut site appears in the soluble domain 4-5 residues C-terminal to the WWDxG motif. In the *C.lari* PglB structure, this site does not stand out as obviously accessible to solvent as it appears to be shielded by a nearby helix (**Figure 10D**). Interestingly, however, the analogous site in the structure of the *P. furiosus* Stt3 soluble domain shows this site immediately preceding a coiled region of high disorder which was not solved with the rest of the structure (**Figure 10D**, [2]).

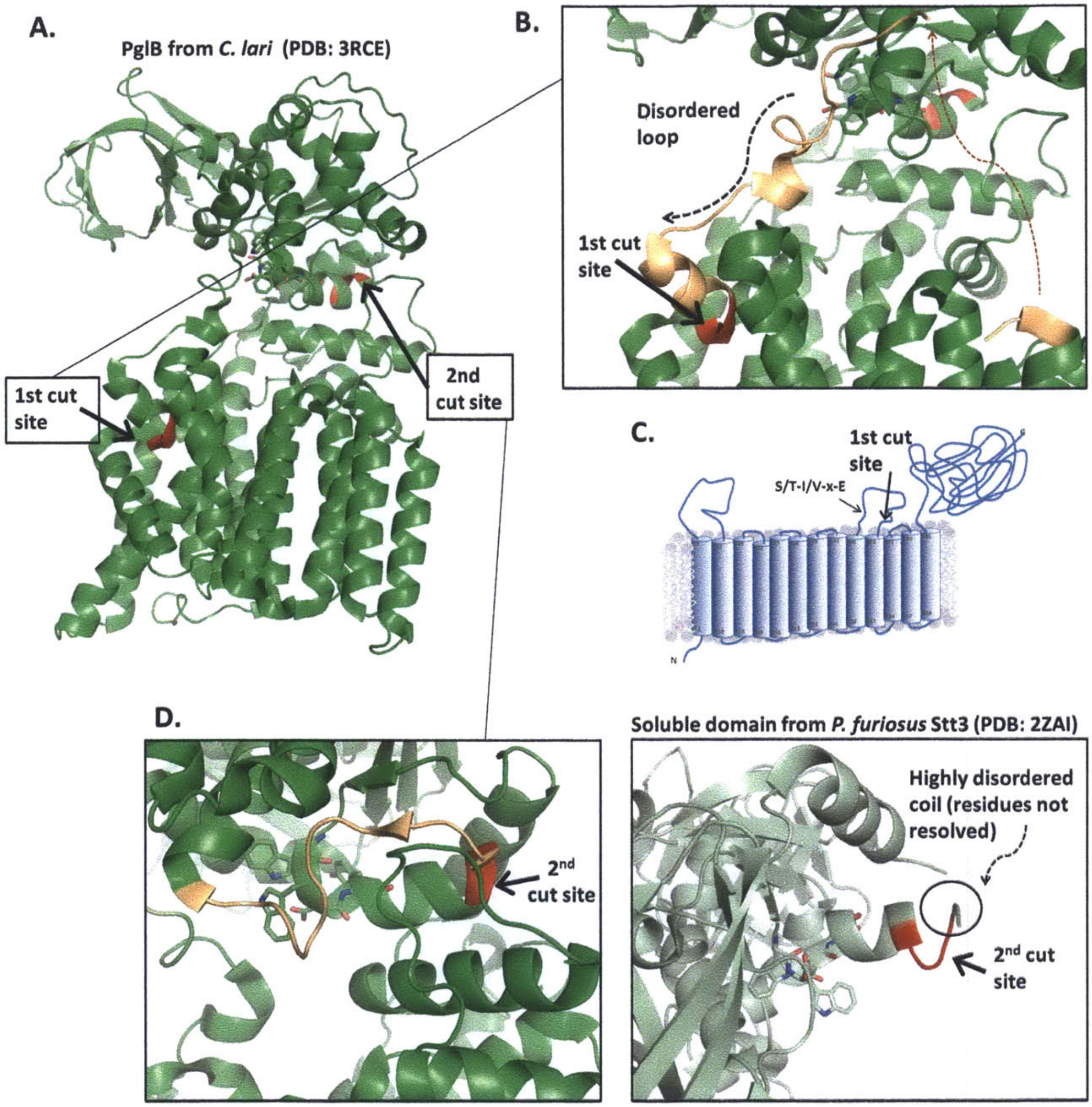

A.
PglB from C. lari (PDB: 3RCE)
1st cut site
2nd cut site
B.
Disordered loop
1st cut site
C.
1st cut site
S/T-I/V-x-E
N
D.
2nd cut site
Soluble domain from P. furiosus Stt3 (PDB: 2ZAI)
Highly disordered coil (residues not resolved)
2nd cut site

Figure 10: N-terminal sequencing of degradation products reveals sites of protease susceptibility.

Predicted cut sites are colored red and the WWDxG motif residues are shown as sticks for reference. *A.* Model of crystal structure of PglB from *C. lari* (PDB: 3RCE, [4]) highlighting predicted locations of digestion. *B.* The region containing the first cut site is expanded for clarity. The loop residues N-terminal to the 1st cut site are colored light orange. The dotted line denotes the unresolved residues in the loop. *C.* Cartoon topology model of PglB, highlighting the general location of the first cut site. *D. (left)* Region containing the second cut site in PglB. The loop C-terminal to the cut site is colored light orange. Shown to the right is the analogous position in a model of the crystal structure of the soluble domain of Stt3 from *P. furiosus* (PDB: 2ZAI, [2]).

The structure of the *C. lari* PglB posits this observation in a structural context and correlates susceptibility to proteolysis with the proposed flexibility and exposure of these regions. The first proteolysis site appears just after the exposed loop 5, designated as EL5, which is noted as highly disordered in the *C. lari* structure [4]. Interestingly, while both of these sites appear in coils adjacent to catalytically important sites, neither site appears prominently exposed in the crystal structure. Nonetheless, the dramatic preference for these proteolysis sites is shown clearly by the discrete banding pattern indicated by western blot. It is possible that these sites become exposed in an alternate conformation of PglB, which may provide insight into structural changes implicated in substrate binding and release. Indeed, it was proposed that this conformational change would involve movement of EL5 [4]. Also, the striking conservation of the WWDxG motif, combined with its proximity to a preferential proteolytic cut site, may indicate that this helix is involved in this conformational change, in addition to the proposed role in peptide/protein substrate binding [4].

The second cut site occurs just after the WWDxG motif, N-terminal to a coil connecting two helices. As noted above and in **Figure 10D**, this site appears relatively shielded from solvent in the *C. lari* crystal structure, while in the *P. furiosus* soluble-domain structure the predicted cut site borders a highly disordered, unresolved coil. Though highly speculative, this distinction of this region in the *C. lari* and *P. furiosus* structures, combined with apparent proteolytic activity at this site, may suggest involvement of this region in a conformational change. Undoubtedly, since the *P. furiosus* structure is solved as a soluble domain alone, the observed conformations are expected to be less representative of reality than the *C. lari* full-length structure because many of the native intramolecular interactions are unable to form. Nevertheless, the two clearly distinct physical states of this region in two structures, combined with the apparent proteolysis (and thus exposure to solvent) at this site, increases the conceivability that the disordered state represents a native conformation of the enzyme. Conversely, an alternative explanation is that this region only becomes accessible when specific interactions between the soluble and the transmembrane domains are disrupted. Once the first cut in the inter-transmembrane loop is made, a majority of the transmembrane helices are separated from the soluble domain, potentially disrupting these interactions and exposing the second site for proteases to act upon.

Conclusion

In recent years, it has become abundantly clear that the conserved loop segments between transmembrane domains often play a fundamental role in substrate recognition and/or catalysis [12-15]. However, it is particularly difficult to observe sequence conservation in these regions because motifs may be embedded within transmembrane regions that have diverged considerably, thereby complicating sequence alignments. Demonstrated here is a straightforward method for defining these buried regions of conservation applied to the complex integral-membrane OTase PglB. Topology predictions are generated for a list of homologous sequences using freely available software; a designed algorithm then parses each sequence by topological feature. These sequence segments are then more fruitful for detecting regions of conservation through sequence alignment. The systematic nature of the method allows for position-specific sequence analysis of a large number of divergent sequences of a given protein, which is crucial for determining the extent and significance of proposed motifs. These results expose the extraordinary level of conservation that exists in Stt3 homologs from bacteria through humans. Conservation over this evolutionary span implies that these regions play an essential role in the OTase activity. The biochemical data verifies that in PglB, many of these residues are essential for enzyme activity, and through limited proteolysis experiments it is shown that activity loss in mutated PglB is not caused by major structural changes. These combined data indicate a direct involvement of these motifs in protein function.

The publication of a medium-resolution structure of the *C. lari* PglB in 2011 indicates that these motifs are centrally involved in catalysis [4]. Thus, the structural data, combined with the independently acquired alignments and biochemical data, provide compelling evidence for the roles of these motifs in OTase catalysis. Much biochemical work remains to investigate the details of the catalysis. In this context, the limited proteolysis approach will be useful for ascertaining whether the E5 loop is flexible in the presence, as well as the absence, of each substrate, and thus may provide insight into the dynamics of these regions accompanying substrate binding, catalysis, and substrate release. Ultimately, similar studies on eukaryotic OTases will be essential for relating the prokaryotic OTase studies to OTase catalysis universally.

The structural studies now provide an excellent framework for developing new experiments to investigate N-linked glycosylation. Yet, the importance of quantitative, in vitro biochemical assays with defined quantities of pure substrates and precise measurements of enzymatic rates cannot be overstated. Based on the *C. lari* PglB structure a mechanism for nucleophilic activation of the asparagine nitrogen has been proposed [4]. A challenge in the future will be to acquire structural and biochemical data that are complementary and consistent with specific mechanistic proposals. Currently, the structural data provides valuable information on the residues that are likely to be involved in catalysis, however at the present structural resolution of 3.4 Å, it is not feasible to identify specific hydrogen bonding networks or, in fact even distinguish between the nitrogen and carbonyl oxygen of the nucleophilic asparagine amide. Furthermore, due to the crystallization conditions, the *C. lari* PglB structure was acquired at a pH of 9.4, where activity is very low [16]. Lastly, as the structure was solved in the presence of only the peptide substrate, the site of polyprenyldiphosphate-glycan binding, the order of binding and release of substrates and products, and nature of potential conformational changes, remain to be assessed. These data are required for the development of hypotheses concerning the mechanistic details of the reaction. The structural and biochemical data, when in agreement, will provide an important foundation for unraveling the details of this intricate cellular process.

Materials and Methods

Mutant production, expression, and purification

The QuikChange (Agilent) protocol was used to generate mutations in the PglB gene, and all mutant genes were sequenced to verify specificity and accuracy of mutagenesis. All mutants and wild type PglB were expressed in a pET24a(+) vector from Novagen. All constructs were transformed into BL21-CodonPlus-RIL cells (Stratagene) and grown in 1L of culture overnight. Cell cultures were harvested by centrifugation, washed in lysis buffer (50 mM HEPES, pH 7.5, 100 mM NaCl, 10% glycerol), re-pelleted, and frozen at -80 ^{0}C until needed. For the preparation of Cell Envelope Fractions (CEFs), cell pellets of equal weight were resuspended in 40 mL lysis buffer with the addition of 40 μg hen egg-white lysozyme (EMD Chemicals) and 40 μL EDTA-free Protease Inhibitor Cocktail III (CalBiochem). Cell suspensions were incubated at 4^{0}C with gentle rocking for one hour and then lysed using sonication with cooling on ice. Specifically, three one-minute sets of one-second pulses at 50% amplitude, with five-minute intervals between each set were employed. Lysates were centrifuged at 6,000 x g for 30 minutes to remove insoluble debris, and the supernatant was then centrifuged at 100,000 x g for 1 hour to pellet the CEF. The resulting supernatant was discarded and the CEF was homogenized in 35 mL of high salt buffer (50 mM HEPES, pH 7.5, 250 mM NaCl, 250 mM KCl, 20% glycerol), incubated at 4 ^{0}C with gentle rocking for 1 hour, then pelleted again. Washed CEFs were homogenized in 10 mL of lysis buffer and stored at -80^{0}C until further use.

For western blot analysis, 8 μL samples of CEF (diluted 1:40 in 50 mM HEPES, pH 7.5) were mixed with 2 μL of 5X SDS reducing buffer and boiled for 5 minutes. Five μL of this solution was added to each lane of a 4-15% Tris-HCl pre-stacked gradient gel (BioRad) . Gels were run at 150 volts for one hour. Protein was then transferred at 120 volts for 2 hours to a nitrocellulose membrane. Membranes were blocked for at least one hour in a solution of 5% BSA in TBS-T, then incubated for 1 hour in a solution of either: 1) T7 antibody conjugated to alkaline phosphatase diluted 1:10,000 in TBS-T (EMD4Biosciences), washed with three one-minute washes in TBS-T and a single one-minute wash in TBS and then developed using the 1-STEP BCIP/NBT (5-bromo-4-chloro-3-indolyl-phosphate/nitro blue tetrazolium) developing solution (Thermo Scientific) for roughly 10 seconds or until staining became visible, at which point the nitrocellulose membrane was washed thoroughly with deionized water, or 2) Tetra His antibody, BSA-free (Qiagen) at 0.1 μg/ml in TBS-T, followed by a three one-minute washes in TBS-T, one-hour incubation in Anti-mouse alkaline phosphatase-conjugated secondary antibody produced in goat (Sigma Aldrich), three one-minute washes in TBS-T and a single one-minute wash in TBS, and development with the alkaline-phosphatase substrate 1-STEP NBT/BCIP. For quantitative western blot analysis, the process was carried out similarly except with optimization of CEF dilution factor to achieve intensities in the range of those of the purified PglB.

Ultimately, CEFs were diluted 1:100 and compared against a set of pure PglB internal standards, which had been quantified by measuring ultraviolet absorption at 280 nm. After staining, the nitrocellulose blots were allowed to dry for 10 minutes and then were immediately scanned at 1200 dpi and analyzed using Adobe Photoshop densitometry software. To maximize reliability, sample preparation and western blot analysis were repeated in triplicate, the data were combined and the average relative quantities used. A representative western blot and standard graph are shown in **Figure 8**.

PglB Activity Assays

A survey of activity measurements for PglB and all PglB mutants was carried out as described previously [17]. Briefly, 10 µL of DMSO was added to a tube containing 6 pmol of dried radiolabeled undecaprenyl-PP-Bac-[3H]GalNAc of specific activity 15 µCi/nmol. The tube was then vortexed and sonicated (water bath) to resuspend the substrate. Then, 100 µL of 2X assay buffer (280 mM sucrose, 2.4% Triton X-100 (v/v), and 100 mM HEPES at pH 7.5), 2 µL of 1 M MnCl2, and 5 µL of PglB CEF were added and the volume brought to a total of 190 µL with water. Reactions were initiated by the addition of 10 µL of 1 mM peptide substrate (Ac-DQNAT-p-NF-NH2; where p-NF is para-nitro-phenylalanine) dissolved in DMSO [18]. Aliquots of the reaction mixture were removed at specified time intervals and quenched in 1 mL of 3:2 chloroform/methanol + 200 µL of 4 mM $MgCl_2$. The aqueous layer was extracted, and the organic layer was washed twice with 300 µL of pure solvent upper phase (3% chloroform, 49% methanol, and 48% water with 100 mM KCl). The aqueous extracts were combined and mixed with 5 mL of EcoLite scintillation fluid (MP Biomedicals), the organic phase was mixed with 5 mL of OPTI-FLUOR scintillation fluid (Perkin Elmer), and all fractions were subjected to scintillation counting. All assays were carried out in duplicate or triplicate.

Synthesis of lipid-linked sugar substrate at three specific activities

For rate comparison assays at varying concentrations of sugar substrate, radiolabeled undecaprenyldiphosphate-disaccharide (Und-PP-Bac-GalNAc) was synthesized at specific activities of 0.15, 1.5, and 15 µCi/nmol, according to procedures described previously [17]. Each PglB assay performed (using one aliquot of Und-PP-Bac-GalNAc) contained the same level of tritium, but depending on the specific activity a single assay would contain 0.01, 0.1, or 1.0 µM sugar substrate, which represents roughly 0.01x, 0.1x, and 1x K_M (app) of the Und-PP-Bac-GalNAc substrate in the PglB reaction [19]. The three different specific activities of the sugar substrates were desirable because the reactions would contain varied concentrations of total substrate, but the handling of very high or very low levels of radioactivity would be avoided.

The starting substrates for synthesis include the UDP-GalNAc (tritium-labeled), UDP-Bac, Und-OH, and ATP. The enzymes used include the SM kinase [20], PglC (purified to CEF), and PglA. It was determined that roughly 20 assays of lipid-linked disaccharide would be carried out at each of the three log units of varied substrate concentration: 20 reactions at K_M(app) (1 μM), 20 reactions at 0.1x K_M(app) (0.1 μM), and 20 reactions at 0.01x K_M(app) (0.01 μM). The manufacturer-supplied specific activity of the pure radiolabeled stock of UDP-GalNAc was 15 Ci/mmol, or 15 μCi/nmol, with 1 μCi (Curie) = 2.22×10^6 DPM. Since each reaction contains 200 μL total volume, and 20 reactions of each concentration of 1, 0.1, and 0.01 μM substrate was required, it can be calculated that for the 1 μM batch of reactions 40 nmol radiolabeled product is required (200 pmol/reaction vial), for the 0.1 μM batch, 4 nmol substrate is required, and for 0.01 μM batch a total of 0.4 nmol would be required.

For the lowest concentration substrate (0.01x K_M(app) product/tube), the radiolabeled UDP-GalNAc was not diluted with any unlabeled UDP-GalNAc. Thus, 0.4 nmol of UDP-GalNAc (specific activity of 15 μCi/nmol, containing 6 μCi) was collected for each of the three reactions. For the 0.1x K_M(app) reactions, a volume containing 3.6 nmol of a known concentration of unlabeled UDP-GalNAc was added to the hot solution, rendering its specific activity 1.5 μCi/nmol. The same was done with the 1x K_M(app) UDP-GalNAc solution, except this time a volume containing 39.6 nmol unlabeled UDP-GalNAc was added in order to give a specific activity of 0.15 μCi/nmol. Thus, each reaction contained the same level of radiolabeled

UDP-GalNAc, the total UDP-GalNAc concentration was varied. The stoichiometric levels of other reaction components were adjusted accordingly in order to maximize turnover of the UDP-GalNAc solution.

The final product was purified over HPLC. The reaction aliquots for each specific activity batch contained roughly 66,600 DPM, which represented 2 pmol, 20 pmol, or 200 pmol labeled substrate for the reactions performed with 15, 1.5, and 0.15 μCi/nmol specific activities, respectively. When resuspended to the reaction concentration of 200 μL, this provided a concentration of 0.01 μM, 0.1 μM, and 1 μM Und-PP-Bac-GalNac, respectively.

A summary of the specific reaction plan is as follows:

Km for disaccharide ~ 1 μM
Specific activity of undiluted hot GalNAc: 15 μCi/nmol

1. 1X hot UDP-GalNAc (for 10 nM concentration in assay with 100K DPM)
-50 nmol Und-P
-20 μL 5 mM UDP-Bac
-20 μL .025 μM UDP-GalNAc (specific activity 15 μCi/nmol)

2. 10X dilution of hot UDP-GalNAc (for 10X increase in concentration in assay (0.1 μM) with 100K DPM)
-50 nmol Und-P
-20 μL 5 mM UDP-Bac
-20 μL .25 μM UDP-GalNAc (specific activity 1.5 μCi/nmol)

3. 100X dilution of hot UDP-GalNAc (for 100X increase in concentration in assay(1 μM) with 100K DPM)
-50 nmol Und-P
-20 μL 5 mM UDP-Bac
-20 μL 2.5 μM UDP-GalNAc (specific activity 150 nCi/nmol)

Limited Proteolysis

Digestion profiles of wild-type PglB using trypsin, α-chymotrypsin, and proteinase K were compared, with proteinase K showing the clearest production of discrete proteolytic fragments. The incubation time, temperature, and the ratio of PglB to protease were optimized, such that the time-dependent production of discrete proteolytic fragments could be clearly viewed by applying His-tagged western blot analysis. The His-tag antibody was used because the location of the His-tag at the C-terminus of the soluble domain resulted in a more readily identifiable degradation profile. Proteolytic assays were performed at room temperature on CEF fractions that were diluted 1:40 in 50 mM HEPES, pH 7.5. Two μL of 0.1 mg/mL of Proteinase K (in 4 mM $MgCl_2$) were added to 160 μL of diluted CEF. Aliquots of 20 μL were quenched into 3 μL of 100 mM PMSF (phenylmethylsulfonyl chloride) in ethanol, at time points of 0, 5, 10, 30, 60, and 180 minutes.

References

1. Igura M, Kohda D (2011) Selective control of oligosaccharide transfer efficiency for the N-glycosylation sequon by a point mutation in oligosaccharyltransferase. J Biol Chem 286: 13255-13260.
2. Igura M, Maita N, Kamishikiryo J, Yamada M, Obita T, et al. (2008) Structure-guided identification of a new catalytic motif of oligosaccharyltransferase. EMBO J 27: 234-243.
3. Maita N, Nyirenda J, Igura M, Kamishikiryo J, Kohda D (2010) Comparative structural biology of eubacterial and archaeal oligosaccharyltransferases. J Biol Chem 285: 4941-4950.
4. Lizak C, Gerber S, Numao S, Aebi M, Locher KP (2011) X-ray structure of a bacterial oligosaccharyltransferase. Nature 474: 350-355.
5. Liu J, Mushegian A (2003) Three monophyletic superfamilies account for the majority of the known glycosyltransferases. Protein Sci 12: 1418-1431.
6. Maeda Y, Watanabe R, Harris CL, Hong Y, Ohishi K, et al. (2001) PIG-M transfers the first mannose to glycosylphosphatidylinositol on the lumenal side of the ER. EMBO J 20: 250-261.
7. Fontana A, de Laureto PP, Spolaore B, Frare E, Picotti P, et al. (2004) Probing protein structure by limited proteolysis. Acta Biochim Pol 51: 299-321.
8. Leandro J, Leandro P, Flatmark T (2011) Heterotetrameric forms of human phenylalanine hydroxylase: co-expression of wild-type and mutant forms in a bicistronic system. Biochim Biophys Acta 1812: 602-612.
9. Spolaore B, Bermejo R, Zambonin M, Fontana A (2001) Protein interactions leading to conformational changes monitored by limited proteolysis: apo form and fragments of horse cytochrome c. Biochemistry 40: 9460-9468.
10. Spolaore B, Polverino de Laureto P, Zambonin M, Fontana A (2004) Limited proteolysis of human growth hormone at low pH: isolation, characterization, and complementation of the two biologically relevant fragments 1-44 and 45-191. Biochemistry 43: 6576-6586.
11. Vita C, Dalzoppo D, Fontana A (1985) Limited proteolysis of thermolysin by subtilisin: isolation and characterization of a partially active enzyme derivative. Biochemistry 24: 1798-1806.
12. Ashkenazi A, Viard M, Wexler-Cohen Y, Blumenthal R, Shai Y (2011) Viral envelope protein folding and membrane hemifusion are enhanced by the conserved loop region of HIV-1 gp41. FASEB J 25: 2156-2166.
13. Estrada-Mondragon A, Reyes-Ruiz JM, Martinez-Torres A, Miledi R (2010) Structure-function study of the fourth transmembrane segment of the GABArho1 receptor. Proc Natl Acad Sci U S A 107: 17780-17784.
14. McCracken LM, McCracken ML, Gong DH, Trudell JR, Harris RA (2010) Linking of Glycine Receptor Transmembrane Segments Three and Four Allows Assignment of Intrasubunit-Facing Residues. ACS Chem Neurosci 1: 482.
15. Ponsaerts R, De Vuyst E, Retamal M, D'Hondt C, Vermeire D, et al. (2010) Intramolecular loop/tail interactions are essential for connexin 43-hemichannel activity. FASEB J 24: 4378-4395.
16. Sharma CB, Lehle L, Tanner W (1981) N-Glycosylation of yeast proteins. Characterization of the solubilized oligosaccharyl transferase. Eur J Biochem 116: 101-108.

17. Glover KJ, Weerapana E, Numao S, Imperiali B (2005) Chemoenzymatic synthesis of glycopeptides with PglB, a bacterial oligosaccharyl transferase from Campylobacter jejuni. Chem Biol 12: 1311-1315.
18. Chen MM, Glover KJ, Imperiali B (2007) From peptide to protein: comparative analysis of the substrate specificity of N-linked glycosylation in C. jejuni. Biochemistry 46: 5579-5585.
19. Chen MM, Weerapana E, Ciepichal E, Stupak J, Reid CW, et al. (2007) Polyisoprenol specificity in the Campylobacter jejuni N-linked glycosylation pathway. Biochemistry 46: 14342-14348.
20. Hartley MD, Larkin A, Imperiali B (2008) Chemoenzymatic synthesis of polyprenyl phosphates. Bioorg Med Chem 16: 5149-5156.

CHAPTER 5:

DESIGN AND DEVELOPMENT OF LRET-BASED SYSTEM TO INVESTIGATE DYNAMICS OF PGLB CATALYSIS

Acknowledgments

Special thanks to Dr. Andrew Krueger and Michelle Chang of the Imperiali lab, both of whom contributed to data presented in this chapter. Dr. Krueger analyzed effects of various phosphorimeter parameters on the luminescence measurements (**Figure 12**) and Michelle Chang contributed luminescent spectra from titration of LBT-AglB with Tb^{3+} (**Figure 14**). Michelle Chang will be continuing investigations of LBT-PglB and LBT-AglB using the system described in this chapter.

Summary

The following chapter describes the design and assembly of components required for obtaining a system to investigate key features of the OTase enzymatic reaction; these key features include order of substrate binding, the sites on the enzyme involved in substrate-binding and release, and the nature of potential conformational changes occurring upon substrate binding and catalysis. Toward these goals, the improved levels of stable PglB are used as a foundation for developing a system to measure informative distances using LRET. The strategy devised includes expression of PglB with a genetically encoded lanthanide-binding tag (LBT). Cysteine-containing peptide substrates were synthesized, purified, and labeled, which will allow intermolecular LRET measurements between the LBT-Tb^{3+} complex and the peptide substrate. In addition, the site-specific insertion of cysteine allows the expressed enzyme to be labeled with a thiol-specific dye that may act as an LRET acceptor, and the complex of LBT-Tb^{3+} functions as a partner donor. As a result of the following efforts, proof of concept has now shown with the measurement of LRET between LBT-PglB and a acceptor-labeled substrate peptide. The system is now being applied to an archaeal OTase, AglB, and additional comparative studies are expected to continue under the direction of a fellow graduate student, Michelle Chang. In summary, a system has been successfully designed and developed to measure dynamic changes and interactions involved in the OTase reaction, which are unattainable using X-ray crystallography.

Introduction

In the past several years, remarkable advances have been made toward better understanding the reaction catalyzed by OTases. Importantly, significant sequence homology has been revealed in OTases throughout all evolutionary domains of life; as a result, the single-subunit OTase PglB from *C. jejuni* has been firmly established as a prototypical model of the conserved enzymatic reaction. Structural models of the enzyme have provided a wealth of knowledge on the manner in which the conserved regions of the protein join to form a predicted active site.

However, fundamental aspects of the OTase reaction remain to be determined. The structural data available for PglB shows a single conformation of the enzyme at a resolution of 3.5Å, an alkaline pH, and in a detergent micelle, which differs from the native state in the membrane and at neutral or acidic pH. As with any crystal structure, it is difficult to discern from the structural data alone whether the enzyme has adopted a native, active conformation when crystallizing. The substrate-binding requirements are unknown, and accordingly it is unclear whether the substrates bind in a specific order, whether binding order is random, and which chemical groups on the substrates function in binding and catalysis. Also unknown is the location on the enzyme that the glycan substrate binds or whether the peptide bound in the crystal structure is representative of the native state of binding. In the case that the saccharide substrate binds before the peptide substrate, saturation of the enzyme with the peptide substrate alone may facilitate a non-native conformation. While key regions of the enzyme have been defined, the function of each of these regions remains to be determined; it is not known whether the critical residues in these regions are involved in substrate-binding, catalysis, structural organization, or a combination. Other major unresolved factors include the role of the required divalent cation and the nature or existence of conformational changes.

The recent availability of a stable and purified PglB enzyme in reasonable quantities and determination of structural regions composing the OTase active site provide new tools to guide experimental design in studying these factors [1-3]. A series of steps were carried out to develop a system to examine substrate binding and the nature of potential conformational changes involved in PglB catalysis. A luminescent spectroscopic approach was preferred because it allows for sensitive measurements and low perturbation of the active enzyme *in vitro*, which are important conditions for studying dynamic enzymatic processes.

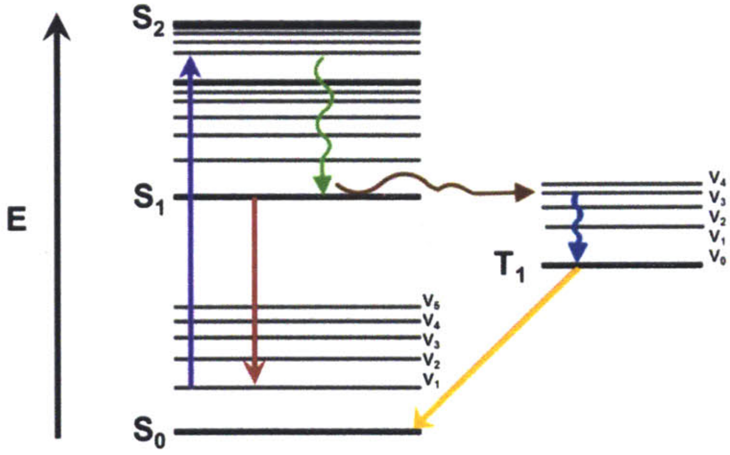


Figure 1: Jablonski diagram representing fluorescent and phosphorescent emission. On the y-axis energy (E) is representative. S_0 through S_2 represent electronic states, T_1 is a triplet state. The purple arrow represents absorbance, the red arrow represents fluorescent decay and the orange arrow represents phosphorescent decay. The squiggly arrows represent non-radiative processes: the green arrow is internal conversion, the brown arrow is intersystem crossing, and the blue arrow is vibrational relaxation.

Specifically, luminescence resonance energy transfer (LRET) will be used to investigate the aforementioned characteristics of the PglB enzymatic reaction. As opposed to fluorescence transfer (FRET), LRET includes emission via phosphorescence as well as fluorescence (hence the all-encompassing term 'luminescence'). The term LRET is often used in the context of lanthanide emission [4]. Lanthanides ions, when excited at the appropriate wavelength, will emit energy via phosphorescent emission. As represented by a Jablonski diagram in **Figure 1**, absorbance of energy results in excitation of an electron form a ground state to an excited state. In addition to non-radiative decay mechanisms (internal conversion and vibrational relaxation), emission of light via luminescence accounts for the loss of energy of the electron as it returns to the ground state. When the decay occurs from the a singlet state to the singlet ground state, the emission of light is defined as fluorescence; when the energy first undergoes intersystem crossing to a triplet state and decays from a triplet state to the singlet ground state, the emission of light is defined as phosphorescence.

There are several differences between LRET and FRET. Lanthanides do not inherently result in improved brightness of the emission relative to that of fluorophores, as quantum yields and extinction are unrelated to the form of emission. However, phosphorescent molecules or atoms (such as lanthanides) have longer Stokes' shifts and lifetimes relative to fluorophores. These qualities result in a dramatically reduced background signal, because the phosphorescent emission can be measured after a time delay during which the shorter-lived auto-fluorescence has already decayed [5]. Furthermore, unlike the polarized mission from fluorogenic molecules, emission from lanthanide ions is radial. The dipole orientation factor κ depends on the relative orientation of the two fluorophore; emission from fluorescent molecules is polarized, which affects the level of energy transfer and thus the Förster radius R_0 (**Equation 1**). Generally, the dipole orientation factor κ is estimated as 2/3 based on an assumption that both molecules are freely rotating in solution causing isotropic emission. Often one or both fluorophores is fixed and is not free to rotate, which makes the assumption of isotropic emission invalid and thus introduces error into calculation of R_0. Thus, the radial (unpolarized) emission from lanthanide ions results in a reduced error in the determination of R_0. In calculation of R_0, Q_0 is the quantum yield of the donor in the absence of acceptor, J is the spectral overlap integral, n is the refractive index of the medium, and N_A is Avogadro's number.

$$R_0^{\,6} = \frac{9\, Q_0\, (ln10)\, \kappa^2 J}{128\, \pi^5 n^4 N_A} \qquad \textbf{(1)}$$

The reduction in background and the improved accuracy of R_0 effectively improves the accuracy of LRET measurements relative to those of FRET. Reduced background gives an improved normalized signal and, because the R_0-value is included in calculation of distance measurements (r-values) from transfer efficiency (E), LRET effectively improves the accuracy of the distances calculated (**Equation 2**). This improved accuracy lends the calculated values to more plausible biological interpretation and results in better agreement with crystallographic measurements.

$$E = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6} \qquad \textbf{(2)}$$

Resonant energy transfer measurements are made using the enzyme in solution. Enzyme activity can be monitored which eliminates any uncertainty as to whether the obtained data is representative of the active enzyme. This gives the resonant energy transfer technique an advantage over X-ray crystallography. LRET is capable of measuring intra- and inter-protein interactions and distances, thus providing structural information that can support, clarify, and supplement crystallographic data. Significantly, LRET is capable of providing time-resolved information about the dynamics of the enzymatic reaction (e.g. conformational changes upon substrate binding) in a single experiment [6,7], which cannot be obtained from crystallographic data.

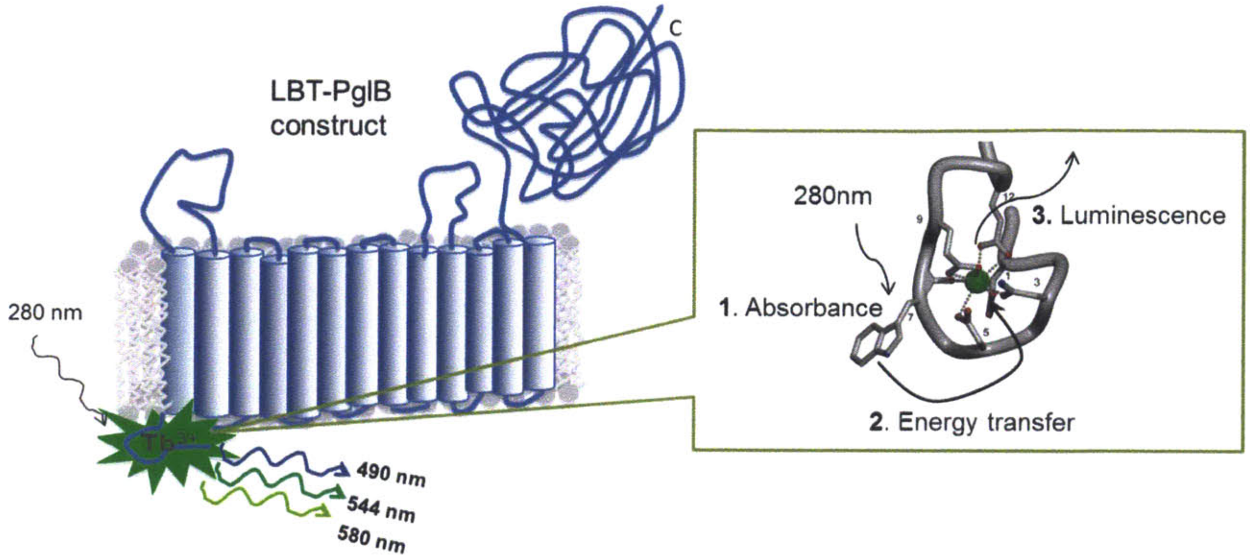


Figure 2: Sensitization of Tb^{3+} by the LBT.
The topology cartoon of PglB shows the LBT bound to the lanthanide ion, Tb^{3+}. The complex is irradiated with UV light, and after a millisecond-lifetime phosphorescent light is emitted. Expansion of the LBT shows the amino acids chelating and sensitizing Tb^{3+}. The tryptophan, when excited by incoming UV light, transfers energy of its excited electronic state to the nearby Tb^{3+}, which then emits light at specific discrete wavelengths. Importantly, the LBT residues protect Tb^{3+} from exposure to water molecules, which quench the luminescence. LBT figure taken from [8].

In order for lanthanide ions in aqueous solution to absorb energy from light, a sensitizer is required; the sensitizer is a molecule that chelates the ion and first absorbs the applied light, ultimately transferring the energy to the ion. The lanthanide-binding tag (LBT) consists of 20 amino acids and binds and sensitizes lanthanide ions (such as Terbium ion (Tb^{3+})), forming a luminescent composite (**Figure 2**) [9,10]. The LBT was chosen to act as the donor in the following LRET experiments for several reasons, in addition to the favorable properties of lanthanide emission discussed above. First, a major challenge for LRET-type experiments is the site-specific labeling of two locations on the protein with different fluorophores. The LBT is genetically encoded, which eliminates half of that challenge. Second, LBT-lanthanide complexes have very narrow and discrete emission bands, enhancing specificity of LRET. Third, LRET between LBT-lanthanide emission and known acceptor fluorophores exhibits a large transfer distance, which ensures detection of an intermembrane region even if it is located at a relatively large distance from the LBT.

In order to measure intermolecular distances between PglB and the peptide substrate, which can provide information on the location and order of substrate binding, a peptide containing the glycosylation consensus sequence was synthesized with a cysteine residue included. The cysteine was then site-specifically labeled using a thiol-reactive LRET acceptor dye. Similarly, cysteine mutations or insertions can be made in locations throughout the enzyme that are predicted to be informative. If the mutants express and show WT-level activity, then thiol-specific labeling can again be used to create a functional donor-acceptor pair. In this way, both inter- and intra-molecular LRET measurements can be made (**Figure 3**).

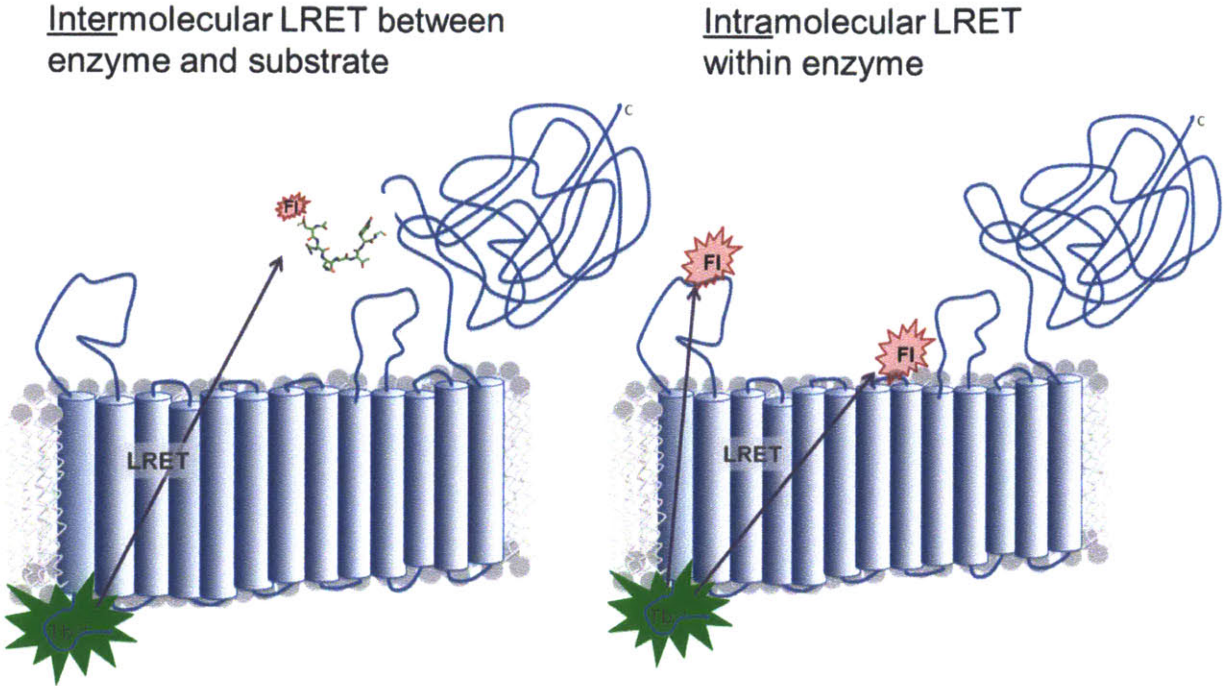

Intermolecular LRET between enzyme and substrate
Intramolecular LRET within enzyme
c
Fl
LRET
Tb
c
Fl
Fl
LRET
Tb

Figure 3: Simple illustration of the planned LRET experiments.
The LBT, genetically encoded on the N-terminus of the PglB gene, acts as a donor when lanthanide ion (Tb^{3+}) is added to the solution. The unpolarized lanthanide emission can engage in LRET with a dye that is specifically labeled via the thiol group on a unique cysteine. For intermolecular LRET measurements, the cysteine is included in the substrate peptide sequence. For intramolecular LRET measurements, cysteine residues are individually mutated into specific locations in the PglB sequence.

Results

Expression of LBT-PglB construct

The first step in developing the LRET system described above involved constructing and expressing a well-behaved LBT-PglB fusion construct. This task was expected to be relatively simple and rapid for several reasons. As seen in Chapter 2 of this thesis, much time had been spent optimizing the expression levels of PglB and the resulting yields were comfortably within a workable range. The LBT was developed and characterized extensively by previous Imperiali lab members who had successfully expressed LBT fusions [8-13]. Additionally, the LBT is only 20 amino acids in length (shown below), making it seem unlikely that fusion to PglB would alter previously established properties of PglB.

LBT amino-acid sequence: **GPGYIDTNNDGWYEGDELLA**

Therefore, it was surprising when LBT-PglB initially showed very poor expression. Despite the small size of the LBT, encoding the LBT sequence at the N-terminus of PglB led to dramatically reduced yields relative to T7-PglB (**Figure 4A**). The construct that was initially tested encoded LBT-PglB-His10 in pET24a(+) and is denoted LBT-PglB-1 (**Table 1**). Troubleshooting started with changing parameters that are known to cause general problems in protein expression. The vectors were re-sequenced to ensure no mutations were at fault. Vectors were re-transformed into a fresh batch of competent cells. Media components involved in making autoinduction media (see Chapter 2) were replaced along with antibiotic solutions. Even a new pET24a(+) carrier vector was obtained and the LBT-PglB sequence was moved to the new vector, but to no avail. All efforts resulted in the same deficient expression of LBT-PglB.

Construct name	**Components & restriction sites**[1]	**Notes**
LBT-PglB-1	NdeI-ATG-LBT-BamHI-PglB-His10-STOP-XhoI	Poor expression
LBT-Ubiq	NdeI-ATG-His6-LBT-BamHI-ubiquitin-STOP-XhoI	
TEV-LBT-PglB	NdeI-ATG-TEV-LBT-BamHI-PglB-His10-STOP-XhoI	
LBT-PglB	NdeI-ATG-LBT(optimized)-BamHI-PglB-His10-STOP-XhoI	Best expression

1. Unless otherwise noted, all constructs are made within a pET24a(+) vector.

Table 1: Summary of major vectors discussed in this chapter.

Upon reviewing earlier studies, it was found that a test expression carried out a year previous had shown reasonable expression of the LBT-PglB; however, this construct had contained a TEV sequence after the start codon but before the LBT sequence due to the cloning strategy. This TEV sequence had since been removed, as it contained no ostensible purpose. Despite a lack of comprehension of any positive impact the resulting from a preceding TEV sequence, few other obvious options remained. (Discussed below, measurements had been made that indicated that the N-terminus was preferable to the C-terminus for LRET measurements, ruling out this option.) Therefore, the TEV-LBT-PglB construct was re-made (**Figure 4C, Table 1**). Surprisingly, the insertion of the TEV sequence before the LBT sequence led to improved expression levels (**Figure 4B**).

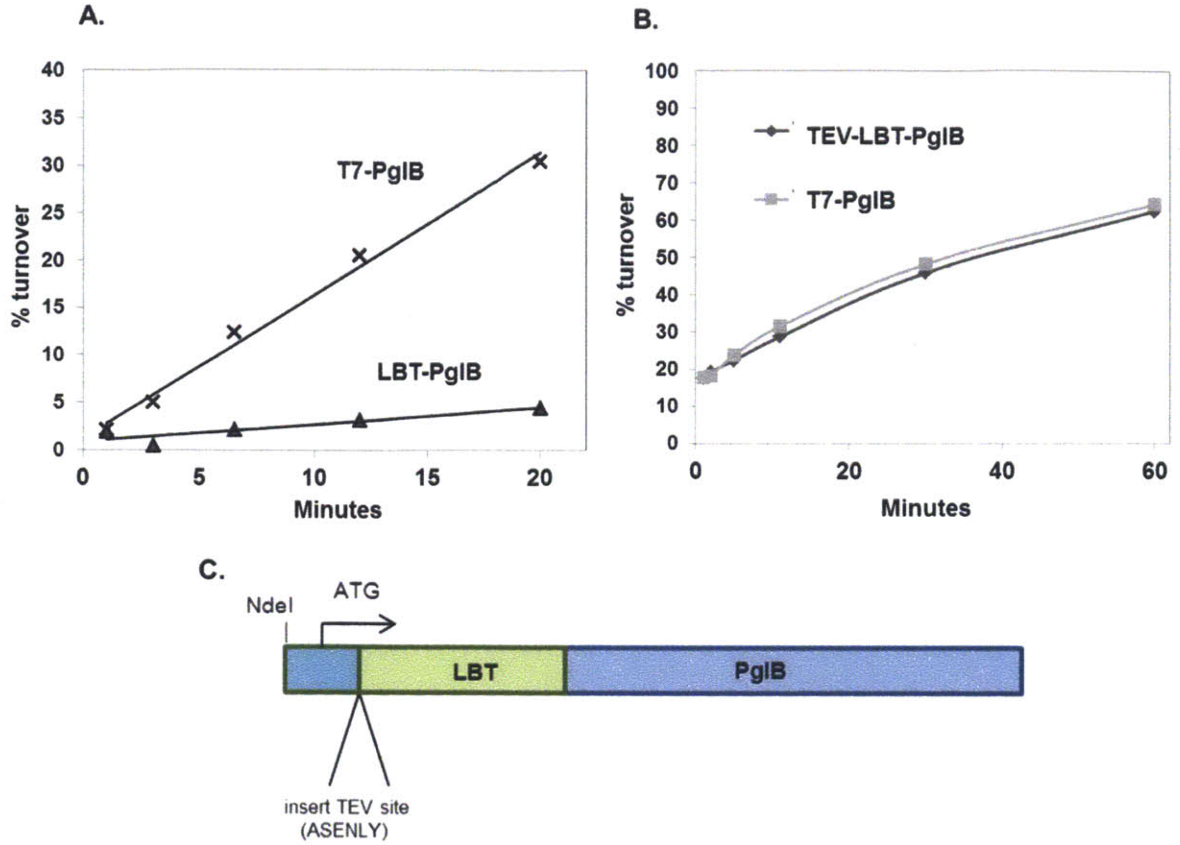


Figure 4: Poor initial expression of LBT-PglB and improvements with N-terminal insertion of TEV sequence.
Activity results of cleared lysates for LBT-PglB (*A*) or TEV-LBT-PglB (*B*) and T7-PglB after growing, lysing, and preparing both samples in parallel. Discrepancy in activity levels in *A* and *B* result from preparation and weight of initial cell pellets used. In each case, T7-PglB was grown and prepared alongside the LBT-PglB or TEV-LBT-PglB construct such that each comparison internally represents a comparison of the same fraction of cell pellet. (*C*) Diagram showing the location of the start codon relative to the LBT-PglB construct and the location of the inserted TEV sequence.

Literature searches led to the development of several hypotheses to explain the observed improvement in expression at the mere insertion of a 6-amino acid TEV sequence (ASENLY). Several studies by the Isaksson group in the Netherlands provided anecdotal and general evidence that several codons immediately following the start codon in a gene significantly affected expression levels in *E. coli*. An early study showed that a simple change of the +3 codon from GAG to GAA (both coding for glutamic acid) in the *glnS* gene results in a four-fold expression improvement [14]. The observation was further investigated, showing that high-expressing genes show a statistically significant overrepresentation of certain codons at early sites after the start codon and mutational analysis implicated ribosome binding and translation initiation as mediators of this effect [14-18]. In *E. coli*, several codons immediately 3' to the start codon (+2, +3, +5, +7) are significantly more likely to be enriched in adenines than guanines [15]. Synthetic constructs with varied codon composition further showed that increasing the adenine content led to improvements in expression of up to 20-fold [18].

Construct	First five codons +1 +2 +3 +4 +5
TEV - LBT - PglB	ATG - GCT - TCT - GAA - AAC
LBT - PglB-1	ATG - GGT - CCG - GGC -TTC
T7 - PglB	ATG - GCT - AGC - ATG - ACT

Table 2: Comparison of nucleotide content of the four codons following the start codon for several constructs discussed.

Table 2 shows the first five codons (in which the start codon is +1) for T7-PglB, LBT-PglB-1, and TEV-LBT-PglB. Comparison of the codon composition of this segment highlights the over-abundance of guanines and lack of adenines following the N-terminal start codon in LBT-PglB-1. In contrast, T7-PglB and TEV-LBT-PglB can be seen to have adenines present and significantly fewer guanines, and no sequential guanines. This observation was encouraging because it indicated that the reason for the low levels of LBT-PglB expression may be due to codon composition in the early codons 3' to the start codon.

Several other lines of reasoning can be used to support this hypothesis. The relatively low publicity received from the papers published by the Isaksson lab can be understood in light of the fact that determinants of expression level are accepted as being diverse, unpredictable, and poorly understood. Heterologous expression of proteins in *E. coli* is accepted as an empirical endeavor. Certain factors may generally influence the expression level of a protein one way or another, such as localization, complex folding, poor solubility, and others; however, these principles are guidelines and there exist many unknown reasons why some proteins are toxic or inefficiently expressed in *E. coli* [19-21]. Therefore, empirical screens of conditions or fusion tags is often the immediate course of action when a protein shows poor heterologous expression.

Fusion tags are frequently used in expressing and purifying proteins in *E. coli* and are often themselves foreign to *E. coli* [22]. Importantly, these tags are commonly developed empirically; therefore, those that are frequently used have been observed to express well in *E. coli* and likely have unrecognized qualities that lead to efficient expression such as high adenine-content in the early codons. One example to consider is the His-tag, which is very frequently used in protein expression and purification and often is placed at the N-terminus of the gene. The two histidine codons, CAC and CAT, are both notably lacking in guanines and each contain an adenosine. Therefore, initiating synthesis with a His-tag would likely shield codon effects to some degree if gene has guanine-rich early codons. In contrast to the empirical success of certain fusion tags (which likely are unknowingly selected for properties such as a beneficial early codon composition), the LBT sequence is based on an amino acid sequence synthesized using SPPS and evolved to optimally bind lanthanides [8]. Therefore, expression of the sequence did not contribute to selection. Expression of the LBT-protein fusions in the past included a His6-tag at the N-terminus and the proteins have been small and soluble and otherwise well-expressed, so early codons, if unfavorable for translation initiation, would not be likely to greatly influence expression [9,11,12].

A final compelling piece of evidence was discovered in a very recent Nature publication from the Weissman lab at the University of California in San Francisco (UCSF) [23]. The study sought to investigate the reasons behind the non-uniform speed at which ribosomes translate a linear RNA substrate. It has been frequently reasoned that relative abundance of tRNAs for each codons causes ribosomes to stall at certain regions of the transcript that have a high concentration of the 'rare codons'. The authors did a bacterial genome-wide analysis using ribosome profiling, or deep sequencing of ribosome-protected mRNA fragments and showed that, surprisingly, codon bias was not associated with ribosome occupation of certain sites. The pauses in translation resulted, rather, when the ribosome encountered Shine-Dalgarno-(SD)-like features within coding sequences [23].

The Shine-Dalgarno (SD)-sequence plays a primary role in translation initiation in bacteria. The sequence is a purine-rich region located roughly 9 nucleotides upstream of the start codon on the mRNA transcript. A sequence at the 3'-end of the 16S rRNA base pairs with part or all of the complementary the SD-sequence. (In *E. coli*, the complementary rRNA sequence is ACCUCCUUA.) The SD-sequence adheres loosely to the motif UAAGGAGGU, but anywhere from 3 to 9 nucleotides base pair with the 16S rRNA 3'terminal sequence. The degree of complementarity affects initiation efficiency, and this factor is used as one way to regulate levels of expression of individual proteins in bacteria. Importantly, if the length of complementarity is very long and precise, ribosomes will bind to mRNA too tightly, disturbing the migration of the ribosome.

```
Pre-optimized
          10        20        30        40        50        60        70
           |         |         |         |         |         |         |
1 ATGGGGCCTGGGTTCATCGATACCAATAACGATGGTTGGATCGAAGGGGATGAACTGCTGGCGGGATCCATG 72
  M  G  P  G  F  I  D  T  N  N  D  G  W  I  E  G  D  E  L  L  A  G  S  M

Optimized
          10        20        30        40        50        60        70
           |         |         |         |         |         |         |
1 ATGAAACTGATTTTCATCGATACCAATAACGATGGCTGGATCGAAGGCGATGAACTGCTGGCCATTGGATCCATG 75
  M  K  L  I  F  I  D  T  N  N  D  G  W  I  E  G  D  E  L  L  A  I  G  S  M
```

Figure 5: LBT before and after optimization of nucleotide content.
The corresponding amino-acid sequence is shown below each of the nucleotide sequences.

Thus, the association of ribosome stalling with appearances of SD-like sequences is not difficult to understand [23]. Generally, the ribosome translates the RNA transcript at steady-state, but upon encountering and binding to a site of significant complementarity, a counter-force is introduced, and the binding energy of the base pairs must be overcome for the ribosome to continue moving along the transcript. The phenomenon is not frame-specific, as the ribosome base-pairing occurs upstream of the start codon (thus upstream of the ribosomal 'A-site'. Additional evidence that SD-like sequences caused ribosome stalling was provided by the ability to change the translation speed of a gene by introducing or eliminating SD-like clusters without changing the encoded amino-acid sequence. Importantly, it was observed and explained that SD-like sequences that appeared closest to the N-terminus resulted in the most suppression of translation [23].

With the above study in mind, the LBT nucleotide sequence was scanned for potentially troublesome regions. In fact, it appeared that there were multiple regions containing clusters of purines (**Figure 5**). It was determined that the nucleotide sequence of the LBT had not been codon optimized when the gene was initially synthesized. The LBT has been fused to multiple protein constructs (generally preceded by a His-tag) and is frequently used in the Imperiali lab. Therefore, it seemed beneficial to optimize the nucleotide composition of the LBT. Optimizing the sequence and re-expressing LBT-PglB tests the hypothesis that the early codon composition of the LBT is the cause of the poor LBT-PglB expression observed. In addition, the optimized vector construct will be helpful to future users of the LBT in the case that optimization successfully improves expression. Indeed, as shown in **Figure 6**, the optimized LBT-PglB showed restored expression levels with comparable activity to the T7-PglB enzyme.

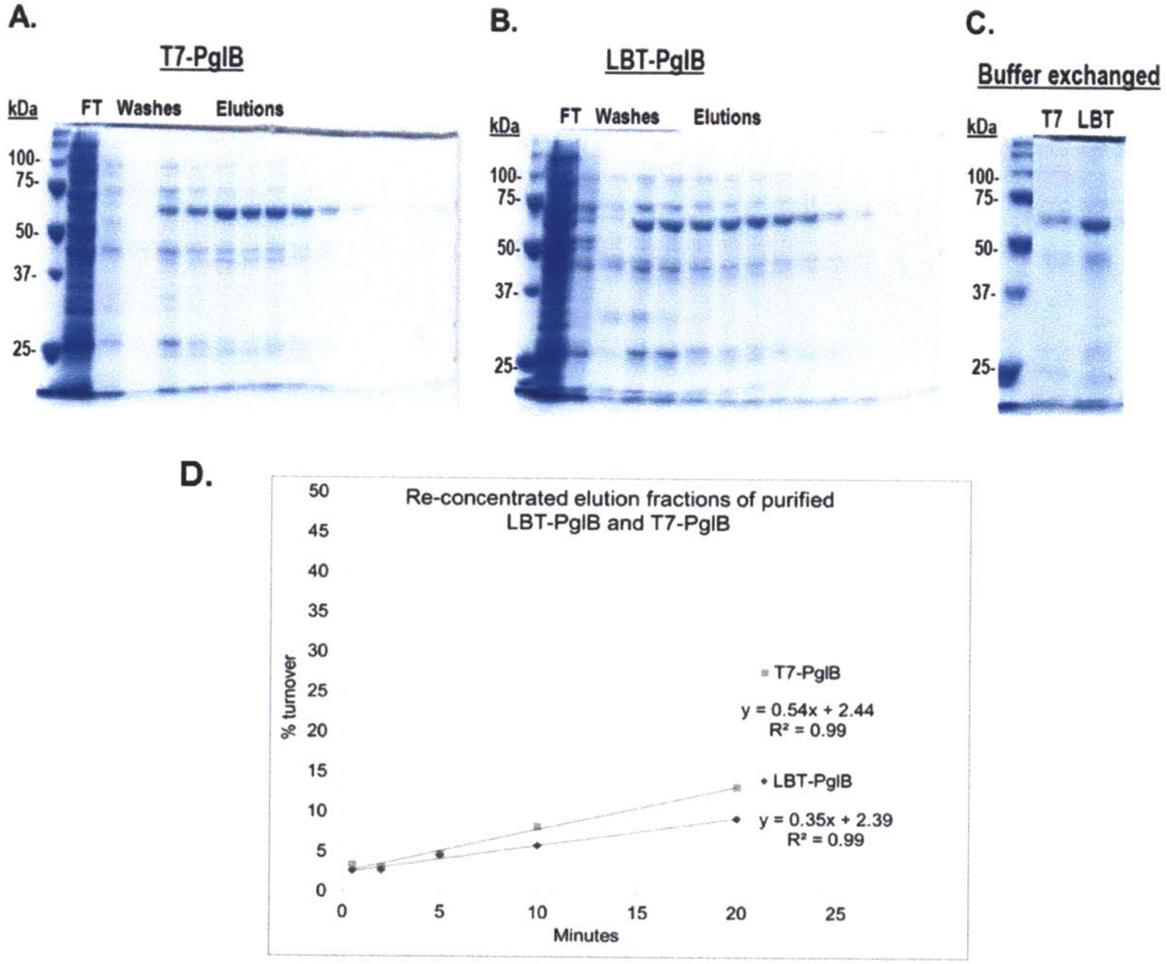


Figure 6: Purification of stocks of T7-PglB and LBT-PglB.
SDS-PAGE (10%, Tris-Glycine) showing *A*. T7-PglB purification fractions, *B*. LBT-PglB purification fractions, and *C*. the final stock solutions after combining protein-containing elutions from each construct, concentration and buffer exchange. The first lane of each gel contains the Precision Plus Protein Unstained Standard ladder (Bio-rad) with labeled molecular weights in kDa. After normalizing the concentration for the fractions in part *C*., activity assays were carried out (*D*). As previously observed, the LBT-PglB construct shows robust activity, which is only slightly less than that of T7-PglB.

Optimization of LBT-PglB luminescence

With a well-expressed, active, and stable LBT-PglB in hand, the luminescent studies were initiated. However, unexpected obstacles once again presented themselves, delaying the ultimate measurements, though providing important and relevant lessons that will be critical to keep in mind in future experiments involving the LBT.

The first task involved validating that the LBT-PglB protein expressed was functioning as expected with regard to luminescence intensity and lifetimes. It was thought that it may be useful to have a well-behaved, soluble LBT-fusion protein for use as an internal comparison when measuring basic binding and luminescent parameters for PglB. Former members of the lab used a LBT-ubiquitin fusion for characterization and development of LBT properties and the protein construct is conveniently well-expressed, soluble, and previously charaterized [11,24]. A large stock of His6-LBT-Ubiquitin was expressed, purified, aliquotted, and stored at -80°C for use as a standard comparison when measuring LBT-PglB photophysical properties. **Figure 7** shows the SDS-PAGE results of the purification. Bands above the 11.6-kDa molecular weight of LBT-ubiquitin stained on the anti-His western blot, appear in multiples of 11.6 kDa, and are apparent on pictures of SDS-PAGE results from earlier purifications by former lab members (see Landon Martin, dissertation); thus these bands are expected to represent higher-order oligomers rather than impurities. The lifetime was determined for this stock of pure LBT-ubiquitin, which agreed with previous measurements of the LBT-Tb^{3+} lifetime (**Figure 8D**) [10].

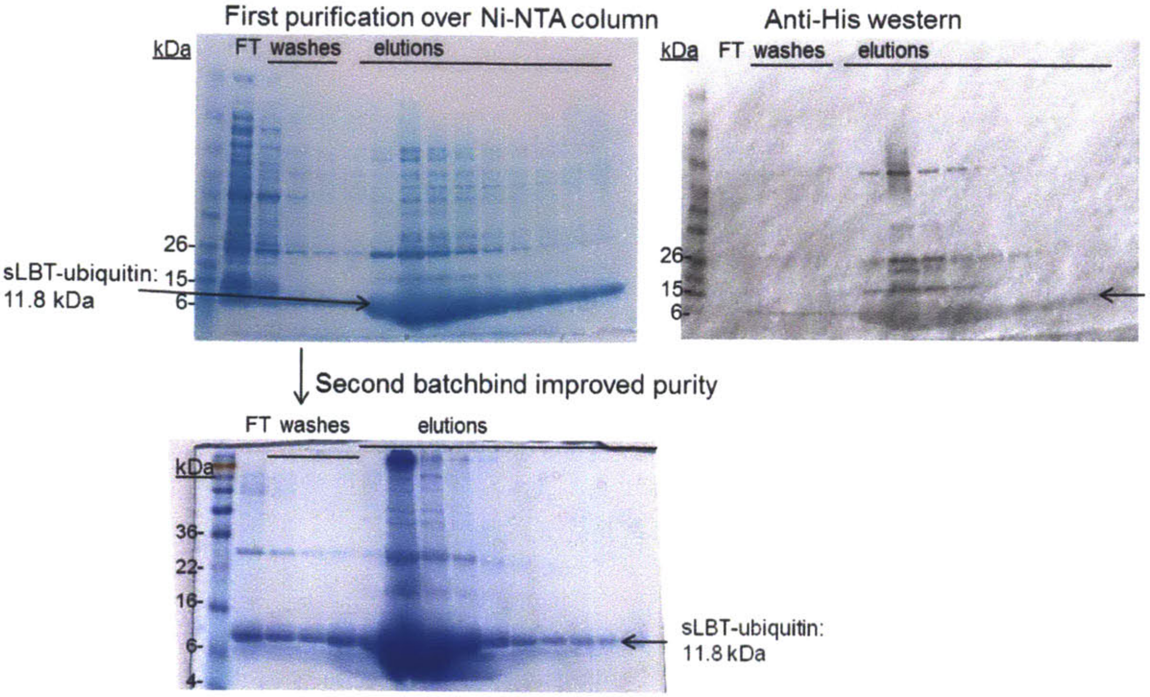

First purification over Ni-NTA column
kDa
FT
washes
elutions
26-
sLBT-ubiquitin: 15-
11.8 kDa
6-
Anti-His western
kDa
FT
washes
elutions
26
15
6
Second batchbind improved purity
FT
washes
elutions
kDa
36-
22-
16-
6-
4-
sLBT-ubiquitin:
11.8 kDa

Figure 7: Expression and purification of an LBT-ubiquitin standard.
First purification on Ni-NTA resin showed impurities in the elutions (*top left*, Coomassie-stained). The anti-His western blot (*top right*) verified that the major band was His6-LBT-ubiquitin. Repeating Ni-NTA batchbind of elutions resulted in improved purity (*bottom left*). Arrows indicate His6-LBT-ubiquitin monomer band.

It was expected that the LBT on two different proteins should give highly similar luminescence intensities upon titration with Tb^{3+}. This property would make it possible to precisely measure the concentration of the enzyme in solution without doing a full titration, as the addition of a given amount of Tb^{3+} should correspond to a specific intensity. A standard internal comparison would be required to control for potential instrument or lamp fluctuations however. The available LBT-ubiquitin was expected to provide this internal comparison in the measurements of concentrations of PglB in pure and semi-pure solutions. Thus, LBT-ubiquitin was titrated with Tb^{3+} at three concentrations and the luminescent data were recorded (**Figure 8**).

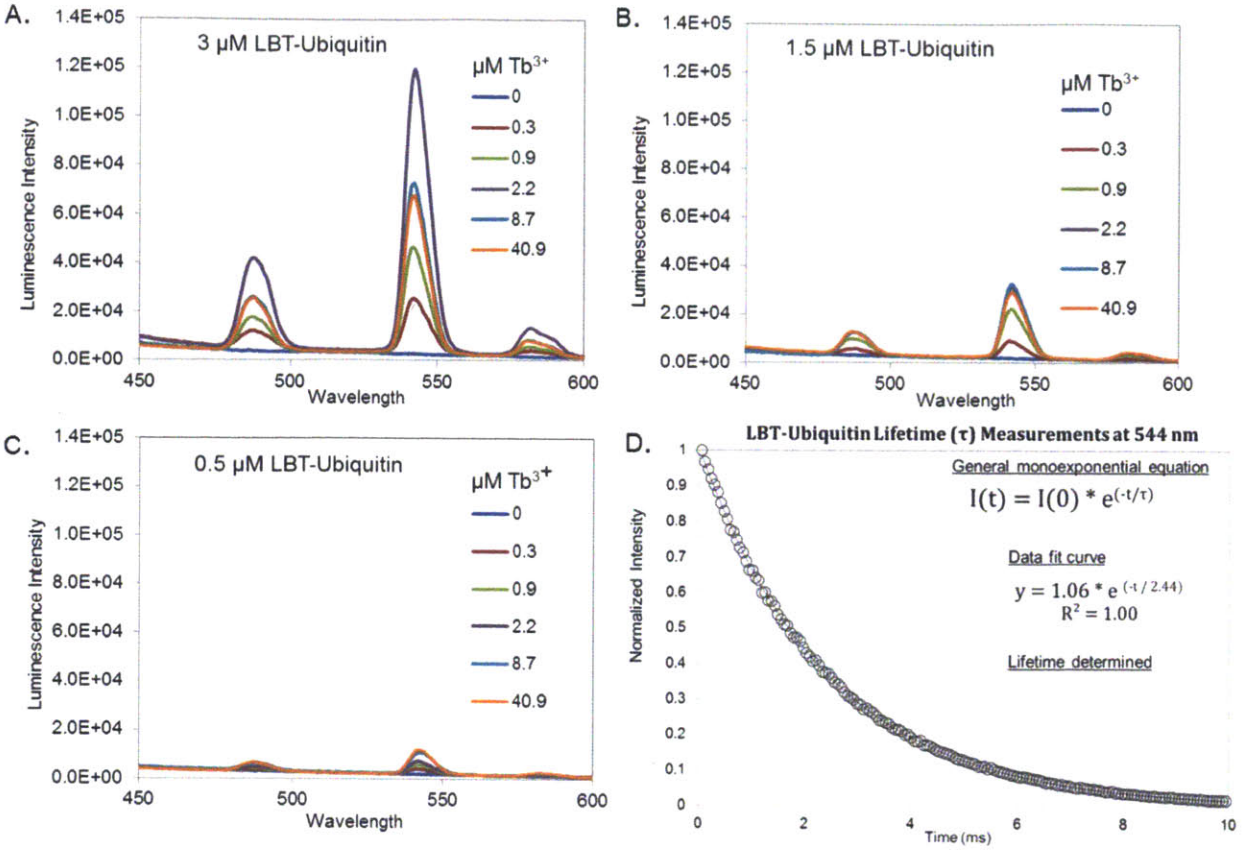


Figure 8: Measurements of LBT-Ubiquitin intensities and lifetime.
Terbium ion titrations of LBT-Ubiquitin at 3 μM *(A)*, 1.5 μM (*B*), and 0.5 μM (*C*). *D:* Lifetime measurement of LBT-ubiquitin bound to Tb^{3+} at 544 nm agrees with previous measurements made independently of the LBT-Tb^{3+} lifetime [10]. The data was fit to a monoexponential equation in which I(t) is the intensity of emission at time t, I(0) is the initial emission intensity, t = time, and τ is the lifetime, determined from the equation fit.

Additionally, LBT-ubiquitin was expected to serve as a standard for determining effects of buffer components on LBT-Tb^{3+} intensities. As PglB is a membrane protein, detergent and potentially additional stabilizing agents and salts may enhance monodispersion of LBT-PglB in solution. As a result, the effects of increasing concentrations of several cations, glycerol, and DDM on LBT-Ubiquitin luminescence were determined (**Figure 9**).

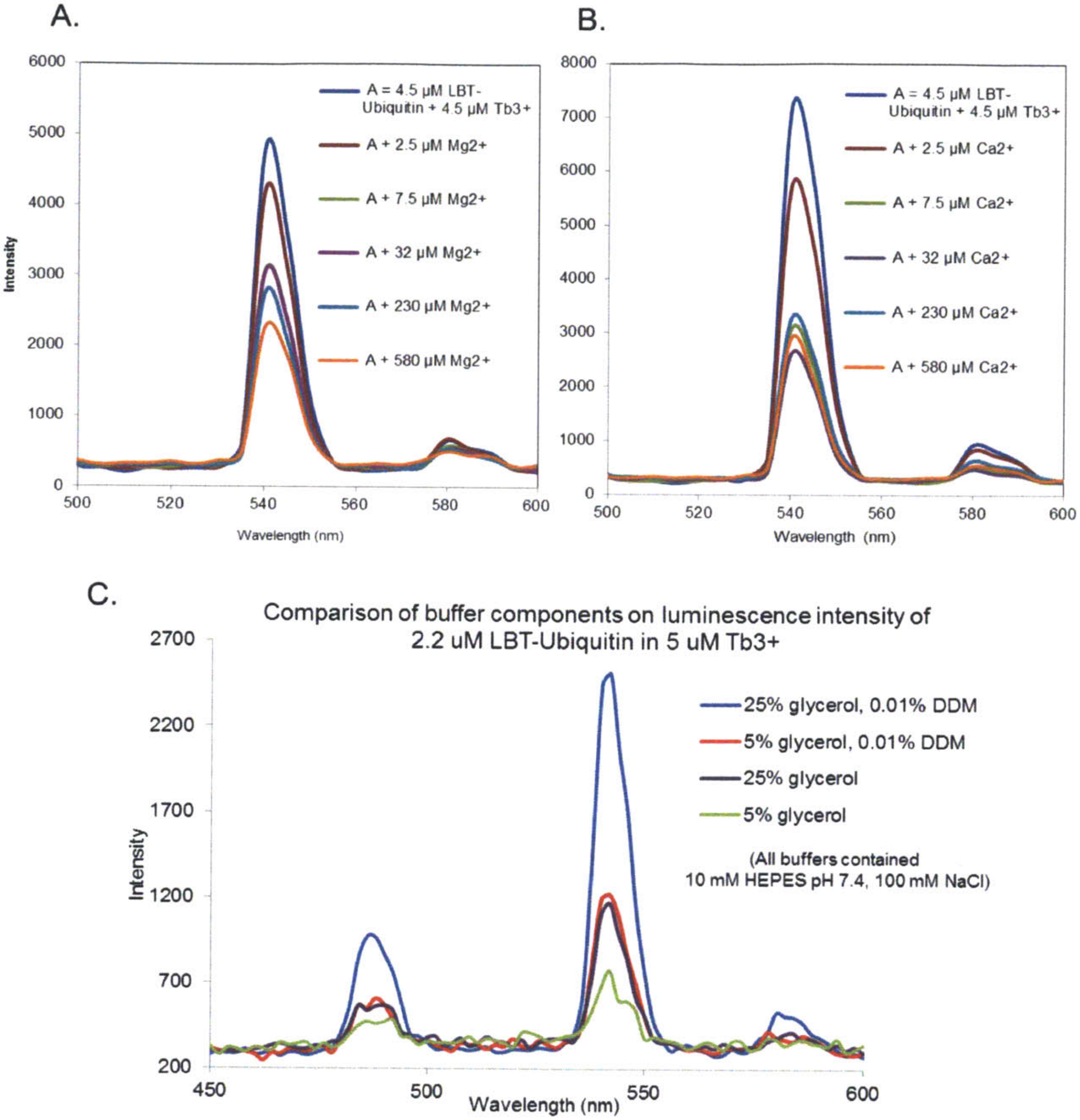


Figure 9: Effects of buffer components on LBT-Ubiquitin luminescence intensities. Addition of increasing Mg^{2+} (*A*) and Ca^{2+} (*B*) into a solution of LBT-Ubiquitin saturated with Tb^{3+} results in steadily decreasing signal intensity. (*C*) Presence of both 0.01% DDM and 25% glycerol in minimal buffer (10 mM HEPES, pH 7.4, 100 mM NaCl) results in an improved LBT-Ubiquitin luminescent signal relative to minimal buffer with levels of either or both DDM and glycerol decreased.

It can be seen in **Figure 9** that addition of higher levels of cations leads to a decrease in luminescence intensities measured for LBT-Ubiquitin. This is expected to be due to competition with Tb^{3+} binding to the LBT when a high level of competing cation relative to Tb^{3+} is added. The addition of non-Tb^{3+} cations was measured because PglB requires a (generally divalent) cation for activity which can be substituted by Tb^{3+} (**Figure 10**). Therefore, it is hoped that providing a low level of divalent cation to solutions of LBT-PglB would out-compete Tb^{3+} for binding to PglB cation-binding site and allow added Tb^{3+} to bind primarily to the LBT site.

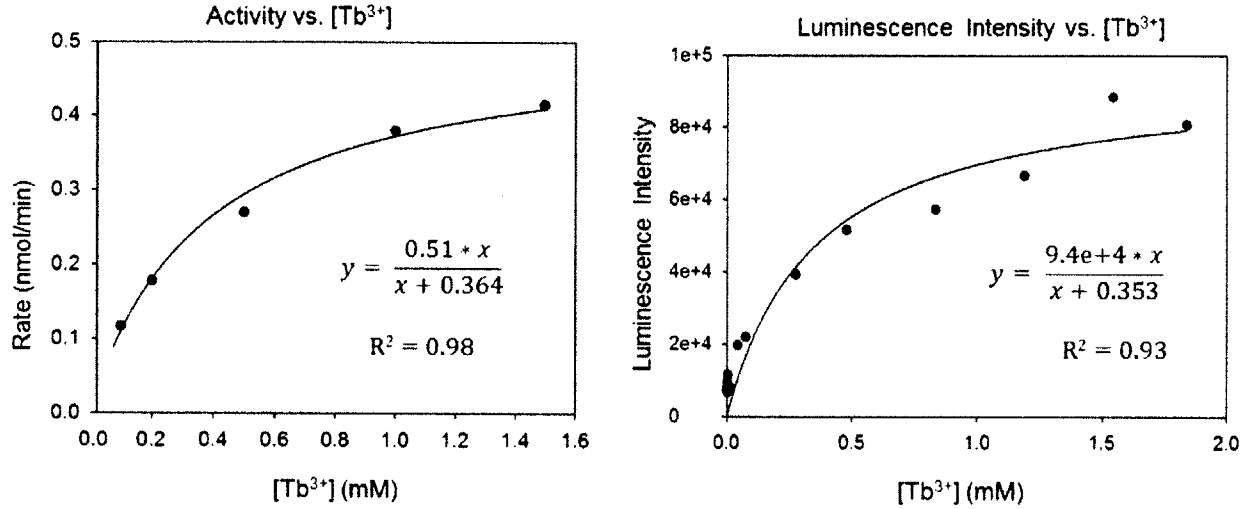


Figure 10: Tb^{3+} can satisfy cation requirement for T7-PglB activity.
Michaelis-Menten binding curves were determined using the activity (*left*) and luminescence intensity (*right*) for a range of Tb^{3+} concentrations. The constant in the denominator represents the estimated binding constant in μM. The above measurements were made using PglB *without* LBT; thus, an amino acid capable of sensitizing Tb^{3+} is most likely present near the cation-binding site.

Upon purifying the optimized LBT-PglB, it was surprising to note that a concentration measured using UV spectroscopy did not give expected luminescence intensities. Confoundingly, the titrations showed that the enzyme was saturated at expected concentrations, ensuring that the concentration measurements from UV absorbance are correct (**Figure 11**). (As Tb^{3+} has low nM binding affinity for LBT, when the LBT is present at micromolar concentrations Tb^{3+} can be titrated into the protein solution, such that each Tb^{3+} atom added binds LBT until all LBT is bound, at which point the luminescence will no longer increase with further addition of Tb^{3+}.) Not only did the intensities fail to correspond to those expected based on the LBT-ubiquitin standards, but the intensities were extremely low. With the low level of intensity observed, luminescent measurements would carry a high degree of error. Therefore, experiments were carried out in order to optimize the luminescent signal measured for LBT-PglB. In addition, investigations were begun to determine the reason for the discrepancy in signals for LBT-PglB and LBT-ubiquitin at the same concentrations of protein and Tb^{3+}.

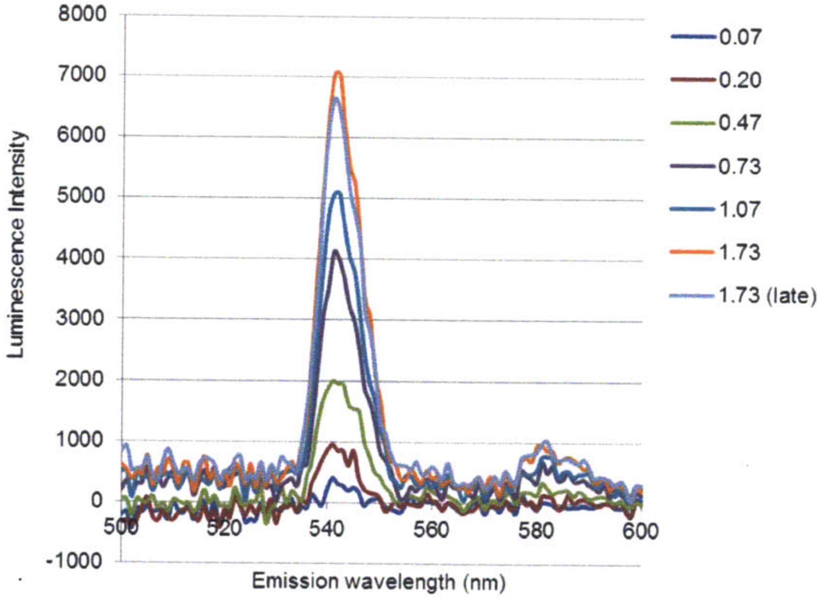


Figure 11: LBT-PglB shows relatively low luminescence intensities.

In order to determine the role of the phosphorimeter program settings on intensities, Dr. Andrew Krueger of the Imperiali lab systematically varied several major parameters and compared the emission intensities (**Figure 12**). A protein construct of EGF (Epidermal Growth Factor) fused to an LBT labeled with a Europium sensitizer was used in these experiments; thus, concentrations of Europium are noted rather than Terbium [25]. Based on the measurements, Dr.

Krueger concluded that changing the delay time does not affect emission intensities, which he reasoned may be due to having flash lamp rather than a pulse laser. Increasing the sample window (amount of time for data collection after flash) increases the intensity of the signal as well as background without any beneficial effect on signal to noise (S/N). Lastly, Dr. Krueger observed that increasing the number of flashes (the number of times pulse is repeated and emission data collected at each point, then combined) improved intensities of signal much more than intensities of noise, thus improving the S/N ratio.

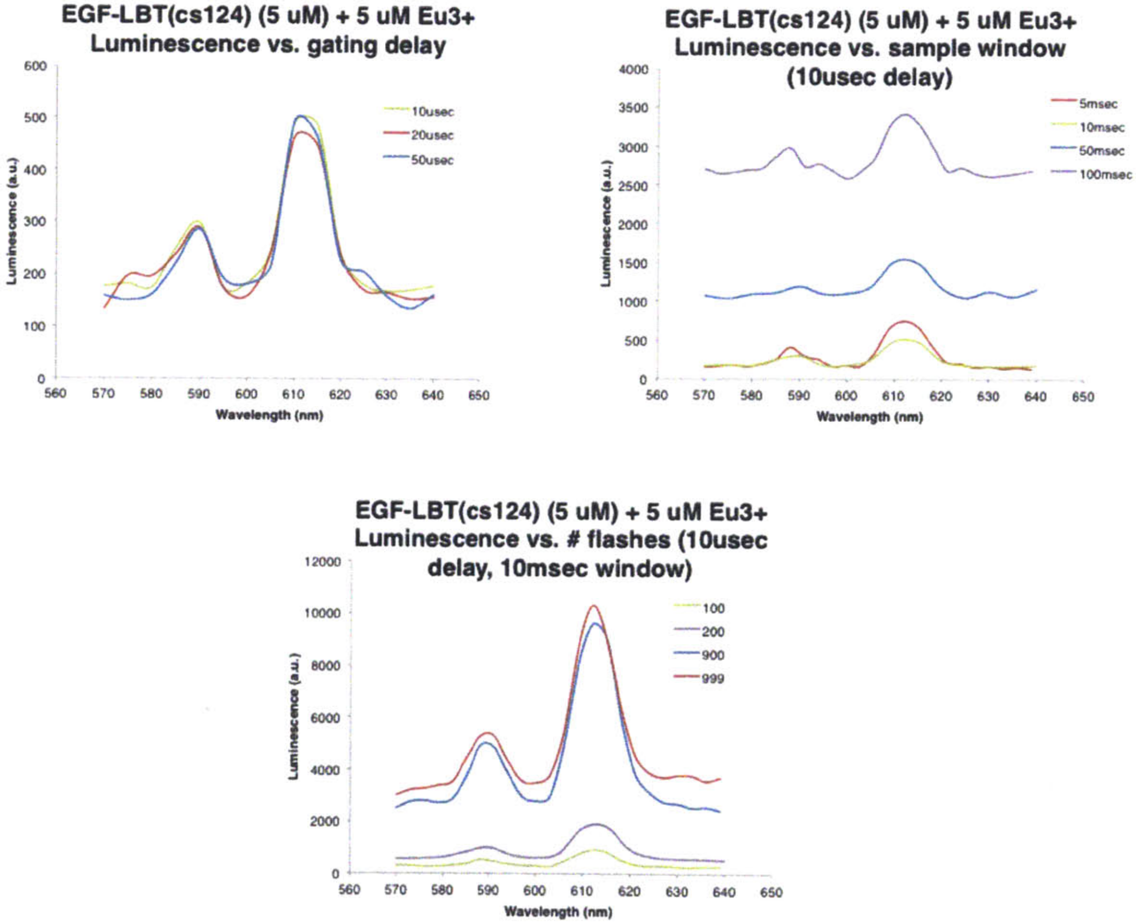


Figure 12: Effects of varying phosphorimeter parameters on luminescence intensity (courtesy of Dr. Andrew Krueger).

LBT-EGF was used to demonstrate effects on intensity when varying gating delay, sample window, and number flashes (clockwise from top left). The (cs124) refers to the LBT labeled with a Europium sensitizer.

Based on measurements made by Dr. Krueger, the number of flashes in a program was considered the most beneficial parameter to increase in order to improve intensities of luminescent intensities measured. Indeed, when this parameter was varied for LBT-Ubiquitin in the presence of saturating Tb^{3+}, the signal-to-noise was improved greatly. The same was observed for LBT-PglB at non-saturating and then saturating Tb^{3+} (**Figure 13**).

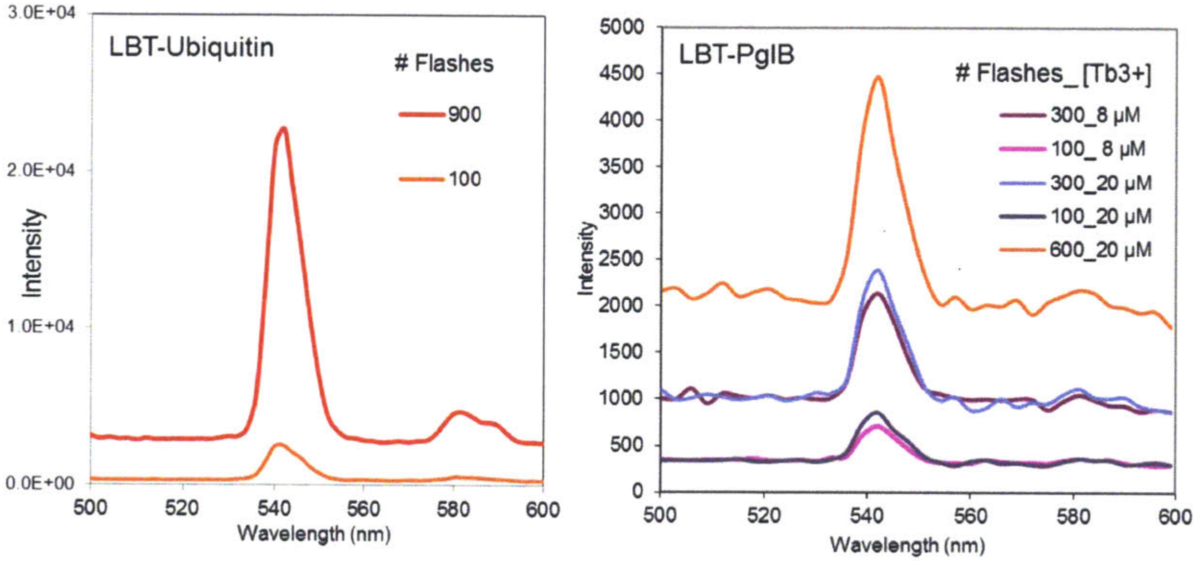


Figure 13: Improvement in luminescent intensities measured for LBT-Ubiquitin and LBT-PglB when the number of flashes is increased.

Increasing the number of flashes used when making luminescent measurements thus appeared to universally improve intensities measured for LBT-fusion proteins in the presence of Tb^{3+}. However, it was still apparent that when instrumental parameters were held constant, luminescent intensities measured for LBT-PglB differed from those of LBT-ubiquitin. At this time, an additional LBT-fusion protein had been introduced: LBT-AglB. AglB, an archaeal OTase, would be expected to work within the LRET framework similarly to PglB, and measurements made using both OTases are expected to provide more information and legitimacy to analysis of data. Another graduate student in the Imperiali lab, Michelle Chang, had expressed and stored solutions of LBT-AglB. Thus, a comparison of luminescent intensities for three LBT fusions was made. Equal concentrations of LBT-AglB, LBT-PglB, and LBT-ubiquitin were made up in identical buffer solutions, and titrations were carried out for each protein identically (**Figure 14**). It can be seen that each LBT-fusion protein approaches saturation of luminescence at roughly 5 μM, verifying that the concentration of each protein is also roughly 5 μM. But the intensities for each fusion protein upon addition of increasing Tb3+ vary tremendously, with LBT-AglB showing the highest, followed by LBT-ubiquitin, and LBT-PglB showing the lowest (**Figure 14**).

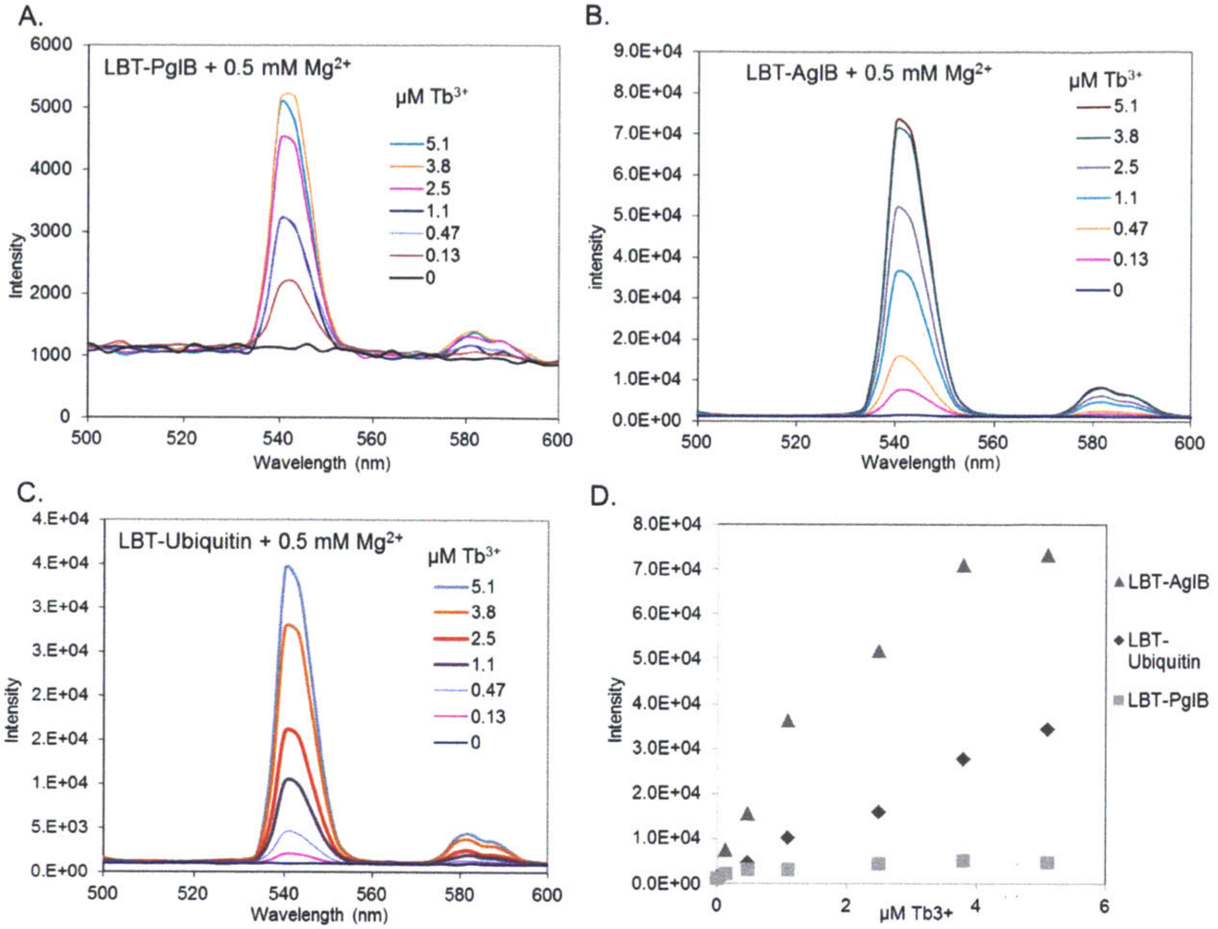


Figure 14: Comparison of luminescent intensities of three proteins fused to LBT. Equimolar concentrations (5 μM) of LBT-PglB (*A*), LBT-AglB (*B*), and LBT-Ubiquitin (*C*) in the presence of 10 mM HEPES, pH 7.4, 100 mM NaCl, 5% glycerol, 0.5 mM $MgCl_2$ were titrated with Tb^{3+}. Titrations were performed on the same day using the same stock solutions of Tb^{3+} to minimize error. Intensity versus [Tb^{3+}] is shown for all three constructs in (*D*) for comparison. Michelle Chang of the Imperiali lab provided data for LBT-AglB.

The data shown in **Figure 14** are extremely important because they overturn the previously held assumption that the LBT luminescence intensity is not affected by the identity of its fusion partner. In fact, the fusion identity matters greatly, though the reasons are not entirely understood. Some effects can be explained by the presence or absence of a cation-binding site within the fusion protein. Indeed, **Figure 15** shows that the addition of Mg^{2+} suppresses LBT-ubiquitin luminescence intensity (also seen in **Figure 9**). In contrast, final addition of 50 μM $MnCl_2$ deeply enhanced the LBT-PglB signal.

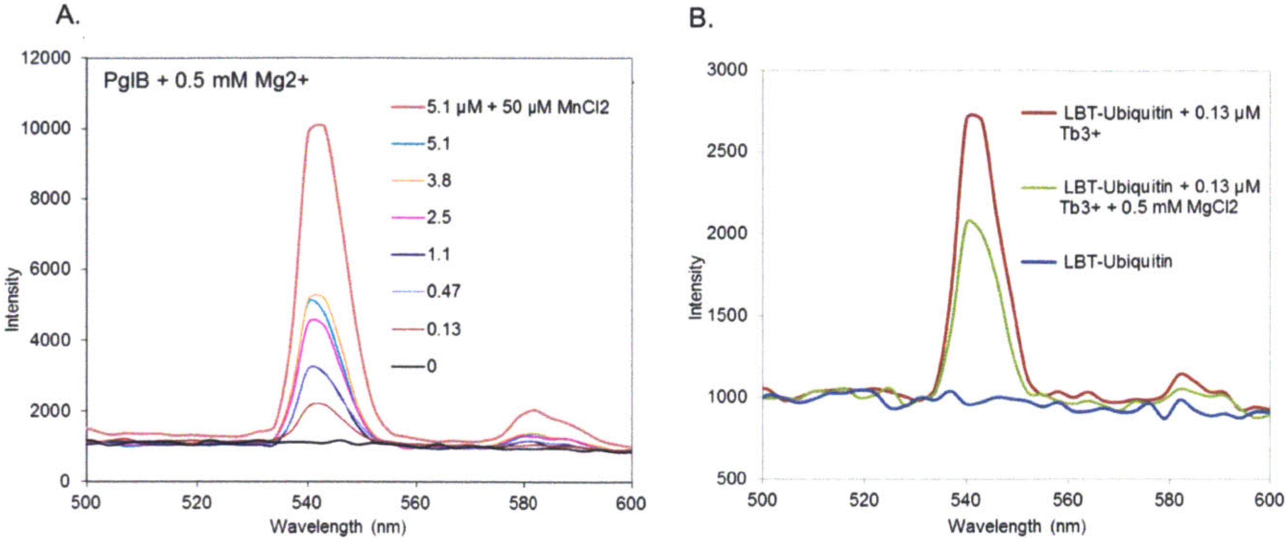


Figure 15: Additional data from luminescent titration experiments shown in Figure 14. (*Left*) LBT-PglB titration is identical to that in **Figure 14** except includes data from a final measurement made upon addition of 50 μM Mn^{2+}. (*Right*) Before performing the titration shown in **Figure 14**, LBT-Ubiquitin luminescence was measured first in minimal buffer (10 mM HEPES pH 7.4, 100 mM NaCl, 5% glycerol), second after addition of 0.13 μM Tb^{3+}, and third after the addition of 0.5 mM $MgCl_2$. (The spectrum measured with both 0.13 μM Tb^{3+} and 0.5 mM $MgCl_2$ is the same as the 0.13 μM Tb^{3+} spectrum shown in **Figure 14** for LBT-ubiquitin).

In view of the above results, it became clear that LBT-Ubiquitin does not provide a reliable standard for determining the effect of buffer components on LBT-PglB luminescence intensities. Therefore, solutions of LBT-PglB were directly used to determine the optimal conditions for LBT-PglB luminescence (**Figures 16, 17, 18**). Ultimately, it was determined that LBT-PglB gives optimal luminescent intenties in the presence of 10 mM HEPES pH 7.4, 100 mM NaCl, 5% glycerol, 0.1% DDM, and 100 μM $MgCl_2$. In these conditions, and using a program with the number of flashes set to at least 300, luminescence intensities were roughly ten-fold higher than before optimization.

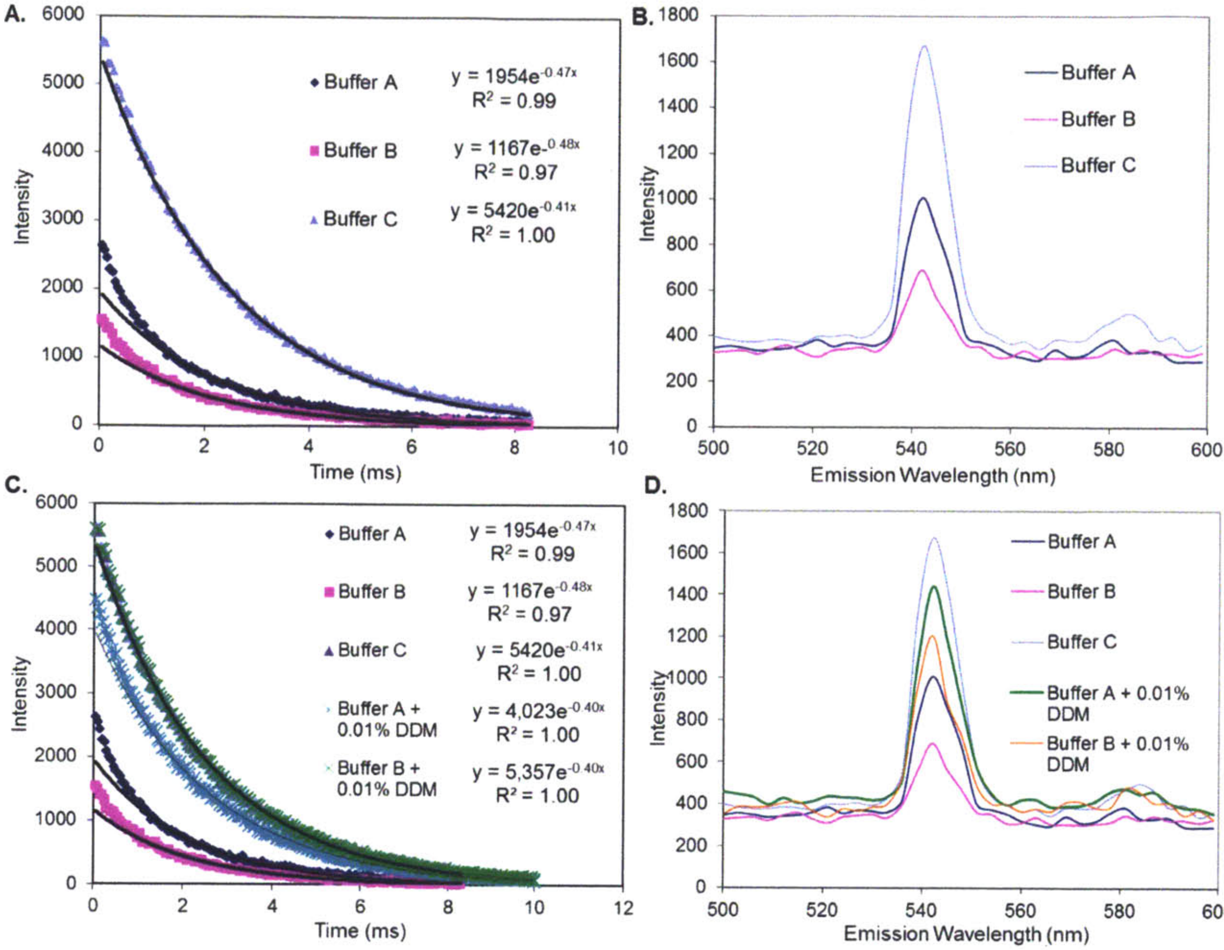


Figure 16: Comparison of LBT-PglB luminescence intensities and lifetime-curve fits in the presence of three buffers.

Buffer A = 10 mM HEPES, pH 7.4, 100 mM NaCl, 5% glycerol; Buffer B = 10 mM HEPES, pH 7.4, 100 mM NaCl, 25% glycerol; Buffer C = 10 mM HEPES, pH 7.4, 100 mM NaCl, 5% glycerol, 0.01% DDM. Buffer C gives the most accurate curve fit for lifetime (*A*) and greatest luminscence intensity (*B*). It is the presence of DDM in Buffer C that accounts for the enhancement, which is verified by the improvement of the signals for Buffer A and Buffer B when 0.01% (final) DDM is directly added to the solutions (*C, D*).

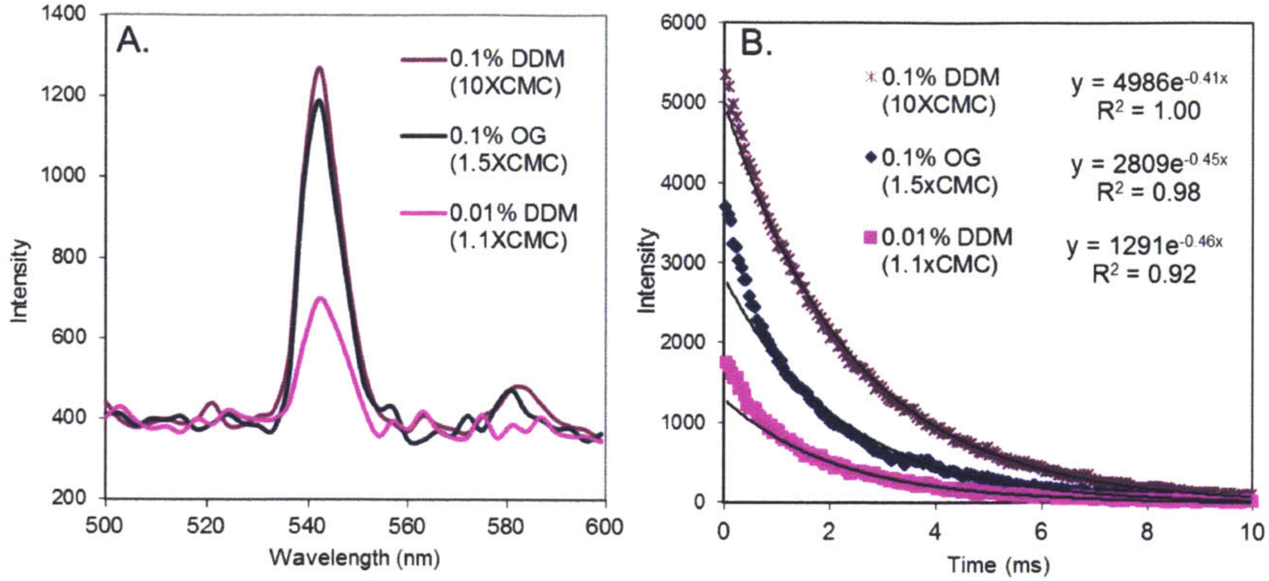


Figure 17: Effect of detergent identity and concentration on luminescence intensity of LBT-PglB in presence of Tb^{3+}.

(*A*) One solution of LBT-PglB with saturating Tb^{3+} was divided into three aliquots and equal volumes were added to each aliquot to give a final detergent concentration of 0.01 % DDM, 0.1 % DDM, or 0.1 % OG. The luminscent intensities and lifetimes (541 nm) were then measured for each solution.

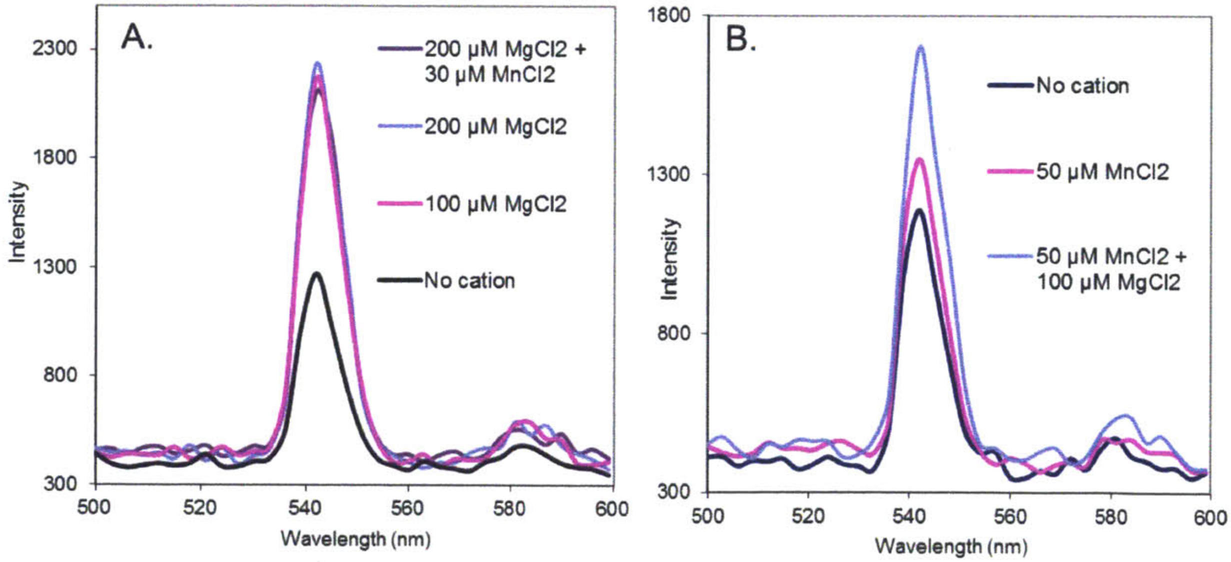


Figure 18: Effect of cation identity and concentration on LBT-PglB luminescence.

(*A*) To a solution of LBT-PglB and saturating Tb^{3+}, aliquots of $MgCl_2$ were added to give a final concentration of 100 µM, followed by 200 µM, and finally an aliquot of $MnCl_2$ was added to give a final cation concentration of 200 µM $MgCl_2$ and 30 µM $MnCl_2$. (*B*) Order of cation addition is reversed to ensure presence of one cation is not influencing effect of second. 50 µM $MnCl_2$ added, followed by 100 µM $MgCl_2$

Insertion of cysteine mutations to allow labeling of PglB with thiol-reactive acceptor dye

The two native, non-conserved cysteines in PglB were mutated to serine so that single cysteines could be inserted at choice locations for site-specific labeling of the enzyme (this construct was designated LBT-PglB-CxxC, representing the LBT-PglB with C28S and C408S mutations). The double-cysteine mutant was shown to be active and expressed at wild-type levels (**Figure 19**).

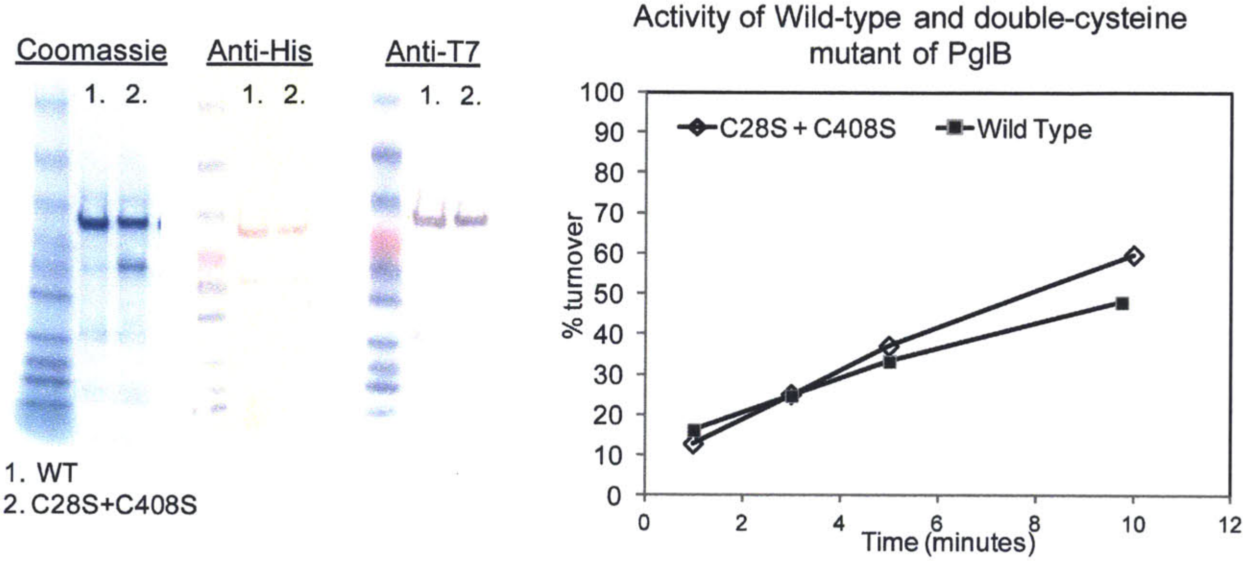


Figure 19: Mutating native cysteines in PglB to serine.
Construction of PglB construct with both native cysteines mutated to serines shows WT-level expression and activity. *Left*: Coomassie-stained SDS-PAGE and western blots with antibodies to N-terminal tag (T7) and C-terminal tag (His_{10}) indicate that the mutations do not significantly alter the recombinant expression levels of PglB produced. *Right*: Activity assay results for equal concentrations of WT-PglB and PglB-CxxC show that the cysteine mutations do not significantly affect PglB activity.

Locations for six cysteine insertions were chosen based on multiple criteria: first, the sites should be accessible to labeling (solvent exposure) and unconstrained (to accommodate Bodipy-dye); second, the predicted measurements at the mutations sites should be informative (in terms of donor-acceptor distance as well as containing regions of conserved loops); third, the precise residues that are mutated or neighboring a cysteine insertion should show low conservation of to avoid significant disruption of activity or structure. Assessment of these criteria was made based on existing structural and biochemical data.

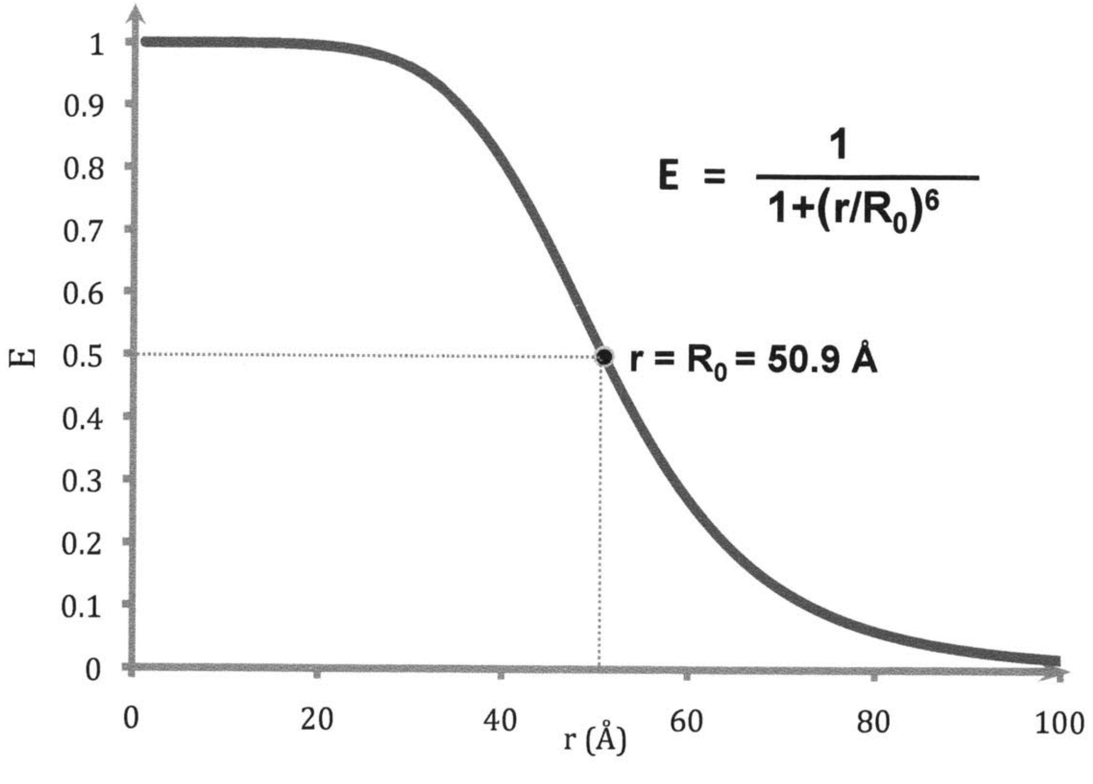


Figure 20: Efficiency (E) of LRET as a function of Tb^{3+} (donor) and Bodipy-TMR (acceptor) distance (r).

The R_0 value between this donor-acceptor pair has been determined to be 50.9 Å. For a narrow range above and below R_0 the curve is approximately linear and the signal can be measured with reasonable precision, as a small change in distance results in a significant change in signal, or efficiency. As the distance from R_0 increases, the precision of the measurement drops off rapidly, as the efficiency becomes less sensitive to small changes in distance.

Additionally, the acceptor dye will ideally be placed within a given range of distances. Accurate measurements can only be made when the acceptor and donor are within a short difference from the R_0-value, or the value at which the LRET efficiency is 50 percent of the maximum, and which is specific to a given donor-acceptor pair. As the difference between R_0 and the acceptor-donor distance increases, the LRET efficiency decreases or increases at a rate of $1/r^6$, where r is the distance between donor and acceptor (**Figure 20**). Importantly, the labeled peptide is useful even if its binding site does not fall within the specified range; while specific distances could not be ascertained, the change in LRET for the case in which the peptide is unbound versus bound is almost certainly great enough to detect the presence of absence of binding. Thus, the requirements for peptide binding or lack thereof may be investigated. However, the thiol-reactive dye chosen to function as an acceptor, Bodipy-TMR maleimide, has been used in combination with the LBT in LRET experiments in the past and the R_0 has been determined as 50.9Å (**Figure 21,** [10]). Prior use of the dyes in an LRET experiment and determination of R_0 simplifies the following analysis and provides confidence that measurements made will be informative. Based on the structural data available, it is possible to estimate an expected range of distances, and the R_0 falls conveniently toward the center of this range.

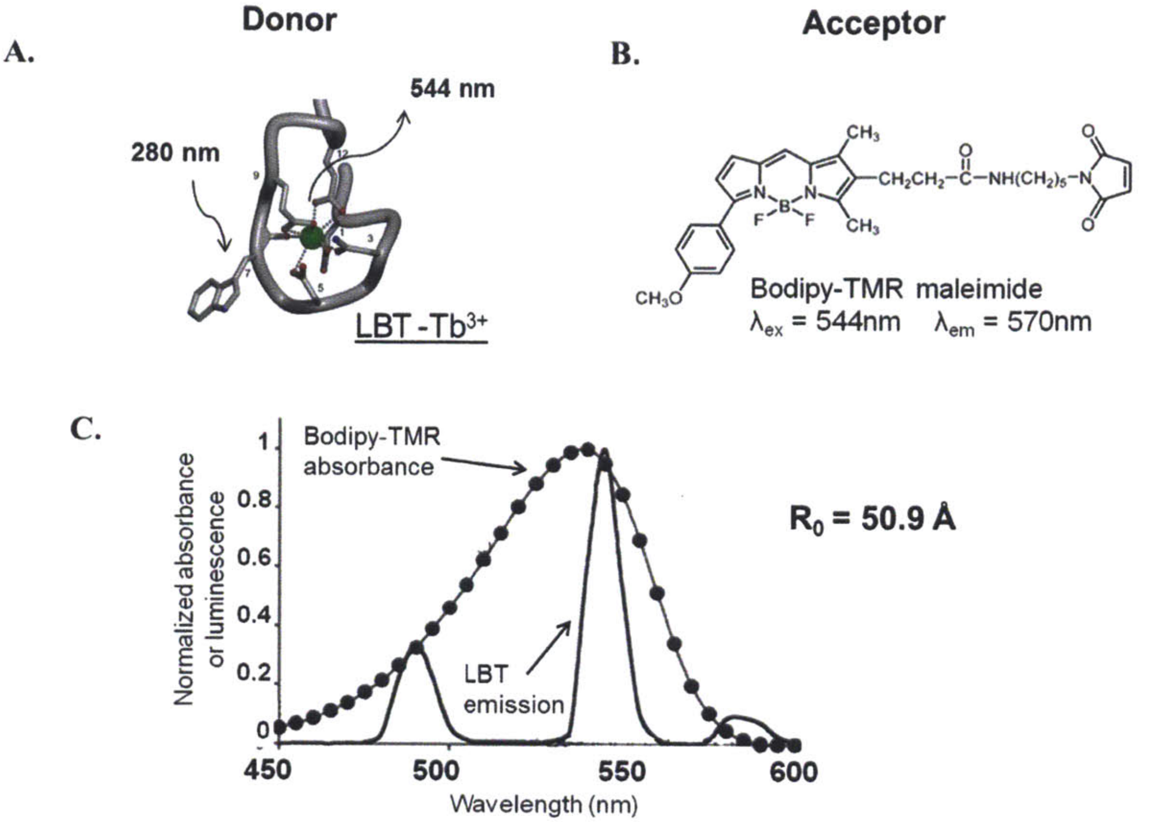
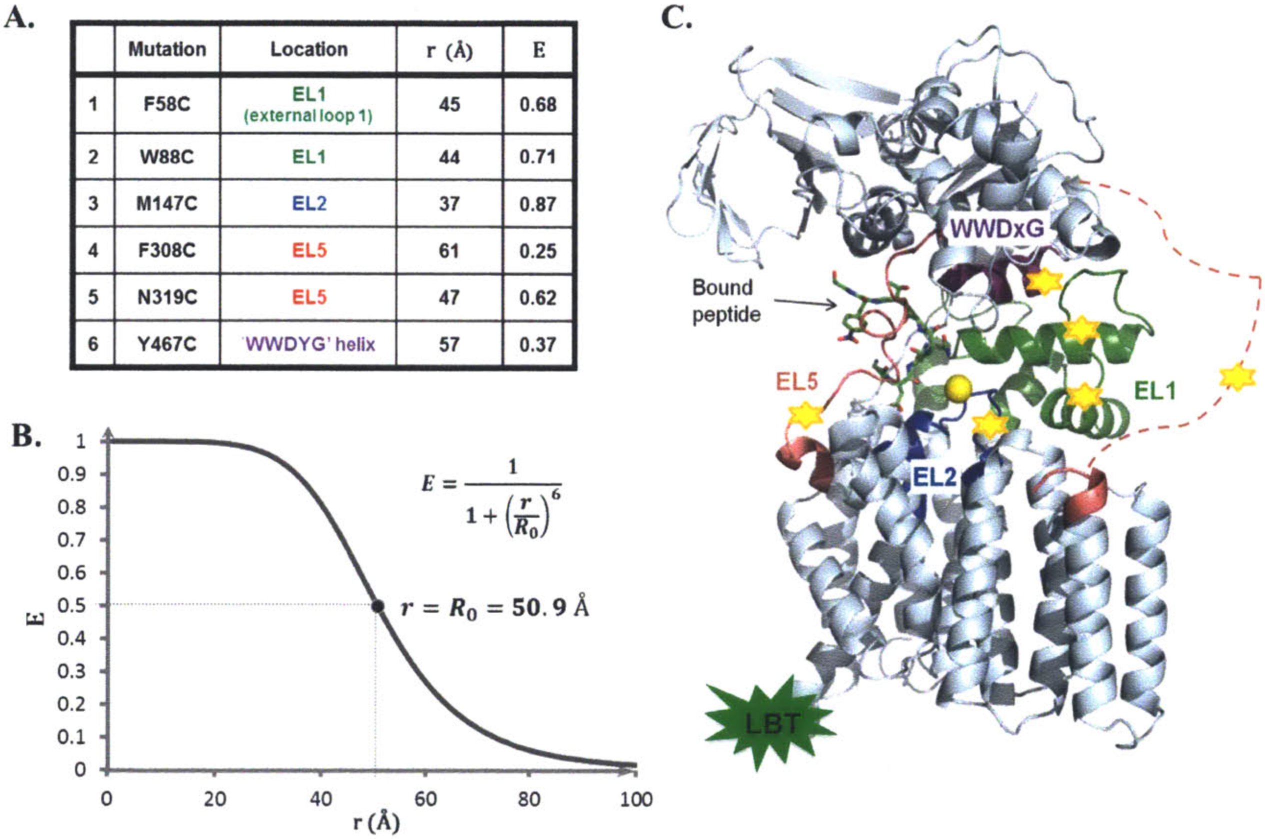


Figure 21: Structure and photophysical properties of LRET pair.
A. Depiction of the LBT bound to Tb^{3+}. *B.* Structure of Bodipy-TMR. *C.* The excitation (absorption) spectrum for Bodipy-TMR (circles), the emission spectrum of LBT-Tb^{3+} (smooth line) and the extent of overlap.

The structure of the *C. lari* PglB was used to ascertain that the Bodipy-TMR label is an appropriate LRET acceptor for the range of distances expected. In addition, the structure was informative in choosing suitable locations for placement of cysteine mutations, which should be close enough to R_0 to make informative measurements (**Figure 22**). However, these measurements are enormously approximate for several reasons. First, the PglB structure is only resolved to 3.4 Å, immediately adding a high degree of error. Additionally, while the sequences of PglB from *C. jejuni* and *C. lari* are highly similar, they are not identical and residues at some of the mutation sites differ between the two proteins; thus, the location of residues in the *C. lari* structure may not accurately represent the exact situation in the *C. jejuni* PglB. Finally, the measurements are made between the site of mutation and the N-terminal LBT extension modeled onto the PglB structure. In fact, the mutation sites provide the unique handles for conjugation to Bodipy-TMR. The error increases as the size of the Bodipy-TMR is the component whose distance from the LBT matters, and it is unknown the degree to which the LBT model represents its actual position when expressed on the N-terminus of PglB.



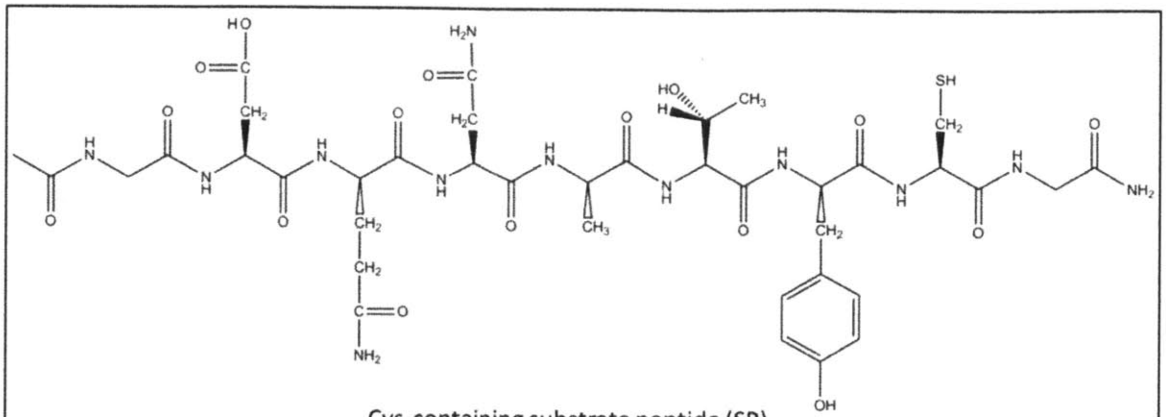
	Mutation	Location	r (Å)	E
1	F58C	EL1 (external loop 1)	45	0.68
2	W88C	EL1	44	0.71
3	M147C	EL2	37	0.87
4	F308C	EL5	61	0.25
5	N319C	EL5	47	0.62
6	Y467C	'WWDYG' helix	57	0.37

Figure 22: Sites of cysteines mutated into PglB to act as handles for LRET acceptor. *(A):* Table listing the residues that were mutated to cysteines (in separate constructs) and the general location of the protein in which they appear. The measurements were made in Pymol between an LBT model at the N-terminus of PglB and each residue. (*B*) Efficiency (E) of LRET as a function of Tb^{3+} (donor) and Bodipy-TMR (acceptor) distance (r), also shown in **Figure 20.** *(C)*: Snapshot of a model of the *C. lari* crystal structure shown in Pymol. The sections of the protein in which the cysteines mutations appear are colored to coordinate with the identifiers listed in the last column of the table on the right.

Nonetheless, the chosen locations span a range of estimated distances within 10-15 Å of the R_0 value, increasing the likelihood that one or several will fall within an informative measurement range. Even when precise distances cannot be accurately obtained, conformational changes can be detected if the helix or loop containing the labeled cysteine undergoes a motion from a distance of relatively low LRET efficiency to one of high efficiency. The cysteine-mutants chosen were constructed by site-specifically mutating the LBT-PglB-CxxC construct, providing unique handles to conjugate Bodipy-TMR.

Synthesis and labeling of cysteine-containing peptide substrates of PglB

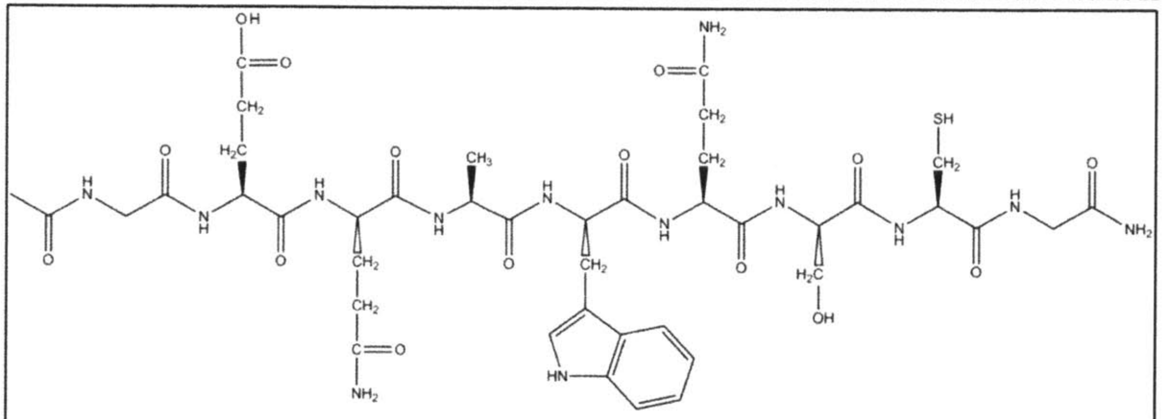
In order to measure intermolecular LRET to gain information on location and order of substrate binding, a substrate peptide (SP) containing a cysteine was synthesized to allow thiol-specific labeling. A negative control (NC) peptide was also synthesized, which also contains a cysteine and similar amino acid content, but lacks the consensus sequence. These peptides are shown in **Figure 23**. The cysteine was added C-terminal to the consensus sequence in SP (and close to the C-terminus in NC) because rough distance measurements of the N-terminus (location of LBT) to the peptide bound in the *C. lari* structure indicates that this distance is closer to the R_0-value than the distance between the N-terminus of the bound peptide and the N-terminus of the PglB structure (**Figure 24**).



Cys-containing substrate peptide (SP)

CH_3CO–GDQNATYCG–$CONH_2$

Exact Mass: 968.5 Da



Cys-containing negative-control peptide (NC)

CH_3CO–GEQAWQSCG–$CONH_2$

Exact Mass: 1005.5 Da

Figure 23: Two peptides synthesized for intermolecular LRET studies.
Top panel shows the substrate peptide, which contains the PglB consensus sequence as well as a cysteine to provide a unique thiol group for labeling. Bottom panel shows the negative control peptide, which also contains a cysteine for labeling. This peptide resembles the substrate peptide in size, charge, and amino acid content, but lacks the consensus sequence required for glycosylation by PglB.

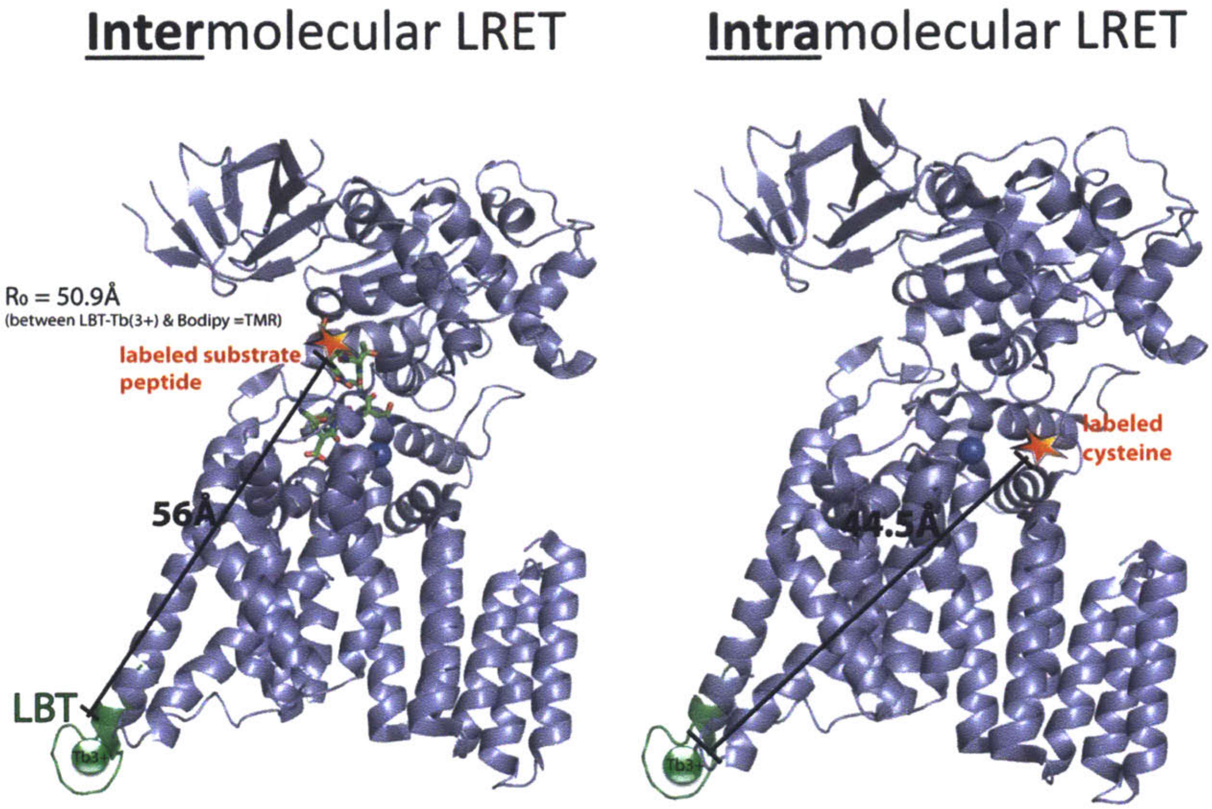

Intermolecular LRET
Intramolecular LRET
Ro = 50.9Å
(between LBT-Tb(3+) & Bodipy =TMR)
labeled substrate peptide
56Å
LBT
Tb3+
labeled cysteine
44.5Å
Tb3+

Figure 24: Distance measurements determine the preferred location of cysteines in substrate peptide and site-specific mutagenesis.
Left: Depiction of the *C.lari* PglB crystal structure showing the bound peptide (N-terminus closest to the bottom in this figure) in green sticks. The structure of the LBT is modeled at the N-terminus of the protein (green) representing where it is genetically encoded. The measurement indicates a distance of 49 Å between the Tb^{3+} and the C-terminus of the peptide. The orange star represents the location the Bodipy-TMR is expected to label a cysteine. Though the measurement is a rough approximation, it falls in the nearby range to the R$_0$ value for the LBT-Tb^{3+} complex and Bodipy-TMR LRET pair, 50.9 Å [10]. *Right:* Analogous figure to left, but in contrast the location of an inserted cysteine for labeling and intramolecular LRET studies is being assessed. Measurement and picture made in Pymol, PglB structure, PDB:3RCE [3], LBT structure, PDB: 1TJB, [8].

These peptides were synthesized by SPPS, purified by HPLC, and verified by MALDI-MS analysis. Pure peptides were subsequently labeled with Bodipy-TMR, re-purified using HPLC, and re-verified with MALDI-MS (**Figures 25, 26**). Luminescent measurements were performed to ensure that these peptides showed predictable behavior with virtually no non-specific emission resulting from 280 nm excitation after a 50-μs delay and high overlap between the LBT-PglB absorbance and the Bodipy-TMR peptide (**Figure 27**). Stocks of each were stored in aliquots at -80°C for upcoming LRET experiments. Control assays showed that the negative control peptide was indeed inactive and that the labeled substrate peptide shows similar activity levels to the unlabeled substrate (**Figure 28**).

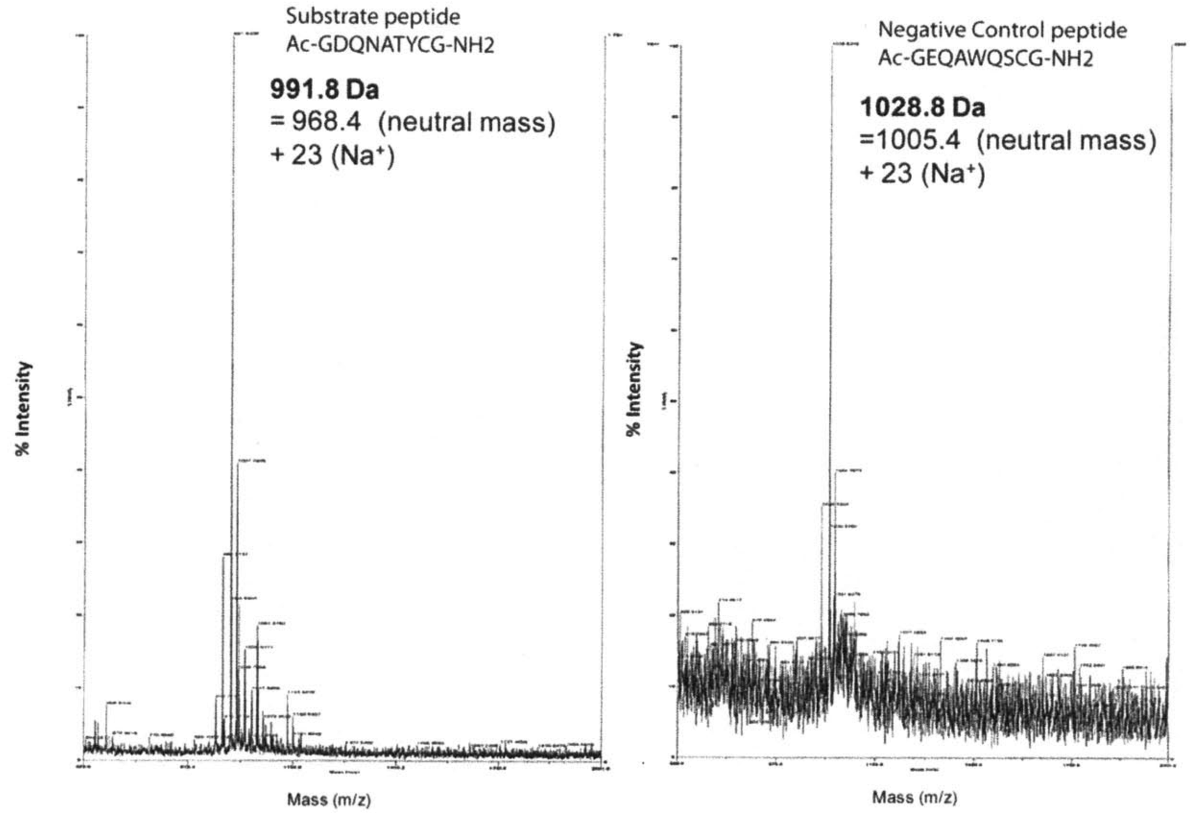


Figure 25: MALDI-MS validation of masses of peptides for LRET studies.
Left: Mass spectrum of substrate peptide solution with a major peak at 991.8 Da. Right: Mass spectrum of the negative-control peptide with a major peak at 1028.8 Da.

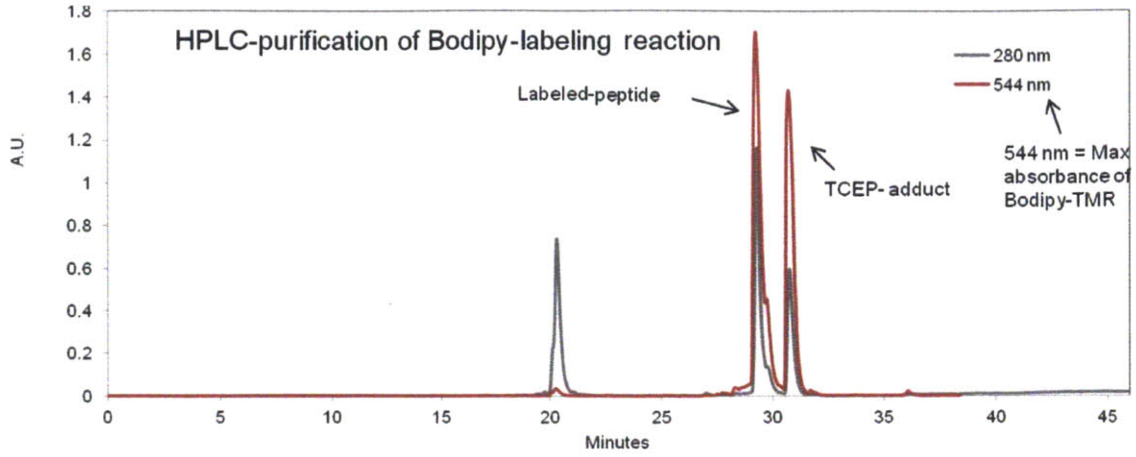


Figure 26: HPLC purification of Bodipy-TMR labeled peptides.
Bodipy-TMR absorbs at 544 nm (red), indicating the peaks containing the Bodipy-TMR. Peptide absorbs at 280 nm (grey). The left-most peak represents unlabeled peptide, the center peak represents labeled peptide, and the right-most peak corresponds to TCEP-adduct with the dye (all determined using MALDI-MS).

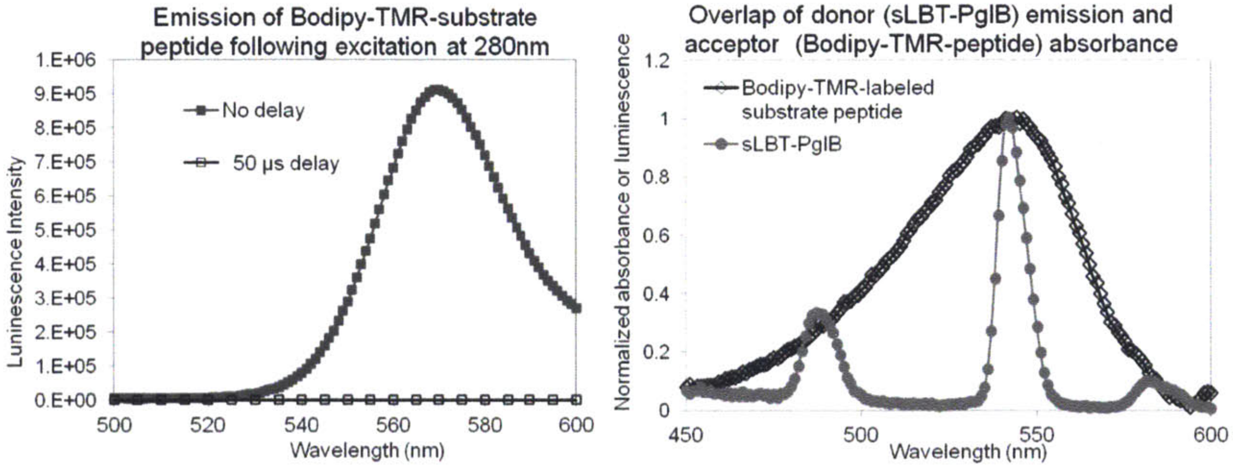


Figure 27: Bodipy-TMR is a favorable LRET acceptor to LBT-Tb^{3+} donor luminescence.
Left: Bodipy-TMR conjugated to a cysteine-containing peptide substrate shows high background luminescence due to non-specific excitation by 280 nm light, which decreases the signal: noise ratio in typical FRET experiments. After a 50 μs time delay, the background signal is virtually eliminated, which demonstrates the benefit of LRET over FRET due to the longer lifetime of several milliseconds. *Right*: Overlap of LBT-Tb^{3+} emission with Bodipy-TMR-thiol conjugate absorbance. The maximum emission of LBT-Tb^{3+} at 544 nm due to excitement at 280 nm overlaps well with the absorbance maximum Bodipy-TMR-thiol conjugate.

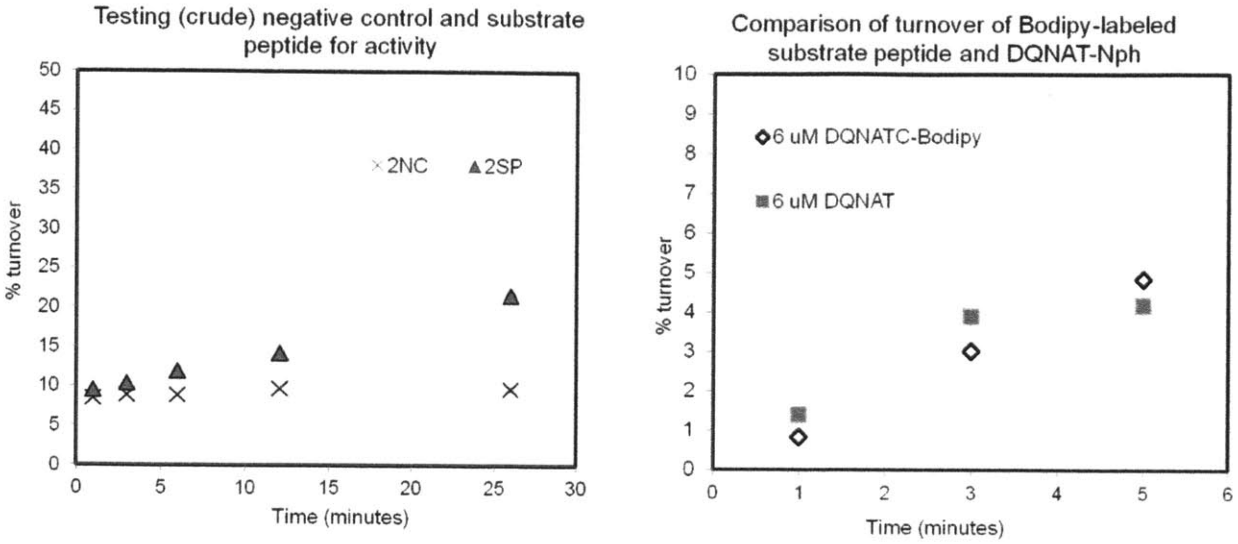


Figure 28: PglB activity assays in presence of substrate and negative-control peptides. *Left:* PglB activity assay shows that the negative-control peptide lacks activity, as expected. *Right:* Bodipy-labeled substrate peptide (SP) shows similar rates to the peptide typically used in PglB assays, indicating that the Bodipy conjugate does not significantly impede binding or catalysis.

Initial LRET measurements

With the above components assembled, initial LRET measurements were made between LBT-PglB and the labeled substrate peptide. The labeled substrate peptide (designated 'SP') was titrated into a solution of LBT-PglB saturated with Tb^{3+}. Emission spectra were recorded and the relative emission at the emission maximum for BODIPY-TMR (572 nm) shows corresponding increases, indicating energy transfer is occurring (**Figure 29**). The increase in emission intensity at 572 nm saturates when all LBT-PglB is bound by labeled peptide, and the intensity at 572 nm can be plotted against the concentration of BODIPY-TMR SP to give a binding constant (**Figure 29**). The binding constant calculated below was in good agreement with activity-based determinations of binding constants for the Ac-DQNAT substrate peptide to PglB, which average 1 μM. The agreement between these values is a good indication that the LRET measured is representing native binding behavior of the substrate peptide to the enzyme.

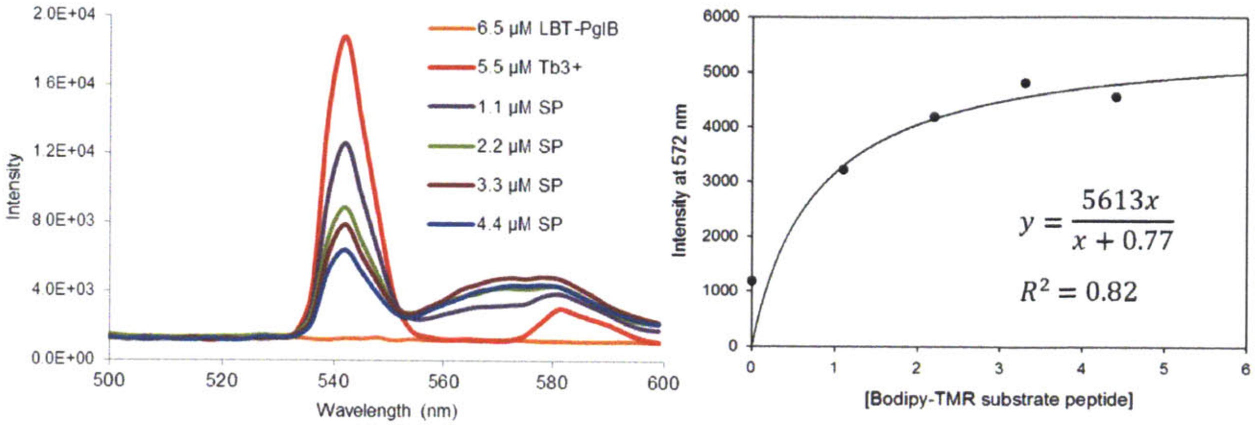


Figure 29: Increase and saturation acceptor emission.
(*Left*) Labeled substrate peptide was added in increasing amounts, resulting in signal saturation at 572 nm (emission maximum for Bodipy-TMR). The increasing intensities can be graphed versus substrate concentration to approximate the binding constant of the peptide to LBT-PglB (*right*). The curve gives a rough fit that estimates the binding constant at 0.77 μM, which agrees with other estimates averaging 1 μM.

In order to translate the LRET measurements into distance measurements, lifetime measurements were made in addition to measuring emission spectra upon addition of increasing concentrations of labeled-SP (**Figure 30**). The corresponding decrease in lifetime is visible and the saturation of the lifetime change can also be observed. By comparing the lifetime measurements for LBT-PglB with saturating Tb^{3+} to LBT-PglB with saturating Tb^{3+} in the presence of labeled SP (with labeled SP contributing roughly half of the signal to the measured lifetime, determined by coefficients for each parameter in the biexponential curve), distance measurements can be obtained (**Figure 30**). The lifetime values are used to calculate a transfer efficiency, which is then used along with the known R_0 value of 50.9 Å to calculate the distance r (see *Methods*). The measurements shown in **Figure 30** result in calculated distance values that agree with expectations based on measurements made using the PDB file for the *C. lari* PglB.

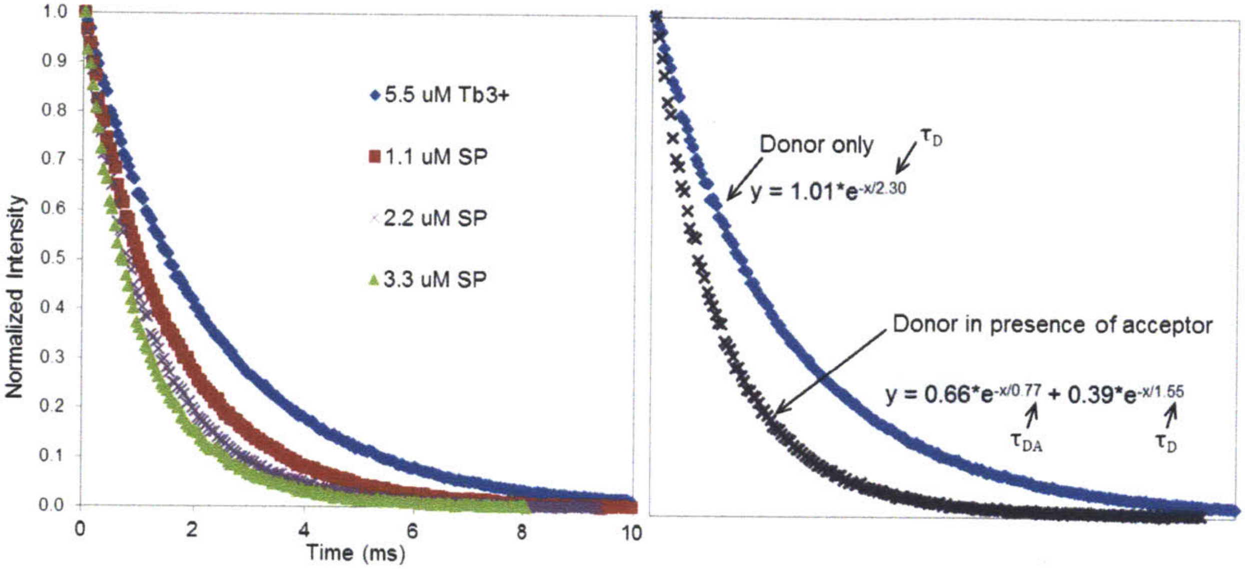


Figure 30: LRET measurements with labeled peptide substrate.
Left graph shows the curves obtained with increasing concentrations of SP (substrate peptide, labeled). The right graph shows only the curves obtained for LBT-PglB with saturating Tb^{3+} (monoexponential) and with the addition of 2.2 μM labeled acceptor peptide (biexponential). Curve fits were made using Sigmaplot.

Conclusion

In order to investigate aspects of the OTase mechanism outlined above, dynamic distances of pure, active PglB in its apo state and upon substrate binding will be measured using intramolecular and intermolecular LRET. This approach utilizes the pure, stable form of PglB, optimized by protocols described in previous progress reports. In contrast to structural data, this strategy is uniquely valuable because it ensures the data derived from the studies is representative of the active states of enzyme. Specifically, a lanthanide-binding tag (LBT) was genetically fused to create the construct LBT-PglB-His$_{10}$. The terbium ion (Tb^{3+})-bound LBT will function as a donor chromophore. Cysteines have been introduced into specific locations of PglB as well as substrate peptide, allowing specific labeling using a thiol-reactive acceptor dye.

Several fundamental tasks have been completed toward development of an LRET system for measurement of dynamic properties of the OTase reaction. These tasks include synthesis, labeling and purification of the substrate and negative-control peptides, the double-cysteine mutant of PglB was constructed, expressed, and measured for activity; the double-cysteine mutant was used as a template for site-directed mutagenesis reactions yielding constructs with cysteines in desirable locations for labeling, the LBT-PglB construct was made in the pET24a(+) vector and the cysteine mutants placed in this construct, and the LBT-Ubiquitin construct was obtained and expressed and purified, and aliquotted for future use. The sequence of the codon sequence of the LBT has been optimized, yielding much-improved expression of LBT-PglB. The important discovery has been made that LBT luminescence intensity is dependent upon the fusion protein and buffer conditions. Finally, LRET was observed between the labeled substrate peptide and LBT-PglB with Tb^{3+}, and the measurement agrees with expectations based on the available structural data. This observation provides strong evidence that the system described is capable of successfully measuring distances between substrate and OTase and potentially between labeled cysteines within the enzyme and the N-terminal LBT. Remaining tasks in the measurement of intermolecular LRET distances include the synthesis of an additional negative-control peptide which is less similar in sequence to the substrate. Unfortunately, the negative-control peptide that was synthesized and labeled (described above) shows LRET in addition to the substrate peptide, indicating some non-specific binding is occurring.

For intramolecular LRET measurements, the cysteine mutants in the LBT-PglB construct must be expressed and measured for activity to determine which are suitable to carry forward into testing, the labeling of the cysteine constructs of LBT-PglB with Bodipy-TMR must be optimized and each active mutant will need to be labeled. These experiments are currently being applied to an archaeal OTase AglB in addition to PglB.

In addition to the goals set forth above and the progress made toward achieving them, there exist additional possibilities for using the LBT-PglB system to investigate further aspects of the reaction. A growing possibility is the attainment of a non-disruptive label that can be specifically conjugated to the Und-PP-glycan substrate. Probe design alone could be highly informative as to which characteristics of the polyprenyl moiety accounts for key interactions and binding specificity. Determining the location of the binding pocket for this substrate and the role the first linked sugar plays in determining specificity would introduce opportunities for glycoengineering to advance. Along these lines, understanding these mechanistic components will have implications for strategies such as directed evolution and rational design of the active site for a desired cause; for example, altering specificity may allow incorporation of new probes *in vivo* that would enable valuable exploration into the role of specific glycans in bacterial pathogenesis. Therefore, establishing in a stepwise manner the principles by which the OTase accomplishes its unique chemical transformation is of vital importance for opening up various avenues for studying N-linked glycosylation in a realm of contexts.

Materials and Methods

Synthesis of substrate and negative-control peptides

Standard Fmoc SPPS was used in production of the cysteine-containing peptides. 200 mg of Novapeg Pink Resin was agitated (with $N_2(g)$) in dichloromethane (CH_2Cl_2) for ten minutes. The Fmoc protecting group was removed from the resin and subsequent amino acids by agitating in 20% piperidine in dimethylformamide (DMF) for five minutes (3X). Washes were done in DMF for 1 minute (5X) and CH_2Cl_2 for 1 minute (3X). The success of each coupling was tested using the TNBS test. Initially, coupling conditions were 4 (resin) equivalents of PyBOP, 4 equivalents of amino acid and 8 equivalents of diisopropyl ethyl amine (DIPEA) were agitated in DMF with the resin for 1 hour. After difficulty coupling 5Y, all subsequent couplings were done with 6 equivalents of PyBOP, 6 equivalents of amino acid and 8 equivalents of diisopropyl ethyl amine (DIPEA), agitated in DMP with the resin for 2 hours. When leaving the incomplete peptide overnight, the reaction was washed, dried in $N_2\,(g)$ and stored at 4°C.

Purification and labeling of substrate-peptide and negative-control peptide

Peptides were purified by reverse phase HPLC with a Waters 600 automated control module on a YMC C18 preparative column eluting with acetonitrile/water containing 0.1% TFA. For detection, a Waters 2487 dual wavelength absorbance detector was used to record at 228 nm and 280 nm. Standard HPLC conditions were 5% acetonitrile for 5 minutes followed by a linear gradient from 5% to 95% acetonitrile for 30 minutes. Purity was confirmed by analytical HPLC and correct mass validated by ESI-MS on a Mariner electrospray mass spectrometer. Bodipy labeling was performed on HPLC purified peptides with N-terminal amines. The lyophilized, pure peptide was dissolved in 0.1 M NaHCO3 and the concentration was determined on a Shimazu Spectrophotometer by recording the absorbance at 280 nm in 6 M Guanidine, using the Extinction Coefficient of tyrosine (εC = 1280 M-1cm-1). 0.52 μmoles of peptide was added to 1.3 μmoles of BODIPY® TMR (Invitrogen), dissolved in 50 μL of DMSO. The reactions were shaken for 4 hours and then purified by preparative HPLC as described above except with monitoring wavelengths at 280 nm and 544 nm (excitation max for BODIPY-TMR). Standard HPLC conditions were a linear gradient from 7% to 100% acetonitrile for 30 minutes. Purity wasconfirmed by analytical HPLC and correct mass validated by ES-MS on a Mariner electrospray mass spectrometer.

Emission spectra

Luminescence titrations were conducted on a Horiba Jobin Yvon Fluoromax-3 equipped with a Spex 1934D3 phosphorimeter in 1 cm path length quartz cells. Sensitization of Tb(III) luminescence was carried out by exciting tryptophan at 280 nm and recording the luminescence at 544 nm. A 315 nm long-pass filter was used to eliminate interference from harmonic doubling. In addition to components explicitly mentioned, all solutions contained 10 mM HEPES, pH 7.4, and 100 mM NaCl. Instrument settings were made using the experiment type "Phosphorimeter Emission Acquisition" with the following settings: Increments = 3 nm, integration time of 1s, scan start of 450 or 500 nm, scan end of 600 nm, excitation wavelength = 280 nm, number of scans = 1, sample window = 10 ms, delay after flash of 0.05 ms, time per flash = 40 ms, number of flashes = 300, signal collected = S, slit widths of 5 nm (excitation) and 5 nm (emission). Aliquots of Tb(III) ions were added, and the luminescence was recorded to obtain a titration curve. These data were fit using the EXCEL or SIGMAPLOT as appropriate.

Lifetime measurements

Lifetime measurements were recorded on a Horiba Jobin Yvon Fluoromax-3 equipped with a Spex 1934D3 phosphorimeter. The intensity at 541 nm was monitored following a lamp pulse at 280 nm from a xenon flash lamp. Lifetime measurements were made using the experiment type "Phosphorimeter Decay Acquisition by Delay" on the Fluoromax-P with the following settings: Initial delay and delay Increments 0.05 ms, max delay of 10 ms, number of scans = 2 (averaged), sample window = 20 ms, time per flash = 70 ms, number of flashes = 20, signal collected = S, slit widths of 5 nm (excitation) and 10 nm (emission).

Calculation of distances using lifetimes

The lifetime decay data was fit to a monoexponential or biexponential curve:

monoexponential: $I(t) = I(0)\exp(-t/\tau)$

biexponential: $I(t) = I(0)_1\exp(-t/\tau_1) + I(0)_2\exp(-t/\tau_2)$

where I(t) luminescence intensity, t = time, I(0) = luminescence at t = 0, τ = lifetime.

SIGMAPLOT was used to determine the best curve fit.The LBT construct without acceptor present is fit to a monoexponential curve (i.e. one lifetime) and the LBT in the presence of an acceptor is fit to a biexpontential (i.e. two lifetimes: LBT influenced and not influenced by the acceptor). These lifetime values can then be used to calculate the distance between the donor and acceptor using the Förster equation:

$r = R_0[(1/E)-1]^{1/6}$

where r = distance, R_0 = Förster radius (50.9 Å for LBT-Tb^{3+} and Bodipy TMR), E = energy transfer. Energy transfer (E) is calculated directly from the lifetimes derived from the curve fits:

$E = (1 - \tau_{DA})/ \tau_D$

where τ_{DA} = lifetime of the donor in the presence of the acceptor, τ_D = lifetime of the donor alone.

References

1. Jaffee MB, Imperiali B (2011) Exploiting topological constraints to reveal buried sequence motifs in the membrane-bound N-linked oligosaccharyl transferases. Biochemistry 50: 7557-7567.
2. Jaffee MB, Imperiali B (2013) Optimized protocol for expression and purification of membrane-bound PglB, a bacterial oligosaccharyl transferase. Protein Expr Purif.
3. Lizak C, Gerber S, Numao S, Aebi M, Locher KP (2011) X-ray structure of a bacterial oligosaccharyltransferase. Nature 474: 350-355.
4. Xiao M, Selvin PR (2001) Quantum yields of luminescent lanthanide chelates and far-red dyes measured by resonance energy transfer. J Am Chem Soc 123: 7067-7073.
5. Hemmila I, Laitala V (2005) Progress in lanthanides as luminescent probes. J Fluoresc 15: 529-542.
6. Heyduk T (2002) Measuring protein conformational changes by FRET/LRET. Curr Opin Biotechnol 13: 292-296.
7. Lohse MJ, Bunemann M, Hoffmann C, Vilardaga JP, Nikolaev VO (2007) Monitoring receptor signaling by intramolecular FRET. Curr Opin Pharmacol 7: 547-553.
8. Nitz M, Sherawat M, Franz KJ, Peisach E, Allen KN, et al. (2004) Structural origin of the high affinity of a chemically evolved lanthanide-binding peptide. Angew Chem Int Ed Engl 43: 3682-3685.
9. Jee J, Ishima R, Gronenborn AM (2008) Characterization of specific protein association by 15N CPMG relaxation dispersion NMR: the GB1(A34F) monomer-dimer equilibrium. J Phys Chem B 112: 6008-6012.
10. Sculimbrene BR, Imperiali B (2006) Lanthanide-binding tags as luminescent probes for studying protein interactions. J Am Chem Soc 128: 7346-7352.
11. Martin LJ, Hahnke MJ, Nitz M, Wohnert J, Silvaggi NR, et al. (2007) Double-lanthanide-binding tags: design, photophysical properties, and NMR applications. J Am Chem Soc 129: 7106-7113.
12. Schmidt HL, Sperling LJ, Gao YG, Wylie BJ, Boettcher JM, et al. (2007) Crystal polymorphism of protein GB1 examined by solid-state NMR spectroscopy and X-ray diffraction. J Phys Chem B 111: 14362-14369.
13. Weinstock DS, Narayanan C, Felts AK, Andrec M, Levy RM, et al. (2007) Distinguishing among structural ensembles of the GB1 peptide: REMD simulations and NMR experiments. J Am Chem Soc 129: 4858-4859.
14. Ding K, Louis JM, Gronenborn AM (2004) Insights into conformation and dynamics of protein GB1 during folding and unfolding by NMR. J Mol Biol 335: 1299-1307.
15. Bouvignies G, Meier S, Grzesiek S, Blackledge M (2006) Ultrahigh-resolution backbone structure of perdeuterated protein GB1 using residual dipolar couplings from two alignment media. Angew Chem Int Ed Engl 45: 8166-8169.
16. Louis JM, Byeon IJ, Baxa U, Gronenborn AM (2005) The GB1 amyloid fibril: recruitment of the peripheral beta-strands of the domain swapped dimer into the polymeric interface. J Mol Biol 348: 687-698.
17. Evans DA, Wales DJ (2004) Folding of the GB1 hairpin peptide from discrete path sampling. J Chem Phys 121: 1080-1090.

18. Byeon IJ, Louis JM, Gronenborn AM (2004) A captured folding intermediate involved in dimerization and domain-swapping of GB1. J Mol Biol 340: 615-625.
19. Jana S, Deb JK (2005) Strategies for efficient production of heterologous proteins in Escherichia coli. Appl Microbiol Biotechnol 67: 289-298.
20. Cornvik T, Dahlroth SL, Magnusdottir A, Flodin S, Engvall B, et al. (2006) An efficient and generic strategy for producing soluble human proteins and domains in E. coli by screening construct libraries. Proteins 65: 266-273.
21. Hartley JL (2006) Cloning technologies for protein expression and purification. Curr Opin Biotechnol 17: 359-366.
22. Esposito D, Chatterjee DK (2006) Enhancement of soluble protein expression through the use of fusion tags. Curr Opin Biotechnol 17: 353-358.
23. Byeon IJ, Louis JM, Gronenborn AM (2003) A protein contortionist: core mutations of GB1 that induce dimerization and domain swapping. J Mol Biol 333: 141-152.
24. Jee J, Byeon IJ, Louis JM, Gronenborn AM (2008) The point mutation A34F causes dimerization of GB1. Proteins 71: 1420-1431.
25. Reynolds AM, Sculimbrene BR, Imperiali B (2008) Lanthanide-binding tags with unnatural amino acids: sensitizing Tb3+ and Eu3+ luminescence at longer wavelengths. Bioconjug Chem 19: 588-591.

Marcie B. Jaffee

Doctor of Biology
Massachusetts Institute of Technology
103 Inman St., Apt 1, Cambridge, MA 02139
mbjaffee@gmail.com 802.752.0189

Skills and expertise:

- Molecular Biology
- Protein expression and purification
- Membrane proteins
- Glycobiology
- Peptide synthesis
- Enzyme kinetics, spectroscopy, radioactivity
- Tissue culture, *D. melanogaster*, *E. coli*
- Coding in Python, Perl
- Bioinformatics: sequence and phylogeny analysis
- Excel, Photoshop, Illustrator, Endnote
- Works well independently and as a team
- Excellent writer
- Experimental design and troubleshooting
- Learns and completes tasks quickly

EDUCATION

Massachusetts Institute of Technology, Cambridge, MA ***2007 - 2013***
Ph.D. Biology

Columbia University, New York, NY ***2003 - 2007***
B.A. Biochemistry with Honors
Cum laude

RESEARCH EXPERIENCE

Graduate Research- Dr. Barbara Imperiali ***2008 – 2013***
Massachusetts Institute of Technology
Departments of Biology and Chemistry

- Biochemical investigation of oligosaccharyl transferase from enteropathogen C. jejuni
- Membrane-protein expression, purification, and stability optimization in E. coli.
- Improved purified recombinant protein yield from .05 to 5 mg protein per liter of culture.
- Bioinformatic sequence analysis and computer-program development; successfully uncovered conserved motifs within divergent membrane-protein sequences.
- Construction and expression of mutant enzymes to investigate bioinformatic predictions.
- Kinetic characterization of wild-type and mutant enzymes; pH profiles and substrate-specificity
- Design and development of LRET system to study conformational changes and substrate binding in enzymatic mechanism of PglB.

Research Intern- Dr. L. Aravind ***Summer 2007***
National Institute of Health
National Center for Biotechnology Information

- Bioinformatic analysis of E2-ligase folds using phylogenetic, structure and sequence data.
- Created and manipulated sequence alignments using Blast and various text editors.
- Analyzed structural data and constructed figures for the resulting manuscript using Pymol.

Undergraduate Research- Dr. Daniel Kalderon ***2006 - 2007***
Columbia University
Department of Biological Sciences

- Genetic analysis of Hedgehog signal transduction in Drosophila somatic stem-cells
- Worked collaboratively to maintain a screen to identify genes involved in ovarian somatic stem-cell development and regulation: included passing large collections of fly cultures and separating flies based on phenotypic features.
- Observed and recorded phenotypic changes in mutants isolated in screen.

Research Intern- Dr. Daniel Reich ***Summer 2005***
Johns Hopkins University
Department of Physics and Astronomy

- Research into biophysical applications of magnetic nanoparticles in human cells.
- Fusion of ferromagnetic nanowires into human cell lines to direct localization and culture.
- Designed and initiated assembly of a 3-dimensional cell-culture matrix for use as an improved model of cell growth in vivo.

TEACHING EXPERIENCE

Teaching Assistant

- Introduction to Experimental Biology, LAB 7.02, MIT (2011)
- Physical Chemistry of Biomolecular System, LECTURE 7.10, MIT (2009)
- General Chemistry, LAB, Columbia University (2006)
- Volunteer ESL teacher, Community Impact Center, Columbia University (2004)

Mentorship and Tutoring

- MIT: Undergraduate Women's Mentoring Program (2012)
- Private tutor for local high-school student in Chemistry I, Boston, MA (2010)
- Volunteer tutor in Molecular and Cell Biology, Columbia University (2006-2007)

RELATED WORK EXPERIENCE

Part-time research assistant- Dr. Rachael Neve ***2009– 2013***
MIT Viral Gene Transfer Facility

- Interpreted molecular biology vector manipulation and constructed visually informative vector maps using DNA and graphic software programs.

Environmental Health and Safety (EHS) Committee Member ***2009 - 2011***
MIT Department of Chemistry

- Participated in EHS committee meetings and lab safety inspections.

Contract technical writer ***2004 – 2007***
New World Encyclopedia Project

- Performed research on varied topics and wrote articles for online encyclopedia entries.

PUBLICATIONS

Jaffee, MB. & Imperiali, B. Optimized protocol for expression and purification of membrane-bound PglB, a bacterial oligosaccharyl transferase *Prot. Exp. Purif.* **2013**, 89(2):241-50

Jaffee, MB.& Imperiali, B. Exploiting topological constraints to reveal buried sequence motifs in the membrane-bound N-linked oligosaccharyl transferases *Biochemistry* **2011** 50(35):7557-67

Burroughs AM, Jaffee M, Iyer LM, Aravind L. Anatomy of the E2 ligase fold: implications for enzymology and evolution of ubiquitin/Ub-like protein conjugation *J. Struc. Biol.* **2008** 162(2):205-18

PRESENTATIONS AND POSTERS

Jaffee, M.B. *Ethical dilemmas in biomedical sciences* Northeastern University, Invited lecture in Course ESC 153: Science, Technology, & Society. Boston, MA, February 14th 2013. (Oral)

Jaffee, M.B. and Imperiali, B. *N-linked glycosylation in bacteria: Tractable systems for biochemical & biophysical analysis*. MIT Annual Retreat for Building 68, Cape Cod, MA. June 11-12, 2012. (Poster); **and** MIT Protein Structure & Function Supergroup, Cambridge, MA, May 10th 2012. (Oral)

Jaffee, M.B. and Imperiali, B. *Investigating the mechanism of the oligosaccharyl transferase.* MIT Graduate Biology Journal Club, Cambridge, MA, August 24th 2009. (Oral)

Jaffee, M.B. *Biochemical techniques to investigate the mechanism of an enzyme*. Undergraduate Biological Chemistry Club of Wellesley College, Wellesley, MA, May 5th 2009. (Oral)

Felton, J., Jaffee, M., Reich, D. *Directed Cell Assembly with Magnetic Nanowires* American Physical Society March Meeting, Baltimore, MD, December 5th 2005. (Poster)