IMPROVEMENT OF INTEGRAL MEMBRANE PROTEIN EXPRESSION VIA OPTIMIZATION OF SIMULATED INTEGRATION EFFICIENCY

Thesis by

Stephen Sandell Marshall

In Partial Fulfillment of the Requirements for

the Degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2018

(Defended August 28, 2017)

Stephen Marshall
ORCID: 0000-0003-2263-6854

ACKNOWLEDGEMENTS

Outside of the Clemons lab, Dr. Khalid Salaita and the members of his group at Emory University, most notably Dr. Daniel Stabley, were highly influential in my training and my acceptance to Caltech. Dr. Tom Miller and Michiel Niesen have been amazing collaborators and my research would not be possible without their input. The work we have collectively produced represents the result of many hours of discussion and numerous rewrites. The members of my committee, Dr. Shu-ou Shan, Dr. Bil Clemons, Dr. Tom Miller, and Dr. Steve Mayo, have provided me with extremely valuable advice during my annual meetings. I appreciate them taking time out of their busy schedules to listen to my presentations.

Within the Clemons lab, I would especially like to thank Dr. Bil Clemons for providing me the opportunity to work in his lab and supporting me over the last five years. He has been both a mentor and friend, helping me develop as a researcher, writer, and speaker. Dr. Axel Muller was instrumental in establishing the subgroup I work in and teaching me during my first year. Shyam Saladi helped immensely with analyzing my data and writing code. All other current and former members of the Clemons lab over the years have contributed to my progress. Arun Chandra, Anthony Jones, and Omoshola Aleru worked with me during the summer quarters and were all exemplary volunteers with admirable work ethics and strong intellects.

In my personal life, both my immediate and new in-law families have been extremely supportive during my time at Caltech. My dog Charles has greeted me eagerly ever day as I come home. Most importantly, my new wife Jen has been with me throughout the entire journey, helping to preserve my sanity while pushing me to do my best as a

researcher and as a person. Our time together in Pasadena has been unforgettable and I look forward to our move to New York City and all else the future holds.

ABSTRACT

Integral membrane protein characterization is limited by the low levels of protein obtainable from heterologous overexpression in hosts such as *Escherichia coli*. Differences in the efficiencies of subdomains of the co-translational integration processes of membrane proteins into the membrane could explain the observed variation in the experimental expression of closely related homologs in *E. coli*. We have developed a method to predict and increase the expression of individual membrane proteins by optimizing the efficiency of their translocon-mediated integration into the membrane. The integration efficiency of each component of a membrane protein is calculated using a coarse-grained co-translational simulated integration model. The results of model simulations, experimental expression levels quantified by integral membrane protein-GFP fusion fluorescence, and a novel antibiotic survival test that reports on misintegration *in vivo* are applied to test the relationship between the integration efficiency of specific domains and experimental expression. Changes in simulated integration efficiencies due to sequence modifications agree with the effects on experimental expression *in vivo*. In the case of the TatC protein family, misintegration of the C-tail is found to be a major contributor to expression failure in *E. coli*. Beneficial sequence modifications that improve both simulated integration efficiency and experimental expression levels can be identified using the model. Preliminary evidence shows that simulated integration efficiency could potentially predict the effects of mutations on *Haemophilus influenzae* GlpG experimental expression in *E. coli*. The process described herein allows for the rational overexpression of integral membrane proteins through the identification and mitigation of inefficiencies in the underlying co-translational membrane integration process.

PUBLISHED CONTENT AND CONTRIBUTIONS

Stephen S. Marshall*, Michiel J. M. Niesen*, Axel Müller, Katrin Tiemann, Shyam M. Saladi, Rachel P. Galimidi, Bin Zhang, William M. Clemons, Jr., and Thomas F. Miller, III. *A Link between Integral Membrane Protein Expression and Simulated Integration Efficiency.* Cell Reports, 2016. **16**(8): p. 2169-2177. doi: 10.1016/j.celrep.2016.07.042

*Stephen S. Marshall and Michiel J.M. Niesen are co-first authors.

Stephen S. Marshall designed the research, performed expression and survival assay experiments in *E. coli* and *M. smegmatis*, performed data analysis, and wrote the paper.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Chapter 2

Chapter 3

Chapter 4

Chapter 1

INTRODUCTION

Integral Membrane Proteins Are Important Research Targets

Integral membrane proteins (IMP) act as key relays between the interior and exterior of a cellular or subcellular environment, facilitating the passage of information, cargo, and energy. They represent approximately 26% of human genes and 60% of current drug targets [1, 2]. Though IMPs are attractive targets for structural characterization, they represent only 2% of structures deposited in the PDB [3]. A major contributor to the limited number of membrane protein structures is the difficulty in obtaining sufficient amounts from heterologous overexpression.

There are two types of IMPs found in bacteria: alpha helical membrane proteins that reside in the inner membrane and beta barrel membrane proteins found in the outer membrane [4]. They have distinct structural motifs and biogenesis pathways and inhabit different subcellular environments. The majority of polytopic alpha helical membrane proteins integrate into the inner membrane with assistance of the translocon, while beta barrel proteins are translocated across the inner membrane through the translocon to be inserted in the outer membrane. Hereafter, IMPs will be used to indicate only polytopic alpha helical membrane proteins.

The domains that make up an IMP can be categorized as either loop or transmembrane domains (TMD). TMDs are alpha helices that reside within the membrane and are enriched in hydrophobic residues. In contrast, loop domains are found in the cytoplasmic or periplasmic space in bacteria and are more hydrophilic. The different

natures of these domains contribute to the delivery and final fold of the IMP through their interactions with processing machinery including the SRP and the translocon.

Membrane Protein Biogenesis in *Escherichia coli*

A key hypothesis made here is that the integration efficiency of an IMP directly affects its expression levels in *E. coli.* Experimental expression represents the amount of protein that is translated and properly folded within the inner membrane. IMP biogenesis requires the placement of the constituent loops and TMDs with the correct orientation relative to the membrane. Ribosomes translating IMPs in bacteria are targeted to the translocon in the inner membrane via interaction with the signal recognition particle (SRP) and its receptor (SR) [5, 6]. The SRP recognizes a hydrophobic signal sequence on a nascent IMP early in its translation. Contact of an SRP with the SR leads to the handoff of the ribosome-nascent chain complex to the translocon. At the core of the translocon is the SecYEG complex, a channel with a unique structure, containing both a pore that can allow passage of substrates across the membrane and a lateral gate that can open to allow cargo within the channel to directly interact with and insert into the membrane [7-10]. The translocon aids in the integration of the TMDs into the membrane by the opening of its lateral gate when a TMD is within the channel, facilitating passage of the TMD into the membrane. Loop domains are either translocated through the channel into the periplasm or retained in the cytoplasm by passing through a space between the translocon and the ribosome.

It is important to note that the targeting and integration of IMPs usually occurs co-translationally; only a portion of the polypeptide is exposed beyond the exit tunnel of the

ribosome to interact with the SRP and translocon [6, 7]. Therefore, it is sequentially early hydrophobic domains that interact with the SRP [5, 6]. As well, the translocon is proposed to integrate TMDs soon after their emergence from the exit tunnel such that TMDs are integrated into the membrane in their order on the primary sequence of the IMP (i.e. TMD 1 integrates first, TMD 2 integrates second, etc.) [10-12]. Notable exceptions to the co-translational and sequential model of IMP integration have been found, including examples of large-scale reorientation of IMP domains [13-15].

The establishment of the correct topology is important for the proper folding and function of an IMP integrated by the translocon and is influenced by the biophysical characteristics of the nascent chain. The topology of an IMP refers to the orientation of the TMD and loop domains with respect to the cytoplasm. For example, an IMP with three TMDs and four loops can be integrated with two orientations: the N-terminal loop in the cytoplasm and the C-terminal loop in the periplasm or the N-terminal loop in the periplasm and the C-terminal loop in the cytoplasm. Interactions between the nascent chain, the translocon, and the surrounding microenvironment are key to establishing topology. Two of the most prominent features of the nascent chain that contribute to topogenesis are the hydrophobicity of the TMDs and the distribution of positive charges on cytoplasmic and periplasmic loops. The hydrophobicity of a TMD correlates with its membrane insertion efficiency [10, 16]. Cytoplasmic loops are highly enriched in positively charged residues as compared to periplasmic loops [17]. Changing the placement of positively charges along the IMP sequence can affect the final topology of the IMP [14, 18, 19]. A failure to establish the correct topology due to low integration efficiency prevents proper folding and function of the IMP and leads to its degradation [20].

In the studies described here, the process of the orientation of a loop domain through interaction with the translocon to establish topology is referred to as its integration. Misintegration indicates improper retention or translocation of a loop, ending with its placement in the incorrect subcellular location. Integration efficiency of a loop or TMD represents the proportion of IMPs expressed with the domains in the correct location after interaction with the translocon has ended. Optimization of the integration efficiency through modification of loops or TMDs provides a method of improving expression of an initially inefficiently integrated IMP.

The Membrane Protein Expression Problem

Heterologous expression levels of IMPs in *E. coli* are often insufficient for characterization by structural or biochemical methods, requiring researchers studying a specific protein to test a number of homologs until one provides sufficient yields, often encountering widely different expression values even among closely related proteins with high sequence homology [21-23]. The hydrophobic stretches that make up the TMDs are unstable and prone to aggregation outside of the membrane from which they cannot be easily refolded, and overexpression of IMPs is often toxic and can inhibit cell replication and lead to less final cell mass [24]. These issues contribute to the significant time and resource costs associated with the study of IMPs [20].

One of the methods developed to increase the throughput of IMP expression quantification involves the addition of a C-terminal GFP to the IMP coding sequence [25, 26]. Fluorescence levels from IMP-GFP fusions have been found to correlate strongly with the level of folded protein available for purification [26-31]. Therefore, measuring the

fluorescence of GFP molecules in the whole cell, in the cell lysate, or on a band on an SDS-PAGE gel can be used to quantify expression without IMP purification. The use of C-terminal GFP fusions greatly increases the number of IMPs that can be tested for expression yield.

There is no universally successful method for improving IMP expression and many of the strategies do not identify the cellular mechanism of the initial expression failure. An analysis of a large-scale expression trial of *Escherichia coli* IMP-GFP fusions was unable to identify a single feature that significantly correlated with expression yield [31]. Strategies for improving IMP expression in *E. coli* can be classified as either, (*a*), a modification of the IMP nucleotide and/or amino acid sequence to optimize its processing by the cellular machinery, or, (*b*), a change to the organism in which the IMP is overexpressed to improve the efficiency of the biogenesis pathway. In regard to (*a*), performing error-prone PCR to find mutations that increase stability and yield have been applied with some success [32, 33]. In the case of (*b*), one of the most successful and widely adapted methods for improving IMP expression involves reducing the amount of mRNA of the non-native IMP produced in the cell, which suggests that higher levels in some way overload the normal capacity for producing IMPs [34, 35]. Given the complex biogenesis process IMPs must undergo that is not required for soluble protein expression, inefficiency within the pathway is a likely contributor to poor expression. Both of these methods do not identify the source of expression failure. Researchers would benefit from an approach that would allow for improvement in expression through the understanding and improvement of the underlying suboptimal processes that lead to poor expression.

A coarse-grained molecular dynamics simulation model of translocon-facilitated integration of IMPs (CG model) provides a view of IMP integration and can be used to identify the source of expression failure [36]. The CG model is derived from over 16 μs of molecular dynamics simulations of the translocon, membrane bilayer, and a substrate sequence. The simplified, coarse-grained nature of the model allows for a large number of simulations over biologically relevant timescales to assess the efficiency of the integration process for an IMP and the domains thereof. The proportion of CG model trajectories that terminate with a domain in the correct topological location provides a measure of the efficiency with which it is oriented by the translocon with the correct final topology (simulated integration efficiency). Use of the CG model allows for understanding the underlying mechanism that leads to observed experimental results due to interactions of the nascent IMP with the membrane-translocon-ribosome environment.

The concept explored in later chapters is that the amount of folded, recoverable protein following the overexpression of an IMP in *E. coli* is directly influenced by the integration efficiency of individual domains and can be predicted by simulated integration efficiency calculated using the CG model. In Chapter 2, expression tests of TatC sequences with mutations limited to the C-tail demonstrate that the integration efficiency of the C-terminal loop (C-tail) is a key predictor of experimental expression as misintegration of the C-tail contributes to the poor expression of some TatCs, confirmed using an *in vivo* ampicillin assay. The effect of C-tail sequence changes on experimentally observed expression levels strongly correlates with the simulated integration efficiency obtained from the CG model. Likewise, mutations that improve the simulated integration efficiency increase the experimentally observed expression levels. In Chapter 3, the concept was

expanded to attempt the prediction of the effect of sequence modifications of any loop of TatC. Simulated integration efficiencies, calculated using a coarse-grained simulation approach, robustly predict expression for a set of 140 sequence modifications on TatC homologs, including loop-swap chimeras and single-residue mutations distributed over much of the protein sequence. The simulated integration efficiency and experimental expression of double-loop-swap chimeras is shown to be multiplicative and largely independent with respect to the component single-swap mutations. The evidence again indicates misintegration of the TatC C-tail is a factor in cases poor expression and mutation of the IMP sequence far from the C-terminus can improve the integration efficiency of the C-tail. In Chapter 4, a library of mutated *Haemophilus influenzae* (*Hi*) GlpG sequences is created using error-prone PCR and tested for experimental expression levels. *Hi*GlpG contains loop domains with simulated integration efficiencies that could be predictive of experimental expression improvement. The combined studies demonstrate that experimental expression of IMPs can be improved by using sequence modification to manipulate the integration efficiency of specific subdomains and these effects can be predicted by analyzing CG model simulations of the co-translational integration process. Future development and testing of the strategy will aim to broaden the applicability and simplify the use of the model on new sequence spaces.

C h a p t e r 2

A LINK BETWEEN INTEGRAL MEMBRANE PROTEIN EXPRESSION AND SIMULATED INTEGRATION EFFICIENCY FOR TATC C-TAIL MUTANTS

Adapted from Stephen S. Marshall*, Michiel J. M. Niesen*, Axel Müller, Katrin Tiemann, Shyam M. Saladi, Rachel P. Galimidi, Bin Zhang, William M. Clemons, Jr., and Thomas F. Miller, III. *A Link between Integral Membrane Protein Expression and Simulated Integration Efficiency*. Cell Reports, 2016. **16**(8): p. 2169-2177. doi: 10.1016/j.celrep.2016.07.042

*Stephen S. Marshall and Michiel J.M. Niesen are co-first authors.

Abstract

Integral membrane proteins control the flow of information and nutrients across cell membranes, yet mechanistic studies of membrane proteins are hindered by difficulties in expression. We investigate this issue by addressing the connection between membrane protein sequence and observed expression levels. For homologs of the integral membrane protein TatC, observed expression levels vary widely and are affected by small changes in protein sequence. The effect of sequence changes on experimentally observed expression levels strongly correlates with the simulated integration efficiency obtained from coarse-grained modeling, which is directly confirmed using an *in vivo* assay. Furthermore, mutations that improve the simulated integration efficiency likewise increase the experimentally observed expression levels. Demonstration of these trends in both *Escherichia coli* and *Mycobacterium smegmatis* suggests that the results are general to other expression systems. This work suggests that integral membrane protein integration is a determinant for successful expression, raising the possibility of controlling expression via rational design.

Introduction

The central role of integral membrane proteins (IMPs) in many biological functions motivates structural and biophysical studies that require large amounts of purified protein, often at considerable costs in terms of both materials and labor. A key obstacle is that only a small percentage of IMPs can be overexpressed (i.e., heterologously produced at levels conducive to further study) [23]. While extensive efforts have shown promising results for individual IMPs, including those focusing on expression conditions, host modification, and directed evolution [35, 37, 38], none of these has proven broadly applicable, even among homologs of a given IMP. In general, the determinants for IMP expression are poorly understood, leading to the prevailing opinion that problems in membrane protein expression must be addressed on a case-by-case basis.

Closely related IMP homologs can vary dramatically in the amount of protein available after expression [23], which raises a fundamental question: what differentiates the expression of IMP homologs? The hypothesis raised here is that the efficiency with which an IMP is integrated into the membrane is a key determinant in the degree of observed IMP expression.

A fundamental step in the biosynthesis of most IMPs involves their targeting to and integration into the membrane via the Sec protein translocation channel [7]. Integration of IMP transmembrane domains (TMDs) into the membrane is facilitated primarily through interaction between the nascent chain and SecY, which forms the core of the protein translocation complex, or translocon. Following the co-translational or post-translational insertion of nascent protein sequences into the translocon channel, hydrophobic segments pass through the lateral gate of SecY into the membrane to form TMDs. Factors such as

TMD hydrophobicity [10, 16] and loop charge [17, 39] have been shown to affect the efficiency of TMD integration and topogenesis. For example, TMD hydrophobicity is directly related to the probability with which TMDs partition into the lipid bilayer, while positively charged residues in the loop alter TMD orientation by preferentially occupying the cytosol [10, 17, 39].

In this study, we investigated the connection between observed IMP expression levels and Sec-facilitated IMP integration efficiency (i.e., the probability of membrane integration with the correct multi-spanning topology). Systematic investigation of chimeras within an IMP family led to the identification of sequence elements that modulate expression levels. *In silico* modeling of IMP integration at the Sec translocation channel found that the sequence modifications that increase the calculated IMP integration efficiency correlate with *in vivo* overexpression improvements, suggesting that IMP integration efficiency is a determinant for successful expression. The result was found to be general across distinct expression systems (*E. coli* and *M. smegmatis*). Furthermore, an *in vivo* assay based on antibiotic resistance in *E. coli* experimentally confirmed the model that the integration efficiency of an individual TMD correlates with the observed IMP expression levels. The strong link between the effects of sequence modifications on simulated integration efficiency and experimentally measured expression levels offers future promise for the rational design of IMP systems with increased expression levels.

Results

As a detailed case study, the TatC IMP family was employed for all experimental and computational results reported here. A component of the bacterial twin-arginine translocation pathway, TatC plays a key role in the transport of folded proteins across the cytoplasmic membrane [40]. The employment of TatC was well suited for this study as it is reasonably sized (only six TMDs; Figure 2.1A), non-essential, and found broadly throughout bacteria; furthermore, TatC homologs previously have been observed to exhibit widely varying expression levels in *E. coli* [41], suggesting the importance of sequence-level details in the expression of this IMP.

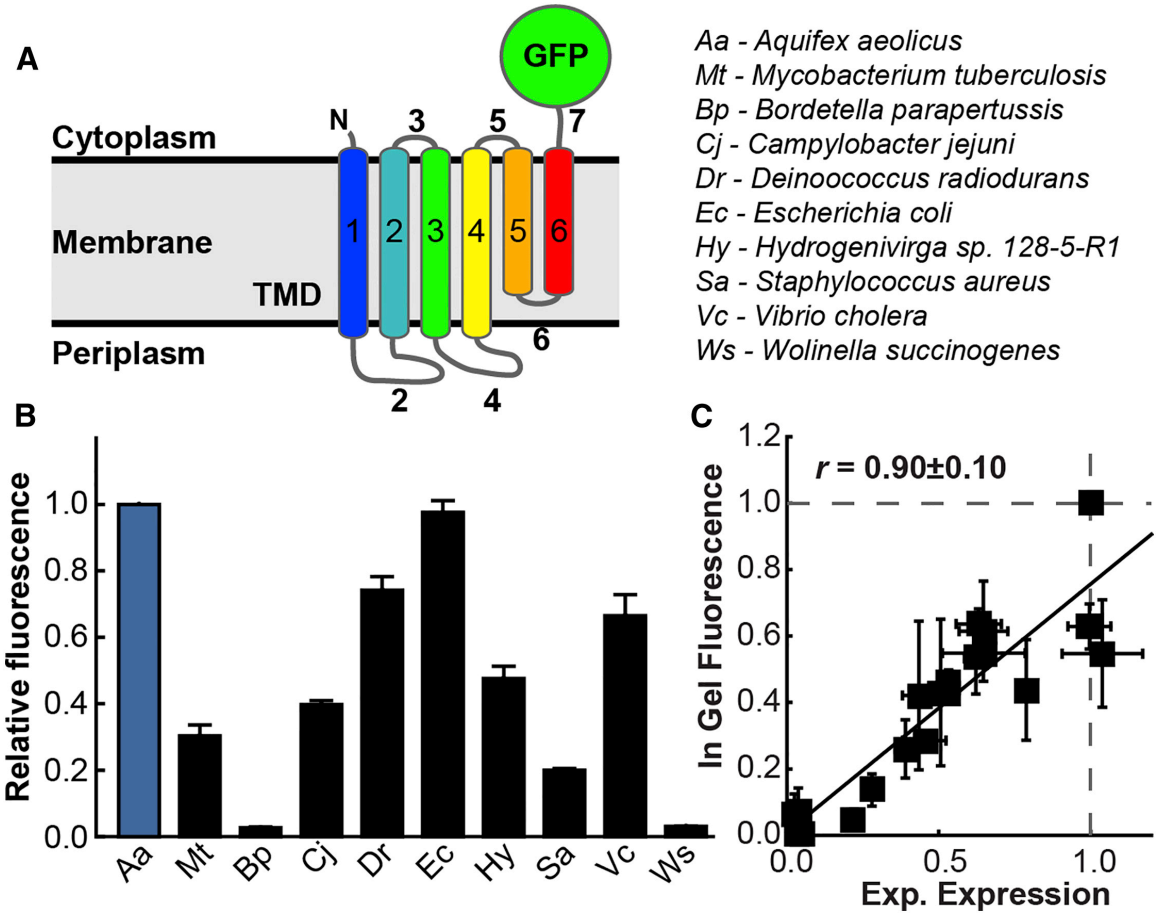
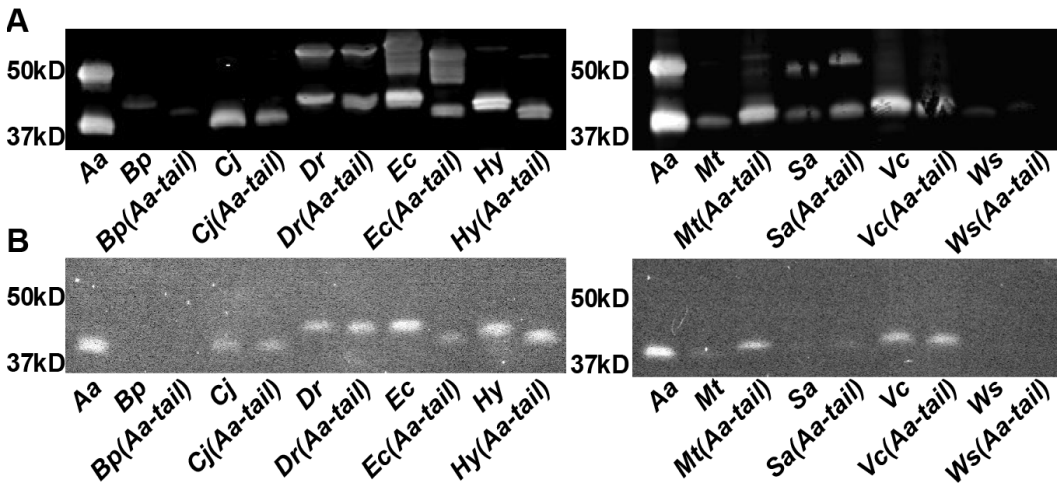


Figure 2.1: Variation in the Expression of TatC Homologs in E. coli. (A) A topology representation of TatC with a GFP C-terminal tag, as used in the expression studies. TMDs and loops are indicated in colors and gray, respectively, and are numbered. (B) Expression levels of various TatC homologs in *E. coli*, measured by TatC-GFP fluorescence, with

expression levels normalized to *Aa*TatC (blue). Error bars indicate the standard errors of the mean.

Wild-Type and Chimeric TatC Expression in E. coli

We first demonstrated that homologs of the IMP TatC exhibit large variance in observed expression levels in *E. coli*. For a quantitative measure of IMP expression, we employed a C-terminal fusion tag of a GFP variant [42] (Figure 2.1A) and measured whole-cell fluorescence by flow cytometry. Whole-cell fluorescence intensity of this fusion tag has been validated in numerous previous studies to correlate strongly with the amount of folded IMP, rather than the total level of IMP translated [26-31]. We further validated the expression levels measured from whole-cell fluorescence (Figure 2.1B) using in-gel fluorescence (Figures 2.1C, 2.2B, and 2.2C; Pearson correlation coefficient, r = 0.9) and western blot analysis (Figure 2.2A). With this approach, expression levels in *E. coli* were experimentally measured for TatC homologs from a variety of bacteria, including *Aquifex aeolicus* (*Aa*), *Bordetella parapertussis* (*Bp*), *Campylobacter jejuni* (*Cj*), *Deinococcus radiodurans* (*Dr*), *Escherichia coli* (*Ec*), *Hydrogenivirga species 128-5-R1* (*Hy*), *Mycobacterium tuberculosis* (*Mt*), *Staphylococcus aureus* (*Sa*), *Vibrio cholera* (*Vc*), and *Wolinella succinogenes* (*Ws*).



C

In Gel Fluorescence

Aa	1	*Hy*	0.4(2)
Bp	0.06(6)	*Hy*(*Aa*-tail)	0.4(2)
Bp(*Aa*-tail)	0.07(7)	*Mt*	0.29(3)
Cj	0.43(3)	*Mt*(*Aa*-tail)	0.528(4)
Cj(*Aa*-tail)	0.46(3)	*Sa*	0.05(3)
Dr	0.5(1)	*Sa*(*Aa*-tail)	0.13(5)
Dr(*Aa*-tail)	0.5(2)	*Vc*	0.6(1)
Ec	0.63(7)	*Vc*(*Aa*-tail)	0.63(5)
Ec(*Aa*-tail)	0.4(2)	*Ws*	n.d.
		Ws(*Aa*-tail)	n.d.

Figure 2.2: Validation of Expression of TatC Variants in E. coli. (A) Anti-GFP western blot results for TatC homologs and the corresponding *Aa*-tail swap chimeras. Two bands were observed for all lanes where TatC-GFP was at high relative concentrations with the lower bands active by in-gel fluorescence and therefore determined to be folded protein. (Waldo et al., 1999) (B) In-gel fluorescence of SDS-PAGE for TatC homologs and the corresponding *Aa*-tail swap chimeras. Bands that exhibit fluorescence represent folded protein. The results exhibit the same trends in expression yield as seen by flow-cytometry. (C) Average in-gel fluorescence quantified across four separate gels. *Ws* and *Ws*(*Aa*-tail) could not be detected (n.d.) by in-gel fluorescence. Values for each band are normalized to the *Aa*TatC band and values in parentheses indicate the standard error of mean.

Figure 2.1B shows the wide range of expression levels that are exhibited by the TatC homologs in *E. coli*. Previous expression trials of TatC homologs identified that *Aa*TatC is readily produced at high levels in *E. coli,* which enabled the solution of its structure [41, 43]. In contrast, low expression is found for both the *Mt*TatC, hereafter

referred to as *Mt*TatC(Wt-tail), and a modified sequence truncating the un-conserved 38-residue sequence of the C-terminal loop, hereafter referred to as *Mt*TatC [41].

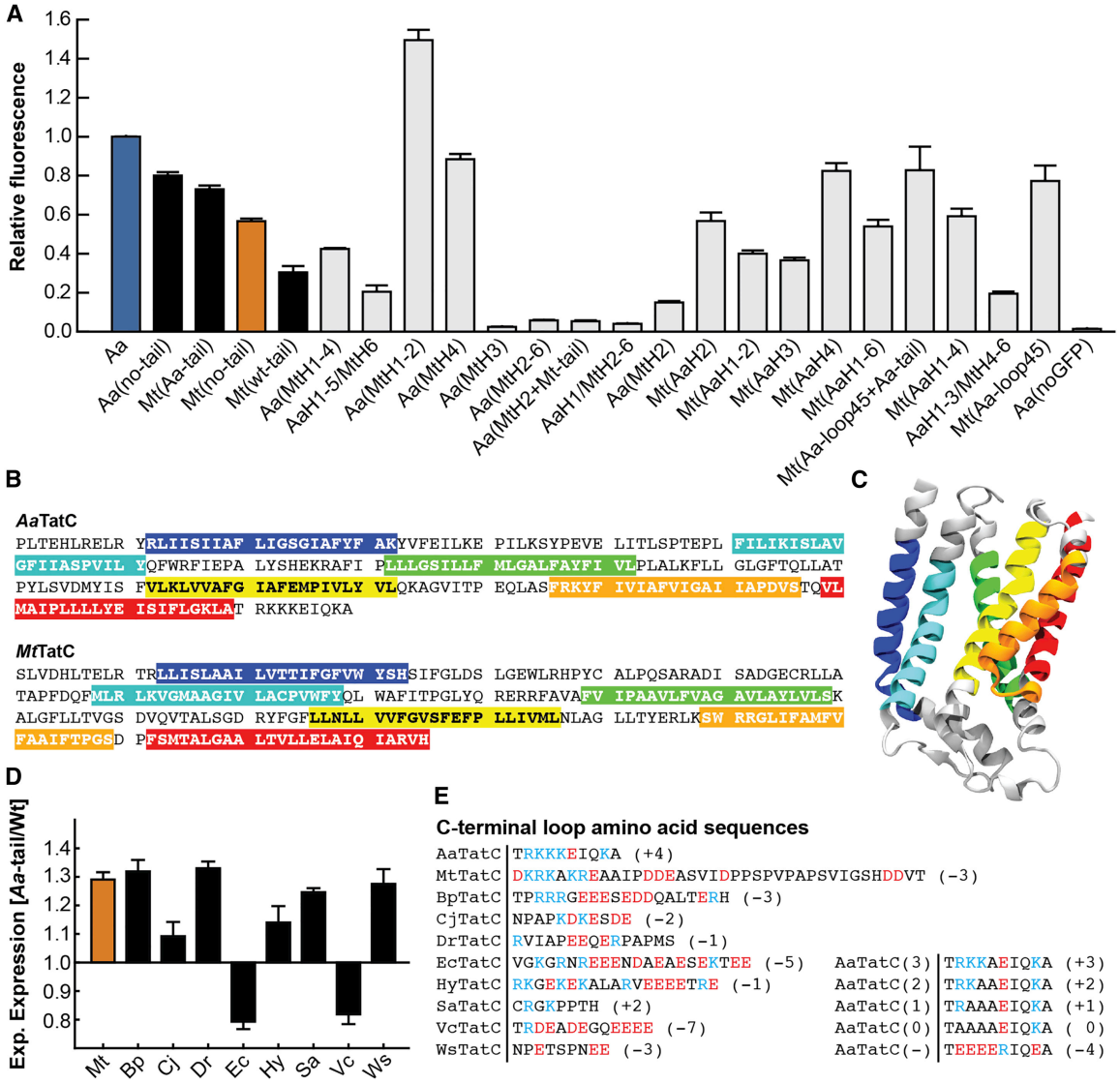


Figure 2.3: Effect of the C-tail on TatC Expression in E. coli. (A) Measured expression levels of the *Aa*TatC and *Mt*TatC chimera proteins, normalized to *Aa*TatC. Shaded bars represent wild-type TatC homologs and mutants with C-tail modifications. (B) Domain definitions used in generating the swap chimeras, with TMDs highlighted, are shown. (C) A ribbons diagram of the structure of *Aa*TatC (RCSB PDB: 4HTS). TMDs are colored according to the highlights used in (B). (D) For each homolog, the ratio of the measured expression level for the *Aa*-tail chimera to that of the corresponding wild-type sequence is shown. (E) TatC wild-type and charge mutant C-tail sequences. Positive residues are in blue and negative residues are in red. The net charge is shown to the right of each sequence. Error bars indicate the standard errors of the mean.

To examine the parts of the protein sequence that affect expression, swap chimeras were generated by exchanging entire loops and TMDs between *Aa*TatC and *Mt*TatC. The TMDs and loops were defined by comparing sequence alignments and membrane topology predictions (Figure 2.3B) [44, 45]. The swap chimeras exhibited a wide range of expression results (Figure 2.3A). The C-terminal loop sequence, referred to as the C-tail and labeled as loop 7 in Figure 2.1A, was found to have a significant effect on expression levels (shaded bars in Figure 2.3A). Removal of the *Mt*TatC C-tail improved expression. Removal of the C-tail from the *Aa*TatC sequence led to a corresponding decrease in expression. Strikingly, swapping the *Aa*TatC C-tail (*Aa*-tail) into the *Mt*TatC sequence led to a significant improvement in expression.

The positive effect of the *Aa*-tail on *Mt*TatC expression raises the question of whether expression can be similarly improved in other TatC homologs by substituting the corresponding C-tail sequence (Figure 2.3E) with that of *Aa*TatC. Swapping the C-tail of the various TatC homologs with the *Aa*-tail improved expression in seven of nine cases (Figure 2.3D). Taken together, the results in Figure 2.3 indicate that the C-tail is a significant factor in determining TatC expression across homologs.

In Silico *Modeling of TatC Integration*

To investigate the mechanistic basis for the experimentally observed effect of the C-tail on expression, we employed a recently developed *in silico* coarse-grained (CG) approach that models co-translational translocation on unbiased biological timescales [36]. The CG model, which is derived from >16 μs of molecular dynamics simulations of the Sec translocation channel, the membrane bilayer, and protein substrates [46, 47], has been

validated for the description of Sec-facilitated membrane integration, including experimentally observed effects of amino acid sequence on the membrane topology of single-spanning IMPs [36] and multi-spanning dual-topology proteins [18]. IMP sequences were mapped onto a Brownian dynamics model of the ribosome/translocation channel/nascent protein system, and the Sec translocon-facilitated integration of the IMP into the lipid bilayer was directly simulated in 1,200 independent minute-timescale trajectories for each TatC (Figure 2.4; Figure 2.5A). This implementation of the CG model did not distinguish between expression systems.

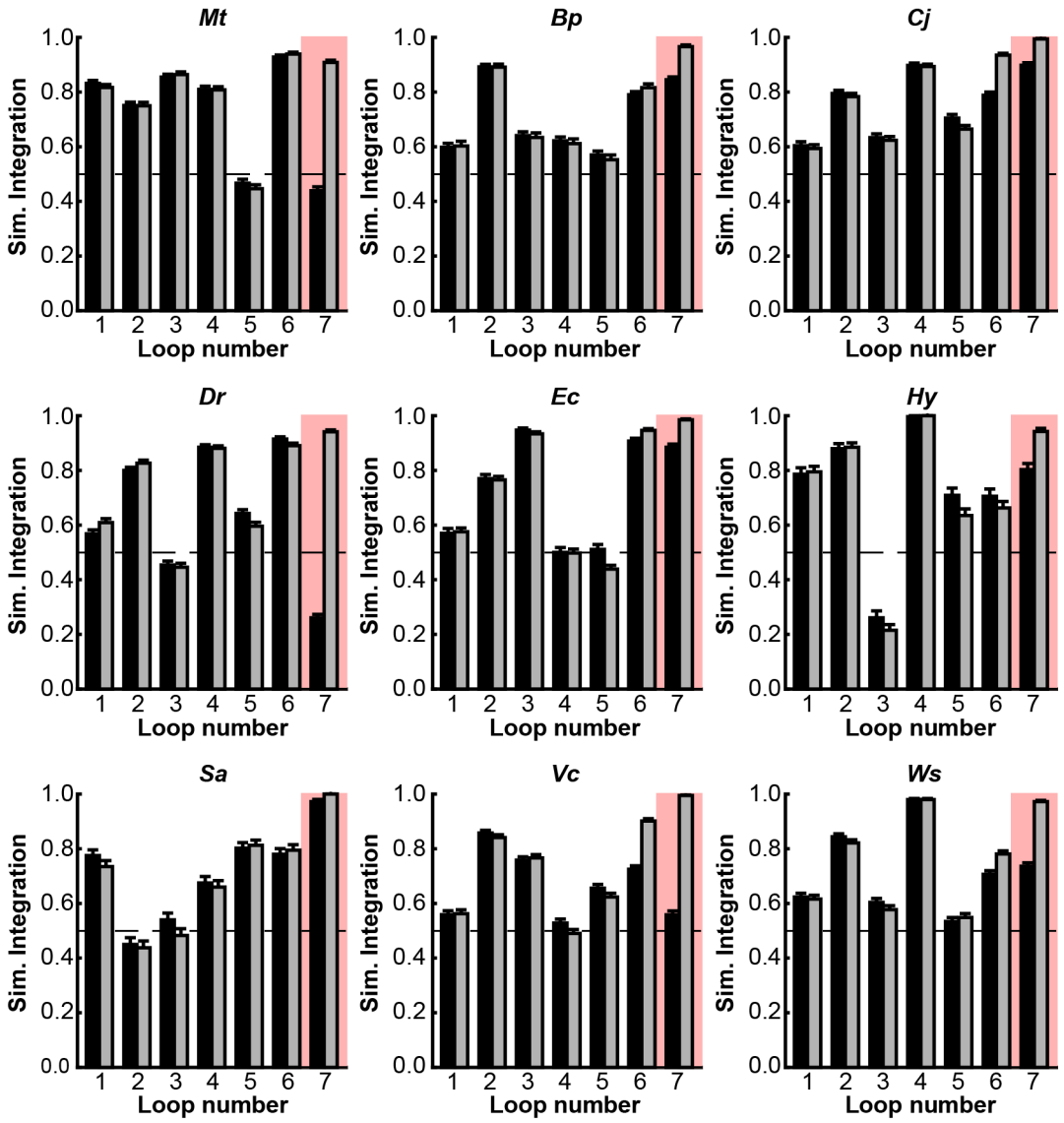
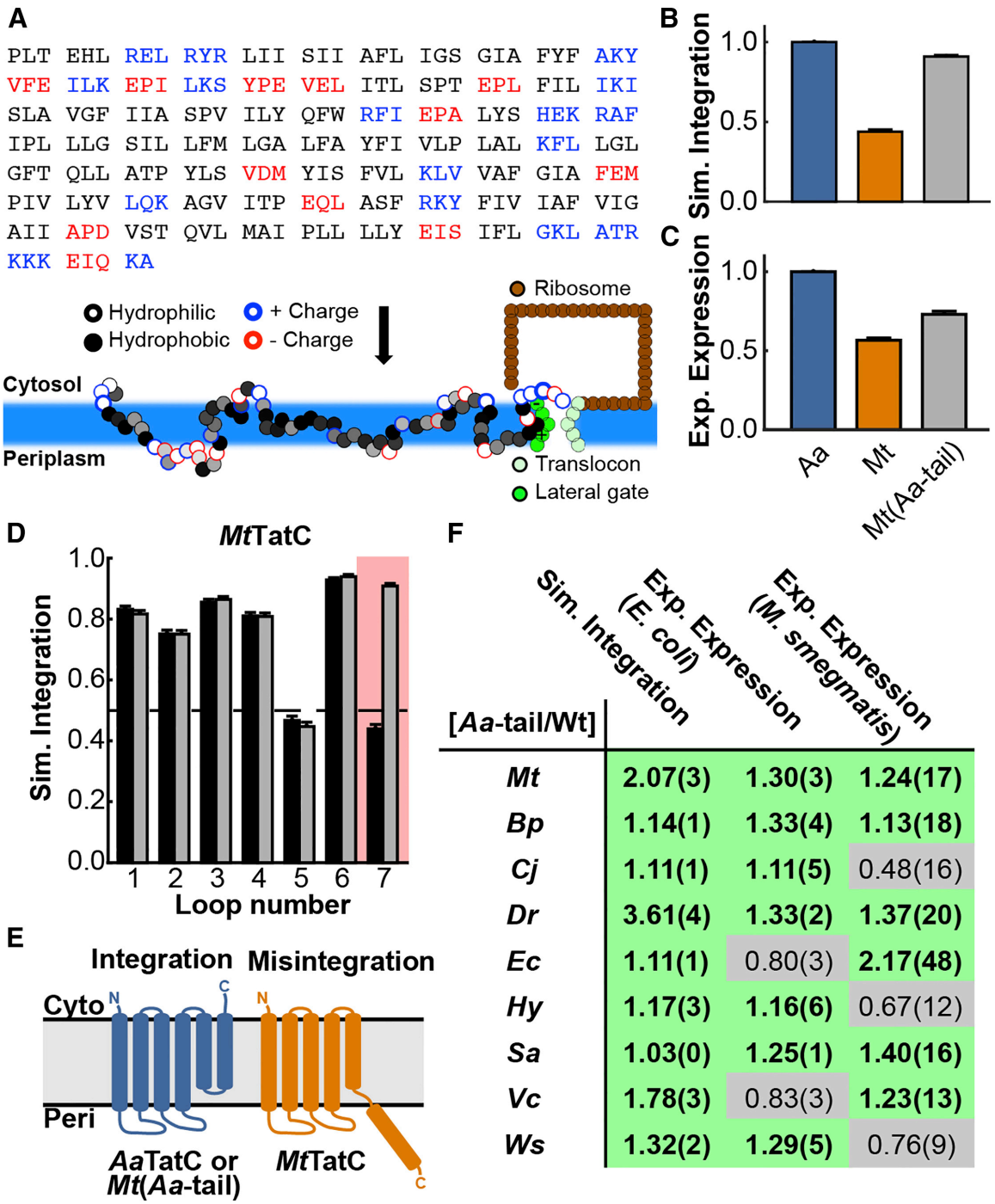


Figure 2.4: Simulated Integration Efficiencies Among All Loops TatC Wild-types and Aa-tail Chimeras. For each considered TatC homolog, the simulated integration efficiency for the individual loops for both the wild-type sequence (black bars) and the *Aa*-tail chimeras (grey bars). It is seen that the *Aa*-tail generally leads to a significant effect on the integration efficiency of loop 7 (highlighted), with smaller effects on the other loops. Error bars indicate the standard error of mean.

Using the results of the CG model, Figure 2.5B presents the simulated integration efficiency, defined to be the fraction of trajectories that led to the correct membrane topology, for several TatC sequences. Unless otherwise specified, we defined membrane

topology in terms of the final orientation of the C-tail. The *Aa*TatC homolog exhibited significantly higher simulated integration efficiency than the *Mt*TatC homolog, which is consistent with the relative experimental expression levels for the two homologs in Figure 2.5C. Figure 2.5B shows that the *Mt*(*Aa*-tail) chimera recovered the high levels of simulated integration efficiency seen for the *Aa*TatC homolog, further mirroring the experimental trends in IMP expression (Figure 2.5C). Figure 2.5D presents an analysis of the orientation of each loop, indicating that only loop 7 was significantly affected swapping the C-tail in the simulations. As is shown schematically in Figure 2.5E, the simulations found that *Mt*TatC exhibits a large fraction of trajectories in which the C-tail resides in the periplasm, such that the C-terminal TMD (TMD 6) fails to correctly integrate into the membrane.



[*Aa*-tail/Wt]	Sim. Integration	Exp. Expression (*E. coli*)	Exp. Expression (*M. smegmatis*)
Mt	**2.07(3)**	**1.30(3)**	**1.24(17)**
Bp	**1.14(1)**	**1.33(4)**	**1.13(18)**
Cj	**1.11(1)**	**1.11(5)**	0.48(16)
Dr	**3.61(4)**	**1.33(2)**	**1.37(20)**
Ec	**1.11(1)**	0.80(3)	**2.17(48)**
Hy	**1.17(3)**	**1.16(6)**	0.67(12)
Sa	**1.03(0)**	**1.25(1)**	**1.40(16)**
Vc	**1.78(3)**	0.83(3)	**1.23(13)**
Ws	**1.32(2)**	**1.29(5)**	0.76(9)

Figure 2.5: Calculation of TatC Integration Efficiencies. (A) Schematic illustration of the CG simulation model that is used to model co-translational IMP membrane integration. The amino acid sequence of the IMP is mapped onto CG beads, with each consecutive trio of amino acid residues in the nascent protein sequence mapped to an associated CG bead; the underlying properties of the amino acid residues determine the interactions of the CG beads, as described in the text. (B) Simulated integration efficiency of the *Aa*TatC, *Mt*TatC, and *Mt*(*Aa*-tail) sequences is shown. Error bars indicate the standard errors of the mean. (C) Experimental expression of the *Aa*TatC, *Mt*TatC, and *Mt*(*Aa*-tail) sequences is shown. Error bars indicate the standard errors of the mean. (D) The simulated integration efficiency

for individual loops of both the wild-type *Mt*TatC sequence (black bars) and the *Aa*-tail swap chimera (gray bars), with loop 7 highlighted, is shown. Error bars indicate the standard errors of the mean. (E) Schematic of the correct and incorrect TatC topologies observed in the simulations. Misintegration of loop 7 and translocation of TMD 6 lead to an incorrect final topology for *Mt*TatC. (F) For each homolog, comparison between the experimental expression levels in *E. coli* and *M. smegmatis* and the simulated integration efficiencies, reporting the ratio of the *Aa*-tail chimera result to that of the corresponding wild-type sequence. Ratios exceeding unity are highlighted in green, indicating enhancement due to the *Aa*-tail. Values in parentheses indicate the standard errors of the mean.

Additional simulations were performed for the full set of the experimentally characterized TatC homologs (Figures 2.4 and 2.5F), allowing comparison of the computationally predicted shifts in IMP integration with those observed experimentally for IMP expression. For each homolog, Figure 2.5F compares the effect of swapping the wild-type C-tail with the *Aa*-tail on both the experimental expression level and the simulated integration efficiency. With the exception of *Vc*TatC and *Ec*TatC, Figure 2.5F shows consistent agreement between the computational and experimental results in *E. coli* upon introducing the *Aa*-tail.

Confirmation of the Predicted Mechanism Using a Translocation Assay

The comparison between simulation and experiment in the previous sections suggests a mechanism in which translocation of the C-tail of TatC into the periplasm leads to a reduction in the observed expression level. To validate this, an experimental *in vivo* assay based on antibiotic resistance in *E. coli* was employed. The C-terminal GFP tag was replaced by β-lactamase, such that an incorrectly oriented C-tail would confer increased resistance to β-lactam antibiotics (Figure 2.6A); an inverse correlation between antibiotic resistance and GFP fluorescence was thus expected. *Aa*TatC, *Mt*, and *Mt*(*Aa*-tail)

constructs containing the β-lactamase tag were expressed using the same protocol as before. Following expression, the cells were diluted to an optical density 600 (OD600) of 0.1 in fresh media without inducing agent, and they were grown to an OD600 of ~0.5 at which point ampicillin was added. Then 1.5 hours after ampicillin treatment, equal amounts of the media were plated on Luria-Bertani (LB) agar plates without ampicillin (Figure 2.6B). The number of observed colonies was used to quantify the relative cell survival (Figure 2.6C, bottom). The survival rate of *Mt*(*Aa*-tail), *Mt*, and *Aa*TatC inversely correlated with the simulated integration efficiency of the C-tail (Figure 2.6C), validating the proposed mechanism.

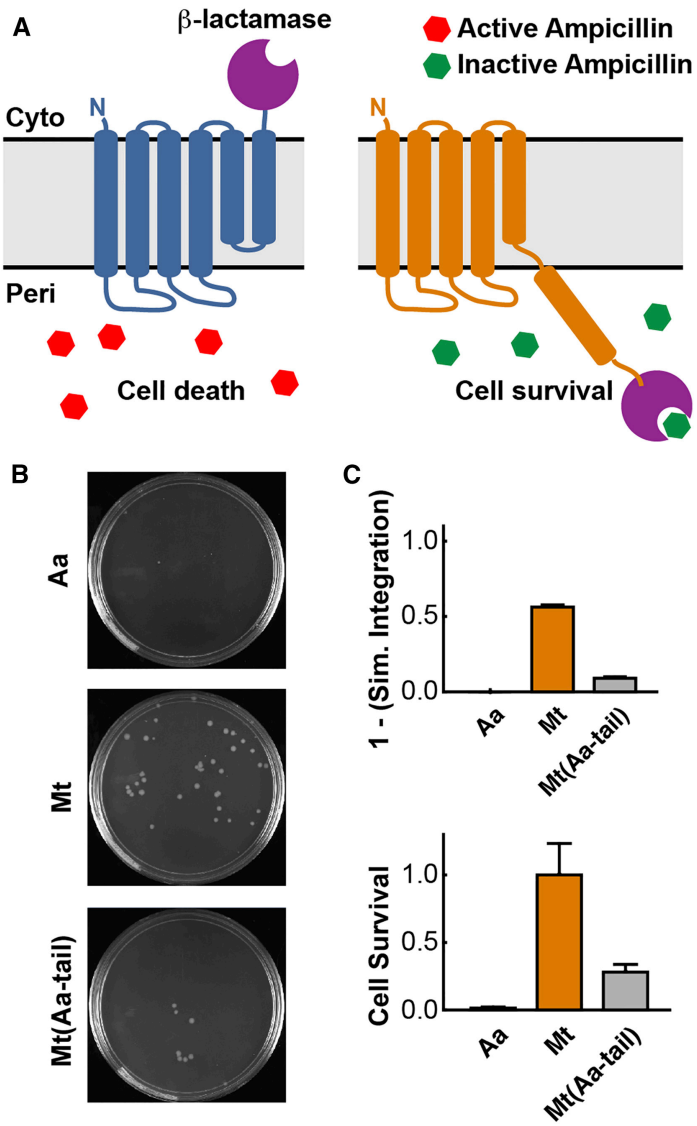


Figure 2.6: Correlation of Antibiotic Resistance to Membrane Topology. (A) Schematic of the cytoplasmic and periplasmic topologies of the TatC C-tail with the fused β-lactamase enzyme. Misintegration of loop 7 leads to periplasmic localization of the β-lactamase, resulting in enhanced antibiotic resistance and cell survival. (B) Representative plates from the ampicillin survival test are shown. (C) Comparison of the simulated integration efficiency (top) and relative ampicillin survival rate (bottom) for *Aa*TatC, *Mt*TatC, and *Mt*(*Aa*-tail). The reported cell survival corresponds to the ratio of counted cells post-treatment versus prior to treatment with ampicillin; all values are reported relative to *Mt*TatC. Error bars indicate the standard errors of the mean.

Tail Charge as an Expression Determinant: Experimental Tests of Computational Predictions

To further establish the connection between the simulated integration efficiencies and the experimentally observed expression levels, we examined the effect of C-tail mutations. We focused on modifications of the C-tail amino acid sequences that involve the introduction or removal of charged residues, which are known to affect IMP topology and stop-transfer efficiency [14, 36, 39].

We began by investigating the generic effect of the C-tail charge magnitude on TatC-simulated integration efficiency. Figure 2.7A presents the results of CG simulations in which the magnitude of the charges on the C-tail of the *Mt*(*Aa*-tail) sequence were scaled by a multiplicative factor, χ, keeping all other aspects of the protein sequence unchanged. The simulations revealed that reducing the charge magnitude on the C-tail led to lower simulated integration efficiency.

To examine the corresponding effect of C-tail charge magnitude on expression levels, Figure 2.7B plots the ratio of experimentally observed expression for each wild-type homolog relative to its corresponding *Aa*-tail swap chimera versus the total charge magnitude on the wild-type C-tail. Without exception in these data, the expression of wild-type homologs with weakly charged C-tails (relative to the *Aa*-tail) was improved upon swapping with the *Aa*-tail, whereas the expression of homologs with strongly charged C-tails was reduced upon swapping with the *Aa*-tail (i.e., all data points in Figure 2.7B fall into the unshaded quadrants).

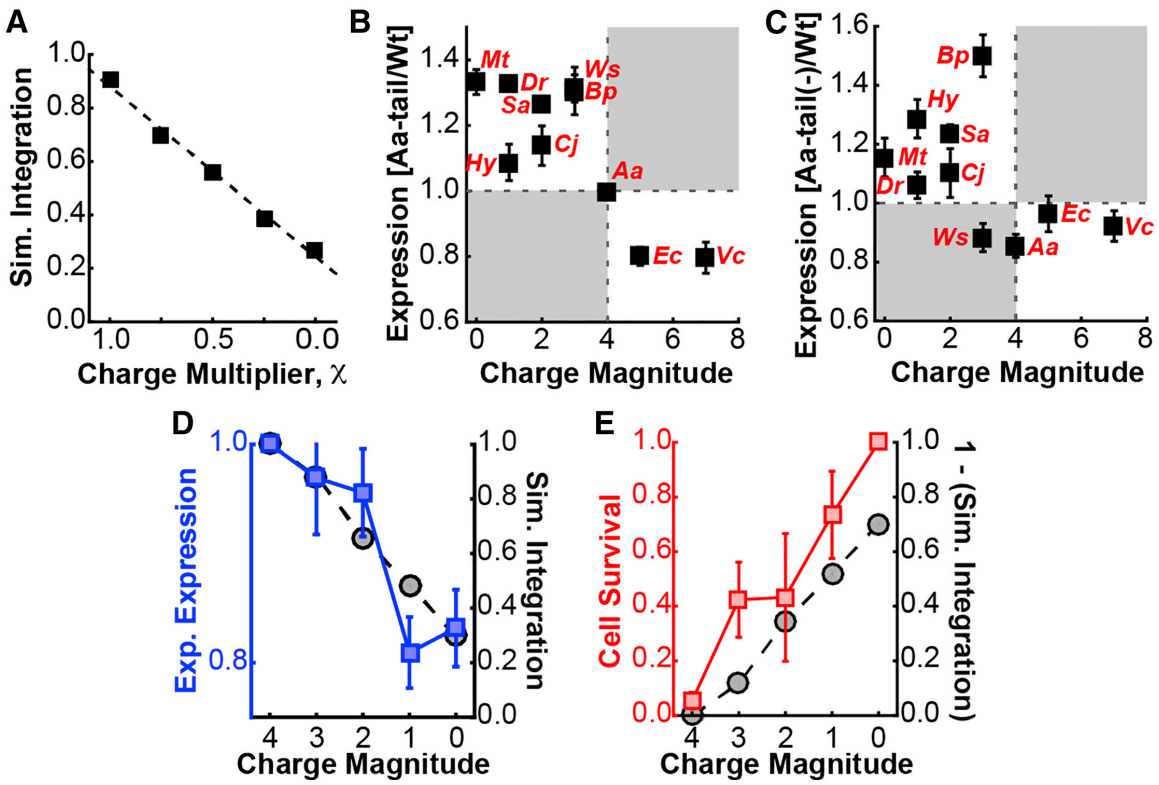


Figure 2.7 Mechanistic Basis Associated with Charged C-tail Residues. (A) Simulated integration efficiency of the *Mt*(*Aa*-tail) chimera, as a function of scaling the charges of the C-tail residues, is shown. (B) Correlation of the ratio of the measured expression for the *Aa*-tail swap chimeras to that of the corresponding wild-type sequence versus the charge magnitude of the wild-type C-tail (data from Figure 2.3E). (Pearson correlation coefficient of r = 0.8 ± 0.2) (C) Correlation of the ratio of the measured expression for the *Aa*-tail(−) swap chimeras to that of the corresponding wild-type sequence versus the charge magnitude of the wild-type C-tail, where the *Aa*-tail(−) swap chimeras include a variant of the *Aa*-tail with net negative charge and the same overall charge magnitude, is shown. (D) Experimental expression levels in *E. coli* (blue, left axis) and simulated integration efficiency (black, right axis) for a series of mutants of the *Mt*(*Aa*-tail) sequence, in which positively charged residues in the *Aa*-tail are mutated to alanine residues. Reported values are normalized to *Mt*(*Aa*-tail). (E) Relative ampicillin survival rate in *E. coli* (red, left axis) and simulated integration efficiency (black, right axis) for a series of mutants of the *Mt*(*Aa*-tail) sequence, in which positively charged residues in the *Aa*-tail are mutated to alanine residues. Simulation results are normalized as in (D), while ampicillin survival is normalized to the highest survival rate (i.e., with zero charge magnitude). Error bars indicate the standard errors of the mean.

Figure 2.7C further illustrates the effect of charge magnitude on expression by presenting the experimentally observed expression levels for *Aa*-tail(−) swap chimeras, in

which the introduced C-tail sequence preserved the charge magnitude of the *Aa*-tail sequence while reversing the net charge (see Figure 2.3E for the C-tail sequences). Despite the complete reversal of the C-tail charge, the observed correlation between expression and C-tail charge magnitude for these two sets of chimeras was strikingly similar (compare Figures 1.7B and 1.7C).

Finally, we considered a series of mutants of the *Mt*(*Aa*-tail) chimera, in which the charge magnitude of the *Aa*-tail was reduced by mutating positively charged residues to alanine residues (see Figure 2.3E for the C-tail sequences). For this series of mutants, Figure 2.7D (black) shows that the simulated integration efficiency decreased with the charge of the C-tail, which predicted a corresponding decrease in the experimental expression levels; indeed, the subsequent experimental measurements confirmed the predicted trend (Figure 2.7D, blue). Again using the antibiotic resistance assay to validate the connection between simulated integration efficiency and observed expression, Figure 2.7E confirms that the simulation results correlated with the relative survival of the *Mt*(*Aa*-tail) alanine mutants with a β-lactamase tag (Figure 2.7E, red). In addition to providing evidence for the connection between simulated integration efficiency and observed expression levels, the results in Figure 2.7 suggest that this link can be used to control IMP expression.

Transferability to M. smegmatis

Beyond the *E. coli* overexpression host, we examined the transferability of the relation between simulated integration efficiency and experimental expression levels. We employed *M. smegmatis*, a genetically tractable model organism that is phylogenetically

distinct from *E. coli*. All coding sequences were transferred into an inducible *M. smegmatis* vector, including the linker and C-terminal GFP, and expressed; expression levels were then measured by flow cytometry and validated by western blot.

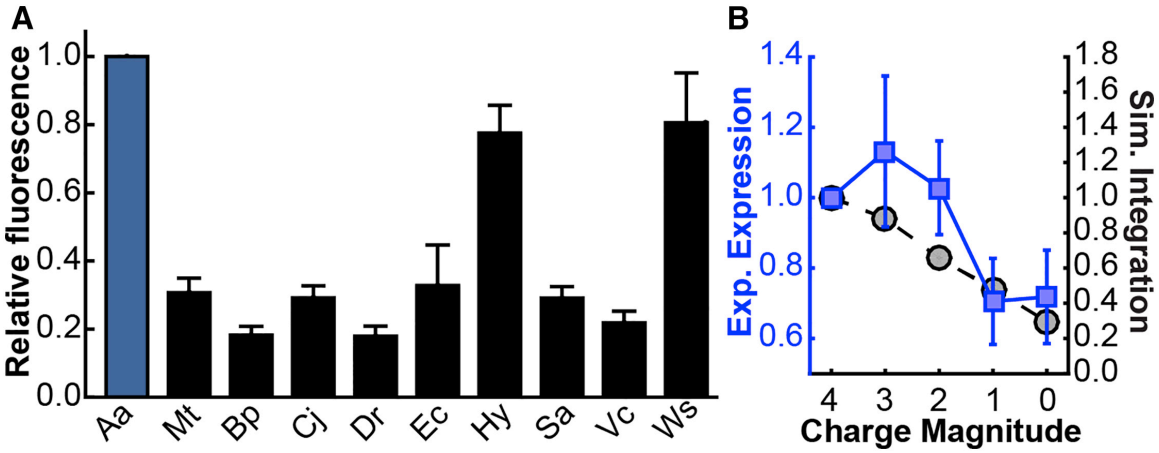


Figure 2.8: M. smegmatis *Expression Tests.* (A) Expression levels of various TatC homologs in *M. smegmatis* were measured by TatC-GFP fluorescence, with expression levels normalized to *Aa*TatC (blue). (B) Simulated integration efficiency (blue, left axis) and measured expression levels in *M. smegmatis* (black, right axis) for a series of mutants of the *Mt*(*Aa*-tail) sequence, in which positively charged residues in the *Aa*-tail are mutated to alanine residues. Error bars indicate the standard errors of the mean.

Figure 2.8A shows that, as in *E. coli*, the experimentally observed expression levels vary widely among the wild-type TatC homologs in *M. smegmatis*. However, comparison of Figure 2.8A with Figure 2.1B reveals that the total expression levels for the homologs in *M. smegmatis* are different from those seen in *E. coli*, although for both systems the *Aa*TatC homolog expresses strongly and *Mt*TatC expresses poorly (which is perhaps surprising, given the close evolutionary link between *M. smegmatis* and *M. tuberculosis*). Figure 2.5F also shows that replacing the wild-type C-tail with the *Aa*-tail in *M. smegmatis* generally increased the experimentally observed expression levels, in general agreement (six of nine homologs) with the previously discussed simulated integration efficiency results.

Figure 2.5F further shows that the subset of homologs, for which the *Aa*-tail swap chimeras led to increased levels of expression in *M. smegmatis*, was overlapping but different from the subset associated with the *E. coli* results. This emphasizes that, although the computed levels of simulated integration efficiency agree with the observed changes in expression levels in both expression systems, the observed expression levels depend on the expression system, while the simulated integration efficiencies calculated using the current implementation of the CG model are independent of the expression system. In short, simulated integration efficiency is a predictor of the expression levels in both systems, but it is not the only factor contributing to the observed expression levels.

Continuing with the *M. smegmatis* expression system, Figure 2.8B repeats the comparison between the simulated integration efficiency and the observed expression levels for the series of mutants of the *Mt*(*Aa*-tail) chimera, in which the positive charge of the *Aa*-tail was reduced by mutating positively charged residues to alanine residues. The simulated integration efficiencies, identical to those in Figure 2.7D, were predicted to decrease as charges were removed. The experimental expression levels for *M. smegmatis* in Figure 2.8B likewise showed a decrease. Taken together, the results obtained for the *M. smegmatis* expression system suggest that the connection between simulated integration efficiency and observed expression levels may be generalizable beyond *E. coli*.

Discussion

The mechanistic picture that emerges from the experimental and theoretical analysis of the TatC IMP family is that the efficiency of Sec-facilitated membrane integration, which is impacted by the IMP amino acid sequence, is a key determinant in the degree of observed protein expression. We observed that TatC homologs had varying levels of expression (Figures 2.1B and 2.8A). Swap chimeras between *Aa*TatC and *Mt*TatC revealed a significant effect of the C-tail in determining expression yields (Figure 2.3A), with the *Aa*-tail having a largely positive effect that was transferrable to other homologs (Figure 2.3F). CG modeling predicted a large, sequence-dependent variation of the simulated integration efficiency for the C-tail (Figure 2.4), suggesting the underlying mechanism by which the *Aa*-tail enhances the expression of other TatC homologs. Validation of this mechanism was experimentally demonstrated using an antibiotic resistance assay (Figure 2.6). Additional point-charge mutations in the C-tail were shown to change the simulated integration efficiency, which in turn predicted changes in the IMP expression levels according to the proposed mechanism; these predictions were experimentally confirmed in both *E. coli* (Figure 2.7) and *M. smegmatis* (Figure 2.8).

The observed correlation between IMP integration efficiency and observed expression levels presented here is consistent with earlier observations that expression can be modulated by mutations of the sequence [48-50], as well as recent work in which misintegrated dual-topology IMPs were shown to be degraded by FtsH [13], However, these earlier studies did not provide a clear mechanistic basis for the relation between IMP sequence modifications and observed expression levels. In the current work, we demonstrate the relation between integration efficiency and observed expression levels,

and we demonstrate a tractable CG approach for computing the simulated integration efficiency and its changes upon sequence modifications. This work also raises the possibility of using simulated integration efficiencies to optimize experimental expression levels, which has been demonstrated here via the computational prediction and subsequent experimental validation of individual charge mutations in the C-tail.

A few comments are worthwhile with regard to the scope of the conclusions drawn here. First, our study focused on comparing protein expression levels among IMP sequences that involve relatively localized changes, such as single mutations or loop-swap chimeras, as opposed to predicting relative expression levels among dramatically different IMP sequences. Second, our study examined experimental conditions for the overexpression of IMPs using the same plasmids, which may be expected to isolate the role of membrane integration in determining the relative expression levels of closely related IMP sequences. The prediction of expression levels among IMPs that involve more dramatic differences in sequence may well require the consideration of other factors, beyond just the simulated integration efficiency. Moving forward, we expect that a useful strategy will be to systematically combine the simulated IMP integration efficiency with other sequence-based properties to predict IMP expression levels [31].

The experimental and computational tools used here are readily applicable to many systems, potentially aiding the understanding and enhancement of IMP expression in many other systems, as well as providing fundamental tools for the investigation of co-translational IMP folding. By demonstrating inexpensive *in silico* methods for predicting protein expression, we note the potential for computationally guided protein expression strategies to significantly impact the isolation and characterization of many IMPs.

Acknowledgements

I would personally like to thank all of the authors for their contributions. The authors thank R.C. Van Lehn and D.C. Rees for comments on the manuscript and D. Daley for helpful discussion of Daley et al. (2005)). Work in the W.M.C. lab is supported by an NIH Pioneer Award to W.M.C. (5DP1GM105385) and an NIH training grant to S.S.M. (NIH/National Research Service Award (NRSA) training grant 5T32GM07616). Work in the T.F.M. group is supported in part by the Office of Naval Research (N00014-10-1-0884), and computational resources were provided by the National Energy Research Scientific Computing Center (NERSC), a (Department of Energy) DOE Office of Science User Facility (DE-AC02-05CH11231).

Methods

Designing and Cloning of TatC Chimeras

The parent plasmid used for cloning, pET28(a+)-GFP-ccdB, was derived from an IMP-GFP vector used by [51]. TatC homologs and chimeras were prepared from genomic DNA, with the exception of wild-type *M. tuberculosis* and *A. aeolicus* TatC genes that were synthesized by primer extension as applied in DNAWorks (NIH) [52]. In most cases, the Gibson assembly cloning protocol was used for cloning [53]. Expression of a vector containing *Aa*TatC with an N-terminal ten-His tag and without the GFP fusion-tag was used as a negative control for in-gel fluorescence, western blot analysis and flow cytometry. For constructs containing the β-lactamase tag, the GFP sequence was removed and replaced with a β-lactamase sequence using Gibson cloning. For generation of *M. smegmatis* compatible plasmids, the entire coding region of the TatC homologs including the entire GFP sequence and the poly-His tag were PCR amplified out of their respective pET28(a+)-GFP-ccdB vector using primers with compatible regions for placement into the pMyNT vector using Gibson assembly [54]. For β-lactamase constructs, the GFP sequence was replaced by a β-lactamase sequence using Gibson assembly.

E. coli *Expression*

Plasmids were transformed into BL21 Gold (DE3) cells and transferred onto LB agar plates containing 50 μg/ml kanamycin plates after one-hour incubation. After overnight incubation at 37°C, colonies were scraped off the plates into 5 mL of LB, resuspended, and the OD600 was determined. These samples were then diluted into 50 mL 2xYT containing 50 μg/ml kanamycin in 125 mL baffled flasks to a starting OD600 of

approximately 0.01. Cultures were grown in an orbital shaker at 37°C until they reached an OD600 of 0.15. The temperature of the orbital shaker was then reduced to 16°C. Upon reaching an OD600 of 0.3, IPTG was added to final concentration of 1mM to induce expression. Cultures were grown for a further 16 hours prior to analysis.

β-Lactamase Survival Test

Plasmids containing the β-lactamase tag were expressed overnight at 16°C as previously described. Cells from each overnight culture were washed with phosphate buffered saline (PBS) to remove IPTG then diluted into fresh 50 mL 2xYT media containing 50 μg/ml kanamycin to a starting OD600 of 0.1 in 125 mL baffled flasks. Cultures were grown at 37°C to an OD600 of approximately 0.5 where a control sample from each culture was taken, diluted 10,000 times in PBS, and 50 μL was plated onto LB agar plates containing 50 μg/ml kanamycin. To each culture, 50 μg/ml ampicillin was added and shaken at 37°C for a further 90 minutes. A sample from each culture was taken and diluted 200 times in PBS, and 50 μL was plated onto LB agar plates containing 50 μg/ml kanamycin. Plates were grown overnight (~16 hours) and the number of colonies on each plate was counted. Colony counts from the second plating were normalized by the colony counts from the first plating to account for variation in the OD600 at which ampicillin was added to determine relative survival. The procedure was performed in triplicate and standard errors of normalized values were calculated. For each plot of relative survival, the values are normalized to the highest survival rate of the samples in the figure.

M. smegmatis *Expression*

For *M. smegmatis* overexpression, constructs were transformed into mc^2155 cells using electroporation and transferred onto Middlebrook 7H11 plates (10.25 g Middlebrook 7H11 Agar Base, 1 vial ADC growth supplement, 2.5 g glycerol, 1 mM CaCl2, 50 µg/mL carbenicillin, 10 µg/mL cyclohexamide, 50 µg/mL hygromycin, and water to 500 mL) after a three hour incubation in 1 mL Middlebrook 7H9 culture media (2.35 g Middlebrook 7H9 Broth Base, 1 vial ADC growth supplement, 0.5 g Tween-80, 1 mM CaCl2, 50 µg/mL carbenicillin, 10 µg/mL cyclohexamide, and water to 500 mL). Plates were grown for three to four days until colonies formed. Single colonies were picked into 5 mL Middlebrook 7H9 culture media containing 50 µg/mL hygromycin. The following day, 50 mL cultures of Middlebrook 7H9 expression media (2.35 g Middlebrook 7H9 Broth Base, 0.25 g Tween-80, 1 g glycerol, 1 g glucose, 1 mM CaCl2, 50 µg/mL carbenicillin, 10 µg/mL cyclohexamide, 50 µg/mL hygromycin, and water to 500 mL) were inoculated at a starting OD600 of 0.005. Cultures were grown at 37°C and expression was induced with 0.2% acetamide at an OD600 of 0.5. Cultures were grown for six hours after induction prior to analysis.

Flow Cytometry

A 200 µL sample of each expression culture was centrifuged at 4000g for 3 minutes to pellet the cells and then the supernatant was removed. Cells were resuspended in 1 mL of PBS and 200 µL of each were dispensed into 96-well plates and kept on ice for analysis. Whole-cell GFP fluorescence was determined using a MACSQuant10 Analyzer. Forward scattering, side scattering, and total fluorescence at 488 nm were considered during

analysis. Measured events were gated based on the negative control sample to contain the lowest 90% of both forward and side scattering values to remove anomalous particles, such as dead or clumped cells. Mean cell fluorescence was calculated for the gated population as a measure of folded TatC. At least four independent expression trials were performed for each sequence tested to ascertain expression variance. Flow cytometry data analysis was performed with FlowJo Software. Flow cytometry data is normalized to a standard for each day data was collected. For example, for '*Aa*-tail/wild-type' data points, the mean fluorescence values of the *Aa*-tail swap chimeras were normalized by the mean fluorescence of their respective homologs containing the wild-type tail for that day's trial. Similarly, for relative fluorescence data points in which wild-type *Aa*TatC was the standard, the mean fluorescence of each sample was normalized by the mean fluorescence of the *Aa*TatC sample for that day's trial. In both cases, final calculated values are averages of the normalized values over at least four trials with error bars representing standard errors of the mean for those normalized values.

In-Gel Fluorescence and Western Blot Analyses

In-gel fluorescence and western blot analyses were used as an alternative measure of total expressed proteins. 5 mL of expression samples were centrifuged and supernatant discarded. Samples were resuspended to an OD600 of 3.0 in PBS. 1 mL of each sample was collected and 250 µL lysis buffer (375 mM Tris-HCl pH 6.8, 6% SDS, 48% glycerol, 9% 2-Mercaptoethanol, 0.03% bromophenol blue) was added. Samples were lysed via freeze fracturing by three rounds of freezing using liquid nitrogen and thawing using room temperature water. 20 µL of each lysed sample was subjected to SDS-PAGE. SDS-PAGE

gels were imaged for fluorescence using a UV gel imager with a filter for GFP fluorescence to determine in-gel fluorescence.

For western blot analysis, the samples were transferred from the gel onto a nitrocellulose membrane using the Trans-Blot Turbo System. The membranes were washed three times with 15 mL TTBS (50 mM Tris pH 7.6, 150 mM NaCl, 0.05% Tween-20), incubated one hour with 15 mL 5% milk powder in TTBS, washed three times with 15 mL of TTBS, and then incubated with 1:5000 anti-GFP Mouse primary antibody (EMD Millipore, Lot # 2483215) in 15 mL 5% milk powder in TTBS overnight. Membranes were washed three times with 15 mL TTBS, incubated with 1:15000 IRDye® 800CW Donkey anti-Mouse secondary antibody (LI-COR, Lot # C31024- 04) in 15 mL 5% milk powder in TTBS for one hour, washed three times with 15 mL TTBS, and then visualized using a Licor IR western blot scanner. ImageJ was used to process the images [55].

The CG Model Overview

The CG model is employed with only minor modifications from [46], all of which are specified below. Key features of the CG model and its implementation are provided here; for a full discussion of the CG model, the reader is referred to [36].

As described in [46], the CG model explicitly describes the configurational dynamics of the nascent-protein chain, conformational gating in the Sec translocon, and the slow dynamics of ribosomal translation. The nascent chain is represented as a freely jointed chain of beads, where each bead represents three amino acids and has a diameter of 8 Å, the typical Kuhn length for polypeptide chains [56, 57]. Bonding interactions between neighboring beads are described using the finite extension nonlinear elastic (FENE)

potential [58], short-ranged nonbonding interactions are modeled using the Lennard-Jones potential, electrostatic interactions are modeled using the Debye-Hückel potential, periplasmic binding is included as described in [36] for BiP, and solvent interactions are described using a position-dependent potential based on the water-membrane transfer free energy for each CG bead; all parameters are the same as used previously [36], unless otherwise stated. The time evolution of the nascent protein is modeled using overdamped Langevin dynamics, with the CG beads confined to a two-dimensional subspace that runs along the axis of the translocon channel and between the two helices of the lateral gate (LG). Conformational gating of the translocon LG corresponds to the LG helices moving out of the plane of confinement for the CG beads, allowing the nascent chain to pass into the membrane bilayer. The rate of stochastic LG opening and closing is dependent on the sequence of the nascent protein CG beads that occupy the translocon channel. Ribosomal translation is directly simulated via growth of the nascent protein at the ribosome exit channel; throughout translation, the C-terminus of the nascent protein is held fixed, and new beads are sequentially added at a rate of 24 residues per second. Upon completion of translation, the C-terminus is released from the ribosome. It has been confirmed that the results presented in the current study are robust with respect to changes in the rate of ribosomal translation (Pearson correlation coefficient between Wt/*Aa*-tail ratios obtained using a rate of translation of 24 residues/sec and 6 residues/sec, $r = 0.99 \pm 0.06$).

The CG Model Implementation Details

Two changes to the protocol for the CG simulation model were introduced in the current study, with respect to the protocol used in[36]. These modifications were included

to remove unphysical artifacts in the simulations, although it is emphasized that conclusions in the main text are qualitatively unchanged by these modifications (Pearson correlation coefficient between Wt/*Aa*-tail ratios obtained with and without the modifications to the simulation protocol, r = 0.97±0.09).

The first change in the CG model is that the ribosome is assumed to remain associated with the translocon following translation of the nascent protein. In the previously implementation of the model, the ribosome was assumed to dissociate from the translocon immediately following stop-translation, which was found in the current study to lead to artifacts for nascent proteins with extremely short C-terminal domains. Furthermore, this modification is consistent with experimental evidence that indicates that the timescale for ribosomal dissociation is slower than the trajectories simulated here [59, 60].

The second change in the CG model relates to the potential energy cost of flipping hydrophilic nascent-protein loops across the lipid membrane at significant distances from the translocon. The Wimley-White water-octanol transfer free energy scale [61] that was used to parameterize the interactions of the CG beads with the membrane is appropriate for describing the transfer of amino acids between an aqueous region and either the phospholipid interface or the region of the membrane interior that is close to the translocon lateral gate [62]. However, the flipping of hydrophilic nascent-protein loops across the membrane at significant distances from the translocon involves moving CG beads through the hydrophilic core of the membrane interior, which will incur a large potential energy barrier [62]. To account for this effect, and to avoid unphysical flipping of short hydrophilic loops across the membrane, an additional potential energy term was included in the

potential energy function that describes the interactions between the CG beads and the membrane. We emphasize that this new term has no noticeable effect on the potential energy function for the CG beads at distances within 8 Å to the translocon channel; it simply affects unphysical flipping of the TM domains across the membrane at larger distances from the channel. This artifact was not observed in the earlier study using the CG model, since only processes involving the translocation or membrane integration of a single TM domain were considered.

The CG Model Bead Mapping

In the current study, amino-acid sequences for the TatC homologs are mapped onto sequences of CG beads as follows. Each consecutive trio of amino acid residues in the nascent protein sequence is mapped to an associated CG bead. The water-membrane transfer free energy for each CG bead is taken to be the sum of the contributions from the individual amino acids; these values are taken from the experimental water-octanol transfer free energies for single residues [61]. The charge for each CG bead is taken to be the sum of the contribution from the individual amino acids. As in [36], positively charged residues (arginine and lysine) were modeled with a +2 charge to capture significant effects on topology due to changes in the nascent protein sequence. Histidine residues were modeled with a +1 charge to account for the partial protonation of these residues, and negatively charged residues (glutamate and aspartate) were modeled with a charge of -1. The mapping procedure for *Aa*TatC is depicted in Figure 2.5A as an example.

The CG Model Calculation Details

The co-translational membrane integration for each TatC sequence is simulated using 1200 independent CG trajectories. As in [36], each CG trajectory is performed with a timestep of 100 ns. All trajectories were terminated 30 seconds after the end of translation for the protein sequence.

Analysis of Simulation Results.

To determine whether a given trajectory leads to integration in the correct multispanning topology, the topology of a nascent protein configuration can be characterized by the location of the soluble loops that connect the TMD. We thus specify a collective variable λi associated with each loop, with i=1 corresponding to the loop that leads TMD 1 in the sequence (i.e. the N-terminal sequence) and i=7 corresponding to the loop that follows TMD 6 (i.e. the C-tail). If loop i is in the cytosol, then $\lambda i = 1$; if loop i is in the periplasm, then $\lambda i = -1$; otherwise, $\lambda i = 0$. Whether a given loop is in the cytosol, in the membrane, or in the periplasm is determined by the tracking position of a representative bead in that loop (Table 2.1). Representative beads were chosen based on having the lowest probability of being inside the lipid region compared to other beads in that loop. A given trajectory is determined to have reached correct IMP integration ($\lambda i = -1$ for periplasmic loops and, $\lambda i = 1$ for cytosolic loops) if a configuration with the loops in the correct orientation is sampled during a time window of 6 seconds taken 25 seconds after the end of translation; the time window of 25 seconds was found sufficient to allow the nascent protein to finish the integration/translocation of TMD 6.

	Loop 1	Loop 2	Loop 3	Loop 4	Loop 5	Loop 6	Loop 7
AaTatC	7-9	43-45	88-90	145-147	181-183	202-204	238-239
Mt	7-9	61-63	112-114	151-153	193-195	220-222	244-246
Mt(Aa-tail)	7-9	61-63	112-114	151-153	193-195	220-222	244-246
Bp	25-27	64-66	112-114	160-162	196-198	220-222	253-255
Bp(Aa-tail)	25-27	64-66	112-114	160-162	196-198	220-222	250-252
Cj	13-15	55-57	100-102	139-141	187-189	208-210	238-240
Cj(Aa-tail)	13-15	55-57	100-102	139-141	187-189	208-210	238-240
Dr	28-30	73-75	118-120	166-168	202-204	229-231	262-264
Dr(Aa-tail)	28-30	73-75	118-120	166-168	202-204	229-231	247-249
Ec	10-12	55-57	103-105	142-144	190-192	211-213	244-246
Ec(Aa-tail)	10-12	55-57	103-105	142-144	190-192	211-213	244-246
Hy	7-9	40-42	94-96	139-141	184-186	205-207	232-234
Hy(Aa-tail)	7-9	40-42	94-96	139-141	184-186	205-207	232-234
Sa	7-9	43-45	91-93	142-144	178-180	199-201	229-231
Sa(Aa-tail)	7-9	43-45	91-93	142-144	178-180	199-201	229-231
Vc	16-18	52-54	103-105	145-147	190-192	211-213	247-249
Vc(Aa-tail)	16-18	52-54	103-105	145-147	190-192	211-213	241-243
Ws	10-12	61-63	97-99	148-150	181-183	205-207	241
Ws(Aa-tail)	10-12	61-63	97-99	148-150	181-183	205-207	235-237

Table 2.1: Loop Definitions Used in Simulation Trajectory Analysis. Each loop is specified in terms of the amino-acid residue sequence numbers (end-points inclusive) associated with the wild-type sequence.

Figure 2.4 shows the fraction of trajectories that exhibit the correct topology for each individual loop for all TatC homologs and chimeras considered in this study. It is clear from Figure 2.4 that the changes to the amino-acid sequence considered in this study largely only impact the topology of the domain where the changes to the amino acid sequence were introduced; the topology of the rest of the protein is not predicted by the CG simulation model to be significantly affected by the sequence changes. The calculated results are robust with respect to the details of the definition of simulated integration efficiency (Pearson correlation coefficient between Wt/Mutant ratios obtained analyzing only the loop that was modified and those obtained analyzing all loops, $r = 0.85 \pm 0.16$); to minimize statistical error, for all simulation results presented in the main text, the topology of the IMP is thus characterized in terms of only the loop of interest.

C h a p t e r 3

A LINK BETWEEN INTEGRAL MEMBRANE PROTEIN EXPRESSION AND SIMULATED INTEGRATION EFFICIENCY OF THE C-TAIL FOR AN EXPANDED POOL OF TATC MUTANTS

Abstract

The heterologous overexpression of integral membrane proteins in *Escherichia coli* often yields insufficient quantities of purifiable protein for applications of interest. The current study leverages a recently discovered link between co-translational membrane integration efficiency and protein expression levels to predict sequence modifications that improve expression. Membrane integration efficiencies, obtained using a coarse-grained simulation approach, robustly predict expression for a set of 140 sequence modifications on the integral membrane protein TatC, including loop-swap chimeras and single-residue mutations distributed throughout the protein sequence. Mutations that improve simulated integration efficiency are found to be almost four-fold enriched with respect to improved experimentally observed expression levels. Furthermore, the effect of double mutations, on both simulated integration efficiency and experimentally observed expression levels, is shown to be largely independent, suggesting that multiple mutations can be introduced to yield higher levels of purifiable protein. This work provides a foundation for a general method for the rational overexpression of integral membrane proteins based on computationally simulated membrane integration efficiencies.

Introduction

Integral membrane proteins (IMPs) play crucial roles in the transport of molecules, energy, and information across the membrane and are an important focus of structural and biophysical studies. However, the production of sufficient levels of IMPs is a limiting factor in their characterization [23]. Even among homologous IMP sequences, expression levels can vary widely [22, 23, 63-65], and the mechanistic basis for this variability is often unclear. Extensive efforts have been committed to identify IMP sequences, expression conditions, and host modifications that yield IMP expression at sufficient levels for further study [20, 37, 38]. Despite these efforts, general guidelines for successful overexpression for IMPs of interest are lacking.

Heterologous overexpression of IMPs in *E. coli* involves multiple steps during biogenesis that are potential bottlenecks for overexpression, including the correct targeting to the inner membrane[5, 6], integration [7, 9, 36, 66-68], and folding [13, 18, 27]. For a given sequence, understanding how each of these steps affects observed expression levels may lead to improved strategies for IMP overexpression.

Previous work indicates that the Sec-facilitated membrane integration step of biogenesis is a limiting factor in the overexpression of the TatC IMP [22]. Sequence changes that alter the efficiency of membrane integration efficiency, determined either from coarse-grained simulations or experimentally, correlate with experimentally observed IMP expression levels. Further work is necessary to explore the generality of this link and its potential for enabling the rational enhancement of IMP expression.

The current study demonstrates the predictive capacity of simulated integration efficiency for experimental expression by examining a wide range of sequence

modifications and TatC homologs. The studied sequence modifications include point mutations, loop-swap chimeras, and double-loop-swap chimeras, and it is shown that the simulated integration efficiency – as predicted by coarse-grained simulations – broadly correlates with IMP expression. An antibiotic-resistance assay is employed to directly validate the simulated integration efficiencies and to confirm the mechanistic interpretation. We further demonstrate multiplicative and largely independent nature of the effect of multiple mutations on both the simulated integration efficiency and the experimentally observed expression levels. Finally, we provide a methodology that can be used to generally identify sequence regions in other IMPs that may exhibit correlations like those elucidated here for TatC, yielding a broadly applicable tool for the computational prediction of sequence modifications that improve IMP overexpression.

Results

TatC Expression Levels Are Changed by Loop Swaps

TatC is an IMP with six transmembrane domains (TMD) and a cytoplasmic N- and C-terminus (Figure 3.1A) that is a component of the bacterial twin-arginine translocation pathway [40]. A representative pool of 111 loop-swap chimeras was generated by replacing a single loop in one of ten wild-type TatC homologs (*Aquifex aeolicus* (*Aa*), *Bordetella parapertussis* (*Bp*), *Campylobacter jejuni* (*Cj*), *Deinococcus radiodurans* (*Dr*), *Escherichia coli* (*Ec*), *Hydrogenivirga species 128-5-R1* (*Hy*), *Mycobacterium tuberculosis* (*Mt*), *Staphylococcus aureus* (*Sa*), *Vibrio cholera* (*Vc*), and *Wolinella succinogenes* (*Ws*)) with the corresponding loop from one of the other nine homologs (Figures 3,1A and 3.2). Loop domains were identified by sequence alignment and membrane topology predictions [45]. Both mutant and wild-type expression levels were determined using a C-terminal GFP tag [26] (see Methods), and the relative effect of each mutation on expression was quantified in terms of the ratio

$$\text{Exp. Expression} = \frac{expression(\text{mutant})}{expression(\text{wild}-\text{type})}, \qquad (1)$$

Values greater than unity (>1.0) indicate improvement in expression due to the sequence modification. The set of loop swaps exhibit a wide range of values for this experimental expression ratio, as shown in Figure 3.1B. The effects of single loop swaps range from 0.02- to 40-fold changes, with 43% of the studied loop swaps yielding improved expression. Control studies were performed to confirm that the C-terminal GFP tag does not substantially alter the experimentally measured expression levels. A set of 11 single-loop-swap chimeras and their corresponding wild-type sequences were cloned into an alternative construct containing an N-terminal Strep tag (WSHPQFEK) with no C-terminal

tag (see Methods). The experimental expression ratio in Equation 1 was measured for each N-terminal Strep tag construct and compared against quantification via C-terminal GFP fluorescence. Figure 3.1C shows this comparison, revealing agreement for all studied cases between measured expression levels using either tag. This result, which is in agreement with extensive studies where IMP-GFP fluorescence was used to quantify expression [26, 27], indicates that the experimental expression outcomes are robust with respect to the means of quantifying the expression levels.

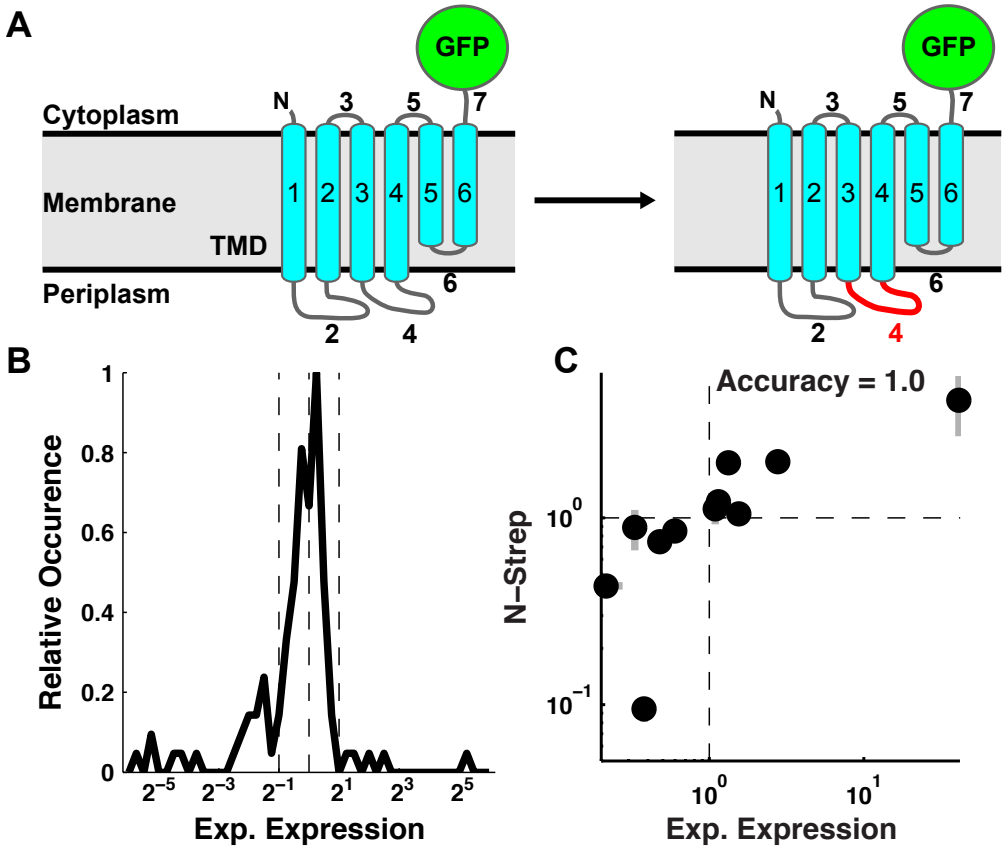
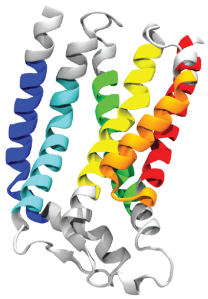


Figure 3.1: TatC Loop-Swap Chimeras Demonstrate a Range of Expression Outcomes. (A) A schematic of a wild-type (left) and loop-swap chimera (right) for the TatC IMP sequence with a C-terminal GFP tag. Homologous loop domains are swapped between TatC homologs to create loop-swap chimera mutants; the figure illustrates a loop-swap chimera of loop 4. (B) The distribution of experimental expression values (mutant/wild-type) representing the pool of 111 single-loop-swap TatC chimeras that are created by swapping loop domains between homologs. Loop-swap mutations have a wide range of effects on

experimental expression. Vertical dashed lines indicate two-fold change, and no change in experimental expression. (C) Correlation between experimental expression levels quantified using a C-terminal GFP tag (Exp. Expression) or using an N-terminal Strep tag (N-strep). This demonstrates that experimental expression outcomes are not influenced by the location or size of the probe.

A



B

>AaTatC
PLTEHLRELRY**RLIISIIAFLIGSGIAFYFAK**YVFEILKEPILKSYPEVELITLSPTEPL**FILIK**
ISLAVGFIIASPVILYQFWRFIEPALYSHEKRAFIP**LLLGSILLFMLGALFAYFIVL**PLALKFLL
GLGFTQLLATPYLSVDMYISF**VLKLVVAFGIAFEMPIVLYVL**QKAGVITPEQLAS**FRKYFIVIAF**
VIGAIIAPDVSTQ**VLMAIPLLLLYEISIFLGKLA**TRKKKEIQKA
>MtTatC
SLVDHLTELRTR**LLISLAAILVTTIFGFVWYSH**SIFGLDSLGEWLRHPYCALPQSARADISADGE
CRLLATAPFDQF**MLRLKVGMAAGIVLACPVWFY**QLWAFITPGLYQRERRFAVA**FVIPAAVLFVAG**
AVLAYLVLSKALGFLLTVGSDVQVTALSGDRYFGF**LLNLLVVFGVSFEFPLLIVML**NLAGLLTYE
RLK**SWRRGLIFAMFVFAAIFTPGS**DP**FSMTALGAALTVLLELAIQIARVH**
>BpTatC
VSQDASNDNPDQQQDSFISHLVELRSRLLK**AAGAVVAVFIVLFLYPGASA**IYDVLAQPMLASLPE
GTRMIATGVITPF**MVPVKVTMMAAFVVALPVVLY**QAWAFVAPGLYKHEKRLALP**LILSSTLLFII**
GMAFCYFFVFRTVFHFIATFAPQSITPAPDIEAYLSF**VMTMFMAFGITFEVPVAVVLL**VKTGIVE
VAKLRAA**RGYVVVGAFVIAAVVTPPDVV**SQ**FMLAVPLCLLYEVGLLCARLV**TPRRRGEEESEDDQ
ALTERH
>CjTatC
MFEELRPHLIELRKR**LFISVACIVVMFIVCFALRSY**ILDILKAPLIAVLPEVAKHVNVIEVQEAL
FTAMKVSFFAAFIFSLPVIFWQFWKFVAPGLYDNEKRLVVP**FVSFASIMFAFGACFCYFVVV**PLA
FKFLINFGLNEDFNPVITIGTYVDF**FTKVVVAFGLAFEMPVIAFFF**AKIGLIDDSFLKRH**FRIAV**
LVIFVFSAFMTPPDVLSQ**FLMAGPLCGLYGLSILIVQKV**NPAPKDKESDE
>DrTatC
TQLPPPEQTVLKPAPPELASAPLFDHLEELRRR**LILSVVFLAVGMVIAFTYRVQ**LIELVKVPLTY
SELYTTGKVQLVTTKLASQL**LLSFNLAFWAGLTLALPFIVW**QIWAFIAPGLYPQERRWGLP**FILG**
AGFAFAAGVVFGYKLVLPTMVPFLIEFLAGTVTQMQDLQEYIGT**VVTFLVAFGVAFELPILAVIL**
TRLGIVNHTMLRQ**GWRFALIGIMILAAVITPTPD**PANM**ALVAVPLYALYELGVVLSRVF**RVIAPE
EQERPAPMS
>EcTatC
MSVEDTQPLITHLIELRKR**LLNCIIAVIVIFLCLVYFAND**IYHLVSAPLIKQLPQGSTMIATDVA
SPF**FTPIKLTFMVSLILSAPVILY**QVWAFIAPALYKHERRLVVP**LLVSSSLLFYIGMAFAYFVVF**
PLAFGFLANTAPEGVQVSTDIASYLSF**VMALFMAFGVSFEVPVAIVLL**CWMGITSPEDLRKK**RPY**
VLVGAFVVGMLLTPPDVFSQT**LLAIPMYCLFEIGVFFSRFY**VGKGRNREEENDAEAESEKTEE
>HyTatC
MPLTEHLRELRTR**LIRSIIAFLIAAGGSFYFARY**VFEFLKEPVVKSYPDVELITLSPTEP**LFILI**
KISLTVGLIIASPVILFEIWRFVEPALYPQEKKLFIP**LLLSSVLLFVMGGVFAYAVVL**PMALKFL
LGLGFSQLAATPYLSVNLYVSF**VLKMLIAFGIAFEMPIFLYML**QRAGVVSQQQLKK**FRRYFIVVA**
FLVGALIAPDVATQ**VLMAIPLLVLYEVSILLGRTV**RKGEKEKALARVEEEETRE
>SaTatC
MGVHFSELRHRLVKI**LLSFVVTVIVVYVSSFWWMT**PFITYITRAHVSLHAFSFTEMI**QIYVMIIF**
FIAFCFISPVMFYQLWAFIAPGLHNNERQFIYK**YSFFSVLLFCAGVAFAFYVGF**PIIIQFALKLS
LTLNISPVIGFKAYLVE**LIRWLFTFGILFQLPILFIGL**AKFGLIDITSLKHY**RKYIYFACFVLAS**
IIAPPDLTLN**ILLTLPLILLFEFSMFIVKFT**CRGKPPTH
>VcTatC
MSSVEQTQPLISHLLELRNR**LLKAVAAVVVIFIGLIYFSN**EIYEFVSKPLVERLPAGATMIATDV
ASPF**FTPLKLTLIAAVFLAVPFILY**QVWAFVAPGLYKHERRLIFP**LLVSSSLLFYCGVAFAYFVV**
FPLVFGFFTAISLGGVEFATDIASYLDF**VLALFLAFGIAFEVPVAIILL**CWTGATTPKSLSEK**RP**
YIIVGAFVVGMLLTPPDMISQ**TLLAIPMCLLFEVGLFFARFY**TRDEADEGQEEEE
>WsTatC
MFEELKPHIQELRKR**LINAVVALFIAFFICFFFWEG**ILDWMIAPLKAALPAGSNVIFTEVGEAF**F**
TAIKVSFFSAFMFSLPVIFWQVWLFVAPGLYQNEKMLVLP**FVFFGTLMFVTGALFAYYVVF**PFGF
TYLINFGSTLFTALPSVGFYVTF**FAKLMIGFGIAFELPVVTFFL**AKLGLVTDKTLRDF**FKYAIII**
IFIVAAILTPPDVITQ**FMMAIPLTFLYWVSILIAKMV**NPETSPNEE

Figure 3.2: TatC Transmembrane and Loop Domain Definitions. (A) A ribbons diagram of the structure of *Aa*TatC (RCSB PDB: 4HTS). (B) Domain definitions used in generating the swap chimeras, with TMDs highlighted, are shown as used in (A).

Simulated Integration Efficiency is Predictive of TatC Expression

Correlation between simulated integration efficiency and experimentally observed expression levels was previously identified in TatC based on a limited set of mutations [22]; here, we systematically test the predictive capacity of simulated integration efficiency for expression in a diverse set of 111 loop-swap chimeras. CG simulations were performed for each chimera and wild-type sequence (see Methods), and the relative effect of each mutation on simulated integration efficiency was quantified in terms of the ratio

$$\text{Sim. Integration} = \frac{P_{Cin}(\text{mutant})}{P_{Cin}(\text{wild-type})}, \qquad (2)$$

where P_{Cin} corresponds to the fraction of simulated trajectories for which the C-tail domain is correctly localized with respect to the cell membrane for each sequence. In a later Results section, we investigate the use of sequence features other than the C-tail for quantifying integration efficiency. Receiver operator characteristic (ROC) curves (Figure 3.3A) [69] provide a statistical measure of the predictive capacity of simulated integration efficiency, with values in excess of 0.5 for the area under the ROC curve (AUC) [69] indicating predictive capacity.

ROC curves in Figure 3.3A are shown for datasets corresponding to all 111 loop-swap chimeras (blue) and to the subset of 82 loop-swap chimeras that exclude C-tail swaps (green). This plot demonstrates the predictive capacity of simulated integration efficiency for experimental expression, with AUC values exceeding 0.5 with 95% statistical confidence. The similarity of the two curves indicates that the predictive capacity of the

simulated integration efficiency is relatively insensitive to whether the loop-swap involves the C-tail domain.

Also, indicated in Figure 3.3A (blue and green dots) are the points along the ROC curve that correspond to the cut-off value (defining positive prediction) for the simulated integration efficiency ratio in Equation 2 that offers the greatest predictive capacity for experimentally observed expression; for both datasets, this optimal value is found to be 1.0, indicating that increases or decreases in the simulated integration efficiency straightforwardly predict the corresponding changes in experimental expression levels.

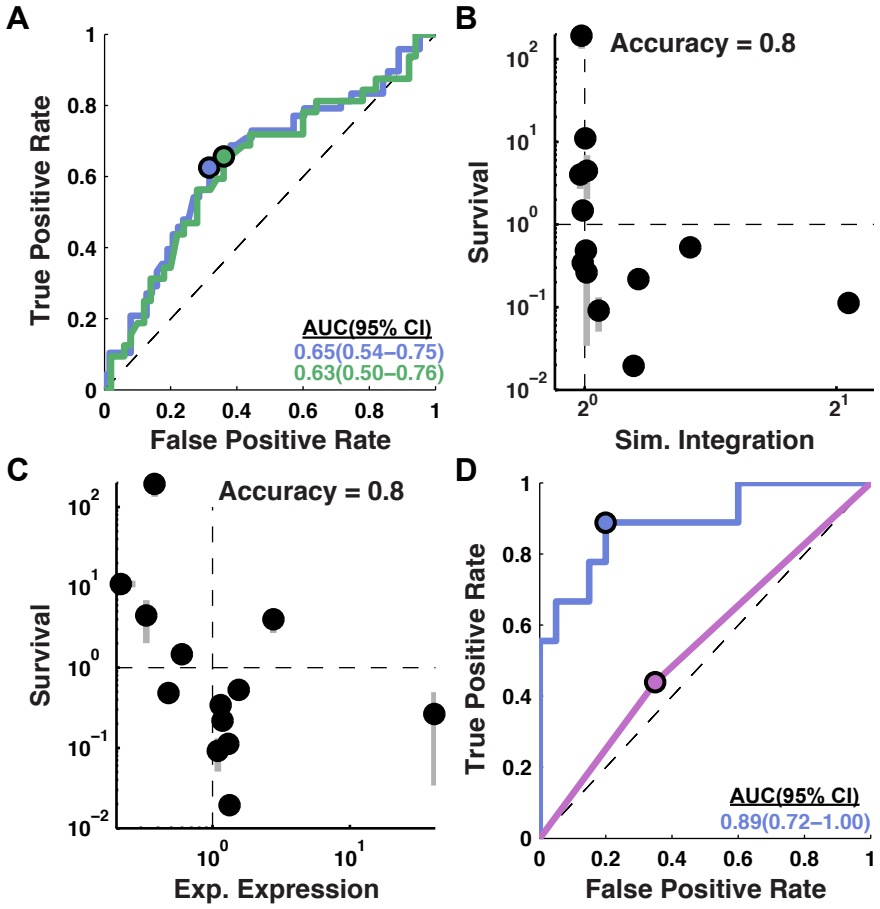


Figure 3.3: C-Tail Localization Is Predictive of Experimental Expression Outcome. (A) The predictive capacity of simulated integration efficiency for experimental expression assessed using a ROC curve for all single-loop-swap chimeras (blue, N=111) and all single-

loop-swap chimeras excluding those in which the C-tail was swapped (green, N=82). Significant predictive capacity is observed for both sets. (B) Comparison of simulated integration efficiency with survival for TatC loop-swap chimeras. A negative correlation between survival and simulated integration efficiency indicates that the C-tail topology predicted by the CG simulations occurs *in vivo*. One sequence tested had a non-observable survival level and was not included in the plot, this datapoint was included in the accuracy calculation. (C) Comparison of experimental expression with relative ampicillin resistance (survival) for TatC loop-swap chimeras. A negative correlation between survival and experimental expression indicates that the C-tail mislocalizes in poorly expressing chimeras, consistent with the mechanism predicted by the CG simulations. One sequence tested had a non-observable survival level and was not included in the plot, this datapoint was included in the accuracy calculation. (D) The predictive capacity of simulated integration efficiency for experimental expression assessed using a ROC curve for TatC point mutants (N=29). Simulated integration efficiency (blue) outperforms prediction of experimental expression by the positive inside rule (purple).

Experimental Confirmation of Simulated Integration Efficiency Values

To experimentally confirm that the *in vivo* integration efficiency is correctly described by the CG simulations, we apply a previously developed antibiotic resistance assay [22] (see Methods). Ampicillin resistance imparted by the expression of TatC sequences containing a C-terminal β-lactamase tag correlates positively with the quantity of proteins integrated with their C-tail in the periplasm (i.e. mislocalized). Therefore, a negative correlation between ampicillin survival and simulated integration is expected if mislocalization of the C-tail occurs *in vivo*, as predicted by CG model simulations.

The survival metric reported in Figure 3.3B is the ratio of colonies observed following ampicillin treatment between a loop-swap chimera and the corresponding wild-type TatC sequence. For a set of 14 loop-swap chimeras, Figure 3.3B compares the relative survival to simulated integration efficiency. For 11 of these 14 cases, the corresponding data points in Figure 3.3B fall into the diagonal quadrants of the plot, indicating good agreement between the experimental and simulated measures of integration efficiency (Accuracy = 0.8±0.2, 95% confidence interval).

Figure 3.3C plots the correlation between ampicillin survival and experimental expression for the same set of loop-swap chimeras. As expected (given the positive correlation between simulated integration efficiency and experimental expression in Figure 3.3A, and given the negative correlation between the simulated integration efficiency and the survival assay in Figure 3.3B), Figure 3.3C indicates strong negative correlation between ampicillin survival and experimental expression, with 11 of the 14 data points falling in the diagonal quadrants (Accuracy = 0.8±0.2, 95% confidence interval). Taken together, Figures 3.3B and C demonstrate that simulated integration is a reliable predictor of the C-tail orientation, which is in turn a reliable predictor of experimental expression.

The Effect of Point Mutations on Integration Efficiency Is Predictive for Expression

Rather than loop-swap mutations, we now consider the effect single-point mutations on both experimental expression and simulated integration efficiency. Point mutants introduce minimal changes to the wild-type sequence and are often used in a protein-sequence design context [70-72]. The blue curve in Figure 3.3D shows the ROC curve for a set of 29 point mutants; each exhibits a single mutation at a position in the wild-type sequence that is not universally conserved across homologs, with the mutation either increasing or decreasing the charge at that position. The blue curve in Figure 3.3D indicates that the simulated integration efficiencies from the coarse-grained method have predictive capacity (AUC = 0.89) that is even higher than was found in Figure 3.3A for loop-swap mutations (AUC = 0.65).

For comparison, the purple curve in Figure 3.3D explores the predictive capacity of a simpler measure of integration efficiency based only on the positive inside rule, which

observes that positively charged residues are more likely to be localized to the cytosolic side of the cell membrane [17] and that modification of the positively charged residues can change IMP topology [13, 18, 22, 27]. As employed here, the positive inside rule simply predicts that a mutation will have increased integration efficiency (and thus a positive effect on expression) if it increases the net charge of the cytosolic loops minus the net charge of the periplasmic loops, and vice versa. It is clear from the Figure 3.3D that in contrast to the prediction of the coarse-grained model (blue), the positive inside rule has little predictive capacity for expression when employed in this way. These results emphasize that the molecular processes and interactions that govern IMP integration are more complex, and they are more completely described using the coarse-grained simulations than by simple analysis of charged residues.

The Effects of Sequence Mutations on Simulated Integration Efficiency and Experimental Expression Are Additive

To determine whether multiple sequence modifications have a combinatory effect on expression and simulated integration efficiency, a set of 12 double-loop-swap chimeras was generated and tested against the corresponding effect of the constituent single-loop-swap mutations. Figure 3.4 shows that for both simulated integration efficiency (part A) and experimental expression (part B) comparison of the fold-change (Equations 1 and 2) observed for the double-loop-swap chimera is strongly correlated with to the product of fold-changes for the corresponding single-loop-swap chimeras (Pearson's correlation coefficient, $r = 0.9$). Linear fits of the data are plotted as solid lines. The slope of the linear fits for both simulated integration efficiency (Figure 3.4A, slope = 0.8) and experimental

expression (Figure 3.4B, slope = 0.7) deviate only slightly from unity, which indicates that the effect of each mutation is largely independent. The results in Figure 3.4 indicate that introducing multiple mutations is a viable strategy for enhancing expression, and that simulated integration efficiency largely captures the effect of these multiple mutations.

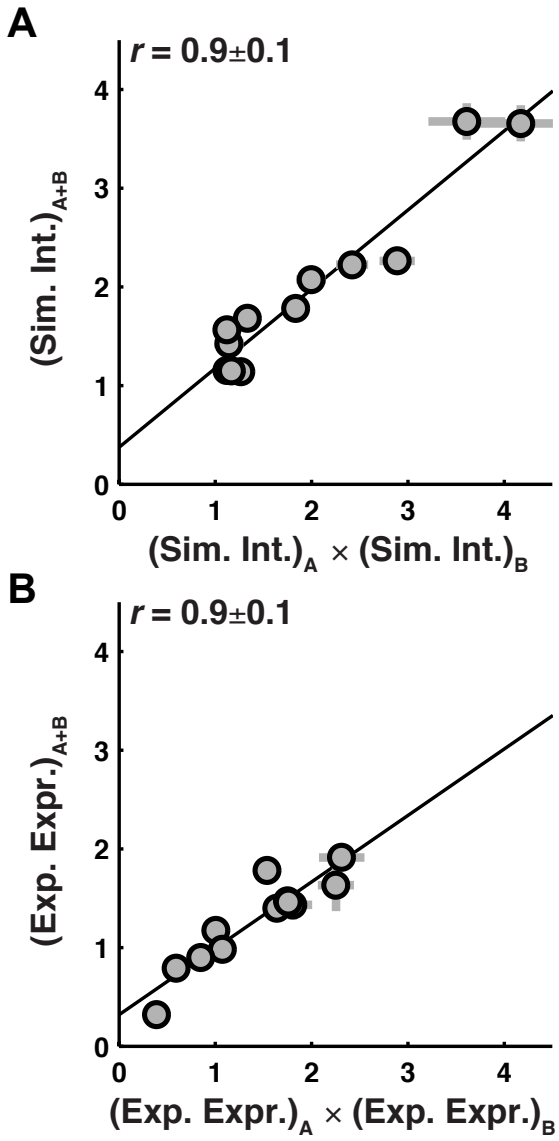


Figure 3.4: Effects of Sequence Modifications on Simulated Integration and Experimental Expression Are Nearly Independent. (A) The simulated integration efficiency of double-loop-swap chimeras (vertical axis) compared to the product of the simulated integration efficiencies of the constituent single-loop-swap chimeras (horizontal axis). There is a strong correlation, with a slope of 0.8, indicating that the effect of loop-swap mutations on

simulated integration efficiency is multiplicative and largely independent. (B) The experimental expression of double-loop-swap chimeras (vertical axis) against the product of the experimental expression values of the constituent single-loop-swap chimeras (horizontal axis). Again, there is strong correlation, with a slope of 0.7, indicating that the effect of loop-swap mutations on experimental expression is multiplicative and largely independent.

TatC Topology Features, Other Than C-tail Localization, Are Not Predictive for Expression

Using the fraction of coarse-grained trajectories for which the TatC C-tail reaches correct localization with the respect to the membrane as the measure of successful IMP integration, the results in Figure 3.3, along with previous work [22], support the conclusion that simulated integration efficiency reliably predicts experimental expression in TatC. However, other features of the TatC topology (such as the localization of other soluble loops) could have been employed to quantify IMP integration from the coarse-grained simulations. We now investigate the predictive capacity of the coarse-grained simulations for experimental expression, using alterative measures of IMP integration.

The alternative measures of IMP integration that are considered include, (1), $p^{(i)}$, the fraction of coarse-grained trajectories for which soluble loop *i* reaches correct localization with the respect to the membrane, (2), $p^{(All)}$, the fraction of coarse-grained trajectories for which all soluble loops reach correct localization, and, (3), $p^{(N)}$, the fraction of coarse-grained trajectories for which correct localization is achieved for the soluble loop that includes the mutation. In this notation, the previously discussed measure of IMP integration based on the C-tail is given by $p^{(7)}$.

Using each of these measures of IMP integration, we obtained ROC curves that compare the simulated integration efficiency with observed experimental expression, and

the corresponding AUC values are presented in Figure 3.5A. In all cases, the ROC curves were obtained for the datasets corresponding to all 140 TatC loop-swap and point mutations discussed in the preceding sections. The AUC for the C-tail measure ($p^{(7)}$) is 0.73, indicating the strong predictive capacity using this measure. However, it is clear that all other measures of integration efficiency fail to offer predictive capacity (yielding AUC values that are within 95% confidence of 0.5). Even when the measure of integration efficiency is based on the localization of the loop in which the mutation occurs (i.e., $p^{(N)}$), the predictive capacity is significant worse than using the C-tail (i.e., $p^{(7)}$).

The results in Figure 3.5A raise the question of the underlying mechanism for the predictive capacity of the C-tail localization for TatC. One hypothesis is that the C-tail acts as "aggregator" of all preceding errors in the IMP integration, providing a cumulative report on the TatC topology. A second hypothesis is that the C-tail is akin to a "canary in the coal mine," particularly sensitive to mutations, regardless of where in the sequence the mutation occurs. Finally, a third hypothesis is that the unique features of the C-tail could make it more amenable to accurate description by the coarse-grained method than the other TatC loops.

We directly test the aggregator hypothesis by investigating the degree to which the C-tail measure of integration efficiency is predictive of the alternative measures. Figure 3.5B presents the resulting AUC values, obtained from ROC curves for $p^{(7)}$ versus the alternative measures, using the full dataset of 140 TatC loop-swap and point mutations. It is clear from the figure that there is no significant correlation between $p^{(7)}$ and the other measures, a finding that is inconsistent with the aggregator hypothesis. Both Figures 3.5A

and B emphasize that the C-tail is a unique reporter of TatC integration efficiency, at least among the diverse set of measures considered here.

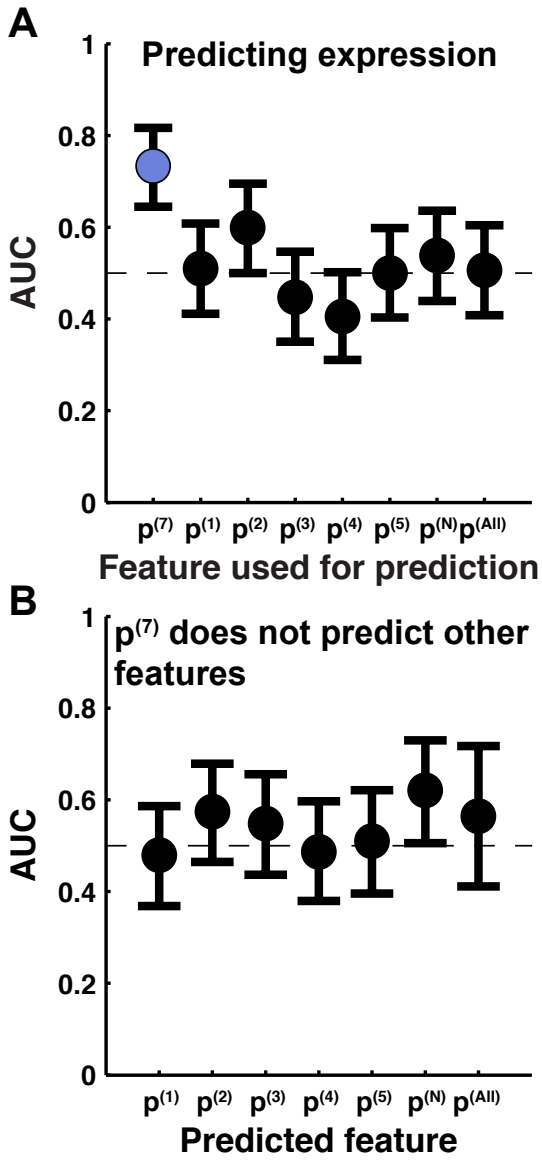


Figure 3.5 Simulated Integration Efficiency of the C-Tail Is the Only Topology Feature That Is Predictive of Experimental Expression for TatC. (A) AUC of the ROC curves for the metrics $p^{(7)}$, $p^{(1)}$, $p^{(2)}$, $p^{(3)}$, $p^{(4)}$, $p^{(5)}$, $p^{(N)}$, and $p^{(All)}$ (defined in the text) predicting the experimental expression of all loop-swap chimeras and point mutants. $P^{(7)}$ (C-tail simulated integration efficiency as used in Figures 3.3 and 3.4) is the only metric that is found have a statistically significant predictive capacity for experimental expression. (B) AUC of the ROC curves for $p^{(7)}$ (C-tail simulated integration efficiency as used in Figures 3.3 and 3.4) predicting $p^{(1)}$, $p^{(2)}$, $p^{(3)}$, $p^{(4)}$, $p^{(5)}$, $p^{(N)}$, and $p^{(All)}$ for all loop-swap chimeras and point mutants. $p^{(7)}$ is not significantly predictive of the localization for any other loop, it is

therefore unlikely an aggregator of upstream mislocalization. Error bars indicate 95% confidence intervals on the AUC value.

The second hypothesis reasons that the C-tail of TatC is particularly sensitive to sequence modification and is thus a useful reporter of integration efficiency, regardless of where in the sequence the mutation occurs. Although this hypothesis is difficult to directly test, it is consistent with the results from the antibiotic resistance assay, which found that C-tail localization was substantially impacted by mutations in other parts of the TatC sequence, even for mutations in other loops. Possibly contributing to the conformational sensitivity of the C-tail is that the preceding TM domains (TM5 and TM6) are relatively short, do not fully span the cell membrane, and are connected by a short turn between the TMDs, for which loop residues are difficult to assign [41, 43]. Furthermore, conformational sensitivity of the C-tail is consistent with the fact that this sequence domain is not conserved across TatC homologs and that it was not resolvable in reported TatC crystal structures [41, 43], indicating flexibility.

With regard to the third hypothesis, we note that the coarse-grained model does not explicitly describe sequence-specific interactions and packing effects among the TM domains; the model is thus expected to be most reliable for describing the topology of TM domains with weak tertiary interactions, such as the C-tail of TatC [41]. This explanation leaves open the possibility that improvements to the coarse-grained model in terms of its description of tertiary IMP interactions could lead to more robust measures of simulated integration efficiency for other loops

The analysis in this section is central to the question of how generally the coarse-grained simulations will be able to predict membrane protein expression for IMPs other

than TatC. It is very possible that for other IMPs, the C-tail localization will not be the most useful measure of IMP integration for predicting expression levels. In the next section, we thus describe a simple strategy for identifying a useful measure of IMP integration, on the basis of limited experimental expression data.

Predictors for Expression Can Be Identified from Training Data

Utilization of simulated integration efficiency to predict IMP expression in other systems requires knowledge of a useful measure of IMP integration to compute from the coarse-grained simulations. The results in Figures 3.3 and 3.4 in this work use the measure of C-tail localization ($p^{(7)}$) for this purpose, but, as is illustrated in Figure 3.5, other reasonable measure of simulated integration efficiency are not predictive for expression. For the study of an arbitrary IMP, we are thus faced with determining, as efficiently as possible, a useful measure of simulated integration efficiency to compute from the coarse-grained method.

Training on a limited dataset could provide one general method for the identification of topology features that are predictive for expression. Here we demonstrate this methodological framework in the context of the previously described set of TatC loop-swap chimeras and point mutants. The predictive capacity (AUC) of topology features is assessed for training sets of varying size, and the feature with the highest AUC is identified. For each training dataset size, 1,000,000 independent samples were taken and the most predictive metric for each of these samples was determined. For each topology feature, i, the probability that it is most predictive, $P^{(i)}$, can then be determined as a function of

training set size, x. We also report the expectation value of the AUC, $\langle \mathrm{AUC} \rangle$, on the full set of 140 sequence modifications calculated using Equation 3.

$$\langle \mathrm{AUC} \rangle(x) = \frac{1}{N_{tf}} \sum_{i \in tf} P^{(i)}(x)\, \mathrm{AUC}^{(i)}, \qquad (3)$$

where N_{tf} is the number of topology features, and tf is the full set of topology features. The expectation value of the AUC gives an indication of the predictive capacity one can expect given a training set size.

Figure 3.6A plots the probability of choosing each simulated integration efficiency metric as the most predictive, $P^{(i)}$, over different training set sizes. The C-tail localization metric, $p^{(7)}$, is correctly identified as most predictive for more than half of the training sets for training set sizes of more than approximately 20. Figure 3.6B shows the expectation value of the AUC for predicting experimental expression on the full dataset, shaded regions indicate 67% (dark) and 95% (light) confidence intervals obtained using bootstrapping. For TatC, as the size of the tested sequence pool increases, there is a greater probability of choosing the simulated integration efficiency of the C-tail as the most important metric (Figure 3.6A) and the significance of the AUC for simulated integration of the C-tail increases with pool size (Figure 3.6B). It is apparent that the full pool tested is not necessary to identify the most predictive metric. These results suggest that for expression data for a small test set of sequence modifications to an IMP sequence can be used to identify simulated integration efficiency features predictive of expression.

The strategy in Figure 3.6 illustrates that for cases in which limited IMP expression data is available, a useful measure of IMP integration from the coarse-grained simulations can be identified without other prior knowledge, thus yielding a general strategy for enhancing IMP expression in systems other than TatC. However, there will be cases in

which even limited IMP expression data is not available. For these cases, a reasonable strategy is to use a measure of IMP integration that involves a sequence domain that is expected to be prone to mislocalization with respect to the cell membrane. Analyses of sequence conservation [73] and residue co-evolution [74] provide reasonable strategies for identifying such sequence domains.

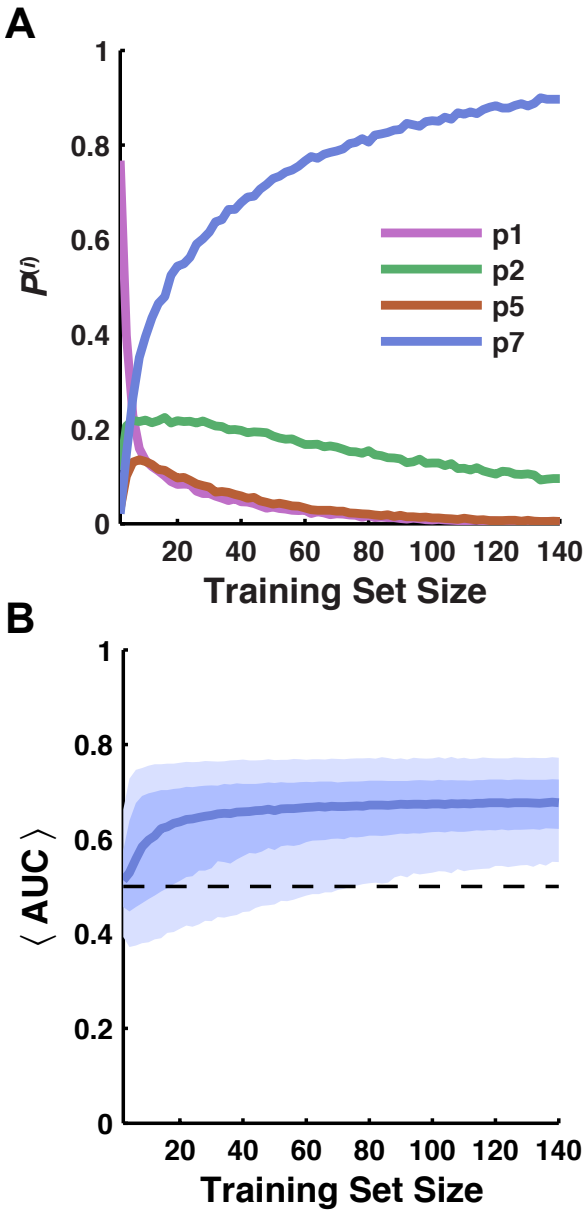


Figure 3.6: Topology Features Predictive of Expression Can Be Determined Based on Limited Training Data. (A) The probability that a topology feature is most predictive ($P^{(i)}$, highest AUC) for expression based on performance on training data of variable size. $p^{(7)}$

(simulated integration efficiency of the C-tail) has a high likelihood of being chosen as the simulated integration efficiency feature with the greatest predictive capacity, even at low training set sizes ($P^{(p7)}$>0.5 at training set size 18, and the probability of identifying $p^{(7)}$ as a predictor increases with the training set size. For clarity only features with values of $P^{(i)}$ greater than 0.1 for any training set size are shown in the plot. Not shown but included in the analysis are; $p^{(3)}$, $p^{(4)}$, $p^{(N)}$, and $p^{(All)}$. (B) The expectation value of the AUC, $\langle AUC \rangle$, for predicting experimental expression versus training data size. Confidence intervals are displayed at 67% (light blue) and 95% (darker blue).

Discussion

Heterologous expression of closely related IMP homologs in *E. coli* can provide a wide range of yields. IMP misintegration is one source of poor expression outcomes. Here, we utilize the link between the effect of sequence modification on the integration efficiency and experimental expression outcomes [22] to predict sequence modifications that improve expression for the IMP TatC. The integration efficiency of a given domain is determined by performing CG molecular dynamics simulations of the co-translational integration of the IMP via the Sec-translocon [36]. Simulated integration efficiency of the C-tail as determined by the CG model, and subsequently confirmed *in vivo*, is demonstrated to accurately predict experimental expression outcomes for diverse a set of 140 sequence modifications including loop-swap chimeras and point mutants. When simulated integration efficiency is used to predict experimental expression of the combined the point mutant and loop-swap chimera datasets, a sequence predicted to increase in integration efficiency is almost four times more likely to increase in experimental expression than a sequence predicted to decrease in integration efficiency, as determined from the diagnostic odds ratio [75] taken over the set of 140 sequences.

The relationship between changes in simulated integration efficiency and IMP experimental expression due to sequence mutations provides a promising tool for predicting expression. Integration into the cell-membrane in the correct multi-spanning topology is a key step in IMP biogenesis. Our work demonstrates that the efficiency of integration in the correct topology can be affected by sequence modification, with a corresponding effect on IMP expression. In particular, for the IMP TatC, localization of the C-tail, quantified both using CG simulations and *in vivo*, is found to be sensitive to

sequence modifications throughout the coding sequence, and is shown to be predictive of experimental expression. The effect of sequence modifications on simulated integration efficiency and expression levels was found to be largely independent, enabling the design of larger sequence modifications with further enhanced expression. Broad applicability of simulated integration efficiency as a predictor of expression for other IMPs has yet to be established. However, we demonstrate that a small pool of test-data would have been sufficient to identify the predictive capacity of the C-tail for TatC, and similar methodology could be used to identify predictive topology features in other IMPs. The workflow established here enables IMP expression via rational improvement of co-translational integration, a key step in IMP biogenesis.

For any IMP that has a domain prone to mislocalize *in vivo*, simulating the effect of sequence modifications on the domain can be used as a forward predictor of expression outcomes and sequence modifications that aid localization may improve expression. In contrast to previous attempts to boost the expression of IMPs [20, 32, 37], the current study is able to identify and improve a specific step in IMP biogenesis. By simulating and enhancing the process of translocon-mediated integration *in silico*, the determinant of expression enhancement can be identified and directly addressed *in vivo*.

Acknowledgments

I would personally like to thank Michiel J. M. Niesen, William M. Clemons, Jr., and Thomas F. Miller, III for contributing to the project. Work for this project was supported by NIH-NIGMS grant 1R01GM125063 to TFM and WMC. Work in the Clemons lab was supported by an NIH Pioneer Award to WMC (5DP1GM105385) and funds from Caltech's Center for Environmental Microbial Interactions and an NIH training grant (NIH/NRSA training grant 5T32GM07616) to SSM. Work in the Miller group is supported in part by the Office of Naval Research (N00014-10-1-0884) and computational resources were provided by the National Energy Research Scientific Computing Center (NERSC) a DOE Office of Science User Facility (DE-AC02-05CH11231) and the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575.

Methods

Cloning

All TatC coding sequences were created using either primer extension or were synthesized by Twist Bioscience. Loop-swap chimeras were limited to those swapping loops 1-5 and 7, avoiding the short loop 6. The pool of 111 loop-swap chimera sequences were chosen from all 540 possible combinations. Each wild-type homolog is used between 6 to 15 times as a parent, between 7 to 19 times as a source for the mutant loop, and each loop is mutated between 8 to 29 times. Point mutants were chosen to affect a change in charge through mutation of neutral residues to charged residues or through mutation of charged residues to the opposite charge. Each loop-swap chimera coding sequences was cloned into the pET28(a+)-GFP-ccdB vector [22, 51] using the Gibson cloning protocol [53], resulting in each IMP possessing a C-terminal GFP tag. For constructs containing the β-lactamase tag, the GFP sequence was replaced with a β-lactamase sequence using Gibson cloning. For constructs containing the N-terminal Strep tag, the GFP and poly-His sequence was removed during PCR and the Strep tag was added using primer extension; the final vectors were constructed using Gibson cloning.

Heterologous Expression in E. coli

Heterologous expression of IMPS in *E. coli* was performed as previously described [22]. In short, IMPs were expressed in BL21 Gold (DE3) cells at 16°C for approximately 16 hours prior to either flow cytometry, western blot, or ampicillin survival analysis.

Flow Cytometry

Flow cytometry was performed as previously described [22]. In short, cultures of cells expressing TatC IMPs with a C-terminal GFP tag were resuspended in PBS and subjected to flow cytometry. Whole cell fluorescence from the B1/FITC channel was measured using a MACSQuant10 Analyzer. Mean fluorescence values are calculated using FlowJo.

Western Blot

All samples of cells expressing IMPs with an N-terminal Strep tag were subjected to the following protocol for western blot analysis. Samples were normalized to an OD_{600} of 3.0 in PBS and subjected to three freeze thaw cycles using liquid nitrogen and applied to 10% SDS-PAGE followed by western blotting. Relative protein levels were determined by incubation of the western blot membrane with an anti-Strep tag primary rabbit antibody followed by incubation with an IRDye® 800CW Donkey anti-rabbit secondary antibody and visualization using a LI-COR IR western blot scanner. Relative band intensities were quantified using ImageJ [55].

Description of the CG Simulations

We apply a previously developed coarse grained simulation approach [22, 36], capable of simulating the minute-timescale dynamics of co-translational integration via the Sec translocon. The CG model is applied and implemented exactly as described in [22], and key features of the CG model are provided here; for more a more extensive description, the reader is referred to [22].

The CG simulations explicitly describe the configurational dynamics of the nascent chain (NC), conformational gating of the Sec translocon lateral gate, and ribosomal translation (at 24 residues/second). The nascent chain (NC) is represented as a freely jointed chain of CG beads, where each CG bead represents three amino acids and has a diameter of 8Å, equal to the Kuhn length of a polypeptide chain [56, 57]. To avoid a frameshift in the mapping of amino acids to CG beads upon a loop-swap sequence modification, dummy atoms were introduced to keep mapping consistent, as done previously [22]. Bonding interactions between neighboring CG beads are described using the finite extension nonlinear elastic (FENE) potential [58], short-ranged non-bonding interactions are modeled using a Lennard-Jones potential, and electrostatic interactions are modeled using the Debye-Hückel potential. Periplasmic binding is included as described in [36, 47] for BiP, and solvent interactions are described using a position-dependent potential based on the water-membrane transfer free energy for each CG bead [22].

The configuration of the NC is time evolved using overdamped Langevin dynamics, with the CG beads confined to a two-dimensional subspace that runs along the axis of the translocon channel and between the two helices of the LG. Conformational gating of the LG corresponds to the LG helices moving out of the place of confinement for the NC, allowing the NC to pass into the membrane bilayer. The rate of stochastic LG opening and closing is dependent on the sequence of the CG beads that occupy the translocon channel [36, 47]. Ribosomal translation is directly simulated via growth of the NC at the ribosomal exit channel; throughout translation, the C-terminus of the NC is held fixed, and new beads are sequentially added at a rate of 24 residues per second. Upon completion of translation, the C-terminus is released from the ribosome.

Trajectories use a step-size of 100 ns for time integration and are terminated 31s after the end of translation. For each protein sequence, at least 400 independent trajectories are calculated.

Determination of Topology Features

The topology features for a protein sequence are determined as described previously [22]. The topology of a protein is analyzed over the last 6s of the CG simulation trajectories, starting 25s after the end of protein translation by the ribosome. For each loop, i, the location of the loop during this time-window is described by a variable λ_i, where λ_i=1 if the loop is in the cytosol, λ_i=-1 if the loop is in the periplasm, and λ_i=0 otherwise. For each trajectory we assess if, during the analysis time-window, a given topology feature is observed. The topology features used in this work are either; ($p^{(1)}$- $p^{(7)}$, $p^{(N)}$) localization of single loops consistent with the known TatC topology (Figure 3.1A), or ($p^{(All)}$) simultaneous localization for all loops consistent with the known topology.

Ampicillin Survival Assay

The ampicillin survival assay was performed as previously described [22]. In short, cells that had expressed IMPs with a C-terminal β-lactamase-tag overnight at 16°C were resuspended to an OD_{600} of 0.1 and grown to an OD_{600} of 0.5, after which ampicillin was added and cells were incubated for a further 1.5 hours, followed by plating on kanamycin LB agar plates. The relative number of observed colonies between loop-swap chimera and wild-type was used to determine the change in C-tail translocation, with a ratio greater than

one representing an increase in translocation of the C-tail to the periplasm due to the sequence modification.

Statistical Significance Calculations

Experimental expression, survival, and N-Strep values reported represent average values over at least 3 independent trials, error bars indicate the standard error of the mean unless otherwise noted. Simulated integration values represent the average outcome of at least 400 independent CG simulations trajectories, error bars indicate the standard error of the mean. Confidence intervals on AUC values are determined by bootstrapping [76]. 1,000,000 samples of simulated integration and expression pairs, with size equal to the set of sequence modification, are drawn with replacement from the set of sequence modifications. An AUC value is calculated for each sample, and the relevant percentile of the resulting AUC value distribution determines the confidence intervals.

C h a p t e r 4

APPLICATION OF SIMULATED INTEGRATION EFFICIENCY TO THE PREDICTION OF ERROR-PRONE MUTANT HAEMOPHILUS INFLUENZAE GLPG EXPERIMENTAL EXPRESSION

Abstract

Integral membrane proteins are key targets for biochemical and structural characterization, but heterologous overexpression often provides insufficient yield, in many cases without an indication as to the source of the failure. Our earlier work shows that we can improve the expression of TatC proteins in *E. coli* by increasing the efficiency of topogenesis, as predicted by a coarse-grained co-translational integration simulation model, by preventing misintegration of the TatC C-tail (Chapters 2 and 3). To assess the predictive power of the model which respect to changes in expression levels due to sequence mutation in a different protein family, a library of mutated *Haemophilus influenzae* (*Hi*) GlpG sequences is created using error-prone PCR and tested for experimental expression levels of membrane protein-GFP fusions and simulated integration efficiency effects. *Hi*GlpG, like TatC, contains loop domains with simulated integration efficiencies that may be predictive of experimental expression improvement, with the reservation that the cutoff most predictive of expression improvement is not consistent and not always 1.0 as expected. Preliminary evidence suggests that the application of model can be expanded to predict the expression of GlpG and other integral membrane proteins, but further testing is needed to clarify the existing issues.

Introduction

For the integral membrane protein (IMP) TatC, C-tail integration efficiency was determined using *in silico* coarse-grained modeling of IMP integration at the translocon (CG model) and changes in the simulated integration efficiency of the C-tail due to sequence modifications, including loop-swap chimeras and point mutants, was predictive of experimental expression. A number of questions remained unresolved. Could the CG model predict the expression effects of mutations on another IMP? Would simulated integration efficiency of the C-tail be predictive for another IMP family or would another loop be more predictive? Our previous work demonstrated that simulated integration efficiency was a strong predictor of experimental expression (Chapters 2 and 3), but a study utilizing an alternative IMP was needed to determine its broader relevancy.

To expand upon previous links between simulated integration efficiency and experimental expression, GlpG, a rhomboid protease widely found in bacteria that catalyzes intramembrane proteolysis [77], was chosen for analysis. It represents an ideal choice due to its size of only six transmembrane domains (TMD), the absence of large N- or C-terminal loop domains, and the availability of *in vivo* assays for measuring activity [77]. Instead of point mutants and loop swaps between homologs as previously used, a single homolog, *Haemophilus influenzae* (*Hi*) GlpG, is chosen for error-prone PCR mutagenesis and subsequent expression testing due to the ambiguity of loop and TMD domain identification. The error-prone PCR *Hi*GlpG library exhibits a wide range of sequence modifications and resulting experimental expression levels. The area under the curve (AUC) of the receiver operating characteristic (ROC) curve for the simulated integration efficiency of loop 4 is significant but $p^{(4)}$ predicts poorly at a cutoff of 1.0, while

loop 1 simulated integration efficiency AUC is not significant but $p^{(1)}$ predicts better than $p^{(4)}$ at a cutoff of 1.0. Further experiments are needed to clarify the relationship between simulated integration efficiency features and experimental expression, determine if *Hi*GlpG activity is uncompromised by mutations that increase integration efficiency, and assess whether the simulated integration efficiencies of the same loops are predictive of sequence modification in other GlpG homologs. Confirmation of all these outstanding issues will provide convincing proof of the utility of simulated integration efficiency in predicting and improving heterologous overexpression of a wider range of IMPs in *E. coli*.

Results

Wild-type Expression Levels

Figure 4.1B shows the measurement of the expression levels of five GlpG homologs compared to *Aa*TatC, which has been previously identified as a high-expressing IMP, needed to identify an ideal target for integration optimization that has low starting experimental expression. Experimental expression is quantified by the mean whole-cell fluorescence measured using flow cytometry as previously applied [22]. Unlike for TatC (Figure 2.1B), expression is consistently high among all GlpG homologs.

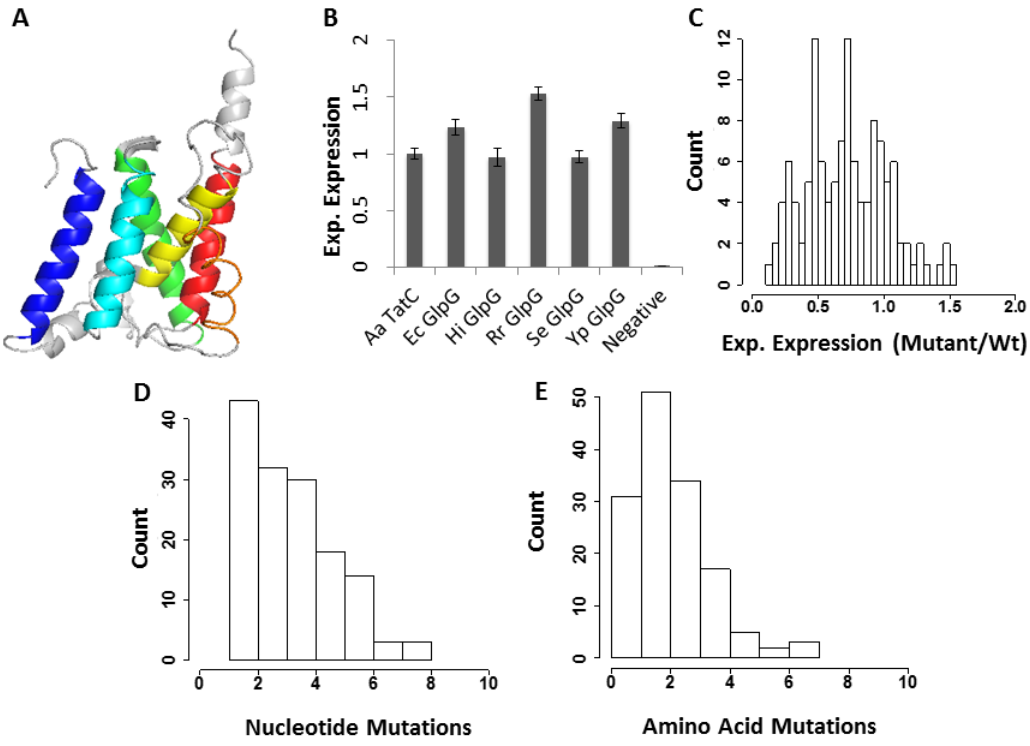


Figure 4.1: Hi*GlpG Experimental Expression.* (A) A ribbons diagram of the structure of *Hi*GlpG (RCSB PDB: 2NR9) with TMDs colored. (B) Experimental expression levels of wild-type GlpGs normalized to *Aa*TatC expression levels. Expression levels for all wild-type GlpGs tested are relatively high and consistent compared to the range observed for TatCs. (C) The distribution of experimental expression values (mutant/wild-type) for the

*Hi*GlpG error-prone PCR library. The majority of mutations have a negative effect on expression. (D) The distribution of the number of nucleotide mutations in the error-prone PCR library, excluding sequences with stop codons. (E) The distribution of the number of amino acid mutations in the error-prone PCR library, excluding sequences with stop codons. Those that do not affect a change in amino acid sequence are shown here but not included for further analysis.

Figure 4.2 indicates that, for TatC homologs, improvement in experimental expression (mutant/wild-type > 1.0, data from Chapter 3) is negatively correlated with the wild-type expression levels (wild-type/*Aa*TatC) (Pearson correlation = -0.2±0.2, Spearman rank correlation = -0.4±0.2), indicating that high expressing IMPs have a lower capacity to improve in expression. If the condition holds for GlpG homologs, it is expected that using any of the wild-type GlpG homologs tested in Figure 4.2B for the generation of a mutant library will lead to a smaller proportion of mutants improving in expression due to the high initial expression levels. This could be due to the high *in vivo* integration efficiency of the wild-type IMP subdomains, which may provide less opportunity for improvement by sequence mutation. *Hi*GlpG was chosen as the best candidate for expression enhancement, even given the high starting experimental levels observed, because a high-resolution structure is available (Figure 4.1A) [78] and it is among the lowest expressing GlpGs tested.

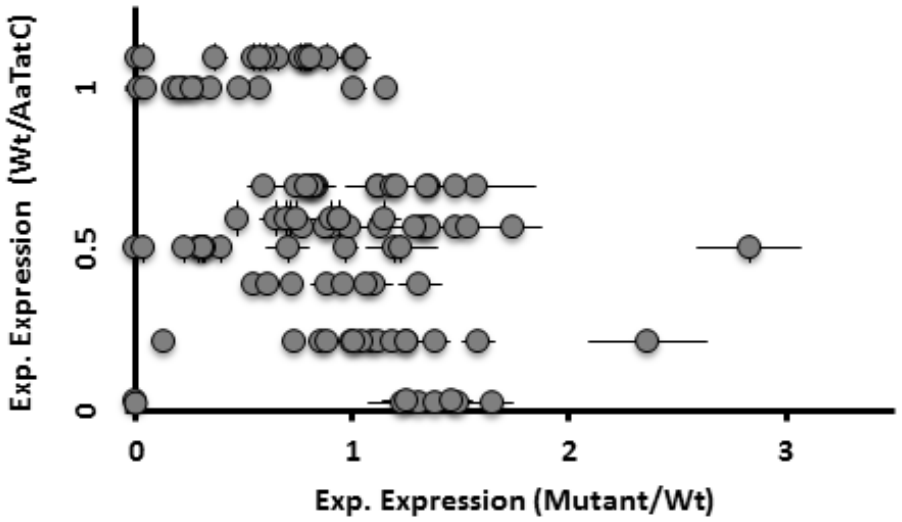


Figure 4.2: Correlation Between Expression Improvement and Wild-type Expression Levels for TatC. Wild-type expression levels (Wt/*Aa*TatC) and loop-swap chimera experimental expression levels (Mutant/Wt) are negatively correlated (Pearson correlation = -0.2±0.2, Spearman rank correlation = -0.4±0.2). Similarly, as wild-type expression increases, the proportion of mutants for a given homolog that increases in expression decreases. Experimental expression (mutant/wt) values greater than 1.0 indicate an increase in expression while those that are less than 1.0 represent a decrease in expression due to mutation.

Hi*GlpG Error-Prone PCR Library Generation and Expression*

The loop domains of *Hi*GlpG are difficult to define given that the loops identified by the OPM database lie at least partially within the membrane [78], making the synthesis of loop-swap chimeras among multiple homologs difficult. As an alternative, error-prone PCR is used to create a library of *Hi*GlpG mutants with one or more point mutations that, in contrast to loop-swap chimeras and other mutants created for testing TatC in Chapters 2 and 3, are not limited to the loop domains but can also occur in TMDs. A library of over 100 error-prone PCR mutated *Hi*GlpG sequences provides expression data that could be used to determine which simulated integration efficiency features (e.g. the localization of soluble loops) correlate with experimental expression. Figures 4.1D and 4.1E show the distribution of number of nucleotide and amino acid mutations, respectively, in the library,

excluding sequences that contained internal stop codons. Figure 4.1C displays the distribution of the experimental expression levels of the mutant sequences normalized to wild-type *Hi*GlpG levels. The majority of mutations exhibit drop in experimental expression compared to the wild-type sequence. This could be due to a higher initial integration efficiency for *Hi*GlpG, observable by the initial high wild-type expression level. In this case, mutations are not limited to loop domains and are observed in all domains of the sequence and, as seen in the library expression distribution (Figure 4.1C), there are significant perturbations to the wild-type expression yield by the introduced mutations.

Evaluating Simulated Integration Efficiencies for Predicting Expression

The receiver operating characteristic (ROC) curves of the two most predictive features for experimental expression: the simulated integration efficiency of loop 1 (Figure 4.3A) and of loop 4 (Figure 4.3B) provide a measure of the capacity for the simulated integration efficiency of loop 1 and loop 4 to predict mutant *Hi*GlpG experimental expression over a range of simulated integration efficiency cutoffs. Cutoffs are labeled inside the curve. The simulated integration efficiency of the wild-type and *Hi*GlpG sequences are determined using the CG model as previously described [22] and compared to the experimental expression data to find which simulated integration efficiency features correlate with experimental expression. As was the case for TatC, the integration efficiency of the *Hi*GlpG loop 6 was not calculated due to its small size [22, 78]

The area under the curve (AUC) of the ROC curves for the full set of simulated integration efficiency features in Figure 4.3C establishes that loop 1 ($p^{(1)}$) and loop 4 ($p^{(4)}$) are most predictive of experimental expression effects. The alternative measures of IMP

integration that are considered include $p^{(i)}$, the fraction of coarse-grained trajectories for which soluble loop i reaches correct localization with the respect to the membrane, and $p^{(All)}$, the fraction of coarse-grained trajectories for which all soluble loops reach correct localization. The simulated integration efficiency of loop 1 appears to be predictive at a simulated integration efficiency cutoff of 1.0, but the AUC of the ROC curve is not statistically significant. Conversely, the AUC of $p^{(4)}$ is statistically significant at a 95% confidence interval, but at a cutoff of 1.0 it does not appear to have predictive power, apparent by the proximity of the ROC curve to the midline at the 1.0 cutoff. The cutoff of 1.0 for simulation integration efficiency would be epected if the CG model correctly idetnfies cases of experimental expression success or failure, as a change in simulated integration effiency would signal a corresponding change in expeirmetal expression in the same direction (increase or decrease), though not necessarily with the same magnitude. The presence of loops that have the potential to be predictive of integration efficiency in another IMP demonstrates that there is opportunity for a broader application of the CG model for the prediction of the effects of sequence mutation on experimental expression by optimizing integration efficiency, but the evidence is not as conclusive and consistent at that seen for the TatC C-tail.

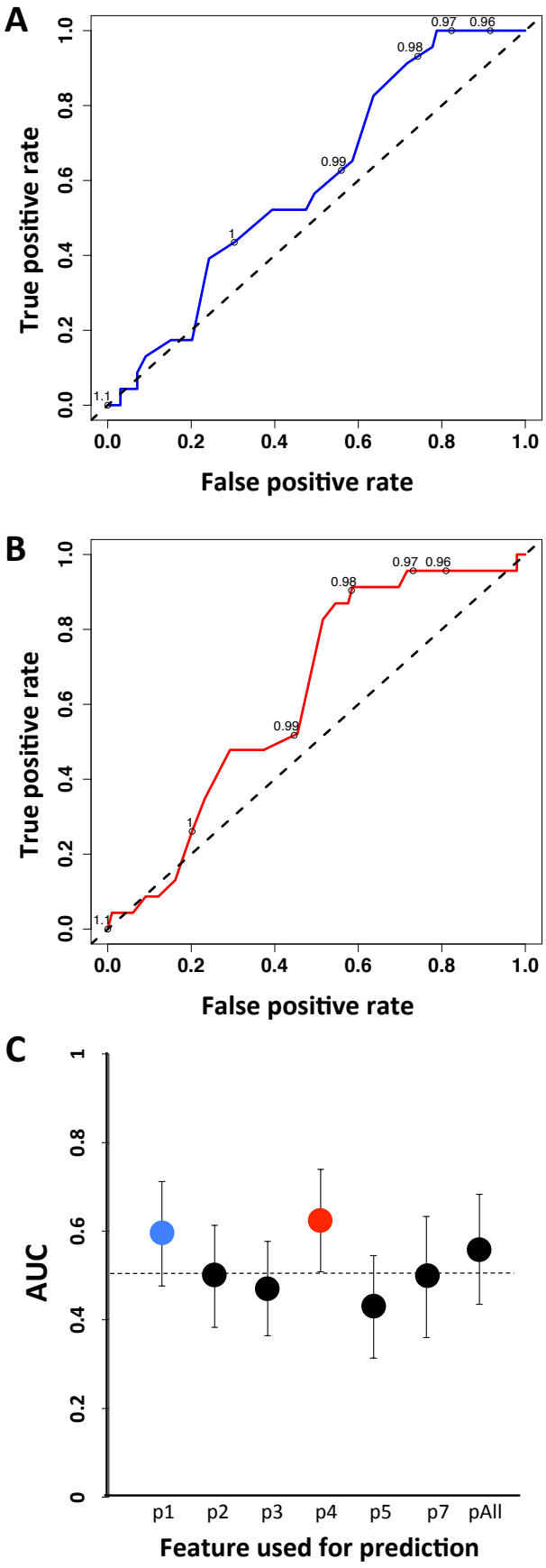

A
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.97
0.96
0.98
0.99
1
1.1
B
True positive rate
False positive rate
0.97
0.96
0.98
0.99
1
1.1
C
AUC
0
0.2
0.4
0.6
0.8
1
p1
p2
p3
p4
p5
p7
pAll
Feature used for prediction

Figure 4.3: Predictive Capacity of Simulated Integration Efficiency Features on Hi*GlpG Experimental Expression.* (A) An ROC curve displaying the capacity for the simulated integration efficiency of loop 1 ($p^{(1)}$, blue) to predict *Hi*GlpG mutant experimental expression over a range of simulated integration efficiency cutoffs. Cutoffs are labeled inside the curve. (B) An ROC curve displaying the capacity for the simulated integration efficiency of loop 4 ($p^{(4)}$, red) to predict *Hi*GlpG mutant experimental expression over a range of simulated integration efficiency cutoffs. Cutoffs are labeled inside the curve. (C) AUC of the ROC curves for the features $p^{(1)}$, $p^{(2)}$, $p^{(3)}$, $p^{(4)}$, $p^{(5)}$, $p^{(7)}$, and $p^{(All)}$ predicting *Hi*GlpG mutant experimental expression effects. $p^{(1)}$ and $p^{(4)}$ are the most significantly predictive features. Error bars represent 95% confidence intervals for each AUC value. The sequences considered contained at least one amino acid mutation and did not contain internal stop codons

Discussion

The results of comparing experimental expression and CG model derived simulated integration efficiencies demonstrate a link between the effects of random mutagenesis on the simulated integration efficiencies of loops 1 and 4 and the amount of *Hi*GlpG correctly folded. However, the AUC of $p^{(1)}$ is not significant at a 95% confidence interval and the cutoff of 1.0 is not predictive for $p^{(4)}$. These issues make it difficult to interpret the observed relationship. The relatively high expression level of wild-type *Hi*GlpG could limit the potential increase in expression and indicate high initial integration efficiency, limiting the potential for improvement. This is supported by the observation that *Hi*GlpG exhibits a smaller proportion of mutated sequences that improve in expression (19%) than the average of the TatCs single-loop-swap chimeras tested (45%). The small proportion of improved expression levels for the tested *Hi*GlpG library limits the statistical significance of the predictive capacity of simulated integration efficiency features.

In the future, several changes could be implemented to better assess the predictive power of simulated integration efficiency for experimental expression of IMPs other than TatC. More GlpG homologs or another protein family could be tested to determine if there is a different wild-type homolog that expresses poorly, ideally due to a lower starting integration efficiency that can be more predictably improved through mutation. Additionally, we could create a library of mutated sequences for a different GlpG and determine if the simulated integration efficiency of loop 1 and loop 4 are consistently predictive of experimental expression even though they are suspect in *Hi*GlpG. We expect that further investigation will provide additional evidence to support the link between IMP

simulated integration efficiency and experimental expression established in Chapters 2 and 3 and the generalizability to other IMP families.

Acknowledgments

I would like to thank Dr. William M. Clemons, Dr. Thomas F. Miller, and Michiel J. M. Niesen for their substantial contributions to the development of the project idea and analysis of resulting data. Michiel J. M. Niesen performed the CG model simulations and calculated the resulting simulated integration efficiencies. Shyam Saladi and Matthew Zimmer provided helpful comments. Dr. Yoshinori Akiyama supplied resources and expertise in reproducing an *in vivo* GlpG activity assay. Work for this project was supported by NIH-NIGMS grant 1R01GM125063 to TFM and WMC. Work in the Clemons lab was supported by an NIH Pioneer Award to WMC (5DP1GM105385) and funds from Caltech's Center for Environmental Microbial Interactions and an NIH training grant (NIH/NRSA training grant 5T32GM07616) to SSM. Work in the Miller group is supported in part by the Office of Naval Research (N00014-10-1-0884) and computational resources were provided by the National Energy Research Scientific Computing Center (NERSC) a DOE Office of Science User Facility (DE-AC02-05CH11231) and the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575.

Methods

Design and Synthesis of Hi*GlpG Wild-type Sequences and Error-Prone Library*

Wild-type GlpG sequences were created by PCR amplification of the genes from genomic DNA followed by Gibson assembly for insertion into pET28(a+)-GFP-ccdB [42]. The *Hi*GlpG sequence was subjected to error-prone PCR using the GeneMorph II kit from Agilent, which allows for the average number of nucleotide mutations to be tuned by adjusting the template concentration. To create a pool of sequences with a moderate number of mutations as shown in Figures 4.1D and 4.1E, 50 ng of the template, pET28(a+)-GFP containing the wild-type *Hi*GlpG sequence (approximately 6500 nucleotides or 2000 kDa), was used. Error-prone *Hi*GlpG mutant sequences were cloned into the pET28(a+)-GFP-ccdB plasmid [22, 51] using Gibson assembly[53], which resulted in a pool of approximately 3600 unique sequences.

E. coli *Expression*

The plasmid library was transformed into BL21 Gold (DE3) cells and grown overnight on LB agar plates containing 50 µg/ml kanamycin after one-hour incubation. After overnight growth at 37°C, individual colonies were picked into a starter culture containing 200 µL 2xYT media (16 g Tryptone, 10 g Yeast Extract, and 5 g NaCl per liter H_20) with 50 µg/ml kanamycin in a 96 well plate (2.0 mL deep well block) and grown at 37°C with shaking. After the OD600 of the cultures reached approximately 1.0, 20 µL of each starting culture was added to an expression culture containing 1 mL 2xYT with kanamycin in a 96 well block. 200 µL of sterile 50% glycerol was added to each starter culture well and the block was saved at -80°C. The expression culture was grown to an

OD600 of 0.15 at 37°C with shaking, then grown at 16°C to an OD600 of 0.3, after which IPTG was added to a final concentration of 1 mM to induce IMP expression. Induced cultures were grown overnight for a further 16 hours prior to IMP quantification via flow cytometry. Expression of each mutant sequence was performed and quantified alongside the wild-type *Hi*GlpG sequence and a negative control expressing *Aa*TatC with no C-terminal GFP tag.

Flow cytometry

A 96 well plate was filled with 150 µL of 1x PBS in each well, to which 20 µL of overnight expression culture was added and subjected to flow cytometry. Whole cell fluorescence was measured using a MACSQuant10 Analyzer. Fluorescence at 488 nm was used as the measure of expression yield. Flow cytometry data analysis was performed with FlowJo Software. Expression values used for experimental expression calculation are mean whole cell fluorescence from flow cytometry.

Plasmid Purification

Stabbings from the saved -80°C starter culture were used to inoculate 1 mL 2xYT with kanamycin in a 96 well block and grown overnight for 16 hours at 37°C with shaking. Cells were pelleted via centrifugation and plasmid was purified via the Macherey-Nagel 96 well NucleoSpin plasmid purification kit using a Tecan Freedom EVO liquid handling robot prior to sequencing. Sequences were analyzed to assure they contained at least one amino acid mutation and did not contain internal stop codons. Only sequences fitting these criteria were considered for analysis in Figure 4.3.

Data Analysis

Experimental expression values (mutant/wild-type) were calculated by dividing the mutant fluorescence value by the wild-type *Hi*GlpG fluorescence value for that experiment on that day. Confidence intervals were calculated using bootstrapping [76]. ROC curves and associated AUCs were calculated using the ROCR package in R [79].

Coarse-grained simulations and simulated integration efficiency calculations are performed using the same protocol described in Chapter 3 Methods [22].

FUTURE DIRECTIONS

The combined results of Chapters 2, 3, and 4 demonstrate that simulated integration efficiency calculated from a coarse-grained co-translational simulated integration model (CG model) can be used to predict changes in experimental expression resulting from mutation of an integral membrane protein (IMP) sequence; those that prevent misintegration enhance expression. However, the process we used to determine the simulated integration efficiency feature with significant predictive capacity, such as the C-tail for TatC, requires a large testing set of experimental outcomes to which integration features are compared ex post. Changes in simulated integration efficiency due to sequence modification correlate with the effect on relative experimental expression (mutant/wild-type) but do not agree with the absolute change in expression levels, indicating the CG model may not be properly calibrated to calculate these effects. These shortcomings limit the potential wider application of the model but provide an opportunity for further development.

Ideally, simulated integration efficiency could be adapted to yield three additional functions: the ability to, (*a*), predict the wild-type expression level of an IMP as compared to other homologs, (*b*), to determine which integration features are significantly predictive without the need experimental expression levels from a set of mutant sequences, and, (*c*), provide that the degree of the change in simulated integration efficiency scales with the fold change in experimental expression. Previous efforts to find a method to reach these goals have been unsuccessful. For example, analysis of the IMP structure could not be used to identify the most predictive loop *a priori* for TatC and GlpG [41, 43, 78]. While the short loop 6 of TatC is a possible source of the predictive capacity of the simulated

integration efficiency of loop 7, no such obvious adverse conformation is apparent in *Hi*GlpG that could be the source of the sensitivity of loops 1 and 4. *Hi*GlpG contains a short loop 6 that is similar to the short TatC loop 6, but the simulated integration efficiency of loop 7 is not predictive in *Hi*GlpG. Wild-type TatC C-tail integration efficiency does not correlate with the predictive capacity of the C-tail for that homolog and wild-type simulated integration efficiencies do not correlate with wild-type expression levels. Nonetheless, the current method provides a rational approach to increase the expression of IMPs through the identification and improvement of weak underlying processes in co-translational IMP integration and the predictive power of the CG model indicates that it does significantly capture the effects of mutations on expression through integration effects.

A deeper understanding of the precise molecular interactions that lead to misintegration is needed to perform predictions using the CG model without any experimental results available with which to interpret the effects of mutations on simulated integration efficiency. The addition of new features to the CG model could accomplish this. Currently, the movement of beads within the CG model as implemented herein is mostly limited to two dimensions. Development is underway to create a CG model that simulates co-translational integration in a 3-dimensional space. This will allow for a more detailed modeling of the ribosome and translocon shape and for the nascent chain to move in two dimensions within the plane of the membrane, instead of the one currently allowed. A complementary parameter that assesses the effect of mutations on the targeting of the IMP to the membrane would help form a more complete view of the biogenesis pathway. Also, while the SecYEG complex represents the core of the translocon and is necessary for

the integration of many IMPs, other chaperones such as YidC and SecDF can assist in co- and post-translational folding and effects can be implicitly or explicitly incorporated into the CG model [80-82]. The simplifications inherent to the current CG model can also be simulated in a more accurate manner, including the explicit modeling of the solvent and the membrane, using a less coarse-grained approach such as single amino acids per bead, using a more precise model of SecYEG and the ribosome, and incorporating co- and post-translational IMP folding. Further development of the CG model and more experimental data with which to test its effectiveness have the potential of further expanding the CG model and allow for computational methods to drive and predict mutations for the purpose of improving expression of an IMP, rather than using the CG model to assess the mechanism behind previously determined experimental expression levels.

BIBLIOGRAPHY

1. Fagerberg, L., et al., *Prediction of the human membrane proteome.* Proteomics, 2010. **10**(6): p. 1141-1149.

2. Yildirim, M.A., et al., *Drug-target network.* Nature Biotechnology, 2007. **25**(10): p. 1119-1126.

3. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-42.

4. Zhou, C., Y. Zheng, and Y. Zhou, *Structure Prediction of Membrane Proteins.* Genomics, Proteomics & Bioinformatics, 2004. **2**(1): p. 1-5.

5. Zhang, X. and S.O. Shan, *Fidelity of cotranslational protein targeting by the signal recognition particle.* Annual Review Biophysics, 2014. **43**: p. 381-408.

6. Akopian, D., et al., *Signal recognition particle: an essential protein-targeting machine.* Annual Review Biochemistry, 2013. **82**: p. 693-721.

7. Rapoport, T.A., *Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes.* Nature, 2007. **450**(7170): p. 663-669.

8. Berg, B.v.d., et al., *X-ray structure of a protein-conducting channel.* Nature, 2004. **427**(6969): p. 36-44.

9. Cymer, F., G. von Heijne, and S.H. White, *Mechanisms of integral membrane protein insertion and folding.* Journal of Molecular Biology, 2015. **427**(5): p. 999-1022.

10. Hessa, T., et al., *Recognition of transmembrane helices by the endoplasmic reticulum translocon.* Nature, 2005. **433**(7024): p. 377-381.

11. Boel, G., et al., *Codon influence on protein expression in E. coli correlates with mRNA levels.* Nature, 2016. **529**(7586): p. 358-63.

12. Do, H., et al., *The cotranslational integration of membrane proteins into the phospholipid bilayer is a multistep process.* Cell, 1996. **85**(3): p. 369-78.

13. Woodall, N.B., Y. Yin, and J.U. Bowie, *Dual-topology insertion of a dual-topology membrane protein.* Nature Communications, 2015. **6**: p. 8099.

14. Seppala, S., et al., *Control of membrane protein topology by a single C-terminal residue.* Science, 2010. **328**(5986): p. 1698-700.

15. Tu, L., et al., *Transmembrane Biogenesis of Kv1.3.* Biochemistry, 2000. **39**(4): p. 824-836.

16. Harley, C.A., et al., *Transmembrane Protein Insertion Orientation in Yeast Depends on the Charge Difference across Transmembrane Segments, Their Total Hydrophobicity, and Its Distribution.* Journal of Biological Chemistry, 1998. **273**(38): p. 24963-24971.

17. Heijne, G., *The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology.* EMBO J, 1986. **5**(11): p. 3021-7.

18. Van Lehn, R.C., B. Zhang, and T.F. Miller, *Regulation of multispanning membrane protein topology via post-translational annealing.* eLife, 2015. **4**: p. e08697.

19. Ismail, N., et al., *Charge-driven dynamics of nascent-chain movement through the SecYEG translocon.* Nature Structural and Molecular Biology, 2015. **22**(2): p. 145-9.

20. Wagner, S., et al., *Rationalizing membrane protein overexpression.* Trends Biotechnology, 2006. **24**(8): p. 364-71.

21. Huang, Y., et al., *Structure and Mechanism of the Glycerol-3-Phosphate Transporter from Escherichia coli.* Science, 2003. **301**(5633): p. 616-620.

22. Marshall, Stephen S., et al., *A Link between Integral Membrane Protein Expression and Simulated Integration Efficiency.* Cell Reports, 2016. **16**(8): p. 2169-2177.

23. Lewinson, O., A.T. Lee, and D.C. Rees, *The funnel approach to the pre-crystallization production of membrane proteins.* Journal of Molecular Biology, 2008. **377**(1): p. 62-73.

24. Dumon-Seignovert, L., G. Cariot, and L. Vuillard, *The toxicity of recombinant proteins in Escherichia coli: a comparison of overexpression in BL21(DE3), C41(DE3), and C43(DE3).* Protein Expression and Purification, 2004. **37**(1): p. 203-206.

25. Drew, D., et al., *GFP-based optimization scheme for the overexpression and purification of eukaryotic membrane proteins in Saccharomyces cerevisiae.* Nature Protocols, 2008. **3**(5): p. 784-798.

26. Drew, D., et al., *A scalable, GFP-based pipeline for membrane protein overexpression screening and purification.* Protein Science : A Publication of the Protein Society, 2005. **14**(8): p. 2011-2017.

27. Fluman, N., et al., *mRNA-programmed translation pauses in the targeting of E. coli membrane proteins.* eLife, 2014. **3**: p. e03440.

28. Wang, Z., et al., *Optimizing expression and purification of an ATP-binding gene gsiA from Escherichia coli k-12 by using GFP fusion.* Genetics and Molecular Biology, 2011. **34**: p. 661-668.

29. Guglielmi, L., et al., *Selection for intrabody solubility in mammalian cells using GFP fusions.* Protein Engineering, Design and Selection, 2011. **24**(12): p. 873-881.

30. Geertsma, E.R., et al., *Quality control of overexpressed membrane proteins.* Proceedings of the National Academy of Sciences, 2008. **105**(15): p. 5722-5727.

31. Daley, D.O., et al., *Global Topology Analysis of the Escherichia coli Inner Membrane Proteome.* Science, 2005. **308**(5726): p. 1321-1323.

32. Scott, D.J., et al., *Stabilizing membrane proteins through protein engineering.* Current Opinions in Chemical Biology, 2013. **17**(3): p. 427-35.

33. Heydenreich, F.M., et al., *Stabilization of G protein-coupled receptors by point mutations.* Frontiers in Pharmacology, 2015. **6**: p. 82.

34. Miroux, B. and J.E. Walker, *Over-production of proteins in Escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels.* Journal of Molecular Biology, 1996. **260**(3): p. 289-98.

35. Wagner, S., et al., *Tuning Escherichia coli for membrane protein overexpression.* Proceedings of the National Academy of Sciences, 2008. **105**(38): p. 14371-14376.

36. Zhang, B. and T.F. Miller, *Long-timescale dynamics and regulation of Sec-facilitated protein translocation.* Cell reports, 2012. **2**(4): p. 927-937.

37. Schlegel, S., et al., *Revolutionizing membrane protein overexpression in bacteria.* Microbial Biotechnology, 2010. **3**(4): p. 403-411.

38. Scott, D.J., et al., *Stabilizing membrane proteins through protein engineering.* Current Opinion in Chemical Biology, 2013. **17**(3): p. 427-435.

39. Goder, V. and M. Spiess, *Molecular mechanism of signal sequence orientation in the endoplasmic reticulum.* The EMBO Journal, 2003. **22**(14): p. 3645-3653.

40. Bogsch, E.G., et al., *An Essential Component of a Novel Bacterial Protein Export System with Homologues in Plastids and Mitochondria.* Journal of Biological Chemistry, 1998. **273**(29): p. 18003-18006.

41. Ramasamy, S., et al., *The glove-like structure of the conserved membrane protein TatC provides insight into signal sequence recognition in twin-arginine translocation.* Structure, 2013. **21**(5): p. 777-88.

42. Waldo, G.S., et al., *Rapid protein-folding assay using green fluorescent protein.* Nature Biotechnology, 1999. **17**(7): p. 691-695.

43. Rollauer, S.E., et al., *Structure of the TatC core of the twin-arginine protein transport system.* Nature, 2012. **492**(7428): p. 210-214.

44. Sievers, F., et al., *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* Molecular Systems Biology, 2011. **7**: p. 539.

45. Tsirigos, K.D., et al., *The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides.* Nucleic Acids Research, 2015. **43**(W1): p. W401-7.

46. Zhang, B. and T.F. Miller, *Direct Simulation of Early-Stage Sec-Facilitated Protein Translocation.* Journal of the American Chemical Society, 2012. **134**(33): p. 13700-13707.

47. Zhang, B. and T.F. Miller, *Hydrophobically stabilized open state for the lateral gate of the Sec translocon.* Proceedings of the National Academy of Sciences, 2010. **107**(12): p. 5399-5404.

48. Sarkar, C.A., et al., *Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity.* Proceedings of the National Academy of Sciences, 2008. **105**(39): p. 14808-14813.

49. Grisshammer, R., R. Duckworth, and R. Henderson, *Expression of a rat neurotensin receptor in Escherichia coli.* Biochemical Journal, 1993. **295**(2): p. 571-576.

50. Warne, T., et al., *Structure of a [bgr]1-adrenergic G-protein-coupled receptor.* Nature, 2008. **454**(7203): p. 486-491.

51. Drew, D.E., et al., *Green fluorescent protein as an indicator to monitor membrane protein overexpression in Escherichia coli.* FEBS Letters, 2001. **507**(2): p. 220-4.

52. Hoover, D.M. and J. Lubkowski, *DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis.* Nucleic Acids Research, 2002. **30**(10): p. e43.

53. Gibson, D.G., *Enzymatic assembly of overlapping DNA fragments.* Molecular Systems Biology, 2011. **498**: p. 349-61.

54. Noens, E.E., et al., *Improved mycobacterial protein production using a Mycobacterium smegmatis groEL1ΔCexpression strain.* BMC Biotechnology, 2011. **11**(1): p. 27.

55. Schindelin, J., et al., *The ImageJ ecosystem: an open platform for biomedical image analysis.* Molecular reproduction and development, 2015. **82**(7-8): p. 518-529.

56. Hanke, F., et al., *Stretching single polypeptides: The effect of rotational constraints in the backbone.* EPL (Europhysics Letters), 2010. **92**(5): p. 53001.

57. Staple, D.B., et al., *Model for Stretching and Unfolding the Giant Multidomain Muscle Protein Using Single-Molecule Force Spectroscopy.* Physical Review Letters, 2008. **101**(24): p. 248301.

58. Kremer, K. and G. S. Grest, *Dynamics of entangled linear polymer melts: A molecular‐dynamics simulation*. Vol. 92. 1990. 5057-5086.

59. Schaletzky, J. and T.A. Rapoport, *Ribosome Binding to and Dissociation from Translocation Sites of the Endoplasmic Reticulum Membrane.* Molecular Biology of the Cell, 2006. **17**(9): p. 3860-3869.

60. Potter, M.D. and C.V. Nicchitta, *Endoplasmic reticulum-bound ribosomes reside in stable association with the translocon following termination of protein synthesis.* Journal of Biological Chemistry, 2002. **277**(26): p. 23314-20.

61. Wimley, W.C., T.P. Creamer, and S.H. White, *Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides.* Biochemistry, 1996. **35**(16): p. 5109-24.

62. MacCallum, J.L. and D.P. Tieleman, *Hydrophobicity scales: a thermodynamic looking glass into lipid-protein interactions.* Trends in Biochemical Science, 2011. **36**(12): p. 653-62.

63. Gordon, E., et al., *Effective high-throughput overproduction of membrane proteins in Escherichia coli.* Protein Expression and Purification, 2008. **62**(1): p. 1-8.

64. Korepanova, A., et al., *Cloning and expression of multiple integral membrane proteins from Mycobacterium tuberculosis in Escherichia coli.* Protein Science, 2005. **14**(1): p. 148-158.

65. Lundstrom, K., *Latest development in drug discovery on G protein-coupled receptors.* Current Protein Peptide Science, 2006. **7**(5): p. 465-70.

66. Driessen, A.J. and N. Nouwen, *Protein translocation across the bacterial cytoplasmic membrane.* Annual Reviews in Biochemistry, 2008. **77**: p. 643-67.

67. Shao, S. and R.S. Hegde, *Membrane protein insertion at the endoplasmic reticulum.* Annual Review in Cellular Developmental Biology, 2011. **27**: p. 25-56.

68. Niesen, M.J., et al., *Structurally detailed coarse-grained model for Sec-facilitated co-translational protein translocation and membrane integration.* PLoS Computational Biology, 2017. **13**(3): p. e1005427.

69. Swets, J.A., R.M. Dawes, and J. Monahan, *Psychological Science Can Improve Diagnostic Decisions.* Psychological Science in the Public Interest, 2000. **1**(1): p. 1-26.

70. Serrano-Vega, M.J., et al., *Conformational thermostabilization of the β1-adrenergic receptor in a detergent-resistant form.* Proceedings of the National Academy of Sciences, 2008. **105**(3): p. 877-882.

71. Magnani, F., et al., *Co-evolving stability and conformational homogeneity of the human adenosine A2a receptor.* Proceedings of the National Academy of Sciences, 2008. **105**(31): p. 10744-9.

72. Schlinkmann, K.M., et al., *Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations.* Proceedings of the National Academy of Sciences, 2012. **109**(25): p. 9810-9815.

73. Altschul, S.F., et al., *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **215**(3): p. 403-10.

74. Marks, D.S., T.A. Hopf, and C. Sander, *Protein structure prediction from sequence variation.* Nature Biotechnology, 2012. **30**(11): p. 1072-1080.

75. Glas, A.S., et al., *The diagnostic odds ratio: a single indicator of test performance.* Journal of Clinical Epidimiology, 2003. **56**(11): p. 1129-35.

76. Carpenter, J. and J. Bithell, *Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians.* Statistical Medicine, 2000. **19**(9): p. 1141-64.

77. Maegawa, S., K. Ito, and Y. Akiyama, *Proteolytic action of GlpG, a rhomboid protease in the Escherichia coli cytoplasmic membrane.* Biochemistry, 2005. **44**(41): p. 13543-52.

78. Lemieux, M.J., et al., *The crystal structure of the rhomboid peptidase from Haemophilus influenzae provides insight into intramembrane proteolysis.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(3): p. 750-754.

79. Sing, T., et al., *ROCR: visualizing classifier performance in R.* Bioinformatics, 2005. **21**(20): p. 3940-1.

80. Tsukazaki, T. and O. Nureki, *The mechanism of protein export enhancement by the SecDF membrane component.* Biophysics (Nagoya-shi), 2011. **7**: p. 129-133.

81. Tsukazaki, T., et al., *Structure and function of a membrane component SecDF that enhances protein export.* Nature, 2011. **474**(7350): p. 235-8.

82. Kiefer, D. and A. Kuhn, *YidC as an essential and multifunctional component in membrane protein assembly.* Internal Review of Cytology, 2007. **259**: p. 113-38.