Understanding the GAS_{right} motif:

sequence, structure, and stability

By

Samantha Marie Anderson

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Biochemistry)

at the

UNIVERSITY OF WISCONSIN – MADISON

2019

Date of final oral examination: October 18, 2019

The dissertation is approved by the following members of the Final Oral Committee:

Alessandro Senes, Associate Professor, Biochemistry
Katherine Henzler-Wildman, Associate Professor, Biochemistry
Srivatsan Raman, Assistant Professor, Biochemistry
Sushmita Roy, Associate Professor, Biostatistics and Medical Informatics
Julie Mitchell, Director, Biosciences Division, Oak Ridge National Laboratory

To Gram and Grump

Loving and supportive grandparents

Rose Newman	Ralph Newman
November 10, 1939 – May 6, 2018	August 20, 1940 – May 8, 2019

Table of Contents

Acknowledgments

There are many people that have contributed to my graduate school journey that I would like to acknowledge here. First and foremost, I would like to thank my advisor, Alessandro Senes. Alessandro has trained me to become a scientist in all of its forms including speaker, writer, graphic designer, programmer, experimentalist, and thinker. He has invested countless hours and funds into my experiments and it was incredibly flattering that he wrote part of a grant on experiments I designed. I will be forever grateful for the times he has challenged me to be a better scientist and supported me in my journey.

I would also like to thank my current and former committee members: Qiang Cui, Katie Henzler-Wildman, Julie Mitchell, Vatsan Raman, Sushmita Roy, and Doug Weibel. In particular, Katie and Vatsan are close mentors and constantly challenge me to think more deeply about my science.

The entire Senes lab has been critical to my success as a scientist. The lab environment is collaborative and friendly and I couldn't have asked for better colleagues for the last five years. This includes Claire Armstrong, Cai Kai, Beth Caselle, Sam Craven, Gladys Díaz-Vázquez, Rika Khadria, Loren LaPointe, Gilbert Loiseau, Deena-al Mahbuba. I would like to thank my collaborator and mentor, Ben Mueller for his support and training when I joined the lab and years later. I would also like to acknowledge all of my undergraduate and high school trainees who have supported my work throughout the years: Junda Chen, Maria Cisler, Amanda Cook, Lucy Jiang, Evan Lange, and Collin McFadden. Joshua Choi is my newest colleague and I am excited to train him to take on the projects outlined in Chapter 3 and beyond. Most of all, I would like to thank Samson Condon for his constant support as a friend and fellow scientist. We have been through some difficult personal circumstances that mirror each other and I have found great comfort in having him as a bay mate. He has been my closest collaborator and mentor, and he has always challenged me to be, and do, better.

Outside of the lab, I would like to thank the helpful community in the Raman lab that has taken me under their wing to train me in the methods outlined in Chapter 3, especially Megan Leander. In fact, the entire Integrated Program in Biochemistry (IPiB) community has been nothing but helpful over my time here. That includes Kate Ryan, the office staff, the media lab and IT department, and all of the graduate students I have come across while working in Madison.

While I have been in graduate school, I have expanded my interests beyond science to the realms of science policy and science diplomacy which is the direction my career will take after leaving Madison. The student groups of Catalysts for Science Policy (CaSP) and the National Science Policy Network (NSPN) were my main support in these interests. I have made so many friends and found role models in these networks including CP Frost, Caitlin Warlick-Short, and Avital Percher who invested in me. Without the funding of the Computation and Informatics in Biology and Medicine traineeship, I would not have been able to afford to attend many of the events associated with CaSP and NSPN, so I have sincere gratitude for that funding source.

I have many mentors in science, but two of the most important have been Sharon Crary and Dan Gurnon, professors at DePauw University. Both have been continuously supportive even after I left the university, including phone calls, recommendation letters, and emails. Sharon in particular is the person I owe much of my success to and many of my liberal arts interests. She is the reason I selected the University of Wisconsin at Madison for my summer research experience and thus why I was accepted to the IPiB program.

Graduate school would have been impossible to survive without the friendship of the 2014 IPiB class, Chris (and Ruth) Brandon, Allie Canales, Alex DeHaven, Kasia Dubiel (and Brian Duckmann), Zack Kemmerer, Mark Klein, Tina Lynch, Deena Al Mahbuba, Andy Voter, Erin Weisenhorn (and Andy Schrage), and Allyson Yake. We have had so much fun working and commiserating together and supporting each other with scientific discussion, mental health, and outside activities. I would especially like to thank Tina Lynch, my confidant and buddy for being

substantial emotional support and entertainment over the last five years and I hope our friendship continues well into our future. One of the most important people I have to thank is Kasia Dubiel. Kasia and I started in the same summer undergraduate research program at UW. She was my roommate for my entire graduate school experience and a dear friend. Always the person that I could come home to, to celebrate, to commiserate, to watch a TV show with. She taught me about new foods and evaluated my conference and date outfits. I spent more time with her than anyone else in the last few years and I couldn't have imagined my life without her.

I want to thank Zed Fashena, my partner over the last eight months, for providing me the necessary emotional support and stress relief during the difficult time leading up to my defense.

Finally, I need to thank my family for the unconditional support they have provided throughout my educational career. My cousins, Kelly and Bryan Yamakawa have been supportive my whole life, always checking in on me when I felt alone. I see their children, Gracie, Jimi, and Charlie, as an inspiration for why I work so hard for the future. My parents, Kathie and Dean Anderson, have always supported my dreams, in whatever form they took. They always trusted me to make the best decision for my future and laid my foundational values. My siblings, Ariane and Dino have become my best friends over the last five years. As a tell-tale sign of age, I learned that family is the most important connection I will ever have, my partners for life. Last, and most important, I would like to give my eternal thanks to my grandparents Ralph (Grump) and Rose Newman and Ken Blohm. My grandparents are the most proud of me and their support has been indispensable.

It took a village to get me here and I can only hope and can be part of all of their villages in the future.

Summary for the Tax Payers

This thesis is primarily focused on computational prediction and experimental validation of membrane protein structure. Membrane proteins make up a quarter of all genomes and half of all the drug targets currently on the market, yet only 3% of the solved protein structures are membrane proteins. This knowledge gap exists is because for decades, scientific techniques have been developed for soluble proteins. Like how vinegar and oil do not mix, neither do soluble and membrane proteins, rendering many established techniques futile.

To solve this problem, I used a computational method to predict possible structures of a specific set of membrane proteins to set up a testable hypothesis. This algorithm, CATM, was previously developed in our lab and accurately predicted a set of known protein structures. I further used CATM to predict the structure of an unknown plant protein important for plant hormore regulation. I anticipated, however, that it could be more powerful.

I hypothesized that CATM could predict the strength of association between two membrane proteins. Using an experimental assay, TOXCAT, I evaluated the strength of protein-protein interactions in the membrane. We assayed dozens of proteins and found that the energy scores used to rank the CATM models correlate with stability measurements. Furthermore, we found sequence, structure, and energetic trends that correspond with those stability measurements, rendering CATM a powerful tool for researchers.

The trouble with that set of experiments was that it took years to evaluate all of the proteins. To look at more of these proteins, and there are indeed many more to test, the existing assay would not be feasible. Therefore, in the second half of my graduate work, I developed a method that would increase the throughput of TOXCAT. I collaborated with another lab to bring sort-seq to the field of membrane proteins. Sort-seq brings together the new technologies of fluorescence-activated cell sorting and next-generation sequencing to evaluate tens of thousands of sequences at a time. I believe this method will be used in the Senes Lab and

elsewhere to test not only membrane protein structure, but also co-evolution hypotheses, protein design, and protein energetics.

Abbreviations

BiFC: bimolecular fluorescence complementation,

CAT: chloramphenicol acetyl transferase, **CATM**: C-alpha transmembrane, **CKX**: cytokinin oxidase/dehydrogenase, **co-IP**: co-immunoprecipitation,

DDM: dodecylmaltoside, **DMS**: deep mutational scanning, **dsTβL**: deep-sequencing TOXCAT-β lactamase

ER: endoplasmic reticulum, **ERAD**: endoplasmic reticulum-associated degradation,

FACS: Fluorescence-activated cell sorting, **FlAsH**: fluorescein arsenical helix, **FP**: fluorescent protein, **FRET**: Förster resonance energy transfer

GFP: green fluorescent protein, **GpA**: glycophorin A,

MBP: maltose binding protein, **MD**: molecular dynamics, **MSL**: molecular software library, **MP**: membrane protein

NEB: New England Biolabs, **NGS**: next-generation sequencing, **NMR**: nuclear magnetic resonance

PBS: phosphate buffer saline**, PCR**: polymerase chain reaction, **PDB**: Protein Data Bank

RFP: red fluorescent protein, **RFU**: relative fluorescence units, **RMSD**: root means square deviation **RTK**: receptor tyrosine kinase

SDS-PAGE: sodium dodecyl sulfate polyacrylamide gel electrophoresis, **SE-AUC**: sedimentation equilibrium analytical ultra-centrifugation, **sfGFP**: superfolder green fluorescent protein, **SPMP**: single-pass membrane protein

TM: transmembrane, **TMD**: transmembrane domain

WT: wild-type

Chapter 1: Introduction

1.1 Introduction to single-pass membrane proteins

Membrane proteins (MP) are a class of proteins that localize or interact with cellular membranes. MPs make up approximately 25-35% of most proteomes and constitute over half of all drug targets currently on the market (Klabunde and Hessler, 2002). There are two main types of MPs: peripheral MPs are primarily soluble and transiently associate with the membrane while integral MPs are embedded in the membrane. Integral MPs include the three classes of single-pass, multi-pass, and β-barrel. Multi-pass MPs cross the bilayer with two or more α-helices and β-barrel MPs are made of multiple β-strands that form a closed cylinder-like sheet, creating an open pore in the membrane. Here, I focus on the most prevalent type of integral MPs, the single-pass membrane proteins (SPMP)–i.e. those that span the membrane bilayer with a single transmembrane (TM) α-helix– that constitute 50% of all MPs (Fagerberg et al., 2010; UniProt Consortium, 2015).

MPs commonly act as mediators between cells and their environment by serving as gatekeepers, transporters, receptors, and enzymes. Despite their abundance and biological significance, MPs make up only 3.7% of the structures deposited in the Protein Data Bank (Sept 2019; PDB). Experimental characterization of these interactions remains difficult primarily due to the challenges of insoluble proteins and the necessity for membrane mimetics. As a result, structure determination and protein folding studies lag behind those of their soluble counterparts. Despite the overall simplistic character of a SPMP, to date, not a single full-length SPMP structure has been solved. This gap is likely due to the mixed hydrophobic and hydrophilic nature of full-length SPMPs. Without MP structures, biological experimentation and pharmaceutical discovery have moved at a much slower pace. It is therefore necessary to find alternative methods to understand MP structures.

The majority of my graduate work was spent developing new experimental techniques and validating existing computational methods to study SPMPs. In this introductory chapter, I begin

by explaining the biological importance of SPMPs. I then introduce the importance of sequence and structure motifs of SPMPs—in particular the sequence and structural motifs, GxxxG and GAS_{right}. I will then review the current methods for studying SPMPs and end with a review of emerging technologies that can be adapted to the MP problem.

1.1.1 Biological importance of single-pass membrane proteins

Historically, the TM domains (TMD) of SPMPs were thought to simply provide a hydrophobic anchor for the more important globular domains inside and outside the cell. Research has shown, however, that many TMDs are functional actors and that they can be critically involved in many diseases (Hubert et al., 2010). This notion is supported by the fact that the TMDs are the most conserved regions of SPMPs, even given that their amino acid library is overwhelmingly limited to hydrophobic amino acids (Zviling et al., 2007).

A specific structural feature of many SPMPs is their ability to oligomerize with other proteins embedded in the membrane, a process that is frequently driven by their TMDs. Acting in cooperation with the soluble domains, TMDs can mediate and modulate various oligomerizing systems. In the human proteome, there are more than 2,300 SPMPs with TMDs are annotated to range from 11 to 41 amino acids in length, providing high density of functionality in a short sequence (Bugge et al., 2016; Teese and Langosch, 2015; UniProt Consortium, 2015). SPMPs include oligomerizing systems such as receptor tyrosine kinases (Anbazhagan et al., 2010; Bocharov et al., 2008a, 2012a; Chung et al., 2010; Mineev et al., 2010a), cytokine receptors (Matthews et al., 2011; Vilar et al., 2009), integrins (Li et al., 2005; Yin et al., 2006), cadherins (Lai and Xu, 2007), apoptotic regulators (Bocharov et al., 2007; Lawrie et al., 2010; Sulistijo and MacKenzie, 2006), enzymes (Khadria et al., 2014a), and immunological complexes (Dixon et al., 2006). Misregulation of these associations is linked to diseases such as cancer, viral infection, and neurodegeneration (Roskoski, 2014). Therefore, it is necessary to understand not only how TMD oligomerization contributes to the functionality of proteins, but also what these oligomers look like to predict the effect of disease-causing mutations, design pharmaceutical

agents, and gain an understanding of the biological systems that mediate the interactions at the intra-extracellular interface.

1.1.2 Studying proteins through motifs

One compelling way to approach the problem of MP structure determination and folding is to study structural association motifs. Motifs are recurrent elements that include a signature sequence and a distinct structure. Biologically, motifs are extremely important because they correspond to structures that are particularly suited for common tasks (association, ligand or metal binding, catalysis, etc.). Understanding the rules that govern the folding and the variation of frequent motifs is key to understanding and predicting stability, function, recognition, and regulation in a variety of biological systems. Furthermore, once a motif is well-understood, it can be used as a scaffold for computational design or to predict the structure, and thus, the presumed function of uncharacterized proteins.

There have been many examples of motifs that have been characterized and subsequently used as a foundation for further experimentation. These include coiled coils (Lupas and Bassler, 2017; Truebestein and Leonard, 2016), zinc fingers (Liu et al., 2015), TIM barrels (Wierenga, 2001), and the DNA-binding helix-turn-helix motif (Aravind et al., 2005). Each of these motifs are now taught in basic biochemistry courses because they are building blocks for the study of a vast set of protein families with different functions.

This thesis is focused on understanding a particular motif that is responsible for helix-helix interactions in the membrane. The GAS_{right} structural motif (and consequently the GxxxG sequence motif) contain the important features of a motif including biological abundance, a signature sequence, and a distinct structure. Each of these features will be expanded upon throughout the rest of this chapter.

1.2 GxxxG: A prevalent sequence motif

1.2.1 Foundational studies

The GxxxG motif is a simple sequence, yet it has come to strongly indicate the presence of helix-helix interactions. The GxxxG motif is defined by the presence of two glycine amino acids, spaced four residues apart. Given the standard 3.6 amino acids per turn of an alpha helix, this places the glycines on the same face of the helix, creating a space where two helices can come very close together and create strong van der Waals packing (Javadpour et al., 1999). Sometimes, the GxxxG motif is expanded to include other small residues at the interface including alanine and serine, known as Sm-xxx-Sm or GxxxG-like.

The prevalence and importance of the GxxxG motif was first identified by a pair of publications from the Engelman lab in 2000. Senes et al. analyzed the presence of pairs and triplets of amino acids in a database of predicted TM helices, finding multiple patterns that defined the TM proteome. The most over-represented pair was GxxxG, by over 30% and a *p*-value of 6.4×10^{-34}. IGxxL was the most over-represented triplet and, it was postulated that these flanking β-branched residues reduce the entropic cost of folding because they lack rotameric freedom (Senes et al., 2000).

The concurrent paper from the Engelman lab assigned a potential function to this GxxxG motif using an association study. A set of standardized TM helices was designed so that one face of the helix was mutated to a random set of hydrophobic amino acids and the other side was fixed and non-associating. This is the same standardized, two face helix I use in Chapter 2. Using antibiotic selection in the TOXCAT assay, the authors isolated strongly associating TM helices and found that nearly 80% of them contained a GxxxG motif. Furthermore, the GxxxG was often flanked by the same β-branched residues found by Senes et al (Russ and Engelman, 2000). The TOXCAT assay will be expanded upon in Section 1.4.4.

Following these two foundational studies, and the realization that the GxxxG motif was found in the strongly dimerizing and well-studied Glycophorin A (GpA) (Lemmon et al., 1992), it became synonymous with dimerization (Liu et al., 2002; Teese and Langosch, 2015). The GxxxG and GxxxG-like motifs were identified in a variety of biological systems in which SPMPs dimerized or even in multi-pass MPs including receptors, integrins, ATP synthase, G-protein coupled receptors, and Notch signaling receptors (Senes et al., 2004). In the next section, I will discuss the caveats that negate the idea that the presence of a GxxxG sequence automatically implies association.

1.2.2 Misconceptions about the GxxxG motif

With the growing discovery of GxxxG motifs in associating TM helices (Fig 1.1), it became clear that the rules were much more complicated than the presence of a GxxxG sequence equates to homodimerization (Senes et al., 2004). As mentioned above, GxxxG is more prevalent than expected in TM helices, but furthermore, they are found in 12% of all TM helices and Sm-xxx-Sm motifs are found in 57% of them (Teese and Langosch, 2015). It is unlikely that all of these helices associate with themselves or other proteins because many function primarily as monomers (Bugge et al., 2016). A study of human TMDs found that the presence of a Sm-xxx-Sm motif did not necessarily mean the proteins would self-associate. Furthermore, even when the helices do self-associate, mutagenesis indicates that the Sm-xxx-Sm motif is not always important for the interaction (Kirrbach et al., 2013). To tease

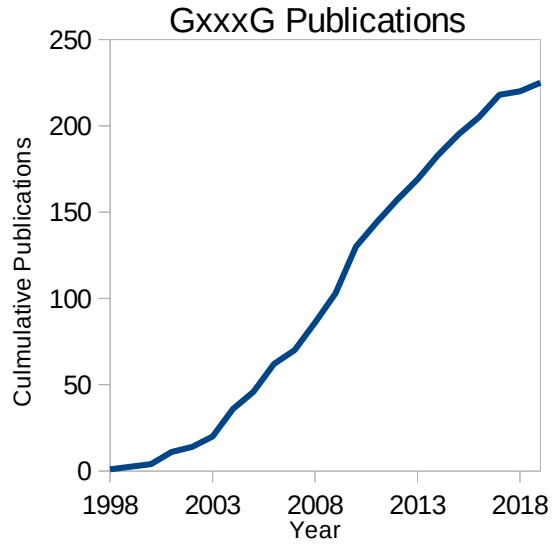


Figure 1.1 Publications containing GxxxG as a keyword. Cumulative GxxxG publications since the sequence motif was discovered until August 2019.

apart the rules governing the GxxxG motif, many studies have focused on the GxxxG associated dimer, GpA.

There are numerous studies on GpA, a model protein for studying helix-helix association. Many of these studies indicate that the sequence context that surrounds the GxxxG motif is critical for dimerization. Although mutagenesis of the GxxxG residues highly disrupt dimerization, those residues are not sufficient for wild-type level dimerization (Doura and Fleming, 2004; Doura et al., 2004; Lemmon et al., 1992; Melnyk et al., 2004; Schneider and Engelman, 2004). The sequence context surrounding the GxxxG motif is also important for other proteins, including the receptor tyrosine kinases (RTK) and integrins (Cymer et al., 2012; He et al., 2011).

Even if we understand that a GxxxG and the surrounding residues induce TM helix dimerization, that does not mean the residues are functionally significant. There are examples in which a full-length protein dimerizes and has a GxxxG motif, but point mutations in the motif do not disrupt protein function (Corver et al., 2007; Melnyk et al., 2004). Therefore, more studies are needed to understand the rules that surround the GxxxG motif and how the sequence, structure, and energetics modulate helix-helix dimerization.

1.3 GAS_{right}: A helix-helix association motif

1.3.1 GAS_{right} definition

GxxxG is the sequence motif associated with the GAS_{right} motif, one of the most frequent structural association motifs in TM proteins. The GAS_{right} motif is named for three structural features: 1) GAS for the small residues Gly, Ala, and Ser that define the sequence motif at the interface (G/A/S)xxx(G/A/S) and resulting in a short interhelical distance, and 2) the right-handed (-40º) crossing angle (Fig. 1.2a). Walters and DeGrado found that 28.8% of helix-helix interactions contain a GAS_{right} motif. These data showed that GAS_{right} is the most prevalent parallel association motif (Walters and DeGrado, 2006).

1.3.2 Prevalence in biological systems

Even before Walters and DeGrado coined the motif name, GpA (the human sialoglycoprotein, glycophorin A) was found to dimerize via a GAS_{right} motif (Lemmon et al., 1992; MacKenzie et al., 1997). As structure determination of MPs started to gain momentum, so did the prevalence of SPMPs mediated by GAS_{right} motifs. Of the 23 unique SPMP structures

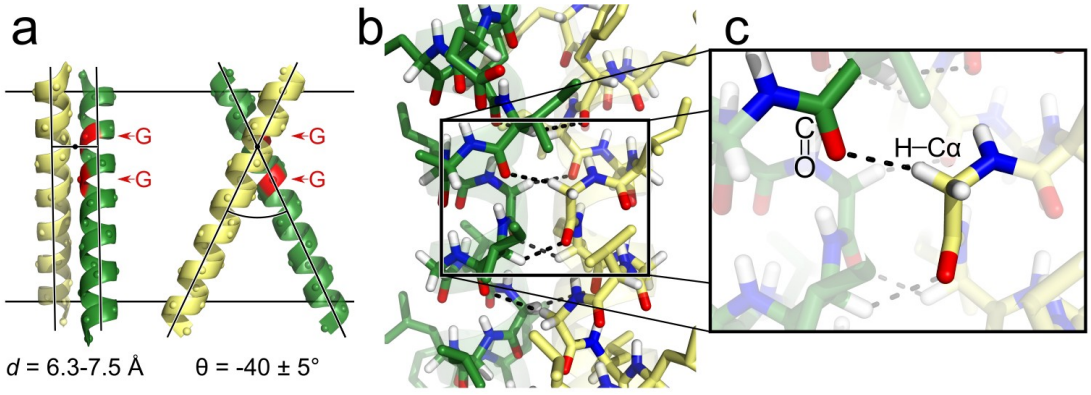


Figure 1.2 The GAS_{right} association motif. a) The GAS_{right} motif is a right-handed helical dimer with a short inter-helical distance *d* and a right-handed crossing angle θ of approximately −40°. The GxxxG sequence pattern at the crossing point (red) allows the backbones to come into contact. b) The contact enables the formation of networks of inter-helical H-bonds between Cα–H donors and carbonyl oxygen acceptors (shown in detail in c).

deposited in the PDB, ten of them are GAS_{right} structures (Lomize et al., 2006). The structures include Bcl-2 nineteen-kDa interacting protein 3 (Bocharov et al., 2007; Sulistijo and Mackenzie, 2009), Glycophorin A (MacKenzie et al., 1997; Trenker et al., 2015), integrin alpha-IIb-beta-3, a proapoptotic protein, cytochrome c nitrite reductase complex (Rodrigues et al., 2006), and several RTKs (Bocharov et al., 2008c, 2008b, 2012b; Bragin et al., 2016; Endres et al., 2013; Mineev et al., 2010b). A more complete list of GAS_{right} structures identified through computational modeling and mutagenesis data, which includes syndecans and major histocompatability complex proteins (Teese and Langosch, 2015). This list of proteins span the biological fields of immunology, metabolism, and cancer. It is clear that GAS_{right} mediated protein association is foundational to many biological systems, and thus, there must be a physical basis for the presence of the motif in various protein families.

1.3.3 Physical forces governing GAS_{right} transmembrane association

The GAS_{right} motif is characterized by a unique structural feature. An early study by Senes and Engelman showed that the close interhelical distance in GpA, a model system for GAS_{right} homodimerization, allowed the two backbones to come into contact and form characteristic networks of weak hydrogen bonds in which the donors are Cα carbons and the acceptors are carbonyl oxygens on the opposed helix (Cα–H···O=C, referred to as “Cα–H bonds” from this point forward; Fig. 1.2c). It was hypothesized that Cα–H bonds are a major driving force for GAS_{right} association (Senes et al., 2001).

Typically, carbon atoms are not thought of as hydrogen bond donors because carbon is less electronegative than nitrogen and oxygen. However, when a carbon is flanked by the electron-withdrawing groups of the peptide backbone, it’s relative electronegativity increases. Quantum mechanics calculations indicate that the energy of Cα–H bonds may be as much as one third to one half of that of N–H donors in vacuum (Scheiner et al., 2001; Vargas et al., 2000). Therefore, they are likely to be stabilizing factors in proteins embedded in the hydrophobic milieu of the membrane, particularly when they occur in multiple instances at the same interface, as in the

GAS_{right} motif (Fig. 1.2b; Senes et al., 2001). An IR-based investigation of the CD_2 stretching mode of a Cα–H donor in the TM domain of GpA produced an estimated interaction of −0.88 kcal/mol for that hydrogen bond (Arbely and Arkin, 2004). Conversely, a folding study of the multi-span MP bacteriorhodopsin, in which a side chain hydroxyl acceptor was mutated, concluded that a particular Cα–H bond did not appear to be stabilizing (Yohannan et al., 2004). A major hurdle to studying these Cα–H bonds is that a mutation strategy is not straightforward to implement when both donor and acceptor groups are part of the backbone (Fig. 1.2c). Subsequent theoretical work suggested that the orientation of the groups may determine whether Cα–H···O interactions may be strongly favorable or even unfavorable (Mottamal and Lazaridis, 2005; Park et al., 2008). In spite of these efforts, the exact contribution of backbone Cα–H bonds to TM helix association has not yet been directly addressed.

To address this contribution, Mueller *et al.* performed a detailed geometric analysis of Cα—H bonding (Fig. 1.3). They found that the GAS_{right} motif is optimized for the formation of networks

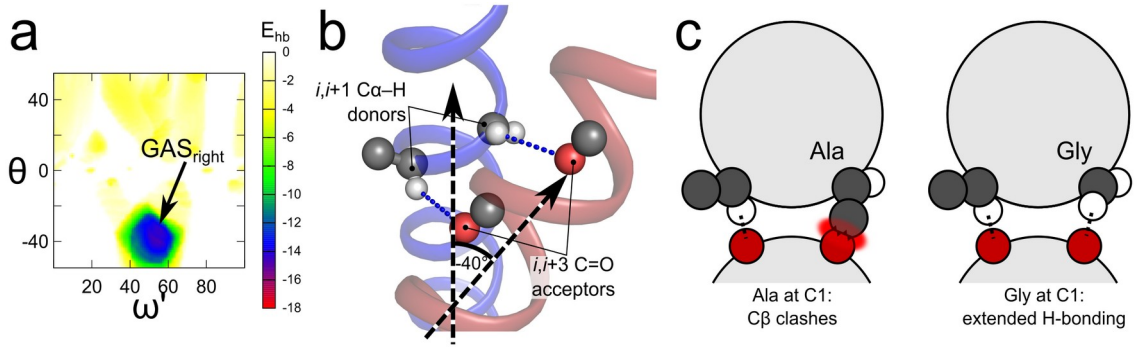


Figure 1.3 GAS_{right} is optimized for Cα–H H-bonding. a) The GAS_{right} motif corresponds to the only deep minimum in the Cα–H bond energy landscape in homodimers. The figure shows the total energy of Cα–H···O=C interactions as a function of dimer geometry (projected in 2 dimensions: crossing angle θ and axial rotation ω). b) The -40° crossing angle of GAS_{right} aligns Cα–H donors at *i*,*i*+1 and carbonyl acceptors at *i*,*i*+3 on opposed helices. c) Gly at the right side positions (N1, C1, C5) of the GAS_{right} interface solves two problems: it prevents clashing and it augments the H-bonding network with the second H group of Gly's Cα. Adapted from Mueller et al. *JACS,* 2017.

of interhelical Cα–H···O. That is, not only does GAS_{right} correspond to the global minimum in the geometric energy landscape of a parallel helical dimer, but in fact GAS_{right} corresponds to the only conformation that promotes the formation of such networks (Mueller et al., 2014). This discovery established a potential "causal link" between the Cα–H bonding and the frequency of the GAS_{right} motif: if Cα–H bonds are indeed particularly useful for mediating TM helix interaction, then it is not surprising that GAS_{right} is so frequently observed in nature. From that data, Mueller *et al.* developed an algorithm that predicts the structures of known parallel GAS_{right} homodimers to near-atomic precision using only Cα–H bonding and van der Waals interaction (Mueller et al., 2014; Fig 1.4). This algorithm, CATM, was later applied to predict the homodimeric structure of ADCK3, a previously unknown interaction (Khadria et al., 2014b). The notion that Cα–H bonds are a primary driving force for GAS_{right} association is logical and compelling, but it remains unproven.

In Chapter 2, I provide the first experimental evidence that supports the hypothesis that Cα–H bonds are important factors for GAS_{right} association. I use statistical analysis of twenty-six constructs to identify the sequence and structural patterns that lead to good Cα–H bonding and

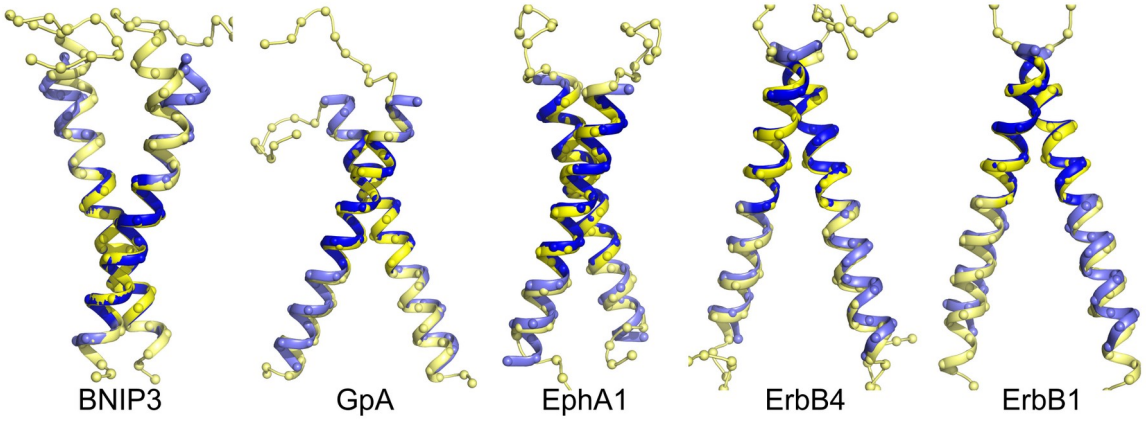


Figure 1.4 CATM predicts known GAS_{right} structures at near-atomic level. NMR structure of five homodimeric GAS_{right} structures (yellow) superimposed with their CATM prediction (blue). Backbone RMSDs: BNIP3 0.6 Å; GpA 1.1 Å; EphA1 1.3Å; ErbB4 0.8 Å; ErbB1 0.8 Å. Adapted from Mueller et al. *JACS,* 2017.

strong overall association. CATM produced energy scores that correlated with experimental dimerization, but with only twnety-six data points, it would be impossible to improve, or train, CATM without over-fitting. In the next section I discuss the alternative methods that have been used to understand helix-helix association

1.4 Methods used to study helix-helix association

In Section 1.1, I discussed that the physical study of MPs is quite challenging, especially in the non-native membrane environments required to isolate the MPs. Furthermore, full-length SPMPs tend to be complicated in that they require both hydrophobic and hydrophilic environments in the same system to account for the globular domains and the TM helix. To simplify this system and since TMDs are independently folded domains, they are often separated and studied in isolation. This section will describe the diverse methods that have been used to understand how TMDs interact.

1.4.1 Computational methods

A wide variety of valuable methods exists that can be used to assist experimental characterization of MPs. Molecular dynamics (MD) simulations can evaluate the structure and energetics of helix-helix association with precision. For example, a predicted structure can be evaluated by the increase of root means square deviation (RMSD) over time from the initial prediction. Association energies can be evaluated by calculating the free energy of association or the potential mean force through methods like window exchange umbrella sampling and free energy perturbation (Li et al., 2014; Park and Im, 2013; Park et al., 2012). The association energy of GpA has been tested in many studies that began with membrane mimetics (Hénin et al., 2005; Zhang and Lazaridis, 2006) and continue to be used to evaluate lipids and new MD techniques. Computational results tend to agree with the trends of experimental measurements, but it is important to remember that the environment strongly contributes to energetics (Janosi et al., 2010; Sengupta and Marrink, 2010). MD simulations are incredibly valuable tools to elucidate protein structure as well as energetic information. They evaluate interactions with the environment and the particular forces important for association, however they are incredible expensive techniques, meaning they require a lot of computational power. One way to utilize

computational techniques without this enormous expense is to use a limited number of functions to predict the structure, rather than the dynamics, of a protein.

Computational modeling of TM helices started in the early 1990s with the prediction of the structure of the GpA model system dimer (Adams et al., 1996; Treutlein et al., 1992) and of the pentameric phospholamban (Adams et al., 1995). Over the next decade, several groups improved on these algorithms based on the extremely limited number of available experimental structures, primarily on GpA (Fleishman and Ben-Tal, 2002; Kim et al., 2003; Park et al., 2004). As more experimental structures became available, the prediction algorithms continued to improve for these SPMP dimers. Using hydrophobicity plots, PREDDIMER identifies interfacial residues of a helix and uses it to predict the dimer structures (Polyansky et al., 2012, 2014). Subsequently, the Senes lab developed the CATM algorithm based on the use of Cα–H bonds and van der Waals parameterized by CHARMM. CATM outperformed PREDDIMER for GAS_{right} dimers, reducing average RMSD values to known structures from 3 Å to sometimes less than 1 Å (Mueller et al., 2014). The following year, EFDock-TM used co-evolutionary restraints to predict models, outperforming the previous methods for non-GAS_{right} dimers (Wang and Barth, 2015). EFDock-TM is based on homology modeling performed using RosettaMembrane (Barth et al., 2007, 2009). With growing interest in predicting membrane protein structures, Lomize and Pogozheva created a web-based server, TMDOCK, to predict TM dimers using several energetic scores (Lomize and Pogozheva, 2017). This algorithm performed better than PREDDIMER, but it was not compared to EFDock-TM and performed worse than CATM for GAS_{right} dimers. The most recent TM dimer structure prediction algorithm is TMDIM which uses cluster-based candidate selection and packing classification (Cao et al., 2017). Though the authors only compared their algorithm to PREDDIMER and PARK (Park et al., 2004), it performed quite well. As outlined here, there are a number of TM dimer prediction algorithms that have different foundational theories and as more structures become available, they will become more accurate and versatile in their prediction abilities.

In this thesis, I will focus solely on the CATM algorithm for both structure prediction and stability prediction for GAS_{right} dimers. I use CATM to predict the stability and analyze sequence, structure, and energetic trends of twenty-six TMDs in Chapter 2. Chapter 3 uses CATM as a basis for TMD selection for a high-throughput mutagenesis screen. In Chapter 4, I use CATM to predict the structure of an unknown dimer, a plant cytokinin oxidase/dehydrogenase.

1.4.2 Structural methods

The oldest method for elucidating protein structure is x-ray crystallography. Even though, SPMPs are notoriously difficult to crystallize due to their amphipathic nature GpA has been solved by lipid-cubic phase crystallography (Trenker et al., 2015). Nevertheless, the method of choice for protein structure determination is nuclear magnetic resonance (NMR). NMR has been used to solve several thousand MP structures, including the SPMP structures listed in Section 1.3.2. In 1997, the first SPMP structure was solved using solution-state NMR (MacKenzie et al., 1997). A handful of other SPMP dimer structures have been solved since then and have been extensively reviewed (Bugge et al., 2016). Solid-state NMR is an alternative that provides an opportunity to use more realistic membrane mimetics like liposomes and nanodisks. It has been used to solve the structure of numerous MPs that have been reviewed elsewhere, but none have been bitopic dimers (Ladizhansky, 2017; Mandala et al., 2018; McDermott, 2009; Patching, 2015).

The cryo-electron microscopy (cryo-EM) revolution has been particularly impactful for MP structures. Cryo-EM does not require absolute sample homogeneity or sample crystallization, the latter of which is a challenging bottleneck for MPs. Cryo-EM accounts for a larger percentage of MP structures deposited into the PDB every year (Cheng, 2018). Even so, small MPs, including SPMPs, remain challenging because it is difficult to obtain accurate image alignment. The smallest MPs solved as of 2018 are G-protein/G-protein coupled receptor complexes which have seven TMDs (Zhang et al., 2017). There may be potential for SPMP cryo-EM structures if the stable and folded soluble domains are included in sample preparation

rather than the traditional domain isolation method. Because of the lack of 3D structures that are available for bitopic proteins, researchers have turned to other methods to elucidate their oligomeric states.

1.4.3 Quantitative methods to measure association and stability

There are many methods that have been used to study MP stability *in vitro* and the largest classifier is the use of fluorescence. The simplest of these methods is co-localization. When two proteins are fused to non-overlapping fluorescent proteins (FP), if the FPs move to the same place, it is an indicator that the proteins of interest may be interacting. Quantitative imaging Förster resonance energy transfer (FRET), attaches a donor and acceptor FP to proteins of interest. If they co-localize and strongly associate in blebbed cells, FRET occurs between the two FPs. Because the FPs fluoresce regardless of association, monomer and dimer concentrations can be measured and dissociation constants calculated (Chen et al., 2010; Li et al., 2008). Another fluorescent method is bimolecular fluorescence complementation (BiFC) in blebbed cells. Instead of two different FPs attached to the proteins of interest, BiFC uses a split FP assay where the N-terminal domain of an FP is fused to one protein and the C-terminal domain is fused to another. When the two proteins associate, the FP refolds and fluoresces (Wang et al., 2017). The most quantiative fluorescent assay is FRET performed with synthesized peptides. In FRET, the peptides are labeled with fluorescent donor and acceptor pairs where the emission spectrum of the donor fluorophore overlaps with the excitation spectrum of the acceptor fluorophore. The dissociation constant is calculated from the donor/acceptor signal ratio throughout a lipid:protein serial dilution (Fisher and Ryan, 1999). Bulk FRET has been used to study GpA in numerous studies (Adair and Engelman, 1994; Fisher et al., 2003) as well as many other proteins (Khadria and Senes, 2015).

Sedimentation equilibrium analytical ultra-centrifugation (SE-AUC) is used to directly measure the mass of a protein complex allowing researchers to determine whether the protein is a monomer, dimer, or higher-order oligomer (Fleming et al., 1997). Like every other method

interrogating TM association, SE-AUC was tested on GpA and it's monomerizing mutants (Doura and Fleming, 2004; Doura et al., 2004; Fleming, 2002).

Steric trapping is a new method that calculates dissociation constants in lipid bilayers through a competitive biotin/streptavidin binding assay (Hong et al., 2010, 2013) derived from a protein unfolding assay (Blois et al., 2009). Another method currently being developed in the Senes lab is single-molecule photobleaching for SPMPs. The method has successfully been used for other MPs (Chadda et al., 2016).

The final method I will describe does not utilize FPs, but instead takes a proteome-wide approach. Tandem affinity chromatography mass spectrometry for MPs tags a bait protein of interest with an affinity tag and then pulls out every protein that associates with it. This huge set of proteins is then sent for mass spectrometry and evaluated for which proteins bound to the bait protein (Babu et al., 2012). Native mass spectrometry has also been used for intact MP complexes to elucidate their stoichiometry (Laganowsky et al., 2013).

Working in *in vitro* systems is best for studying a specific protein or interaction, but rarely captures the entire picture. In every *in vitro* system, the membrane mimetic choice dramatically impacts the results (Zhou and Cross, 2013). Proteins never exist in a vacuum and always interact with the rest of the cellular environment. Therefore, studying MPs *in vivo* better approximates the native environment of the protein.

1.4.4 ToxR derived genetic reporter assays

Genetic reporter assays have been used in a variety of systems to detect protein-protein interactions, including to evaluate helix-helix interactions in the membrane, mostly in *E. coli*. Table 1.1 is a comprehensive list of genetic reporter assays that were derived from or modeled off of the original ToxR assay (Langosch et al., 1996). Russ and Engelman swapped out the original β-galactosidase reporter gene with chloramphenicol acetyl-transferase (CAT) to make the widely used TOXCAT system (Russ and Engelman, 1999). In this assay and the majority of the derivatives, the dimerization of TM helices is measured by the expression of a reporter

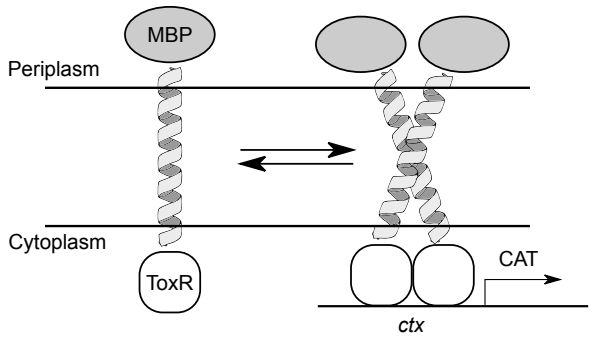


Figure 1.5 TOXCAT genetic reporter assay schematic. TM helices are connected to the transcription factor ToxR and MBP. When the helices dimerize, the ToxR moieties bind to the *ctx* promoter and turn on the expression of CAT.

gene. This reporter gene is turned on by ToxR, a dimerization-dependent transcription factor attached to the periplasmic ends of each helix. This assay correlates the amount of reporter gene expression to the amount of helical dimer present. Additionally, MP insertion is ensured through the use of maltose binding protein (MBP) through a subsequent complementation assay (Fig. 1.5).

Many groups have derived variants of the TOXCAT assay, implementing a variety of reporter genes including fluorescent proteins (Armstrong and Senes, A.; Berger et al., 2010) and luciferase (Bennasroune et al., 2005). Different association types can be measured including heterodimers (Berger et al., 2010; Julius et al., 2017; Ouellette et al., 2017; Schanzenbach et al., 2017; Schneider and Engelman, 2003; Su and Berger, 2012, 2013) and multi-pass proteins (Joce et al., 2011). Other versions of the assay cloned the reporter gene into the *E. coli* genome (Gurezka and Langosch, 2001), measured insertion through antibiotic resistance (Lis and Blumenthal, 2006), or used split reporter genes (Julius et al., 2017; Schanzenbach et al., 2017).

These genetic reporter assays are some of the simplest ways to begin probing a new system. They have two main advantages: the interactions are being measured in biological membranes, and mutagenesis screens can be performed relatively rapidly to identify the interface and critical residues for association. This latter step is only limited by the cloning capability. Furthermore, the genetic reporter assays can be utilized to perform both screening and selection on TM libraries, making them incredibly versatile tools.

1.4.5 Screening and selection for membrane protein structure

Screening is defined here as evaluating each variant in a library individually. Conversely, selection is used to isolate the variants that match a certain criterion, sorting the highest or lowest fluorescent variants or the ones that survive in a growth assay. Most of the genetic reporter assays described in Section 1.4.4 can be applied to either one or both of these applications. In screening, the researcher evaluates the dimerization of TM helices in a given library through the expression of a reporter gene, which can be determined via its fluorescence, enzymatic activity, antibiotic resistance, or another method specific to the reporter gene's nature. Screening has been performed by the Senes lab on the GAS_{right} motif (Anderson et al., 2017; Khadria et al, 2014), divisome proteins (Armstrong and Senes, A.; LaPointe et al., 2013), and other systems (Gromek et al., 2014; Hsin et al., 2011). Other protein systems that have been tested for dimerization using TOXCAT screening include BNIP3 (Lawrie et al., 2010), RTKs (Finger et al., 2009), and integrins (Li et al., 2005). Screening has also been used to elucidate the energetic properties of association and protein structure. For example, one group analyzed hydrophobic mismatch on TMD packing (Grau et al., 2017) and another evaluated the importance of electrostatic interactions for TMD association (Schanzenbach et al., 2017). Chapter 2 of this thesis uses TOXCAT-based screening to evaluate the importance of van der Waals and Cα—H bonding in the GAS_{right} motif. Altogether, screening has provided invaluable information. However, it is currently limited by the number of constructs that can practically be tested because traditional methods are not coupled to high-throughput systems to evaluate hundreds or thousands of variants.

Alternatively, selection methods can test a nearly unlimited amount of samples to find qualities of interest. Since the reporter gene of TOXCAT produces the antibiotic resistance protein CAT, TOXCAT can be applied to selection of large libraries. This kind of selection assay identified the GxxxG motif (Russ and Engelman, 2000) and membrane insertion propensity (Lis and Blumenthal, 2006). Selection assays isolate desired properties out of large libraries, but

they lack the desired quality of quantifying dimerization and insertion of individual variants directly.

1.4.6 The need for new methods

Using all the methods described above, the field has made considerable strides in understanding the structure and function of TM dimers. We have learned about interaction motifs, including GAS_{right}, that drive oligomerization. A number of TM binding partners for proteins critical for human and agricultural health have been identified. Structures have been elucidated and drug targets designed. The field is beginning to understand the physical forces that drive protein-protein association including van der Waals, hydrogen bonding, protein-lipid interactions, and how those forces contribute to association strength.

Nevertheless, MP biology lags far behind that of soluble proteins because they are significantly more challenging to work with. There have been new technological developments in recent years that can be applied to this problem by working at a high-throughput or “-omic” level. High-throughput methodologies provide an opportunity to evaluate proteins at the level of thousands, or even millions, of variants at a time. They can be used to study the energy landscape of a protein, allosteric networks, trends in entire classes of proteins, or explore alternative evolutionary pathways because these questions require measuring a multitude of variants. In the following section, I will describe these improvements and how they have been applied to MP challenges.

1.5 Emerging technology: FACS and NGS

1.5.1 Deep mutational scanning and sort-seq in the literature

Deep mutational scanning (DMS) is a catch-all term used to describe the mass mutagenesis of a protein or other DNA-encoded element that is assayed using next-generation sequencing (NGS) and some type of selection assay (Araya and Fowler, 2011). DMS ideally characterizes a library that covers the entire sequence space. A small TMD that averages 21 amino acids in length translates to a library of 2.1×10^{27}. Supposing that we limit TMDs to small or hydrophobic amino acids (A, C, F, G, I, L, S, T, V, W, Y), the library is still 7.4×10^{21} sequences. A mutational library that size is too large to screen meaningfully, and as a result, the majority of these DMS libraries can only be subjected to selection assays.

These selection assays work by calculating the enrichment of variants before and after selection to evaluate protein fitness. DMS has been performed in *E. coli*, yeast, and mammalian cells as well as *in vitro* systems through ribosome and bacteriophage display through ligand binding, growth rate, and enzymatic activity. Applications of DMS include protein model discrimination, epitope mapping, binding and stability measurements, and structure and phenotype prediction. This work has been extensively reviewed elsewhere (Araya and Fowler, 2011; Fowler and Fields, 2014; Gupta and Varadarajan, 2018).

Recently, a variant of DMS has combined NGS with fluorescence-activated cell sorting (FACS) to evaluate, or screen, a large number of sequences. Sort-seq reconstructs the original fluorescence value of each variant in a library after sorting subpopulations of cells and identifying the clones in their pools by NGS. The sort-seq method was developed in 2010 (Kinney et al., 2010) and the name was coined in 2014 (Peterman et al., 2014). There are less than 20 publications so far that have used this method and they primarily use genetic reporter assays to evaluate protein-protein interactions (Table 1.2). Performed in both yeast and *E. coli*, the sort-seq methods have evaluated promoter systems (Kinney et al., 2010; Rohlhill et al.,

2017; Sharon et al., 2012), RNA regulation (Holmqvist et al., 2013; Peterman et al., 2014), splicing variants (Cheung et al., 2019), and protein-ligand specificity (McLaughlin et al., 2012) among several others (Müller et al., 2014; Peterman and Levine, 2016; Starr et al., 2017). Sort-seq is also sometimes known as FACS-seq in medical articles for transcriptome analysis. Though these assays do not reach the level of saturation mutagenesis, they can however, evaluate tens of thousands of sequences in a single sort-seq experiment, where the main limitations that determine the library size are the cloning capacity, statistical analysis, and meaningful interpretation (Peterman and Levine, 2016).

1.5.2 High-throughput membrane protein applications

There are few published attempts to apply DMS to MP problems, and to the best of my knowledge, there are none that apply sort-seq (Table 1.3). The Sanders lab ventured into medium-throughput methods that measure the expression, insertion, and electrophysiology of the *KCNQ1* voltage-gated potassium channel in mammalian cells (Huang et al., 2018; Vanoye et al., 2018). The Plückthun lab used DMS assays to evaluate stability, codon preference, and production yield of G-protein coupled receptors (GPCR) (Sarkar et al., 2008; Schlinkmann et al., 2012; Schütz et al., 2016). An interesting variant of DMS for MPs used liposome display in a cell-free system to evaluate transporter functionality (Fujii et al., 2014). One of the more recent applications of DMS is to study helix-helix interactions using a variant of the TOXCAT assay to measure TM insertion and dimerization (Elazar et al., 2016). Many of the DMS methods listed above can be adapted to sort-seq to better quantify the effects of specific mutations.

1.5.3 Sort-Seq for understanding GAS_{right} structure

In Chapter 3, I describe a new technique that will expand the use of genetic reporter assays to test an unprecedented number of TM homodimers for self-association. The goal of the assay is to be able to test large libraries of TM sequences and evaluate not just the very strong or very weak dimers, but measure dimerization propensity across the sequence space of the GAS_{right}

motif. It is difficult to extrapolate findings obtained from individual protein systems, such as GpA, to understand the properties of an entire motif which is why I want to expand the analysis to a larger sequence space. Though outside the scope of this thesis, the sort-seq method I describe in Chapter 3 will be used to formulate a structure-based hypothesis of the determinant of stability of the GAS_{right} motif, addressing the question of the relative contribution of Cα–H bonds, packing, and other factors to the stability of the motif. Sort-seq can also be used for many TM systems including comprehensive mutagenesis of TMDs critical in human health and disease.

1.6 Overview of this thesis

My graduate work has focused on understanding the details of the GAS_{right} motif. I used computational prediction and experimental analysis to probe how the sequence context around GxxxG-like sequence motifs modulates the geometric and energetic properties of GAS_{right} dimers. I created a new experimental tool that, in combination with high-throughput computing, will enable the analysis of large TM libraries to probe their association, structure, and stability.

In Chapter 2, I discuss the experimental evaluation of a library of TMDs which revealed sequence, structure, and energetic trends within the GAS_{right} motif. I analyzed the entire human genome of SPMPs with the CATM algorithm, finding a large number had the potential to associate via the GAS_{right} motif. After testing 26 constructs, I found that there are certain positions where a glycine residue is more favorable for association and leads to closer and narrower helices. This work provides experimental evidence that Cα—H bonds are major contributors to the free energy of association and that van der Waals forces are optimized in stronger associating helices. I found that these trends were paralleled in the entire human genome, not just the experimentally tested structures. This work was published in the Journal of the American Chemical Society in 2017.

In Chapter 3, I will discuss my current work on developing a high-throughput method to understand helix-helix association in the membrane. Previous ToxR-based assays required long days to test a limited number of samples. I have increased capacity by combining FACS with NGS on libraries cloned using oligo pool technology. This has increased our testing capacity to tens of thousands of TMDs. I tested human wild-type sequences with extensive mutagenesis to identify new dimerizing proteins including new GAS_{right} dimers. This method was able to identify insertion capability, dimerization propensity, and interfacial residues. I will be publishing these results in combination with an evaluation of the CATM algorithm within the next year.

In Chapter 4, I include the collaboration with Michael Niemann and Tomáš Werner of the University of Graz, who performed *in vitro* and *in vivo* assays that showed the *Arabidopsis thaliana* protein cytokinin dehydrogenase/oxidase 1 TMD forms homodimers. I created a model of this GAS_{right} dimer using the CATM algorithm combined with independent proline modeling. I was able to identify mutations in the GxxxG-like sequence that disrupt the helix-helix interface. They confirmed that these mutations disrupt ER localization and its activity *in vivo*. This work was published in Plant Physiology in 2018.

In Chapter 5, I describe continuing work and future directions of the GAS_{right} projects in the lab. One of these projects evaluates GAS_{right} association across model organism proteomes which will provide structural insight for biologists. High-throughput experimental evaluation and computational learning of the CATM algorithm will lend insight into the energetic forces that drive GAS_{right} association. CATM currently only evaluates GAS_{right} homodimers, but a continuing side project is to expand it to include heterodimers. I also discuss possible improvements for the TOXGREEN assay that will increase the capability to quantify insertion rates that are critical to association constants of TM dimers.

Table 1.1 Genetic assays that measure TM helix association. Ordered by publication year

Assay Name	Dimerization Reporter Gene	Transcription Factor	Insertion Gene	Homo/ Hetero	Single/ Multi	Reference
ToxR	lacZ	ToxR	MBP	Homo	Single	(Langosch et al., 1996)
TOXCAT	CAT	ToxR	MBP	Homo	Single	(Russ and Engelman, 1999)
POSSYCAT	CAT[a]	ToxR	MBP	Homo	Single	(Gurezka and Langosch, 2001)
GALLEX	lacZ	LexA	MBP	Hetero	Single	(Schneider and Engelman, 2003)
ToxLux	Luciferase	ToxR	MBP	Homo	Single	(Bennasroune et al., 2005)
βLac	CAT	ToxR	βLac	Homo	Single	(Lis and Blumenthal, 2006)
DN-ToxRed	RFP	ToxR	MBP	Hetero	Single	(Berger et al., 2010)
Multi-Tox	lacZ	ToxR	MBP	Homo	Multi	(Joce et al., 2011)
AraTM	GFP	AraC	MBP	Homo/ Hetero	Single	(Su and Berger, 2012, 2013)
TOXGREEN	sfGFP	ToxR	MBP	Homo	Single	(Armstrong and Senes, A.)
BlaTM[b]	Split βLac	N/A	Split βLac	Hetero	Single	(Julius et al., 2017; Schanzenbach et al., 2017)
BACTH	Multiple	Catabolite activator protein	None	Hetero	Single	(Ouellette et al., 2017)

a: CAT is on the *E. coli* genome.

b: Expression is measured through the intra-cellular globular domain sfGFP

Table 1.2 Publications that include the sort-seq method

Reference	Macromolecule evaluated	Assay	System	Finding
(Kinney et al., 2010)	DNA promoters	lac genetic reporter	DNA footprinting	Method development, binding energy evaluated for each mutant
(Sharon et al., 2012)	DNA promoters	Dual genetic reporter (yeast)	Transcription factor combined with promoter effects	Changing a TF binding site by a few bps has a large effect on gene expression.
(McLaughlin et al., 2012)	Protein	Genetic reporter	Coevolution	Identified sectors of a domain tolerant and resistant to mutation; ligand specificity switching
(Holmqvist et al., 2013)	mRNA	Genetic reporter	csgD mRNA is regulated by a sRNA	Differentiated the positions in the mRNA that are important for translation
(Kosuri et al., 2013)	DNA/RNA/ protein	Dual genetic reporter	Promoters and ribosome binding sites	Method development to quantify transcription and translation rates simultaneously
(Müller et al., 2014)	DNA	Genetic reporter (yeast)	Genome replication	Method development, haploid and diploid cells have identical replication profiles
(Peterman et al., 2014)	sRNA	Genetic reporter	DsrA and RyhB sRNAs	Binding specificity can be enhanced by rigid molecular structure
(Noderer et al., 2014)	mRNA*	Genetic reporter	Translation Initiation	Optimized an enhanced start codon recognition motif
(Sharon et al., 2014)	DNA promoters	Dual genetic reporter (yeast)	Noise vs expression level for designed promoters	Nucleosome-disfavoring promoters result in less noise and more transcription factor binding sites results in more noise
(Mirzadeh et al., 2015)	mRNA*	Genetic reporter	Vector-inset cloning scars	Low GC content and relaxed mRNA stability help for high protein expression
(Peterman and Levine, 2016)	Protein	Simulated data	Evaluating sort-seq design choices	Limit the number of sort gates, include a reference reporter, variability and statistics must be applied with care.
(Rohlhill et al., 2017)	DNA promoter	Dual genetic reporter	Formaldehyde inducible promoter	Designed a promoter with predictable expression based on formaldehyde concentration
(Ahn et al., 2017)	RNA*	Sort human cells	Transcriptome analysis	Different skin cell types have different gene expression
(Jobe et al., 2017)	RNA*	GFP tagged TF (mouse)	Transcriptome analysis	Methyl-CpG-binding domain 1 is important for neural stem cell integrity
(Starr et al., 2017)	Protein	Dual genetic reporter (yeast)	Alternate evolution pathways	There are many evolutionary paths that could lead to the same protein function, even with many different sequences
(Singh et al., 2018)	RNA*	GFP tagged protein from forelimbs (mouse)	Transcriptome analysis	Pax3 has immunological derivatives, not just skeletal
(Džunková et al., 2018)	DNA*	Sort bacterial cells	Pathogen identifcation	Medical test development
(Cheung et al., 2019)	DNA	Dual genetic reporter (yeast)	Splicing	Large-effect splice disruption variants are not in canonical splice sites

*: Uses FACS-seq instead of sort-seq

Table 1.3 High-throughput assay publications that focus on membrane proteins

Reference	Protein evaluated	Mutants Evaluated	Assay	Host Type	Selection	Finding
(Sarkar et al., 2008)	GPCRs	Estimated 10^7	Bacterial display/ ligand binding	*E. coli*	MP stability/ fluorescence	Method development, greater GPCR stability
(Schlinkmann et al., 2012)	GPCRs	~24,000	(Sarkar et al., 2008)	*E. coli*	MP stability/ fluorescence	Evaluate codon preference at each position
(Fujii et al., 2014)	α-hemolysin	Estimated 10^7	Liposome display/ ligand binding	Cell-free translation	MP activity/ fluorescence	Method development, if a transporter works, the fluorescent ligand is moved into the liposome, fluorescence indicates protein activity
(Elazar et al., 2016)	GpA, ErbB, L-selectin	1,320	Genetic reporter assay (dsTbL)	*E. coli*	MP insertion and dimerization/ Growth	Method development, quantify ddGs of association for mutants
(Schütz et al., 2016)	GPCRs	Estimated 10^7	Yeast display/ ligand binding	Yeast	MP expression/ fluorescence	Method development, if a GPCR is correctly inserted and folded, it will bind the fluorescent ligand, used for directed evolution
(Huang et al., 2018)	KCNQ1	51	Bacterial display/ ligand binding	HEK293	MP expression and insertion/ fluorescence	Method development, >50% mutations examined were seen to destabilize KCNQ1, accompanied by mistrafficking and degradation by the proteasome
(Vanoye et al., 2018)	KCNQ1	78	Patch-clamp	CHO	N/A	Method development, reclassifying loss of function mutations

1.7 References

Adair, B.D., and Engelman, D.M. (1994). Glycophorin A helical transmembrane domains dimerize in phospholipid bilayers: a resonance energy transfer study. Biochemistry *33*, 5539–5544.

Adams, P.D., Arkin, I.T., Engelman, D.M., and Brünger, A.T. (1995). Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. Nat. Struct. Biol. *2*, 154–162.

Adams, P.D., Engelman, D.M., and Brünger, A.T. (1996). Improved prediction for the structure of the dimeric transmembrane domain of glycophorin A obtained through global searching. Proteins *26*, 257–261.

Ahn, R.S., Taravati, K., Lai, K., Lee, K.M., Nititham, J., Gupta, R., Chang, D.S., Arron, S.T., Rosenblum, M., and Liao, W. (2017). Transcriptional landscape of epithelial and immune cell populations revealed through FACS-seq of healthy human skin. Sci. Rep. *7*, 1343.

Anbazhagan, V., Munz, C., Tome, L., and Schneider, D. (2010). Fluidizing the membrane by a local anesthetic: phenylethanol affects membrane protein oligomerization. J. Mol. Biol. *404*, 773–777.

Anderson, S.M., Mueller, B.K., Lange, E.J., and Senes, A. (2017). Combination of Cα-H Hydrogen Bonds and van der Waals Packing Modulates the Stability of GxxxG-Mediated Dimers in Membranes. J. Am. Chem. Soc. *139*, 15774–15783.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., and Iyer, L.M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol. Rev. *29*, 231–262.

Araya, C.L., and Fowler, D.M. (2011). Deep mutational scanning: assessing protein function on a massive scale. Trends Biotechnol. *29*, 435–442.

Arbely, E., and Arkin, I.T. (2004). Experimental measurement of the strength of a C alpha-H...O bond in a lipid bilayer. J. Am. Chem. Soc. *126*, 5362–5363.

Armstrong, C.R., and Senes, A. TOXGREEN, a high-throughput variant of the TOXCAT assay for helix self-association in biological membranes. Submitt. Publ.

Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B.D.M., Burston, H.E., Vizeacoumar, F.J., Snider, J., Phanse, S., et al. (2012). Interaction landscape of membrane-protein complexes in Saccharomyces cerevisiae. Nature *489*, 585–589.

Barth, P., Schonbrun, J., and Baker, D. (2007). Toward high-resolution prediction and design of transmembrane helical protein structures. Proc. Natl. Acad. Sci. U. S. A. *104*, 15682–15687.

Barth, P., Wallner, B., and Baker, D. (2009). Prediction of membrane protein structures with complex topologies using limited constraints. Proc. Natl. Acad. Sci. U. S. A. *106*, 1409–1414.

Bennasroune, A., Gardin, A., Auzan, C., Clauser, E., Dirrig-Grosch, S., Meira, M., Appert-Collin, A., Aunis, D., Crémel, G., and Hubert, P. (2005). Inhibition by transmembrane peptides of chimeric insulin receptors. Cell. Mol. Life Sci. CMLS *62*, 2124–2131.

Berger, B.W., Kulp, D.W., Span, L.M., DeGrado, J.L., Billings, P.C., Senes, A., Bennett, J.S., and DeGrado, W.F. (2010). Consensus motif for integrin transmembrane helix association. Proc. Natl. Acad. Sci. U. S. A. *107*, 703–708.

Blois, T.M., Hong, H., Kim, T.H., and Bowie, J.U. (2009). Protein unfolding with a steric trap. J. Am. Chem. Soc. *131*, 13914–13915.

Bocharov, E.V., Pustovalova, Y.E., Pavlov, K.V., Volynsky, P.E., Goncharuk, M.V., Ermolyuk, Y.S., Karpunin, D.V., Schulga, A.A., Kirpichnikov, M.P., Efremov, R.G., et al. (2007). Unique dimeric structure of BNip3 transmembrane domain suggests membrane permeabilization as a cell death trigger. J. Biol. Chem. *282*, 16256–16266.

Bocharov, E.V., Mineev, K.S., Volynsky, P.E., Ermolyuk, Y.S., Tkach, E.N., Sobol, A.G., Chupin, V.V., Kirpichnikov, M.P., Efremov, R.G., and Arseniev, A.S. (2008a). Spatial structure of the dimeric transmembrane domain of the growth factor receptor ErbB2 presumably corresponding to the receptor active state. J. Biol. Chem. *283*, 6950–6956.

Bocharov, E.V., Mineev, K.S., Volynsky, P.E., Ermolyuk, Y.S., Tkach, E.N., Sobol, A.G., Chupin, V.V., Kirpichnikov, M.P., Efremov, R.G., and Arseniev, A.S. (2008b). Spatial structure of the dimeric transmembrane domain of the growth factor receptor ErbB2 presumably corresponding to the receptor active state. J. Biol. Chem. *283*, 6950–6956.

Bocharov, E.V., Mayzel, M.L., Volynsky, P.E., Goncharuk, M.V., Ermolyuk, Y.S., Schulga, A.A., Artemenko, E.O., Efremov, R.G., and Arseniev, A.S. (2008c). Spatial structure and pH-dependent conformational diversity of dimeric transmembrane domain of the receptor tyrosine kinase EphA1. J. Biol. Chem. *283*, 29385–29395.

Bocharov, E.V., Mineev, K.S., Goncharuk, M.V., and Arseniev, A.S. (2012a). Structural and thermodynamic insight into the process of “weak” dimerization of the ErbB4 transmembrane domain by solution NMR. Biochim. Biophys. Acta *1818*, 2158–2170.

Bocharov, E.V., Mineev, K.S., Goncharuk, M.V., and Arseniev, A.S. (2012b). Structural and thermodynamic insight into the process of “weak” dimerization of the ErbB4 transmembrane domain by solution NMR. Biochim. Biophys. Acta *1818*, 2158–2170.

Bragin, P.E., Mineev, K.S., Bocharova, O.V., Volynsky, P.E., Bocharov, E.V., and Arseniev, A.S. (2016). HER2 Transmembrane Domain Dimerization Coupled with Self-Association of Membrane-Embedded Cytoplasmic Juxtamembrane Regions. J. Mol. Biol. *428*, 52–61.

Bugge, K., Lindorff-Larsen, K., and Kragelund, B.B. (2016). Understanding single-pass transmembrane receptor signaling from a structural viewpoint-what are we missing? FEBS J. *283*, 4424–4451.

Cao, H., Ng, M.C.K., Jusoh, S.A., Tai, H.K., and Siu, S.W.I. (2017). TMDIM: an improved algorithm for the structure prediction of transmembrane domains of bitopic dimers. J. Comput. Aided Mol. Des.

Chadda, R., Krishnamani, V., Mersch, K., Wong, J., Brimberry, M., Chadda, A., Kolmakova-Partensky, L., Friedman, L.J., Gelles, J., and Robertson, J.L. (2016). The dimerization equilibrium of a ClC Cl(-)/H(+) antiporter in lipid bilayers. ELife *5*.

Chen, L., Novicky, L., Merzlyakov, M., Hristov, T., and Hristova, K. (2010). Measuring the energetics of membrane protein dimerization in mammalian membranes. J. Am. Chem. Soc. *132*, 3628–3635.

Cheng, Y. (2018). Membrane protein structural biology in the era of single particle cryo-EM. Curr. Opin. Struct. Biol. *52*, 58–63.

Cheung, R., Insigne, K.D., Yao, D., Burghard, C.P., Wang, J., Hsiao, Y.-H.E., Jones, E.M., Goodman, D.B., Xiao, X., and Kosuri, S. (2019). A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. Mol. Cell *73*, 183-194.e8.

Chung, I., Akita, R., Vandlen, R., Toomre, D., Schlessinger, J., and Mellman, I. (2010). Spatial control of EGF receptor activation by reversible dimerization on living cells. Nature *464*, 783–787.

Corver, J., Broer, R., van Kasteren, P., and Spaan, W. (2007). GxxxG motif of severe acute respiratory syndrome coronavirus spike glycoprotein transmembrane domain is not involved in trimerization and is not important for entry. J. Virol. *81*, 8352–8355.

Cymer, F., Veerappan, A., and Schneider, D. (2012). Transmembrane helix-helix interactions are modulated by the sequence context and by lipid bilayer properties. Biochim. Biophys. Acta *1818*, 963–973.

Dixon, A.M., Stanley, B.J., Matthews, E.E., Dawson, J.P., and Engelman, D.M. (2006). Invariant chain transmembrane domain trimerization: a step in MHC class II assembly. Biochemistry *45*, 5228–5234.

Doura, A.K., and Fleming, K.G. (2004). Complex interactions at the helix-helix interface stabilize the glycophorin A transmembrane dimer. J. Mol. Biol. *343*, 1487–1497.

Doura, A.K., Kobus, F.J., Dubrovsky, L., Hibbard, E., and Fleming, K.G. (2004). Sequence context modulates the stability of a GxxxG-mediated transmembrane helix-helix dimer. J. Mol. Biol. *341*, 991–998.

Džunková, M., Moya, A., Chen, X., Kelly, C., and D'Auria, G. (2018). Detection of mixed-strain infections by FACS and ultra-low input genome sequencing. Gut Microbes 1–5.

Elazar, A., Weinstein, J., Biran, I., Fridman, Y., Bibi, E., and Fleishman, S.J. (2016). Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. ELife *5*, e12125.

Endres, N.F., Das, R., Smith, A.W., Arkhipov, A., Kovacs, E., Huang, Y., Pelton, J.G., Shan, Y., Shaw, D.E., Wemmer, D.E., et al. (2013). Conformational coupling across the plasma membrane in activation of the EGF receptor. Cell *152*, 543–556.

Fagerberg, L., Jonasson, K., von Heijne, G., Uhlén, M., and Berglund, L. (2010). Prediction of the human membrane proteome. Proteomics *10*, 1141–1149.

Finger, C., Escher, C., and Schneider, D. (2009). The single transmembrane domains of human receptor tyrosine kinases encode self-interactions. Sci. Signal. *2*, ra56.

Fisher, C.A., and Ryan, R.O. (1999). Lipid binding-induced conformational changes in the N-terminal domain of human apolipoprotein E. J. Lipid Res. *40*, 93–99.

Fisher, L.E., Engelman, D.M., and Sturgis, J.N. (2003). Effect of detergents on the association of the glycophorin a transmembrane helix. Biophys. J. *85*, 3097–3105.

Fleishman, S.J., and Ben-Tal, N. (2002). A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. J. Mol. Biol. *321*, 363–378.

Fleming, K.G. (2002). Standardizing the free energy change of transmembrane helix-helix interactions. J. Mol. Biol. *323*, 563–571.

Fleming, K.G., Ackerman, A.L., and Engelman, D.M. (1997). The effect of point mutations on the free energy of transmembrane alpha-helix dimerization. J. Mol. Biol. *272*, 266–275.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat. Methods *11*, 801–807.

Fujii, S., Matsuura, T., Sunami, T., Nishikawa, T., Kazuta, Y., and Yomo, T. (2014). Liposome display for in vitro selection and evolution of membrane proteins. Nat. Protoc. *9*, 1578–1591.

Grau, B., Javanainen, M., García-Murria, M.J., Kulig, W., Vattulainen, I., Mingarro, I., and Martínez-Gil, L. (2017). The role of hydrophobic matching on transmembrane helix packing in cells. Cell Stress *1*, 90–106.

Gromek, K.A., Suchy, F.P., Meddaugh, H.R., Wrobel, R.L., LaPointe, L.M., Chu, U.B., Primm, J.G., Ruoho, A.E., Senes, A., and Fox, B.G. (2014). The oligomeric states of the purified sigma-1 receptor are stabilized by ligands. J. Biol. Chem. *289*, 20333–20344.

Gupta, K., and Varadarajan, R. (2018). Insights into protein structure, stability and function from saturation mutagenesis. Curr. Opin. Struct. Biol. *50*, 117–125.

Gurezka, R., and Langosch, D. (2001). In Vitro Selection of Membrane-spanning Leucine Zipper Protein-Protein Interaction Motifs Using POSSYCCAT. J. Biol. Chem. *276*, 45580–45587.

He, L., Hoffmann, A.R., Serrano, C., Hristova, K., and Wimley, W.C. (2011). High-throughput selection of transmembrane sequences that enhance receptor tyrosine kinase activation. J. Mol. Biol. *412*, 43–54.

Hénin, J., Pohorille, A., and Chipot, C. (2005). Insights into the recognition and association of transmembrane alpha-helices. The free energy of alpha-helix dimerization in glycophorin A. J. Am. Chem. Soc. *127*, 8478–8484.

Holmqvist, E., Reimegård, J., and Wagner, E.G.H. (2013). Massive functional mapping of a 5'-UTR by saturation mutagenesis, phenotypic sorting and deep sequencing. Nucleic Acids Res. *41*, e122.

Hong, H., Blois, T.M., Cao, Z., and Bowie, J.U. (2010). Method to measure strong protein-protein interactions in lipid bilayers using a steric trap. Proc. Natl. Acad. Sci. U. S. A. *107*, 19802–19807.

Hong, H., Chang, Y.-C., and Bowie, J.U. (2013). Measuring transmembrane helix interaction strengths in lipid bilayers using steric trapping. Methods Mol. Biol. Clifton NJ *1063*, 37–56.

Hsin, J., LaPointe, L.M., Kazy, A., Chipot, C., Senes, A., and Schulten, K. (2011). Oligomerization state of photosynthetic core complexes is correlated with the dimerization affinity of a transmembrane helix. J. Am. Chem. Soc. *133*, 14071–14081.

Huang, H., Kuenze, G., Smith, J.A., Taylor, K.C., Duran, A.M., Hadziselimovic, A., Meiler, J., Vanoye, C.G., George, A.L., and Sanders, C.R. (2018). Mechanisms of KCNQ1 channel dysfunction in long QT syndrome involving voltage sensor domain mutations. Sci. Adv. *4*, eaar2631.

Hubert, P., Sawma, P., Duneau, J.-P., Khao, J., Hénin, J., Bagnard, D., and Sturgis, J. (2010). Single-spanning transmembrane domains in cell growth and cell-cell interactions: More than meets the eye? Cell Adhes. Migr. *4*, 313–324.

Janosi, L., Prakash, A., and Doxastakis, M. (2010). Lipid-modulated sequence-specific association of glycophorin A in membranes. Biophys. J. *99*, 284–292.

Javadpour, M.M., Eilers, M., Groesbeek, M., and Smith, S.O. (1999). Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. Biophys. J. *77*, 1609–1618.

Jobe, E.M., Gao, Y., Eisinger, B.E., Mladucky, J.K., Giuliani, C.C., Kelnhofer, L.E., and Zhao, X. (2017). Methyl-CpG-Binding Protein MBD1 Regulates Neuronal Lineage Commitment through Maintaining Adult Neural Stem Cell Identity. J. Neurosci. Off. J. Soc. Neurosci. *37*, 523–536.

Joce, C., Wiener, A.A., and Yin, H. (2011). Multi-Tox: application of the ToxR-transcriptional reporter assay to the study of multi-pass protein transmembrane domain oligomerization. Biochim. Biophys. Acta *1808*, 2948–2953.

Julius, A., Laur, L., Schanzenbach, C., and Langosch, D. (2017). BLaTM 2.0, a Genetic Tool Revealing Preferred Antiparallel Interaction of Transmembrane Helix 4 of the Dual-Topology Protein EmrE. J. Mol. Biol. *429*, 1630–1637.

Khadria, A.S., and Senes, A. (2015). Fluorophores, environments, and quantification techniques in the analysis of transmembrane helix interaction using FRET. Biopolymers *104*, 247–264.

Khadria, A.S., Mueller, B.K., Stefely, J.A., Tan, C.H., Pagliarini, D.J., and Senes, A. (2014a). A Gly-zipper motif mediates homodimerization of the transmembrane domain of the mitochondrial kinase ADCK3. J. Am. Chem. Soc. *136*, 14068–14077.

Khadria, A.S., Mueller, B.K., Stefely, J.A., Tan, C.H., Pagliarini, D.J., and Senes, A. (2014b). A Gly-zipper motif mediates homodimerization of the transmembrane domain of the mitochondrial kinase ADCK3. J. Am. Chem. Soc. *136*, 14068–14077.

Kim, S., Chamberlain, A.K., and Bowie, J.U. (2003). A simple method for modeling transmembrane helix oligomers. J. Mol. Biol. *329*, 831–840.

Kinney, J.B., Murugan, A., Callan, C.G., and Cox, E.C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc. Natl. Acad. Sci. U. S. A. *107*, 9158–9163.

Kirrbach, J., Krugliak, M., Ried, C.L., Pagel, P., Arkin, I.T., and Langosch, D. (2013). Self-interaction of transmembrane helices representing pre-clusters from the human single-span membrane proteins. Bioinforma. Oxf. Engl. *29*, 1623–1630.

Klabunde, T., and Hessler, G. (2002). Drug design strategies for targeting G-protein-coupled receptors. Chembiochem Eur. J. Chem. Biol. *3*, 928–944.

Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Endy, D., and Church, G.M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. *110*, 14024–14029.

Ladizhansky, V. (2017). Applications of solid-state NMR to membrane proteins. Biochim. Biophys. Acta Proteins Proteomics *1865*, 1577–1586.

Laganowsky, A., Reading, E., Hopper, J.T.S., and Robinson, C.V. (2013). Mass spectrometry of intact membrane protein complexes. Nat. Protoc. *8*, 639–651.

Lai, M.-D., and Xu, J. (2007). Ribosomal Proteins and Colorectal Cancer. Curr. Genomics *8*, 43–49.

Langosch, D., Brosig, B., Kolmar, H., and Fritz, H.-J. (1996). Dimerisation of the Glycophorin A Transmembrane Segment in Membranes Probed with the ToxR Transcription Activator. J. Mol. Biol. *263*, 525–530.

LaPointe, L.M., Taylor, K.C., Subramaniam, S., Khadria, A., Rayment, I., and Senes, A. (2013). Structural organization of FtsB, a transmembrane protein of the bacterial divisome. Biochemistry *52*, 2574–2585.

Lawrie, C.M., Sulistijo, E.S., and MacKenzie, K.R. (2010). Intermonomer hydrogen bonds enhance GxxxG-driven dimerization of the BNIP3 transmembrane domain: roles for sequence context in helix-helix association in membranes. J. Mol. Biol. *396*, 924–936.

Lemmon, M.A., Flanagan, J.M., Hunt, J.F., Adair, B.D., Bormann, B.J., Dempsey, C.E., and Engelman, D.M. (1992). Glycophorin A dimerization is driven by specific interactions between transmembrane alpha-helices. J. Biol. Chem. *267*, 7683–7689.

Li, E., Placone, J., Merzlyakov, M., and Hristova, K. (2008). Quantitative measurements of protein interactions in a crowded cellular environment. Anal. Chem. *80*, 5976–5985.

Li, P.-C., Miyashita, N., Im, W., Ishido, S., and Sugita, Y. (2014). Multidimensional umbrella sampling and replica-exchange molecular dynamics simulations for structure prediction of transmembrane helix dimers. J. Comput. Chem. *35*, 300–308.

Li, W., Metcalf, D.G., Gorelik, R., Li, R., Mitra, N., Nanda, V., Law, P.B., Lear, J.D., Degrado, W.F., and Bennett, J.S. (2005). A push-pull mechanism for regulating integrin function. Proc. Natl. Acad. Sci. U. S. A. *102*, 1424–1429.

Lis, M., and Blumenthal, K. (2006). A modified, dual reporter TOXCAT system for monitoring homodimerization of transmembrane segments of proteins. Biochem. Biophys. Res. Commun. *339*, 321–324.

Liu, Y., Engelman, D.M., and Gerstein, M. (2002). Genomic analysis of membrane protein families: abundance and conserved motifs. Genome Biol. *3*, research0054.

Liu, Y., Ma, D., and Ji, C. (2015). Zinc fingers and homeoboxes family in human diseases. Cancer Gene Ther. *22*, 223–226.

Lomize, A.L., and Pogozheva, I.D. (2017). TMDOCK: An Energy-Based Method for Modeling α-Helical Dimers in Membranes. J. Mol. Biol. *429*, 390–398.

Lomize, M.A., Lomize, A.L., Pogozheva, I.D., and Mosberg, H.I. (2006). OPM: orientations of proteins in membranes database. Bioinforma. Oxf. Engl. *22*, 623–625.

Lupas, A.N., and Bassler, J. (2017). Coiled Coils - A Model System for the 21st Century. Trends Biochem. Sci. *42*, 130–140.

MacKenzie, K.R., Prestegard, J.H., and Engelman, D.M. (1997). A transmembrane helix dimer: structure and implications. Science *276*, 131–133.

Mandala, V.S., Williams, J.K., and Hong, M. (2018). Structure and Dynamics of Membrane Proteins from Solid-State NMR. Annu. Rev. Biophys. *47*, 201–222.

Matthews, E.E., Thévenin, D., Rogers, J.M., Gotow, L., Lira, P.D., Reiter, L.A., Brissette, W.H., and Engelman, D.M. (2011). Thrombopoietin receptor activation: transmembrane helix dimerization, rotation, and allosteric modulation. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *25*, 2234–2244.

McDermott, A. (2009). Structure and dynamics of membrane proteins by magic angle spinning solid-state NMR. Annu. Rev. Biophys. *38*, 385–403.

McLaughlin, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138–142.

Melnyk, R.A., Kim, S., Curran, A.R., Engelman, D.M., Bowie, J.U., and Deber, C.M. (2004). The affinity of GXXXG motifs in transmembrane helix-helix interactions is modulated by long-range communication. J. Biol. Chem. *279*, 16591–16597.

Mineev, K.S., Bocharov, E.V., Pustovalova, Y.E., Bocharova, O.V., Chupin, V.V., and Arseniev, A.S. (2010a). Spatial structure of the transmembrane domain heterodimer of ErbB1 and ErbB2 receptor tyrosine kinases. J. Mol. Biol. *400*, 231–243.

Mineev, K.S., Bocharov, E.V., Pustovalova, Y.E., Bocharova, O.V., Chupin, V.V., and Arseniev, A.S. (2010b). Spatial Structure of the Transmembrane Domain Heterodimer of ErbB1 and ErbB2 Receptor Tyrosine Kinases. J. Mol. Biol. *400*, 231–243.

Mirzadeh, K., Martínez, V., Toddo, S., Guntur, S., Herrgård, M.J., Elofsson, A., Nørholm, M.H.H., and Daley, D.O. (2015). Enhanced Protein Production in Escherichia coli by Optimization of Cloning Scars at the Vector-Coding Sequence Junction. ACS Synth. Biol. *4*, 959–965.

Mottamal, M., and Lazaridis, T. (2005). The contribution of C alpha-H...O hydrogen bonds to membrane protein stability depends on the position of the amide. Biochemistry *44*, 1607–1613.

Mueller, B.K., Subramaniam, S., and Senes, A. (2014). A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα-H hydrogen bonds. Proc. Natl. Acad. Sci. U. S. A. *111*, E888-895.

Müller, C.A., Hawkins, M., Retkute, R., Malla, S., Wilson, R., Blythe, M.J., Nakato, R., Komata, M., Shirahige, K., de Moura, A.P.S., et al. (2014). The dynamics of genome replication using deep sequencing. Nucleic Acids Res. *42*, e3.

Noderer, W.L., Flockhart, R.J., Bhaduri, A., Diaz de Arce, A.J., Zhang, J., Khavari, P.A., and Wang, C.L. (2014). Quantitative analysis of mammalian translation initiation sites by FACS-seq. Mol. Syst. Biol. *10*, 748.

Ouellette, S.P., Karimova, G., Davi, M., and Ladant, D. (2017). Analysis of Membrane Protein Interactions with a Bacterial Adenylate Cyclase-Based Two-Hybrid (BACTH) Technique. Curr. Protoc. Mol. Biol. *118*, 20.12.1-20.12.24.

Park, S., and Im, W. (2013). Two Dimensional Window Exchange Umbrella Sampling for Transmembrane Helix Assembly. J. Chem. Theory Comput. *9*, 13–17.

Park, H., Yoon, J., and Seok, C. (2008). Strength of Calpha-H...O=C hydrogen bonds in transmembrane proteins. J. Phys. Chem. B *112*, 1041–1048.

Park, S., Kim, T., and Im, W. (2012). Transmembrane helix assembly by window exchange umbrella sampling. Phys. Rev. Lett. *108*, 108102.

Park, Y., Elsner, M., Staritzbichler, R., and Helms, V. (2004). Novel scoring function for modeling structures of oligomers of transmembrane alpha-helices. Proteins *57*, 577–585.

Patching, S.G. (2015). Solid-state NMR structures of integral membrane proteins. Mol. Membr. Biol. *32*, 156–178.

Peterman, N., and Levine, E. (2016). Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. BMC Genomics *17*, 206.

Peterman, N., Lavi-Itzkovitz, A., and Levine, E. (2014). Large-scale mapping of sequence-function relations in small regulatory RNAs reveals plasticity and modularity. Nucleic Acids Res. *42*, 12177–12188.

Polyansky, A.A., Volynsky, P.E., and Efremov, R.G. (2012). Multistate organization of transmembrane helical protein dimers governed by the host membrane. J. Am. Chem. Soc. *134*, 14390–14400.

Polyansky, A.A., Chugunov, A.O., Volynsky, P.E., Krylov, N.A., Nolde, D.E., and Efremov, R.G. (2014). PREDDIMER: a web server for prediction of transmembrane helical dimers. Bioinforma. Oxf. Engl. *30*, 889–890.

Rodrigues, M.L., Oliveira, T.F., Pereira, I.A.C., and Archer, M. (2006). X-ray structure of the membrane-bound cytochrome c quinol dehydrogenase NrfH reveals novel haem coordination. EMBO J. *25*, 5951–5960.

Rohlhill, J., Sandoval, N.R., and Papoutsakis, E.T. (2017). Sort-Seq Approach to Engineering a Formaldehyde-Inducible Promoter for Dynamically Regulated Escherichia coli Growth on Methanol. ACS Synth. Biol. *6*, 1584–1595.

Roskoski, R. (2014). The ErbB/HER family of protein-tyrosine kinases and cancer. Pharmacol. Res. *79*, 34–74.

Russ, W.P., and Engelman, D.M. (1999). TOXCAT: a measure of transmembrane helix association in a biological membrane. Proc. Natl. Acad. Sci. U. S. A. *96*, 863–868.

Russ, W.P., and Engelman, D.M. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. J. Mol. Biol. *296*, 911–919.

Sarkar, C.A., Dodevski, I., Kenig, M., Dudli, S., Mohr, A., Hermans, E., and Plückthun, A. (2008). Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. Proc. Natl. Acad. Sci. U. S. A. *105*, 14808–14813.

Schanzenbach, C., Schmidt, F.C., Breckner, P., Teese, M.G., and Langosch, D. (2017). Identifying ionic interactions within a membrane using BLaTM, a genetic tool to measure homo- and heterotypic transmembrane helix-helix interactions. Sci. Rep. *7*, 43476.

Scheiner, S., Kar, T., and Gu, Y. (2001). Strength of the Calpha H..O hydrogen bond of amino acid residues. J. Biol. Chem. *276*, 9832–9837.

Schlinkmann, K.M., Honegger, A., Türeci, E., Robison, K.E., Lipovšek, D., and Plückthun, A. (2012). Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations. Proc. Natl. Acad. Sci. U. S. A. *109*, 9810–9815.

Schneider, D., and Engelman, D.M. (2003). GALLEX, a measurement of heterologous association of transmembrane helices in a biological membrane. J. Biol. Chem. *278*, 3105–3111.

Schneider, D., and Engelman, D.M. (2004). Motifs of two small residues can assist but are not sufficient to mediate transmembrane helix interactions. J. Mol. Biol. *343*, 799–804.

Schütz, M., Schöppe, J., Sedlák, E., Hillenbrand, M., Nagy-Davidescu, G., Ehrenmann, J., Klenk, C., Egloff, P., Kummer, L., and Plückthun, A. (2016). Directed evolution of G protein-coupled receptors in yeast for higher functional production in eukaryotic expression hosts. Sci. Rep. *6*, 21508.

Senes, A., Gerstein, M., and Engelman, D.M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. J. Mol. Biol. *296*, 921–936.

Senes, A., Ubarretxena-Belandia, I., and Engelman, D.M. (2001). The Calpha ---H...O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. Proc. Natl. Acad. Sci. U. S. A. *98*, 9056–9061.

Senes, A., Engel, D.E., and DeGrado, W.F. (2004). Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. Curr. Opin. Struct. Biol. *14*, 465–479.

Sengupta, D., and Marrink, S.J. (2010). Lipid-mediated interactions tune the association of glycophorin A helix and its disruptive mutants in membranes. Phys. Chem. Chem. Phys. PCCP *12*, 12987–12996.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

Sharon, E., van Dijk, D., Kalma, Y., Keren, L., Manor, O., Yakhini, Z., and Segal, E. (2014). Probing the effect of promoters on noise in gene expression using thousands of designed sequences. Genome Res. *24*, 1698–1706.

Singh, A.J., Chang, C.-N., Ma, H.-Y., Ramsey, S.A., Filtz, T.M., and Kioussi, C. (2018). FACS-Seq analysis of Pax3-derived cells identifies non-myogenic lineages in the embryonic forelimb. Sci. Rep. *8*, 7670.

Starr, T.N., Picton, L.K., and Thornton, J.W. (2017). Alternative evolutionary histories in the sequence space of an ancient protein. Nature *549*, 409–413.

Su, P.-C., and Berger, B.W. (2012). Identifying Key Juxtamembrane Interactions in Cell Membranes Using AraC-based Transcriptional Reporter Assay (AraTM). J. Biol. Chem. *287*, 31515–31526.

Su, P.-C., and Berger, B.W. (2013). A Novel Assay for Assessing Juxtamembrane and Transmembrane Domain Interactions Important for Receptor Heterodimerization. J. Mol. Biol. *425*, 4652–4658.

Sulistijo, E.S., and MacKenzie, K.R. (2006). Sequence dependence of BNIP3 transmembrane domain dimerization implicates side-chain hydrogen bonding and a tandem GxxxG motif in specific helix-helix interactions. J. Mol. Biol. *364*, 974–990.

Sulistijo, E.S., and Mackenzie, K.R. (2009). Structural basis for dimerization of the BNIP3 transmembrane domain. Biochemistry *48*, 5106–5120.

Teese, M.G., and Langosch, D. (2015). Role of GxxxG Motifs in Transmembrane Domain Interactions. Biochemistry *54*, 5125–5135.

Trenker, R., Call, M.E., and Call, M.J. (2015). Crystal Structure of the Glycophorin A Transmembrane Dimer in Lipidic Cubic Phase. J. Am. Chem. Soc. *137*, 15676–15679.

Treutlein, H.R., Lemmon, M.A., Engelman, D.M., and Brünger, A.T. (1992). The glycophorin A transmembrane domain dimer: sequence-specific propensity for a right-handed supercoil of helices. Biochemistry *31*, 12726–12732.

Truebestein, L., and Leonard, T.A. (2016). Coiled-coils: The long and short of it. BioEssays News Rev. Mol. Cell. Dev. Biol. *38*, 903–916.

UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204-212.

Vanoye, C.G., Desai, R.R., Fabre, K.L., Gallagher, S.L., Potet, F., DeKeyser, J.-M., Macaya, D., Meiler, J., Sanders, C.R., and George, A.L. (2018). High-Throughput Functional Evaluation of KCNQ1 Decrypts Variants of Unknown Significance. Circ. Genomic Precis. Med. *11*, e002345.

Vargas, R., Garza, J., Dixon, D.A., and Hay, B.P. (2000). How Strong Is the Cα−H···OC Hydrogen Bond? J Am Chem Soc *122*, 4750–4755.

Vilar, M., Charalampopoulos, I., Kenchappa, R.S., Simi, A., Karaca, E., Reversi, A., Choi, S., Bothwell, M., Mingarro, I., Friedman, W.J., et al. (2009). Activation of the p75 neurotrophin receptor through conformational rearrangement of disulphide-linked receptor dimers. Neuron *62*, 72–83.

Walters, R.F.S., and DeGrado, W.F. (2006). Helix-packing motifs in membrane proteins. Proc. Natl. Acad. Sci. U. S. A. *103*, 13658–13663.

Wang, Y., and Barth, P. (2015). Evolutionary-guided de novo structure prediction of self-associated transmembrane helical proteins with near-atomic accuracy. Nat. Commun. *6*, 7196.

Wang, S., Ding, M., Chen, X., Chang, L., and Sun, Y. (2017). Development of bimolecular fluorescence complementation using rsEGFP2 for detection and super-resolution imaging of protein-protein interactions in live cells. Biomed. Opt. Express *8*, 3119–3131.

Wierenga, R.K. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. FEBS Lett. *492*, 193–198.

Yin, H., Litvinov, R.I., Vilaire, G., Zhu, H., Li, W., Caputo, G.A., Moore, D.T., Lear, J.D., Weisel, J.W., Degrado, W.F., et al. (2006). Activation of platelet alphaIIbbeta3 by an exogenous peptide corresponding to the transmembrane domain of alphaIIb. J. Biol. Chem. *281*, 36732–36741.

Yohannan, S., Faham, S., Yang, D., Grosfeld, D., Chamberlain, A.K., and Bowie, J.U. (2004). A C alpha-H...O hydrogen bond in a membrane protein is not stabilizing. J. Am. Chem. Soc. *126*, 2284–2285.

Zhang, J., and Lazaridis, T. (2006). Calculating the free energy of association of transmembrane helices. Biophys. J. *91*, 1710–1723.

Zhang, Y., Sun, B., Feng, D., Hu, H., Chu, M., Qu, Q., Tarrasch, J.T., Li, S., Sun Kobilka, T., Kobilka, B.K., et al. (2017). Cryo-EM structure of the activated GLP-1 receptor in complex with a G protein. Nature *546*, 248–253.

Zhou, H.-X., and Cross, T.A. (2013). Influences of membrane mimetic environments on membrane protein structures. Annu. Rev. Biophys. *42*, 361–392.

Zviling, M., Kochva, U., and Arkin, I.T. (2007). How important are transmembrane helices of bitopic membrane proteins? Biochim. Biophys. Acta *1768*, 387–392.

Chapter 2: Combination of Cα – H Hydrogen Bonds and van der Waals Packing Modulates the Stability of GxxxG-Mediated Dimers in Membranes

This chapter was prepared for publication as:

Samantha M. Anderson*, Benjamin K. Mueller*, Evan J. Lange, and Alessandro Senes "Combination of Cα – H Hydrogen Bonds and van der Waals Packing Modulates the Stability of GxxxG-Mediated Dimers in Membranes" J Am Chem Soc. 2017 139 15774-83

*This work was completed with equal contributions from myself and Benjamin K. Mueller

2.1 Abstract

The GxxxG motif is frequently found at the dimerization interface of a transmembrane structural motif called GAS_{right}, which is characterized by a short interhelical distance and a right-handed crossing angle between the helices. In GAS_{right} dimers, such as glycophorin A (GpA), BNIP3, and members of the ErbB family, the backbones of the helices are in contact, and they invariably display networks of 4 to 8 weak hydrogen bonds between Cα–H carbon donors and carbonyl acceptors on opposing helices (Cα–H···O═C hydrogen bonds). These networks of weak hydrogen bonds at the helix–helix interface are presumably stabilizing, but their energetic contribution to dimerization has yet to be determined experimentally. Here, we present a computational and experimental structure-based analysis of GAS_{right} dimers of different predicted stabilities, which show that a combination of van der Waals packing and Cα–H hydrogen bonding predicts the experimental trend of dimerization propensities. This finding provides experimental support for the hypothesis that the networks of Cα–H hydrogen bonds are major contributors to the free energy of association of GxxxG-mediated dimers. The structural comparison between groups of GAS_{right} dimers of different stabilities reveals distinct sequence as well as conformational preferences. Stability correlates with shorter interhelical distances, narrower crossing angles, better packing, and the formation of larger networks of Cα–H hydrogen bonds. The identification of these structural rules provides insight on how nature could modulate stability in GAS_{right} and finely tune dimerization to support biological function.

2.2 Introduction

Oligomerization is critical for the biological function of many membrane proteins. In particular, oligomerization is important for the bitopic or “single-pass” proteins [i.e., those that span the membrane bilayer with a single transmembrane (TM) helix], which are the largest class of integral membrane proteins (Arkin and Brunger, 1998; Hubert et al., 2010; Wallin and von Heijne, 1998). Over 2300 single-pass proteins are predicted to exist in the human proteome alone, including oligomerizing systems such as receptor tyrosine kinases (Anbazhagan et al., 2010; Bocharov et al., 2008a, 2012; Chung et al., 2010; Mineev et al., 2010), cytokine receptors (Matthews et al., 2011; Vilar et al., 2009), integrins (Li et al., 2005; Yin et al., 2006), cadherins (Lai and Xu, 2007), apoptotic regulators (Bocharov et al., 2007; Lawrie et al., 2010; Sulistijo and MacKenzie, 2006), enzymes (Khadria et al., 2014), immunological complexes (Dixon et al., 2006), and many more (Teese and Langosch, 2015). The TM helices often have a critical role in driving and modulating the oligomerization of these systems, frequently acting in cooperation with the proteins’ soluble domains. Deciphering the rules that govern TM helix oligomerization in these systems is critical to understanding function and mechanisms of disease in a broad array of biological events.

The oligomerization of TM helices is often mediated by structural motifs that are evolutionarily optimized for protein–protein interactions (Walters and DeGrado, 2006; Zhang et al., 2015). One of the most prevalent dimerization motifs for single-pass proteins is the fold of the glycophorin A dimer (GpA), which is named GAS_{right} from the right-handed crossing angle between the helices (near −40°), and the presence of small amino acids (Gly, Ala, Ser: GAS) (Walters and DeGrado, 2006). These small residues are arranged to form GxxxG and GxxxG-like sequence motifs (GxxxG, GxxxA, SxxxG, etc.) (Brosig and Langosch, 1998; Russ and Engelman, 2000; Senes et al., 2000) typically found at the GAS_{right} dimerization interface (Figure 2.1a). As extensively reviewed by Teese and Langosch, GxxxG sequence motifs are

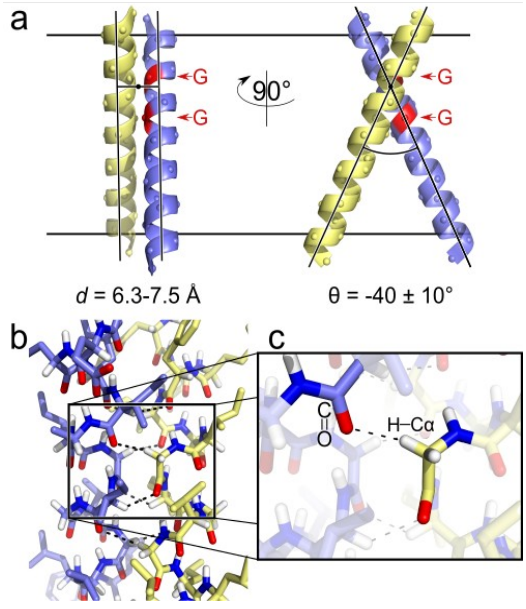


Figure 2.1 The GAS_{right} dimerization motif. (a) The GAS_{right} motif is a right-handed helical dimer with a short interhelical distance (6.3–7.5 Å) and a right-handed crossing angle of approximately −40°. The GxxxG sequence pattern at the crossing point (red) allows the backbones to come into contact. (b) The contact enables formation of networks of weak interhelical H bonds between Cα–H donors and carbonyl oxygen acceptors (shown in detail in c).

prevalent in biology, and they are frequently associated with parallel, right-handed GAS_{right} structures (although GxxxG can also be found in antiparallel or left-handed dimers and even at lipid-binding sites) (Teese and Langosch, 2015). The sequence context surrounding the GxxxG motif can modulate stability (Doura and Fleming, 2004; Li et al., 2012), and thus, the versatile GAS_{right} motif can be found both in proteins that form very stable “structural” dimers (such as GpA (MacKenzie et al., 1997) and BNIP3 (Sulistijo and MacKenzie, 2006)), as well as in weaker and dynamic systems in which changes in conformation or oligomerization state are necessary for supporting function (such as signaling in members of the ErbB receptor tyrosine kinase (Bocharov et al., 2008b, 2008a, 2012; Bragin et al., 2016; Endres et al., 2013; Mineev et al., 2010) and integrin families (Lau et al., 2009; Li et al., 2003, 2004)). Despite its common occurrence and importance, however, the fundamental physical rules that determine the strength of GAS_{right} dimerization are yet not well understood.

The major unknown is the contribution of weak hydrogen bonds that occur at the interface of GAS_{right} dimers to the free energy of dimerization. GAS_{right} invariably displays networks of

hydrogen bonds formed by Cα–H carbon donors and carbonyl acceptors (Cα–H···O═C), occurring in four to eight instances between atoms on opposing helices at the association interface (Figure 2.1, panels b and c) (Senes et al., 2001). In general, hydrogen bonding can be a stabilizing force in membrane proteins, and it has been shown that “canonical” hydrogen bonds (i.e., those formed by oxygen or nitrogen donors) can drive the interaction of TM helices (Bowie, 2011; Choma et al., 2000; Gratkowski et al., 2001; Zhou et al., 2000, 2001). Carbon is a weaker donor than oxygen or nitrogen, but Cα–H groups are activated by the flanking electron-withdrawing amide groups in the peptide backbone, and thus the strength of Cα–H hydrogen bonds has been estimated to be as much as one-half of the N–H donors in vacuum (Scheiner et al., 2001; Vargas et al., 2000). Therefore, it is plausible that multiple Cα–H hydrogen bonds occurring at the dimerization interface would contribute significantly to the free energy of association in GAS_{right} dimers (Mueller et al., 2014; Senes et al., 2001). Nevertheless, experimental demonstration of this hypothesis has, so far, remained elusive.

A major technical challenge in measuring the contribution of Cα–H hydrogen bonds to TM helix association in GAS_{right} dimers is the fact that both the donor and acceptor groups are part of the backbone, making a rational mutation strategy difficult to implement. To date, there have been only two experimental studies that have probed the contribution of Cα–H hydrogen bonds in membrane proteins. One of these studies was not performed on a GAS_{right} dimer but rather on the 7-TM helix membrane protein bacteriorhodopsin (Yohannan et al., 2004a). The study focused on the interaction between a Cα–H hydrogen bond donor and a threonine hydroxyl group acceptor and found that the removal of the side-chain acceptor group by mutation did not destabilize folding. However, it should be noted that the study targeted one isolated Cα–H hydrogen bond that occurs in the context of a large, multispan membrane protein. A second study investigated the energy of interaction of a Cα–H hydrogen bond in a GAS_{right} dimer by IR spectroscopy, estimating a favorable interaction energy of −0.88 kcal/mol between the Cα–H

donor of Gly 79 and the carbonyl of Ile 76 of GpA (Arbely and Arkin, 2004). This result supports the notion that Cα–H hydrogen bonds are likely significantly stabilizing. However, it is understood that geometry can play a significant role in determining the strength of Cα–H hydrogen bonds (Park et al., 2008), and this study is limited to a single specific bond among the many found in GpA. Moreover, the study measured hydrogen bonding strength but not its contribution to the free energy of dimerization, which has not been yet directly assessed.

The hypothesis that Cα–H hydrogen bonds are major contributors to the free energy of GAS_{right} dimerization remains compelling, particularly given by the unique ability of the structural motif to form this unusual feature. In fact, among all possible symmetric homodimeric configurations, GAS_{right} is the only one that promotes the formation of a large number of concurrent Cα–H hydrogen bonds (Mueller et al., 2014). This ability arises from three unique aspects of the geometry of GAS_{right}: (1) a crossing angle that precisely aligns Cα–H donors and carbonyl acceptors across two helices, (2) the presence of Gly at certain specific positions (producing the GxxxG pattern), where they are necessary to prevent clashing between the close helices, and (3) the ability of those same Gly residues to increment the number of Cα–H bonds by donating their second Hα. Therefore, GAS_{right} appears to be a structural motif optimized for the formation of Cα–H hydrogen bond networks.

We found that an algorithm (CATM) based on the simultaneous optimization of van der Waals forces and Cα–H hydrogen bonding was able to predict the small database of known three-dimensional structures of GAS_{right} homodimers to near atomic precision (Mueller et al., 2014), another finding that indirectly reinforces the importance of these forces in dimerization. The CATM algorithm was later successfully applied to predict the interface of a previously uncharacterized GxxxG-containing dimer, ADCK3, a mitochondrial protein that plays an essential role in the biosynthesis of coenzyme Q (Khadria et al., 2014). CATM can capture, with

remarkable accuracy, the structural features of a variety of GAS_{right} dimers. The success in predicting structure raises the question of whether the underlying energetic model can also capture, at least in part, the energetics of GAS_{right} dimerization.

To address this question, here we have combined CATM with a high-throughput biological assay to examine the relationship between structure and stability of GAS_{right} dimers of various geometries. We have applied CATM to the over 2300 sequences of TM domains of single-pass proteins present in the human genome, predicting the structure of hundreds of potential GAS_{right} dimers. We then selected candidates that represent a range of predicted dimerization stabilities and assessed their association propensity with TOXCAT, a widely used *in vivo* reporter assay that is sensitive to the relative association of TM dimers in a biological membrane (Russ and Engelman, 1999). After several steps of experimental validation, we obtained computational and experimental measurements for 26 well-behaved candidate GAS_{right} homodimers. We observe a significant correlation in the overall trend of energies predicted computationally and the dimerization propensities measured experimentally. These data provide the first experimental evidence for a model in which a combination of van der Waals forces and Cα–H hydrogen bonding acts as a primary source of stability, modulating the strength of GAS_{right} association.

2.3 Results and Discussion

2.3.1 Structural Prediction of GAS_{right} Homodimers

The CATM algorithm is designed to predict the structure of potential GAS_{right} homodimers from the amino acid sequence of a TM domain by docking the two helices and simultaneously optimizing van der Waals interactions, weak and canonical hydrogen bondings, and an implicit membrane solvation model. The algorithm only considers potential GAS_{right} conformations, and it does not explore the entire conformational range of a generic TM helix dimer, which makes it efficient and capable of searching for potential GAS_{right} dimers in high-throughput in large databases of TM sequences.

To create a diverse set of predicted GAS_{right} dimer structures to be tested experimentally, we drew sequences from the human proteome. The Uniprot database of annotated protein sequences currently identifies 2383 human proteins containing a single TM domain (The UniProt Consortium, 2017). When these TM domain sequences were run through CATM, they produced 1141 potential GAS_{right} dimers with a negative (i.e., favorable) energy score (dimer energy–monomer energy). The CATM scores assume a broad range of association energies, from −70 to 0 kcal/mol, with a skewed bell distribution (Figure 2.2a). The left tail of the distribution contains sequences enriched in well-packed structures with extensive Cα–H hydrogen bonding networks that are predicted to be very stable (Figure 2.2, panels b and c). The top 10% of the predicted structures form an average of 6.0 ± 1.7 Cα–H bonds. The predicted structures represent a rich repertoire of potential GAS_{right} dimers covering a wide range of predicted stabilities for follow-up experimental analysis.

For the subsequent experimental phase, we did not consider any sequence whose dimer interface contained strongly polar residues (Asp, Glu, Arg, Lys, His, Asn, and Gln) or Pro. We chose to exclude these residues because proline has a tendency to form kinks in helices that

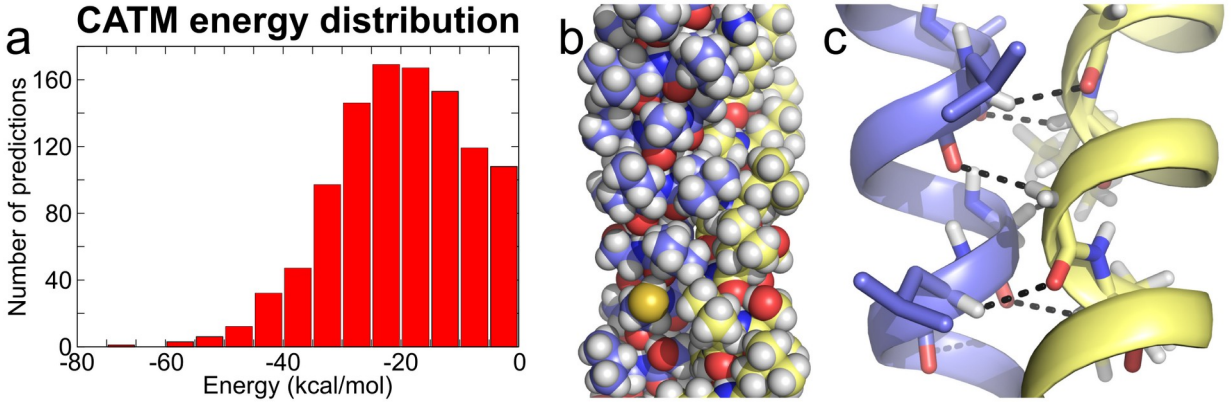


Figure 2.2 Energy distribution of CATM predicted GAS_{right} dimers in human single-pass sequences. (a) Histogram of calculated energies of human GAS_{right} dimers. CATM identified 1141 sequences that produced a model with a negative (favorable) energy of association. (b) Extensive complementary packing, as well as (c) the characteristic networks of Cα–H hydrogen bonds displayed by the lowest energy structures, chondroitin sulfate glucuronyltransferase (Uniprot accession Q9P2E5).

are difficult to predict (Senes et al., 2004; Yohannan et al., 2004b), while strongly polar residues have a propensity to drive TM association through the formation of interhelical hydrogen bonds (Gratkowski et al., 2001; Hong et al., 2013; Zhou et al., 2000, 2001). Their inclusion would have increased the probability of dimers mediated by nonspecific interfaces, breaking the desired structural correspondence between the model predicted by CATM and the constructs in experimental conditions. These exclusions reduced the number of available sequences to 668. We also excluded sequences with predicted marginal stability (a score higher than −5 kcal/mol). From the remaining 604 sequences, we randomly selected 65 diverse candidates for experimental analysis (Tables S2.1–S2.4).

2.3.2 Experimental Strategy: TOXCAT Assay Using Standardized Sequences

To experimentally assess the dimerization of the 65 predicted GAS_{right} dimers and their mutants, we used TOXCAT, a widely adopted assay that measures TM homo-oligomerization in biological membranes (Russ and Engelman, 1999). This system is based on the *in vivo*

expression of a chimeric protein in the inner membrane of *Escherichia coli* in which the TM domain of interest is fused to the ToxR transcriptional activator. Dimerization of the TM helices brings together two ToxR subunits, which bind to a specific promoter, activating the expression of the reporter gene chloramphenicol acetyltransferase (CAT). Quantification of CAT thus provides an indication of the extent of TM helix–helix association in a biological membrane (Figure 2.3a).

The general relationship between reporter gene expression in TOXCAT and thermodynamic stability of any given dimer is likely complex, but reasonable correlation has been found for collections of point mutants of GpA and their energy of dimerization in detergents (Duong et al., 2007; Elazar et al., 2016). In these studies, the constructs are homogeneous, having identical length of the TM region, nearly identical sequence, and comparable hydrophobicity. Because

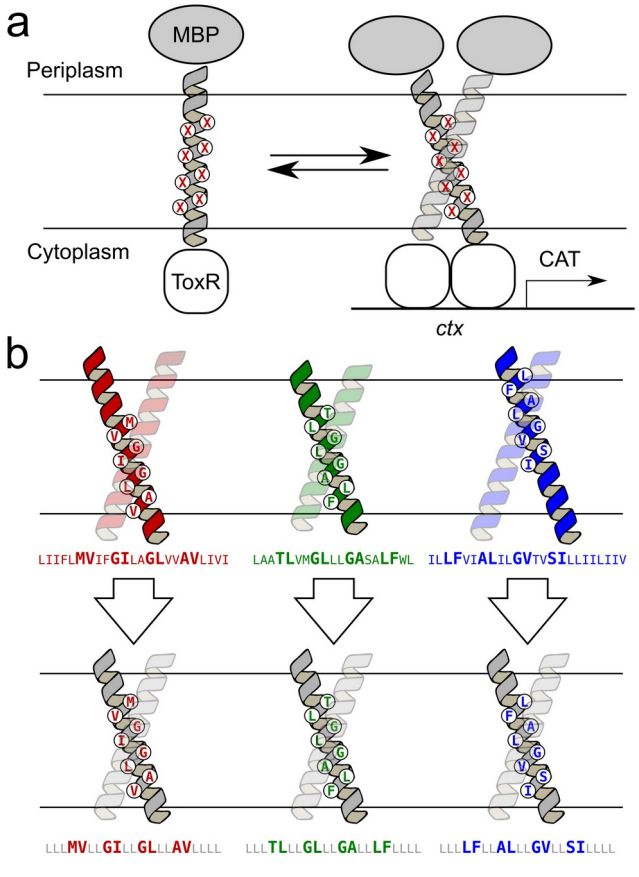


Figure 2.3 Experimental design. (a) TOXCAT is an *in vivo* assay based on a construct in which the TM domain under investigation is fused to the ToxR transcriptional activator. TM association results in the expression of a reporter gene in *E. coli* cells, which can be quantified. (b) To reduce variability in TOXCAT, the eight interfacial amino acids identified by CATM in the wild-type sequences (top) were “stitched” into a standardized poly-Leu sequence (bottom). Standardization of the predicted constructs retains the geometry of the interface while controlling the length of the TM helix, the position of the crossing point, and the hydrophobicity of the TM segment.

TOXCAT's response may be dependent on these variables (Johnson et al., 2006; Kirrbach et al., 2013; Lawrie et al., 2010), controlling them is likely to simplify the comparison between constructs. The predicted lengths of the TM domains of the 2383 human single-pass sequences in Uniprot range widely (Senes et al., 2000), and their estimated Δ*G* of membrane insertion ranges from −6.7 to +11.9 kcal/mol (using the biological ΔG_{app} predictor (Hessa et al., 2007)). To reduce heterogeneity as much as possible, we adopted a strategy of "stitching" the 8 positions predicted by CATM to be at the helix–helix interface of a standardized TM helix of 21 amino acids consisting of a poly-Leu backbone (LLLxxLLxxLLxxLLxxLILI, where the x represents the variable interfacial positions).

As illustrated in Figure 2.3b, this stitching strategy ensures that all constructs have the same TM domain length and that the predicted interface is centered in the middle of the membrane. Perhaps most importantly, the standardized sequence reduces the variability in hydrophobicity. Because the noninterfacial residues in all the constructs remain constant, the ΔG_{app} range for membrane insertion is reduced to −6.6 to −2.9 kcal/mol, likely leading to a more consistent expression of the constructs in the *E. coli* membrane. Another important reason for standardizing all noninterfacial positions is that the strategy removes potential alternative dimerization interfaces that may be present within the wild-type sequence because only the amino acids involved in the predicted GAS_{right} interfaces are carried over into the standardized constructs.

There is an existing precedent for such a strategy with GAS_{right} homodimers: it has been shown that the interfacial residues of GpA in a leucine backbone behave similarly to the wild-type sequence (Russ and Engelman, 1999). In addition, a pure poly-Leu sequence has a relatively low propensity for self-association in TOXCAT (Ruan et al., 2004; Zhou et al., 2000, 2001), which is important for reducing the risk of alternate interfaces. To ensure that the interfaces of the standardized sequences were consistent with those initially predicted for the

wild-type sequences, the standardized sequences were also evaluated with CATM (Tables S2.1–S2.4). We found that CATM consistently predicts nearly identical interfaces for wild-type and standardized constructs. The computed energies that we report for our analysis below correspond to those calculated using the standardized poly-Leu construct and not the original wild-type sequences.

2.3.3 Experimental Validation of Predicted Structures

To partially validate the predicted structural models, we adopted a mutagenesis strategy. Saturation mutagenesis has been commonly used to identify or confirm the interface of TM dimers (Khadria et al., 2014; LaPointe et al., 2013; Lawrie et al., 2010; Wei et al., 2011, 2013). Because it would be impractical to perform saturation mutagenesis of all 65 candidate constructs, we opted to introduce in each construct a single mutation predicted to be highly detrimental, selecting the most sensitive interfacial position of GAS_{right} homodimers, the so-called “C1” position, as defined in our previous work (Mueller et al., 2014). The C1 position is one of the residues near the crossing point of the helical dimer. In GAS_{right} homodimers, C1 is required to be occupied by Gly in order to allow contact between the backbones of the two helices (Mueller et al., 2014). Substitution of Gly at C1 with a large hydrophobic amino acid, such as Ile, would push the helices apart and completely eliminate any potential association mediated by the predicted interface. We computationally verified that all models of $C1_{Gly\rightarrow Ile}$ variants contained significant clashes. Introduction of this control enabled the removal of constructs that retained significant association in TOXCAT after the $C1_{Gly\rightarrow Ile}$ mutation, since these results suggest that the dimerization observed experimentally was not mediated by the predicted GAS_{right} structural model (or, alternatively, that a second possible dimerization interface is also present in the construct, which is not disrupted by the $C1_{Gly\rightarrow Ile}$ mutation).

To confirm proper membrane insertion, each of the 65 constructs and their $C1_{Gly→Ile}$ variants was tested for its ability to support growth in minimal medium containing maltose as the only carbon source, as standard practice in TOXCAT (Russ and Engelman, 1999). A total of 15 constructs (wild type or C1 variant) did not fully grow in these conditions (Table S2.2). These constructs were not further considered in the study. We then eliminated constructs whose TOXCAT signal was below the minimal threshold of a pure poly-Leu construct because we would not be able to differentiate specific GAS_{right}-mediated dimerization from background association. A pure poly-Leu construct displays approximately 30% of the CAT expression level of the GpA standard, therefore any construct below the 30% threshold was eliminated (10 constructs, Table S2.3).

Finally, any constructs whose $C1_{Gly→Ile}$ control variant scored above 30% of relative CAT expression level were also eliminated from the analysis because they did not match our expected model, as explained earlier (14 constructs, Table S2.4). As an exception to this rule, if a $C1_{Gly→Ile}$ mutation reduced the “wild-type” CAT activity by at least 75% we retained it for analysis, even if it was above the 30% threshold, because of the dramatic reduction in dimerization (3 constructs). The final 26 GAS_{right} constructs are listed in Table S2.1. Their predicted structural models are illustrated in Figure S2.1. The progression from the 2328 genomic sequence to the final 26 experimental constructs is summarized in Table S2.5. We verified the expression of the ToxR-TM-MBP chimeras of the 26 constructs by Western blots: the constructs displayed rather homogeneous levels of expression, with a standard deviation of 22% (Figure S2.2).

2.3.4 Cα–H Hydrogen Bonds and vdW Predict Experimental Association Propensities

The comparison of association energies calculated with CATM and dimerization propensities assessed by TOXCAT for the 26 selected constructs is shown in Figure 2.4a. The plot shows a statistically significant correlation (R^2 = 0.441, p < 0.0005, t test of linear regression slope). One clear outlier is present in the plot (the TNR12 construct, TOXCAT 119%, CATM −6 kcal/mol, highlighted in gray): if this point is excluded, the R^2 increases to 0.647 (p < 0.000005). The correlation is also statistically significant by rank order correlation coefficient analysis, which does not assume a linear model (r = −0.683, p < 0.005, and r = −0.827, p < 0.000001, with and without TNR12, respectively) (Spearman, 1904). Some of the variance is likely due to the biological nature of the TOXCAT assay, some to imprecision by CATM in predicting the structures, and the remaining variance can be attributed to the limitations of the energy model, which was constructed solely on its ability to predict structure. However, the energetic model is clearly able to capture the trend of dimerization propensities observed experimentally.

2.3.5 Structural and Sequence Analysis of Groups with Distinct Stability

Interesting differences in structural and sequence features are observed among constructs with different dimerization propensities. To appreciate these structural and energetic properties that distinguish strong from weaker dimers, we grouped the data according to five levels of TOXCAT signal, using five 25%-wide bins, from very weak (25–50% GpA) to very strong apparent dimerization (>125% GpA). We first confirmed that the energy model is sensitive enough to distinguish between the five stability groups. Indeed, proportionality is retained after TOXCAT and CATM values are averaged within each groups (Figure 2.4b). Linear regression of these averaged values produces a significant fit (p < 0.01), with a R^2 value of 0.931 if the TNR12 outlier is excluded, and a R^2 value of 0.883 when TNR12 is included (p < 0.05, Figure S2.3). The regression analyses of Figure 2.4 (panels a and b) produce two distinct equations of the line, which is an expected mathematical outcome of averaging. However, it should be noted that a linear relationship is likely not the correct physical model and is not necessarily expected

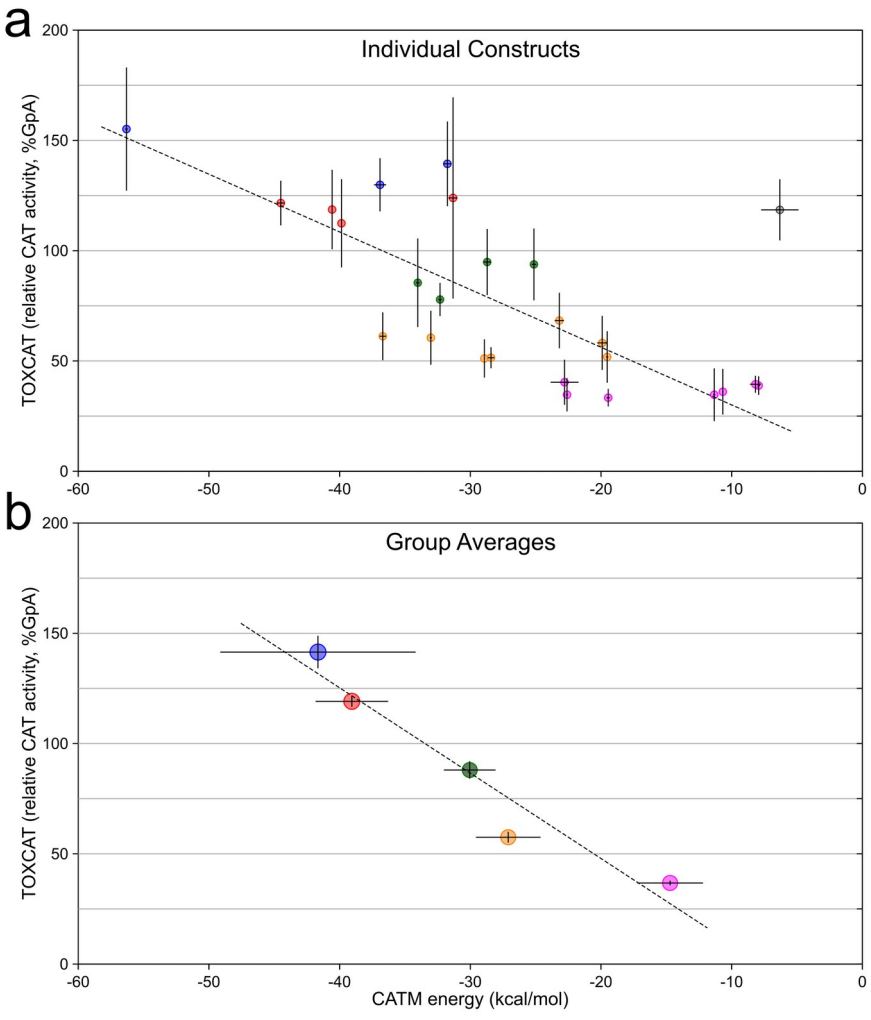


Figure 2.4 Comparison of CATM energies with apparent TOXCAT dimerization. (a) Comparison of CATM energy score of 26 sequences and their TOXCAT signal (measured as the enzymatic activity of the reporter gene CAT). The points are color-coded according to the grouping in (b). The error bars represent the standard deviation among replicates. The dashed line represents the linear regression fit of the data, with the exclusion of the outlier point highlighted in gray (R^2 = 0.647, p < 0.000005). (b) Same data as in (a), grouped and averaged in five bins based on CAT activity from weak (>25%, magenta) to very strong (>125%, blue), in 25% intervals. The error bars represent the standard error of the average. The dashed line is the linear regression of the data (R^2 = 0.931, p < 0.01). The groups are the base of the analysis reported in Figure 2.5.

(MacKenzie and Fleming, 2008). What is important is that there is proportionality between TOXCAT and CATM outcomes, and that the energetic model is able to clearly differentiate among the five sets of constructs. Therefore, the grouping is suitable for a comparative analysis of sequence and structural features that characterize constructs with increasing apparent stability. Statistical analysis of the trends independent of grouping is also provided.

2.3.6 Stability Correlates with Sequence Biases

The results of the sequence and structural features of the five groups are summarized in Table 2.1 and Figure 2.5. Some sequence biases at the interface of the predicted dimers were already present in the initial pool of 604 sequences, as expected for a selection of GAS_{right} dimers (Figure S2.4). Most notably, the sequences are enriched with GxxxG and GxxxG-like motifs, and Gly is nearly absolutely preserved at position C1, where this amino acid is required for interhelical backbone contact (the nomenclature of the positions is defined in Figure 2.5b) (Mueller et al., 2014). However, on top of these biases, a number of interesting trends emerged within our experimental pool that correlate statistically with their stability.

The first trend is the frequency of the GxxxG motif, which increases from the least to the most stable groups (Figure 2.5c, orange symbols, and Table 2.1). In particular, the three more stable groups (>75%, >100%, >125%) contain GxxxG motifs in all but one construct, formed by the C1 Gly and a second Gly either at N1 (the position at *i*-4 from C1) or at C5 (the position at *i*+4); conversely, in the two less stable groups (>50% and >25%) GxxxG is found in just 43% of the sequences. The biased distribution of GxxxG containing sequences is confirmed, independently from the grouping scheme, using Point Biserial correlation statistics, which measures correlation between a continuous variable (TOXCAT signal) and a binary variable (occurrence of GxxxG) (correlation coefficient $r = +0.63$, $p < 0.001$) (Tate, 1954). The fact that GxxxG is present in the most stable constructs is not surprising. However, it should be noted that some low-stability sequences also contain GxxxG, further demonstrating that the presence

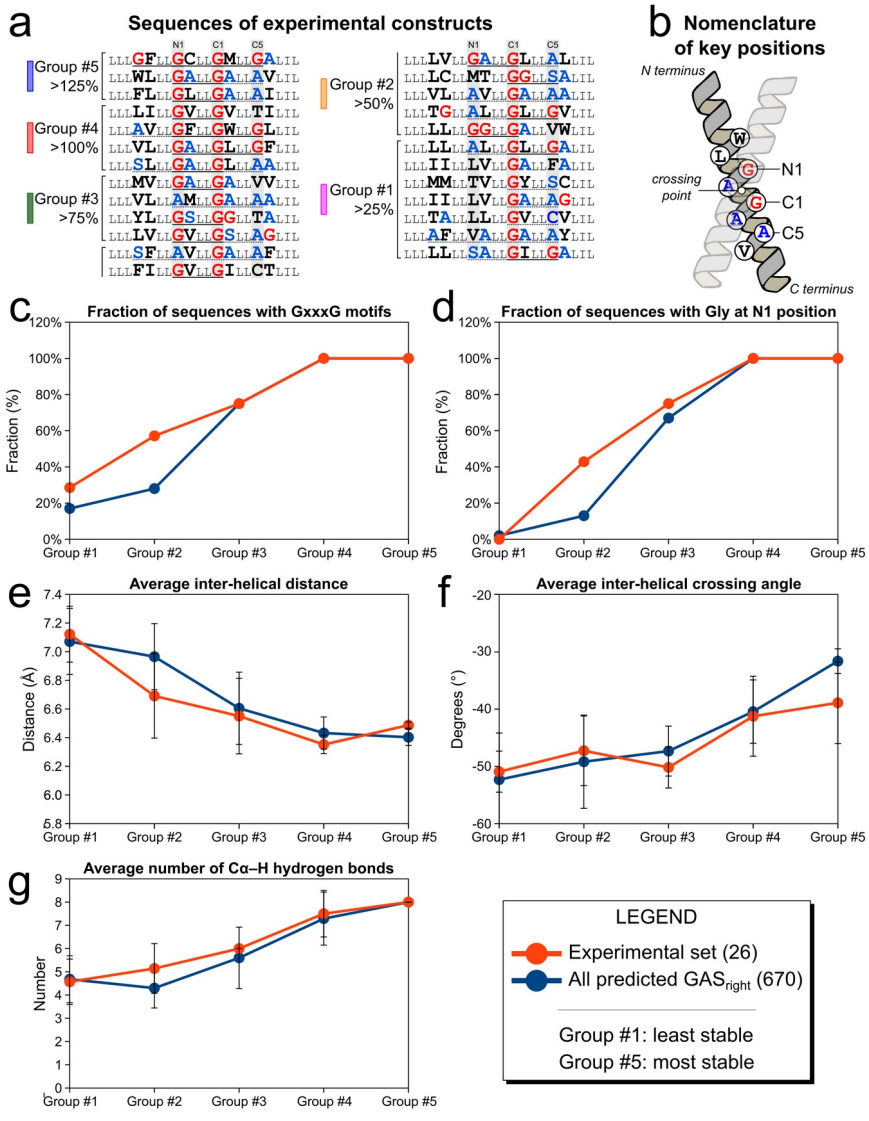


Figure 2.5 Sequence and structural bias occur in groups with different stabilities. (a) Sequences of the 26 constructs ranked by TOXCAT signal showing the groups, as defined in Figure 2.4. GxxxG motifs are underlined with a solid line, GxxxG-like motifs with a dotted line. Color coding as in Figure 2.b. (b) Nomenclature of the interfacial positions, as defined previously (Mueller et al., 2014). The sequence and structural biases of the groups of experimental constructs (orange symbols) are illustrated for (c) the number of Cα–H hydrogen bonds, which increases with stability, (d) the interhelical distance, and (e) crossing-angle, which decrease with stability, and the fraction of sequences containing (f) GxxxG and (g) Gly at the N1 position, which also increase. Data also reported in Table 2.1. The same trends are observed in groups of different stabilities computed from the entire data set of 670 structures predicted from the human proteome (blue symbols).

Table 2.1 Energetic and geometric properties of groups of constructs of different apparent dimerization and statistical significance of the distributions

TOXCAT range (% GpA)[1]	25-50%	50-75%	75-100%	100-125%	125+%	Correlation with TOXCAT
Number of constructs	7	7	4	4	3	
Average TOXCAT (%GpA)[3]	37±1	58±2	88±4	118±2	141±7	
CATM energy score (kcal/mol)[3]	-14.7±2.5	-27.1±2.5	-30.0±2.0	-39.1±2.8	-41.7±7.5	$p < 0.000001$[4]
Van der Waals (kcal/mol)	-26.2±5.3	-33.7±4.5	-33.6±2.1	-39.3±2.4	-39.0±11.1	$p < 0.005$[4]
Cα-H hydrogen bonding (kcal/mol)	-5.2±1.1	-8.0±1.9	-9.7±0.5	-12.0±2.3	-13.0±0.8	$p < 0.000001$[4]
Solvation (kcal/mol)	16.7±1.9	14.2±1.9	13.3±2.0	11.7±2.4	10.6±2.7	$p < 0.00005$[4]
Crossing angle (°)	-51±4	-47±6	-49±2	-41±7	-39±7	$p < 0.01$[4]
Number of Cα-H bonds	4.6±1.0	5.1±1.1	6.0±0.0	7.5±1.0	8.0±0.0	N/A[5]
Interface surface area (Å^2)	4810±490	4660±500	4630±190	4770±540	4510±280	–
Inter-helical distance (Å)	7.1±0.2	6.7±0.3	6.5±0.1	6.4±0.1	6.5±0.0	$p < 0.00005$[4]
Van der Waals/Interface surface area (kcal/(mol Å^2))	-0.0054±0.0012	-0.0073±0.0015	-0.0073±0.0005	-0.0083±0.0008	-0.0086±0.0021	$p < 0.001$[4]
Sequences with GxxxG	2/7	4/7	3/4	4/4	3/3	$p < 0.01$[6]
Sequences with Sm-xxx-Sm[1]	6/7	7/7	4/4	4/4	3/3	–
Sequences with Gly at N1	0/7	3/7	3/4	4/4	3/3	$p < 0.0001$[6]
Sequences with Gly at C1	7/7	7/7	4/4	4/4	3/3	–
Sequences with Gly at C5	2/7	1/7	0/4	2/4	1/3	–

[1]All values are reported as averages ± standard deviation. The outlier TNR12 was excluded from the 100-125% group.

[2]Sm-xxx-Sm are defined by any combinations of Gly, Ala, Ser and Cys at the first and last position.

[3]Values are reported as averages ± standard error as in Figure 2.4.

[4]Rank Order (Spearman) Correlation analysis (Spearman, 1904)

[5]Rank correlation statistics not applicable to non-continuous variable

[6]Point Biserial Correlation analysis (Tate, 1954)

of the sequence motif is not the sole determinant of stability (Doura and Fleming, 2004; Li et al., 2012). Notably, all sequences that do not contain a GxxxG motif contain a Small-xxx-Small motif (i.e., GxxxA, SxxxG, etc.), with the exception of one low affinity construct (1A32–2).

The first trend is the frequency of the GxxxG motif, which increases from the least to the most stable groups (Figure 2.5c, orange symbols, and Table 2.1). In particular, the three more stable groups (>75%, >100%, >125%) contain GxxxG motifs in all but one construct, formed by the C1 Gly and a second Gly either at N1 (the position at i-4 from C1) or at C5 (the position at i+4); conversely, in the two less stable groups (>50% and >25%) GxxxG is found in just 43% of the sequences. The biased distribution of GxxxG containing sequences is confirmed, independently from the grouping scheme, using Point Biserial correlation statistics, which measures correlation between a continuous variable (TOXCAT signal) and a binary variable (occurrence of GxxxG) (correlation coefficient $r = +0.63$, $p < 0.001$) (Tate, 1954). The fact that GxxxG is present in the most stable constructs is not surprising. However, it should be noted that some low-stability sequences also contain GxxxG, further demonstrating that the presence of the sequence motif is not the sole determinant of stability (Doura and Fleming, 2004; Li et al., 2012). Notably, all sequences that do not contain a GxxxG motif contain a Small-xxx-Small motif (i.e., GxxxA, SxxxG, etc.), with the exception of one low affinity construct (1A32–2).

The GxxxG motifs in the sequence can be formed by the invariable Gly at C1 together with a second Gly at either N1 or C5. However, the marked increase of GxxxG in the most stable constructs is primarily due to the presence of a Gly at position N1 (Figure 2.5d). In the three most stable groups 10 out of 11 of the sequences have a Gly at N1, whereas Gly occurs rarely at the same position in the lower stability groups ($p < 0.0001$). Conversely, Gly at C1 is rarer and its presence does not correlate with apparent stability.

2.3.7 Stability Correlates with Structural Features

These trends suggest that distinct sequence biases occur among GAS_{right} dimers of different stabilities. To understand their physical basis, we looked at how structural parameters varied as a function of apparent stability. We observed numerous structure-related differences, which are summarized in Table 2.1 and Figure 2.5. The analysis indicates that as stability

increases, (i) the distance between the helices becomes increasingly shorter, (ii) the crossing angle becomes smaller, (iii) the structural models become increasingly better packed, and (iv) they display larger networks of Cα–H hydrogen bonds.

The interhelical distance (measured between the helical axes) progressively decreases from an average of 7.1 Å for the lowest stability set, down to 6.5 Å for the most stable group, which is near the closest two helices can approach before their backbones would sterically clash (Figure 2.5e). The correlation between TOXCAT and interhelical distance is statistically significant (rank order spearman correlation coefficient r = +0.74, p < 0.0005, Table 2.1). We also observe a reduction of the interhelical angle, which progressively decreases toward −40° (p < 0.01, Figure 2.5f). These geometric changes are favored by the presence of Gly at N1, as discussed in the previous section.

The tighter interhelical contact in the most stable constructs leads to an increase of favorable van der Waals interactions between the helices (p < 0.005), which improve by 49% from the lowest to the highest dimerizing groups from −26.2 to −39.0 kcal/mol. These improved van der Waals interactions do not originate from a larger dimer interface (which remains relatively constant across the sets), therefore they are attributable to better packing. The more intimate interhelical contact also favors a very significant change in hydrogen bonding: the number of interhelical Cα–H hydrogen bonds increases from 4.6 to 8.0 on average (Figure 2.5g). Correspondingly, the average contribution of hydrogen bonding to the binding energy more than doubles from −5.2 to −13.0 kcal/mol (p < 0.000001). Finally, we observe a reduction of the cost of desolvating the helices (from +17 kcal/mol to +11 kcal/mol) that also contributes to the better energy score computed in CATM for the more stable dimers (p < 0.00005).

To further investigate these sequence and geometry biases, we performed a similar analysis on the entire set of 604 poly-Leu predicted structures, grouping the results by decreasing CATM energy (blue symbols in Figure 2.5 and Table S2.7). We observe very similar trends across all examined variables. A progressive reduction of the crossing angle and interhelical distance is

observed as the energy score decreases, along with an increased number of Cα–H hydrogen bonds and improvement of the packing. Similarly, presence of a GxxxG motif increases, reaching 100% in the lowest energy groups. The trend mirrors the presence of Gly at position N1, whereas the presence of Gly at C5 (the second position that can form a GxxxG motif with C1) also increases but not as dramatically, topping at 50% in the lowest energy group.

In summary, the model suggests that the stronger interactions tend to be formed by helices that have a closer distance and a smaller crossing angle. These geometries tend to be favored by the presence of a second Gly at N1, forming a GxxxG motif with the Gly at C1, although the precise stability and conformation of each dimer is influenced by its entire sequence context. It is possible that some of the observed results may be influenced by the current experimental conditions. For example, the optimal crossing angle could be sensitive to the thickness of the membrane and the length of the TM helices, which were not varied in either the computational or the TOXCAT experiments. Nevertheless, these biases provide important insight into how the GAS_{right} sequence is able to modulate stability, a feature that is likely important for a structural motif that is found in both stable constitutive dimers, as well as in weaker “dynamic” or transient dimers, where dissociation or conformational change is required for the function of the protein.

2.4 Conclusions

An unusual interaction is at the core of one of the most common transmembrane motifs, and yet the contribution of these Cα–H hydrogen bond networks to the free energy of dimerization has remained uncertain. This is in part due to scarce availability of structures, which poses a serious hurdle to understanding the structural basis of TM helix oligomerization. Structure-based analysis has been possible for a few structurally characterized dimeric systems, such as GpA (Doura and Fleming, 2004; Fleming et al., 2004) and BNIP3 (Sulistijo and MacKenzie, 2006). Conversely, large-scale comparative analyses, based either on combinatorial libraries (Dawson et al., 2003; Herrmann et al., 2010; Ridder et al., 2005; Russ and Engelman, 2000; Schanzenbach et al., 2017; Unterreitmeier et al., 2007), comprehensive protein families(Mendrola et al., 2002),or homology-based clusters of human proteins (Kirrbach et al., 2013) have been performed primarily on sequences of unknown structure. Computational modeling has often been applied in coordination to these approaches to aid in the interpretation of experimental data (Adams et al., 1995; Dixon et al., 2006, 2014; Engelman et al., 1993; Howard et al., 2002; Khadria et al., 2014; LaPointe et al., 2013; Li et al., 2005; Stouffer et al., 2005; Wei et al., 2013; Zhang et al., 2009; Zhu et al., 2009). An advance of the present work is the availability of a reliable structural prediction method, which has enabled the design of an experimental analysis of dimers of diverse stabilities to test pre-existing structural and energetic models.

This analysis addressed the question of whether Cα–H hydrogen bonding and van der Waals forces are predictive of the dimerization propensity of GAS_{right} dimers. The results provide the first experimental support for the hypothesis that Cα–H hydrogen bonds are indeed major determinants of dimerization in GxxxG-mediated dimers. Our data complement the only other experimental report in the literature that has shown that Cα–H hydrogen bonds have the potential to stabilize GAS_{right}, a study by Arbely & Arkin that measured the strength of a single

hydrogen bond interaction in a well-characterized GAS_{right} model system (Arbely and Arkin, 2004); here, we addressed the role of Cα–H hydrogen bonds as contributor to the free energy of dimerization, examined at the level of the entire structural motif.

We found that a simple energy model combining Cα–H hydrogen bonding and van der Waals already forms a good base when tested in biological membranes, albeit in standardized sequence conditions. The present analysis also provides initial insight on how change in the sequence and geometry may modulate these terms and therefore overall stability in GAS_{right}. The results also suggest that, with more data, a similar strategy would likely support the development of a more sophisticated energy function, which would provide further insight into the forces involved in GAS_{right} association as well as improve our ability to accurately predict structure and stability of these dimers from primary sequence data alone.

2.5 Methods

2.5.1 Software

All calculations were implemented and performed using MSL v. 1.1 (Kulp et al., 2012), an open source C++ library that is freely available at http://msl-libraries.org.

2.5.2 Prediction of GAS_{right} Structure and Dimerization Energy

The structure of GAS_{right} dimers was predicted from a database of 2383 human sequences annotated as single-pass membrane proteins in Uniprot (as of November 2, 2016) (The UniProt Consortium, 2017). Structural prediction was performed with the program CATM (Mueller et al., 2014). Side chain mobility was modeled using the energy-based conformer library applied at the 95% level (Subramaniam and Senes, 2012). Energies were determined using the CHARMM 22 van der Waals function (MacKerell et al., 1998), the IMM1 membrane implicit solvation model (Lazaridis, 2003), and the hydrogen bonding function of SCWRL 4 (Krivov et al., 2009), as implemented in MSL (Kulp et al., 2012), with the following parameters for Cα donors, as reported previously: B = 60.278, D_0 = 2.3 Å, σ_d = 1.202 Å, α_{max} = 74.0°, and β_{max} = 98.0° (Mueller et al., 2014).

The CATM algorithm was described in detail previously (Mueller et al., 2014). Briefly, the sequence of interest is threaded into a set of different registers at each of 463 representative geometries. If sequence-based filtering rules are met, the sequence is built on the backbone in all atoms and the helices are docked by reducing the interhelical distance in steps. At each step, the side chains are optimized and the interaction energy is evaluated until a minimum energy is found. To further optimize the dimer, the geometry is then subjected to Monte Carlo backbone perturbation cycles in which all interhelical parameters (distance, Z shift, axial rotation, and crossing angle) are locally varied. If the final interaction energy (calculated as the energy of the dimer minus the energy of two monomers separated at long distance) is negative, the solution is

accepted. The solutions are then clustered using an RMSD criterion to produce a series of distinct models. The computation produced 1141 structures of predicted GAS_{right} homodimers. These structures are available at http://seneslab.org/CATM.

2.5.3 Cloning and Expression of Chimeric Proteins in MM39 Cells and MalE Complementation Assay

DNA sequences containing the transmembrane region of interest were cut with NheI and DpnII restriction enzymes and cloned into the NheI-*Bam*HI restriction sites of the pccKAN vector as previously described (Khadria et al., 2014; LaPointe et al., 2013). The TOXCAT constructs were transformed into MM39 cells. A freshly streaked colony was inoculated into 3 mL of LB broth containing 100 μg/mL ampicillin and grown overnight at 37 °C. 50 μL of overnight cultures were inoculated into 3 mL of LB broth and grown to an OD_{600} of 0.8–1.0 at 37 °C. After recording the optical density, 1 mL of cells was spun down for 15 min at 17000*g* and resuspended in 500 mL of sonication buffer (25 mM Tris-HCl, 2 mM EDTA, pH 8.0). Cells were lysed by probe sonication at medium power for 10 s over ice. An aliquot was removed from each sample and stored in SDS–PAGE loading buffer for immunoblotting. The lysates were then cleared by centrifugation at 17000*g*, and the supernatant was kept on ice for chloramphenicol acetyltransferase (CAT) activity assay.

To confirm proper membrane insertion and orientation of the TOXCAT constructs, overnight cultures were plated on M9 minimal medium plates containing 0.4% maltose as the only carbon source and grown at 37 °C for 48–72 h. The variants that did not grow in these conditions were not considered for this study.

2.5.4 Chloramphenicol Acetyltransferase (CAT) Spectrophotometric Assay

CAT activity was measured as described (LaPointe et al., 2013; Shaw, 1975). Briefly, 750 μL of buffer containing 0.4 mg/mL 5,5′- dithiobis(2-nitrobenzoic acid) or Ellman's reagent and 0.1 M Tris-HCl, pH 7.8, were mixed with 250 μL of 0.4 mM acetyl CoA and 40 μL of cleared cell

lysates, and the absorbance at 412 nm was measured for 2 min to establish basal enzyme activity rate. After addition of 40 μL of 2.5 mM chloramphenicol in 10% ethanol, the absorbance was measured for an additional 2 min to determine CAT activity. The basal CAT activity was subtracted, and the value was normalized by the cell density measured as OD_{600}. All measurements were determined by at least four independently cultured biological replicates, each of which was measured with two technical replicates.

2.5.5 Quantification of Expression by Immunoblotting

Protein expression was confirmed by immunoblotting. The cell lysates were normalized by cell density and loaded onto a NuPAGE 4–12% bis-tris SDS–PAGE gel (Invitrogen) and then transferred to PVDF membranes (VWR) for 1 h at 100 millivolts. Blots were blocked using 5% bovine serum albumin (US Biologicals) in TBS-Tween buffer (50 mM Tris, 150 mM NaCl, 0.05% Tween 20) for overnight at 4 °C, incubated with goat biotinylated anti-maltose binding protein antibodies (Vector laboratories) for 2 h at room temperature, followed by peroxidase-conjugated streptavidin antigoat secondary antibodies (Jackson ImmunoResearch) for 2 h at 4 °C. Blots were developed with the Pierce ECL Western Blotting Substrate Kit; 1 mL of ECL solution was added to the blot and incubated for 90 s. Chemiluminescence was measured using an ImageQuant LAS 4000 (GE Healthsciences). Individual bands were quantified by ImageJ.

2.6 Supporting Information

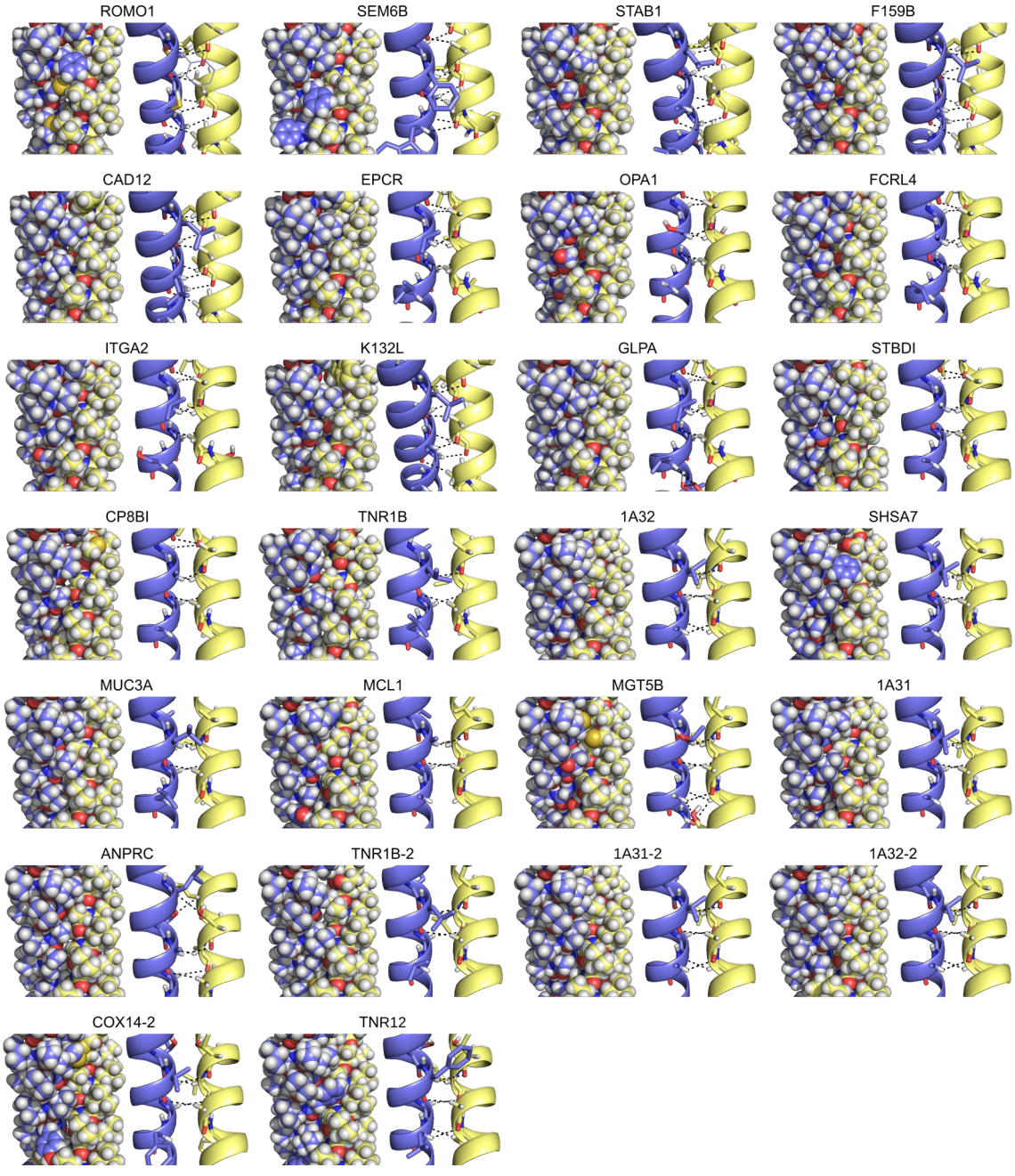


Figure S2.1 Structural models of the 26 final experimental constructs. The constructs are sorted left to right by CATM energy score. For each construct, a space filling representation is shown on the left to illustrate the packing, and a cartoon representation is shown on the right to illustrate the Cα–H hydrogen bond network.

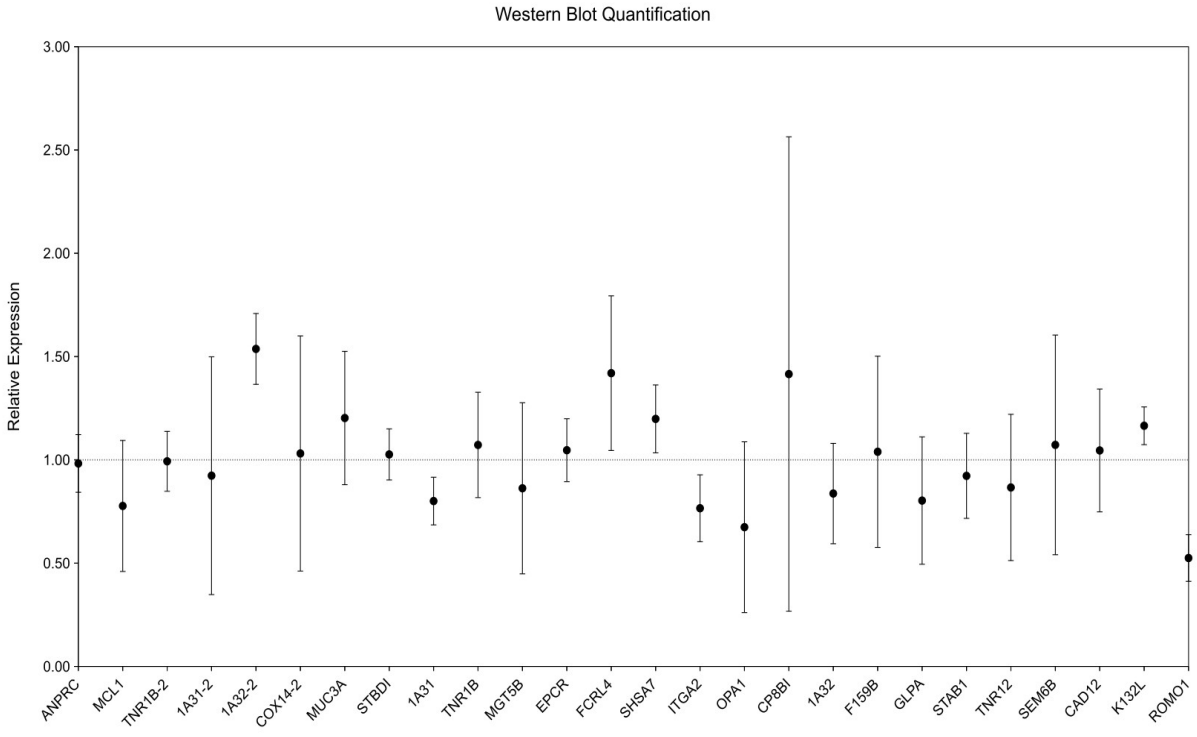


Figure S2.2 Quantification of expression of the TOXCAT construct using immunoblotting. Normalized expression. The individual construct expression range is 0.52-1.54 fold of the average expression (dashed line) with a standard deviation of the relative expression of the 26 samples of 0.22. The error bars represent the standard deviation of four biological replicates of each construct.

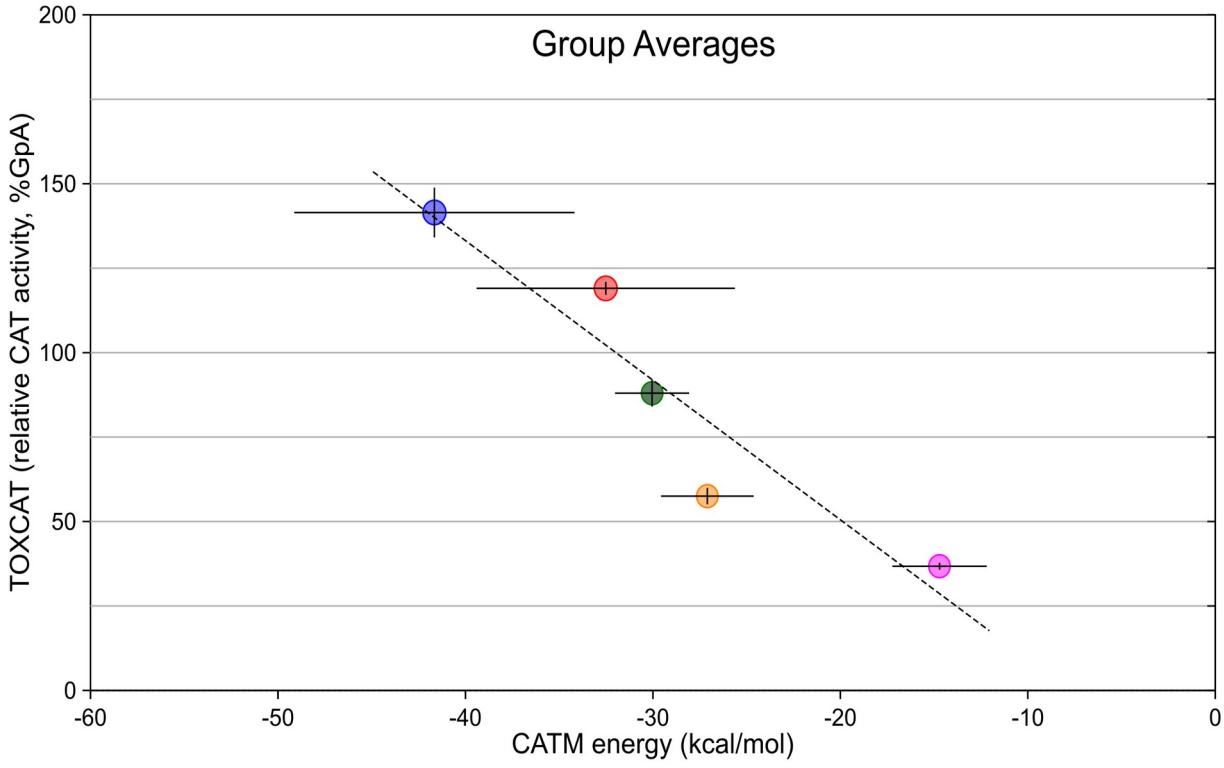


Figure S2.3 Group averages containing the TNR12 outlier. Same data as in Fig. 4a, with the inclusion the outlier point included in the >100% group. Each point corresponds to the group averaged of five bins based on CAT activity from weak (>25%, magenta) to very strong (>125%, blue), in 25% intervals. The error bars represent the standard error of the average. The dashed line is the linear regression of the data (R^2 = 0.883, p < 0.05).

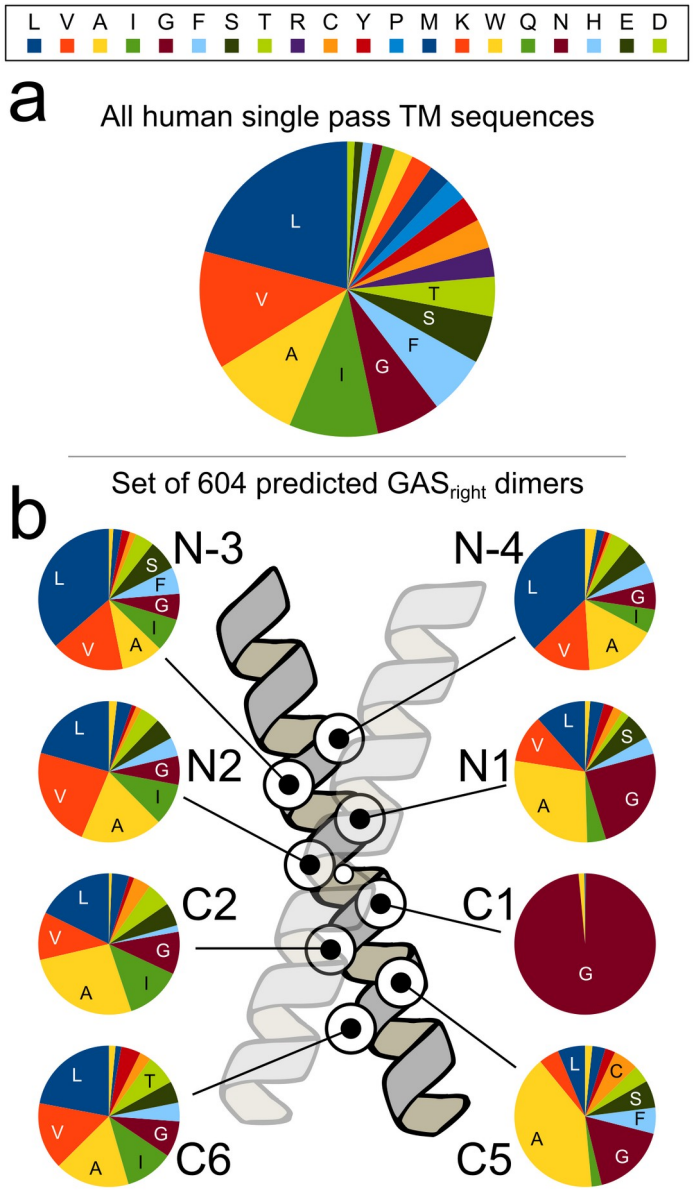


Figure S2.4 Composition of the interfacial positions of the 604 predicted GAS$_{right}$ dimers. a) For reference, overall composition of all single pass human TM domains. b) Composition of the interfacial position of the final 604 predicted GAS$_{right}$ sequences. These sequences exclude any sequence that contained polar amino acids or proline residues at the eight interfacial positons. The observed biases are consistent with those expected for GAS$_{right}$ motifs, with Gly almost invariably present at position C1, and small amino acids frequently present at positions N1 and C5.

Table S2.1 Final set of TOXCAT constructs and their C1 mutants

Uniprot AC[1]	Name	Wild-type Sequence	Poly-Leu Construct	TOXCAT[2]	CATM (kcal/mol)[3]
P60602	ROMO1[4]	VKM**GF**VM**GC**AV**GM**AA**GA**LFGTFSCLRIGM	RASLLL**GF**LL**GC**LL**GM**LL**GA**LILI	155%±28%	-56.3±0.1
P60602	ROMO1-G12I		RASLLL**GF**LL**GC**LL**IM**LL**GA**LILI	80%±42%	-24.3±0.1
A8MWY0	K132L	**WL**KV**GA**GV**GA**FT**AV**LLVALTCYFWKKN	RASLLL**WL**LL**GA**LL**GA**LL**AV**LILI	139%±19%	-31.8±0.3
A8MWY0	K132L_G12I		RASLLL**WL**LL**GA**LL**IA**LL**AV**LILI	22%±10%	N.M.
P55289	CAD12	**FL**AV**GL**ST**GA**LI**AI**LLCIVILLAIVVLYVALRRQ	RASLLL**FL**LL**GL**LL**GA**LL**AI**LILI	130%±12%	-36.9±0.5
P55289	CAD12_G12I		RASLLL**FL**LL**GL**LL**IA**LL**AI**LILI	34%±9%	N.M.
P02724	GLPA[4]	EAEIT**LI**IF**GV**MA**GV**IG**TI**LLISYGIRRL	RASLLL**LI**LL**GV**LL**GV**LL**TI**LILI	124%±46%	-31.3±0.3
P02724	GLPA_G12I		RASLLL**LI**LL**GV**LL**IV**LL**TI**LILI	32%±4%	N.M.
Q9H3T3	SEM6B	TSSVAAFVVG**AV**VS**GF**SV**GW**FV**GL**RER	RASLLL**AV**LL**GF**LL**GW**LL**GL**LILI	122%±10%	-44.5±0.3
Q9H3T3	SEM6B_G12I		RASLLL**AV**LL**GF**LL**IW**LL**GL**LILI	27%±5%	N.M.
Q9NY15	STAB1	VAAGVGA**VL**AA**GA**LL**GL**VA**GA**LYLRAR	RASLLL**VL**LL**GA**LL**GL**LL**GA**LILI	119%±18%	-40.6±0.0
Q9NY15	STAB1_G12I		RASLLL**VL**LL**GA**LL**IL**LL**GA**LILI	22%±4%	N.M.
Q9NP84	TNR12	LWAILGG**AL**SL**TF**VL**GL**LS**GF**LVWRRC	RASLLL**AL**LL**TF**LL**GL**LL**GF**LILI	119%±14%	-6.3±1.4
Q9NP84	TNR12_G12I		RASLLL**AL**LL**TF**LL**IL**LL**GF**LILI	29%±5%	N.M.
A6NKW6	F159B[4]	**SL**SI**GA**LI**GL**GI**AA**LVLLAFVISVCVL	RASLLL**SL**LL**GA**LL**GL**LL**AA**LILI	112%±20%	-39.9±0.0
A6NKW6	F159B_G12I		RASLLL**SL**LL**GA**LL**IL**LL**AA**LILI	32%±7%	-5.2±0.9
Q9UNU6	CP8B1	**MV**LW**GA**VL**GA**LL**VV**IAGYLCLAGM	RASLLL**MV**LL**GA**LL**GA**LL**VV**LILI	95%±15%	-28.7±0.3
Q9UNU6	CP8B1_G12I		RASLLL**MV**LL**GA**LL**IA**LL**VV**LILI	27%±6%	N.M.
P10314	1A32	IAIVGIIAGL**VL**FG**AM**FA**GA**VV**AA**VRWRRK	RASLLL**VL**LL**AM**LL**GA**LL**AA**LILI	94%±16%	-25.1±0.2
P10314	1A32_G12I		RASLLL**VL**LL**AM**LL**IA**LL**AA**LILI	20%±5%	N.M.
O60313	OPA1	ATRLLKLR**YL**IL**GS**AV**GG**GY**TA**K	RASLLL**YL**LL**GS**LL**GG**LL**TA**LILI	85%±20%	-34.0±0.1
O60313	OPA1_G12I		RASLLL**YL**LL**GS**LL**IG**LL**TA**LILI	21%±3%	-15.0±0.6
P17301	ITA2	**V**AT**GV**II**GS**II**AG**ILLLLALVAILWKLG	RASLLL**LV**LL**GV**LL**GS**LL**AG**LILI	78%±8%	-32.3±0.1
P17301	ITA2_G12I		RASLLL**LV**LL**GV**LL**IS**LL**AG**LILI	22%±4%	N.M.
A6NL88	SHSA7	STAYVVCGVI**SF**AL**AV**GV**GA**KV**AF**SKA	RASLLL**SF**LL**AV**LL**GA**LL**AF**LILI	68%±13%	-23.2±0.4
A6NL88	SHSA7_G12I		RASLLL**SF**LL**AV**LL**IA**LL**AF**LILI	17%±4%	N.M.
Q9UNN8	EPCR	YTSLVLGVLVGS**FI**IA**GV**AV**GI**FL**CT**G	RASLLL**FI**LL**GV**LL**GI**LL**CT**LILI	61%±11%	-36.7±0.2
Q9UNN8	EPCR_G12I		RASLLL**FI**LL**GV**LL**II**LL**CT**LILI	12%±8%	N.M.
Q96PJ5	FCRL4	DG**LV**AA**GA**TG**GL**LS**AL**LLAVALLFICW	RASLLL**LV**LL**GA**LL**GL**LL**AL**LILI	61%±12%	-33.0±0.1
Q96PJ5	FCRL4_G12I		RASLLL**LV**LL**GA**LL**IL**LL**AL**LILI	25%±5%	N.M.
Q3V5L5	MGT5B	FRLFVLGIGFFT**LC**FL**MT**SL**GG**QF**SA**R	RASLLL**LC**LL**MT**LL**GG**LL**SA**LILI	58%±12%	-19.9±0.4
Q3V5L5	MGT5B_G12I		RASLLL**LC**LL**MT**LL**IG**LL**SA**LILI	29%±6%	-10.8±0.1
P16189	1A31	IAIVGIIAGL**VL**FG**AV**FA**GA**VV**AA**VRWRRK	RASLLL**VL**LL**AV**LL**GA**LL**AA**LILI	52%±12%	-19.5±0.1
P16189	1A31_G12I		RASLLL**VL**LL**AV**LL**IA**LL**AA**LILI	25%±6%	N.M.
P20333	TNR1B	**TG**DF**AL**AV**GL**IV**GV**TALGLLIIGVVNCVIMTQVKKK	RASLLL**TG**LL**AL**LL**GL**LL**GV**LILI	51%±5%	-28.4±0.2
P20333	TNR1B G12I		RASLLL**TG**LL**AL**LL**IL**LL**GV**LILI	18%±8%	N.M.
O95210	STBD1	VWSA**LL**VG**GG**LA**GA**LF**VW**LLRGG	RASLLL**LL**LL**GG**LL**GA**LL**VW**LILI	51%±9%	-28.9±0.0
O95210	STBD1_G12I		RASLLL**LL**LL**GG**LL**IA**LL**VW**LILI	21%±4%	N.M.
Q02505	MUC3A	WR**AL**VG**GL**TA**GA**ALLVLLLLALGVRAV	RASLLL**LL**LL**AL**LL**GL**LL**GA**LILI	40%±10%	-22.8±1.1
Q02505	MUC3A_G12I		RASLLL**LL**LL**AL**LL**IL**LL**GA**LILI	16%±5%	N.M.
P10314	1A32-2	IAIVG**II**AG**LV**LF**GA**MFAGAVVAAVRWRRK	RASLLL**II**LL**LV**LL**GA**LL**FA**LILI	39%±4%	-8.2±0.4
P10314	1A32-2_G12I		RASLLL**II**LL**LV**LL**IA**LL**FA**LILI	4%±9%	N.M.
Q96I36	COX14-2	YKTFSTS**MM**LL**TV**YG**GY**LC**SV**RVYHYFQW	RASLLL**MM**LL**TV**LL**GY**LL**SC**LILI	39%±4%	-7.9±0.1
Q96I36	COX14-2_G12I		RASLLL**MM**LL**TV**LL**IY**LL**SC**LILI	28%±2%	N.M.
P16189	1A31-2	IAIVG**II**AG**LV**LF**GA**VF**AG**AVVAAVRWRRK	RASLLL**II**LL**LV**LL**GA**LL**AG**LILI	36%±10%	-10.7±0.0
P16189	1A31-2_G12I		RASLLL**II**LL**LV**LL**IA**LL**AG**LILI	19%±13%	N.M.
P20333	TNR1B-2	TGDFALAVGLIVGV**TA**LG**LL**II**GV**VN**CV**IMTQVKKK	RASLLL**TA**LL**LL**LL**GV**LL**CV**LILI	35%±12%	-11.3±0.1
P20333	TNR1B-2_G12I		RASLLL**TA**LL**LL**LL**IV**LL**CV**LILI	15%±2%	N.M.
Q07820	MCL1	EGGIRNVLL**AF**AG**VA**GV**GA**GL**AY**LIR	RASLLL**AF**LL**VA**LL**GA**LL**AY**LILI	35%±8%	-22.6±0.1
Q07820	MCL1-G12I		RASLLL**AF**LL**VA**LL**IA**LL**AY**LILI	19%±4%	N.M.
P17342	ANPRC	**L**EE**SA**VT**GI**VV**GA**LLGAGLLMAFYFFRKK	RASLLL**LL**LL**SA**LL**GI**LL**GA**LILI	34%±4%	-19.4±0.1
P17342	ANPRC-G12I		RASLLL**LL**LL**SA**LL**II**LL**GA**LILI	24%±1%	N.M.

1) Uniprot Accession
2) Relative CAT activity compared to GpA (average ± standard deviation)
3) CATM energy of the poly-Leu construct (average ± standard deviation). N.M. = no model predicted.
4) Constructs with C1 values >30% but at least a 75% reduction

Table S2.2 Constructs removed because of growth defect on maltose media

Uniprot AC[1]	Name	Wild-type Sequence	Poly-Leu Construct	TOXCAT[2]	CATM (kcal/mol)[3]
Q15116	PDCD1	**TL**VV**GV**VG**GL**LG**SL**VLLVWVLAVICSR	RASLLL**TL**LL**GV**LL**GL**LL**SL**LILI	215%±5%	-35.6±0.6
Q15116	PDCD1_G12I		RASLLL**TL**LL**GV**LL**IL**LL**SL**LILI	44%±10%	N.M.
Q96LC7	SIG10	**F**SN**GA**FL**GI**GI**TA**LLFLCLALIIMKIL	RASLLL**LF**LL**GA**LL**GI**LL**TA**LILI	214%±10%	-38.9±0.9
Q96LC7	SIG10_G12I		RASLLL**LF**LL**GA**LL**II**LL**TA**LILI	36%±2%	N.M.
Q15904	VAS1	DCAS**FF**SA**GI**WM**GL**LT**SL**FMLFIFTYG	RASLLL**FF**LL**GI**LL**GI**LL**SL**LILI	165%±19%	-44.5±0.8
Q15904	VAS1-G12I		RASLLL**FF**LL**GI**LL**II**LL**SL**LILI	20%±9%	N.M.
Q96D42	HAVCR1	TK**GI**YA**GV**CI**SV**LVLLALLGVIIAKKY	RASLLL**LL**LL**GI**LL**GV**LL**SV**LILI	160%±5%	-37.5±0.9
Q96D42	HAVCR1_G12I		RASLLL**LL**LL**GI**LL**IV**LL**SV**LILI	20%±5%	N.M.
Q6P7N7	TMM81	VAS**AL**GI**GI**AI**GV**VG**GV**LVRIVLCALR	RASLLL**AL**LL**GI**LL**GV**LL**GV**LILI	145%±6%	-44.8±0.7
Q6P7N7	TMM8I_G12I		RASLLL**AL**LL**GI**LL**IV**LL**GV**LILI	29%±2%	N.M.
Q8NCU8	YB039	ERTLQL**SV**LV**AF**AS**GV**LL**GW**QAN	RASLLL**SV**LL**AF**LL**GV**LL**GW**LILI	135%±45%	-23.2±0.5
Q8NCU8	YB039_G12I		RASLLL**SV**LL**AF**LL**IV**LL**GW**LILI	-2%±1%	N.M.
Q9Y6N1	COX11	KTTL**TY**VA**AV**AV**GM**LG**AS**YAAVALY	RASLLL**TY**LL**AV**LL**GM**LL**AS**LILI	128%±9%	-33.3±0.3
Q9Y6N1	COX11_G12I		RASLLL**TY**LL**AV**LL**IM**LL**AS**LILI	25%±5%	N.M.
P13591	NCAM1	TS**GL**ST**GA**IV**GI**LIVIFVLLLVVVDIT	RASLLL**LL**LL**GL**LL**GA**LL**GI**LILI	127%±1%	-32.8±0.6
P13591	NCAM1-G12I		RASLLL**LL**LL**GL**LL**IA**LL**GI**LILI	34%±2%	N.M.
O75354	ENTP6	SLR**VA**KV**AY**AL**GL**CV**GV**FIYVAYIKWH	RASLLL**VA**LL**AY**LL**GL**LL**GV**LILI	106%±7%	-25.3±0.1
O75354	ENTP6-G12I		RASLLL**VA**LL**AY**LL**IL**LL**GV**LILI	73%±2%	N.M.
Q9NVM1	EVA1B	ESFGL**YF**VL**GV**CF**GL**LL**TL**CLLVISIS	RASLLL**YF**LL**GV**LL**GL**LL**TL**LILI	105%±10%	-38.0±0.5
Q9NVM1	EVA1B_G12I		RASLLL**YF**LL**GV**LL**IL**LL**TL**LILI	26%±6%	N.M.
Q86X52	CHSS1	GRRAWL**SV**LL**GL**VL**GF**VL**AS**RLVLARA	RASLLL**SV**LL**GL**LL**GF**LL**AS**LILI	93%±7%	-35.8±0.1
Q86X52	CHSS1-G12I		RASLLL**SV**LL**GL**LL**IF**LL**AS**LILI	17%±5%	N.M.
Q9UKU0	ACSL6	FRSL**SA**TT**LV**SM**GA**LA**AI**LAYWFTHRA	RASLLL**SA**LL**LV**LL**GA**LL**AI**LILI	59%±17%	-13.3±0.1
Q9UKU0	ACSL6-G12I		RASLLL**SA**LL**LV**LL**IA**LL**AI**LILI	24%±4%	N.M.
Q93038	TNR25	WRQM**FW**VQ**VL**LA**GL**VV**AL**LLGATLTYT	RASLLL**FW**LL**VL**LL**GL**LL**AL**LILI	56%±9%	-17.5±0.3
Q93038	TNR25_G12I		RASLLL**FW**LL**VL**LL**IL**LL**AL**LILI	56%±8%	N.M.
P05026	AT1B1	WFKILLFY**VI**FY**GC**LA**GI**FI**GT**IQVMLLTISEFK	RASLLL**VI**LL**GC**LL**GI**LL**GT**LILI	39%±3%	-33.7±0.3
P05026	AT1B1_G12I		RASLLL**VI**LL**GC**LL**II**LL**GT**LILI	24%±8%	N.M.
Q8N6P7	I22R1	**T**WT**YS**FS**GA**FL**FS**MGFLVAVLCYLSYR	RASLLL**LT**LL**YS**LL**GA**LL**FS**LILI	31%±9%	-21.9±0.1
Q8N6P7	I22R1_G12I		RASLLL**LT**LL**YS**LL**IA**LL**FS**LILI	39%±11%	N.M.

1) Uniprot Accession

2) Relative CAT activity compared to GpA (average ± standard deviation)

3) CATM energy of the poly-Leu construct (average ± standard deviation). N.M. = no model predicted.

Table S2.3 Constructs removed because the TOXCAT signal of the WT sequence was <30%

Uniprot AC[1]	Name	Wild-type Sequence	Poly-Leu Construct	TOXCAT[2]	CATM (kcal/mol)[3]
Q6UXN7	TO201	LLR**LL**AA**AA**AC**GA**FA**FL**GYCIYLNRK	RASLLL**LL**LL**AA**LL**GA**LL**FL**LILI	28%±4%	-26.6±1.3
Q6UXN7	TO20L_G12I		RASLLL**LL**LL**AA**LL**IA**LL**FL**LILI	29%±2%	N.M.
Q8TEM1	PO210	SYQVMF**FT**LF**AL**LA**GT**AV**MI**IAYHTVC	RASLLL**FT**LL**AL**LL**GT**LL**MI**LILI	28%±5%	-25.7±0.2
Q8TEM1	PO210_G12I		RASLLL**FT**LL**AL**LL**IT**LL**MI**LILI	40%±11%	N.M.
Q9BRQ8	AIFM2	QVSVESGALHV**VI**VG**GG**FG**GI**AA**AS**QL	RASLLL**VI**LL**GG**LL**GI**LL**AS**LILI	26%±8%	-29.0±0.1
Q9BRQ8	AIFM2_G12I		RASLLL**VI**LL**GG**LL**II**LL**AS**LILI	22%±10%	N.M.
Q8IVU1	IGDCC3	TT**GI**VI**GI**HI**GV**TCIIFCVLFLLFGQR	RASLLL**LL**LL**GI**LL**GI**LL**GV**LILI	25%±6%	-41.1±0.5
Q8IVU1	IGDCC3_G12I		RASLLL**LL**LL**GI**LL**II**LL**GV**LILI	22%±4%	N.M.
O15197	EPHB6	RL**SL**VI**GS**IL**GA**LAFLLLAAITVLAVV	RASLLL**LL**LL**SL**LL**GS**LL**GA**LILI	25%±3%	-21.6±0.2
O15197	EPHB6_G12I		RASLLL**LL**LL**SL**LL**IS**LL**GA**LILI	18%±5%	N.M.
Q13586	STIM1	LKDFMLV**VS**IV**IG**VG**GC**WF**AY**IQNRYS	RASLLL**VS**LL**IG**LL**GC**LL**AY**LILI	24%±6%	-20.4±0.4
Q13586	STIM1_G12I		RASLLL**VS**LL**IG**LL**IC**LL**AY**LILI	27%±8%	-20.2±1.9
B6SEH8	ERVV1	KRA**LG**LI**LA**GM**GA**AI**GM**IAAWGGFTYH	RASLLL**LG**LL**LA**LL**GA**LL**GM**LILI	24%±8%	-19.5±0.2
B6SEH8	ERVV1_G12I		RASLLL**LG**LL**LA**LL**IA**LL**GM**LILI	22%±5%	-9.2±0.4
Q8N3T1	GLT15	HRACRLQ**FL**LL**LL**ML**GC**VL**MM**VAMLHAAH	RASLLL**FL**LL**LL**LL**GC**LL**MM**LILI	19%±1%	-18.1±0.4
Q8N3T1	GLT15_G12I		RASLLL**FL**LL**LL**LL**IC**LL**MM**LILI	37%±7%	N.M.
P05067	A4	**S**NK**GA**II**GL**MV**GG**VVIATVIVITLVMLKKK	RASLLL**LS**LL**GA**LL**GL**LL**GG**LILI	18%±4%	-35.2±0.1
P05067	A4_G12I		RASLLL**LS**LL**GA**LL**IL**LL**GG**LILI	16%±3%	N.M.
Q96I36	COX14	YKTFST**SM**ML**LT**VY**GG**YL**CS**VRVYHYFQW	RASLLL**SM**LL**LT**LL**GG**LL**CS**LILI	14%±2%	-21.1±0.1
Q96I36	COX14_G12I		RASLLL**SM**LL**LT**LL**IG**LL**CS**LILI	27%±7%	-5.1±0.7

[1]Uniprot Accession

[2]Relative CAT activity, compared to GpA (average ± standard deviation)

[3]CATM energy of the poly-Leu construct (average ± standard deviation). N.M. = no model predicted.

Table S2.4 Constructs discarded because of TOXCAT signal >30% for the $C1_{Gly \rightarrow Ile}$ variant

Uniprot AC[1]	Name	Wild-type Sequence	Poly-Leu Construct	TOXCAT[2]	CATM (kcal/mol)[3]
P43489	TNR4	GRAVAA**IL**GL**GL**VL**GL**LG**AL**AILLALY	RASLLL**IL**LL**GL**LL**GL**LL**AL**LILI	102%±11%	-39.7±0.1
P43489	TNR4_G12I		RASLLL**IL**LL**GL**LL**IL**LL**AL**LILI	109%±13%	N.M.
Q9UPZ6	THS7A	K**TW**VY**GV**AA**GA**FVLLIFIVSMIYLACK	RASLLL**LL**LL**TW**LL**GV**LL**GA**LILI	87%±37%	-3.2±0.1
Q9UPZ6	THS7A_G12I		RASLLL**LL**LL**TW**LL**IV**LL**GA**LILI	76%±9%	N.M.
Q7L8C5	STY13	SVA**VI**AL**GA**TL**GT**AT**SI**LALCGVTCLCRH	RASLLL**VI**LL**GA**LL**GT**LL**SI**LILI	79%±5%	-31.0±0.1
Q7L8C5	STY13_G12I		RASLLL**VI**LL**GA**LL**IT**LL**SI**LILI	45%±6%	N.M.
Q08ET2	SIG14	LVL**TL**IR**GA**LM**GA**GF**LL**TYGLTWIYYTRC	RASLLL**TL**LL**GA**LL**GA**LL**LL**LILI	77%±7%	-26.3±0.1
Q08ET2	SIG14_G12I		RASLLL**TL**LL**GA**LL**IA**LL**LL**LILI	39%±14%	N.M.
Q8TDF5	NETO1	S**GT**VI**GV**TS**CI**VIILIIISVIVQIKQA	RASLLL**LL**LL**GT**LL**GV**LL**CI**LILI	72%±5%	-32.9±0.3
Q8TDF5	NETO1_G12I		RASLLL**LL**LL**GT**LL**IV**LL**CI**LILI	38%±%	N.M.
Q99795	GPA33	**M**NV**AL**YV**GI**AV**GV**VAALIIIGIIIY?C	RASLLL**LM**LL**AL**LL**GI**LL**GV**LILI	68%±17%	-23.3±0.0
Q99795	GPA33_G12I		RASLLL**LM**LL**AL**LL**II**LL**GV**LILI	38%±7%	N.M.
Q9BY71	LRRC3	TTD**VA**ML**VT**MF**GW**FA**MV**IAYVVYYVRH	RASLLL**VA**LL**VT**LL**GW**LL**MV**LILI	68%±10%	-8.3±0.4
Q9BY71	LRRC3_G12I		RASLLL**VA**LL**VT**LL**IW**LL**MV**LILI	57%±12%	N.M.
Q30201	HFE	**TL**VI**GV**IS**GI**AV**FV**VILFIGILFIILRKRQ	RASLLL**TL**LL**GV**LL**GI**LL**FV**LILI	64%±9%	-23.3±0.2
Q30201	HFE-G12I		RASLLL**TL**LL**GV**LL**II**LL**FV**LILI	53%±5%	N.M.
Q5SSG8	MUC21	AWEIFLITLVSV**VA**AV**GL**FA**GL**FF**CV**R	RASLLL**VA**LL**GL**LL**GL**LL**CV**LILI	60%±6%	-32.3±0.4
Q5SSG8	MUC21_G12I		RASLLL**VA**LL**GL**LL**IL**LL**CV**LILI	43%±%	N.M.
Q9P2S2	NRXN2	GMVVGIVAAAALCILILLYAM	RASLLL**LS**LL**GM**LL**GI**LL**AA**LILI	54%±2%	-33.7±0.2
Q9P2S2	NRXN2_G12I		RASLLL**LS**LL**GM**LL**II**LL**AA**LILI	39%±8%	N.M.
Q2M385	MPEG1	SG**GA**AA**GV**TV**GV**TTILAVVITLAIYGT	RASLLL**LL**LL**GA**LL**GV**LL**GV**LILI	34%±6%	-33.9±0.9
Q2M385	MPEG1_G12I		RASLLL**LL**LL**GA**LL**IV**LL**GV**LILI	33%±1%	N.M.
Q8N967	LRTM2	MGT**VI**IA**GV**VC**GV**VC**IM**MVVAAAYGCI	RASLLL**VI**LL**GV**LL**GV**LL**IM**LILI	32%±5%	-17.1±0.3
Q8N967	LRTM2_G12I		RASLLL**VI**LL**GV**LL**IV**LL**IM**LILI	31%±13%	N.M.
Q9H3N1	TMX1	WGSYTV**FA**LA**TL**FS**GL**LL**GL**CMIFVADCL	RASLLL**FA**LL**TL**LL**GL**LL**GL**LILI	31%±3%	-9.0±0.4
Q9H3N1	TMX1_G12I		RASLLL**FA**LL**TL**LL**IL**LL**GL**LILI	32%±5%	N.M.
Q9H6B4	CLMP	**M**VA**GA**VT**GI**VA**GA**LLIFLLVWLLIRRK	RASLLL**LM**LL**GA**LL**GI**LL**GA**LILI	30%±7%	-33.8±0.1
Q9H6B4	CLMP_G12I		RASLLL**LM**LL**GA**LL**II**LL**GA**LILI	33%±4%	N.M.

[1]Uniprot Accession

[2]Relative CAT activity, compared to GpA (average ± standard deviation)

[3]CATM energy of the poly-Leu construct (average ± standard deviation). N.M. = no model predicted.

Table S2.5 Progression of number of constructs from the computational analysis to the final set of experimental constructs

Human Single Pass Proteins in Uniprot	2,383
Wild-type with energy score below 0 kcal/mol	1,141
Poly-Leu sequences with energy score below 0 kcal/mol	1,020
Subset of sequences with non-polar interface	668
Poly-Leu sequences with energy score below -5 kcal/mol	604
Sequences selected for experimental analysis	65
Sequences that passed the maltose growth test (-15)	50
Sequences with TOXCAT >30% (-10)	40
Sequences with C1 mutation <30% and final set (-14)	26

Table S2.6 Uniprot Accession and description of the 26 final constructs applied for the analysis

Uniprot Accession	Uniprot Entry Name	Protein Name
A6NKW6	F159B_HUMAN	Membrane protein FAM159B
A6NL88	SHSA7_HUMAN	Protein shisa-7
A8MWY0	K132L_HUMAN	UPF0577 protein KIAA1324-like
O60313	OPA1_HUMAN	Dynamin-like 120 kDa protein, mitochondrial
O95210	STBD1_HUMAN	Starch-binding domain-containing protein 1
P02724	GLPA_HUMAN	Glycophorin-A
P10314	1A32_HUMAN	HLA class I histocompatibility antigen, A-32 alpha chain
P16189	1A31_HUMAN	HLA class I histocompatibility antigen, A-31 alpha chain
P17301	ITA2_HUMAN	Integrin alpha-2
P17342	ANPRC_HUMAN	Atrial natriuretic peptide receptor 3
P20333	TNR1B_HUMAN	Tumor necrosis factor receptor superfamily member 1B
P55289	CAD12_HUMAN	Cadherin-12
P60602	ROMO1_HUMAN	Reactive oxygen species modulator 1
Q02505	MUC3A_HUMAN	Mucin-3A
Q07820	MCL1_HUMAN	Induced myeloid leukemia cell differentiation protein Mcl-1
Q3V5L5	MGT5B_HUMAN	Alpha-1,6-mannosylglycoprotein 6-beta-N-acetylglucosaminyltransferase B
Q96I36	COX14_HUMAN	Cytochrome c oxidase assembly protein COX14
Q96PJ5	FCRL4_HUMAN	Fc receptor-like protein 4
Q9H3T3	SEM6B_HUMAN	Semaphorin-6B
Q9NP84	TNR12_HUMAN	Tumor necrosis factor receptor superfamily member 12A
Q9NY15	STAB1_HUMAN	Stabilin-1
Q9UNN8	EPCR_HUMAN	Endothelial protein C receptor
Q9UNU6	CP8B1_HUMAN	7-alpha-hydroxycholest-4-en-3-one 12-alpha-hydroxylase

Table S2.7 Energetic and geometric properties of groups of CATM scores.

CATM Score Range[1]	-5 to -15	-15 to-25	-25 to -35	-35 to -45	-45 and below
Number of models	210	254	89	45	6
CATM energy score (kcal/mol)	-10.1±3.0	-19.5±2.7	-29.4±3.1	-39.0±2.9	-47.6±4.0
Van der Waals (kcal/mol)	-24.2±4.2	-29.9±3.7	-35.2±3.2	-39.7±3.7	-39.9±5.3
Cα-H hydrogen bonding (kcal/mol)	-3.4±4.9	-4.9±3.9	-7.6±3.2	-10.0±2.9	-13.2±0.8
Solvation (kcal/mol)	18.2±2.1	16.1±2.5	14.1±2.1	11.1±2.8	5.5±1.1
Crossing angle (°)	-52.3±8.1	-49.2±8.1	-47.3±4.4	-40.4±5.5	-31.6±2.1
Number of Cα-H bonds	4.7±1.0	4.3±0.9	5.6±1.3	7.3±1.1	8.0±0.0
Interface surface area (Å^2)	4730±910	4790±910	4730±710	4400±580	4070±490
Inter-helical distance (Å)	7.1±0.2	7.0±0.2	6.6±0.3	6.4±0.1	6.4±0.1
Van der Waals/Interface surface area (kcal/(mol Å^2))	-0.0051± 0.0013	-0.0063± 0.0014	-0.0076± 0.0013	-0.0091± 0.0010	-0.0098± 0.0006
Sequences with GxxxG	17%	28%	75%	100%	100%
Sequences with Sm-xxx-Sm	93%	96%	100%	100%	100%
Sequences with Gly at N1	2%	13%	67%	100%	100%
Sequences with Gly at C1	99%	97%	100%	98%	100%
Sequences with Gly at C5	17%	14%	16%	36%	50%

[1]All values are reported as averages ± standard deviation.

[2]Sm-xxx-Sm are defined by any combinations of Gly, Ala, Ser and Cys at the first and last position.

2.7 Acknowledgments

The work was supported by National Science Foundation Grants CHE-1415910 and CHE-1710182. B.K.M. and S.M.A. acknowledge the support ofthe NLM training grant 5T15LM007359 to the CIBM Training Program. S.M.A. also acknowledges the support of the Dr. James Chieh-Hsia Mai Wisconsin Distinguished Graduate Fellowship. B.K.M. also acknowledges support by a fellowship in Informatics from the PhRMA Foundation. E.J.L. acknowledges the support of a Hilldale Undergraduate Research Fellowship. We are grateful to Dr. Sabareesh Subramaniam for contributions to the development of CATM, to Samson Condon and Claire Armstrong for helpful suggestions and discussion, and to Elizabeth Caselle for critical reading of the manuscript.

2.8 References

Adams, P.D., Arkin, I.T., Engelman, D.M., and Brünger, A.T. (1995). Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. Nat. Struct. Biol. *2*, 154–162.

Anbazhagan, V., Munz, C., Tome, L., and Schneider, D. (2010). Fluidizing the membrane by a local anesthetic: phenylethanol affects membrane protein oligomerization. J. Mol. Biol. *404*, 773–777.

Arbely, E., and Arkin, I.T. (2004). Experimental measurement of the strength of a C alpha-H...O bond in a lipid bilayer. J. Am. Chem. Soc. *126*, 5362–5363.

Arkin, I.T., and Brunger, A.T. (1998). Statistical analysis of predicted transmembrane alpha-helices. Biochim. Biophys. Acta *1429*, 113–128.

Bocharov, E.V., Pustovalova, Y.E., Pavlov, K.V., Volynsky, P.E., Goncharuk, M.V., Ermolyuk, Y.S., Karpunin, D.V., Schulga, A.A., Kirpichnikov, M.P., Efremov, R.G., et al. (2007). Unique dimeric structure of BNip3 transmembrane domain suggests membrane permeabilization as a cell death trigger. J. Biol. Chem. *282*, 16256–16266.

Bocharov, E.V., Mineev, K.S., Volynsky, P.E., Ermolyuk, Y.S., Tkach, E.N., Sobol, A.G., Chupin, V.V., Kirpichnikov, M.P., Efremov, R.G., and Arseniev, A.S. (2008a). Spatial structure of the dimeric transmembrane domain of the growth factor receptor ErbB2 presumably corresponding to the receptor active state. J. Biol. Chem. *283*, 6950–6956.

Bocharov, E.V., Mayzel, M.L., Volynsky, P.E., Goncharuk, M.V., Ermolyuk, Y.S., Schulga, A.A., Artemenko, E.O., Efremov, R.G., and Arseniev, A.S. (2008b). Spatial structure and pH-dependent conformational diversity of dimeric transmembrane domain of the receptor tyrosine kinase EphA1. J. Biol. Chem. *283*, 29385–29395.

Bocharov, E.V., Mineev, K.S., Goncharuk, M.V., and Arseniev, A.S. (2012). Structural and thermodynamic insight into the process of “weak” dimerization of the ErbB4 transmembrane domain by solution NMR. Biochim. Biophys. Acta *1818*, 2158–2170.

Bowie, J.U. (2011). Membrane protein folding: how important are hydrogen bonds? Curr. Opin. Struct. Biol. *21*, 42–49.

Bragin, P.E., Mineev, K.S., Bocharova, O.V., Volynsky, P.E., Bocharov, E.V., and Arseniev, A.S. (2016). HER2 Transmembrane Domain Dimerization Coupled with Self-Association of Membrane-Embedded Cytoplasmic Juxtamembrane Regions. J. Mol. Biol. *428*, 52–61.

Brosig, B., and Langosch, D. (1998). The dimerization motif of the glycophorin A transmembrane segment in membranes: importance of glycine residues. Protein Sci. Publ. Protein Soc. *7*, 1052–1056.

Choma, C., Gratkowski, H., Lear, J.D., and DeGrado, W.F. (2000). Asparagine-mediated self-association of a model transmembrane helix. Nat. Struct. Biol. *7*, 161–166.

Chung, I., Akita, R., Vandlen, R., Toomre, D., Schlessinger, J., and Mellman, I. (2010). Spatial control of EGF receptor activation by reversible dimerization on living cells. Nature *464*, 783–787.

Dawson, J.P., Melnyk, R.A., Deber, C.M., and Engelman, D.M. (2003). Sequence context strongly modulates association of polar residues in transmembrane helices. J. Mol. Biol. *331*, 255–262.

Dixon, A.M., Stanley, B.J., Matthews, E.E., Dawson, J.P., and Engelman, D.M. (2006). Invariant chain transmembrane domain trimerization: a step in MHC class II assembly. Biochemistry *45*, 5228–5234.

Dixon, A.M., Drake, L., Hughes, K.T., Sargent, E., Hunt, D., Harton, J.A., and Drake, J.R. (2014). Differential transmembrane domain GXXXG motif pairing impacts major histocompatibility complex (MHC) class II structure. J. Biol. Chem. *289*, 11695–11703.

Doura, A.K., and Fleming, K.G. (2004). Complex interactions at the helix-helix interface stabilize the glycophorin A transmembrane dimer. J. Mol. Biol. *343*, 1487–1497.

Duong, M.T., Jaszewski, T.M., Fleming, K.G., and MacKenzie, K.R. (2007). Changes in apparent free energy of helix-helix dimerization in a biological membrane due to point mutations. J. Mol. Biol. *371*, 422–434.

Elazar, A., Weinstein, J., Biran, I., Fridman, Y., Bibi, E., and Fleishman, S.J. (2016). Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. ELife *5*.

Endres, N.F., Das, R., Smith, A.W., Arkhipov, A., Kovacs, E., Huang, Y., Pelton, J.G., Shan, Y., Shaw, D.E., Wemmer, D.E., et al. (2013). Conformational coupling across the plasma membrane in activation of the EGF receptor. Cell *152*, 543–556.

Engelman, D.M., Adair, B.D., Brünger, A., Flanagan, J.M., Hunt, J.F., Lemmon, M.A., Treutlein, H., and Zhang, J. (1993). Dimerization of glycophorin A transmembrane helices: mutagenesis and modeling. Soc. Gen. Physiol. Ser. *48*, 11–21.

Fleming, K.G., Ren, C.-C., Doura, A.K., Eisley, M.E., Kobus, F.J., and Stanley, A.M. (2004). Thermodynamics of glycophorin A transmembrane helix dimerization in C14 betaine micelles. Biophys. Chem. *108*, 43–49.

Gratkowski, H., Lear, J.D., and DeGrado, W.F. (2001). Polar side chains drive the association of model transmembrane peptides. Proc. Natl. Acad. Sci. U. S. A. *98*, 880–885.

Herrmann, J.R., Fuchs, A., Panitz, J.C., Eckert, T., Unterreitmeier, S., Frishman, D., and Langosch, D. (2010). Ionic interactions promote transmembrane helix-helix association depending on sequence context. J. Mol. Biol. *396*, 452–461.

Hessa, T., Meindl-Beinker, N.M., Bernsel, A., Kim, H., Sato, Y., Lerch-Bader, M., Nilsson, I., White, S.H., and von Heijne, G. (2007). Molecular code for transmembrane-helix recognition by the Sec61 translocon. Nature *450*, 1026–1030.

Hong, H., Chang, Y.-C., and Bowie, J.U. (2013). Measuring transmembrane helix interaction strengths in lipid bilayers using steric trapping. Methods Mol. Biol. Clifton NJ *1063*, 37–56.

Howard, K.P., Lear, J.D., and DeGrado, W.F. (2002). Sequence determinants of the energetics of folding of a transmembrane four-helix-bundle protein. Proc. Natl. Acad. Sci. U. S. A. *99*, 8568–8572.

Hubert, P., Sawma, P., Duneau, J.-P., Khao, J., Hénin, J., Bagnard, D., and Sturgis, J. (2010). Single-spanning transmembrane domains in cell growth and cell-cell interactions: More than meets the eye? Cell Adhes. Migr. *4*, 313–324.

Johnson, R.M., Rath, A., and Deber, C.M. (2006). The position of the Gly-xxx-Gly motif in transmembrane segments modulates dimer affinity. Biochem. Cell Biol. Biochim. Biol. Cell. *84*, 1006–1012.

Khadria, A.S., Mueller, B.K., Stefely, J.A., Tan, C.H., Pagliarini, D.J., and Senes, A. (2014). A Gly-zipper motif mediates homodimerization of the transmembrane domain of the mitochondrial kinase ADCK3. J. Am. Chem. Soc. *136*, 14068–14077.

Kirrbach, J., Krugliak, M., Ried, C.L., Pagel, P., Arkin, I.T., and Langosch, D. (2013). Self-interaction of transmembrane helices representing pre-clusters from the human single-span membrane proteins. Bioinforma. Oxf. Engl. *29*, 1623–1630.

Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778–795.

Kulp, D.W., Subramaniam, S., Donald, J.E., Hannigan, B.T., Mueller, B.K., Grigoryan, G., and Senes, A. (2012). Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL). J. Comput. Chem. *33*, 1645–1661.

Lai, M.-D., and Xu, J. (2007). Ribosomal Proteins and Colorectal Cancer. Curr. Genomics *8*, 43–49.

LaPointe, L.M., Taylor, K.C., Subramaniam, S., Khadria, A., Rayment, I., and Senes, A. (2013). Structural organization of FtsB, a transmembrane protein of the bacterial divisome. Biochemistry *52*, 2574–2585.

Lau, T.-L., Kim, C., Ginsberg, M.H., and Ulmer, T.S. (2009). The structure of the integrin alphaIIbbeta3 transmembrane complex explains integrin transmembrane signalling. EMBO J. *28*, 1351–1361.

Lawrie, C.M., Sulistijo, E.S., and MacKenzie, K.R. (2010). Intermonomer hydrogen bonds enhance GxxxG-driven dimerization of the BNIP3 transmembrane domain: roles for sequence context in helix-helix association in membranes. J. Mol. Biol. *396*, 924–936.

Lazaridis, T. (2003). Effective energy function for proteins in lipid membranes. Proteins *52*, 176–192.

Li, E., Wimley, W.C., and Hristova, K. (2012). Transmembrane helix dimerization: beyond the search for sequence motifs. Biochim. Biophys. Acta *1818*, 183–193.

Li, R., Mitra, N., Gratkowski, H., Vilaire, G., Litvinov, R., Nagasami, C., Weisel, J.W., Lear, J.D., DeGrado, W.F., and Bennett, J.S. (2003). Activation of integrin alphaIIbbeta3 by modulation of transmembrane helix associations. Science *300*, 795–798.

Li, R., Gorelik, R., Nanda, V., Law, P.B., Lear, J.D., DeGrado, W.F., and Bennett, J.S. (2004). Dimerization of the transmembrane domain of Integrin alphaIIb subunit in cell membranes. J. Biol. Chem. *279*, 26666–26673.

Li, W., Metcalf, D.G., Gorelik, R., Li, R., Mitra, N., Nanda, V., Law, P.B., Lear, J.D., Degrado, W.F., and Bennett, J.S. (2005). A push-pull mechanism for regulating integrin function. Proc. Natl. Acad. Sci. U. S. A. *102*, 1424–1429.

MacKenzie, K.R., and Fleming, K.G. (2008). Association energetics of membrane spanning alpha-helices. Curr. Opin. Struct. Biol. *18*, 412–419.

MacKenzie, K.R., Prestegard, J.H., and Engelman, D.M. (1997). A transmembrane helix dimer: structure and implications. Science *276*, 131–133.

MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B *102*, 3586–3616.

Matthews, E.E., Thévenin, D., Rogers, J.M., Gotow, L., Lira, P.D., Reiter, L.A., Brissette, W.H., and Engelman, D.M. (2011). Thrombopoietin receptor activation: transmembrane helix dimerization, rotation, and allosteric modulation. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *25*, 2234–2244.

Mendrola, J.M., Berger, M.B., King, M.C., and Lemmon, M.A. (2002). The single transmembrane domains of ErbB receptors self-associate in cell membranes. J. Biol. Chem. *277*, 4704–4712.

Mineev, K.S., Bocharov, E.V., Pustovalova, Y.E., Bocharova, O.V., Chupin, V.V., and Arseniev, A.S. (2010). Spatial structure of the transmembrane domain heterodimer of ErbB1 and ErbB2 receptor tyrosine kinases. J. Mol. Biol. *400*, 231–243.

Mueller, B.K., Subramaniam, S., and Senes, A. (2014). A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα-H hydrogen bonds. Proc. Natl. Acad. Sci. U. S. A. *111*, E888-895.

Park, H., Yoon, J., and Seok, C. (2008). Strength of Calpha-H...O=C hydrogen bonds in transmembrane proteins. J. Phys. Chem. B *112*, 1041–1048.

Ridder, A., Skupjen, P., Unterreitmeier, S., and Langosch, D. (2005). Tryptophan supports interaction of transmembrane helices. J. Mol. Biol. *354*, 894–902.

Ruan, W., Lindner, E., and Langosch, D. (2004). The interface of a membrane-spanning leucine zipper mapped by asparagine-scanning mutagenesis. Protein Sci. Publ. Protein Soc. *13*, 555–559.

Russ, W.P., and Engelman, D.M. (1999). TOXCAT: a measure of transmembrane helix association in a biological membrane. Proc. Natl. Acad. Sci. U. S. A. *96*, 863–868.

Russ, W.P., and Engelman, D.M. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. J. Mol. Biol. *296*, 911–919.

Schanzenbach, C., Schmidt, F.C., Breckner, P., Teese, M.G., and Langosch, D. (2017). Identifying ionic interactions within a membrane using BLaTM, a genetic tool to measure homo- and heterotypic transmembrane helix-helix interactions. Sci. Rep. *7*, 43476.

Scheiner, S., Kar, T., and Gu, Y. (2001). Strength of the Calpha H..O hydrogen bond of amino acid residues. J. Biol. Chem. *276*, 9832–9837.

Senes, A., Gerstein, M., and Engelman, D.M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. J. Mol. Biol. *296*, 921–936.

Senes, A., Ubarretxena-Belandia, I., and Engelman, D.M. (2001). The Calpha ---H...O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. Proc. Natl. Acad. Sci. U. S. A. *98*, 9056–9061.

Senes, A., Engel, D.E., and DeGrado, W.F. (2004). Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. Curr. Opin. Struct. Biol. *14*, 465–479.

Shaw, W.V. (1975). Chloramphenicol acetyltransferase from chloramphenicol-resistant bacteria. Methods Enzymol. *43*, 737–755.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. Am. J. Psychol. *15*, 72–101.

Stouffer, A.L., Nanda, V., Lear, J.D., and DeGrado, W.F. (2005). Sequence determinants of a transmembrane proton channel: an inverse relationship between stability and function. J. Mol. Biol. *347*, 169–179.

Subramaniam, S., and Senes, A. (2012). An energy-based conformer library for side chain optimization: improved prediction and adjustable sampling. Proteins *80*, 2218–2234.

Sulistijo, E.S., and MacKenzie, K.R. (2006). Sequence dependence of BNIP3 transmembrane domain dimerization implicates side-chain hydrogen bonding and a tandem GxxxG motif in specific helix-helix interactions. J. Mol. Biol. *364*, 974–990.

Tate, R.F. (1954). Correlation Between a Discrete and a Continuous Variable. Point-Biserial Correlation. Ann. Math. Stat. *25*, 603–607.

Teese, M.G., and Langosch, D. (2015). Role of GxxxG Motifs in Transmembrane Domain Interactions. Biochemistry *54*, 5125–5135.

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *45*, D158–D169.

Unterreitmeier, S., Fuchs, A., Schäffler, T., Heym, R.G., Frishman, D., and Langosch, D. (2007). Phenylalanine promotes interaction of transmembrane domains via GxxxG motifs. J. Mol. Biol. *374*, 705–718.

Vargas, R., Garza, J., Dixon, D.A., and Hay, B.P. (2000). How Strong Is the Cα−H···OC Hydrogen Bond? J. Am. Chem. Soc. *122*, 4750–4755.

Vilar, M., Charalampopoulos, I., Kenchappa, R.S., Simi, A., Karaca, E., Reversi, A., Choi, S., Bothwell, M., Mingarro, I., Friedman, W.J., et al. (2009). Activation of the p75 neurotrophin receptor through conformational rearrangement of disulphide-linked receptor dimers. Neuron *62*, 72–83.

Wallin, E., and von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein Sci. Publ. Protein Soc. *7*, 1029–1038.

Walters, R.F.S., and DeGrado, W.F. (2006). Helix-packing motifs in membrane proteins. Proc. Natl. Acad. Sci. U. S. A. *103*, 13658–13663.

Wei, P., Liu, X., Hu, M.-H., Zuo, L.-M., Kai, M., Wang, R., and Luo, S.-Z. (2011). The dimerization interface of the glycoprotein Ibβ transmembrane domain corresponds to polar residues within a leucine zipper motif. Protein Sci. Publ. Protein Soc. *20*, 1814–1823.

Wei, P., Zheng, B.-K., Guo, P.-R., Kawakami, T., and Luo, S.-Z. (2013). The association of polar residues in the DAP12 homodimer: TOXCAT and molecular dynamics simulation studies. Biophys. J. *104*, 1435–1444.

Yin, H., Litvinov, R.I., Vilaire, G., Zhu, H., Li, W., Caputo, G.A., Moore, D.T., Lear, J.D., Weisel, J.W., Degrado, W.F., et al. (2006). Activation of platelet alphaIIbbeta3 by an exogenous peptide corresponding to the transmembrane domain of alphaIIb. J. Biol. Chem. *281*, 36732–36741.

Yohannan, S., Faham, S., Yang, D., Grosfeld, D., Chamberlain, A.K., and Bowie, J.U. (2004a). A C alpha-H...O hydrogen bond in a membrane protein is not stabilizing. J. Am. Chem. Soc. *126*, 2284–2285.

Yohannan, S., Yang, D., Faham, S., Boulting, G., Whitelegge, J., and Bowie, J.U. (2004b). Proline substitutions are not easily accommodated in a membrane protein. J. Mol. Biol. *341*, 1–6.

Zhang, S.-Q., Kulp, D.W., Schramm, C.A., Mravic, M., Samish, I., and DeGrado, W.F. (2015). The Membrane- and Soluble-Protein Helix-Helix Interactome: Similar Geometry via Different Interactions. Struct. Lond. Engl. 1993 *23*, 527–541.

Zhang, Y., Kulp, D.W., Lear, J.D., and DeGrado, W.F. (2009). Experimental and computational evaluation of forces directing the association of transmembrane helices. J. Am. Chem. Soc. *131*, 11341–11343.

Zhou, F.X., Cocco, M.J., Russ, W.P., Brunger, A.T., and Engelman, D.M. (2000). Interhelical hydrogen bonding drives strong interactions in membrane proteins. Nat. Struct. Biol. *7*, 154–160.

Zhou, F.X., Merianos, H.J., Brunger, A.T., and Engelman, D.M. (2001). Polar residues drive association of polyleucine transmembrane helices. Proc. Natl. Acad. Sci. U. S. A. *98*, 2250–2255.

Zhu, J., Luo, B.-H., Barth, P., Schonbrun, J., Baker, D., and Springer, T.A. (2009). The structure of a receptor with two associating transmembrane domains on the cell surface: integrin alphaIIbbeta3. Mol. Cell *34*, 234–249.

Chapter 3: Development of sort-seq for helix-helix association

3.1 Abstract

Transmembrane helix-helix association is critical to the function of many receptors, enzymes, transporters, and channels. Having structural information about these complexes greatly increases the ability for researchers to study them biologically. The methods used to obtain atomic level structures, however, are time- and cost-intensive. Another avenue to infer structure-level information is to perform mutagenesis of a transmembrane domain to identify the residues critical for association at the interface. Here, I propose a method that combines fluorescence-activated cell-sorting (FACS) with next-generation sequencing and the TOXGREEN assay. This sort-seq derivative can evaluate high-throughput mutagenesis data for the self-association of transmembrane domains. Specifically, I evaluate the dimerization propensity of point mutants for human transmembrane domains and identify new dimers and their interfaces.

3.2 Introduction

The association of membrane proteins (MP) is critical for many biological processes and misregulation of association leads to a variety of diseases that include cancer and Alzheimer's (Hubert et al., 2010; Roskoski, 2014). Association is particularly important for single-pass MPs (SPMP), which are proteins that span a lipid bilayer with one alpha helix and often have globular, water-soluble domains on either side. Though it was initially believed that the transmembrane (TM) domains of MPs simply anchored the important globular domains to the cell surface, it has become well known that TM domains (TMD) often drive the oligomerization of the entire protein sometimes leading to protein activation (Bocharov et al., 2017). Understanding TMD association will lend insight into how biological processes work and what drives disease phenotypes.

The methods that are used to obtain association information range from computational predictions to experimental measurements. Researchers have used quantitative methods like nuclear magnetic resonance (NMR; MacKenzie et al., 1997), sedimentation equilibrium analytical ultracentrifugation (SE-AUC; Doura and Fleming, 2004; Fleming et al., 1997), Förster resonance energy transfer (FRET; Adair and Engelman, 1994; Fisher et al., 2003), and molecular dynamics (MD; Hénin et al., 2005; Zhang and Lazaridis, 2006) to measure stoichiometry, stability, and structure of SPMP oligomers. These methods, however, are time- and cost-intensive and challenging. For example, efforts to obtain high resolution structures of SPMPs have unfortunately only produced 23 unique protein structures (Berman et al., 2000).

To obtain more structural information, researchers can identify an interaction interface to structurally model protein-protein interactions. Single-pass TMDs can interact in a relatively limited number of ways because they are constrained by the lipid bilayer. Finding an interface narrows the possible 3D structure space tremendously. Knowing the interface of a TM domain therefore gives insight into the overall structure. One way to identify the interface of a protein-

protein interaction, without solving the structure, is to perform mutagenesis across the protein (Elazar et al., 2016; Khadria et al., 2014; Langosch et al., 1996; LaPointe et al., 2013; Russ and Engelman, 1999; Sulistijo et al., 2003). The mutations that affect association are likely to be at the interface and the mutations that do not affect association are likely on the outward facing side of the helix.

Many groups have turned to *in vivo* systems to determine the interfacial residues of a TMD. Genetic reporter assays are more powerful tools than the quantitative methods listed above for analyzing a large number of mutations because of their higher throughput. These assays have evaluated helix-helix association in hundreds, if not thousands, of proteins from human, bacterial, plant, and viral systems. One generic reporter assay, TOXCAT was developed to specifically evaluate homodimer interactions (Russ and Engelman, 1999). In this assay, a TMD of interest is fused to the cytoplasmic ToxR transcription factor and the periplasmic maltose binding protein (MBP). When the helices dimerize, the transcription factors bind to the *ctx* promoter and induce expression of the reporter gene chlomamphenicol acetyl transferase (CAT). TOXCAT can check if the TMD is inserted into the membrane through a MalE complementation assay. MBP is natively translocated to the periplasm where it can bind a maltose sugar and bring it to the maltose transporter. In this assay, the plasmids are transformed into *E. coli* cells that lack native MBP (MM39), without which they do not grow on minimal maltose media. If the TOXCAT construct, which contains a fused MBP, is properly inserted into the membrane, the MM39 cells can now survive on minimal maltose media. Many versions of the TOXCAT assay exist (Bennasroune et al., 2005; Berger et al., 2010; Joce et al., 2011; Julius et al., 2017; Langosch et al., 1996; Lis and Blumenthal, 2006; Ouellette et al., 2017; Su and Berger, 2012), and the Senes lab has previously reported the robustness of one version, TOXGREEN (Armstrong and Senes, 2016).

TOXGREEN only slightly differs genetically from the TOXCAT assay in that the reporter gene is switched to superfolder green fluorescence protein (sfGFP) instead of CAT (Fig. 3.1).

The quantification of dimerization using CAT, performed as an enzymatic assay in Anderson et al., 2017, is a relatively laborious process, limiting responsible experimentation to eight samples per day. The fluorescence reporter gene allows the a more direct measurement of dimerization in unprocessed cell cultures on a plate reader. This advancement increased the throughput of the assay to hundreds of constructs in a day using 96-well plates in a plate reader.

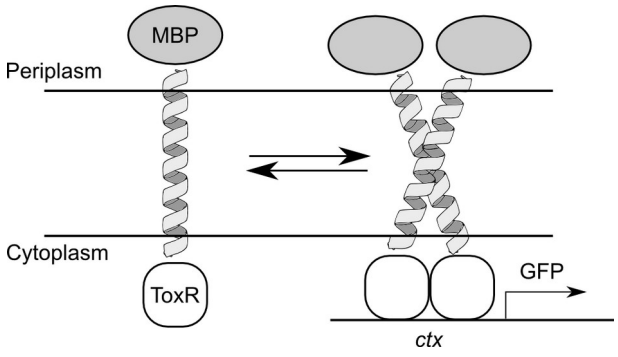


Fig. 3.1 TOXGREEN assay schematic. The transmembrane domain of interest is fused to a ToxR transcription factor on the N-terminus and a maltose-binding protein on the C-terminus. When the helices dimerize, so does ToxR and it can bind the *ctx* promoter and turn on the sfGFP reporter gene.

Unfortunately, the complex sequence space of proteins requires more mutagenesis than site-specific mutagenesis cloning can attain, even at hundreds of mutations. This challenge is why some groups have resorted to deep mutational scanning (DMS) to fully characterize the sequence space of single, or sometimes multiple, mutations. DMS is a technique that combines selection assays with next-generation sequencing (NGS) to identify variants enriched in certain properties, such as fluorescence or antibiotic resistance (Araya and Fowler, 2011; Fowler and Fields, 2014). Though the idea of selection and screening is not new (Russ and Engelman, 2000), the throughput has greatly increased due to sequencing technologies. DMS is a powerful new technology, and it has previously been applied to MPs. One group developed liposome display as an MP DMS technique (Fujii et al., 2014) and the Plückthun lab applied such an approach to evaluate GPCR stability (Sarkar et al., 2008; Schlinkmann et al., 2012; Schütz et al., 2016). However, these studies have not been applied to SPMPs.

Most relevant to my work, a deep-sequencing TOXCAT-β lactamase (dsTβL) assay was recently developed by Fleishman et al. The base genetic reporter assay differs from the

TOXCAT assay in that the MBP moiety is substituted with β-lactamase (Lis and Blumenthal, 2006). β-lactamase measures MP insertion through survival against ampicilin. Fleishman et al. adapted the assay to create the first DMS method for evaluating helix-helix interaction (Elazar et al., 2016). The authors tested their method on the well-characterized SPMP homodimers glycophorin A (GpA) and receptor tyrosine kinase, ErbB. They quantiied insertion and dimerization propensity of each mutant to identify the interface of the protein-protein interaction. The benefit of this assay is that it can differentiate between mutants that do not dimerize due to either interface disruption or improper insertion. The drawback of this selection assay, as with many ToxR assay variants, is that the readout is dependent on survival, which is an indirect way of measuring both insertion and dimerization.

Furthermore, a disadvantage for this, and most DMS assays, is that they evaluate fitness through an enrichment score. An enrichment score evaluates the population before and after selection and use the difference between the two as the output. Enrichment scores are prone to faulty interpretation due to the loss of global results. For example, a ten-fold increase result may be statistically significant, but a pre-enrichment starting value of 1 or 10,000 has dramatic effects on the biological relevance of the measurement. Moreover, for enrichment scores to be statistically sound, an assumption of independence between variant readouts must be made. For survival assays, though, this may not be a valid assumption because of limited available resources (Hibbing et al., 2010). A much more direct and useful measurement would be evaluating each protein independently in a single experiment.

The method described here addresses those challenges through a combination of the TOXGREEN assay with a variant of DMS called sort-seq (Fig. 3.2). To evaluate the function of many protein variants at a time, sort-seq calculates the fluorescence of individual cells using a cell sorter, each containing a single variant in a library of clones. The library is sorted into bins based on their fluorescence, each bin is sent for NGS, and the count of each variant in each bin is recorded. Using statistical inference, the fluorescence profile of each variant is reconstructed

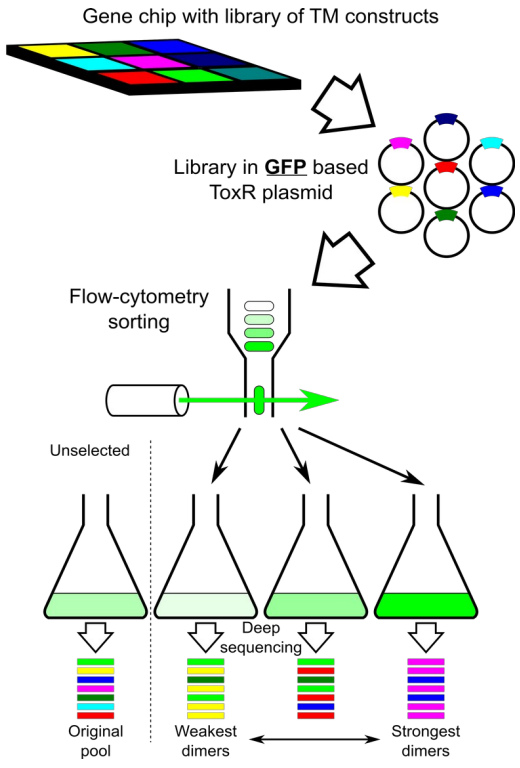


Figure 3.2 TOXGREEN sort-seq schematic. In the TOXGREEN sort-seq assay, I begin with an oligo pool library of different TM sequences that are cloned into the pccGFPKan plasmid. The sfGFP genetic reporter gene is expressed in accordance with the dimerization propensity of the TMD. Based on this fluorescence, the cells are sorted into different bins and the DNA is pooled and sent for NGS. NGS returns a count of each variant in each bin, which is used to reconstruct the dimerization propensity of the variants.

from its relative count in each bin. In sort-seq, variants are more independent than in other DMS assays because the functional readout does not affect survival. A major advantage of sort-seq compared to selection is that this method is able to evaluate the dimerization propensity of weak and strong dimers with equal confidence as opposed to enrichment scores.

Sort-seq was recently developed in 2010 to perform DNA footprinting on a promoter (Kinney et al., 2010). It has primarily been used to evaluate transcription regulatory elements (Cheung et al., 2019; Holmqvist et al., 2013; Kosuri et al., 2013; Müller et al., 2014; Peterman et al., 2014; Rohlhill et al., 2017; Sharon et al., 2012), though it has also been used to investigate protein evolution (McLaughlin et al., 2012; Starr et al., 2017). This chapter pioneers the TOXGREEN sort-seq method to evaluate TM helix dimerization through the expression of sfGFP at the level of tens of thousands of variants in a single experiment. As described below, I have successfully optimized the sorting and sequencing parameters to match flow cytometry

measurements of individual variants. As proof-of-principle, I have applied this method to identify the dimerization propensity and important interfacial residues of 100 human TM helices.

3.3 Results and Discussion

3.3.1 Library design and construction

For the development of the TOXGREEN sort-seq method, I designed a mutagenesis library of wild-type human TMDs predicted to form GAS_{right} homodimers. Schematically, out of the 2,383 single-pass proteins in the human genome annotated in UniProt on 2016.11.02 (UniProt Consortium, 2019), 1,141 were predicted by the CATM algorithm to potentially form GAS_{right} homodimers of various stabilities (Anderson et al., 2017). From these sequences, I removed any that contained proline to prevent helix kinking. I also removed sequences with more than one strongly polar residue (D, E, H, K, N, Q) because excessive hydrogen bonding would likely overwhelm the Cα—H bonds that mediate the GAS_{right} models. To standardize my constructs I chose TM sequences that were exactly 21 residues long as annotated by UniProt, the most common length found in the database. I also discarded any sequence in which part of the interface fell outside of the TMD. From the remaining sequences, I selected the 100 top scoring TMDs in CATM (Table 3.1).

Because ToxR assays – such as TOXGREEN – can be quite sensitive to the length of the TM helix, we tested each sequence in three different lengths (Lawrie et al., 2010). In addition to the 21 amino acid wild-type sequences, I truncated these proteins to lengths of 19 and 17 residues. These truncations were performed either from the N or C terminus so that the predicted crossing point moved toward the center of the membrane. The libraries are referred to as L21, L19, and L17 signaling the different lengths.

To perform extensive mutagenesis, each position in these helices was mutated to three hydrophobic residues: Ala, Leu, and Ile. If the wild-type residue was one of these, the third mutation would be Phe, another hydrophobic, but large residue. Together, each protein was tested with 174 single variants (3 WT sequences, plus 21×3 + 19×3 + 17×3 point mutants), resulting in a library of approximately 18,000 variants. These sequences were ordered as an

oligo pool library and cloned into MM39 (MBP deficient) cells. A semi-random subset of this library was isolated and clonally measured for sort-seq verification (see Methods Section 3.5.9). When combined, this subset is known as the “spike-in” library.

3.3.2 Flow cytometry measurements accurately represents dimerization propensity

To evaluate the feasibility of measuring TOXGREEN signal with a fluorescence-activated cell sorter, we decided to benchmark a set of known standards. The difference between bulk fluorescence measurements and cell sorting is that the former measures the average fluorescence of a population of cells of one protein variant in a single well while the latter measures the distribution of fluorescence that occurs in many individual cells (Peterman and Levine, 2016). I performed the TOXGREEN assay for 95 test variants from the full library with a plate reader and with a flow cytometer (Fig. 3.3). Since the two instruments measure fluorescence in relative fluorescence units (RFU), linear correlation between them is expected.

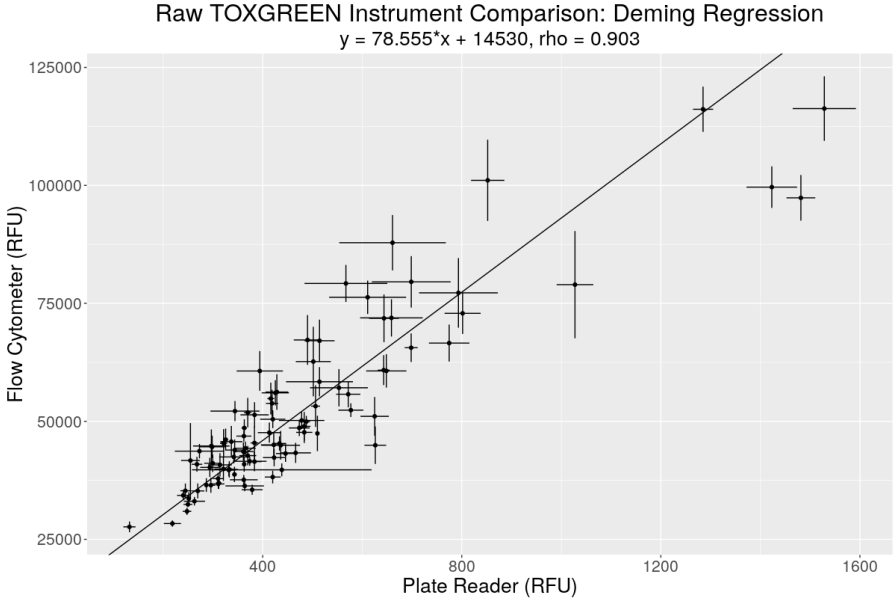


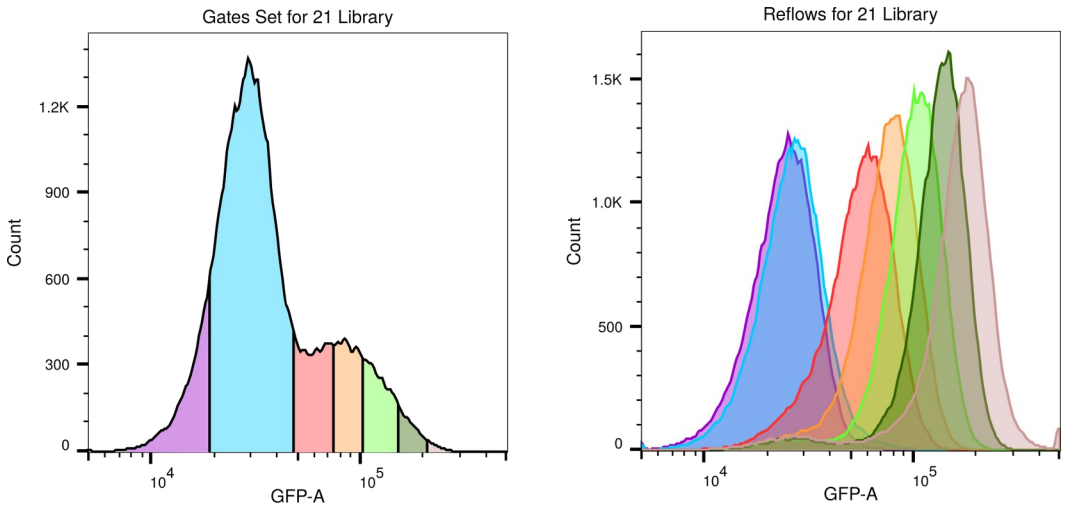
Figure 3.3 TOXGREEN instrument comparison. Running TOXGREEN on a plate reader or a flow cytometer results in similar relative measurements for the sum of variants. N = 95, error bars are standard error of 3 replicates in the x direction and at least 5 (up to 12) replicates in the y direction.

Given that both variables have error, I performed a Deming linear regression (Deming, 1943) between the average fluorescence on the plate reader and the average median of the distribution on the flow cytometer. A Pearson correlation coefficient was used to evaluate how well the data match each other. As the overall fluorescence increases, so does the divergence from each other. However, overall the fit is very good ($\rho = 0.9$), demonstrating the reproducibility of TOXGREEN across instruments.

3.3.3 Sort-seq method

Once it was demonstrated that it is possible to differentiate between the selected variants by flow cytometry, I implemented the sort-seq method on the spike-in library of constructs (Kinney et al., 2010). Briefly, a sort-seq method takes a library of constructs, uses a fluorescence activated cell sorter to pool and sort variants within desired ranges of fluorescence, and uses NGS data to reconstruct the fluorescence value for each individual construct (Fig. 3.2). Each of these steps has operative decisions that must be validated and optimized to most accurately represent the system. After gating away cell debris and doublets (two cells that flow together), sfGFP fluorescence thresholds (called “gates”) are chosen to separate the library into pools. A comprehensive study based on simulated data aided me in making these choices for my system (Peterman and Levine, 2016).

However, this simulated data does not take into account a very important detail: sorting error from the FACS instrument. An important part of performing sort-seq is to ensure that the set gates from which fluorescence statistics of the sorted pool are calculated, match the population distribution that are actually sorted. Fig. 3.4 shows how this is not necessarily true. Each reflowed bin is shifted toward the center of the initial population and in the higher fluorescence bins have a long tail to the left. As I discuss below, the differences between these set gates and reflow distributions are important when setting up the parameters for individual construct fluorescence reconstruction.



	Subset Name	Freq. of Parent	Median GFP
	C6	0.75	212514
	C5	2.52	161346
	C4	6.55	117904
	C3	8.27	85483
	C2	9.71	61078
	C1	60.2	29377
	C0	10.1	16062

	Sample Name	Median GFP
	C6	171446
	C5	136921
	C4	101643
	C3	76393
	C2	56710
	C1	26224
	C0	24376

Figure 3.4 Example reflow distributions. (Left) Representative distribution of the 21 length library that spans ~1.5 logs. Seven bins were gated for sorting with median GFP values from 16k to 213k, a 13 fold difference. After sorting, the resulting cell pools were flowed again on the cytometer resulting in the distributions on the right. (Right) Gates C0 and C1 overlap with each other meaning the FACS machine was unable to differentiate between the two. The rest of the gates clearly separated with a wide range of medians, yet all the gate medians shifted toward the center of the distribution.

I used the NGS reads in each bin to reconstruct the fluorescence level of each protein in the spike-in library. To do this construction, I used a weighted average that takes into account the NGS sampling level of each bin and percentage of the fluorescence distribution that sorts into each bin (Kosuri et al., 2013). Each gate was characterized by the median fluorescence. I started by comparing the median fluorescence of each of the individually characterized constructs to reconstructed RFU from the spike-in sort-seq (Fig. 3.5). In an ideal scenario, the data would create an x = y line, and in this effort, the data come incredibly close (Pearson correlation coefficient = 0.903). I compared both the individual measurements and the spike-in

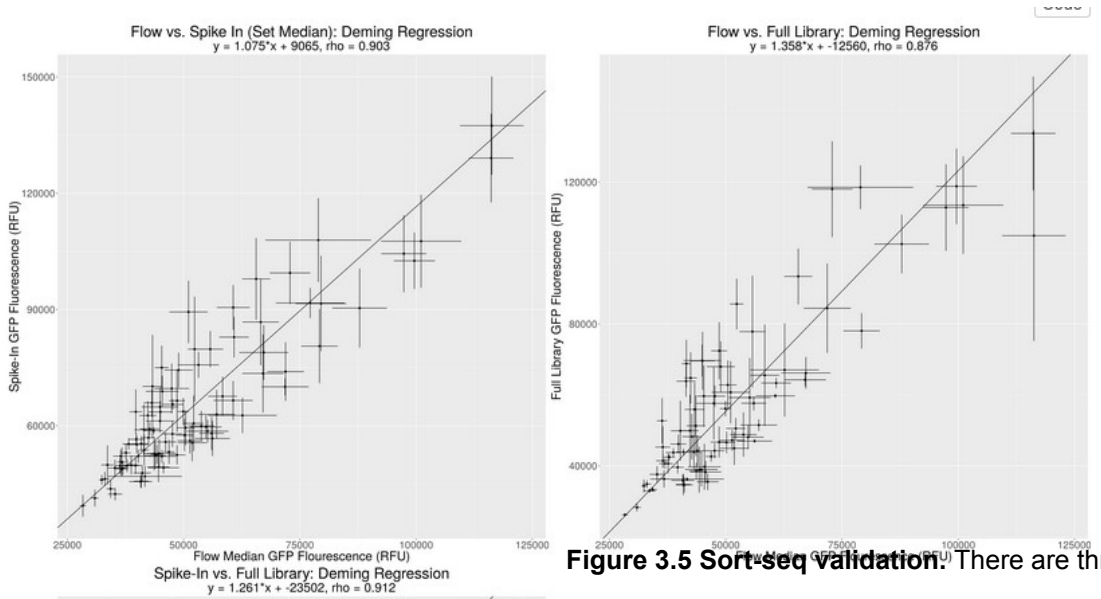


Figure 3.5 Sort-seq validation. There are three sets of measurements on the same constructs: clonal flow cytometry (at least 5 replicates), spike-in sort-seq (3 replicates), and full library sort-seq (3 replicates). Error bars represent standard error of the mean. An ideal line would be x = y and all of the measurements significantly align, demonstrating that sort-seq can accurately recreate individual TOXGREEN measurements.

reconstruction values to the entire library reconstruction values and obtained a very good correlation. In sum, this means that the sort-seq reconstructed RFU values can be used to evaluate the dimerization propensity of many individual constructs within a library.

3.3.4 Evaluating protein insertion into the membrane

By definition, quantifying the association strength of a protein-protein interaction requires a measurement of the relative concentrations of monomer to dimer. However, MPs have the added complexity of membrane insertion. According to the two-stage model of MP folding (Popot and Engelman, 1990), the dimerization of helices does not occur until after insertion, so

only the amount of protein in the membrane is relevant for stability, or K_d, calculations. It has been argued that the TOXCAT assay can quantify the amount of protein inserted into the membrane using Western blots and thresults were compared to SE-AUC calculations (Duong et al., 2007; Elazar et al., 2016), but the correlation is weak. Currently, exact protein concentration and insertion rates are not accessible with the ToxR methods.

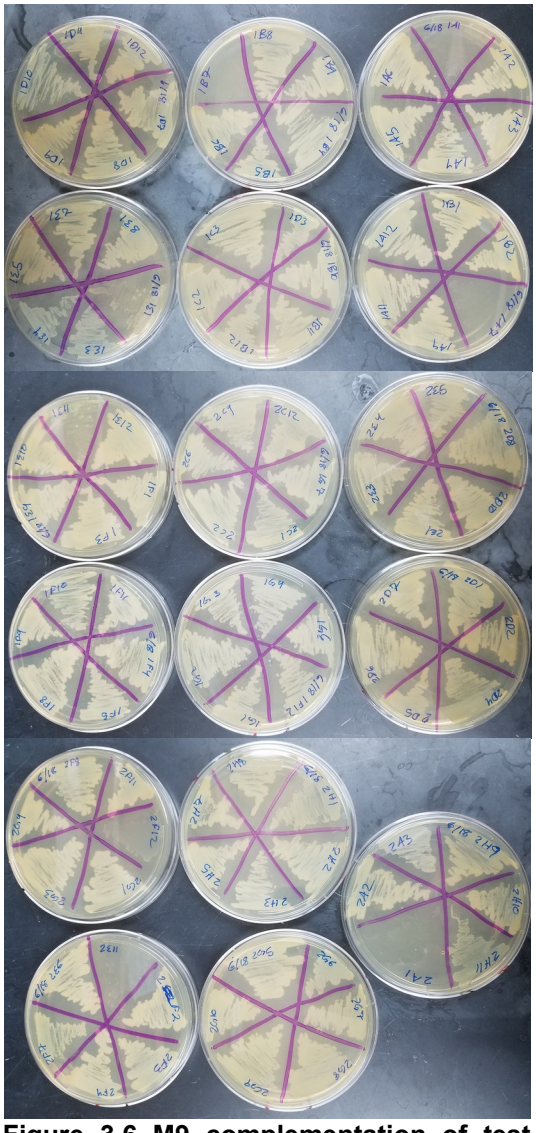


Figure 3.6 M9 complementation of test clones. Stationary phase cultures of each construct were streaked out on minimal M9 plates and incubated at 37 C for 72 hours. Five constructs did not grow indicating they are not inserted into the membrane (1A01 [also 2A01], 1A07, 1B07, 1E11, 2F12).

Nevertheless, in low-throughput ToxR based assays, protein expression is estimated via Western blots and membrane insertion is estimated through a MalE complementation assay. I performed these tests for the 105 individual constructs (Fig. 3.6) that resulted in binary insertion results and variable protein expression levels (data not shown). It would be impossible to perform individual Western blots and MalE complementation assays for each of the 18,000 sequences in the library. Instead, to evaluate protein insertion at the library level, I needed to apply a different evaluation procedure.

One study assayed protein concentration in the membrane through antibiotic resistance selection enrichment scores (Elazar et al., 2016). A variant of the ToxR assay in which the MBP moiety has been switched for β-lactamase confers resistance to ampicillin (Lis and Blumenthal, 2006) Elazar et al. used this assay to measure insertion through a

survival enrichment score by measuring NGS counts before and after selection in ampicillin. When I attempted replicating these results for my controls, there was a reduced range of resistance to the dimerization reporter gene, CAT (Fig. S3.1) and I could not differentiate between the positive and negative controls with the sfGFP reporter gene (Fig. S3.2).

Another group developed a method that measures both protein expression and membrane insertion using a myc epitope tag (Huang et al., 2018). In this method, a mammalian cell is stained with one fluorescent antibody that labels all of the properly inserted KCNQ1. The membrane is then permeabilized, and the cell is stained with another a different fluorescent antibody to label the intracellular proteins. The cells are passed through a two-channel flow cytometer to measure the fluorescence of intracellular and membrane-inserted KCNQ1. Mammalian cells are much more robust to permeabilization, but it is possible to permeabilize *E. coli* cells with some technical difficulty (Ranjit and Young, 2013).

Instead of performing the standard plate-based complementation assay, I scaled up by implementing a liquid culture version. I hypothesized that over the course of a library growth in maltose minimal media, at some point, all of the poorly inserting constructs would be competed out due to poor uptake of the carbon source. To test this hypothesis, I performed a growth curve of 97 constructs through sequencing. After 36 hours of growing in the M9 maltose media, constructs that do not grow on M9 maltose media plates had disappeared from the pool (Fig. 3.7). Two additional constructs also disappeared that I knew grew on the M9 plates, out-competed by constructs with better insertion. A practical approach is to grow the libraries for 36 hours before sequencing: this growth time enables me to filter out any constructs that would not insert into the membrane, even at the cost of losing some good constructs. In the conclusions, I discuss some future quantitative alternatives for total protein expression.

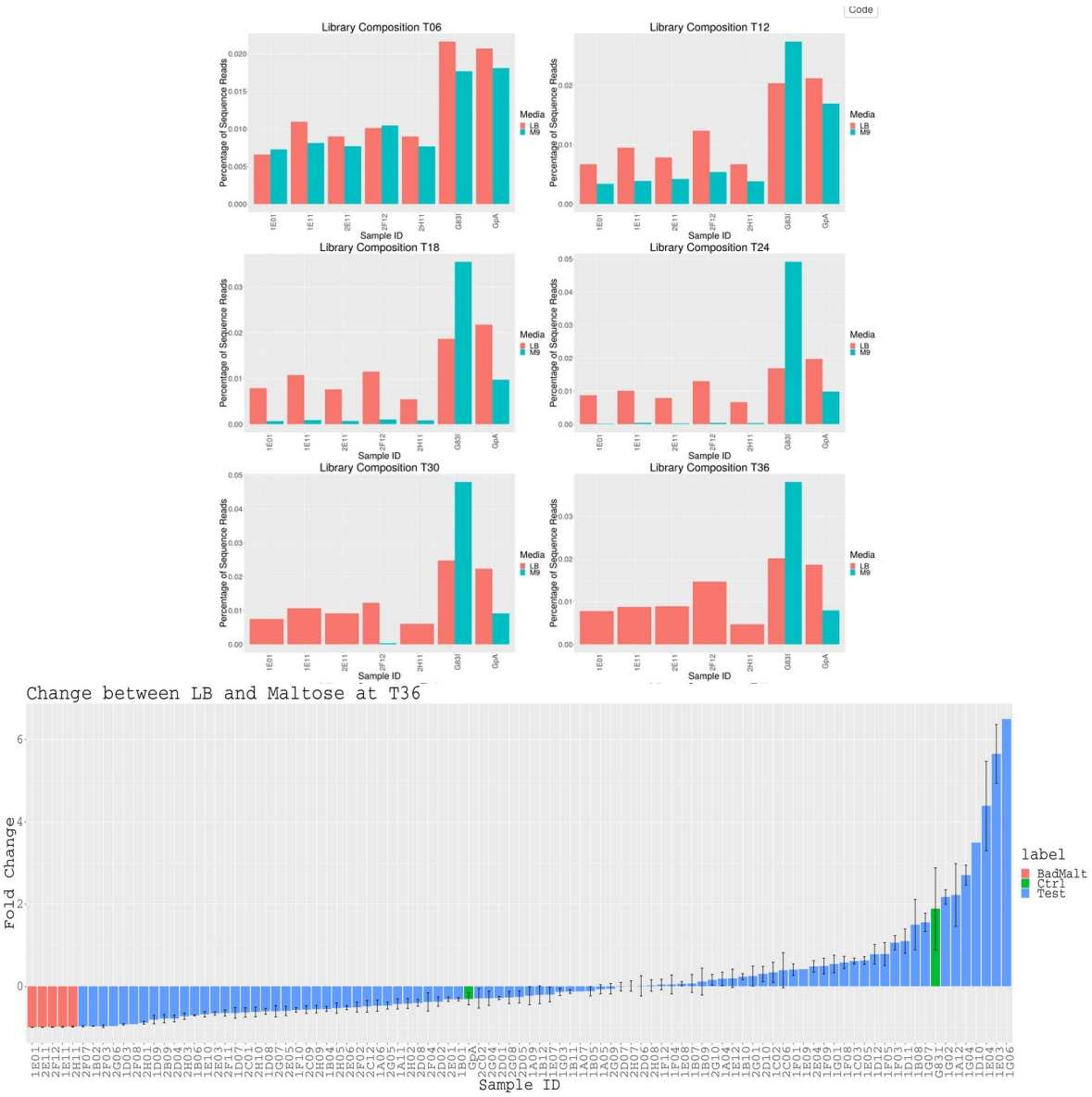
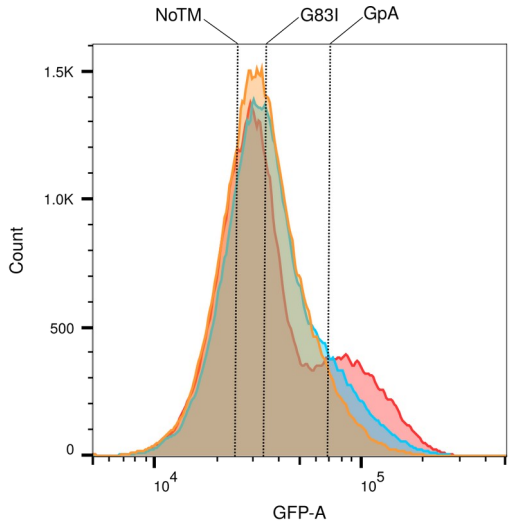


Figure 3.7 Insertion testing of spike-in library. (Top) Pink bars show the percentage of sequence reads for each construct in LB media and blue bars are in M9 media. The first five constructs poorly insert into the *E. coli* membrane according to the individual maltose test. GpA and G83I are positive insertion controls (Fig. 3.6). After six hours of growth, the constructs are represented similarly in each media, but for each time point afterwards, the poor inserters have less representation in the M9 media. (Bottom) The difference after 36 for the entire spike-in library is shown. Error bars are the standard error of three replicates. Red bars sequences that poorly inserted individually, green are dimerization controls, and blue bars represent sequences I know insert (Fig. 3.6).

3.3.5 Sort-seq on a library of human transmembrane dimers

I demonstrate applicability of the sort-seq method with extensive mutagenesis of human TMDs, designed in Section 3.3.1. Each library of different TMD lengths was cloned, sorted, and processed separately. Fig. 3.8 shows the fluorescence distribution of the three different libraries, L17, L19, L21. Three controls were used in my experiments: “NoTM”, “GpA”, and “GpA-G83I”. NoTM is an empty vector, which has an average median GFP fluorescence of 25,500 ± 1,100. GpA is the TMD of glycophorin A (GpA), a stable dimer: it has an average median GFP fluorescence of 63,100 ± 4,900. GpA-G83I is a monomeric mutant of GpA which has an average median GFP fluorescence of 31,700 ± 1,300. Though the majority of the variants in each library overlay with the empty vector, there is a high fluorescence tail that extends beyond this. The positive tail are variants that can dimerize better than the negative controls. This was expected because only a few constructs are wild-type and predicted to be dimers, while many of the mutations are likely to decrease the stability of the dimers. There are even many variants that extend beyond GpA control’s fluorescence, indicating that there are very strong dimers within the libraries.



Sample Name	Count	Median GFP	% above GpA
17 TM Length	88366	32142	6.81
19 TM Length	85624	34269	12.5
21 TM Length	85292	32652	19.5

Figure 3.8 Example distribution of various TM libraries. The three libraries of different TM length, 17, 19, and 21 are overlaid with each other. The amount of the library that is above the median GpA value is shown as a percentage of the total library.

Figure 3.9 (top) shows the distribution of dimerization propensity of the 87 wild-type proteins cloned in the 21 TM length library. The distribution shows a full range from 7-240% GpA and 60 of the constructs are considered dimers at above 40% GpA. Applying the same cutoff to the

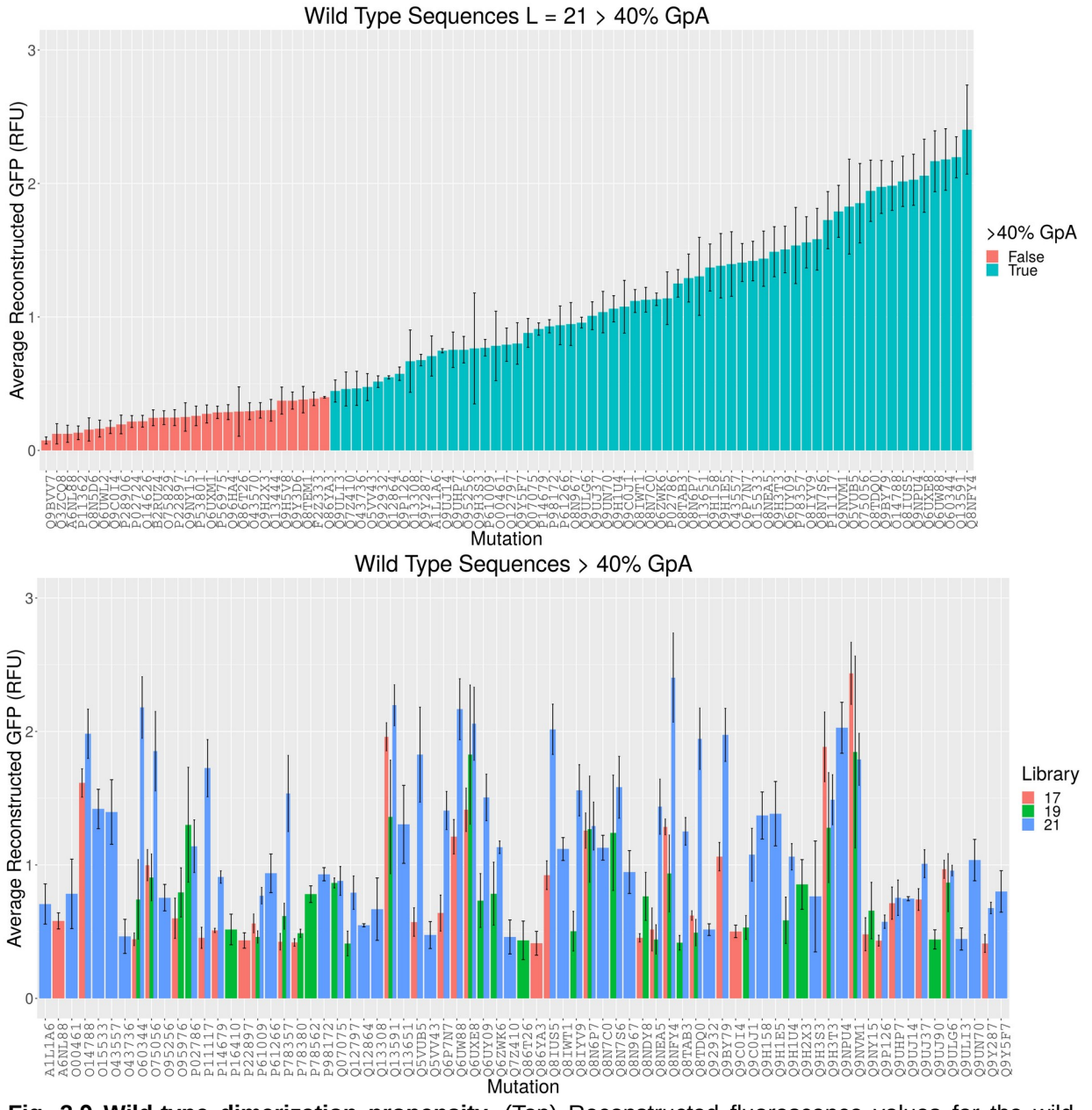


Fig. 3.9 Wild-type dimerization propensity. (Top) Reconstructed fluorescence values for the wild-types that are 21 amino acids long. In pink are wild-types that are less than 40% of the dimerizing control GpA, or non-dimers. In blue are the dimers identified by sort-seq. (Bottom) Reconstructed fluorescence values for the wild-types that are above 40% of GpA of the three different lengths, 17, 19, and 21 amino acids long. N = 3 and error bars are standard error.

truncated libraries, L19 ranges 4-184% GpA and has 32/58 dimers, L17 ranges 0-244% GpA and has 33/101 dimers. Fig 3.9 (bottom) shows all of the dimers colored by TMD length. It is

interesting to note that the L21 library has more dimers than L17 and L19, and of the proteins that have dimers of different lengths, the L21 variant is often more stable. Only seven of these 132 dimers (73 unique proteins) did not pass the MalE complementation test meaning the vast majority of the TMDs properly inserted into the membrane (Table 3.2). The list of dimeric proteins contains mulitple members of the serine protease, tumor necrosis factor, semaphorin, cadherin, interkeukin receptor, and C-type lectin domain families. Several of the proteins are unnamed and uncharacterized, but others are known as homo- or hetero-dimers.

In addition to characterizing potential new dimers in the human genome, I identify critical residues for most dimers. For some dimers, I was even able to identify an interface with multiple critical residues along one face of the helix. The following UniProt IDs have clear interfaces that include a Sm-xxx-Sm motif that is indicative of a GAS_{right} motif, though further studies will be needed to confirm this hypothesis: O75056 (Syndecan-3), P02786 (Transferrin receptor protein 1), P61009 (Signal peptidase complex subunit 3), Q07075 (Glutamyl aminopeptidase), Q13591 (Semaphorin-5A), Q6P7N7 (Transmembrane protein 81), Q6UW88 (Epigen), Q6ZWK6 (Transmembrane protease serine 11F), Q6N6P7 (Possible EA31 gene protein, phage lambda), Q8NEA5 (Uncharacterized protein C19orf18), Q8NFY4 (Semaphorin-6D), Q8TDQ0 (Hepatitis A virus cellular receptor 2), Q9H3T3 (Semaphorin-6B), Q9NVM1 (Protein eva-1 homolog B), Q9UJ37 (Alpha-N-acetylgalactosaminide alpha-2,6-sialyltransferase 2).

In Figure 3.10, I show one example with a clear mutational profile, UniProt ID O75056, L17. O75056 codes for a cell surface adhesion molecule, Syndecan-3 and the mutational profile shows a GxxxGxxxA motif that is critical for dimerization because multiple amino acids significantly decrease the dimerization propensity. When the average disruption is plotted by residue on a helical wheel, the critical residues all lie on one face of the helix. The combination of the GxxxG-like motif and the single face indicate that this TMD is likely to form a GAS_{right} dimer. These data support a previous study that identified the same GxxxG sequence as critical for dimerization (Dews and Mackenzie, 2007). Identifying a known dimer with a matching

mutational profile confirms that this TOXGREEN sort-seq assay can accurately recapitulate dimerization propensity.

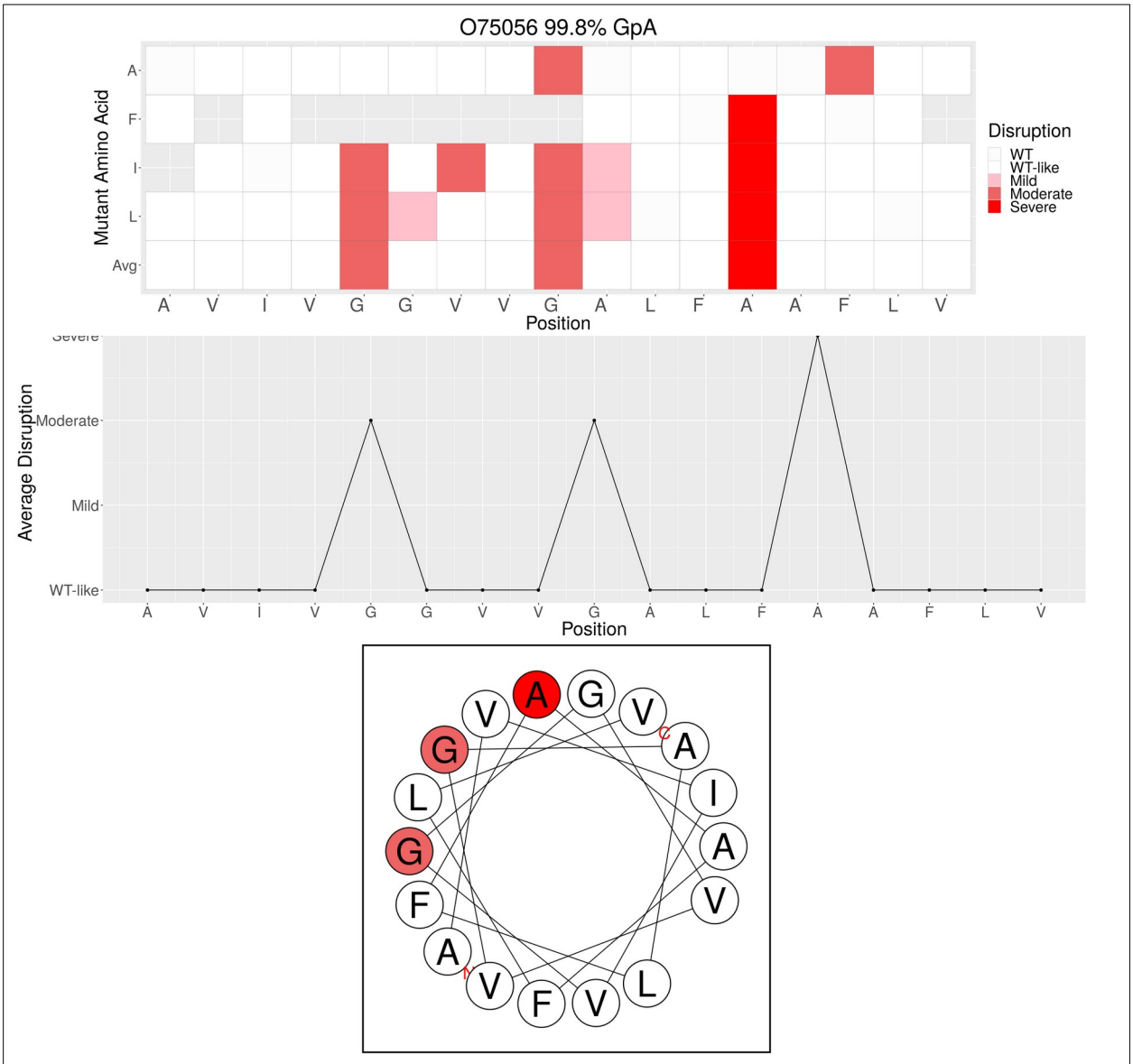


Figure 3.10 Mutational profile of O75056. An example of mutational analysis to identify an interface. (Top) The heat map shows disruption per mutation as compared to the wild-type sequence. WT-like >75% WT, mild >50% WT, moderate >25% WT, severe < 25% WT. The bottom row averages the percent disruption and colors accordingly. (Middle) Average disruption is plotted by residue. (Bottom) A helical wheel colored by average disruption per residue. All of the disruptive mutations are along one face of the helix.

3.4 Conclusions

I have demonstrated that the sort-seq method accurately recreates the clonal TOXGREEN measurements and that it greatly increases the number of samples that can be measured in a single experiment. Furthermore, I can measure insertion rates using an enrichment score for the MalE complementation assay. I have shown that the fluorescence range of the TOXGREEN assay is able to characterize the raw fluorescence value of thousands of sequences. Overall, the sort-seq TOXGREEN assay provides an unprecedented high-throughput screening ability for TM dimers. Combined with the ability to construct a large number of specific mutations, this assay allows for the testing of hypotheses without the noise of random mutagenesis.

Here, I tested a library of human TMDs with extensive mutagenesis at different lengths. I identified 73 oligomers and the residues important for their association, many of them previously uncharacterized. Future work on this project will include further analysis of the mutational data to identify experimental interfaces and compare them to the interfaces predicted by the CATM algorithm. This comparison will enable me to train the CATM algorithm for better structure and stability predication capability. Furthermore, I will use the Molecular Software Library to model the constructs that do not match the algorithm's predictions.

A variety of TMD questions can be answered using sort-seq including the characterization of mutational patterns of TMDs, protein design, co-evolutionary analysis, and stability calculations. The sort-seq method can be adapted to any ToxR assay that use a fluorophore as the reporter gene (Grau et al., 2017). Future improvements should focus on adding a quantitative measurement of expression and insertion through fluorescence rather than selection. Another challenge to the TOXGREEN assay is the limited dynamic range of reporter gene expression. The dynamic range of an assay affects the resolution of the assay, therefore, with an increased reporter range, we can better discriminate between variants, transitioning from mild, moderate, and severe to percent change. Potential options are discussed in Section 5.3.

3.5 Methods

3.5.1 Software

All calculations were implemented and performed using MSL v. 1.1, an open source C++ library that is freely available at http://msl-libraries.org (Kulp et al., 2012). Flow cytometry analysis was performed in SONY Cell Sorter Cytometer and Becton, Dickinson & Company FlowJo. Statistical analysis and graphing was implemented in Rstudio Version 1.1.456 (RStudio Team, 2016). Relevant packages include:

tidyverse: (Wickham, 2017)

ggplot2: (Wickham, 2016)

reshape2: (Wickham, 2007)

dplyr: (Wickham et al., 2019)

data.table: (Dowle and Srinivasan, 2019)

mcr: (Manuilova et al., 2014)

gridExtra: (Auguie, 2017)

plotly: (Sievert, 2018)

scales: (Wickham, 2018)

helixvis: (Wadhwa et al., 2018)

remotes: (Hester et al., 2019)

heliquest: (Gautier et al., 2008)

gsubfn: (Grothendieck, 2018)

plyr: (Wickham et al., 2019)

3.5.2 Prediction of GAS_{right} structure

Structural prediction was performed with the program CATM (Mueller et al., 2014). Side chain mobility was modeled using the Energy-Based conformer library applied at the 95% level (Subramaniam and Senes, 2012). Energies were determined using the CHARMM 22 van der

Waals function (MacKerell et al., 1998), the IMM1 membrane implicit solvation model (Lazaridis, 2003), and the hydrogen bonding function of SCWRL 4 (Krivov et al., 2009), as implemented in MSL, with the following parameters for Cα donors, as reported previously (Mueller et al., 2014): B=60.278; D_0=2.3 Å; σ_d=1.202 Å; α_{max}=74.0°; β_{max}=98.0°.

The CATM algorithm was described in detail in previously (Mueller et al., 2014). Briefly, the sequence of interest is threaded into a set of different registers at each of 463 representative geometries. If sequence-based filtering rules are met, the sequence is built on the backbone in all atoms and the helices are docked by reducing the inter-helical distance in steps. At each step the side chains are optimized and the interaction energy is evaluated until a minimum energy is found. To further optimize the dimer, the geometry is then subjected to Monte Carlo backbone perturbation cycles in which all inter-helical parameters (distance, Z shift, axial rotation, and crossing angle) are locally varied. If the final interaction energy (calculated as the energy of the dimer minus the energy of two monomers separated at long distance) is negative, the solution is accepted. The solutions are then clustered using an RMSD criterion to produce a series of distinct models. The wild-type structures are available at http://seneslab.org/CATM.

3.5.3 Cloning of chimeric library

The oligo pool containing the sequences of interest was synthesized by Twist Bioscience and delivered as a pool of lyophilized single-stranded DNA. The oligo pool was resuspended to 1 ng/uL. Subpools were amplified using primers specific to each subpool (Table 3.4) from each oligonucleotide pool by means of quantitative PCR (SYBR qPCR master mix, KAPA Biosystems; 20 µl reaction volume; 1 ng oligonucleotide pool template) until the second inflection point on a real-time plot of cycle number versus well fluorescence indicated amplification saturation was beginning, following the protocol outlined in (Kosuri et al., 2010). Another traditional PCR reaction was performed using the qPCR reaction as a template for 15 cycles to obtain sufficient DNA for cloning. An aliquot was saved to confirm amplicon size. After

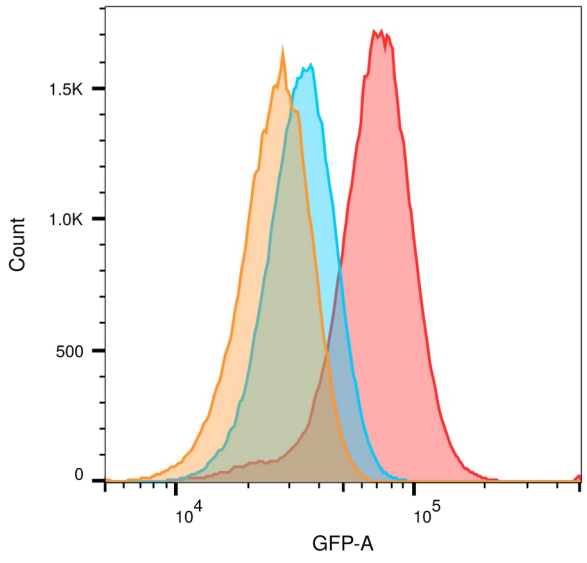
a PCR clean-up, the amplified subpools were double digested with NheI-HF and DpnII in CutSmart Buffer from New England Biolabs (NEB). Since DpnII works at 25% in CutSmart, I used 4x as much DpnII as NheI-HF for an overnight digest at 37°C. The digested subpools were cleaned up with Qiagen's Nucleotide Cleanup kit and a 4% DNA gel was run to determine if the DpnII digest was successful.

100 ng of the subpools were ligated into 100 ng of NheI, BamHI, and CIP digested pccGFPKan (Armstrong and Senes, 2016) with NEB's ElectroLigase for 2 hours at room temperature, followed by heat-inactivation at 65°C for 10 minutes. 5 uL of the ligation mixture was electroporated into 50 uL of NEB's DH10β cells. The cells were recovered in 1 mL SOM media for 1 hour at 37°C and then 4 mL LB with ampicillin was added for overnight growth. To estimate the number of transformants, x100 and x1,000 dilutions of the recovered cells were plated onto LB plates containing 100μg/mL ampicillin and grown overnight at 37°C. The rest of the cells were grown overnight at 37°C. If the colony forming units were at least ten times the number of sequences in the subpool, the overnight culture was miniprepped.

2 uL of the miniprepped plasmid libraries were transformed into 100 uL of electrocompetent MM39 cells with the same recovery and plating procedure. If the colony forming units were at least ten times the number of sequences in the subpool, the overnight culture was saved as a glycerol stock for future experiments.

3.5.4 Selection of individual clones

The MM39 libraries were combined by length of transmembrane protein and sorted by a SONY LE-SH800 into three gates: low, medium, and high fluorescence. The low gate was drawn with no lower bound, but with an upped bound at the median fluorescence of the monomerizing mutant of GpA, G83I. The medium gate's lower bound was this same value and the upper bound was drawn at the median fluorescence value of GpA. The high gate was defined as anything above this value (Fig. 3.11). The sorted libraries were plated on LB/Amp and incubated at 37°C overnight. 18-36 colonies were chosen from each one of these pools.



	Sample Name	Count	Median GFP
	NoTM	72,808	26,550
	GpA:G83I	72,007	33,772
	GpA	81,493	69,900

Fig. 3.11 Example control construct distributions. I collected 100,000 events for each construct and gated out poor events. The median GFP value for the remaining events for each construct are displayed. There is a 2.6 fold difference between the negative construct with no TMD and GpA, the positive control.

Colonies were sent for sequencing the eliminate doubles and frame shifted variants, resulting in 98 variants and three controls.

3.5.5 TOXGREEN

TOXGREEN constructs were transformed into *E. coli* MM39 cells. sfGFP expression was quantified in stationary phase in LB media. A freshly streaked colony was inoculated into 3 mL of LB broth containing 100µg/mL ampicillin and grown overnight at 37 °C. To reduce background for samples grown in LB, 1.5 mL of cells were collected by centrifugation at 17,000g and concentrated three-fold by re-suspending them in 0.5 mL in PBS solution (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 2 mM KH_2PO_4, pH 7.4), prior to fluorescence measurements. Aliquots were removed and stored in SDS-PAGE loading buffer for immunoblotting. 300µL of each cell sample was transferred to a 96-well black walled, clear bottom plate (Fisher Scientific).

Plate reader fluorescence measurements were performed using an BioTek Synergy | HTZ multi-mode plate reader, using an excitation wavelength of 485 nm and emission wavelength of 528 nm. The relative sfGFP expression (TOXGREEN signal) was calculated by normalizing the fluorescence emission at 512 nm to the optical density of the sample at 600 nm that was measured individually on a Agilent Technologies Cary 60 UV-Vis. The normalized fluorescence

of each sample was then subtracted of the normalized fluorescence of cells that contained the no-TM control plasmid pccGFPKAN to remove non-specific background.

Flow cytometry measurements were performed in a SONY LE-SH800 with a SONY 70 μM chip. The threshold was set at 0.05% and the FL1 (GFP) gain was set at 90%. Back scatter gain was set at 30%. The FCS files were analyzed using FlowJo. The first gate was drawn to eliminate doublets on a forward scatter area vs height graph. The second gate was drawn to eliminate cell debris on a forward scatter area vs. side scatter area graph. Median GFP fluorescence was calculated from gated cells for each variant.

3.5.6 MalE complementation assays

To confirm proper membrane insertion and orientation of the individual TOXGREEN constructs, overnight cultures were plated on M9 minimal medium plates containing 0.4% maltose as the only carbon source and grown at 37 °C for 48 - 72 h.

MalE complementation assays for the libraries were performed in liquid media. Each construct or library was grown was transformed into *E. coli* MM39 cells. A freshly streaked colony was inoculated into 3 mL of LB broth containing 100 μg/mL ampicillin and grown overnight at 37 °C. The cells were back diltuted and normalized to OD600. 0.01 ODs of each of constructs for the spike-in culture were mixed together. Two flasks of 500 mL of LB and M9 maltose media cultures were started for each of the 3 libraries and the spike-in library with 0.00125 total ODs. Samples were taken every six hours, miniprepped, and sent for NGS. If the ratio of the fraction of the population in M9 was -0.95 fold that of the population in LB, I classified it as non-inserting.

3.5.7 Immunoblotting for individual constructs.

Samples were grown overnight, spun down and resuspended in sonication buffer (25mM Tris-HCl, 2mM ethylenediaminetetraacetic acid, pH 8.0). 500 uL of the sample was lysed by bath sonication in a Qsonia cuphorn. Cell lysates were normalized by total protein concentration and loaded onto a NuPAGE 4-12% Bis-Tris SDS-PAGE gel (Invitrogen) and then transferred to

polyvinylidene difluoride membranes (VWR) for 1 hour at 100 millivolts. Blots were blocked using 5% Bovine serum albumin (US Biologicals) in TBS-Tween buffer (50 mM Tris, 150 mM NaCl, 0.05% Tween 20) for overnight at 4 °C, incubated with goat biotinylated anti-Maltose Binding Protein antibodies (Vector labs) for 2 hours at room temperature, followed by peroxidase-conjugated streptavidin anti-goat secondary antibodies (Jackson ImmunoResearch) for two hours at 4 °C. Blots were developed with the Pierce ECL Western Blotting Substrate Kit, 1 mL of enhanced chemiluminescence solution was added to the blot and incubated for 90 s. Chemiluminescence was measured using an iBright CL1000 Imaging System (Invitrogen). Individual bands were quantified by ImageJ (Schneider et al., 2012).

3.5.9 Spike-in procedure

Individual constructs were grown from colonies picked into 96-well plates overnight. In the morning, 10 uL of each overnight culture were combined and the cells were diluted into PBS at approximately 1:50 and sorted.

3.5.10 Fluorescence-activated cell sorting

FACS was performed in a SONY LE-SH800with a SONY 70 μM chip. The threshold was set at 0.05% and the FL1 (GFP) gain was set at 90%. Back scatter gain was set at 30%. The first gate was drawn to eliminate doublets on a forward scatter area vs height graph. The second gate was drawn to eliminate cell debris on a forward scatter area vs. side scatter area graph. 6-7 gates were drawn to span of the entire library to maximize reflow accuracy. 100,000 cells from each gate were sorted into PBS and reflowed until 60,000 events were collected to estimate the distribution of the sorted population. 500,000 events were sorted into 2 mL LB broth containing 100 μg/mL ampicillin and grew up in a total of 5 mL LB broth containing 100 μg/mL ampicillin until an OD600 of approximately 0.1 and miniprepped.

3.5.11 Deep sequencing

The TMDs were amplified for sequencing using a specific polymerase chain reaction (PCR) procedure. For 12 cycles, the plasmid was mixed with 15 nM forward and reverse primers specific to the TMD. For MiSeq runs, an additional 8 cycles were performed with 1.45 μM stem primers, PCR reactions were loaded on agarose gels to confirm the addition of the stem primer, and successful reactions were cleaned up using Qiagen's Nucleotide Cleanup kit. The samples were submitted to the University of Wisconsin Biotechnology Center DNA Sequencing Facility. I ran 2x150 runs on an Illumina MiSeq or NovaSeq600.

3.5.12 Sequence Analysis

The quality of a NGS sequence is determined by the Phred score (Q) of each base, measuring the probability that the base is incorrect. Often the average Phred score of the entire sequence as a measure a measure quality. I have instead decided to use the expected number of incorrect bases in a sequence which is logarithmically related to Q. As an example, I will describe two different reads. The first read has 140 bases at a quality of 35, but 10 bases have a quality of 2. The average Q value of this read is 33, but the expected number of incorrect bases is 6.4. The second read has all 150 bases at a quality of 25. The average Q value of this read is 25, but the expected number of incorrect bases is 0.5. I argue the latter metric for excluding sequences from the analysis. I implemented a cut-off of 1 incorrect base pair per sequence. Finally, I had to decide how many copies of a sequence had to be present in a NGS run to be confident that it was indeed there. I set this cut off to be 10. After filtering, each DNA sequence was translated to amino acids and matched to it's appropriate TM in length, protein ID, and mutation. This work was all completed using custom Perl programs.

The statistical inference calculations were performed in R. Reconstructed GFP levels were calculated as a weighted average (Kosuri et al., 2013). This method normalizes read per construct per bin with the fraction of the population that is in that bin. The normalized fractional contribution of each bin (j) for each sequence (i), a_{ij} is calculated as:

$$a_{ij} = \frac{\frac{f_j \cdot c_{ij}}{\sum_i c_{ij}}}{\sum_j \frac{f_j \cdot c_{ij}}{\sum_i c_{ij}}}$$

so that $\Sigma_j a_{ij} = 1$. Once the contribution of each bin was calculated, I used the median fluorescence level in each bin (m_j) as the value for all observations in that bin.

$$p_i = \sum_j a_{ij} \cdot m_j$$

f_j = fraction of the population that sorts to bin j c_{ij} = # of reads of construct i in bin j

a_{ij} = fraction of weighted construct i found in j p_i = reconstructed GFP value of construct i

3.5.13 Identifying experimental interfaces

The dimerization propensity for all sequences was reconstructed as above. Each mutation was calculated as a fraction of the wild-type association strength. Mutations were characterized as WT-like if they measured >75% WT, mild >50% WT, moderate >25% WT, severe < 25% WT. The percent disruption for each available mutation was averaged and characterized the same way. The helical wheel graphing function was adapted from heliquest (Gautier et al., 2008).

Table 3.1 Selection of wild-type sequences for experimental characterization

Human Single Pass Proteins in Uniprot	2,383
Wild-type with energy score below 0 kcal/mol	1,141
Sequences without proline	609
Sequences with the most common length (21)	454
Wild-type with energy score below -10 kcal/mol	363
Sequences with a “good” pattern (8 interfacial residues)	122
Sequences with only 1 or 0 polar residues	110
Sequences selected for experimental analysis (Best 100 energies)	100

Table 3.2 TMDs that insert poorly into the membrane

UniProt ID	Length	Sequence	No. of Replicates	Start Residue	Fold Difference*	Std Err
Q7Z410	17	VVATSLVVLTLGVLLAF	2	33	-1.00	0.00
Q9BVV7^	17	LIVVLFGISITGGLFYT	3	111	-1.00	0.00
P11362	17	CTGAFLISCMVGSVIVY	1	381	-1.00	NA
Q9Y5F7	17	VSLVAICFVSFGSFVAL	3	695	-1.00	0.00
A6NL88	19	VCGVISFALAVGVGAKVAF	1	189	-1.00	NA
Q6UY09^	19	IGILAVIAVASELGYFLCI	3	453	-0.99	0.00
O94898	19	IVIIVVVCCVVGTSLIWVI	3	808	-0.97	0.01
Q9NVM1^	21	GLYFVLGVCFGLLLTLCLLVI	3	29	-0.97	0.02
O14788^	21	MFVALLGLGLGQVVCSVALFF	3	48	-0.98	0.01
Q6UXE8^	21	ILLGLLCGALCGVVMGMIIVF	3	238	-1.00	0.00
O15533^	21	GLFLSAFLLLGLFKALGWAAV	3	415	-1.00	0.00
Q9H3T3^	21	VAAFVVGAVVSGFSVGWFVGL	2	604	-1.00	0.00
Q6UXM1	21	VVIIAVVCCVVGTSLVWVVII	3	810	-1.00	0.00

***(**Percent of sequence reads in M9 media after 36 hours of growth - Percent of sequence reads in LB media after 36 hours of growth) / Percent of sequence reads in LB media after 36 hours of growth

^Constructs that a dimerization value above 40% GpA,

Table 3.3 TMDs that dimerize organized al[alphabetically by UniProt ID

TM	UniProt ID	Length	Reconstructed RFU	%GpA	Std Err
LLAAGILGAGALIAGMCFIII	A1L1A6	21	33,414	70.74%	15.07%
GVISFALAVGVGAKVAF	A6NL88	17	27,429	58.07%	6.03%
IVVAMTAVGGSICVMLVVICL	F2Z333	21	18,261	38.66%	5.10%
IFQTLLLLTVVFGFLYGAMLY	O00461	21	36,969	78.27%	26.01%
MFVALLGLGLGQVVCSV	O14788	17	76,244	161.41%	10.57%
MFVALLGLGLGQVVCSVALFF	O14788	21	93,651	198.27%	18.43%
GLFLSAFLLLGLFKALG	O15533	17	16,612	35.17%	2.12%
GLFLSAFLLLGLFKALGWAAV	O15533	21	67,022	141.89%	14.79%
VGLGLLLLLMGAGLAVQGWFL	O43557	21	65,958	139.64%	24.21%
IILSLALAGILGICIVV	O43570	17	15,580	32.98%	0.61%
MLTLLGLSFILAGLIVGGACI	O43736	21	21,956	46.48%	12.78%
GASLLLAALLLGCLVAL	O60344	17	20,925	44.30%	4.75%
LAGASLLLAALLLGCLVAL	O60344	19	35,026	74.15%	29.79%
LVLAGASLLLAALLLGCLVAL	O60344	21	102,958	217.97%	23.07%
AVIVGGVVGALFAAFLV	O75056	17	47,158	99.84%	11.67%
AVIVGGVVGALFAAFLVTL	O75056	19	42,824	90.66%	17.53%
AVIVGGVVGALFAAFLVTLLI	O75056	21	87,496	185.24%	29.73%
GVVLLYILLGTIGTLVAVLAA	O95256	21	35,651	75.48%	9.94%
ALVSLLSVYVTGVCVAF	O95976	17	28,341	60.00%	15.08%
TALVSLLSVYVTGVCVAFI	O95976	19	37,514	79.42%	18.42%
GTIAVIVFFLIGFMIGY	P02786	17	18,870	39.95%	0.57%
YGTIAVIVFFLIGFMIGYL	P02786	19	61,445	130.08%	43.03%
YGTIAVIVFFLIGFMIGYLGY	P02786	21	53,858	114.02%	19.73%
VIVALAVCGSILFLLIV	P11117	17	21,440	45.39%	7.89%
VIVALAVCGSILFLLIVLLLT	P11117	21	81,485	172.51%	21.40%
WLLGAAMVGAVLTALLA	P14679	17	24,054	50.92%	1.57%
WLLGAAMVGAVLTALLAGLVS	P14679	21	42,945	90.92%	4.61%
FLLWILAAVSSGLFFYSFL	P16410	19	24,387	51.63%	11.49%
VVIIVILLILTGAGLAA	P22897	17	20,492	43.38%	5.78%
IITMSVVGGTLLLGIAICC	P53801	19	18,111	38.34%	19.49%
FSLSVMAALTFGCFITT	P61009	17	26,540	56.19%	7.04%
AFSLSVMAALTFGCFITTA	P61009	19	21,739	46.02%	4.56%
FAFSLSVMAALTFGCFITTAF	P61009	21	36,358	76.97%	6.21%
IMIIICCVVLGVVLASSIGGT	P61266	21	44,268	93.72%	14.47%

TM	UniProt ID	Length	Reconstructed RFU	%GpA	Std Err
FLVAFLLLGLVGMLVLF	P78357	17	20,054	42.46%	6.19%
LGFLVAFLLLGLVGMLVLF	P78357	19	29,083	61.57%	9.73%
ILLGFLVAFLLLGLVGMLVLF	P78357	21	72,515	153.52%	28.55%
LAAATLGVLCLGLVVTI	P78380	17	19,778	41.87%	2.83%
LAAATLGVLCLGLVVTIMV	P78380	19	23,054	48.81%	3.04%
ALVVFVGGTLVLGTILFLV	P78562	19	36,881	78.08%	6.28%
VALFAAVGAGCVIFLLIIIFL	P98172	21	43,938	93.02%	4.94%
ILCAVVVGVGLIVGLAVGL	Q07075	19	40,857	86.50%	3.69%
VAILCAVVVGVGLIVGLAVGL	Q07075	21	41,580	88.03%	10.81%
FFTWFMVIALLGVWTSV	Q12797	17	17,487	37.02%	5.35%
FFTWFMVIALLGVWTSVAV	Q12797	19	19,420	41.11%	9.26%
FFTWFMVIALLGVWTSVAVVW	Q12797	21	37,414	79.21%	12.50%
AVGILLTTLLVIGIILAVVFI	Q12864	21	25,858	54.74%	1.16%
IGLSVGAAVAYIIAVLG	Q13308	17	14,549	30.80%	4.14%
IGLSVGAAVAYIIAVLGLMFY	Q13308	21	31,591	66.88%	23.36%
GLLLSLLVLLVLVMLGASYWY	Q13444	21	14,210	30.08%	8.08%
MIAVGLSSSILGCLLTL	Q13591	17	92,583	196.01%	10.53%
HMIAVGLSSSILGCLLTLL	Q13591	19	64,219	135.96%	42.58%
FHMIAVGLSSSILGCLLTLLV	Q13591	21	103,753	219.65%	15.36%
VIIFFAFVLLLSGALAYCLAL	Q13651	21	61,595	130.40%	29.20%
VFLLAILGGMAFILLVL	Q5VUB5	17	27,014	57.19%	10.67%
VFLLAILGGMAFILLVLLCLL	Q5VUB5	21	86,275	182.65%	35.56%
IFYVTVLAFTLIVLTGGFTWL	Q5VV43	21	22,413	47.45%	10.12%
ALGIGIAIGVVGGVLVR	Q6P7N7	17	30,196	63.93%	13.46%
ALGIGIAIGVVGGVLVRIVLC	Q6P7N7	21	66,493	140.77%	14.25%
YIAIGIGVGLLLSGFLV	Q6UW88	17	57,211	121.12%	12.92%
YIAIGIGVGLLLSGFLVIFYC	Q6UW88	21	102,332	216.64%	22.81%
ILLGLLCGALCGVVMGM	Q6UXE8	17	66,750	141.31%	16.28%
ILLGLLCGALCGVVMGMII	Q6UXE8	19	86,283	182.67%	52.12%
ILLGLLCGALCGVVMGMIIVF	Q6UXE8	21	97,199	205.78%	27.34%
ILAVIAVASELGYFLCI	Q6UY09	17	14,862	31.46%	4.28%
IGILAVIAVASELGYFLCI	Q6UY09	19	34,632	73.32%	20.16%
IVIGILAVIAVASELGYFLCI	Q6UY09	21	71,102	150.53%	17.42%
LAIVAIIGIAIGIVTHF	Q6ZWK6	17	15,644	33.12%	2.23%
TLAIVAIIGIAIGIVTHFV	Q6ZWK6	19	37,041	78.42%	23.59%
FTLAIVAIIGIAIGIVTHFVV	Q6ZWK6	21	53,488	113.24%	4.66%

TM	UniProt ID	Length	Reconstructed RFU	%GpA	Std Err
SIGVVATSLVVLTLGVLLAFL	Q7Z410	21	21,740	46.03%	12.77%
IFIFLGVAAILGVTIGLLV	Q86T26	19	20,545	43.50%	14.45%
LIQSLIASGIAGSMIGV	Q86YA3	17	19,502	41.29%	8.91%
LIQSLIASGIAGSMIGVITLY	Q86YA3	21	18,815	39.83%	0.59%
LLFWSLVYCYCGLCASI	Q8IUS5	17	43,593	92.29%	10.73%
SLLFWSLVYCYCGLCASIHLL	Q8IUS5	21	95,231	201.61%	18.84%
LIILAVVGGVIGLLILI	Q8IWT1	17	17,956	38.01%	3.38%
TLIILAVVGGVIGLLILILLI	Q8IWT1	21	52,862	111.91%	8.59%
GLLICGSLALITGLTFAIF	Q8IYV9	19	23,785	50.36%	14.85%
LLGLLICGSLALITGLTFAIF	Q8IYV9	21	73,646	155.91%	19.17%
YSFSGAFLFSMGFLVAV	Q8N6P7	17	59,386	125.73%	13.22%
YSFSGAFLFSMGFLVAVLC	Q8N6P7	19	59,889	126.79%	39.81%
YSFSGAFLFSMGFLVAVLCYL	Q8N6P7	21	61,004	129.15%	17.96%
YIFLLLIGFCIFAAGTVAAWL	Q8N7C0	21	53,326	112.89%	9.29%
LTALLAVSFHSIGVVIMTS	Q8N7S6	19	58,597	124.05%	43.22%
CILTALLAVSFHSIGVVIMTS	Q8N7S6	21	74,729	158.21%	23.13%
VIIAGVVCGVVCIMMVVAAAY	Q8N967	21	44,717	94.67%	16.08%
ILLAVLLLLLCGVTAGC	Q8NDY8	17	21,430	45.37%	3.10%
LILLAVLLLLLCGVTAGCV	Q8NDY8	19	36,099	76.43%	17.99%
SSVAFSIALICGMAISY	Q8NEA5	17	24,393	51.64%	16.13%
LISSVAFSIALICGMAISY	Q8NEA5	19	20,867	44.18%	11.00%
VILISSVAFSIALICGMAISY	Q8NEA5	21	67,846	143.63%	20.66%
LITCVFAAFVLGAFIAG	Q8NFY4	17	60,694	128.49%	5.92%
VLITCVFAAFVLGAFIAGV	Q8NFY4	19	44,204	93.58%	28.82%
VLITCVFAAFVLGAFIAGVAV	Q8NFY4	21	113,551	240.40%	33.38%
LIFIIALGSIAGILFVTMI	Q8TAB3	19	19,697	41.70%	5.47%
LIFIIALGSIAGILFVTMIFV	Q8TAB3	21	59,119	125.16%	10.30%
IYIGAGICAGLALALIF	Q8TDQ0	17	29,324	62.08%	3.55%
IYIGAGICAGLALALIFGA	Q8TDQ0	19	23,284	49.29%	9.92%
IYIGAGICAGLALALIFGALI	Q8TDQ0	21	91,890	194.54%	22.98%
VMFFTLFALLAGTAVMIIAYH	Q8TEM1	21	17,972	38.05%	9.88%
IALTLVSLACILGVLLASG	Q92932	19	15,479	32.77%	6.44%
IALTLVSLACILGVLLASGLI	Q92932	21	24,325	51.50%	4.36%
LLLSSLLLLLLGLLVAI	Q9BY79	17	50,167	106.21%	10.65%
VLLLSSLLLLLLGLLVAIILA	Q9BY79	21	93,255	197.43%	19.85%
IWVYGVSGGAFLIMIFL	Q9C0I4	17	23,703	50.18%	4.66%

TM	UniProt ID	Length	Reconstructed RFU	%GpA	Std Err
LVSYSLAVLLLGCLLFL	Q9C0J1	17	15,367	32.53%	2.62%
CWLVSYSLAVLLLGCLLFL	Q9C0J1	19	25,088	53.11%	8.93%
RLCWLVSYSLAVLLLGCLLFL	Q9C0J1	21	50,846	107.64%	19.76%
LVIALACISFLFLGCLLFFVC	Q9H158	21	64,726	137.03%	17.63%
VFFVIATLVFGLFMGLV	Q9H1E5	17	17,238	36.49%	7.16%
VFFVIATLVFGLFMGLVLVVI	Q9H1E5	21	65,364	138.38%	24.13%
TVIIIVVVLLMGFVGAV	Q9H1U4	17	18,048	38.21%	3.45%
ILTVIIIVVVLLMGFVGAV	Q9H1U4	19	27,645	58.53%	17.37%
IIILTVIIIVVVLLMGFVGAV	Q9H1U4	21	50,126	106.12%	9.81%
LVLQLLSFMLLAGVLVAIL	Q9H2X3	19	40,306	85.33%	18.62%
ILTIVGTIAGIVILSMIIA	Q9H3R2	19	17,884	37.86%	1.69%
AVLGALGLLAGAGVGSW	Q9H3S3	17	14,172	30.00%	2.35%
AVLGALGLLAGAGVGSWLLVL	Q9H3S3	21	36,086	76.40%	41.55%
VVGAVVSGFSVGWFVGL	Q9H3T3	17	89,021	188.46%	26.09%
AFVVGAVVSGFSVGWFVGL	Q9H3T3	19	60,436	127.95%	41.22%
VAAFVVGAVVSGFSVGWFVGL	Q9H3T3	21	70,316	148.86%	18.76%
VILIAAVGGGVLLLSALGLII	Q9H5V8	21	17,630	37.32%	10.12%
ILFAVIILTGASFAHLF	Q9NPU4	17	17,257	36.53%	5.76%
ILFAVIILTGASFAHLFVALF	Q9NPU4	21	95,771	202.76%	19.12%
GLYFVLGVCFGLLLTLC	Q9NVM1	17	115,044	243.56%	23.25%
GLYFVLGVCFGLLLTLCLL	Q9NVM1	19	87,174	184.55%	71.93%
GLYFVLGVCFGLLLTLCLLVI	Q9NVM1	21	84,609	179.12%	19.47%
VGAVLAAGALLGLVAGA	Q9NY15	17	22,693	48.04%	12.39%
GVGAVLAAGALLGLVAGAL	Q9NY15	19	31,078	65.79%	21.15%
ILLILCVGMVVGLVALG	Q9P126	17	20,437	43.27%	4.05%
ALILLILCVGMVVGLVALGIW	Q9P126	21	27,141	57.46%	5.01%
FLIMFLTIIVCGMVAAL	Q9UHP7	17	33,707	71.36%	12.07%
FFLIMFLTIIVCGMVAALS	Q9UHP7	19	16,022	33.92%	5.65%
FFLIMFLTIIVCGMVAALSAI	Q9UHP7	21	35,617	75.40%	13.25%
LTVIVTACLTFATGVTVALVM	Q9UJ14	21	35,273	74.68%	1.53%
WLLLLLTAACSGLLFAL	Q9UJ37	17	34,982	74.06%	8.13%
FFWLLLLLTAACSGLLFALYF	Q9UJ37	21	47,680	100.94%	10.42%
ILLIMIFYACLAGGLILAY	Q9UJ90	19	20,868	44.18%	7.20%
LALVIAISMGFGHFYGT	Q9ULG6	17	45,688	96.72%	6.82%
ILALVIAISMGFGHFYGTI	Q9ULG6	19	40,958	86.71%	21.76%
CVILALVIAISMGFGHFYGTI	Q9ULG6	21	45,245	95.79%	4.00%

TM	UniProt ID	Length	Reconstructed RFU	%GpA	Std Err
ITVVIAAAGGGLLLILGIALI	Q9ULI3	21	21,020	44.50%	8.27%
LLLSLILVSVGFVVTVFGV	Q9UN70	19	16,532	35.00%	5.90%
LLLSLILVSVGFVVTVFGVII	Q9UN70	21	48,945	103.62%	15.50%
CMCFGLAFMLAGVILGG	Q9Y287	17	19,427	41.13%	6.70%
WCMCFGLAFMLAGVILGGAYL	Q9Y287	21	31,988	67.72%	4.30%
LVGMAIVGGMALGVAGL	Q9Y3D6	17	18,244	38.62%	6.29%
LVGMAIVGGMALGVAGLAGLI	Q9Y3D6	21	17,641	37.35%	6.44%
AVSLVAICFVSFGSFVALL	Q9Y5F7	19	18,747	39.69%	22.06%
LAVSLVAICFVSFGSFVALLS	Q9Y5F7	21	37,845	80.12%	15.56%
VILYLMVMIGMFSFIIV	Q9Y6J6	17	15,023	31.80%	4.44%

3.6 Supplemental Information

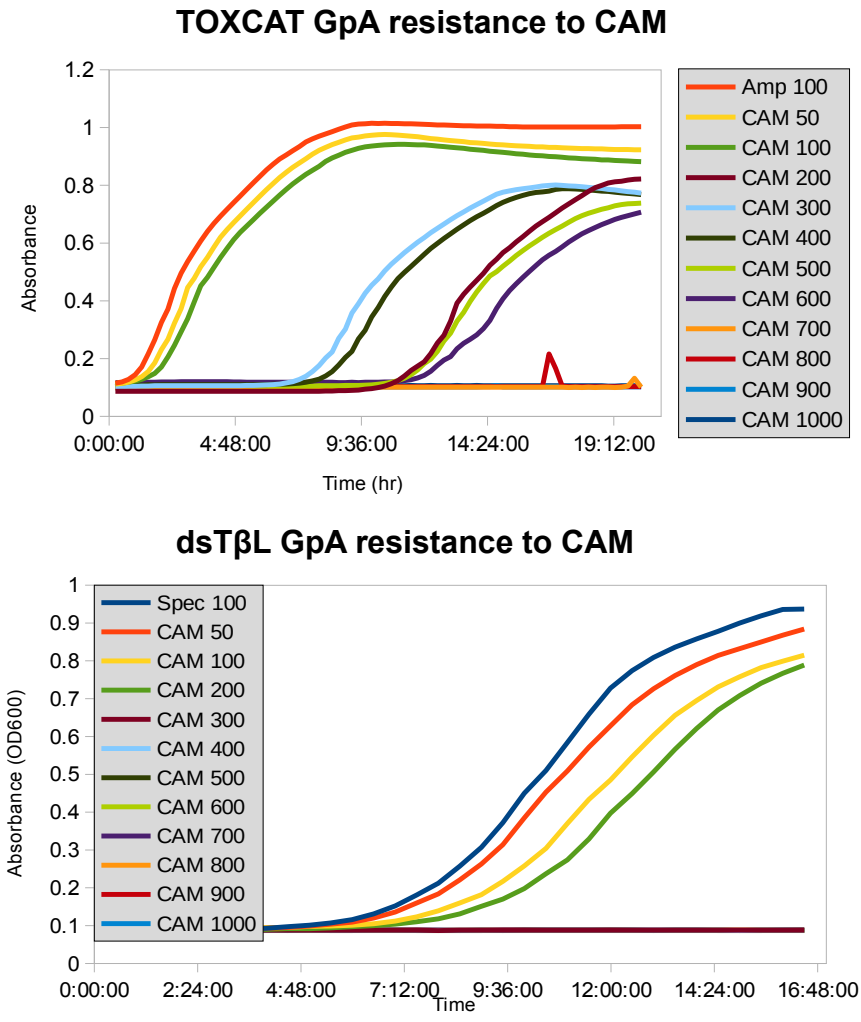
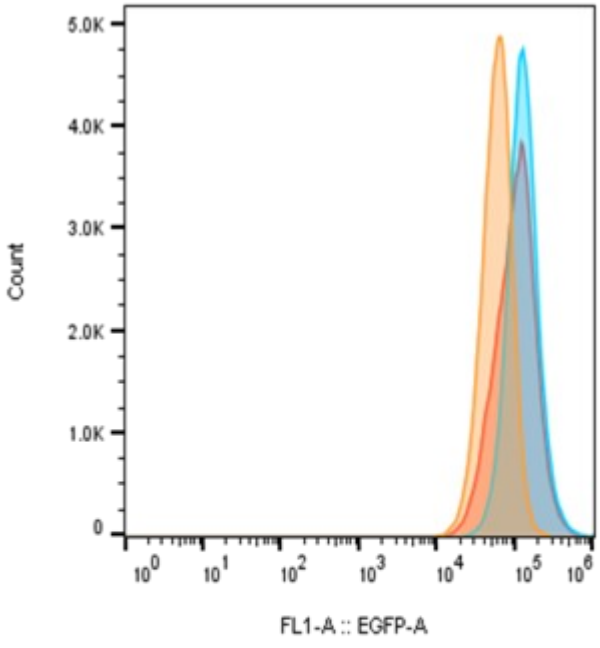


Figure S3.1 Antibiotic resistance ranges for ToxR assays. Growth curves of GpA controls in the TOXCAT and dsTβL systems. (Top) TOXCAT plasmids have ampicillin resistance which is used as the positive control. Increasing concentrations of chloramphenicol (CAM) are added to the growth conditions in each line. The construct can continue to grow at some level all the way up to 600 µg/mL CAM. (Bottom) dsTβL plasmids have ampicillin resistance which is used as the positive control. In this system, GpA can only survive in up to 200 µg/mL CAM, a reduced reporter gene range.



	Sample Name	Median : FL1-A
	NoTM2-Spec - 1_Data Source - 1.fcs	57448
	G83I-1-Spec - 1_Data Source - 1.fcs	120654
	GpA1-Spec - 1_Data Source - 1.fcs	101244

Figure S3.2 Control measurements for the dsTβL assay. After swapping the reporter gene CAT for sfGFP, I performed flow cytometry on each of the control constructs. As displayed here, the positive control GpA has a slightly lower median than the monomerizing control, indicating that the modified dsTβL system could not differentiate between monomers and dimers.

3.7 Acknowledgments

The work was supported by National Science Foundation grants CHE-1415910 and CHE-1710182. S.M.A. acknowledge the support of the NLM training grant 5T15LM007359 to the CIBM Training Program. I are grateful to Samson Condon, Megan Leander, Nick Hoppe, and Kyle Nishikawa for helpful suggestions and discussion.

3.8 References

Adair, B.D., and Engelman, D.M. (1994). Glycophorin A helical transmembrane domains dimerize in phospholipid bilayers: a resonance energy transfer study. Biochemistry *33*, 5539–5544.

Anderson, S.M., Mueller, B.K., Lange, E.J., and Senes, A. (2017). Combination of Cα-H Hydrogen Bonds and van der Waals Packing Modulates the Stability of GxxxG-Mediated Dimers in Membranes. J. Am. Chem. Soc. *139*, 15774–15783.

Araya, C.L., and Fowler, D.M. (2011). Deep mutational scanning: assessing protein function on a massive scale. Trends Biotechnol. *29*, 435–442.

Armstrong, C.R., and Senes, A. (2016). Screening for transmembrane association in divisome proteins using TOXGREEN, a high-throughput variant of the TOXCAT assay. Biochim. Biophys. Acta *1858*, 2573–2583.

Auguie, B. (2017). gridExtra: Miscellaneous Functions for “Grid” Graphics.

Bennasroune, A., Gardin, A., Auzan, C., Clauser, E., Dirrig-Grosch, S., Meira, M., Appert-Collin, A., Aunis, D., Crémel, G., and Hubert, P. (2005). Inhibition by transmembrane peptides of chimeric insulin receptors. Cell. Mol. Life Sci. CMLS *62*, 2124–2131.

Berger, B.W., Kulp, D.W., Span, L.M., DeGrado, J.L., Billings, P.C., Senes, A., Bennett, J.S., and DeGrado, W.F. (2010). Consensus motif for integrin transmembrane helix association. Proc. Natl. Acad. Sci. U. S. A. *107*, 703–708.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Bocharov, E.V., Mineev, K.S., Pavlov, K.V., Akimov, S.A., Kuznetsov, A.S., Efremov, R.G., and Arseniev, A.S. (2017). Helix-helix interactions in membrane domains of bitopic proteins: Specificity and role of lipid environment. Biochim. Biophys. Acta Biomembr. *1859*, 561–576.

Cheung, R., Insigne, K.D., Yao, D., Burghard, C.P., Wang, J., Hsiao, Y.-H.E., Jones, E.M., Goodman, D.B., Xiao, X., and Kosuri, S. (2019). A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions. Mol. Cell *73*, 183-194.e8.

Deming, W.E. (1943). Statistical adjustment of data. (New York: J. Wiley & Sons, Incorporated).

Dews, I.C., and Mackenzie, K.R. (2007). Transmembrane domains of the syndecan family of growth factor coreceptors display a hierarchy of homotypic and heterotypic interactions. Proc. Natl. Acad. Sci. U. S. A. *104*, 20782–20787.

Doura, A.K., and Fleming, K.G. (2004). Complex interactions at the helix-helix interface stabilize the glycophorin A transmembrane dimer. J. Mol. Biol. *343*, 1487–1497.

Dowle, M., and Srinivasan, A. (2019). data.table: Extension of `data.frame`.

Duong, M.T., Jaszewski, T.M., Fleming, K.G., and MacKenzie, K.R. (2007). Changes in apparent free energy of helix-helix dimerization in a biological membrane due to point mutations. J. Mol. Biol. *371*, 422–434.

Elazar, A., Weinstein, J., Biran, I., Fridman, Y., Bibi, E., and Fleishman, S.J. (2016). Mutational scanning reveals the determinants of protein insertion and association energetics in the plasma membrane. ELife *5*, e12125.

Fisher, L.E., Engelman, D.M., and Sturgis, J.N. (2003). Effect of detergents on the association of the glycophorin a transmembrane helix. Biophys. J. *85*, 3097–3105.

Fleming, K.G., Ackerman, A.L., and Engelman, D.M. (1997). The effect of point mutations on the free energy of transmembrane alpha-helix dimerization. J. Mol. Biol. *272*, 266–275.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat. Methods *11*, 801–807.

Fujii, S., Matsuura, T., Sunami, T., Nishikawa, T., Kazuta, Y., and Yomo, T. (2014). Liposome display for in vitro selection and evolution of membrane proteins. Nat. Protoc. *9*, 1578–1591.

Gautier, R., Douguet, D., Antonny, B., and Drin, G. (2008). HELIQUEST: a web server to screen sequences with specific alpha-helical properties. Bioinforma. Oxf. Engl. *24*, 2101–2102.

Grau, B., Javanainen, M., García-Murria, M.J., Kulig, W., Vattulainen, I., Mingarro, I., and Martínez-Gil, L. (2017). The role of hydrophobic matching on transmembrane helix packing in cells. Cell Stress *1*, 90–106.

Grothendieck, G. (2018). gsubfn: Utilities for Strings and Function Arguments.

Hénin, J., Pohorille, A., and Chipot, C. (2005). Insights into the recognition and association of transmembrane alpha-helices. The free energy of alpha-helix dimerization in glycophorin A. J. Am. Chem. Soc. *127*, 8478–8484.

Hester, J., Csárdi, G., Wickham, H., Chang, W., Morgan, M., and Tenenbaum, D. (2019). remotes: R Package Installation from Remote Repositories, Including “GitHub”.

Hibbing, M.E., Fuqua, C., Parsek, M.R., and Peterson, S.B. (2010). Bacterial competition: surviving and thriving in the microbial jungle. Nat. Rev. Microbiol. *8*, 15–25.

Holmqvist, E., Reimegård, J., and Wagner, E.G.H. (2013). Massive functional mapping of a 5’-UTR by saturation mutagenesis, phenotypic sorting and deep sequencing. Nucleic Acids Res. *41*, e122.

Huang, H., Kuenze, G., Smith, J.A., Taylor, K.C., Duran, A.M., Hadziselimovic, A., Meiler, J., Vanoye, C.G., George, A.L., and Sanders, C.R. (2018). Mechanisms of KCNQ1 channel dysfunction in long QT syndrome involving voltage sensor domain mutations. Sci. Adv. *4*, eaar2631.

Hubert, P., Sawma, P., Duneau, J.-P., Khao, J., Hénin, J., Bagnard, D., and Sturgis, J. (2010). Single-spanning transmembrane domains in cell growth and cell-cell interactions: More than meets the eye? Cell Adhes. Migr. *4*, 313–324.

Joce, C., Wiener, A.A., and Yin, H. (2011). Multi-Tox: application of the ToxR-transcriptional reporter assay to the study of multi-pass protein transmembrane domain oligomerization. Biochim. Biophys. Acta *1808*, 2948–2953.

Julius, A., Laur, L., Schanzenbach, C., and Langosch, D. (2017). BLaTM 2.0, a Genetic Tool Revealing Preferred Antiparallel Interaction of Transmembrane Helix 4 of the Dual-Topology Protein EmrE. J. Mol. Biol. *429*, 1630–1637.

Khadria, A.S., Mueller, B.K., Stefely, J.A., Tan, C.H., Pagliarini, D.J., and Senes, A. (2014). A Gly-zipper motif mediates homodimerization of the transmembrane domain of the mitochondrial kinase ADCK3. J. Am. Chem. Soc. *136*, 14068–14077.

Kinney, J.B., Murugan, A., Callan, C.G., and Cox, E.C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc. Natl. Acad. Sci. U. S. A. *107*, 9158–9163.

Kosuri, S., Eroshenko, N., Leproust, E.M., Super, M., Way, J., Li, J.B., and Church, G.M. (2010). Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. Nat. Biotechnol. *28*, 1295–1299.

Kosuri, S., Goodman, D.B., Cambray, G., Mutalik, V.K., Gao, Y., Arkin, A.P., Endy, D., and Church, G.M. (2013). Composability of regulatory sequences controlling transcription and translation in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. *110*, 14024–14029.

Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778–795.

Kulp, D.W., Subramaniam, S., Donald, J.E., Hannigan, B.T., Mueller, B.K., Grigoryan, G., and Senes, A. (2012). Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL). J. Comput. Chem. *33*, 1645–1661.

Langosch, D., Brosig, B., Kolmar, H., and Fritz, H.J. (1996). Dimerisation of the glycophorin A transmembrane segment in membranes probed with the ToxR transcription activator. J. Mol. Biol. *263*, 525–530.

LaPointe, L.M., Taylor, K.C., Subramaniam, S., Khadria, A., Rayment, I., and Senes, A. (2013). Structural organization of FtsB, a transmembrane protein of the bacterial divisome. Biochemistry *52*, 2574–2585.

Lawrie, C.M., Sulistijo, E.S., and MacKenzie, K.R. (2010). Intermonomer hydrogen bonds enhance GxxxG-driven dimerization of the BNIP3 transmembrane domain: roles for sequence context in helix-helix association in membranes. J. Mol. Biol. *396*, 924–936.

Lazaridis, T. (2003). Effective energy function for proteins in lipid membranes. Proteins *52*, 176–192.

Lis, M., and Blumenthal, K. (2006). A modified, dual reporter TOXCAT system for monitoring homodimerization of transmembrane segments of proteins. Biochem. Biophys. Res. Commun. *339*, 321–324.

MacKenzie, K.R., Prestegard, J.H., and Engelman, D.M. (1997). A transmembrane helix dimer: structure and implications. Science *276*, 131–133.

MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B *102*, 3586–3616.

Manuilova, E., Schuetzenmeister, A., and Model, F. (2014). mcr: Method Comparison Regression.

McLaughlin, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138–142.

Mueller, B.K., Subramaniam, S., and Senes, A. (2014). A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα-H hydrogen bonds. Proc. Natl. Acad. Sci. U. S. A. *111*, E888-895.

Müller, C.A., Hawkins, M., Retkute, R., Malla, S., Wilson, R., Blythe, M.J., Nakato, R., Komata, M., Shirahige, K., de Moura, A.P.S., et al. (2014). The dynamics of genome replication using deep sequencing. Nucleic Acids Res. *42*, e3.

Ouellette, S.P., Karimova, G., Davi, M., and Ladant, D. (2017). Analysis of Membrane Protein Interactions with a Bacterial Adenylate Cyclase-Based Two-Hybrid (BACTH) Technique. Curr. Protoc. Mol. Biol. *118*, 20.12.1-20.12.24.

Peterman, N., and Levine, E. (2016). Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. BMC Genomics *17*, 206.

Peterman, N., Lavi-Itzkovitz, A., and Levine, E. (2014). Large-scale mapping of sequence-function relations in small regulatory RNAs reveals plasticity and modularity. Nucleic Acids Res. *42*, 12177–12188.

Popot, J.L., and Engelman, D.M. (1990). Membrane protein folding and oligomerization: the two-stage model. Biochemistry *29*, 4031–4037.

Ranjit, D.K., and Young, K.D. (2013). The Rcs stress response and accessory envelope proteins are required for de novo generation of cell shape in Escherichia coli. J. Bacteriol. *195*, 2452–2462.

Rohlhill, J., Sandoval, N.R., and Papoutsakis, E.T. (2017). Sort-Seq Approach to Engineering a Formaldehyde-Inducible Promoter for Dynamically Regulated Escherichia coli Growth on Methanol. ACS Synth. Biol. *6*, 1584–1595.

Roskoski, R. (2014). The ErbB/HER family of protein-tyrosine kinases and cancer. Pharmacol. Res. *79*, 34–74.

RStudio Team (2016). RStudio: Integrated Development for R. (Boston, MA: RStudio, Inc.).

Russ, W.P., and Engelman, D.M. (1999). TOXCAT: a measure of transmembrane helix association in a biological membrane. Proc. Natl. Acad. Sci. U. S. A. *96*, 863–868.

Russ, W.P., and Engelman, D.M. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. J. Mol. Biol. *296*, 911–919.

Sarkar, C.A., Dodevski, I., Kenig, M., Dudli, S., Mohr, A., Hermans, E., and Plückthun, A. (2008). Directed evolution of a G protein-coupled receptor for expression, stability, and binding selectivity. Proc. Natl. Acad. Sci. U. S. A. *105*, 14808–14813.

Schlinkmann, K.M., Honegger, A., Türeci, E., Robison, K.E., Lipovšek, D., and Plückthun, A. (2012). Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations. Proc. Natl. Acad. Sci. U. S. A. *109*, 9810–9815.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671–675.

Schütz, M., Schöppe, J., Sedlák, E., Hillenbrand, M., Nagy-Davidescu, G., Ehrenmann, J., Klenk, C., Egloff, P., Kummer, L., and Plückthun, A. (2016). Directed evolution of G protein-coupled receptors in yeast for higher functional production in eukaryotic expression hosts. Sci. Rep. *6*, 21508.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

Sievert, C. (2018). plotly for R.

Starr, T.N., Picton, L.K., and Thornton, J.W. (2017). Alternative evolutionary histories in the sequence space of an ancient protein. Nature *549*, 409–413.

Su, P.-C., and Berger, B.W. (2012). Identifying Key Juxtamembrane Interactions in Cell Membranes Using AraC-based Transcriptional Reporter Assay (AraTM). J. Biol. Chem. *287*, 31515–31526.

Subramaniam, S., and Senes, A. (2012). An energy-based conformer library for side chain optimization: improved prediction and adjustable sampling. Proteins *80*, 2218–2234.

Sulistijo, E.S., Jaszewski, T.M., and MacKenzie, K.R. (2003). Sequence-specific dimerization of the transmembrane domain of the “BH3-only” protein BNIP3 in membranes and detergent. J. Biol. Chem. *278*, 51950–51956.

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.

Wadhwa, R., Subramanian, V., and Stevens-Truss (2018). Visualizing alpha-helical peptides in R with helixvis. *3*, 10008.

Wickham, H. (2007). Reshaping Data with the reshape Package. Journal of Statistical Software *21*, 1–20.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag New York).

Wickham, H. (2017). tidyverse: Easily Install and Load the “Tidyverse”.

Wickham, H. (2018). scales: Scale Functions for Visualization.

Wickham, H., François, R., Henry, L., and Müller, K. (2019). dplyr: A Grammar of Data Manipulation.

Zhang, J., and Lazaridis, T. (2006). Calculating the free energy of association of transmembrane helices. Biophys. J. *91*, 1710–1723.

Chapter 4: The cytokinin oxidase/dehydrogenase CKX1 is a membrane-bound protein requiring homooligomerization in the endoplasmic reticulum for its cellular activity

This chapter was prepared for publication as:

Michael C.E. Niemann, Henriette Weber, Tomáš Hluska, Georgeta Leonte, Samantha M. Anderson, Ondřej Novák, Alessandro Senes, and Tomáš Werner. “The cytokinin oxidase/dehydrogenase CKX1 is a membrane-bound protein requiring homooligomerization in the endoplasmic reticulum for its cellular activity” Plant Physiology 2018 179, 2024-39

My contribution is the computational model of CKX1 in figures 4.6 and S4.6 and the identification of mutations that are critical to CKX1 dimerization.

4.1 Abstract

Degradation of the plant hormone cytokinin is controlled by cytokinin oxidase/dehydrogenase (CKX) enzymes. The molecular and cellular behavior of these proteins is still largely unknown. In this study, we show that CKX1 is a type II single-pass membrane protein that localizes predominantly to the endoplasmic reticulum (ER) in Arabidopsis (*Arabidopsis thaliana*). This indicates that this CKX isoform is a bona fide ER protein directly controlling the cytokinin, which triggers the signaling from the ER. By using various approaches, we demonstrate that CKX1 forms homodimers and homooligomers *in vivo*. The amino-terminal part of CKX1 was necessary and sufficient for the protein oligomerization as well as for targeting and retention in the ER. Moreover, we show that protein-protein interaction is largely facilitated by transmembrane helices and depends on a functional GxxxG-like interaction motif. Importantly, mutations rendering CKX1 monomeric interfere with its steady-state localization in the ER and cause a loss of the CKX1 biological activity by increasing its ER-associated degradation. Therefore, our study provides evidence that oligomerization is a crucial parameter regulating CKX1 biological activity and the cytokinin concentration in the ER. The work also lends strong support for the cytokinin signaling from the ER and for the functional relevance of the cytokinin pool in this compartment.

4.2 Introduction

Cytokinin is a plant hormone involved in a wide range of biological processes from cell proliferation and differentiation, tissue patterning and organ initiation, to physiological responses to the environment(Hwang et al., 2012; Werner and Schmülling, 2009). Cytokinin concentrations need to be dynamically adjusted in different cell types to optimize developmental processes and growth. An important mechanism contributing to this regulation is the irreversible metabolic degradation catalyzed by cytokinin oxidase/dehydrogenase (CKX) enzymes encoded by a small gene family comprising seven members in Arabidopsis (*Arabidopsis thaliana*; (Schmülling et al., 2003). Individual CKX isoforms are expressed in different tissues (Bartrina et al., 2011; Köllmer et al., 2014; Werner et al., 2003) and differ partially in their substrate specificities (Galuszka et al., 2007; Kowalska et al., 2010). It has been proposed that individual CKX proteins also control partly different cellular cytokinin pools depending on their subcellular localizations. Whereas CKX7 apparently is the only Arabidopsis isoform localized to the cytosol (Köllmer et al., 2014), CKX1 to CKX6 contain a highly hydrophobic N-terminal domain serving as a target sequence for their import to the endoplasmic reticulum (ER; (Schmülling et al., 2003; Werner et al., 2003).

The ER is the entry compartment of the secretory pathway consisting of multiple organelles with distinct morphologies and functions. Secretory proteins usually enter the ER cotranslationally. Translocation of soluble proteins is directed by a cleavable N-terminal signal peptide. Integral membrane proteins possess one or more hydrophobic transmembrane (TM)-spanning regions that are inserted into the membrane of the ER. Single-pass type I membrane proteins have a cleavable signal peptide and a separate TM domain and are oriented with their N terminus in the lumen of an organelle and their C terminus in the cytoplasm. By contrast, type II membrane proteins have an uncleavable signal anchor close to the N terminus serving both as a targeting signal and as a TM helix. This type of protein exhibits an opposite topology (i.e. the C terminus faces the organellar lumen and the N terminus faces the cytosol; (van Anken and

Braakman, 2005; von Heijne and Gavel, 1988). Soluble and membrane cargo proteins are transported bidirectionally between organelles in membranous vesicles that are generated by cytoplasmic coat proteins (Schekman and Orci, 1996). In the biosynthetic anterograde pathway, proteins shuttle from the ER through the Golgi apparatus to reach the plasma membrane and extracellular space or vacuoles. The retrograde traffic facilitates endocytosis and the recycling of membranes and directs resident proteins back to their original compartments (Rojo and Denecke, 2008). Protein sorting into different compartments is generally mediated by sorting determinants contained within the cargo proteins themselves. These sorting determinants consist of either short conserved amino acid sequences, which are recognized by respective sorting receptors, or of physical properties such as the length and hydrophobicity of the TM span (Chevalier and Chaumont, 2015; Cosson et al., 2013; De Marcos Lousa et al., 2012; Gao et al., 2014; Jurgens, 2004).

The distribution and fate of individual CKX isoforms within the secretory system are not well understood. Some CKX isoforms, such as CKX2, were shown to be secreted from cells in heterologous yeast expression systems (Bilyeu et al., 2001; Werner et al., 2001), providing indirect evidence that these isoforms might be targeted to the apoplast in plants. This hypothesis is supported by direct localization of a CKX homolog from maize (*Zea mays*) to the apoplast (Galuszka et al., 2005). In contrast, the localization of CKX proteins to intracellular compartments of the secretory pathway has been proposed based on the localization of CKX1-GFP and CKX3-GFP reporter proteins to the ER and, occasionally, to vacuoles (Werner et al., 2003). Intriguingly, the ER localization of CKX1 has been indirectly supported by the apparent absence of hybrid and complex *N*-glycans (Niemann et al., 2015), which are synthesized on *N*-glycoproteins upon their arrival to the Golgi apparatus (von Schaewen et al., 1993).

The cytokinin signal is perceived by sensor His kinases (Inoue et al., 2001; Suzuki et al., 2001), which are localized predominantly to the ER membrane and, to a lesser extent, to the plasma membrane (Caesar et al., 2011; Lomin et al., 2011; Wulfetange et al., 2011), indicating

that the steady-state cytokinin concentration in the ER lumen and apoplast might determine cellular responses to the hormone. Given the possibility that the sites of cytokinin degradation and perception might spatially overlap within the secretory system, it is essential to understand the precise distribution of CKX proteins and the underlying sorting mechanisms. Moreover, a recent study revealing the relevance of endoplasmic reticulum-associated degradation (ERAD; (Römisch, 2005) for the control of CKX protein levels and plant development (Niemann et al., 2015) has demonstrated the necessity to explore molecular mechanisms controlling CKX proteins in the secretory system.

Here, we show that Arabidopsis CKX1 is an intrinsic membrane protein localized predominantly to the ER and forming homooligomeric complexes. Our data provide insights into oligomerization mechanisms mediated largely by the TM helices and indicate the functional relevance of complex formation for subcellular localization and protein stability. We propose that some CKX isoforms, including CKX1 as a case example, act as authentic ER-resident proteins to control the cytokinin homeostatic levels in the ER lumen, thereby directly regulating the signaling output from this compartment.

4.3 Results

4.3.1 CKX1 Is a Type II Integral Membrane Protein

Our previous experimental work (Niemann et al., 2015) has suggested that, potentially, not all CKX proteins are soluble proteins, as is generally assumed. In order to directly test the solubility and possible membrane association, we focused on Arabidopsis CKX1 and prepared microsomal membranes from *Nicotiana benthamiana* leaves transiently expressing *35S*:*myc-CKX1* (Niemann et al., 2015). The protein gel-blot analysis showed that myc-CKX1 was detected exclusively in the 100,000*g* pellet fraction but not in the supernatant, indicating that it is either a membrane-associated protein or a soluble luminal protein (Fig. 4.1A). To differentiate between these possibilities, the microsomal vesicles were subjected to differential solubilization. As shown in Figure 4.1A, treatment with Na2CO3 (pH 11), which converts closed microsomal vesicles to open membrane sheets, thereby releasing the soluble luminal proteins and peripheral membrane proteins (Fujiki et al., 1982; Mothes et al., 1997), did not release myc-CKX1 from microsomal vesicles. Treatment with urea, which also extracts peripheral and luminal proteins (Obrdlik et al., 2000; Schook et al., 1979), was similarly ineffective, indicating that myc-CKX1 is not a soluble luminal protein but rather is tightly anchored within the membrane. In accord with this conclusion, myc-CKX1 could only be displaced from membranes by Triton X-100 treatment (Fig. 4.1A), indicating that CKX1 is a true integral membrane protein. Consistent with this experimental result, the SignalP 4.0 algorithm discriminating signal peptides from TM regions (Petersen et al., 2011) predicts the presence of an uncleavable signal anchor for CKX1 rather than a cleavable signal peptide (Table 4.1). The TM-spanning region is predicted to comprise amino acid residues 14 to 34 (Table 4.1). A similar prediction was obtained for CKX3 and CKX6, whereas the other CKX proteins with higher probability possess signal peptides (Table 4.1).

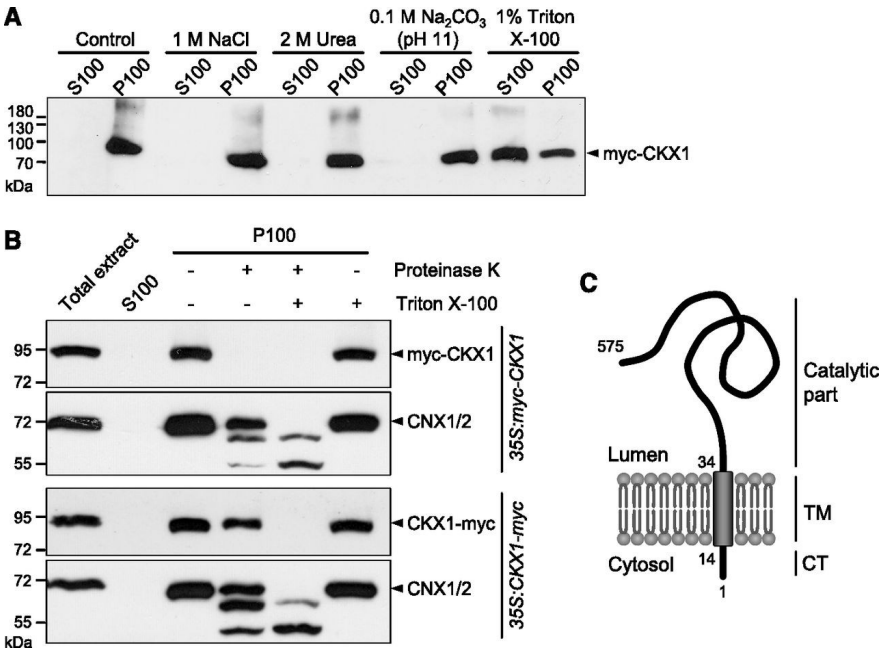


Figure 4.1 Membrane association and membrane topology of the myc-fused CKX1 proteins. A, Association of myc-CKX1 with membranes. Total membranes isolated from *N. benthamiana* leaves transiently expressing *35S*:*myc-CKX1* were incubated in homogenization buffer (control) or treated with 1 m NaCl, chaotropic agent (Urea), alkaline solution, or Triton X-100. Soluble (supernatant; S100) and insoluble (pellet; P100) membrane protein fractions were separated by recentrifugation at 100,000*g*. The proteins were resolved by SDS-PAGE and submitted to immunoblot analysis with anti-myc antibody. B, Proteinase K protection assay to reveal the membrane topology of CKX1: the C terminus of CKX1 is protected from proteinase K digestion. Microsomal membranes of *35S*:*myc-CKX1*- and *35S*:*CKX1-myc*-expressing Arabidopsis plants were incubated with proteinase K in the absence (lane 4) or presence (lane 5) of 1% Triton X-100. The samples were analyzed by SDS-PAGE and immunoblot with anti-myc antibody. Control immunoblot analysis using antibody against the luminal domain of the ER-localized, membrane-bound CALNEXIN1 and CALNEXIN2 (CNX1/2) was performed to monitor the integrity of the microsomal vesicles. C, Model of CKX1 topology in the membrane. Type II membrane protein topology included a predicted short cytosolic tail (CT), a single membrane-spanning helix (TM), and a lumen-oriented catalytic domain. Numbers indicate amino acid residues.

Table 4.1. Bioinformatic prediction to discriminate signal peptide and signal anchor sequences in Arabidopsis CKX proteins.

Protein	SignalP 4.1 (D-Score)[a]	Signal Peptide (P)/ Signal Anchor (A)	TM Domain (Residues)[b]
CKX1 (At2g41510.1)	0.302	A	14–34
CKX2 (At2g19500.1)	0.738	P	–
CKX3 (At5g56970.1)	0.332	A	9–29
CKX4 (At4g29740.1)	0.683	P	–
CKX5[c] (At1g75450.1)	0.729	P	–
CKX5[c] (At1g75450.2)	0.775	P	–
CKX6 (At3g63440.1)	0.261	A	16–36

a) Signal peptide/anchor prediction was performed with SignalP 4.1 (http://www.cbs.dtu.dk/services/SignalP/). D-score indicates the likelihood of signal peptide cleavage. b) TM domains were predicted by ConPred_v2 at the Aramemnon database (Schwacke et al., 2003) c) Different translation starts resulting in different N termini are predicted in TAIR (http://www.arabidopsis.org/).

To determine the CKX1 topology in the membrane, we performed a proteinase K protection assay with microsomal vesicles isolated from Arabidopsis plants stably expressing either the *35S*:*myc-CKX1* gene (Niemann et al., 2015) or CKX1 fused to a myc tag at the C terminus (*35S*:*CKX1-myc*). Proteinase K digestion typically results in the proteolytic removal of polypeptides that are exposed on the cytoplasmic face of microsomal vesicles. Immunoblot analysis revealed that proteinase K treatment of intact 35S:myc-CKX1 microsomes in the absence of detergent led to a complete loss of the myc signal (Fig. 4.1B), indicating that the N-terminally fused myc tag of the myc-CKX1 chimera was localized on the cytosolic membrane face. The proteolysis of ER chaperone proteins, CNX1/2, was used as a positive control in this proteolysis protection assay. Calnexins are conserved single-pass membrane proteins consisting of a large luminal domain and short tail facing the cytosol (Huang et al., 1993). Consistent with their topology, immunoblot analysis of proteinase K-treated microsomes showed that CNX1/2 were largely protected from the proteolysis, indicating high integrity of the

microsomal membrane preparations. In contrast to myc-CKX1, the CKX1-myc fusion protein in microsomes isolated from *35S*:*CKX1-myc*-expressing plants was largely protected from proteinase K activity and, similar to CNX1/2, was fully susceptible to proteolysis only in the presence of the detergent, indicating that the myc-tagged C terminus of CKX1 was localized in the microsomal lumen. Together, our results show that CKX1 is a typical type II membrane protein with topology consisting of a short cytoplasmic tail, a single TM region, and a large, luminally oriented, C-terminal catalytic domain (Fig. 4.1C).

4.3.2 CKX1 Forms Homooligomeric Complexes

Immunoblot analyses of protein extracts from *35S*:*CKX1* plants repeatedly showed that CKX1 partly runs as an SDS-resistant complex of higher molecular mass on a reducing SDS-PAGE gel (Fig. 4.1A), which indicates potential tight interactions with other proteins or protein homooligomerization. To analyze specifically whether CKX1 can homooligomerize, GFP-CKX1 and myc-CKX1 fusion proteins were transiently coexpressed in *N. benthamiana* leaves, and total protein extracts were used for coimmunoprecipitation (Co-IP) assay with an anti-GFP antibody. As shown in Figure 4.2A, myc-CKX1 was detected robustly in the GFP-CKX1 immunocomplex, but it did not coimmunoprecipitate with GFP alone, strongly supporting the notion of CKX1 homodimerization *in vivo*. Interestingly, Figure 4.2A shows that, in contrast to the input samples, the myc-CKX1 signal of high molecular mass was the most prevalent in the Co-IP fraction. This suggests the formation of a higher order oligomer that was not fully resolved under SDS-PAGE conditions.

To examine the CKX1 oligomerization in more detail, we subjected protein extracts from *N. benthamiana* expressing myc-CKX1 to size exclusion chromatography (SEC; Fig. 4.2, B–D). Given the strong membrane association of myc-CKX1 (Fig. 4.1), the membrane proteins were solubilized with the nonionic detergent DDM, and the detergent was included in the chromatography buffer during SEC fractionation at a concentration above the critical micelle concentration. Solubilization of membranes by detergents converts intrinsic membrane proteins

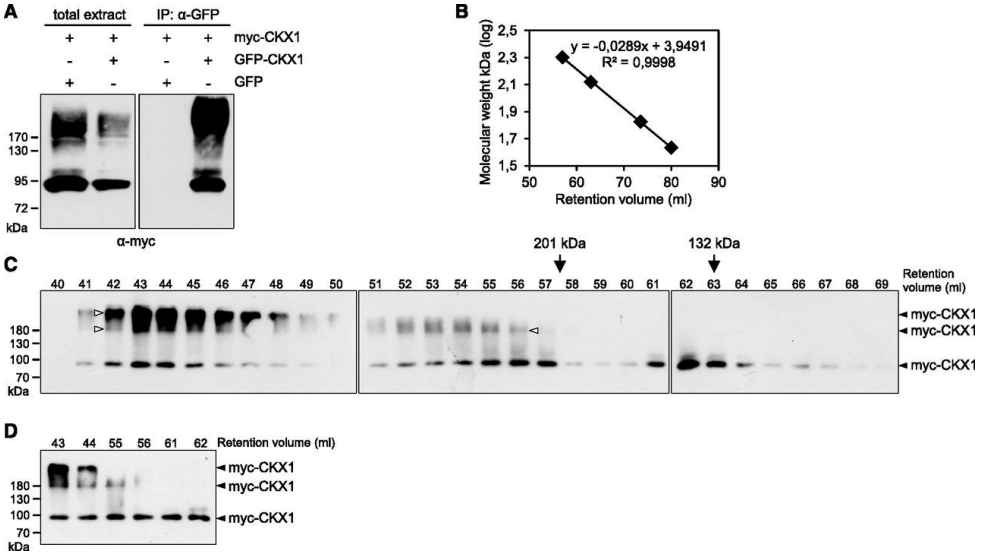


Figure 4.2 Analysis of CKX1 oligomerization. A, *In vivo* oligomerization of CKX1 detected by Co-IP assay. The myc-CKX1 protein was transiently coexpressed with GFP-CKX1 or GFP in *N. benthamiana*, and the protein extracts were used for immunoprecipitations (IP) with anti-GFP antibody followed by immunoblot detection with anti-myc antibody. The left gel shows the input (10 µg of the crude extract used for Co-IP assay); the right gel shows the pellet fractions from the Co-IP assays. The input control for GFP-CKX1 and GFP is shown in Supplemental Figure S4.1. B to D, SEC analysis of CKX1 complex formation. B, For the column calibration, the standard linear regression curve was generated by plotting the log of the molecular mass of calibration proteins against their retention volumes: BSA trimer (201 kD; 57.5 min), BSA dimer (132 kD; 63 min), BSA monomer (67 kD; 73.5 min), and ovalbumin (43 kD; 80 min). C, Microsomal membranes isolated from *N. benthamiana* leaves transiently expressing *35S*:*myc-CKX1* were solubilized with 1% *n*-dodecyl-β-d-maltoside (DDM) and subjected to SEC on a column equilibrated with 0.05% DDM, 50 mm Tris-HCl, pH 7.5, 10% glycerol, and 150 mm NaCl. Eluted fractions were analyzed by SDS-PAGE and immunoblot with anti-myc antibody. Arrows indicate peak elutions of molecular mass markers (BSA trimer and dimer). White arrowheads indicate the resistant dimeric and higher oligomeric myc-CKX1 forms. D, Six elution fractions from the experiment shown in C were reanalyzed in parallel on one western blot.

into complexes composed of protein, lipid, and detergent. The micelle size of the detergent used contributes to the final molecular mass of a given complex, thus influencing the elution volume during SEC (Kunji et al., 2008). Three major peaks of different apparent molecular sizes containing myc-CKX1 were detected. One myc-CKX1 peak eluted late, with an elution volume of 62 to 63 mL corresponding to an apparent molecular mass of ~130 to 140 kD (Fig. 4.2C). The average molecular mass of a DDM micelle is ~50 kD (Rögner, 2000) and the apparent molecular size of the myc-CKX1 monomer is ~90 kD, as deduced from the protein migration on the SDS-PAGE gel (Fig. 4.2A). Thus, the myc-CKX1 peak with an apparent size of ~130 to 140 kD corresponds to the myc-CKX1 monomer. The second myc-CKX1 peak eluted with a retention volume of 55 to 56 mL corresponding to an apparent molecular mass range of 215 to 230 kD and most probably represented the myc-CKX1 homodimeric form. The third peak, with a retention volume of 43 to 44 mL and apparent mass of around 470 to 510 kD, represented a higher oligomeric form of myc-CKX1 protein. Interestingly, the immunoblot analysis revealed that the apparent myc-CKX1 homodimer and higher oligomer were partly stable under our SDS-PAGE conditions. Intriguingly, when the SEC experiment was performed under reducing conditions with 5 mm β-mercaptoethanol included in the chromatography buffer, similar results were obtained and the different oligomeric forms of myc-CKX1 were detected (Supplemental Fig. S4.2).

To test CKX1 oligomerization independently and determine whether the protein-protein interaction also can occur in planta, oligomerization was tested using the optimized single vector bimolecular fluorescence complementation (BiFC) system, which utilizes monomeric Venus split at residue 210 (Gookin and Assmann, 2014). For this, *CKX1* was cloned in two expression cassettes of the double open reading frame expression vector pDOE-08, and by this, the N termini of two individual CKX1 proteins were fused to the N- and C-proximal halves of Venus (NVen and CVen, respectively). To monitor the nonspecific assembly of NVen and CVen, the parent vector expressing NVen-CKX1 and unfused CVen was used as a control. The vector

used also contains an integrated Golgi-localized mTurquoise2 marker (Golgi-mTq2) for the specific identification of transformed cells. We performed transient transformations of *N. benthamiana* leaves and examined the fluorescence by confocal laser scanning microscopy. We identified expressing epidermal cells by monitoring the Golgi-mTq2 fluorescence, and, as illustrated in Figure 4.3A, all transformed cells showed very strong Venus fluorescence,

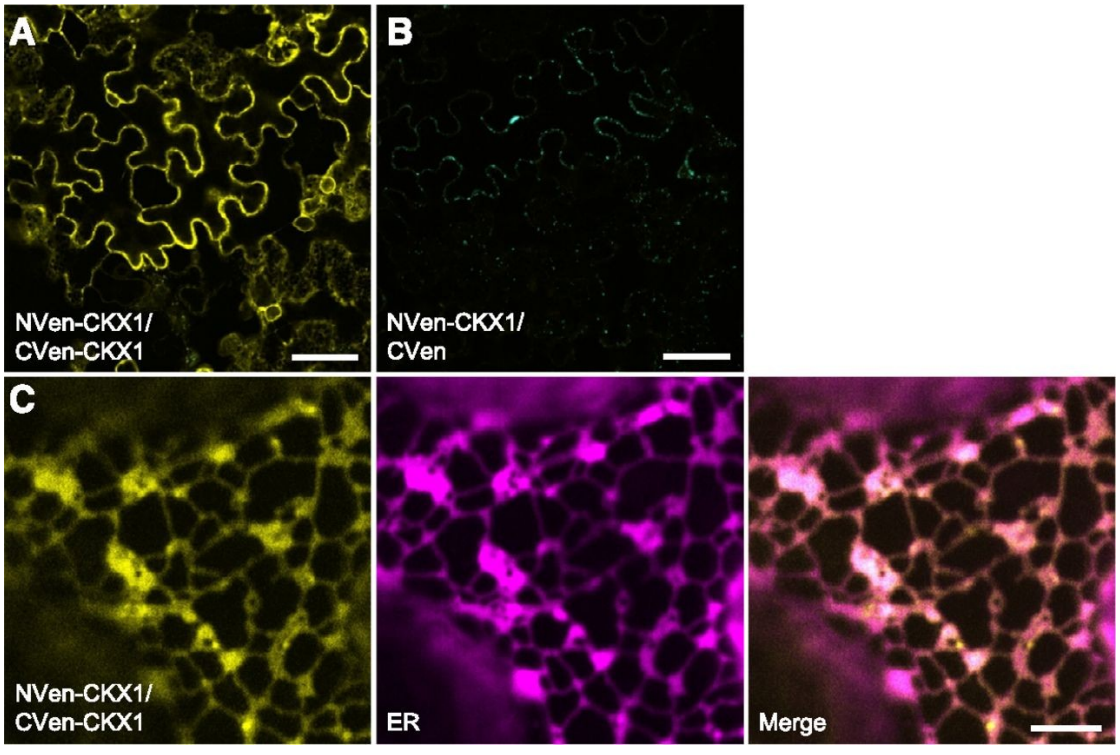


Figure 4.3 CKX1 homodimerizes in a BiFC assay. A, Confocal microscopy analysis of BiFC in *N. benthamiana* epidermal leaf cells reveals interaction between NVen-CKX1 and CVen-CKX1, as apparent from the reconstitution of the Venus-derived fluorescence (yellow). The microscopy was performed 2 d after infiltration (DAI). B, The NVen-CKX1/CVen parent vector shows no background BiFC signal 2 DAI. Successful transformation is evident from the cyan mTq2 signal in the Golgi. No BiFC signal was observed even after prolonged incubation (3 DAI). Images in A and B were captured using identical confocal settings. C, NVen-CKX1/CVen-CKX1 BiFC fluorescence signal localizes predominantly to the ER. Yellow, Venus BiFC; magenta, RFP-p24. Bars = 50 μm (A and B) and 5 μm (C).

indicating BiFC between NVen-CKX1 and CVen-CKX1. By contrast, no BiFC was detected when NVen-CKX1 was coexpressed with the untagged CVen fragment, although the analyzed cells displayed strong Golgi-mTq2 fluorescence (Fig. 4.3B), thus proving to be a genuine negative control. From these results, we conclude that CKX1 can assemble into a homooligomeric complex in planta.

4.3.3 CKX1 Is an ER-Resident Protein

Further detailed microscopic analysis showed that the NVen-CKX1/CVen-CKX1 BiFC signal was distributed in a reticular pattern characteristic of the cortical ER network (Fig. 4.3C). To verify this, we cotransformed the CKX1-BiFC construct with an ER marker protein (Lerich et al., 2011). The colocalization of the BiFC and ER marker signals indicated that the putative CKX1 homodimer localizes to the ER. This localization would be only partially consistent with the previously published data, which have shown that CKX1-GFP localized largely to the ER but occasionally also to the vacuole when expressed stably in Arabidopsis under the control of the 35S promoter (Werner et al., 2003). These earlier experiments, however, might not have been fully conclusive, due to possible overexpression artifacts (Werner et al., 2003). Indeed, (Niemann et al., 2015) recently showed that CKX1 apparently does not contain complex *N*-glycans, which is consistent with the idea that CKX1 could be an ER-resident protein. To examine the CKX1 subcellular localization and avoid strong overexpression effects, we tagged CKX1 with GFP at the N terminus (recapitulating the topology of the chimeric proteins in the BiFC assay), expressed the fusion protein transiently in *N. benthamiana* leaves, and performed confocal imaging at early time points after infiltration. As shown in Figure 4.4A, early after infiltration, GFP-CKX1 was localized exclusively to the ER in cells moderately expressing the fusion protein. In cells with higher expression levels, the GFP-CKX1 fluorescence signal also was localized to bright puncta of varying sizes (Supplemental Fig. S4.3). However, a similar shift of the fluorescence signal from the ER network into punctate structures also was often observed for the ER membrane marker RFP-p24 (Lerich et al., 2011) when expressed to higher levels

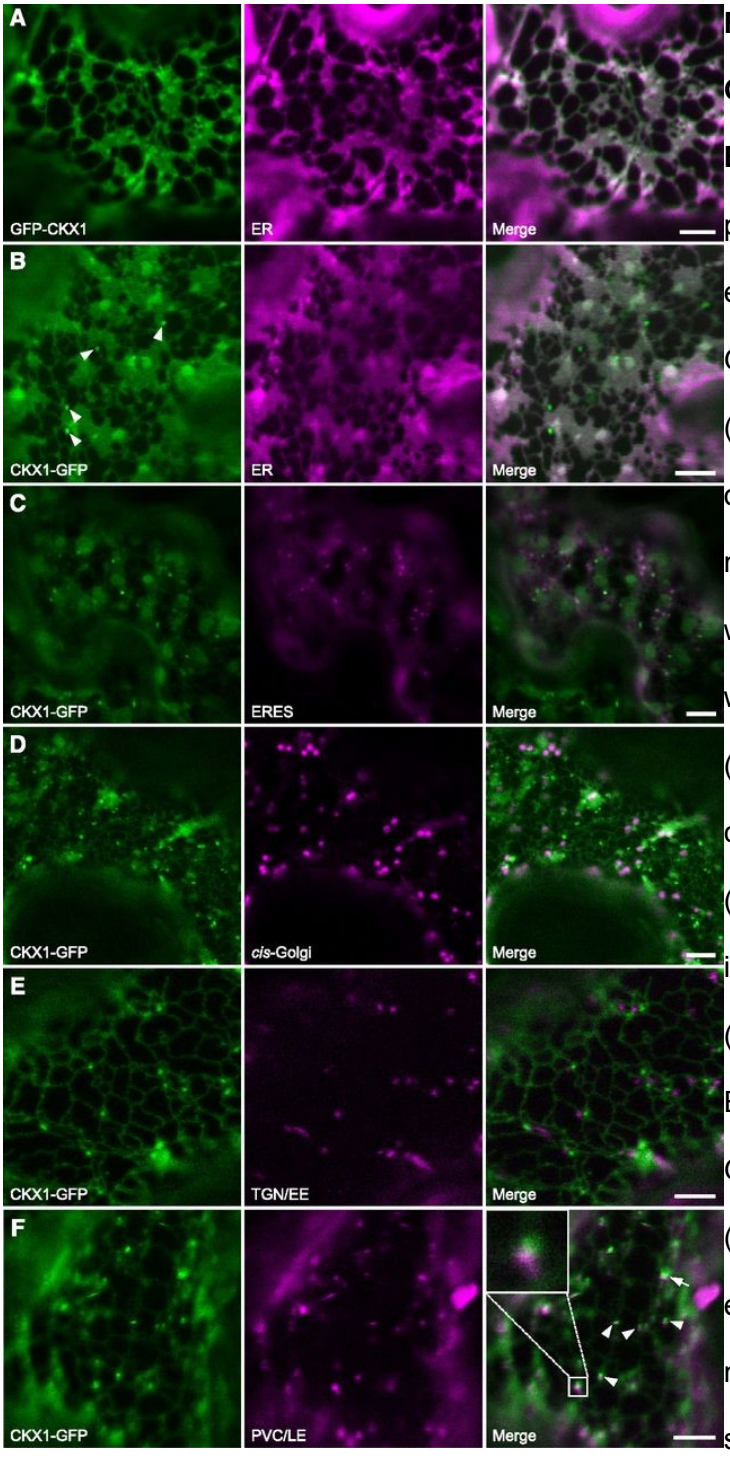


Figure 4.4 CKX1 fusion proteins to GFP localize predominantly to the ER. Confocal microscopy analysis was performed on *N. benthamiana* leaf epidermal cells coexpressing different GFP-fused CKX1 chimeric proteins (left column; green) and the indicated organelle markers (middle column; magenta). A, GFP-CKX1 colocalizes with the ER marker protein RFP-p24 when expressed 1 d after infiltration (DAI). B to E, CKX1-GFP largely colocalizes with the ER marker protein (B) and shows additional localization in small punctate structures (arrowheads), which are distinct from ERES labeled by YFP-Sec24 (C), Golgi bodies labeled by ERD2-YFP (D), and trans-Golgi network/early endosome (TGN/EE) labeled by mCherry-SYP61 (E). F, CKX1-GFP signal localizes mostly close to prevacuolar compartments/late endosomes (PVC/LE) labeled by ARA6-mCherry (arrowheads and magnified in inset). Occasionally, CKX1-GFP and ARA6-mCherry signals overlapped (arrow). The microscopy was performed 2 DAI. Bars = 5 μm.

(Supplemental Fig. S4.3A), which suggests either an aberrant protein localization due to exceeded ER retention capacity or general changes in the ER morphology. The latter assumption was further supported by a frequent colocalization of the strongly expressed GFP-CKX1 with the tobacco mosaic virus movement protein (MP-RFP; (Sambade et al., 2008); Supplemental Fig. S4.3B), marking ER-associated inclusions whose formation is associated with rearrangements of the ER membrane.

To analyze the possibility that the N-terminal GFP fusion masked an important targeting signal in the cytoplasmic tail of CKX1, we also transiently expressed CKX1 fused C-terminally to GFP under the control of the *35S* promoter (CKX1-GFP). Compared with the CKX1-GFP construct reported previously (Werner et al., 2003), this new chimeric protein includes a short linker to provide flexibility between the fused proteins (Miyawaki et al., 2003). Figure 4.4B shows that the fusion protein was localized predominantly to the ER with additional GFP signal associated with small endogenous, motile compartments (arrowheads). We first tested whether the punctate signal might represent endoplasmic reticulum export sites (ERES; (Hanton et al., 2006) by coexpression of the ERES marker protein YFP-Sec24 (Stefano et al., 2006); however, we observed no colocalization (Fig. 4.4C). Next, we analyzed colocalization with markers for some of the well-defined post-ER compartments. The cis-Golgi marker ERD2-YFP (Brandizzi et al., 2002) did not colocalize with the punctate CKX1-GFP signal (Fig. 4.4D). CKX1-GFP also did not colocalize with the trans-Golgi network/early endosome marker mCherry-SYP61 (Gu and Innes, 2011; Uemura et al., 2004); Fig. 4.4E). Intriguingly, upon coexpression with ARA6-mCherry, which labels prevacuolar compartments/late endosomes (Gu and Innes, 2012; Ueda et al., 2001), we observed that the punctate CKX1-GFP signal mostly localized in very close proximity to the ARA6-mCherry signal and occasionally colocalized with it (Fig. 4.4F). From the above experiments, we conclude that, similar to GFP-CKX1, the CKX1-GFP protein is localized mainly to the ER and to its closely associated punctate structures of not fully resolved nature.

4.3.4 The N-Terminal Part of CKX1 Is Required and Sufficient for Homooligomerization and Targeting to the ER

Interestingly, we observed no homodimerization in a yeast two-hybrid assay when truncated CKX1 protein without N-terminal signal anchor sequence (CKX135-575) was used, although this mutant form was capable of interacting with other proteins in yeast (H. Weber, unpublished data). This indicates that the CKX1 N terminus might be relevant for the homooligomerization. In order to test this hypothesis, we generated a chimeric reporter construct consisting of the first 79 N-terminal amino acid residues (comprising the cytoplasmic tail, TM domain, and putative stem region) of CKX1 fused to GFP (CKX11-79-GFP). To test the capacity of this short N-terminal peptide to mediate the homooligomerization, CKX11-79-GFP was coexpressed transiently together with myc-CKX1 in *N. benthamiana* leaves and used as bait in Co-IP experiments. Immunoblot analysis revealed that myc-CKX1 copurified robustly with CKX11-79-GFP (Fig. 4.5A; Supplemental Fig. S4.4). In parallel, we performed a control Co-IP assay with the full-length CKX1 protein tagged C terminally with GFP (CKX1-GFP). Interestingly, the quantity of copurified myc-CKX1 was comparable to that of CKX11-79-GFP (Fig. 4.5A), suggesting that the examined N-terminal region of CKX1 is primarily responsible for homooligomeric complex formation.

Next, we questioned the role of the CKX1 N terminus in directing the subcellular localization of the protein and compared the subcellular localization of CKX11-79-GFP with that of the full-length reporter CKX1-GFP. Transient expression and confocal imaging of CKX11-79-GFP revealed very similar subcellular localization comparable to that of the full-length reporter CKX1-GFP (Fig. 4.5B), suggesting that the N-terminal peptide is sufficient for the targeting and retrieval to the ER.

4.3.5 The CKX1 TM Domain Is Required for Protein Homooligomerization

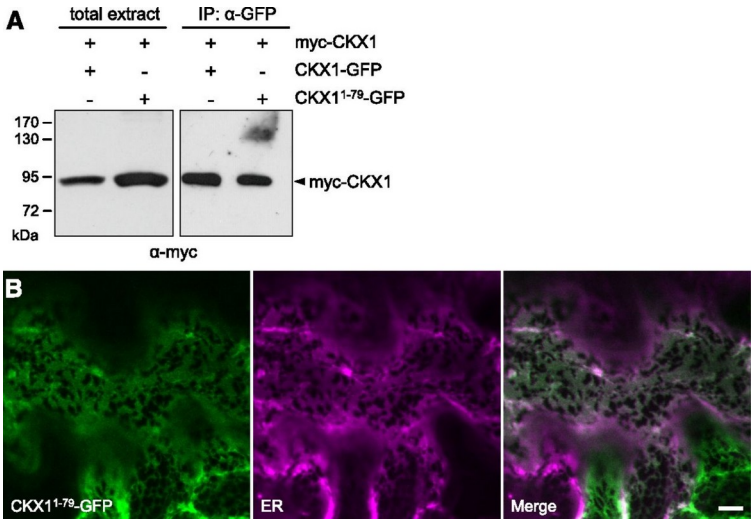


Figure 4.5 The CKX1 N terminus directs the homooligomerization and targeting to the ER. A, Co-IP assay for the detection of homooligomerization mediated by the CKX1$^{1-79}$ N-terminal fragment. The myc-CKX1 protein was coexpressed transiently with CKX1$^{1-79}$-GFP or CKX1-GFP in *N. benthamiana*, and the protein extracts were used for immunoprecipitations (IP) with anti-GFP antibody followed by immunoblot detection with anti-myc antibody. The left gel shows the input (10 µg of the crude extract used for Co-IP assay); the right gel shows the pellet fractions from the Co-IP assays. The input control for CKX1$^{1-79}$-GFP and CKX1-GFP is shown in Supplemental Figure S4.4. B, Confocal microscopy analysis of *N. benthamiana* leaf epidermal cells coexpressing CKX1$^{1-79}$-GFP (green) with the ER marker RFP-p24 (magenta). The microscopy was performed 2 DAI. Bars = 5 µm.

Showing the relevance of the N-terminal part of CKX1 for homooligomerization and subcellular localization, we further aimed to delimitate the functional motifs relevant for these processes. Several reports have shown that TM helices of type II membrane proteins can mediate protein oligomerization involving different interaction mechanisms (Tu and Banfield, 2010). CKX1 contains a potential GxxxG-like interaction motif (SxxxG) formed by residues Ser-27 and Gly-31 (Fig. 4.6A). GxxxG-like motifs consist of small amino acids (Gly, Ala, and Ser)

arranged to form GxxxG (where x represents any amino acid) and GxxxG-like patterns (Russ and Engelman, 2000; Senes et al., 2000). These interaction motifs often are found at the interface of GAS_{right} dimers, a frequently occurring TM association motif (Walters and DeGrado, 2006) characterized by the close proximity of the two TM helices and the formation of characteristic networks of carbon hydrogen bonds (Senes et al., 2001). We employed the computational structure prediction program CATM (Anderson et al., 2017; Mueller et al., 2014) to investigate whether the TM helices of CKX1 proteins may associate by forming a GAS_{right} dimer.

As shown in Figure 4.6A, CATM predicts a plausible GAS_{right} dimer for the TM sequence of CKX1. The resulting model is characterized by favorable complementary packing and mediated by the SxxxG motif. It should be noted that the software assumes that the TM domain is in regular helical conformation. The sequence of CKX1 contains a Pro residue at position 30, an amino acid that could potentially kink the helices, even though Pro also is compatible with straight or nearly straight conformation in TM helices (Senes et al., 2004). The Pro residue occurs on the opposite face of the dimeric contact; thus, it does not participate in the predictedinterface. The amino acids that are involved at the dimer interface are highlighted in the sequence in Figure 4.6A.

Computational mutational analysis indicated that introduction of the large Ile in place of the small amino acids of the SxxxG motif (Ser-27Ile and Gly-31Ile) would create significant steric clashes in the model; thus, it should prevent any dimerization mediated by the predicted association interface (Supplemental Fig. S4.6). To test the prediction, we introduced two Ile residues (Ser-27Ile and Gly-31Ile) in the full-length CKX1 protein and analyzed the homodimerization ability of the resulting mutant protein (CKX1) in a BiFC assay. Compared with the nonmutagenized control, BiFC between NVen-CKX1 and CVen-CKX1 was reduced severely (Fig. 4.6, B and C), suggesting that the interaction was mediated largely by the TM domain and that the identified residues are functionally relevant. Importantly, CKX1-GFP was glycosylated in

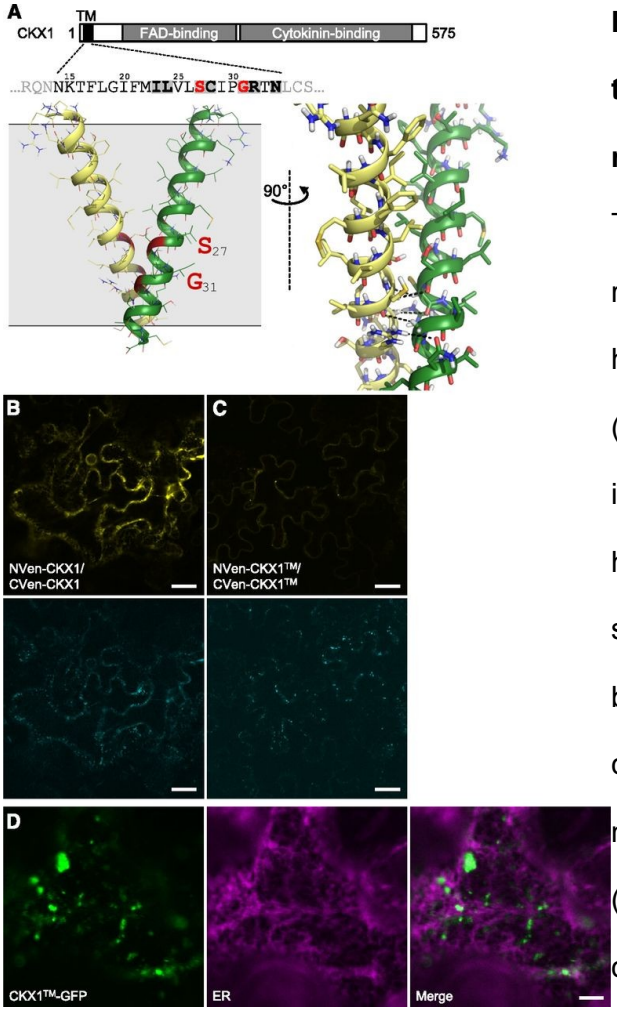


Figure 4.6 The CKX1 TM domain mediates the protein homooligomerization and ER retention. A, Structural model of the CKX1 TM dimer predicted by CATM. From left to right, ribbon representation of the entire TM helix (front view) and detail of the interface (side view). CATM predicts a well-packed interface mediated by the amino acids highlighted in the sequence. The SxxxG sequence pattern (marked in red) allows the backbones to come into close contact at the crossing point, enabling the formation of networks of interhelical hydrogen bonds (dashed lines) between Cα-H donors and carbonyl oxygen acceptors. Additionally, an interhelical hydrogen bond between the side chains of Arg-32 and Asn-34 also is observed. B and C, Confocal microscopy analysis of BiFC in *N. benthamiana* epidermal leaf cells reveals strongly reduced interaction between NVen-CKX1 and CVen-CKX1 (C) in comparison with the interaction of NVen-CKX1 and CVen-CKX1 (B), as apparent from the reconstitution of the Venus-derived fluorescence (yellow; top images). Comparable expression levels are apparent from the activity of the control gene Golgi-mTq2 (cyan; bottom images). The microscopy was performed 2 d after infiltration (DAI). Identical confocal settings were used to capture respective images in B and C. D, Confocal microscopy analysis of *N. benthamiana* leaf epidermal cells coexpressing CKX1-GFP (green) with the ER marker RFP-p24 (magenta). The microscopy was performed 2 DAI. Bars = 25 μm (B and C) and 5 μm (D).

a similar fashion to the control, and the mutant protein was detected exclusively in the membrane protein fraction (Supplemental Fig. S4.5), indicating that the introduced mutations in the TM region neither altered the capacity of the signal anchor to translocate the protein into the ER nor compromised the general helix structure and anchoring to the membrane.

4.3.6 CKX1 Oligomerization Is Indispensable for Its Biological Activity

We further examined whether the altered ability of CKX1-GFP to oligomerize affects the cellular behavior and biological activity of the protein. Confocal microscopy showed that, in comparison with the CKX1-GFP control (Fig. 4.4B), the fluorescence signal of CKX1-GFP was completely absent from the ER and accumulated in vesicles and vesicular aggregates of varying sizes (Fig. 4.6D). In addition, the overall CKX1-GFP fluorescence signal was reduced greatly in comparison with that of CKX1-GFP, together suggesting that the TM-mediated oligomerization regulates CKX1 retention/localization to the ER and/or contributes to the control of protein abundance in the ER. To address the biological relevance of the CKX1 oligomerization, we analyzed Arabidopsis plants stably expressing *CKX1-GFP* or *CKX1-GFP* under the control of the 35S promoter. The shoots of plants expressing the *35S*:*CKX1-GFP* transgene displayed strong phenotypic changes typical for cytokinin deficiency (Fig. 4.7A; (Holst et al., 2011; Werner et al., 2003), with a frequency of 45% among T1 plants. In comparison, we only detected very subtle phenotypic alterations among 90 T1 individuals transformed with the *35S*:*CKX1-GFP* construct. To rule out the possibility that the observed differences were due to different transgene expression levels, we identified homozygous lines with comparable transcript levels of the respective transgene (Fig. 4.7C; Supplemental Fig. S4.7). This analysis confirmed that the strong cytokinin deficiency phenotype was associated only with the expression of the *35S*:*CKX1-GFP*, but not the *35S*:*CKX1-GFP*, transgene (Fig. 4.7, A and B). Moreover, immunoblot analysis revealed that the levels of the CKX1-GFP protein were strongly diminished in comparison with CKX1-GFP (Fig. 4.7D; Supplemental Fig. S4.7), which was consistent with the absence of strong phenotypical changes in plants expressing the mutant protein. In line with

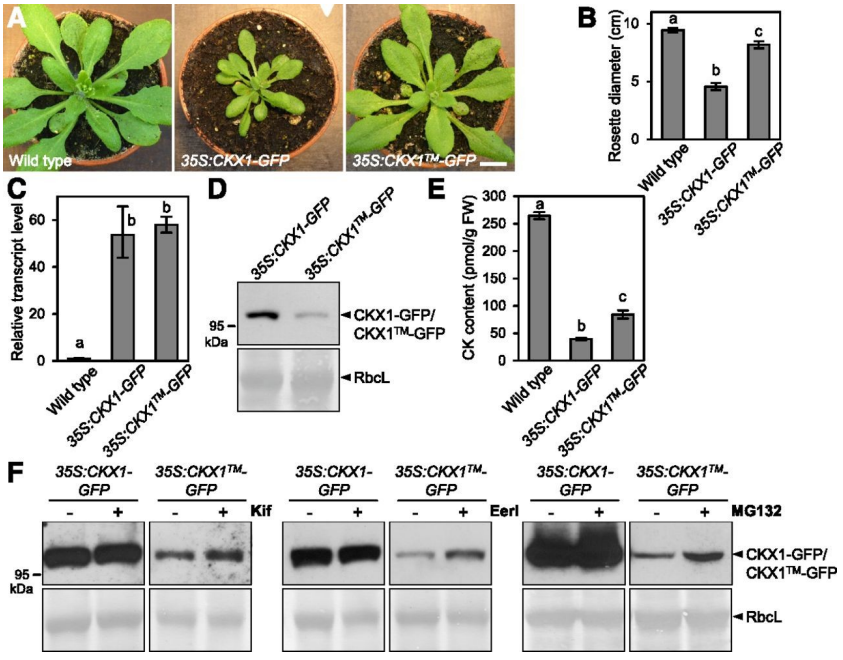


Figure 4.7 TM-mediated CKX1 homooligomerization regulates protein stability. A, Shoot phenotypes of the soil-grown wild-type control and plants expressing *35S*:*CKX1-GFP* (line 1) and *35S*:*CKX1-GFP* (line 14) 4 weeks after germination. Homozygous T4 plants are shown. Bar = 1 cm. B, Rosette diameters of the plants shown in A. Values are means ± sd ($n \geq 5$). C and D, Comparison of the *CKX1* transcript levels (C) and the protein abundances (D) in shoots of the *35*:*CKX1-GFP* and *35*:*CKX1-GFP* plants shown in A. Transcript levels were determined by quantitative real-time PCR. Means ± sd ($n = 4$) are shown in C. For the protein abundance analysis, 50 µg of the crude protein extracts was analyzed by immunoblot using anti-GFP antibody. Coomassie Blue staining of Rubisco large subunit (RbcL) was used as a loading control in D. E, Total cytokinin (CK) contents of the 3-week-old plants shown in A. Means ± sd ($n = 3$) are shown. FW, Fresh weight. F, Analysis of the effects of the ERAD inhibitors EerI and Kif, and of the proteasome inhibitor MG132, on CKX1-GFP and CKX1-GFP protein abundances. Arabidopsis seedlings grown in liquid cultures for 7 d were treated for 24 h with 50 µm Kif or 20 µm EerI and for 9 h with 100 mm MG132. Proteins were analyzed as described in D. In B, C, and E, different letters indicate statistically significant differences (Student's *t* test, $P < 0.05$).

this, direct determination of endogenous cytokinins revealed that their levels were significantly weaker in *35S*:*CKX1-GFP* lines in comparison with plants expressing the nonmutated *35S*:*CKX1-GFP* construct (Fig. 4.7E; Supplemental Fig. S4.7).

It was shown previously that several secretory CKX proteins are regulated by the ERAD pathway (Niemann et al., 2015). Therefore, we reasoned that the low levels of the oligomerization-deficient CKX1-GFP protein variant could be due to enhanced ERAD. To test this assumption, we analyzed CKX1-GFP levels upon treatment with Eeyarestatin I (EerI) and Kifunensin (Kif), which are specific inhibitors of the ERAD pathway (Fiebiger et al., 2004; Tokunaga et al., 2000; Wang et al., 2010). Figure 4.7D shows that the CKX1-GFP levels increased significantly upon both treatments. Similarly, CKX1-GFP levels were enhanced significantly by the proteasome inhibitor MG132. Together, these results suggest that the lower CKX1-GFP steady-state levels were caused by increased ERAD and that the CKX1 oligomerization is an important factor regulating its stability in the ER.

4.4 Discussion

Taking CKX1 from Arabidopsis as a case example, this study draws attention to several new molecular and cellular aspects of CKX-mediated cytokinin degradation and provides results that are relevant for better understanding of the functional modality of this metabolic pathway in controlling cytokinin activity in plants.

First, we demonstrated that CKX1 is not a soluble protein but an integral single-pass membrane protein with a type II architecture comprising a short N-terminal cytoplasmic tail, a TM helix, and a luminally oriented catalytic domain. A similar topology is typical for proteins such as Golgi- and ER-resident glycosyltransferases and glycosidases (Tu and Banfield, 2010). As signal peptides and N-terminal TM helices of the secretory pathway proteins generally are difficult to discriminate (Petersen et al., 2011), the possible membrane association of CKX proteins has been neglected previously (Schmülling et al., 2003) and clearly needs to be determined experimentally for individual CKX isoforms. Several CKX proteins have been shown convincingly to be soluble proteins containing cleavable signal peptides (Galuszka et al., 2005; Houba-Hérin et al., 1999), which, together with our results, suggests that two different subtypes of CKX isoforms, soluble and membrane bound, operate in the secretory system.

Along with the difference in this basic protein feature, different cellular behavior of the individual CKX isoforms can be expected as, for example, different sorting mechanisms for soluble and membrane proteins operate in the secretory pathway. Unlike for soluble proteins, sorting of membrane proteins is additionally determined by motifs located in their cytosolic domains and by the structure of the TM domain (Brandizzi et al., 2002; Gao et al., 2014). It is well established that the default destination for soluble proteins lacking positive sorting information is the apoplast (Rojo and Denecke, 2008). Indeed, several CKX isoforms shown to be soluble or having strongly predicted signal peptides have been demonstrated to be secreted to the apoplast (Bilyeu et al., 2001; Galuszka et al., 2005; Houba-Hérin et al., 1999). These

findings correlate with the fact that all putative soluble CKX proteins lack obvious sorting determinants, such as the (H/K)DEL ER retention signal. Thus, it appears that soluble CKX isoforms may generally follow the default secretory route to the apoplast. In contrast, CKX1, defined in this work as a prototypic membrane-bound CKX isoform, localized predominantly to the ER. CKX1 retention in the ER is well consistent with its apparent modification by high-Man *N*-glycans (Niemann et al., 2015). This finding is highly relevant because recent reports have revealed that AHK cytokinin sensor His kinases are localized predominantly in the ER (Caesar et al., 2011; Lomin et al., 2011; Wulfetange et al., 2011). Hence, CKX1, as an authentic ER protein, presumably coincides with the ER-localized AHK proteins and controls cytokinin concentrations directly perceived by the hormone receptors in the ER lumen. This active control of the cytokinin pool in the ER by CKX lends more support to the functional relevance of cytokinin receptor-mediated signaling from this cellular compartment.

Evidence regarding CKX1 localization beyond the ER is ambiguous. For example, prolonged expression of GFP-CKX1 caused, in addition to ER localization, the accumulation of GFP signal in larger bodies that coincide with ER-associated inclusions formed, for example, upon the expression of MP-RFP (Sambade et al., 2008). GFP-CKX1 signals in these structures, therefore, may reflect changes in ER structure that can be caused by strong, transient overexpression of an ER-resident protein (Niehl et al., 2012); Supplemental Fig. S4.3A) rather than by the normal cellular distribution of GFP-CKX1. Additionally, in the case of the C-terminal CKX1-GFP fusion, the predominant ER signal was accompanied by localization to very small puncta, which often were positioned in direct proximity of prevacuolar compartments/late endosomes labeled by ARA6-mCherry. However, these showed only limited colocalization. The identity of these CKX1-GFP-labeled structures will require further clarification. Importantly, both analyzed CKX1 fusion proteins were not detected in the vacuole. Together, the analysis does not support our previous hypothesis that CKX1 might be actively targeted to the lytic vacuole (Werner et al., 2003). It is possible that the occasional vacuolar targeting observed previously

(Werner et al., 2003) was an overexpression artifact due to saturated ER retention capacity. CKX1-GFP escaping ER retention mechanisms might passively reach the vacuole, which has been discussed as the default compartment for some membrane proteins (Barrieu and Chrispeels, 1999; Langhans et al., 2008). Accordingly, although we showed in this study that CKX1 is bound exclusively to membrane, the CKX1-GFP signal reported by (Werner et al., 2003) did not label the tonoplast but the vacuolar lumen, which indicates the formation of a soluble degradation product. Taken together, there is currently no clear experimental evidence supporting the function of CKX proteins in the vacuole. However, we note that cytokinin has been detected in vacuoles (Fusseder and Ziegler, 1988; Jiskrová et al., 2016; Kiran et al., 2012), but its biological significance in this organelle remains obscure.

A surprising outcome of our study is that CKX1 forms homodimeric and oligomeric complexes *in vivo*. Most interestingly, complex formation was mediated mainly by a strong interaction between the TM domains. Oligomerization of the TM helices of bitopic membrane proteins can be important for the structural assembly of stable protein complexes as well as play functional roles when association or conformational changes are critical for modulating signaling and regulation (Moore et al., 2008). Classic examples are the receptor Tyr kinase and cytokine receptor families of type I membrane proteins, for which dimerization and structural rearrangement involving the TM region play critical roles in activation (Li and Hristova, 2006; Maruyama, 2015). For animal type II membrane proteins, including several Golgi glycosyltransferases, it was shown previously that protein oligomerization can be determined by the luminal juxtamembrane region and/or the TM-spanning region (Tu and Banfield, 2010).

A variety of physical forces have been implicated in the promotion of TM helix interactions (Li et al., 2012; Senes et al., 2004), from van der Waals packing (MacKenzie et al., 1997) to hydrogen bonding between polar amino acids (Choma et al., 2000; Zhou et al., 2000) and aromatic π-π and cation-π interactions (Johnson et al., 2007). A particularly important class of TM helix interaction motifs is the GAS_{right} dimer, which is stabilized by unusual networks of

hydrogen bonds that are formed by Cα-H donors and backbone carbonyl oxygen acceptors on the opposite helix (Cα-H···O=C bonds; (Senes et al., 2001)). The signature sequence pattern of GAS_{right} is the presence of small residues (Gly, Ala, and Ser) arranged in motifs such as GxxxG or variants thereof that facilitate close interhelical contact and carbon-hydrogen bond formation between TM helices (Mueller et al., 2014). The mutation and Co-IP analyses presented in this work showed that the identified GxxxG-like motif (SxxxG) in the TM domain of CKX1 is largely required for CKX1 homooligomerization. Currently, little is known about GxxxG-mediated protein-protein interactions and their functions in plants. A TM domain containing a GxxxG motif has been reported to occur in many receptor-like kinases and receptor-like proteins mediating plant immune responses (Fritz-Laylin et al., 2005), but only two studies have addressed the function of the GxxxG motif in protein-protein interactions and signaling responses to pathogens (Bi et al., 2016; Zhang et al., 2010). In addition to dimerization, our SEC fractionations also suggested higher order oligomerization of CKX1, which is in accord with several previous reports describing the assembly of GxxxG dimers into higher oligomeric complexes (Dews and Mackenzie, 2007; Hoang et al., 2015; Kwon et al., 2015; Xu et al., 2007). Although the underlying assembly mechanisms are mostly unclear, they may involve TM domain interactions as well as interfaces in the soluble domain.

It will be important to understand whether the described protein features are conserved among CKX homologs. Our sequence analysis revealed only a related AxxxA motif (Gimpelev et al., 2004) in the TM domain of CKX6, indicating that the sequence of TM domains is not conserved and that the SxxxG motif is unique to CKX1. However, given that TM domains do not need to contain specific sequence motifs to oligomerize (Moore et al., 2008), it is currently not possible to conclude whether oligomerization is a shared mechanism in the CKX family, and individual proteins will need to be analyzed experimentally in the future.

Ultimately, it is important to understand the significance of CKX1 homooligomerization for its cellular activity. One possibility is that the CKX1 oligomeric state and its enzymatic activity would

be coupled. Examples of such a structure-activity relation for type II membrane proteins are known (Chung et al., 2010; Tu and Banfield, 2010). However, heterologous expression of a chimeric CKX1 protein with the N terminus replaced by a cleavable yeast secretion signal yielded a relatively high enzyme activity preparation (Kowalska et al., 2010), suggesting that the TM-mediated oligomerization may not be required for CKX1 enzyme activity per se. Further experiments are still needed to test this possibility more rigorously. In contrast, our analysis demonstrated that mutations rendering CKX1 monomeric cause (1) a loss of its ER localization, resulting in an unspecified cellular redistribution, and (2) a reduction of its overall cellular levels. The first suggests that the CKX1 oligomerization status may represent an important determinant for its ER retention and, consequently, for the cytokinin concentration and signaling activity in the ER. ER retention mechanisms based on TM-mediated protein dimerization were proposed earlier (e.g. for the type II TM chaperone COSMS; (Sun et al., 2011). It should be noted, however, that the ER residency of membrane proteins often can be determined by the combined activity of different retention and retrieval signals (Boulaflous et al., 2009). Therefore, it will be interesting to analyze whether the ER localization of CKX1 is eventually controlled by additional sorting signals (Cosson et al., 2013; Gao et al., 2014). Second, our analysis demonstrated that plants expressing the monomeric CKX1 mutant variant accumulated the protein to levels considerably lower than those detected in plants expressing the wild-type form. The reduced protein levels were correlated with the lack of a prominent cytokinin deficiency phenotype in the respective transgenic lines, suggesting that the capacity of the mutant protein to regulate the cytokinin concentration in the ER was impaired. Less severe reduction of endogenous cytokinin levels in *35S*:*CKX1*-*GFP*-expressing lines corroborated this conclusion. It is interesting that the cytokinin levels in these lines were still significantly lower in comparison with the wild type, which is in line with the notion that the cytokinin signal must be reduced below a certain threshold to trigger strong growth alterations (Werner et al., 2010).

We have recently shown that CKX1 as well as apparently other CKX isoforms targeted to the secretory pathway are regulated by the proteasome-dependent ERAD pathway (Niemann et al., 2015), which represents a conserved cellular route to withdraw proteins from the ER that fail to attain their native conformation (Römisch, 2005). Therefore, it is conceivable that the unassembled monomeric CKX1 is prone to increased degradation by ERAD. Consistent with this hypothesis, the CKX1-GFP protein levels were significantly restored by treatments with ERAD inhibitors, indicating that CKX1 oligomerization is a crucial parameter determining its ERAD and, hence, the protein abundance in the ER. The exact mechanisms underlying ERAD of CKX1 and other CKX proteins are currently unknown; however, it can be hypothesized that the assembly of individual subunits into multimeric complexes can enhance protein folding or conformational stability, which can prevent proteolytic degradation (Vembar and Brodsky, 2008). Although it needs to be studied in more detail, it is interesting that CKX1-GFP levels were not fully rescued by the ERAD inhibition, suggesting that, eventually, other mechanisms may be involved in CKX1 removal from the ER as well.

It should be further noted that detailed genetic studies will be required in the future to complement the data presented here and to identify biological processes involving the molecular mechanisms described in this work.

4.5 Materials and Methods

4.5.1 Plasmid Construction

The *35S*:*myc-CKX1* construct was described previously (Niemann et al., 2015). To generate *35S*:*GFP-CKX1*, the CKX1 cDNA from pDONR221-CKX1 (Niemann et al., 2015) was subcloned into pK7WGF2 (Karimi et al., 2002) by Gateway LR recombination (Invitrogen). For *35S*:*CKX1-GFP*, the *CKX1* cDNA was PCR amplified in two steps by using primer pairs 1/2 and 3/4 (Supplemental Table S4.1), and the final amplicon was cloned into the vector pDONR221 (Invitrogen) and subsequently pK7FWG2 (Karimi et al., 2002). The *CKX11-79-GFP* fusion gene was created by overlapping PCR. In the first step, the *CKX1* fragment and GFP-coding sequence were amplified by using primer pairs 3/5 and 6/7 and pDONR221-CKX1 and pK7WGF2 as templates, respectively. These two fragments were combined and amplified with primers 3 and 4, and the final amplicon was cloned successively in pDONR221 and pK7FWG2 to generate *35S*:*CKX11-79-GFP*. The CKX1-GFP construct was generated by site-directed mutagenesis (Eurofins Genomics).

For protein-protein interaction by BiFC, the *CKX1* cDNA was PCR amplified using primer pairs 8/9 and 10/11, and the resulting fragments were cloned into pJet vector. First, CKX1 cDNA was subcloned into the MCS1 *Bam*HI site of pDOE-08 (Gookin and Assmann, 2014), resulting in pDOE-08-CKX1 parent vector expressing CKX1 N-terminally tagged with the N-terminal fragment of monomeric Venus split at residue 210 (NVen-CKX1) and unfused C-terminal Venus fragment (CVen). This vector was used as a negative control. In the next step, the second CKX1 cDNA fragment was subcloned into the *Kfl*I site within MCS3 of the pDOE-08-CKX1 parent vector, resulting in vector expressing NVen-CKX1/CVen-CKX1 used for the homodimerization test. Mutated full-length *CKX1* cDNA was used in a similar cloning approach to generate the vector encoding NVen-CKX1/CVen-CKX1.

Single-copy transgenic Arabidopsis (*Arabidopsis thaliana*) lines harboring *35S*:*CKX1-GFP* and *35S*:*CKX1-GFP* were used in this study.

4.5.2 Transient Expression in *Nicotiana benthamiana* and Confocal Laser Scanning Microscopy

Infiltration was performed as described previously (Niemann et al., 2015; Sparkes et al., 2006) using *Agrobacterium tumefaciens* strain GV3101:pMP90 and 6-week-old *N. benthamiana* plants. For coexpression, the *A. tumefaciens* cultures harboring different expression constructs were mixed in infiltration medium to a final OD600 of 0.1 for the CKX1 fusions and 0.01 to 0.05 for the marker constructs. *35S*:*p19* was included in all infiltrations at OD600 = 0.1. The following binary constructs were used in this work: pH7MP:RFP (Boutant et al., 2010), RFP-p24 (Lerich et al., 2011), ERD2-YFP (Brandizzi et al., 2002), YFP-Sec24 (Stefano et al., 2006), mCherry-SYP61 (Gu and Innes, 2011), and ARA6-mCherry (Gu and Innes, 2012). Confocal imaging analysis was performed using a Leica TCS SP5 laser scanning confocal microscope 1 to 3 d after infiltration. mTq2, GFP, YFP, RFP, and mCherry were excited at 458, 488, 514, and 561 nm, and the fluorescence emissions were detected at 461 to 488, 498 to 538, 520 to 556, 600 to 630, and 590 to 640 nm, respectively. In cases where GFP and YFP were analyzed simultaneously, GFP and YFP were detected at 490 to 507 and 557 to 585 nm, respectively.

4.5.3 Preparation of Microsomal Membranes and Membrane Association Analysis

N. benthamiana leaves (1 g) were homogenized in 5 mL of homogenization buffer (25 mm Tris-HCl, pH 7.5, 300 mm Suc, 1 mm EDTA, 1 mm 1,4-dithioerythritol, and complete protease inhibitor cocktail without EDTA [Roche]) using a mortar and pestle. The homogenate was passed through one layer of Miracloth (Calbiochem) and centrifuged at 10,000*g* for 10 min at 4°C to remove the debris. The microsomal membrane fraction was pelleted by ultracentrifugation at 100,000*g* for 90 min at 4°C. Pellets were resuspended in 5 mL of

homogenization buffer or homogenization buffer supplemented with 1 m NaCl, 2 m urea, 0.1 m Na2CO3, pH 11, or 1% Triton X-100.

For the protease digestion assay, the microsomal membranes were isolated from rosettes of 14-d-old soil-grown Arabidopsis plants expressing *35S*:*myc-CKX1* (Niemann et al., 2015) and *35S*:*CKX1-myc*. The 100,000*g* pellet was resuspended in proteinase inhibitor-free homogenization buffer and incubated with 10 µg mL−1 proteinase K at room temperature for 45 min in the presence or absence of 1% Triton X-100. A concentration of 6 mm phenylmethanesulfonyl fluoride (Sigma-Aldrich) was used to terminate the protease digestions. After 15 min of incubation on ice, the membranes were solubilized with 2× SDS-PAGE sample buffer (125 mm Tris-HCl, pH 6.8, 4% SDS, 20% glycerol, 10% β-mercaptoethanol, and 0.01% Bromphenol Blue).

Protein samples were resolved by SDS-PAGE and blotted on PVDF membranes (Millipore). Membranes were blocked with 5% skim milk in PBS containing 0.1% Tween 20. A mouse monoclonal anti-myc antibody (clone 4A6; Millipore; dilution 1:1,000) followed by a goat anti-mouse antibody coupled to horseradish peroxidase (sc-2005; Santa-Cruz; dilution 1:2,000) were used to detect myc-CKX1. For immunodetection of Arabidopsis calnexins, the blots were stripped (2 × 10 min; 1.5% Gly, 0.1% SDS, and 1% Tween 20, pH 2.2) and reprobed by using anti-CNX1/2 antibody (Agrisera; dilution 1:10,000) and horseradish peroxidase-conjugated goat anti-rabbit antibody (Calbiochem; dilution 1:2,000). Bound antibodies were visualized with SuperSignal West Pico chemiluminescent substrate (Thermo Scientific).

4.5.4 Co-IP Assays

GFP and myc fusion proteins were coexpressed in *N. benthamiana* leaves, which were ground in liquid nitrogen and homogenized in extraction buffer (50 mm Tris-HCl, pH 7.5, 150 mm NaCl, 0.3% Triton X-100, 0.2% Igepal, 1 mm phenylmethanesulfonyl fluoride, and complete protease inhibitor cocktail [Roche]). Samples were cleared by 10 min of centrifugation at 4°C and 6,000*g*. Supernatants (1.4 mL) were adjusted to 2.8 mg mL−1 protein and incubated with

20 µL of GFP-Trap-A beads (Chromotek) for 3 to 4 h at 4°C. Beads were washed five times with the extraction buffer, mixed with 20 µL of 2× SDS-PAGE sample buffer, incubated for 5 min at 95°C, and cleared by centrifugation. The proteins were subjected to SDS-PAGE and immunoblot analysis using anti-myc or anti-GFP antibody (clone JL-8; Clontech; dilution 1:2,500).

4.5.5 SEC

Microsomal membranes were isolated according to the protocol by (Abas and Luschnig, 2010). Briefly, *N. benthamiana* leaves were homogenized in 1 volume of extraction buffer (100 mm Tris-HCl, pH 7.5, 300 mm NaCl, 25% Suc, and 5% glycerol). The homogenate was kept on ice for 20 min and centrifuged at 600*g* for 3 min. After an additional 20 min of incubation on ice, the supernatant was diluted with 1 volume of water, divided into 200-µL aliquots in 1.5-mL tubes, and centrifuged at 16,000*g* for 2.5 h. The membranes from 7 g of leaves were solubilized in 1 volume of buffer (50 mm Tris-HCl, pH 7.5, 150 mm NaCl, 20% glycerol, 15 mm β-mercaptoethanol, and 1% DDM) overnight at 4°C. Membrane proteins were concentrated using Amicon Ultra-15 centrifugal filter units (50-kD cutoff; Millipore). The whole protein extract was loaded on the HiLoad 16/60 Superdex 200 column (GE Healthcare) equilibrated with at least 3 column volumes of running buffer (50 mm Tris-HCl, pH 7.5, 150 mm NaCl, 10% glycerol, and 0.05% DDM). Chromatography was performed with the ÄKTA FPLC system (GE Healthcare) at a flow rate of 1 mL min−1. Elution fractions of 1 mL were collected and subjected to SDS-PAGE followed by western blotting and immunodetection. The Superdex 200 column was calibrated using the following proteins as standards: BSA trimer (201 kD), BSA dimer (132 kD), BSA monomer (67 kD), and ovalbumin (43 kD).

4.5.6 RNA Extraction, cDNA Synthesis, and Quantitative PCR

RNA extraction from shoots of single plants, cDNA synthesis, and quantitative PCR were done as described before using *UBC10* for normalization (Niemann et al., 2015). The primers

used for *CKX1* amplification in the quantitative PCR were CKX1-fw (5′-ATGGATCAGGAAACTGGCAA-3′) and CKX1-rev (5′-AGATGAAAACAAAGTGGATGGAA-3′).

4.5.7 Treatments with ERAD Inhibitors

Seedlings were grown in liquid cultures for 7 d followed by 24 h of treatment with 50 µm Kif dissolved in water, 20 µm EerI dissolved in DMSO, and the respective mocks. A total of 50 µg of protein extracts was analyzed by immunoblot analysis using anti-GFP antibody as described above. Loading was verified by Coomassie Blue staining after immunoblot detection according to (Welinder and Ekblad, 2011).

4.5.8 Determination of Cytokinin Content

The cytokinin content in shoots of 3-week-old soil-grown plants was determined by ultra-performance liquid chromatography-electrospray-tandem mass spectrometry as described by (Svačinová et al., 2012), including modifications described by (Antoniadi et al., 2015).

4.5.9 Computational Modeling

The structure of CKX1-TM was predicted from its sequence (11-RQNNKTFLGIFMILVLSCIAGRTNLCS-37) using CATM (Mueller et al., 2014). Side chain mobility was modeled using the energy-based conformer library applied at the 95% level (Subramaniam and Senes, 2012). Energies were determined using the CHARMM 22 van der Waals function (MacKerell et al., 1998) and the hydrogen bonding function of SCWRL 4 (Krivov et al., 2009), as implemented in MSL (Kulp et al., 2012), with the following parameters for Cα donors, as reported previously: B = 60.278; D0 = 2.3 Å; σd = 1.202 Å; αmax = 74°; and βmax = 98° (Mueller et al., 2014). The relative energy of the Ser-27Ile, Gly-31Ile mutant was calculated as

$$\Delta E_{mut} = \left(E_{mut,dimer} - E_{mut,monomer}\right) - \left(E_{WT,dimer} - E_{WT,monomer}\right)$$

where *EWT,dimer* and *Emut,dimer* are the energies of the wild-type and mutant sequences, respectively, in the dimeric state and *EWT,monomer* and *Emut,monomer* are the energies of the

wild-type and mutant sequences, respectively, in a side chain-optimized monomeric state with the same sequence.

4.5.10 Accession Numbers

Sequence data from this article can be found in the GenBank/EMBL data libraries under accession number At2G41510 (*CKX1*).

4.6 Supporting Information

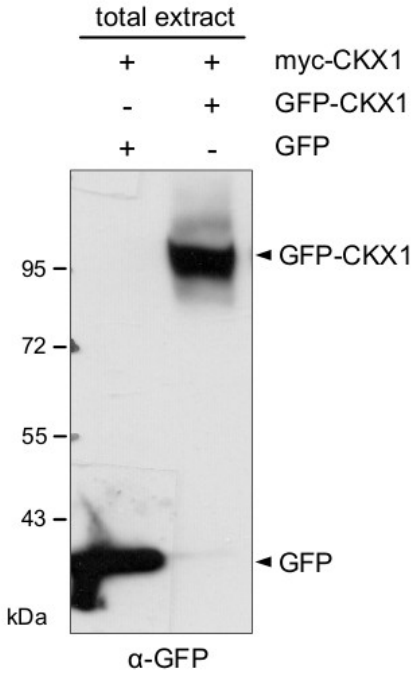


Figure S4.1 Coimmunoprecipitation (Co-IP) detection of CKX1 oligomerization. Supplementary information to the Figure 4.2A showing comparable expression of GFP-CKX1 and GFP bait proteins detected by immunoblot assay using anti-GFP antibody.

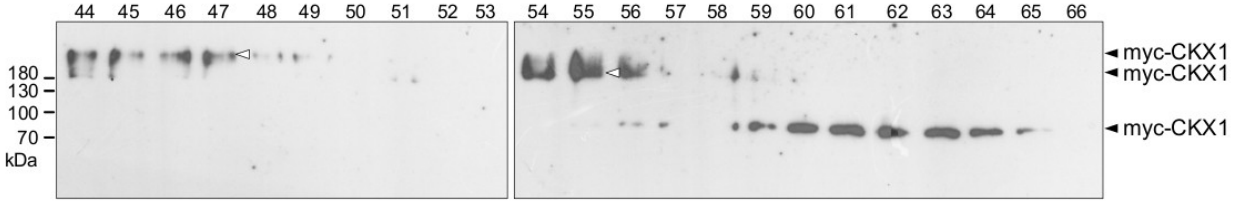


Figure S4.2 Size exclusion chromatography analysis of CKX1 complex formation. Total proteins were extracted from N. benthamiana leaves by using buffer containing 15 mM β-mercaptoethanol. Size exclusion chromatography was performed under identical conditions as described in Figure 4.2, except the chromatography was performed under reducing conditions with 5 mM β-mercaptoethanol included in the chromatography buffer. Eluted fractions were analyzed by SDS-PAGE and immunoblot using anti-myc antibody. White arrowheads indicate the dimeric (fractions 54-56) and higher oligomeric (fractions 44-47) myc-CKX1 forms.

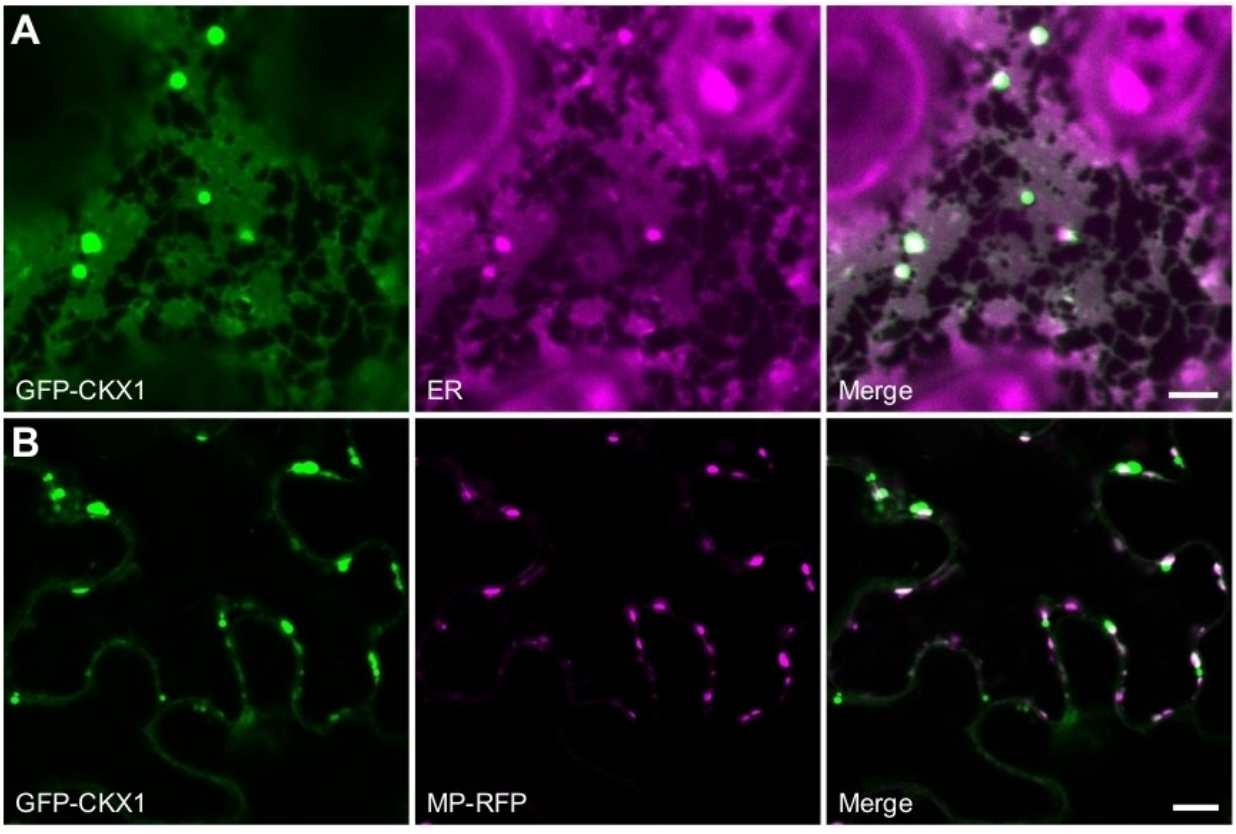


Figure S4.3 Strong overexpression of GFP-CKX1 alters the ER morphology. A and B, GFP- CKX1 localizes to the ER network and bright puncta of varying sizes in N. benthamiana leaf epidermal cells expressing high levels of GFP-CKX1 2 DAI. In cells with high level of expression of the marker protein, these GFP-CKX1 puncta frequently colocalize with the ER marker (A) and
transiently expressed tobacco mosaic virus movement protein (MP-RFP) (B). Scale bars = 5 μm (A) and 10 μm (B).

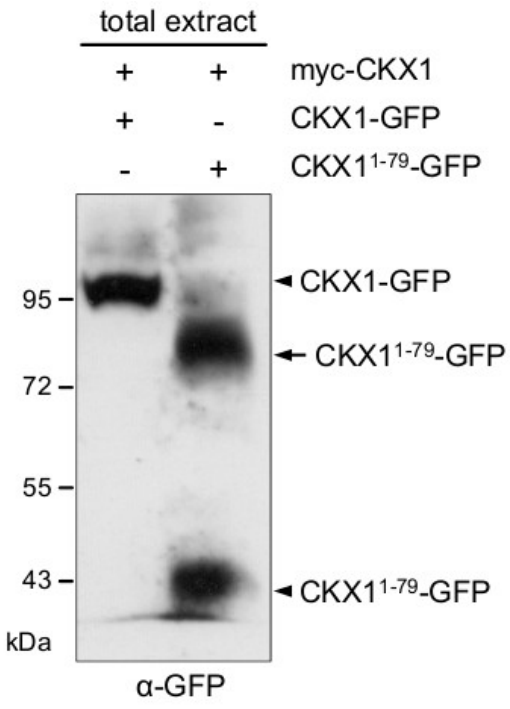


Figure S4.4 Coimmunoprecipitation (Co-IP) detection of homooligomerization mediated by the CKX1 1-79 N-terminal fragment. Supplementary information to the Figure 4.5A showing comparable expression of CKX1-GFP and CKX1 1-79 -GFP bait proteins detected by immunoblot assay using anti-GFP antibody. Note that CKX1 1-79 -GFP migrates in the SDS-PAGE gel as a monomer (predicted molecular mass of approximately 38 kDa; arrowhead) and as an apparent dimer (arrow).

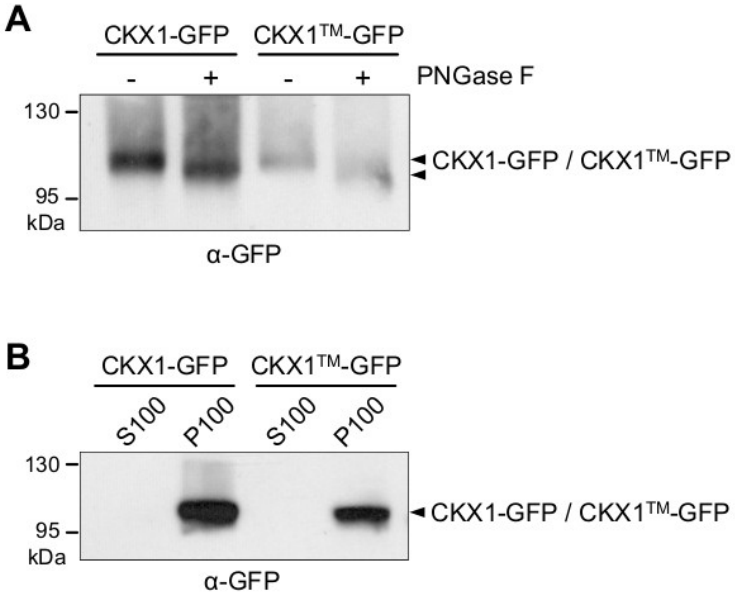


Figure S4.5 CKX1 TM -GFP is N-glycosylated and membrane-associated. A, The N-glycosylation status of CKX1 TM -GFP is comparable to that of CKX1-GFP. The fusion proteins were transiently expressed in N. benthamiana, the protein extracts treated with PNGase F and the electrophoretic mobility shift analyzed by SDS-PAGE and immunoblot with anti-GFP antibody. B, The membrane association of the CKX1 TM -GFP is not altered. Total membranes (P100) were isolated from precleared protein extracts by centrifugation at 100,000g, resuspended in the same volume of buffer, and compared to the soluble protein fraction (S100) by SDS-PAGE and immunoblot using anti-GFP antibody.

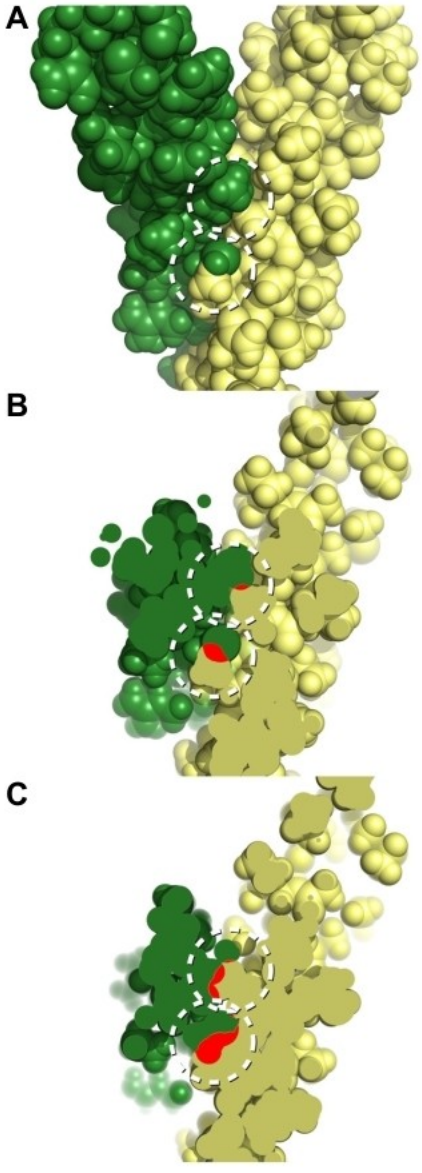


Figure S4.6 Structural model of Ser27Ile, Gly31Ile double mutant of CKX1. Introduction of Ile side chains at positions 27 and 31 causes severe clashes in the predicted structural model of CKX1 TM. A, Space filling model. One helix is represented in green, the other helix in yellow. The dashed circles indicate the locations of the clashes between Ile27 and Leu24 on the opposed chain (upper circle), and between Ile31 and Cys28 (lower circle). B and C, The clashes are evidenced in two different cross-section views of the model shown in (A), in which the regions of overlap of atoms belonging to the two helices are highlighted in red.

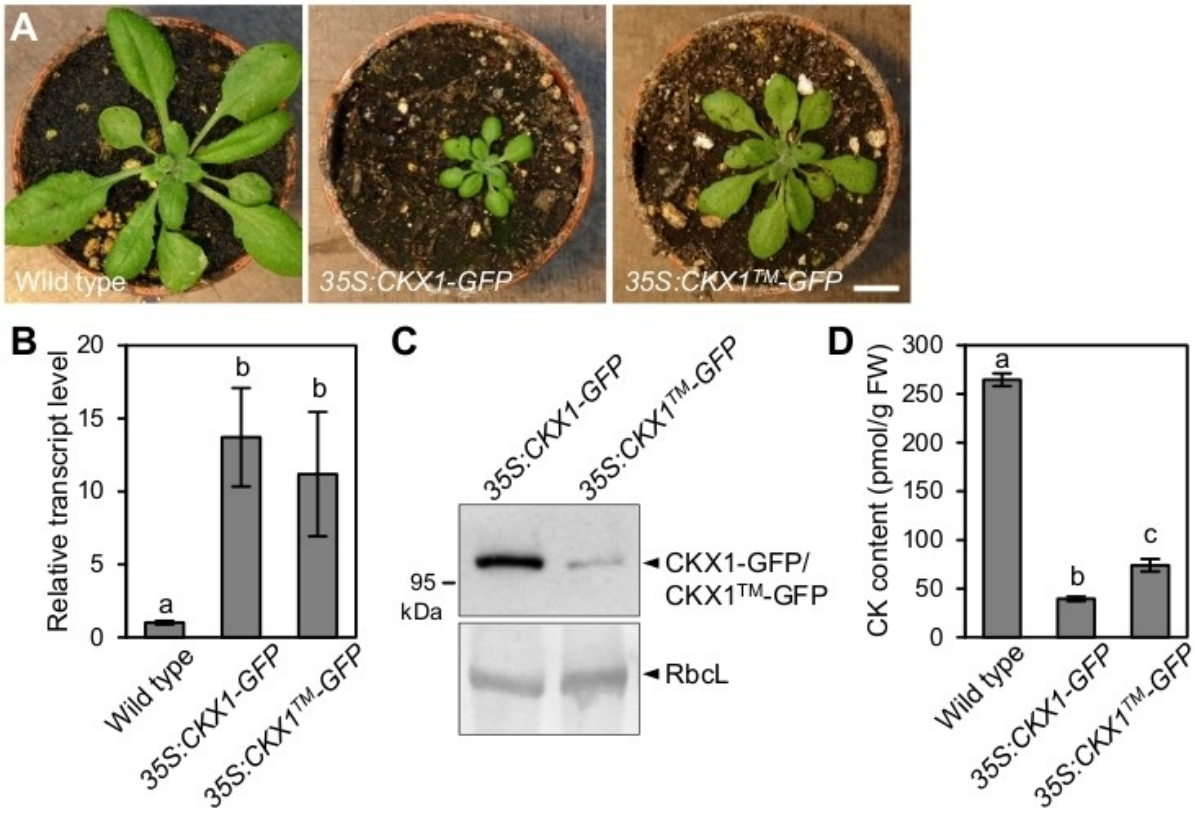


Figure S4.7 Growth and molecular phenotypes of Arabidopsis plants expressing 35S:CKX1-GFP (line 1) and 35S:CKX1 TM -GFP (line 54). A, Shoot phenotypes of soil-grown homozygous T4 plants 3 weeks after germination. Scale bar = 1 cm. B and C, Comparison of the CKX1 transcript levels (B) and the protein abundances (C) in the shoots of the 35:CKX1-GFP and 35:CKX1 TM -GFP plants shown in (A). Transcript levels were determined by quantitative real-time PCR. Means ± SD (n = 3) are shown (B). For the protein abundance analysis, 20 µg of the crude protein extracts were analyzed by immunoblot using α-GFP antibody. Coomassie blue staining of Rubisco large subunit (RbcL) was used as loading control (C). D, Total cytokinin (CK) content of 3- week-old plants. Means ± SD (n = 3) are shown. Different letters indicate statistically significant differences (Student's t test, $P < 0.05$).

Table S4.1 Oligonucleotides used in this study.

Primer	Name	Gene	Fw/ Rev	Sequence (5’ - 3’)
1	CKX1_GW5	CKX1	Fw	aaaaagcaggctttATGGGATTGACCTC
2	CKX1_GW3	CKX1	Rev	agaaagctgggttTACAGTTCTAGGTTTCGG
3	attB1		Fw	GGGGACAAGTTTGTACAAAAAAGCAGGCT
4	attB2		Rev	GGGGACCACTTTGTACAAGAAAGCTGGGT
5	CKX1_D79_rv	CKX1	Rev	GTCCTTGGCCACATTG
6	CKX1-GFP_fw2	GFP	Fw	CACAATGTGGCCAAGGACGTGAGCAAGGGCGAG
7	GFP_att2	GFP	Rev	agaaagctgggttCTAAAGCTTATCTTGTACAGCTCG
8	CKX1 MCS1 fw	CKX1	Fw	agggatccGGATTGACCTCATCC
9	CKX1 MCS1 rev	CKX1	Rev	agggatccTTATACAGTTCTAGG
10	CKX1 MCS3 fw	CKX1	Fw	aggggtcccCGGGATTGACCTCATCC
11	CKX1 MCS3 rev	CKX1	Rev	aggggacccTTATACAGTTCTAGG

Fw, forward; Rev, reverse.

Cloning adaptors are shown in lowercase letters.

4.7 Acknowledgments

We thank Roger Innes for providing the mCherry-SYP61 and ARA6-mCherry constructs, Manfred Heinlein for the MP:RFP construct, and Federica Brandizzi for the ERD2-YFP and YFP-Sec24 constructs. We thank Sören Werner and Thomas Schmülling for providing the 35S:CKX1-myc construct.

4.8 References

Abas, L., and Luschnig, C. (2010). Maximum yields of microsomal-type membranes from small amounts of plant material without requiring ultracentrifugation. Anal. Biochem. *401*, 217–227.

Anderson, S.M., Mueller, B.K., Lange, E.J., and Senes, A. (2017). Combination of Cα-H Hydrogen Bonds and van der Waals Packing Modulates the Stability of GxxxG-Mediated Dimers in Membranes. J. Am. Chem. Soc. *139*, 15774–15783.

van Anken, E., and Braakman, I. (2005). Versatility of the endoplasmic reticulum protein folding factory. Crit. Rev. Biochem. Mol. Biol. *40*, 191–228.

Antoniadi, I., Plačková, L., Simonovik, B., Doležal, K., Turnbull, C., Ljung, K., and Novák, O. (2015). Cell-Type-Specific Cytokinin Distribution within the Arabidopsis Primary Root Apex. Plant Cell *27*, 1955–1967.

Barrieu, F., and Chrispeels, M.J. (1999). Delivery of a secreted soluble protein to the vacuole via a membrane anchor. Plant Physiol. *120*, 961–968.

Bartrina, I., Otto, E., Strnad, M., Werner, T., and Schmülling, T. (2011). Cytokinin regulates the activity of reproductive meristems, flower organ size, ovule formation, and thus seed yield in Arabidopsis thaliana. Plant Cell *23*, 69–80.

Bi, G., Liebrand, T.W.H., Bye, R.R., Postma, J., van der Burgh, A.M., Robatzek, S., Xu, X., and Joosten, M.H.A.J. (2016). SOBIR1 requires the GxxxG dimerization motif in its transmembrane domain to form constitutive complexes with receptor-like proteins. Mol. Plant Pathol. *17*, 96–107.

Bilyeu, K.D., Cole, J.L., Laskey, J.G., Riekhof, W.R., Esparza, T.J., Kramer, M.D., and Morris, R.O. (2001). Molecular and biochemical characterization of a cytokinin oxidase from maize. Plant Physiol. *125*, 378–386.

Boulaflous, A., Saint-Jore-Dupas, C., Herranz-Gordo, M.-C., Pagny-Salehabadi, S., Plasson, C., Garidou, F., Kiefer-Meyer, M.-C., Ritzenthaler, C., Faye, L., and Gomord, V. (2009). Cytosolic N-terminal arginine-based signals together with a luminal signal target a type II membrane protein to the plant ER. BMC Plant Biol. *9*, 144.

Boutant, E., Didier, P., Niehl, A., Mély, Y., Ritzenthaler, C., and Heinlein, M. (2010). Fluorescent protein recruitment assay for demonstration and analysis of in vivo protein interactions in plant cells and its application to Tobacco mosaic virus movement protein. Plant J. Cell Mol. Biol. *62*, 171–177.

Brandizzi, F., Frangne, N., Marc-Martin, S., Hawes, C., Neuhaus, J.-M., and Paris, N. (2002). The destination for single-pass membrane proteins is influenced markedly by the length of the hydrophobic domain. Plant Cell *14*, 1077–1092.

Caesar, K., Thamm, A.M.K., Witthöft, J., Elgass, K., Huppenberger, P., Grefen, C., Horak, J., and Harter, K. (2011). Evidence for the localization of the Arabidopsis cytokinin receptors AHK3 and AHK4 in the endoplasmic reticulum. J. Exp. Bot. *62*, 5571–5580.

Chevalier, A.S., and Chaumont, F. (2015). Trafficking of plant plasma membrane aquaporins: multiple regulation levels and complex sorting signals. Plant Cell Physiol. *56*, 819–829.

Choma, C., Gratkowski, H., Lear, J.D., and DeGrado, W.F. (2000). Asparagine-mediated self-association of a model transmembrane helix. Nat. Struct. Biol. *7*, 161–166.

Chung, I., Akita, R., Vandlen, R., Toomre, D., Schlessinger, J., and Mellman, I. (2010). Spatial control of EGF receptor activation by reversible dimerization on living cells. Nature *464*, 783–787.

Cosson, P., Perrin, J., and Bonifacino, J.S. (2013). Anchors aweigh: protein localization and transport mediated by transmembrane domains. Trends Cell Biol. *23*, 511–517.

De Marcos Lousa, C., Gershlick, D.C., and Denecke, J. (2012). Mechanisms and concepts paving the way towards a complete transport cycle of plant vacuolar sorting receptors. Plant Cell *24*, 1714–1732.

Dews, I.C., and Mackenzie, K.R. (2007). Transmembrane domains of the syndecan family of growth factor coreceptors display a hierarchy of homotypic and heterotypic interactions. Proc. Natl. Acad. Sci. U. S. A. *104*, 20782–20787.

Fiebiger, E., Hirsch, C., Vyas, J.M., Gordon, E., Ploegh, H.L., and Tortorella, D. (2004). Dissection of the dislocation pathway for type I membrane proteins with a new small molecule inhibitor, eeyarestatin. Mol. Biol. Cell *15*, 1635–1646.

Fritz-Laylin, L.K., Krishnamurthy, N., Tör, M., Sjölander, K.V., and Jones, J.D.G. (2005). Phylogenomic analysis of the receptor-like proteins of rice and Arabidopsis. Plant Physiol. *138*, 611–623.

Fujiki, Y., Hubbard, A.L., Fowler, S., and Lazarow, P.B. (1982). Isolation of intracellular membranes by means of sodium carbonate treatment: application to endoplasmic reticulum. J. Cell Biol. *93*, 97–102.

Fusseder, A., and Ziegler, P. (1988). Metabolism and compartmentation of dihydrozeatin exogenously supplied to photoautotrophic suspension cultures of Chenopodium rubrum. Planta *173*, 104–109.

Galuszka, P., Frébortová, J., Luhová, L., Bilyeu, K.D., English, J.T., and Frébort, I. (2005). Tissue localization of cytokinin dehydrogenase in maize: possible involvement of quinone species generated from plant phenolics by other enzymatic systems in the catalytic reaction. Plant Cell Physiol. *46*, 716–728.

Galuszka, P., Popelková, H., Werner, T., Frébortová, J., Pospíšilová, H., Mik, V., Köllmer, I., Schmülling, T., and Frébort, I. (2007). Biochemical Characterization of Cytokinin Oxidases/Dehydrogenases from Arabidopsis thaliana Expressed in Nicotiana tabacum L. J. Plant Growth Regul. *26*, 255–267.

Gao, C., Cai, Y., Wang, Y., Kang, B.-H., Aniento, F., Robinson, D.G., and Jiang, L. (2014). Retention mechanisms for ER and Golgi membrane proteins. Trends Plant Sci. *19*, 508–515.

Gimpelev, M., Forrest, L.R., Murray, D., and Honig, B. (2004). Helical packing patterns in membrane and soluble proteins. Biophys. J. *87*, 4075–4086.

Gookin, T.E., and Assmann, S.M. (2014). Significant reduction of BiFC non-specific assembly facilitates in planta assessment of heterotrimeric G-protein interactors. Plant J. Cell Mol. Biol. *80*, 553–567.

Gu, Y., and Innes, R.W. (2011). The KEEP ON GOING protein of Arabidopsis recruits the ENHANCED DISEASE RESISTANCE1 protein to trans-Golgi network/early endosome vesicles. Plant Physiol. *155*, 1827–1838.

Gu, Y., and Innes, R.W. (2012). The KEEP ON GOING protein of Arabidopsis regulates intracellular protein trafficking and is degraded during fungal infection. Plant Cell *24*, 4717–4730.

Hanton, S.L., Matheson, L.A., and Brandizzi, F. (2006). Seeking a way out: export of proteins from the plant endoplasmic reticulum. Trends Plant Sci. *11*, 335–343.

von Heijne, G., and Gavel, Y. (1988). Topogenic signals in integral membrane proteins. Eur. J. Biochem. *174*, 671–678.

Hoang, T., Kuljanin, M., Smith, M.D., and Jelokhani-Niaraki, M. (2015). A biophysical study on molecular physiology of the uncoupling proteins of the central nervous system. Biosci. Rep. *35*.

Holst, K., Schmülling, T., and Werner, T. (2011). Enhanced cytokinin degradation in leaf primordia of transgenic Arabidopsis plants reduces leaf size and shoot organ primordia formation. J. Plant Physiol. *168*, 1328–1334.

Houba-Hérin, N., Pethe, C., d'Alayer, J., and Laloue, M. (1999). Cytokinin oxidase from Zea mays: purification, cDNA cloning and expression in moss protoplasts. Plant J. Cell Mol. Biol. *17*, 615–626.

Huang, L., Franklin, A.E., and Hoffman, N.E. (1993). Primary structure and characterization of an Arabidopsis thaliana calnexin-like protein. J. Biol. Chem. *268*, 6560–6566.

Hwang, I., Sheen, J., and Müller, B. (2012). Cytokinin signaling networks. Annu. Rev. Plant Biol. *63*, 353–380.

Inoue, T., Higuchi, M., Hashimoto, Y., Seki, M., Kobayashi, M., Kato, T., Tabata, S., Shinozaki, K., and Kakimoto, T. (2001). Identification of CRE1 as a cytokinin receptor from Arabidopsis. Nature *409*, 1060–1063.

Jiskrová, E., Novák, O., Pospíšilová, H., Holubová, K., Karády, M., Galuszka, P., Robert, S., and Frébort, I. (2016). Extra- and intracellular distribution of cytokinins in the leaves of monocots and dicots. New Biotechnol. *33*, 735–742.

Johnson, R.M., Hecht, K., and Deber, C.M. (2007). Aromatic and cation-pi interactions enhance helix-helix association in a membrane environment. Biochemistry *46*, 9208–9214.

Jurgens, G. (2004). Membrane trafficking in plants. Annu. Rev. Cell Dev. Biol. *20*, 481–504.

Karimi, M., Inzé, D., and Depicker, A. (2002). GATEWAY vectors for Agrobacterium-mediated plant transformation. Trends Plant Sci. *7*, 193–195.

Kiran, N.S., Benková, E., Reková, A., Dubová, J., Malbeck, J., Palme, K., and Brzobohatý, B. (2012). Retargeting a maize β-glucosidase to the vacuole--evidence from intact plants that zeatin-O-glucoside is stored in the vacuole. Phytochemistry *79*, 67–77.

Köllmer, I., Novák, O., Strnad, M., Schmülling, T., and Werner, T. (2014). Overexpression of the cytosolic cytokinin oxidase/dehydrogenase (CKX7) from Arabidopsis causes specific changes in root growth and xylem differentiation. Plant J. Cell Mol. Biol. *78*, 359–371.

Kowalska, M., Galuszka, P., Frébortová, J., Šebela, M., Béres, T., Hluska, T., Šmehilová, M., Bilyeu, K.D., and Frébort, I. (2010). Vacuolar and cytosolic cytokinin dehydrogenases of Arabidopsis thaliana: heterologous expression, purification and properties. Phytochemistry *71*, 1970–1978.

Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778–795.

Kulp, D.W., Subramaniam, S., Donald, J.E., Hannigan, B.T., Mueller, B.K., Grigoryan, G., and Senes, A. (2012). Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL). J. Comput. Chem. *33*, 1645–1661.

Kunji, E.R.S., Harding, M., Butler, P.J.G., and Akamine, P. (2008). Determination of the molecular mass and dimensions of membrane proteins by size exclusion chromatography. Methods San Diego Calif *46*, 62–72.

Kwon, M.-J., Choi, Y., Yun, J.-H., Lee, W., Han, I.-O., and Oh, E.-S. (2015). A unique phenylalanine in the transmembrane domain strengthens homodimerization of the syndecan-2 transmembrane domain and functionally regulates syndecan-2. J. Biol. Chem. *290*, 5772–5782.

Langhans, M., Marcote, M.J., Pimpl, P., Virgili-López, G., Robinson, D.G., and Aniento, F. (2008). In vivo trafficking and localization of p24 proteins in plant cells. Traffic Cph. Den. *9*, 770–785.

Lerich, A., Langhans, M., Sturm, S., and Robinson, D.G. (2011). Is the 6 kDa tobacco etch viral protein a bona fide ERES marker? J. Exp. Bot. *62*, 5013–5023.

Li, E., and Hristova, K. (2006). Role of receptor tyrosine kinase transmembrane domains in cell signaling and human pathologies. Biochemistry *45*, 6241–6251.

Li, E., Wimley, W.C., and Hristova, K. (2012). Transmembrane helix dimerization: beyond the search for sequence motifs. Biochim. Biophys. Acta *1818*, 183–193.

Lomin, S.N., Yonekura-Sakakibara, K., Romanov, G.A., and Sakakibara, H. (2011). Ligand-binding properties and subcellular localization of maize cytokinin receptors. J. Exp. Bot. *62*, 5149–5159.

MacKenzie, K.R., Prestegard, J.H., and Engelman, D.M. (1997). A transmembrane helix dimer: structure and implications. Science *276*, 131–133.

MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B *102*, 3586–3616.

Maruyama, I.N. (2015). Activation of transmembrane cell-surface receptors via a common mechanism? The “rotation model.” BioEssays News Rev. Mol. Cell. Dev. Biol. *37*, 959–967.

Miyawaki, A., Sawano, A., and Kogure, T. (2003). Lighting up cells: labelling proteins with fluorophores. Nat. Cell Biol. *Suppl*, S1-7.

Moore, D.T., Berger, B.W., and DeGrado, W.F. (2008). Protein-protein interactions in the membrane: sequence, structural, and biological motifs. Struct. Lond. Engl. 1993 *16*, 991–1001.

Mothes, W., Heinrich, S.U., Graf, R., Nilsson, I., von Heijne, G., Brunner, J., and Rapoport, T.A. (1997). Molecular mechanism of membrane protein integration into the endoplasmic reticulum. Cell *89*, 523–533.

Mueller, B.K., Subramaniam, S., and Senes, A. (2014). A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα-H hydrogen bonds. Proc. Natl. Acad. Sci. U. S. A. *111*, E888-895.

Niehl, A., Amari, K., Gereige, D., Brandner, K., Mély, Y., and Heinlein, M. (2012). Control of Tobacco mosaic virus movement protein fate by CELL-DIVISION-CYCLE protein48. Plant Physiol. *160*, 2093–2108.

Niemann, M.C.E., Bartrina, I., Ashikov, A., Weber, H., Novák, O., Spíchal, L., Strnad, M., Strasser, R., Bakker, H., Schmülling, T., et al. (2015). Arabidopsis ROCK1 transports UDP-GlcNAc/UDP-GalNAc and regulates ER protein quality control and cytokinin activity. Proc. Natl. Acad. Sci. U. S. A. *112*, 291–296.

Obrdlik, P., Neuhaus, G., and Merkle, T. (2000). Plant heterotrimeric G protein beta subunit is associated with membranes via protein interactions involving coiled-coil formation. FEBS Lett. *476*, 208–212.

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

Rögner, M. (2000). Size exclusion chromatography. J. Chromatogr. Libr. *61*, 89–145.

Rojo, E., and Denecke, J. (2008). What is moving in the secretory pathway of plants? Plant Physiol. *147*, 1493–1503.

Römisch, K. (2005). Endoplasmic reticulum-associated degradation. Annu. Rev. Cell Dev. Biol. *21*, 435–456.

Russ, W.P., and Engelman, D.M. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. J. Mol. Biol. *296*, 911–919.

Sambade, A., Brandner, K., Hofmann, C., Seemanpillai, M., Mutterer, J., and Heinlein, M. (2008). Transport of TMV movement protein particles associated with the targeting of RNA to plasmodesmata. Traffic Cph. Den. *9*, 2073–2088.

von Schaewen, A., Sturm, A., O’Neill, J., and Chrispeels, M.J. (1993). Isolation of a mutant Arabidopsis plant that lacks N-acetyl glucosaminyl transferase I and is unable to synthesize Golgi-modified complex N-linked glycans. Plant Physiol. *102*, 1109–1118.

Schekman, R., and Orci, L. (1996). Coat proteins and vesicle budding. Science *271*, 1526–1533.

Schmülling, T., Werner, T., Riefler, M., Krupková, E., and Bartrina y Manns, I. (2003). Structure and function of cytokinin oxidase/dehydrogenase genes of maize, rice, Arabidopsis and other species. J. Plant Res. *116*, 241–252.

Schook, W., Puszkin, S., Bloom, W., Ores, C., and Kochwa, S. (1979). Mechanochemical properties of brain clathrin: interactions with actin and alpha-actinin and polymerization into basketlike structures or filaments. Proc. Natl. Acad. Sci. U. S. A. *76*, 116–120.

Schwacke, R., Schneider, A., van der Graaff, E., Fischer, K., Catoni, E., Desimone, M., Frommer, W.B., Flügge, U.-I., and Kunze, R. (2003). ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. Plant Physiol. *131*, 16–26.

Senes, A., Gerstein, M., and Engelman, D.M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. J. Mol. Biol. *296*, 921–936.

Senes, A., Ubarretxena-Belandia, I., and Engelman, D.M. (2001). The Calpha ---H...O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. Proc. Natl. Acad. Sci. U. S. A. *98*, 9056–9061.

Senes, A., Engel, D.E., and DeGrado, W.F. (2004). Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. Curr. Opin. Struct. Biol. *14*, 465–479.

Sparkes, I.A., Runions, J., Kearns, A., and Hawes, C. (2006). Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. Nat. Protoc. *1*, 2019–2025.

Stefano, G., Renna, L., Chatre, L., Hanton, S.L., Moreau, P., Hawes, C., and Brandizzi, F. (2006). In tobacco leaf epidermal cells, the integrity of protein export from the endoplasmic reticulum and of ER export sites depends on active COPI machinery. Plant J. Cell Mol. Biol. *46*, 95–110.

Subramaniam, S., and Senes, A. (2012). An energy-based conformer library for side chain optimization: improved prediction and adjustable sampling. Proteins *80*, 2218–2234.

Sun, Q., Ju, T., and Cummings, R.D. (2011). The transmembrane domain of the molecular chaperone Cosmc directs its localization to the endoplasmic reticulum. J. Biol. Chem. *286*, 11529–11542.

Suzuki, T., Miwa, K., Ishikawa, K., Yamada, H., Aiba, H., and Mizuno, T. (2001). The Arabidopsis sensor His-kinase, AHk4, can respond to cytokinins. Plant Cell Physiol. *42*, 107–113.

Svačinová, J., Novák, O., Plačková, L., Lenobel, R., Holík, J., Strnad, M., and Doležal, K. (2012). A new approach for cytokinin isolation from Arabidopsis tissues using miniaturized purification: pipette tip solid-phase extraction. Plant Methods *8*, 17.

Tokunaga, F., Brostrom, C., Koide, T., and Arvan, P. (2000). Endoplasmic reticulum (ER)-associated degradation of misfolded N-linked glycoproteins is suppressed upon inhibition of ER mannosidase I. J. Biol. Chem. *275*, 40757–40764.

Tu, L., and Banfield, D.K. (2010). Localization of Golgi-resident glycosyltransferases. Cell. Mol. Life Sci. CMLS *67*, 29–41.

Ueda, T., Yamaguchi, M., Uchimiya, H., and Nakano, A. (2001). Ara6, a plant-unique novel type Rab GTPase, functions in the endocytic pathway of Arabidopsis thaliana. EMBO J. *20*, 4730–4741.

Uemura, T., Ueda, T., Ohniwa, R.L., Nakano, A., Takeyasu, K., and Sato, M.H. (2004). Systematic analysis of SNARE molecules in Arabidopsis: dissection of the post-Golgi network in plant cells. Cell Struct. Funct. *29*, 49–65.

Vembar, S.S., and Brodsky, J.L. (2008). One step at a time: endoplasmic reticulum-associated degradation. Nat. Rev. Mol. Cell Biol. *9*, 944–957.

Walters, R.F.S., and DeGrado, W.F. (2006). Helix-packing motifs in membrane proteins. Proc. Natl. Acad. Sci. U. S. A. *103*, 13658–13663.

Wang, Q., Shinkre, B.A., Lee, J., Weniger, M.A., Liu, Y., Chen, W., Wiestner, A., Trenkle, W.C., and Ye, Y. (2010). The ERAD inhibitor Eeyarestatin I is a bifunctional compound with a membrane-binding domain and a p97/VCP inhibitory group. PloS One *5*, e15479.

Welinder, C., and Ekblad, L. (2011). Coomassie staining as loading control in Western blot analysis. J. Proteome Res. *10*, 1416–1419.

Werner, T., and Schmülling, T. (2009). Cytokinin action in plant development. Curr. Opin. Plant Biol. *12*, 527–538.

Werner, T., Motyka, V., Strnad, M., and Schmülling, T. (2001). Regulation of plant growth by cytokinin. Proc. Natl. Acad. Sci. U. S. A. *98*, 10487–10492.

Werner, T., Motyka, V., Laucou, V., Smets, R., Van Onckelen, H., and Schmülling, T. (2003). Cytokinin-deficient transgenic Arabidopsis plants show multiple developmental alterations indicating opposite functions of cytokinins in the regulation of shoot and root meristem activity. Plant Cell *15*, 2532–2550.

Werner, T., Nehnevajova, E., Köllmer, I., Novák, O., Strnad, M., Krämer, U., and Schmülling, T. (2010). Root-specific reduction of cytokinin causes enhanced root growth, drought tolerance, and leaf mineral enrichment in Arabidopsis and tobacco. Plant Cell *22*, 3905–3920.

Wulfetange, K., Lomin, S.N., Romanov, G.A., Stolz, A., Heyl, A., and Schmülling, T. (2011). The cytokinin receptors of Arabidopsis are located mainly to the endoplasmic reticulum. Plant Physiol. *156*, 1808–1818.

Xu, J., Peng, H., Chen, Q., Liu, Y., Dong, Z., and Zhang, J.-T. (2007). Oligomerization domain of the multidrug resistance-associated transporter ABCG2 and its dominant inhibitory activity. Cancer Res. *67*, 4373–4381.

Zhang, Y., Yang, Y., Fang, B., Gannon, P., Ding, P., Li, X., and Zhang, Y. (2010). Arabidopsis snc2-1D activates receptor-like protein-mediated immunity transduced through WRKY70. Plant Cell *22*, 3153–3163.

Zhou, F.X., Cocco, M.J., Russ, W.P., Brunger, A.T., and Engelman, D.M. (2000). Interhelical hydrogen bonding drives strong interactions in membrane proteins. Nat. Struct. Biol. *7*, 154–160.

Chapter 5: Future Directions and Continuing Work

5.1 CATM predictions for model organism genomes

In Chapters 2 and 3, I discussed the use of the human genome as a template for generating potential GAS_{right} transmembrane (TM) domains (TMD). Theoretically, we could have used random combinations of hydrophobic amino acids that contain a GxxxG-like motif to analyze experimentally, but we chose to select proteins derived from the human genome because they are more likely to have biological significance. Membrane proteins, and in particular single pass membrane proteins (SPMP) are important therapeutic targets especially in the case of cell receptors (see Chapter 1.1.1 for further discussion). However, the vast majority of biological experimentation that leads to therapeutic design occurs in non-human organisms when human experimentation is infeasible or unethical. Therefore, it is important to understand the GAS_{right} structures found in a variety of model organisms. To this aim, I performed similar computational analysis of the proteomes of several model organisms.

The model organisms selected for analysis were *Escherichia coli, Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Danio rerio, Xenopus laevis,* and *Homo sapiens*. The genomes were extracted from the Uniprot database on 11/02/2016 and filtered for proteins annotated to have a single TMD. Though individual proteome size varies by genome complexity, there were approximately 5,550 TMDs in total to pass through the CATM algorithm (Table 5.1). Like the human genome, described in Chapter 2, all of the model organisms had 1/3 to ½ of their SPMP proteomes create potential GAS_{right} dimers, indicated by negative CATM scores. The score distribution of each model organism can be found in Figure 5.1 which indicates that the scores, regardless of the organism, follow a rough bell-curve distribution with peaks between -15 and -25 kcal/mol.

I implemented a filtering system to select TM sequences suitable for experimental characterization. Following the same standardization scheme as found in Chapter 2, stitched the interfacial residues predicted by CATM into a poly-Leucine backbone, and ran the new

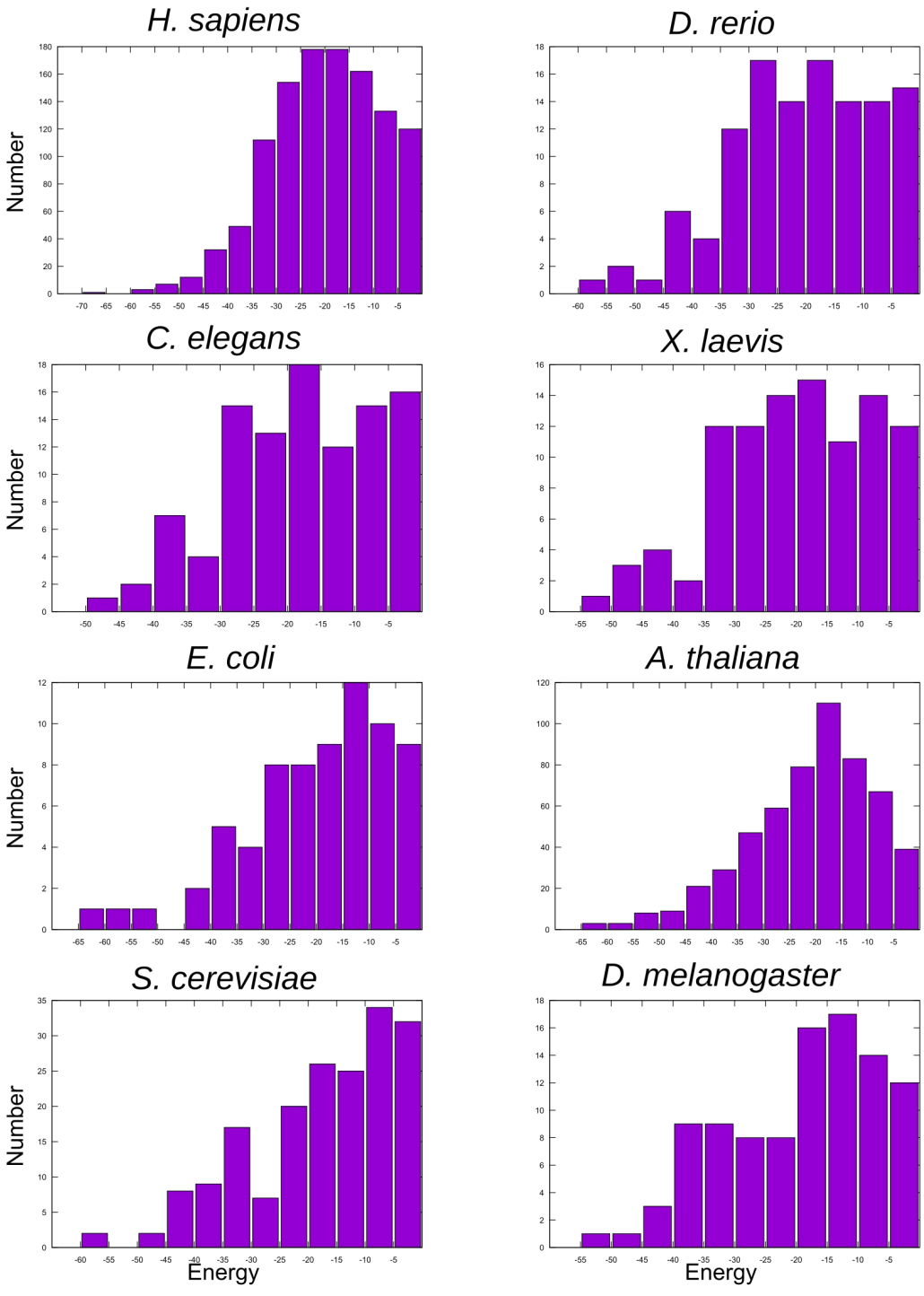
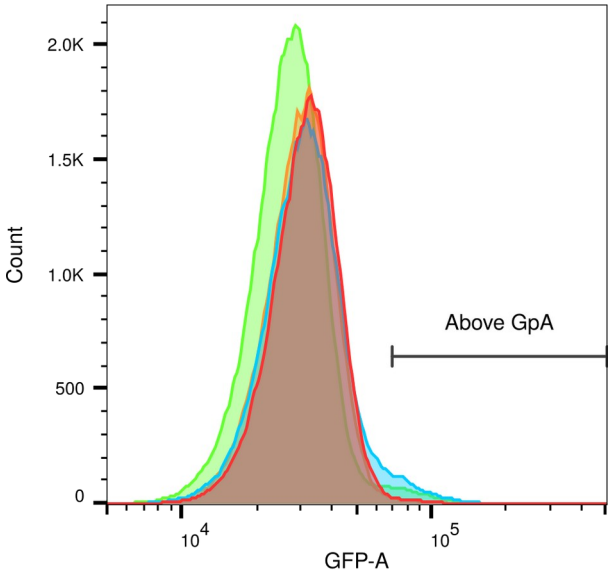


Figure 5.1 CATM score distributions for model organisms. For each model organism, the SPMP proteome, as annotated by Uniprot in 2016, was run through the CATM algorithm. The histograms show the distributions of CATM scores for each genome. Note that the scales for each of these graphs differ because the size of the genomes vary greatly.

sequence through CATM. Some of the redesigned TM sequences were no longer predicted to assocatiate, though many of these were originally at the bottom of the distribution (closer to 0 kcal/mol; Table 5.1). From the remaining 2,124 sequences, I filtered out sequences that had polar residues (932), that had 100% sequence redundancy (144), were previously tested (70), and in which the interface identified in the poly-Leucine backbone stretched into the base TOXGREEN construct (118). This resulted in 860 constructs to test experimentally along with their C1 and N1 monomerizing mutants. These constructs were divided into five segments ordered in an oligo pool and 4/5 of these have been cloned and are ready to experimentally test via the sort-seq method described in Chapter 3. Preliminary flow cytometry results of the segments demonstrate a low, narrow distribution of fluorescence with a slightly extended high fluorescence tail (Figure 5.2).



	Sample Name	Median GFP	% above GpA
	2.4	31676	0.81
	2.3	31181	3.39
	2.5	30385	0.84
	2.7	27215	1.82

Figure 5.2 Preliminary genome distributions. The flow cytometry distributions are overlaid and each segment is denoted as 2.X The amount of the library that is above the median GpA value is shown as a percentage of the total library.

The evolutionary connections between these various organisms may provide insight into GAS_{right} structures. A protein that shares homology, but not a predicted interface could be a way to implicitly filter out false positives from the CATM algorithm. Alternatively, if several homologs share the ability to dimerize, this may provide evidence that it is biologically important to the function of the protein. Extensive testing of the proteomes of model organisms will provide experimental information on evolutionary comparisons and biological processes.

5.2 Training the CATM algorithm

The focus of Chapter 2 is understanding the sequence, structure, and energetic trends that underlay GAS_{right} dimerization propensity. An original aim of the paper was to use the TOXCAT results to train the CATM algorithm to better predict the dimerization propensity of transmembrane sequences. With only 26 points, however, any type of training would have overfitted the limited amount of data we had. The assay described in Chapter 3, will aid in this endeavor by providing us with a larger number of experimental measurements to perform a robust training of CATM.

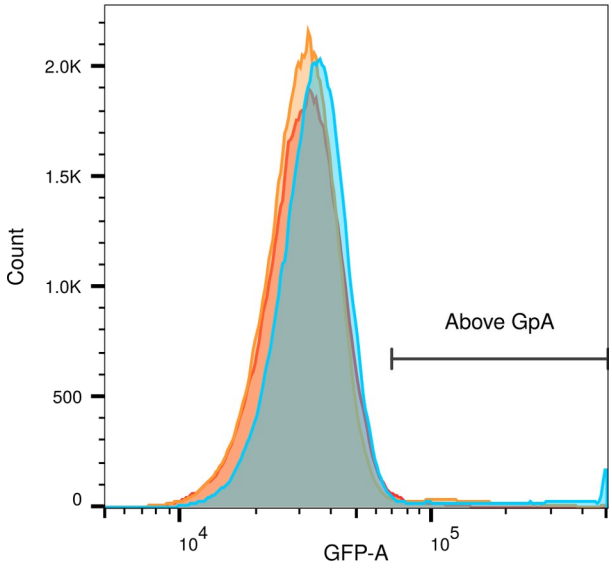
The set of experimental constructs that will be used to train the CATM algorithm are derived from the standardized constructs found in Chapter 2. I selected 40 of these transmembrane domains to analyze experimentally and computationally. I chose the 26 successful sequences, as well as some that were removed as described in the supplemental material. I used the same normalization strategy by stitching the interfaces into a poly-Leucine backbone (Fig 2.3), but I also included three other backbone sequences:

```
RASLIV..LL..IL..VV..LILI
RASLIL..VI..VV..LL..LILI
RASVIL..LV..VI..LL..LILI
```

These backbone sequences were designed by evaluating the average composition of the common nonpolar amino acids isoleucine (22%), leucine (48%), and valine (30%) in transmembrane domains and using a proportionate number of those residues. These values were obtained by calculating the distribution of amino acids in the human SPMP proteome and looking at the proportion of the hydrophobic ratios. There are 4 leucines, 2 isoleucines, and 3 valines in each of the tested sequences. 21 different sequence combinations were run through the CATM algorithm with the GpA interface stitched into the backbone (Table 5.2). I selected the sequences with the top three CATM scores for further analysis.

For each sequence, I also performed 49 mutations: 4 at each interface position and 17 total for the other positions. The mutants were chosen by substituting various hydrophobic and small

residues (A, C, F, G, I, L, M, S, V) at each interfacial position and ranking them by CATM score. To restrict the library to a manageable number of mutations, we selected four of those mutants that were evenly spread across the CATM range of scores. The non-interfacial positions were mutated to alaine and leucine. When the original amino acid was leucine, it was mutated to alanine and isoleucine. These constructs were ordered in an oligo pool and cloned into the pccGFPKan backbone to analyze using the sort-seq method. The successfully cloned segments are ready to experimentally test via the sort-seq method described in Chapter 3. Preliminary flow cytometry results of two of the segments demonstrate a low, narrow distribution of fluorescence with a slightly extended high fluorescence tail (Figure 5.3).



	Sample Name	Median GFP	% above GpA
	2.8	34774	4.02
	2.9	31111	2.96
	2.17	31534	1.37

Figure 5.3 Preliminary library distributions. The flow cytometry distributions are overlaid and each segment is denoted as 2.X. The amount of the library that is above the median GpA value is shown as a percentage of the total library.

Once the constructs are experimentally assessed, the CATM training can be performed. The current energy function of CATM (Mueller et al., 2014) is relatively simple, including an unweighted sum of van der Waals interactions (CHARMM 22 (MacKerell et al., 1998)), hydrogen bonding (SCRWL 4 (Krivov et al., 2009a)), and solvation (IMM1 (Lazaridis, 2003)). Its expansion may include terms such as electrostatics, potential energy functions for membrane insertion and tilting (Hessa et al., 2007; Lomize et al., 2006; Schramm et al., 2012; Senes et al., 2007; Ulmschneider et al., 2005), as well as statistical rotameric energies (eq. 5.1) (Das and Baker, 2008). The objective (eq. 5.2) is to identify the set of energy terms

{$E_1 \ldots E_n$} and relative "weights" {$w_1 \ldots w_n$} that minimize the difference between the experimental apparent free energies and the computational energies calculated with CATM across the entire set of constructs *i*.

$$(1) \quad E^i_{CATM} = \sum_{j=1}^{n} w_j E^i_j = w_{vdw} E^i_{vdw} + w_{Hb} E^i_{Hb} + w_{solv} E^i_{solv} + w_{elec} E^i_{elect} + w_{insert} E^i_{insert} + w_{rot} E^i_{rot} + \ldots + w_n E^i_n$$

$$(2) \quad \min_{\{w_1 \ldots w_n\}} \left(\sum_i \left| \Delta G^{app,i}_{dimerization} - \sum_{j=1}^{n} w_j E^i_j \right| \right)$$

There are a variety of methods to train the algorithm that will be left to others. One method is to perform the fitting in rounds, using rigid-body energy evaluations on the structural models. As the scoring function is updates, the TMD models will also change to find the new optimum. Therefore each round will recalculate and reevaluate the models with the updated force field, until convergence is reached. During the training, a subset of 10-20% of the constructs should be set aside for testing to avoid overfitting the data. Graduate student Joshua Choi will be continuing this project.

5.3 Potential improvements of the TOXGREEN assay

As discussed in Section 1.4.4, there have been a variety of ToxR-based assays to measure helix-helix association in the membrane. Much of Chapter 3 discussed how one version of that assay, TOXGREEN, was scaled up to work for the sort-seq method. Using a fluorescent reporter gene, many of the other versions could be scaled up as well. However, the overall genetic set up of this assay is still somewhat limiting. This section will address two problems with these assays: measuring TM helix expression and insertion and low dynamic range of fluorescence.

The first, and most pressing, problem is to accurately measure both the transmembrane helix expression and insertion rates. This problem is of critical importance because dimerization propensity is directly related to protein concentration, and in particular, the protein concentration in the membrane (Russ and Engelman, 1999). Ensuring that a TMD is inserted into a membrane is also important because a cytoplasmic TM can still induce reporter gene expression if it is aggregated. In the traditional versions of this assay, expression and insertion are measured through Western blots and a MalE complementation assay, respectively (Anderson et al., 2017). Western blots have even been used to measure expression for the calculation of free energies of dimerization (Duong et al., 2007). These methods, however, are made to measure individual proteins and cannot easily be scaled up. In Chapter 3, the MalE complementation assay was used as a selection method to exclude TM sequences that did not insert into the membrane at all. Unfortunately, this method only gives a binary response rather than a quantification of insertion or expression. It is important to quantify both of these aspects because a slight defect in insertion could be made up for by increased expression. I have designed two strategic solutions to this problem and it will be up to the reader to implement these methods in the future (Fig. 5.4).

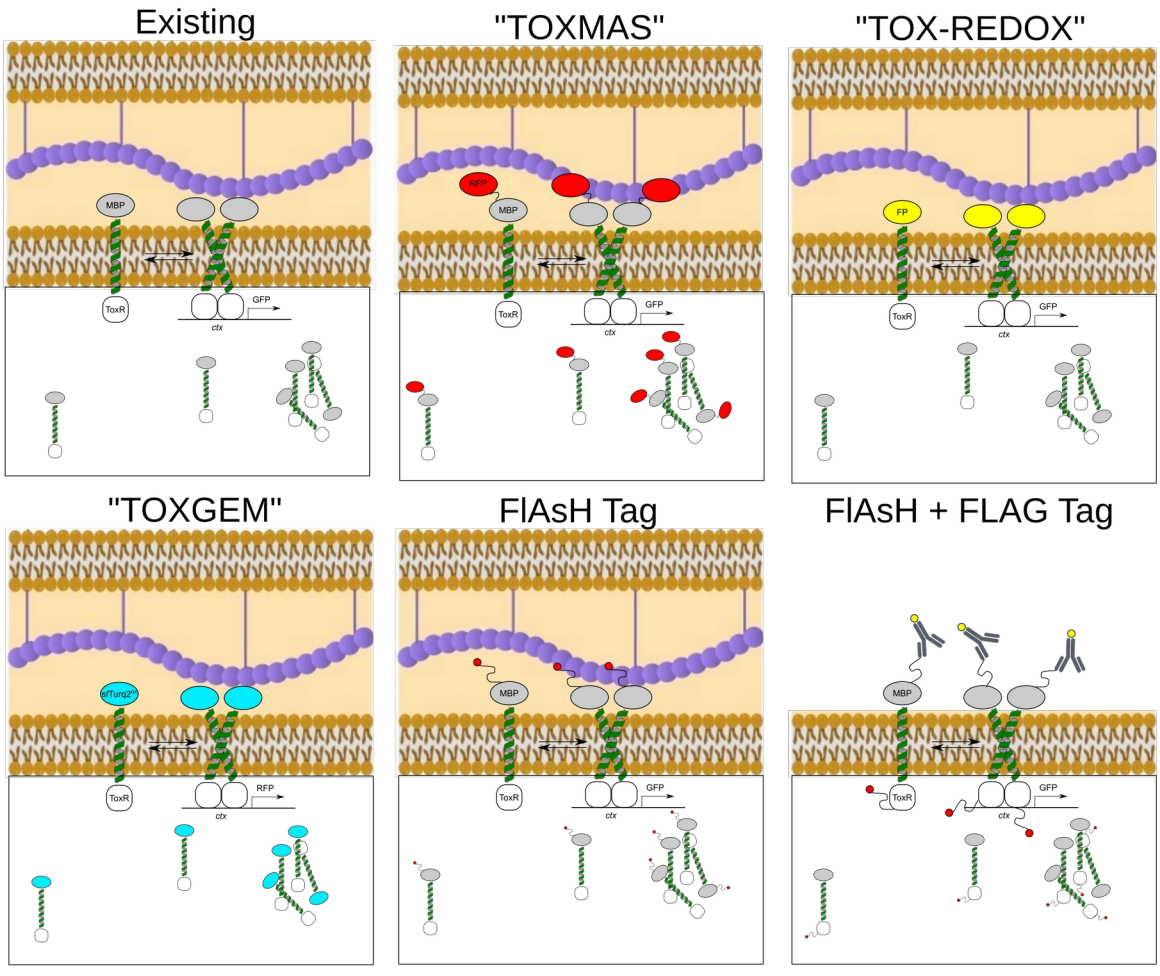


Figure 5.4 Potential derivatives of the TOXGREEN assay. Each panel shows the inner and outer membranes of *E. coli* and the rest of the cell is shown in the white box beneath the inner membrane. In the "Existing" panel, the TM domain on interest is fused to the ToxR transcription factor and a maltose binding protein (MBP). When the TM helices dimerize, ToxR binds to the *ctx* promoter and turns on the reporter gene, GFP. This assay is known as TOXGREEN. The helices in the cytoplasm demonstrate that not all of the protein correctly inserts in the membrane. The following panels demonstrate potential derivatives of this traditional version. In TOXMAS, RFP is fused to MBP in a failed attempt to measure protein expression. TOX-REDOX is a method that uses an FP that turns on in the periplasm, but off in the cytoplasm based on the oxidizing environment. TOXGEM uses the periplasm optimized sfTurq2^{ox} to measure protein expression and alters the reporter gene to RFP. The FlAsH and FLAG tag panels are able to measure expression and insertion respectively.

The simpler of the two strategies is to replace and/or add specific domains in the assay to make them more compatible with flow cytometry. A redox sensitive FP (roGFP) was developed to monitor redox equilibrium in mammalian cells (Dooley et al., 2004). This version of roGFP fluoresces more in the periplasm than in the cytoplasm and it could be used to replace MBP (TOX-REDOX). However, the ratio of periplasmic to cytoplasmic fluorescence could not be elucidated via flow cytometry. Alternatively, a dual reporter could measure total expression of the ToxR construct with a different FP. Previously, Claire Armstrong in the Senes lab attempted to create an assay where a red fluorescent protein was fused to MBP to measure expression (TOXMAS). This method failed because many fluorescent proteins do not fold properly in the periplasm due to the oxidizing environment (Meiresonne et al., 2017). A cyan FP version was recently developed, however that folds and fluoresces as well in the periplasm as it does in the cytoplasm (Meiresonne et al., 2019). I obtained this plasmid and, together with the help of high school student Amanda Cook, was able to clone this superfolder mTurquoise2^{ox} in place of the traditional MBP periplasmic domain for some control samples (TOXGEM). Evaluation of this construct for insertion and expression measurements is currently underway.

The second strategy to simultaneously measure expression and insertion in high-throughput experiments is to add peptide tags to the ToxR construct so that fluorescent bodies can be attached. One possibility would be to use the FlAsH/ReAsH tetracysteine-based protein detection system (Thermo Fisher Scientific). In this system, an epitope tag would be added to the ToxR construct and a membrane-permeable ligand mixed with the cells. When the ligand binds the tag, it would fluoresce, allowing quantification of protein expression. A traditional version of this method is to add an epitope tag, like FLAG, to the end of a protein, permeabilize cells, and add a fluorescent antibody that binds to this tag. This method is primarily performed in mammalian cells that can withstand a certain level of permeabilization and still survive. In *E. coli*, this becomes much more challenging due to the small size of bacterial cells as well as the presence of the outer membrane. Instead, if the outer membrane and periplasm are removed

from the cells, the fluorescent antibodies could bind the exterior facing MBP to measure the concentration of inserted ToxR constructs. By combining these two methods, FlAsH and FLAG, it would be possible to measure expression, insertion, and dimerization in the ToxR system with three separate fluorophores. Though flow cytometers can easily measure three different fluorophores, it will add the complication of compensation. Compensation is required when correcting for spectral overlap between multiple fluorophores.

The second drawback of the current TOXGREEN assay is that the reporter gene expression is quite low, rendering the fluorescence dynamic range lower than is optimal. Currently, the SONY machine is set at a 90% gain for GFP and my positive and negative controls only cover 1.5 logs of the six available. This means there is space in the measurable range to expand our reporter gene's expression. With a greater dynamic range, more fine-grained analysis can be done to evaluate dimerization propensity and mutation severity. One possible explanation for the low reporter expression is because the ToxR construct is derived from *Vibrio cholera*, meaning that the promoter region is not optimized for *E. coli* transcription (Higgins and DiRita, 1996). There are several ways to combat this process. One way is to make the fluorescence signal brighter by putting multiple copies of the sfGFP on the plasmid to make a double sfGFP reporter gene or by inserting a second copy of the *ctx* promoter and reporter gene. Another way to increase brightness is to use a, yet to be created, brighter GFP or choose a FP that does not partially overlap with *E. coli* auto-fluorescence (Mihalcescu et al., 2015). A second method is to optimize the plasmid DNA for *E. coli* by changing the ribosome binding site from a *V. cholera* one to an *E. coli* one. Alternatively, the original TOXCAT version was inducible via the *lac* promoter (Russ and Engelman, 1999). Using this original version would increase the total expression of the ToxR construct.

There are two other small improvements that could increase the usefulness of the TOXGREEN assay. First, insertion could be improved by optimizing the linker regions between the TM domain and ToxR and MBP. The current set up extends the desired TM domain by "LILI,"

affecting C-terminal interactions with lipid headgroups. The BamHI/DpnII cut site combination that is required to keep the TM sequence in frame is troublesome due to binding a four base pair cut site to a six base pair cut site.

Second, a better negative control is needed. Currently, the negative control is the pccGFPKan plasmid which produces ToxR, but no TM domain or MBP. Instead of a 100 base pair TM domain between ToxR and MBP, pccGFPKan has a 1.5 kilobase kanamycin cassette. The NGS primer sites are located on ToxR and MBP and thus, a 2x150 NGS read will not capture the 1.5 kilobase cassette. Therefore, the negative control fluorescence value from the NGS cannot be reconstructed. One solution is to put a stop codon in an existing TM sequence to prevent transcription, but retain DNA length that can be captured by NGS.

There are many ways to improve this assay for future users, but focus should be put on the opportunity to quantify the levels of expression and insertion of the ToxR construct. Precise quantification of expression, insertion, and reporter level will render more accurate calculations of dimerization propensity, perhaps enabling calculation of apparent free energies (Duong et al., 2007).

5.4 Heterodimer modeling

The current CATM algorithm can only measure the energetics and predict the structures of parallel, C2 symmetric homodimers, and one of my long-standing goals has been to expand this capability. Cα—H bonding may be important in other GAS_{right} association contexts including parallel and antiparallel heterodimers as well as higher order oligomers. For example, integrins can form GAS_{right} heterodimers and there is a coronavirus spike protein that has been modeled to form GAS_{right} trimers (Arbely et al., 2006; Yang et al., 2009). Therefore, the logical next step is to expand the CATM algorithm to evaluate heterodimeric sequences. The first step to create this algorithm is to determine the geometric and sequence features that are required to make a stable heterodimer, as they will likely differ from that of the homodimer.

Though this work was pioneered by students before me in the lab, Benjamin Mueller and Sabareesh Subramaniam, I have worked to understand what we call the hetero-universe, or the hydrogen bonding strength present at a variety of geometries. The challenge to the homodimeric work of Mueller et al. (2014), is that the geometry of the hetero-universe is more extensive, increasing the degrees of freedom from 4 to 6. The first two geometric variables are simple: the distance (d) and the crossing angle (θ) between the helices. As opposed to homodimers, heterodimers are made up of two distinct helices whose rotation and vertical shift can vary independently from one another. The other four variables describe the point of closest approach. To identify that crossing point on each helix, the unit cell of the helical lattice is defined by the axial rotation (ω_1, ω_2) and the vertical shift (Z_1, Z_2) of each helix (Fig. 5.5). The addition of two extra degrees of freedom makes complete exploration of the conformational space significantly more expensive. Beginning with a poly-Glycine helix, the entirety of the conformational space was explored, taking into account each of the six variables. For each geometry, the hydrogen bonding propensity was calculated with the SCWRL 4 hydrogen bonding function that was reparameterized to include alpha carbon donors (Krivov et al., 2009b;

Mueller et al., 2014). Preliminary results indicated that weak Cα—H bonds could be formed over a greater spread of the hetero-universe than the homo-universe (data not shown).

Following this initial work, I collaborated with a number of undergraduate students to further our understanding of GAS_{right} heterodimers. I will describe the research done in connection with Collin McFadden and Junda Chen in improving the computational search algorithm that measures Cα—H bonding.

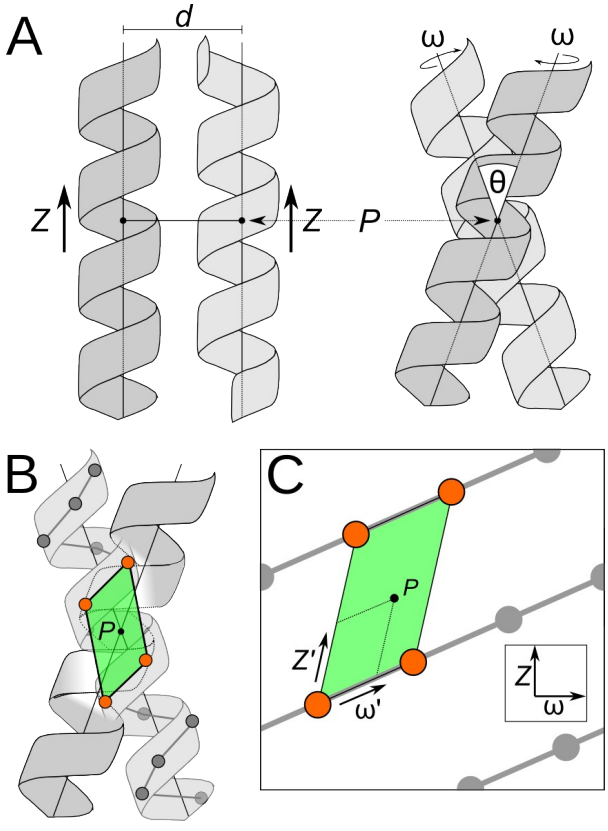


Figure 5.5 Heterodimer interactions are defined by six parameters. A) There are six parameters that define heterodimers: the inter-helical distance d; crossing angle θ; two rotation terms ω; and two vertical positions Z of the point of closest approach P. B) The helix was divided into "unit cells" bounded by 4 alpha-carbons. C) Within this unit cell we can describe P in terms of the rotation along the helical screw (ω') and the vertical shift (Z').

One of the greatest challenges of this project is that a six dimensional search space combined with the sequence space of two transmembrane helices is infinite. Thus, in order to appropriately explore the space, the algorithm must intelligently sample the helix-helix landscape. To perform testing, we must narrow down the conformations we would like to explore. Junda identified an iterative Monte Carlo refinement approach that will segment the hetero-universe into coarse-grained geometries allowing for an increase in parallel computing (Fig 5.6). Briefly, this process starts by coarsely segmenting the geometric space, calculating the energy of a random set of geometries, and averaging the energies for a segment. Based on that average, the segments are prioritized and reprocessed by priority. The chosen segment is more finely

segmented, and the process repeats. This segmentation will create a simple way to visualize and understand a six-dimensional space.

The leading undergraduate on this project was Collin, who identified a way to reduce the energetic calculations by using a database. After reducing the search space, the sequence complexity must be minimized to make a manageable universe. While looking exclusively at backbone Cα—H bonding potential, the only strong amino acid possibilities are glycine and alanine due to their small size and glycine's additional available hydrogen (Mueller et al., 2014). When exploring sequences composed of two amino acids (A = 2), A^2 runs are required to observe all combinations of residue-to-residue pairs. Following these reductions, his algorithm can then be condensed down into several steps. First, poly-Ala and poly-Gly transmembrane domains are run through all geometric conformations and the energies for each pair of residues

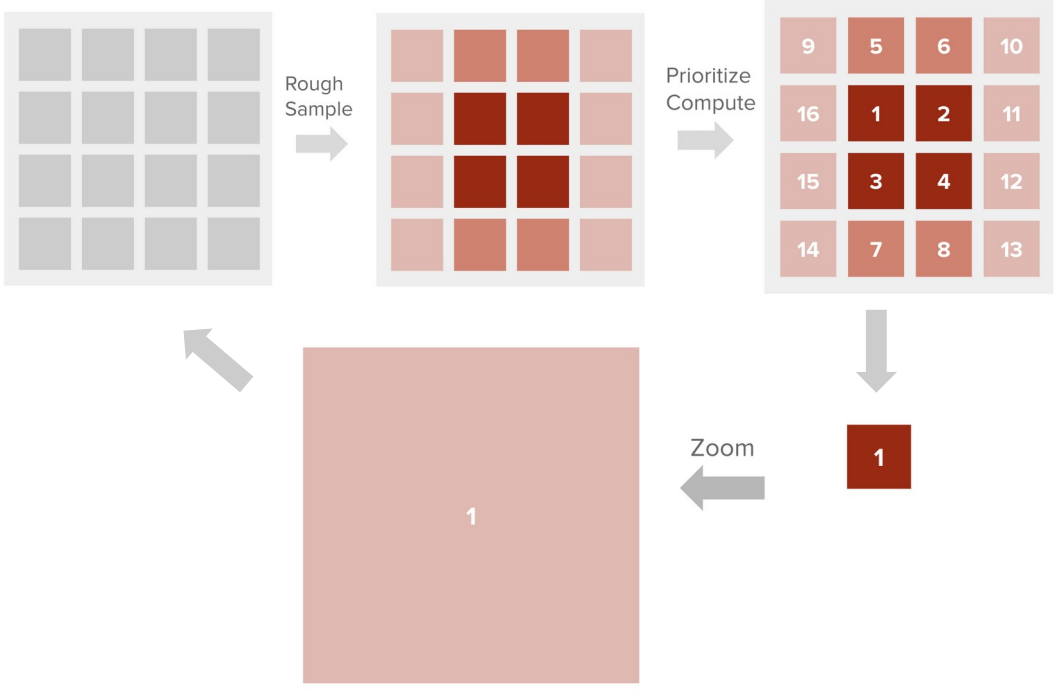


Figure 5.6 Iterative Monte Carlo refinement approach. In this algorithm, a search space is coarsely segmented and randomly sampled for energetic scores. These segments are then prioritized and sub-segmented and the process cycles until sufficient sampling identifies geometries that are important for Cα—H bonding. Figure courtesy of Junda Chen.

are recorded in a database (Fig 5.7). Following these extensive calculations, for any transmembrane sequence composed of alanine and glycine, the individual residue pairs can be selected from the database and added together to evaluate the best conformation for that sequence.

The next steps in this project are to implement these methods in combination to create a fast algorithm that can calculate the energetic variables of many transmembrane sequences. Joshua Choi will be continuing this project as part of his thesis work.

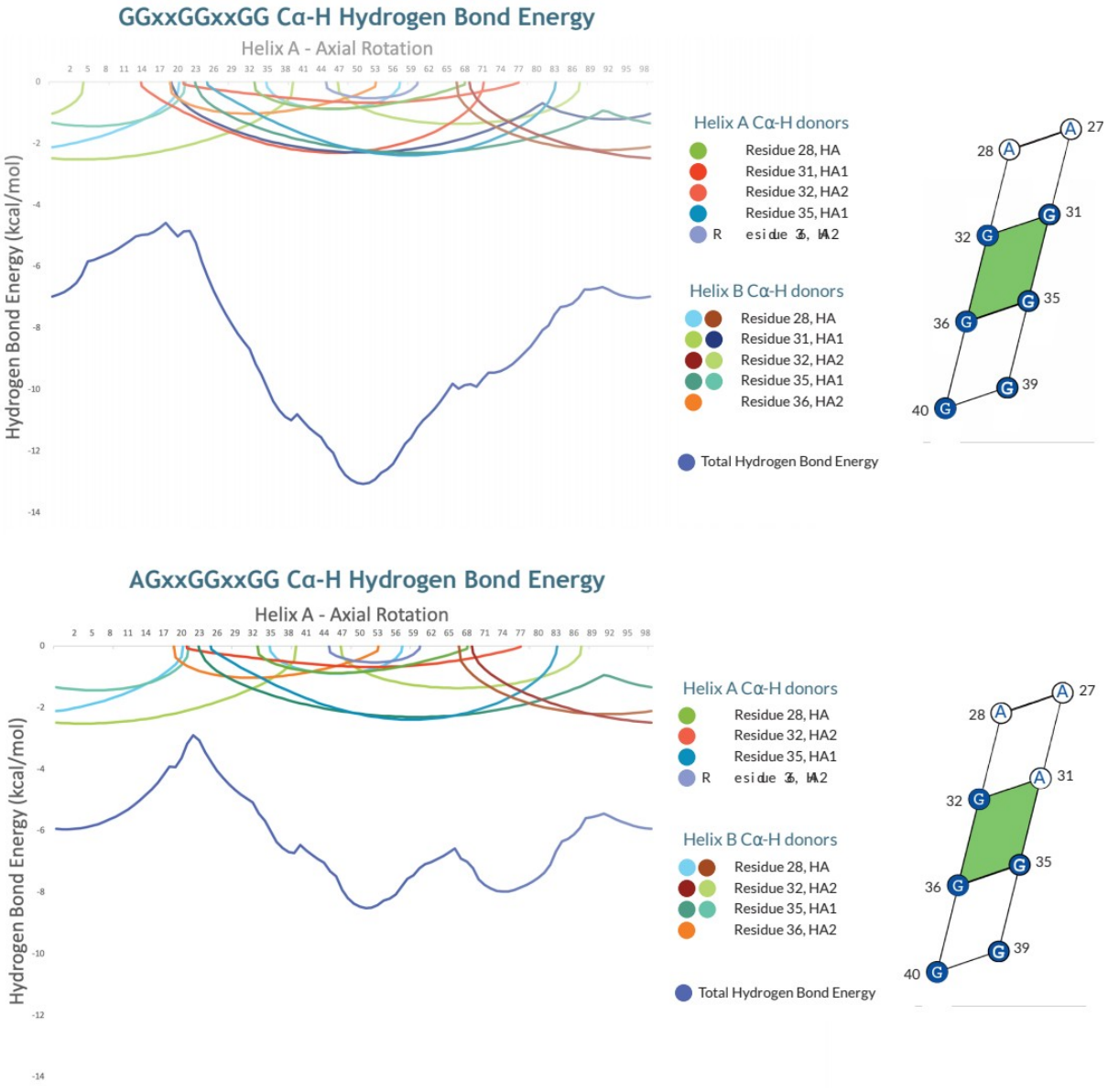


Figure 5.7 Total hydrogen bonding energy at different conformations. With five of six geometric variables fixed, we observe that the total Cα—H bond energy changes drastically and unpredictably when viewing only the sum of the hydrogen bond interactions (deep blue line). The other five variables are interhelical distance = 6.4 Å, vertical shift A = 4.5 Å, vertical shift B = 4.5 Å, axial rotation B = 50°, and crossing angle = -33°. The only sequence difference between top and bottom are two mutated residues (G31A in both helices). Each arc represents the energy of a different hydrogen bond. The arcs that exist in both graphs are identical, but some arcs disappear when G31 is mutated, creating vastly different total hydrogen bond energy curves. Data courtesy of Collin McFadden.

Table 5.1 Model organisms' single-pass membrane proteomes and CATM results.

	H. sapiens	*E. coli*	*A. thanliana*	*C. elegans*	*X. laevis*	*D. rerio*	*D. melanogaster*	*S. cervisiae*	Total
Total SPMPs	2383	163	1379	307	218	260	222	530	5462
Positive scoring models	1141	70	557	103	100	117	98	182	2368
Positive scoring poly-Leu models	1021	62	511	88	87	103	89	163	2124
Nonpolar poly-Leu models	576	33	305	49	44	54	56	75	1192

Table 5.2 CATM Energies for GpA interface stitched into different backgrounds.

Sequence																							XShift	ZShift	AxialRot	CrossAng	Energy
R	A	S	V	I	L	L	I	L	V	G	V	V	I	G	V	L	L	T	I	L	I	L	I	6.31	5.69	-89.05	-50.88	-34.05
R	A	S	L	I	L	L	I	V	I	G	V	V	V	G	V	L	L	T	I	L	I	L	I	6.30	5.67	-89.10	-50.79	-33.85
R	A	S	L	I	V	L	I	L	L	G	V	I	L	G	V	V	V	T	I	L	I	L	I	6.30	5.66	-89.11	-50.79	-33.55
R	A	S	L	V	I	L	I	L	L	G	V	V	I	G	V	L	V	T	I	L	I	L	I	6.31	5.66	-89.21	-50.79	-33.54
R	A	S	L	V	L	L	I	L	V	G	V	V	I	G	V	L	I	T	I	L	I	L	I	6.39	5.66	-89.27	-50.86	-32.92
R	A	S	L	V	V	L	I	L	I	G	V	L	V	G	V	I	L	T	I	L	I	L	I	6.30	5.68	-89.20	-50.86	-32.88
R	A	S	V	V	L	L	I	L	I	G	V	V	I	G	V	L	L	T	I	L	I	L	I	6.30	5.68	-89.05	-50.67	-32.66
R	A	S	I	L	L	L	I	V	L	G	V	V	I	G	V	V	L	T	I	L	I	L	I	6.39	5.68	-89.07	-50.98	-32.64
R	A	S	I	L	I	L	I	L	L	G	V	V	L	G	V	V	V	T	I	L	I	L	I	6.31	5.67	-89.04	-50.90	-32.60
R	A	S	L	L	I	L	I	L	L	G	V	V	I	G	V	V	V	T	I	L	I	L	I	6.32	5.68	-89.07	-50.86	-32.56
R	A	S	L	L	I	L	I	V	V	G	V	L	L	G	V	I	V	T	I	L	I	L	I	6.29	5.66	-89.20	-51.09	-32.40
R	A	S	V	L	V	L	I	L	V	G	V	L	I	G	V	L	I	T	I	L	I	L	I	6.31	5.66	-89.06	-50.67	-31.93
R	A	S	L	L	V	L	I	L	I	G	V	V	V	G	V	I	L	T	I	L	I	L	I	6.32	5.67	-89.22	-51.00	-31.90
R	A	S	I	V	V	L	I	I	L	G	V	L	V	G	V	L	L	T	I	L	I	L	I	6.40	5.67	-89.04	-50.99	-31.78
R	A	S	I	L	V	L	I	L	L	G	V	V	V	G	V	L	I	T	I	L	I	L	I	6.40	5.68	-89.04	-50.93	-31.55
R	A	S	I	V	L	L	I	I	L	G	V	V	V	G	V	L	L	T	I	L	I	L	I	6.40	5.67	-89.04	-50.92	-31.25
R	A	S	L	V	L	L	I	I	L	G	V	L	I	G	V	V	V	T	I	L	I	L	I	6.31	5.68	-89.03	-50.94	-31.25
R	A	S	V	L	I	L	I	L	L	G	V	I	L	G	V	V	V	T	I	L	I	L	I	6.30	5.68	-88.91	-50.82	-30.82
R	A	S	V	L	L	L	I	V	L	G	V	I	V	G	V	I	L	T	I	L	I	L	I	6.31	5.68	-89.11	-50.97	-29.97
R	A	S	V	L	L	L	I	I	L	G	V	I	L	G	V	V	V	T	I	L	I	L	I	6.31	5.66	-89.11	-50.88	-29.55

5.5 References

Anderson, S.M., Mueller, B.K., Lange, E.J., and Senes, A. (2017). Combination of Cα-H Hydrogen Bonds and van der Waals Packing Modulates the Stability of GxxxG-Mediated Dimers in Membranes. J. Am. Chem. Soc. *139*, 15774–15783.

Arbely, E., Granot, Z., Kass, I., Orly, J., and Arkin, I.T. (2006). A trimerizing GxxxG motif is uniquely inserted in the severe acute respiratory syndrome (SARS) coronavirus spike protein transmembrane domain. Biochemistry *45*, 11349–11356.

Das, R., and Baker, D. (2008). Macromolecular modeling with rosetta. Annu. Rev. Biochem. *77*, 363–382.

Dooley, C.T., Dore, T.M., Hanson, G.T., Jackson, W.C., Remington, S.J., and Tsien, R.Y. (2004). Imaging dynamic redox changes in mammalian cells with green fluorescent protein indicators. J. Biol. Chem. *279*, 22284–22293.

Duong, M.T., Jaszewski, T.M., Fleming, K.G., and MacKenzie, K.R. (2007). Changes in apparent free energy of helix-helix dimerization in a biological membrane due to point mutations. J. Mol. Biol. *371*, 422–434.

Hessa, T., Meindl-Beinker, N.M., Bernsel, A., Kim, H., Sato, Y., Lerch-Bader, M., Nilsson, I., White, S.H., and von Heijne, G. (2007). Molecular code for transmembrane-helix recognition by the Sec61 translocon. Nature *450*, 1026–1030.

Higgins, D.E., and DiRita, V.J. (1996). Genetic analysis of the interaction between Vibrio cholerae transcription activator ToxR and toxT promoter DNA. J. Bacteriol. *178*, 1080–1087.

Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L. (2009a). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778–795.

Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L. (2009b). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778–795.

Lazaridis, T. (2003). Effective energy function for proteins in lipid membranes. Proteins *52*, 176–192.

Lomize, A.L., Pogozheva, I.D., Lomize, M.A., and Mosberg, H.I. (2006). Positioning of proteins in membranes: a computational approach. Protein Sci. Publ. Protein Soc. *15*, 1318–1333.

MacKerell, Bashford, D., Bellott, Dunbrack, Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†. J. Phys. Chem. B *102*, 3586–3616.

Meiresonne, N.Y., van der Ploeg, R., Hink, M.A., and den Blaauwen, T. (2017). Activity-Related Conformational Changes in d,d-Carboxypeptidases Revealed by In Vivo Periplasmic Förster Resonance Energy Transfer Assay in Escherichia coli. MBio *8*.

Meiresonne, N.Y., Consoli, E., Mertens, L.M.Y., Chertkova, A.O., Goedhart, J., and den Blaauwen, T. (2019). Superfolder mTurquoise2ox optimized for the bacterial periplasm allows high efficiency in vivo FRET of cell division antibiotic targets. Mol. Microbiol. *111*, 1025–1038.

Mihalcescu, I., Van-Melle Gateau, M., Chelli, B., Pinel, C., and Ravanat, J.-L. (2015). Green autofluorescence, a double edged monitoring tool for bacterial growth and activity in micro-plates. Phys. Biol. *12*, 066016.

Mueller, B.K., Subramaniam, S., and Senes, A. (2014). A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα-H hydrogen bonds. Proc. Natl. Acad. Sci. U. S. A. *111*, E888-895.

Russ, W.P., and Engelman, D.M. (1999). TOXCAT: a measure of transmembrane helix association in a biological membrane. Proc. Natl. Acad. Sci. U. S. A. *96*, 863–868.

Schramm, C.A., Hannigan, B.T., Donald, J.E., Keasar, C., Saven, J.G., Degrado, W.F., and Samish, I. (2012). Knowledge-based potential for positioning membrane-associated structures and assessing residue-specific energetic contributions. Struct. Lond. Engl. 1993 *20*, 924–935.

Senes, A., Chadi, D.C., Law, P.B., Walters, R.F.S., Nanda, V., and Degrado, W.F. (2007). E(z), a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: derivation and applications to determining the orientation of transmembrane and interfacial helices. J. Mol. Biol. *366*, 436–448.

Ulmschneider, M.B., Sansom, M.S.P., and Di Nola, A. (2005). Properties of integral membrane protein structures: derivation of an implicit membrane potential. Proteins *59*, 252–265.

Yang, J., Ma, Y.-Q., Page, R.C., Misra, S., Plow, E.F., and Qin, J. (2009). Structure of an integrin alphaIIb beta3 transmembrane-cytoplasmic heterocomplex provides insight into integrin activation. Proc. Natl. Acad. Sci. U. S. A. *106*, 17729–17734.