The GAS_{right} Motif in Membrane Protein Dimerization and Structural Prediction

By

Benjamin Keymar Mueller

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Biochemistry)

at the

UNIVERSITY OF WISCONSIN – MADISON

2015

Date of final oral examination: 6/11/15

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

Alessandro Senes, Associate Professor, Biochemistry
Ivan Rayment, Professor, Biochemistry
Qiang Cui, Professor, Chemistry
Ann Palmenberg, Professor, Biochemistry and the Institute for Molecular Virology
Julie Mitchell, Professor, Biochemistry and Mathematics

Table of Contents

List of Figures and Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Acknowledgements

Throughout my time as a graduate student many people have contributed to get me to where I am today. While some played larger roles than others, all of them were integral in my development as a researcher and a person. First I would like to thank my advisor and mentor, Alessandro Senes. He has been a great teacher and source of inspiration and guidance. Going forward, I can only hope I continue to have such an invaluable mentor.

I would also like to thank my committee members: George Phillips, John Markley, Julie Mitchell, Ann Palmenberg, Ivan Rayment and Qiang Cui, for their insightful and critical comments over the years. Also I would like to thank the Computation and Informatics in Biology and Medicine training grant for supporting my research.

The help I've received over the years from the lab has been truly exceptional. For all his help in teaching me computational modeling, I would like to thank Sabareesh Subramaniam. He has been a great collaborator over the years. Also thank you to my other lab collaborator Ambalika Khadria, who has also been a great resource. I would like to thank Loren LaPointe for being my go to person for anything lab related, she's helped me out of jams countless times. Thank you to Samantha Anderson, for being my newest collaborator, who will continue my work once I'm gone. Thanks to Samson Condon for always being there to bounce ideas off of. And thank you to the other graduate members of the Senes Lab, Samuel Craven, Gladys Díaz-Vázquez, and Deena-Al Mahbuba for helping as well.

I would also like the thank all the undergraduate researchers who have assisted time and time again in my projects, special thanks must be given to Claire Holesovsky, Yudong Sun, Zixiao Chen, and last, but most importantly Evan Lange.

My friends have also played an important role in my graduate career, both as an

intellectual resource, and for being just some wonderful friends. Special thanks go to my Biochemistry class, we have stayed close through our time here and I've made some really great friends. Also thanks to the group of people I played Ultimate Frisbee with, I've met some great people through that as well. And while all of these people would take too long to list, I need to highlight two people who have really helped me through my time here: Jess Feldman and Matt Mead. These two have been amazing friends, and I hope we stay close far into the future.

I would like to thank my family, especially my Mom, Dad and Sister, for playing an important role in getting me to where I am today. I couldn't have made it without them.

Finally I would like to thank my wife, Klare, for giving me unconditional love and support throughout my graduate studies. I love you very much.

Abstract

The most common transmembrane proteins are the single-pass membrane proteins (SPMPs), which span the membrane by threading a single α-helix through the hydrophobic lipid bilayer. SPMPs are biologically important to the function of the cell. The function of many SPMPs involves the homo-dimerization of their TM domain. One sequence motif that has been shown to be important in SPMP homo-dimerization is the GxxxG motif, or two glycine residues spaced at *i* and *i+4*. This small residue motif is often found at the interface of the GAS_{right} TM dimer structural motif, the association of two parallel transmembrane helices at a right-handed crossing angle of around -40 degrees.

The small residues at the interface allow for the close approach between the two helices, and this tight packing allows for the formation of Cα–H···O=C bonds. Cα–H···O=C bonds form between the α-carbon hydrogen on one helix, and the carbonyl oxygen on the opposing helix. While the energetic contribution of these bonds is still a matter of debate, the bonds are commonly seen in GAS_{right} structures and are predicted to have a favorable contribution to protein folding.

Using computational modeling, I have discovered strong evidence suggesting that interhelical carbon hydrogen bonds which occur between Cα–H donors and backbone carbonyl oxygen acceptors (Cα–H···O=C bonds) are an important force driving in the association of transmembrane proteins.

Using this knowledge, I was able to design an ab initio structural prediction algorithm (CATM) to correctly predict the known structures of transmembrane dimers. From this work, I predicted the structure of a protein whose transmembrane structure has not been solved – an important mitochondrial kinase involved in coenzyme Q biosynthesis.

Currently, I am working to further develop the CATM algorithm to predict the strength of transmembrane dimerization in bacterial membranes. To improve the prediction

capability, I have been experimentally measuring GAS_{right} oligomerization on a large number of predicted sequences. The current model shows a strong correlation between the experimental strength of association and optimization of van der Waals packing and Cα-H hydrogen bonding.

Chapter 1

The GxxxG sequence motif and GAS_{right} structural motif in single-pass membrane proteins

1.1 Introduction to Single-pass Membrane Proteins

Transmembrane (TM) proteins can be divided into three main classes: single-pass, multi-pass and β-barreled membrane proteins. Single-pass membrane proteins span the membrane by threading a single α-helix through the hydrophobic lipid bilayer (Fig. 1.1a). Multi-pass membrane proteins span the membrane two or more times via α-helices (Fig. 1.1b). β-barreled membrane proteins create a pore through the bilayer using a cylindrical β-sheet (Fig. 1.1c). Estimates predict that in most organisms, 20-30% of encoded proteins are membrane proteins. In humans approximately 6,000 different membrane proteins are translated [Wallin, et al. (1998); Krogh, et al. (2001); Käll, et al. (2004)]. Specifically, my work is focused on single-pass membrane proteins (SPMPs). SPMPs are both topologically simple and the most abundant of the transmembrane proteins. Of the approximately 6,000 human membrane proteins there are roughly 2,200 single-pass membrane proteins, as annotated in the Uniprot database [Uniprot Consortium (2014)].

Although membrane proteins comprise a large percentage of the genome, they are understudied. This is due to the difficulties in performing experiments when the protein is bound inside the membrane. As a result, structure determination has also lagged behind soluble proteins. Currently, there are just over 100,000 protein structures in the Protein Data Bank (PDB). Of these only 1,654 are membrane proteins, and only 539 are unique proteins. If around 20-30% of genomes encode for membrane proteins, there should be approximately twenty to thirty thousand membrane protein structures in the

PDB. This limited structural insight hampers our understanding of how membrane proteins fold, interact and perform their biological functions. Fewer structures means more broad and time consuming experiments. Three dimensional structures provide important insights into the critical residues of the protein; therefore, it is important to develop new ways of understanding how membrane proteins fold and determining their three-dimensional structure.

My thesis is focused on answering the question: can we use computational methods to both better understand the forces that guide membrane protein dimerization, and use this knowledge to predict membrane protein structure? In this introductory chapter, I will begin by explaining the biological importance of single-pass membrane proteins. Then I will introduce a commonly found membrane sequence motif, the GxxxG motif, and explain its role in the GAS_{right} helix dimer structural motif. Finally I will explain how the combination of this sequence and structure motif allow for the formation of Cα-H···O=C hydrogen bonds, and postulate the importance of their role in dimerization.

1.2 Biological Importance of Single-pass Membrane Proteins

The transmembrane domains of SPMPs have historically been thought of as “anchors” tethering the soluble portion of the protein in place near one of the membranes in the cell. However, numerous studies have shown that SPMPs, and more importantly their TM domains, play an active role in both cellular processes as well as disease.

A subset of SPMPs that are critically important to human health are integrins, which are located in the cellular membrane of all metazoa, and are responsible for cell-to-cell adhesion. Integrins are a heterodimer of two SPMPs, composed of one α and one β subunit. A ligand-mediated equilibrium between the heterodimeric and monomeric state regulates the cellular adhesion signal [Arnaout, et al. (2005)]. Mutations to integrins have been found to cause prolonged blood coagulation (Glanzmann's thrombasthenia), recurring infections (leukocyte adhesion deficiency type I) and skin blistering disorder (pyloric atresia) [Winograd-Katz, et al. (2014)]. Specific studies have found that a laboratory designed integrin binding TM peptide can disrupt integrin heterodimerization forcing the complex into its "on" state. This action induces platelet adhesion, indicating that the TM region is central to integrin regulation [Yin, et al. (2007)].

Another important and well-studied group of SPMPs are the Receptor Tyrosine Kinase (RTK) proteins – a diverse class of signaling receptors [He, et al. (2012)]. RTKs are targeted by extracellular growth factors, and have been implicated in many human diseases [Ullrich, et al. (1990); Blume-Jensen, et al. (2001)]. The human RTK family includes 58 different proteins divided into 20 classes, which perform and assist a diverse range of functions including: cellular proliferation, differentiation, survival and migration, as well as metabolism. The RTKs act as a conduit, relaying the external signals to the interior of the cell. The architecture is nearly identical across the 20 families and consists of: an extracellular ligand binding domain, a TM region and an intracellular tyrosine kinase domain [Lemmon, et al. (2010)]. The general method of

signaling involves a ligand binding to the RTK extracellular domain. This, in turn, drives a conformational shift and protein dimerization which causes intracellular tyrosine autophosphorylation. This phosphorylation is responsible for interaction with downstream signaling molecules [Hubbard, et al. (2007)]. While mutations to the ligand binding and tyrosine kinase domain result in altered function of the protein, it has also been observed that mutations to the TM domain can cause symptoms such as Thanatophoric dysplasia, Achondroplasia, and Crouzon syndrome. Most of these disorders stemming from increased dimerization of the TM domain [He, et al. (2012)].

Due to their prevalence and perceived importance, the dimeric state of RTKs have been a continued target for structure determination via NMR. Currently, there are at least eight homo- and hetero-dimeric structures of receptor tyrosine kinases, listed as follows by their familial classification. The epidermal growth factor receptor (ErbB) family is composed of 4 members, all four of which have solved homo-dimeric structures: Erbb1 (also commonly known as Epidermal Growth Factor Receptor (EGFR)) [Endres, et al. (2013)], Erbb2 [Bocharov, et al. (2008)], Erbb3 [Mineev, et al. (2011)], and Erbb4 [Bocharov, et al. (2012)]. The heterodimeric structure of Erbb1 (EGFR) and Erbb2 has also been solved [Mineev, et al. (2010)]. The Erythropoietin-Producing Hepatocellular Receptors (Eph) family has two solved homo-dimeric structures: EphA1 [Bocharov, et al. (2008)] and EphA2 [Bocharov, et al. (2010)]. The EphA1 dimer is of great interest, as it has been solved at two different pH values, leading to two different configurations of the dimer interface. As RTKs are thought to undergo a conformational change

mediated through their transmembrane domain, this finding may have physiological implications [Bocharov, et al. (2008)]. The Fibroblast Growth Factor Receptor (FGFR) family contains one dimeric structure member FGFR3 [Bocharov, et al. (2013)]. Finally, the last family represented is the Vascular Endothelial Growth Factor Receptors, which contains one member, the homo-dimer, VEGFR2 [Manni, et al. (2014)]. As demonstrated above, the NMR structures of the TM regions of RTKs have shown diversity in both structural configuration and residue composition – yielding an important set of proteins for continuing studies.

The investigation into both integrins and RTKs have shown how SPMPs help to regulate eukaryotic cellular processes. Additionally, SPMPs play a role in viral infection. The HIV envelope glycoprotein (Env), which shares similar structural homology to other enveloped virus proteins, is central in the process of infection. The Env complex is composed of two subunits: gp41 and gp120. The gp41 subunit is an SPMP that fuses the viral and cellular membranes [Weiss (2003)]. When the gp41 TM domain is replaced with a topologically similar, yet distinct in sequence TM domain, it results in decreased fusion activity, indicating the importance of the TM region to this interaction [Kondo, et al. (2010)]. Therefore it is, once again, the specific TM sequence that is important for SPMP biological activity, not just the presence of a hydrophobic anchor.

The regulation and control of SPMPs are important as well, especially in the case of Alzheimer’s disease. One protein of particular interest to researchers studying this

disease is the β–amyloid precursor protein (APP). APP has multiple cleavage products, one of which has been implicated in synaptic damage and neuron loss, Aβ42 [Zhang, et al. (2012)]. APP is a SPMP, and the dimerization strength of its TM domain has been suggested as a possible regulator of Aβ42 generation [Munter, et al. (2007)].

1.3 Using Single-pass Transmembrane Proteins to Understand Protein Folding

SPMP are not only biologically important, they also provide an excellent system for determining the amino acids and forces that mediate membrane protein folding. Much of the study of how transmembrane helices interact in the membrane have been based on the Popot and Engelman “Two-stage model”. In stage one, the stable α-helix is inserted into the membrane and in stage two, the helices associate into the fully-folded protein. In the hydrophobic environment of the membrane the cost of breaking a backbone carbonyl to amine bond (*i to i*+4) is too great; therefore, the alpha helix will be maintained [Popot, et al. (1990)]. When studying membrane protein folding, the focus is on the lateral interaction in the bilayer, and for SPMPs there are a wide range of assays available for this type of study.

The biophysical characterization of the folding and interaction of membrane proteins is generally challenging. One of the main benefits of studying SPMPs is that a multitude of assays are available for the study of how their TM helices interact, which can be divided into two main classes: *in vivo* and *in vitro* assays.

The TM domain dimerization *in vivo* assays began with the development of the ToxR system by Langosch, et. al. in 1996 [Langosch, et al. (1996)]. This assay relates the dimerization of TM proteins in the bacterial membrane to the expression of a reporter gene. The relative strength of TM domain dimerization is then measured by the activity of expressed β-galactosidase. The ToxR system was further developed into the more widely used TOXCAT system by Russ and Engelman in 1999 [Russ, et al. (1999)]. The two assays are methodologically similar, the main difference being the reporter gene is changed to Chloramphenicol Acetyl-Transferase (CAT). This change allows for two methods to determine relative dimerization. One method is by a selection assay. Using this method, cells with the dimeric TM domains express CAT, and are able to grow in the presence of chloramphenicol. The other method works by lysing the cells and determining the level of expression of CAT by a colormetric assay. Unfortuantely, both these assays are limited as they only measure homo-dimerization. However a similar assay, GALLEX, has been developed to study hetero-dimerization. Two helices must be brought together for β-galactosidase repression to occur; therefore the amount of repression is an indirect measure of dimerization activity [Schneider, et al. (2003)]. While they are only qualitative, there are two important benefits of these assays: dimerization is measured in natural membranes, and mutagenesis screens can be easily accomplished to determine the amino acids at the interface of interaction.

In vitro assays cover a diverse range of techniques and hydrophobic environments and include, but are not limited to, sodium dodecyl sulfate polyacrylamide gel

electrophoresis (SDS PAGE), Förster resonance energy transfer (FRET) and Sedimentation Equilibrium Analytical Ultracentrifugation (SE AUC). The use of SDS PAGE to probe non-covalent membrane protein interaction in a hydrophobic environment has been used since the 1970's [Furthmayr, et al. (1976)]; however it took almost 20 years for a systematic probing of the TM domain to occur using this technique. While SDS denatures soluble proteins, TM domains will maintain their helicity in the detergent, and will oligomerize as they would in lipid bilayers. TM domains of interest are usually paired with a larger monomeric protein (for better resolution), and can be run out on a gel with distinct monomer/dimer bands observed [Lemmon, et al. JBC (1992); Lemmon, et al. Biochemistry (1992)]. However the results are not quantitative, and the detergent SDS does not truly mimic a biological environment.

Förster resonance energy transfer, or FRET, is another method used to probe membrane protein association *in vitro*. Using this method, peptides are synthesized and labeled with FRET donor/acceptor pairs. FRET is measured by a ratio of donor to acceptor signal, and the dissociation constant determined by a serial peptide dilution while maintaining the other component concentration constant [Fisher, et al. (1999)]. FRET allows for more quantitative calculations than any of the biological assays or SDS-PAGE; however, the stability of the dimer is heavily influenced by the type of detergent used. FRET can also be performed in liposomes, but care must be taken to determine if FRET activity is due to actual dimerization or the incorporation of two FRET pairs into the same liposome. The benefit of this system is that FRET is measured in a

more native environment [You, et al. (2005)].

Sedimentation Equilibrium Analytical Ultracentrifugation (SE AUC) has also been developed as an *in vitro* quantitative assay for determining membrane protein oligomerization [Fleming, et al. (1997)]. The main benefit of SE AUC is that the mass of the protein complex is measured, either in a monomer, dimer, or higher order oligomer. This is a benefit over FRET, as an exact oligomeric state can be determined. Also, SE AUC is explicitly able to show that a protein is monomeric. This is in contrast to experiments such as FRET or TOXCAT, where a lack of signal is interpreted as a negative result, which could be due to a multitude of factors, and not just a absence of interaction [Burgess, et al. (2008)].

While there are many methods for determining the oligomeric state, dimerization energy, and the residues involved in the helix-helix interaction, it is important to understand that while overall the methods will agree with one another, many details can vary from assay to assay, due to the method used to the conditions of the assay. Therefore, it is always important to understand the system being used and how it may impact results.

1.4 A Frequently Occurring TM Sequence Motif: GxxxG

In order to understand how SPMP, and by extension all membrane proteins both oligomerize and fold, it is critical to understand the amino acids that guide this

interaction. While every protein contains a unique set of amino acids which creates a unique fold, some sequence patterns exist across a wide range of protein families to facilitate a common interaction and/or function. The most common and most studied amino acid pattern in membrane proteins is the GxxxG motif, or two glycine residues spaced four positions apart (Fig 1.2). Due to the geometry of the α-helix (~3.6 residues per turn), the four residue spacing places the glycines on the same face of the helix. The importance of the two glycine residue motif was first fully understood in 2000 by Senes, et. al. They analyzed the distribution of pairs and triplets of amino acids in TM domain sequences to find frequently occurring patterns [Senes, et al. (2000)]. While other motifs are more common, such as LxxxL, the GxxxG motif is observed the most over an expected occurrence, 31.6% above the expected value. In a concurrent paper from the same laboratory the GxxxG motif was studied *in vivo* using the TOXCAT assay [Russ, et al. (2000)]. A TM domain of a SPMP composed of leucine or alanine residues was randomly mutated at positions (spaced at 1, 2, 5, 6, 9, 10 and 13) to 9 possible amino acids (glycine, alanine, valine, leucine, isoleucine, serine, threonine, proline and arginine). The TMs were then screened for their ability to allow for bacterial colony growth in the presence of chloramphenicol. The TOXCAT assay confers increasing chloramphenicol resistance to bacteria based on increasing TM region dimerization. The results closely matched the natural amino acid distribution noted by Senes, Gerstein and Engelman. GxxxG was the most common motif in the chloramphenicol resistant bacteria. While found at a much lower frequency, other small amino acid motifs spaced four residues apart were observed, such as SxxxG and SxxxS. These small residue

motifs, including GxxxG, are often found with adjacent large aliphatic residues such as isoleucine, leucine and valine, forming a motif such as GVxxGV. These large amino acids can create a “ridge” that can pack into the “groove” of the adjacent helix, allowing for tight packing at the interface [MacKenzie, et al. (1997)]. The GxxxG motif and related [G,A,S]xxx[G,A,S] sequences occur in both multi-pass and single-pass membrane proteins. In single-pass proteins their occurrence is frequent, with over 60% of non-redundant TMs containing at least one [G,A,S]xxx[G,A,S] motif. Even when the pattern is constrained to contain at least one glycine, the motif exists in 42% of all SPMP TM domains [Senes, et al. JMB (2000)].

1.5 GxxxG Sequence Motif Is Commonly Found In the GAS_{right} Structure Motif

The GxxxG motif and related [G,A,S]xxx[G,A,S] motifs have been found to play an intergral role in a common membrane protein structural motif [Walters, et al. (2006); Zhang, et al. (2015)]. In 2006, Walters and DeGrado analyzed the Protein Data Bank (PDB) and found 31 representative membrane protein structures, which contained 445 helical pairs. They clustered these helix pairs by root mean squared deviation (RMSD), and found that 74% of all helical dimer space could be represented by only 5 clusters. It was found that 12.8% of the total were parallel, right-handed crossing dimers – which is overall the third most common fold, and the most common parallel structure. The right-handed crossing angle is in a Gaussian distribution centered around -40 degrees (Fig. 1.3). What was most interesting about this motif was the amino acid composition at the helical interface – in many cases it was the highly observed sequence motif GxxxG. In

addition to the GxxxG motif, other small amino acids, such as alanine and serine, were also observed, further confirming the Senes and Russ findings [Senes, et al. (2000); Russ, et al. (2000)]. The parallel, right-handed crossing motif was designated as GAS_{right}. The “GAS” comes from the interfacial residues Gly, Ala, Ser and the “right” from the right-handed crossing angle. The small amino acid motif seen in GAS_{right} structures allows for the close approach of two helices and tight amino acid packing. Prior to Walters and DeGrado formalizing the GAS_{right} motif in 2006, only one GAS_{right} mediated SPMP dimer structure had been published: the Glycophorin A dimer. Since then it has been found that many SPMP dimers fold via the GAS_{right} motif.

1.6 GAS_{right} Structural Motif Is Found In Many Commonly Studied Single-Pass Membrane Proteins

The most famous and well-studied SPMP dimer is Glycophorin A (GpA), a sialoglycoprotein found in human erythrocyte membranes. GpA was first discovered in 1975 [Furthmayr, et al. (1975)] and the following year it was determined that GpA forms a stable dimer in the presence of the detergent SDS, presumably through an interaction of the hydrophobic region of the peptide. A competition experiment showed that a hydrophobic peptide derived from full GpA was able to bind to the monomeric form of the protein, but not the dimer. This lead the researchers to correctly surmise that the interaction of the GpA monomers occurred due to their hydrophobic domains [Furthmayr, et al. (1976)].

It took researchers an additional fifteen years to determine that the helices of GpA alone could cause dimerization. This was done by fusing Staphylococcal nuclease (a known soluble monomeric protein) with the TM region of GpA. The resulting chimera ran as a dimer when SDS gel electrophoresis was performed [Lemmon, et al. (1992)]. The researchers continued by determining the exact sequence motif responsible for dimerization. Extensive mutagenesis was performed on the 23 residue hydrophobic region of GpA, and the mutants were analyzed for dimerization via SDS-PAGE.

The residues most susceptible to the effects of mutation were the seven bolded residues: **LI**xx**GV**xx**GV**xx**T**. The researchers noted that while these seven positions were the most sensitive, all but one had tolerable mutants that would not completely disrupt dimerization. Position 83 (the second glycine), when mutated to anything other than the wild-type glycine, will result in a monomeric form of GpA. A helical wheel projection was used to map the sequence of GpA and, as was expected, all the positions sensitive to mutation fell on the same side of the helix, thereby giving an initial glimpse into the interaction interface of GpA [Lemmon, et al. (1992)].

These findings were confirmed when the structure of GpA was published in 1997 [MacKenzie, et al. (1997)]. A 40 residue stretch of GpA containing the TM domain was solubilized in detergent micelles and the structure determined by NMR. The researches saw that the mutation sensitive residues did form the interface between the two GpA peptides. The two glycines (G79 and G83) formed a “groove” that the “ridge” of two

valines (V80 and V84) tightly packed into, forming excellent van der Waals contacts. The symmetrical helix dimer formed a -40° (right-handed) crossing angle – a canonical GAS_{right} structure.

Glycophorin A is not the only solved NMR structure of a membrane protein dimer to fold via a GAS_{right} motif and have a GxxxG motif at the interface. Bcl-2 Nineteen-kDa interacting protein 3 (BNIP3) is an apoptotic Bcl-2 protein, and another well studied SPMP. It contains a C-terminal GxxxG motif with both positions highly sensitive to mutation [Sulistijo, et al. (2006)]. In the center of the TM region there are two polar residues: a histidine and serine, both of which are also sensitive to mutation. When the structure of the protein was determined by NMR, it was found that the BNIP3 dimer forms a GAS_{right} structure with the GxxxG motif at the center of the interaction [Bocharov, et al. (2007)] and the serine and histidine form an interhelical hydrogen bond [Sulistijo, et al. (2009)].

As discussed previously, the Receptor Tyrosine Kinases comprise the largest set of solved TM dimer structures, many of which contain a GxxxG motif, and associate via a GAS_{right} fold. Erbb1 (EGFR) has a TxxxG motif at the interface. While threonine is not formally included in the GAS_{right} definition, it is a small polar amino acid with similar chemical and structual properties to serine. The TxxxG motif allows Erbb1 to fold in a right-handed GAS_{right} fashion [Endres, et al. (2013)]. Erbb2 has both a SxxxG and a GxxxG motif. The N-terminal SxxxG motif is at the interface, while the more C-terminal

GxxxG motif is unused in the homodimer. This structure also has a right-handed crossing angle of around -41°, a canonical GAS_{right} structure [Bocharov, Mineev, et al. (2008)]. The NMR structure of Erbb4 also folds as a GAS_{right} homodimer, which associates at a right-handed crossing angle at an N-terminal GxxxG motif. [Bocharov, et al. (2012)].

The dimeric structure of EphA1 has been solved at 2 different pH values, at which the interfacial contacts slide along a GxxxAxxxGxxxG motif in a GAS_{right} fold. The structural change at varying pH values may indicate that these extended motifs have a biological and functional role [Bocharov, Mayzel, et al. (2008)]. The GAS_{right} mediated structure of the Erbb1/2 heterodimer also associates via a GxxxG-like motif, wherein Erbb1 places TxxxGxxxG at the interface while Erbb2 contributes the sequence SxxxGxxxA [Mineev, et al. (2010)]. However, not all GxxxG containing membrane protein dimer NMR structures associate via their GxxxG motif. The structure of EphA2 places a GxxxG motif on the backside of its interacting interface, and associates with a left-handed crossing angle instead of the usual right-handed crossing of most RTKs [Bocharov, et al. (2010)].

1.7 The GAS_{right} motif is mediated by Cα-H hydrogen bonds

The placement of the small glycine residues at the interface of a GAS_{right} fold allows the dimerizing helices to pack together tightly. This tight interhelical distance allows for the formation of non-canonical Cα-H hydrogen bonds. Cα-H hydrogen bonds occur

between the hydrogen on the α carbon on one helix and the carbonyl oxygen on the opposing helix (Fig. 1.4). It is hypothesized that these weak hydrogen bonds in the hydrophobic and low dielectric membrane environment may drive, or at least assist in, the oligomerization and folding of α-helices in the membrane [Senes, et al. (2001)].

Linus Pauling was the first to write about C-H···O bonds when he noted the large difference in boiling temperature between acetylchloride and trifluoroacetylchloride. He hypothesized that the CH_3 group of acetylchloride, unlike the CF_3 group of trifluoroacetylchloride, could act as a hydrogen bond donor [Pauling (1960)]. However, it took more than 30 years until physical evidence for their existence in proteins was discovered. C to O contacts were evaluated, and it was observed that close inter-atomic distances matched the geometry of canonical hydrogen bonds [Derewenda, et al. (1994); Derewenda, et al. (1995)]. C-H···O bonds are rarely included in the definition of hydrogen bonds due to the fact that carbon is much less electronegative than nitrogen or oxygen; however, the context of the carbon is important, as the surrounding atoms can increase the electronegativity of the atom [Gu, et al. (1999)]. In the context of a protein the α carbon is positioned between the electronegative C=O and the N-H of the adjacent amide groups. Additionally, in the apolar environment of the membrane these electronegative Cα-H groups will not be shielded by water atoms. While this interaction is difficult to study experimentally as there are no simple mutations that can be performed to replace a backbone hydrogen or carbonyl oxygen, many studies have attempted to determine the strength of Cα-H hydrogen bonds and their importance to

protein interaction.

Initial studies of the strength of the Cα-H hydrogen bond were done as *ab initio* quantum calculations, one of the first of which was performed with the model molecule *N*,*N*-dimethylformamide (DMF) [Vargas, et al. (2000)]. In this calculation, multiple stable geometries occur between DMF dimers with multiple C-H···O bonds formed. In almost all cases, the hydrogen to oxygen distances are less than the sum of the van der Waal radii – evidence of a true C-H···O bond. These calculations estimated the strength of the Cα-H hydrogen bonds to be approximatively half the strength of a "canonical" N-H···O=C hydrogen bonds. While weaker than canonical hydrogen bonds, C-H···O bonds are more tolerant of deviations from the ideal hydrogen bond geometry [Gu, et al. (1999)]. Quantum calculations with methane and its fluorinated derivatives (CFH_3, CF_2H_2, and CF_3H) showed that a wider energy minimum existed when varying the O,H,C angle away from the ideal 120° angle than a canonical hydrogen bond. Also, the O to H distance energy minima occurs 0.3 to 0.4 angstroms further out than the canonical, with the strength of the interaction dying off far less quickly [Gu, et al. (1999)]. While these were important initial calculations, DMF and fluoromethane are not true amino acid mimics. Further quantum calculations were done in 2001 with an array of amino acids in their $NH_2CHRCOOH$ nonzwitterionic state (matching their neutral state in the interior of a protein), and with water used as the proton acceptor [Scheiner, et al. (2001)]. It was found that the binding energy of a hydrophobic amino acid Cα-H bond to a water molecule was in the range of -1.9 to -2.5 kcals/mol, approximately half of the

predicted energy (-4.5 kcals/mol) of a water to water hydrogen bond.

In 2004, the first non-computational measurements were made to determine the energetics of a non-canonical hydrogen bond. Arbely and Arkin measured the strength of a Cα-H hydrogen bond in a GpA peptide, using Fourier transform infrared spectroscopy (FTIR) [Arbely, et al. (2004)]. A deuterated glycine was added at position 79 and the CD_2 asymmetric stretching mode was compared between the wild type sequence and a variant that is known to abolish GpA dimerization (G83I) [Lemmon, et al. (1992)]. The difference in the frequency of the stretching mode allowed the calculation of an estimated ΔG of -0.88 kcals/mol for the hydrogen bond interaction. While this number is less than that of the obtained quantum gas phase calculations, the contribution is meaningful, especially considering there are six predicted Cα-H bonds in GpA. These networks of non-canonical bonds are consistently found in other structures [Senes, et al. (2001)].

A different approach was used by the Bowie lab to experimentally address the contribution of Cα-H···O to the folding of a membrane protein. They used an SDS unfolding assay to probe a Cα-H···O bond in bacteriorhodopsin (bR) (Cα-H of Ala51 to the O_γ of Thr24). Mutations of the threonine to alanine or valine were designed to remove a backbone-to-sidechain Cα-H···O bond. The alanine mutation was found to be stabilizing by 0.6 kcals/mol while the valine mutation was only destabilizing by 0.2 kcals/mol. A further mutation to serine found it to be stabilizing as well (by 0.3

kcals/mol), but the crystal structure of this mutant demonstrated that the serine did not maintain an orientation compatible with the formation of the Cα-H bond [Yohannan, et al. (2004)]. The authors suggested better vdw packing was obtained by moving the serine hydroxyl group, and the energetic gain in packing was more substantial than the hydrogen bond contribution.

A year after the FTIR and SDS unfolding studies, work done by Mottamal and Lazaridis helped to explain their apparent discrepancy [Mottamal, et al. (2005)]. They used the molecular modeling force field CHARMM with the implicit membrane model IMM1 to calculate the energies of the Cα-H bonds in structures based on the NMR structure of GpA and the crystal structure of bacteriorhodopsin. Based on this analysis both group's measurements appeared to be correct, but the geometry of the two proteins caused the difference in their conclusions – specifically the positioning of the hydrogen bond acceptor oxygen. The positioning of the oxygen acceptor was closer to the donor Cα in GpA, whereas in bR the O is closer to the adjacent nitrogen. It is this positioning of the O that makes the Thr to Ala “hydrogen bond” unstable. However, their work did note that other purported Cα-H bonds in bR are stabilizing.

The next major development occured in 2007. Park, et. al. noted that previous quantum mechanical calculations had been only done with optimized structures, rather than natural NMR or crystallographic structures [Park, et al. (2008)]. The researchers analyzed 263 Cα-H···O=C contacts from α-helical TM proteins, extracting the exact

geometries from their structures. The results showed that 89% of backbone-to-backbone hydrogen bonds were stabilizing, with 13% being stabilizing by more than 3 kcals/mol. The authors also looked at whether, as had previously been suggested, Cα-H···O=C bonds would be more frequent in membrane proteins than in soluble proteins. They found that there were, on average, almost triple the number of Cα-H···O=C contacts in membrane proteins. This suggests that these polar backbone-to-backbone bonds may be more important in protein folding in the hydrophobic environment of the membrane than in an aqueous environment.

Further evidence for the formation of Cα-H···O=C hydrogen bonds in membrane proteins was obtained by NMR in the SPMP BNIP3, solved in the detergent dodecylphosphocholine (DPC). The authors specifically looked for chemical shifts that corresponded to Cα-H···O=C hydrogen bond formation. They noted that a chemical shift of around 1.5 ppm occured for hydrogens involved in a non-canonical hydrogen bond based on quantum calculations. The C-terminal Hα2 glycine in the interfacial GxxxG motif is shifted downfield 0.91 ppm from the Hα2/3 average of the n-terminal glycine. Because of this shift the authors concluded that this is evidence of Cα-H···O=C hydrogen bonding [Sulistijo, et al. (2009)].

1.8 Overview of this Thesis

The forces that govern the folding and association of membrane proteins are not fully understood. In the hydrophobic environment of the lipid bilayer, the hydrophobic effect is

no longer a driving force of protein folding. Instead a combination of van der Waals, electrostatics and hydrogen bonding work together to mediate amino acid interactions. However, the relative contributions of these three forces, and the prevalence of the weak, Cα-H···O=C hydrogen bond is not well understood. By understanding association of single-pass membrane proteins, it is likely that these rules will extend across all membrane protein interactions.

Across all membrane proteins, a common motif, both in multi-pass and single-pass membrane proteins, is the GxxxG motif. The GxxxG motif and other related small residue motifs are often found in tandem with the GAS_{right} motif, a right-handed crossing of two parallel helices. Due to the close contact afforded by the small interfacial residues, backbone to backbone Cα-H···O=C bonds can form between the helices. While quantum calculations have shown that Cα-H···O=C bonds could be one-third to one-half the strength of a canonical hydrogen bond, there have been few *in vivo/vitro* tests to demonstrate if these bonds are an appreciable force in membrane protein folding.

My thesis work has focused on understanding the role of the small interfacial residues, especially glycine, and the their relation to the geometry of the GAS_{right} motif. Because of this understanding, I was able to computationally model GAS_{right} mediated protein dimers. While single-pass membrane proteins are the most common type of membrane protein, few structures have been solved. Therefore, being able to computationally

model a subset of SPMP interactions allows for more guided research and better understanding of the TM region in SPMP.

Chapter 2 poses the question of what aspects of the geometry of GAS_{right} motif allow the formation of networks of Cα-H···O=C bonds. By analyzing the entire universe of geometric positions that symmetrical homo-dimers can associate in, I found that there is a singular “hot-spot” of Cα-H···O=C hydrogen bond potential. This potential corresponds to the precise geometry of the GAS_{right} motif. The placement of a glycine at the interface is required for Cα-H···O=C hydrogen bond formation. This is due to the limited steric repulsion of the smallest amino acid, and the addition of a second hydrogen bond donor in the form of the hydrogen at the R group position. Adding a second glycine four residues away from the first glycine (creating a GxxxG motif) greatly enhances the Cα-H···O=C hydrogen bond formation. Using the knowledge gained from understanding the geometries required for Cα-H···O=C hydrogen bonds, I developed a *de novo* algorithm (CATM – Cα TransMembrane) to predict GAS_{right} structures. The CATM algorithm is able to correctly predict, to atomic accuracy, the structure of known GAS_{right} homo-dimers. This work highlights that by better understanding the forces governing protein interaction better structural predictions can be made.

In the Chapter 2 I show that the CATM algorithm can correctly predict the known structure of GAS_{right} proteins. In Chapter 3 I show that CATM can successfully predict the structure of a protein of unknown structure ADKC3. ADCK3 is a mitochondrial

kinase involved in the biosynthesis of the redox active lipid ubiquinone. ADCK3 is predicted to be a SPMP, and contains an extended GxxxGxxxG motif. CATM predicts the protein to dimerize in a GAS_{right} fold with the glycine motif at the interface. Extensive *in vivo* mutagenesis was done which confirmed the placement of the motif at the interface. This work is the first to show the correct prediction of a protein by CATM.

Chapters 2 and 3 demonstrate the geometric understanding of Cα-H···O=C hydrogen bonds, and how the application of this information can be used to predict both known and unknown GAS_{right} mediated homo-dimers. Here in Chapter 4, I address the problem of relating the CATM energy scores to the relative protein dimerization seen in nature. To do so, I associated the computational analysis to an experimental assay (TOXCAT), which can test for the relative strength of TM homo-dimerization in natural membranes. While TOXCAT is not as quantitative as other methods *in vitro*, it is higher-throughput, and sequence mutations can be easily performed. Using rationally designed constructs, I show that a strong correlation between TOXCAT and CATM score exists. This high degree of correlation can help guide future *in vitro* and *in vivo* research by allowing researchers to understand how and to what degree proteins associate.

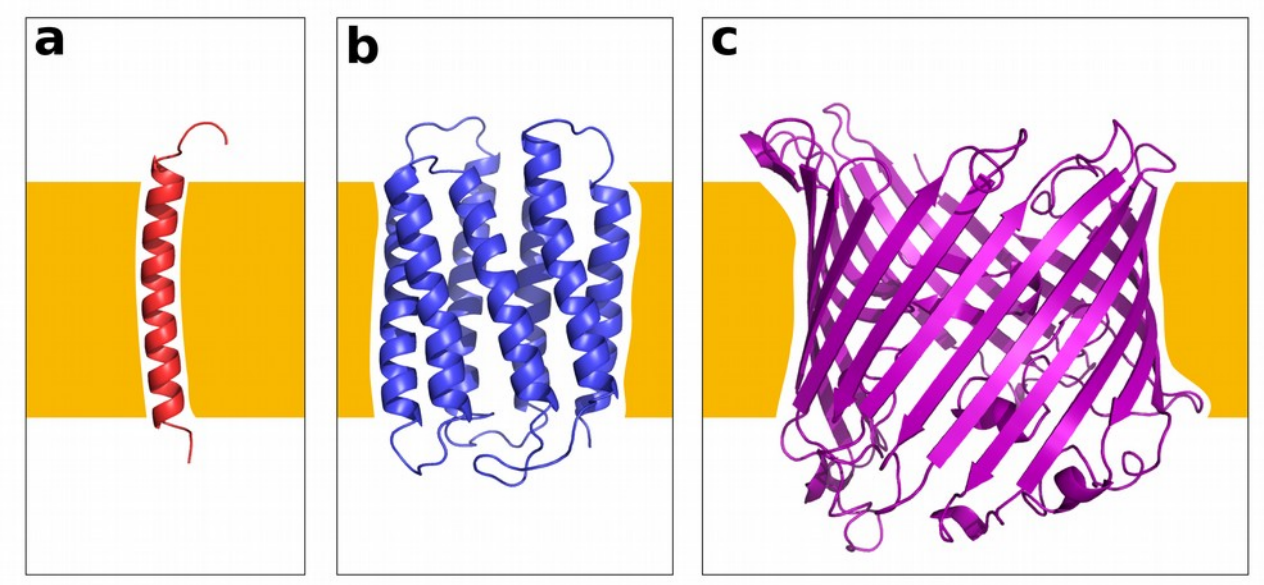


Fig. 1.1 Single-pass, Multi-pass, and Beta-barrel Membrane Proteins. a) Single-pass membrane proteins thread a single α-helix through the lipid bilayer. b) Multi-pass membrane proteins span the membrane two or more times via α-helices. c) β-barreled membrane proteins create a pore through the bilayer using a cylindrical β-sheet.

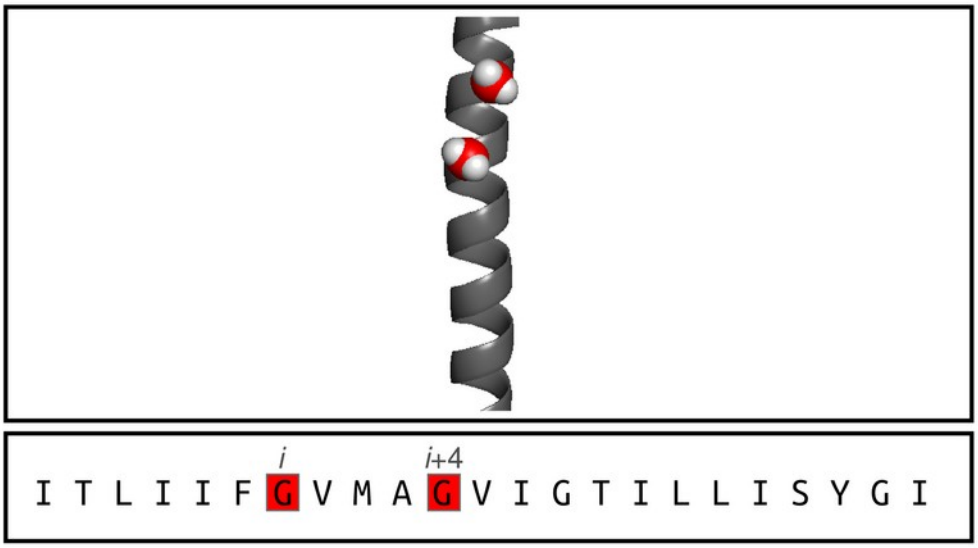


Fig. 1.2 The GxxxG motif. The GxxxG motif, is a common transmembrane sequence motif with two glycine residues spaced at *i* and *i*+4. Due to the periodic nature of the helix (~3.6 residues per turn), the two small amino acids are placed on the same face of the helix.

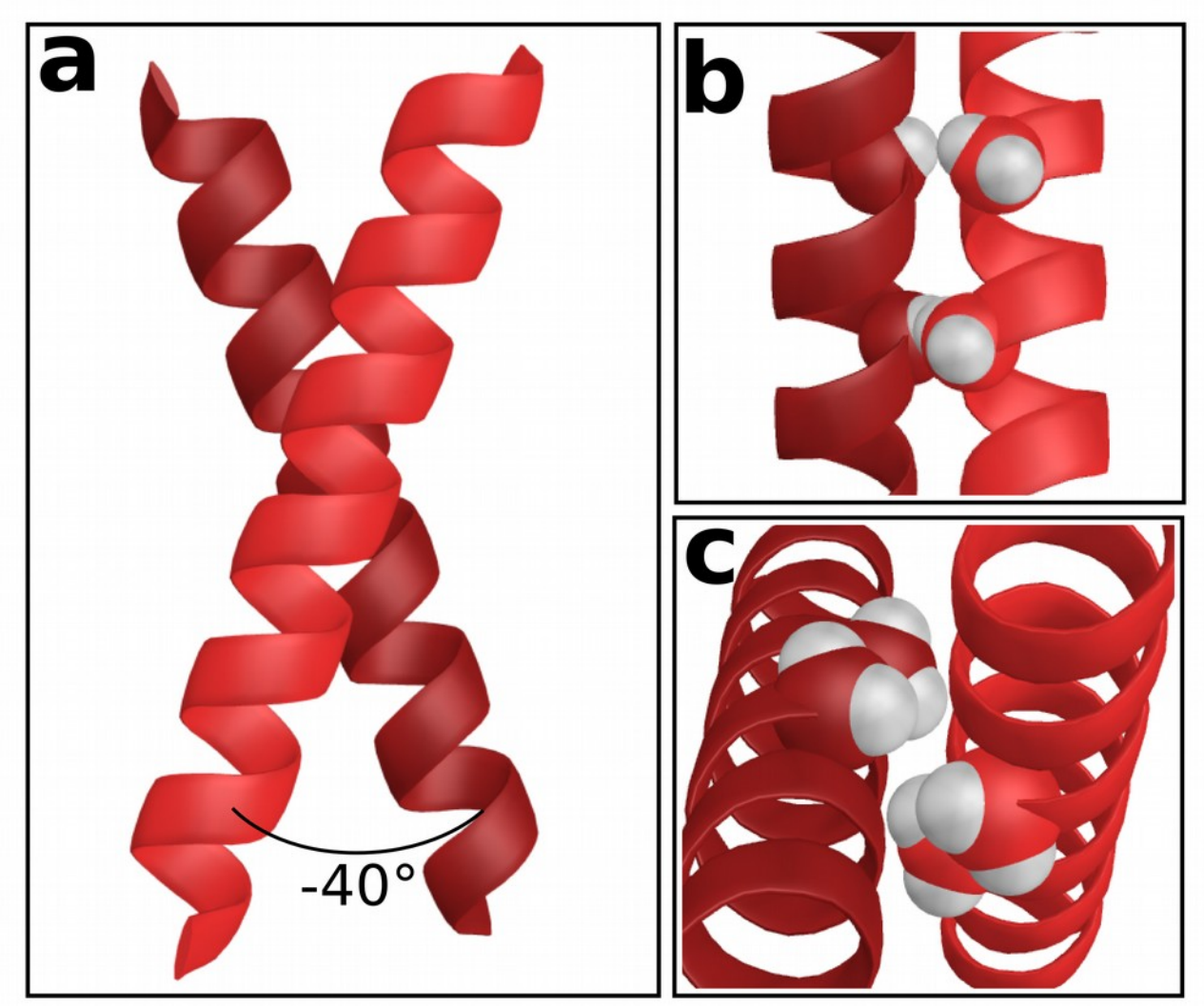


Fig. 1.3 GAS_{right} structural motif. a) The GAS_{right} motif is the association of two parallel transmembrane helices at a right-handed crossing angle of around -40 degrees. b&c) The structural motif places Glycine, Alanine, or Serine (GAS) residues at the interface. In the figure the common GxxxG motif is placed at the interface, with the glycine residues placed at *i* and *i*+4, the two residues are at the same face of the helix.

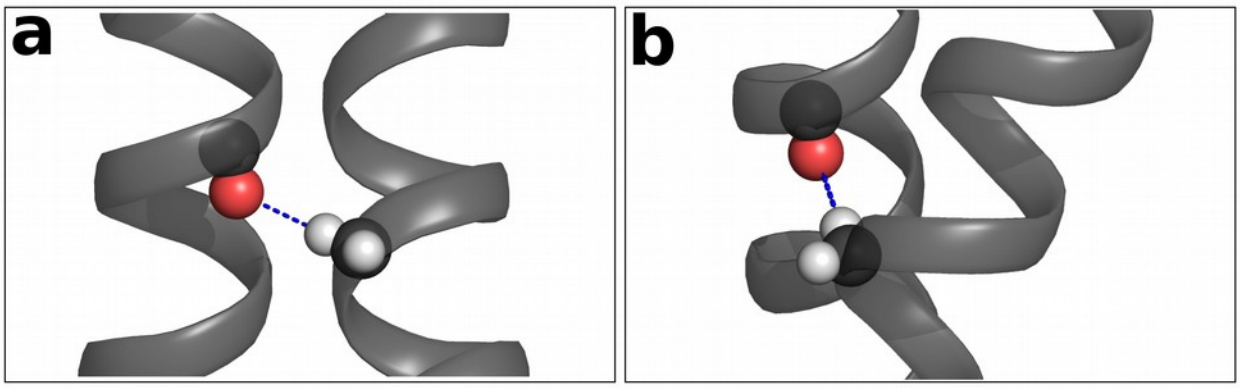


Fig. 1.4 Cα-H···O=C Hydrogen Bonds. Cα-H···O=C hydrogen bonds form between the α-carbon hydrogen on one helix to the carbonyl oxygen on the opposing helix. The presence of small residues, most importantly glycine, allow for the formation of these bonds. The bonds are commonly seen in GAS_{right} structures.

1.9 References

Arbely, Eyal, and Isaiah T. Arkin. "Experimental Measurement of the Strength of a Cα−H···O Bond in a Lipid Bilayer." *Journal of the American Chemical Society* 126.17 (2004): 5362-363.

Arnaout, M.A., B. Mahalingam, and J.-P. Xiong. "Integrin Structure, Allostery, And Bidirectional Signaling." *Annual Review of Cell and Developmental Biology* 21.1 (2005): 381-410.

Blume-Jensen, Peter, and Tony Hunter. "Oncogenic Kinase Signalling." *Nature* 411.6835 (2001): 355-65.

Bocharov, E. V., Y. E. Pustovalova, K. V. Pavlov, P. E. Volynsky, M. V. Goncharuk, Y. S. Ermolyuk, D. V. Karpunin, A. A. Schulga, M. P. Kirpichnikov, R. G. Efremov, I. V. Maslennikov, and A. S. Arseniev. "Unique Dimeric Structure of BNip3 Transmembrane Domain Suggests Membrane Permeabilization as a Cell Death Trigger." *Journal of Biological Chemistry* 282.22 (2007): 16256-6266.

Bocharov, Eduard V., Konstantin S. Mineev, Marina V. Goncharuk, and Alexander S. Arseniev. "Structural and Thermodynamic Insight into the Process of “weak” Dimerization of the ErbB4 Transmembrane Domain by Solution NMR." *Biochimica Et Biophysica Acta (BBA) - Biomembranes* 1818.9 (2012): 2158-170.

Bocharov, Eduard V., Maxim L. Mayzel, Pavel E. Volynsky, Konstantin S. Mineev, Elena N. Tkach, Yaroslav S. Ermolyuk, Alexey A. Schulga, Roman G. Efremov, and Alexander S. Arseniev. "Left-Handed Dimer of EphA2 Transmembrane Domain: Helix Packing Diversity among Receptor Tyrosine Kinases." *Biophysical Journal* 98.5 (2010): 881-89.

Bocharov, E. V., M. L. Mayzel, P. E. Volynsky, M. V. Goncharuk, Y. S. Ermolyuk, A. A. Schulga, E. O. Artemenko, R. G. Efremov, and A. S. Arseniev. "Spatial Structure and pH-dependent Conformational Diversity of Dimeric Transmembrane Domain of the Receptor Tyrosine Kinase EphA1." *Journal of Biological Chemistry* 283.43 (2008): 29385-9395.

Bocharov, E. V., K. S. Mineev, P. E. Volynsky, Y. S. Ermolyuk, E. N. Tkach, A. G. Sobol, V. V. Chupin, M. P. Kirpichnikov, R. G. Efremov, and A. S. Arseniev. "Spatial Structure of the Dimeric Transmembrane Domain of the Growth Factor Receptor ErbB2 Presumably Corresponding to the Receptor Active State." *Journal of Biological Chemistry* 283.11 (2008): 6950-956.

Bocharov, Eduard V., Dmitry M. Lesovoy, Sergey A. Goncharuk, Marina V. Goncharuk, Kalina Hristova, and Alexander S. Arseniev. "Structure of FGFR3 Transmembrane Domain Dimer: Implications for Signaling and Human Pathologies." *Structure* 21.11

(2013): 2087-093.

Burgess, Nancy K., Ann Marie Stanley, and Karen G. Fleming. "Determination of Membrane Protein Molecular Weights and Association Equilibrium Constants Using Sedimentation Equilibrium and Sedimentation Velocity." *Biophysical Tools for Biologists, Volume One: In Vitro Techniques Methods in Cell Biology* (2008): 181-211.

Derewenda, Z.S., U. Derewenda, and P.M. Kobos. "(His)Cε-H···O=C." *Journal of Molecular Biology* 241.1 (1994): 83-93.

Derewenda, Zygmunt S., Linda Lee, and Urszula Derewenda. "The Occurence of C–H · · · O Hydrogen Bonds in Proteins." *Journal of Molecular Biology* 252.2 (1995): 248-62.

Endres, Nicholas F., Rahul Das, Adam W. Smith, Anton Arkhipov, Erika Kovacs, Yongjian Huang, Jeffrey G. Pelton, Yibing Shan, David E. Shaw, David E. Wemmer, Jay T. Groves, and John Kuriyan. "Conformational Coupling across the Plasma Membrane in Activation of the EGF Receptor." *Cell* 152.3 (2013): 543-56.

Fisher, Lillian E., Donald M. Engelman, and James N. Sturgis. "Detergents Modulate Dimerization, but Not Helicity, of the Glycophorin A Transmembrane Domain." *Journal of Molecular Biology* 293.3 (1999): 639-51.

Fleming, Karen G., Anne L. Ackerman, and Donald M. Engelman. "The Effect of Point Mutations on the Free Energy of Transmembrane α-helix Dimerization." *Journal of Molecular Biology* 272.2 (1997): 266-75.

Furthmayr, H., and V. T. Marchesi. "Subunit Structure of Human Erythrocyte Glycophorin A." *Biochemistry* 15.5 (1976): 1137-144.

Furthmayr, H., M. Tomita, and V.T. Marchesi. "Fractionation of the Major Sialoglycopeptides of the Human Red Blood Cell Membrane." *Biochemical and Biophysical Research Communications* 65.1 (1975): 113-21.

Gu, Yanliang, Tapas Kar, and Steve Scheiner. "Fundamental Properties of the CH···O Interaction: Is It a True Hydrogen Bond?" *Journal of the American Chemical Society* 121.40 (1999): 9411-422.

He, Lijuan, and Kalina Hristova. "Physical–chemical Principles Underlying RTK Activation, and Their Implications for Human Disease." *Biochimica Et Biophysica Acta (BBA) - Biomembranes* 1818.4 (2012): 995-1005.

Hubbard, Stevan R., and W. Todd Miller. "Receptor Tyrosine Kinases: Mechanisms of Activation and Signaling." *Current Opinion in Cell Biology* 19.2 (2007): 117-23.

Käll, Lukas, Anders Krogh, and Erik L.l Sonnhammer. "A Combined Transmembrane Topology and Signal Peptide Prediction Method." *Journal of Molecular Biology* 338.5 (2004): 1027-036.

Kondo, N., K. Miyauchi, F. Meng, A. Iwamoto, and Z. Matsuda. "Conformational Changes of the HIV-1 Envelope Protein during Membrane Fusion Are Inhibited by the Replacement of Its Membrane-spanning Domain." *Journal of Biological Chemistry* 285.19 (2010): 14681-4688.

Krogh, Anders, Björn Larsson, Gunnar Von Heijne, and Erik L.l Sonnhammer. "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes." *Journal of Molecular Biology* 305.3 (2001): 567-80.

Langosch, Dieter, Bettina Brosig, Harald Kolmar, and Hans-Joachim Fritz. "Dimerisation of the Glycophorin A Transmembrane Segment in Membranes Probed with the ToxR Transcription Activator." *Journal of Molecular Biology* 263.4 (1996): 525-30.

Lemmon, Mark A., John M. Flanagan, John F. Hunt, Brian D. Adair, Barbara-Jean Bormann, Christopher E. Dempsey, and Donald M. Engelman. “Glycophorin A Dimerization Is Driven by Specific Interactions between Transmembrane α-Helices.” *Journal of Biological Chemistry* 267.11 (1992): 7683-7689.

Lemmon, Mark A., John M. Flanagan, Herbert R. Treutlein, Jian Zhang, and Donald M. Engelman. "Sequence Specificity in the Dimerization of Transmembrane α-helixes." *Biochemistry* 31.51 (1992): 12719-2725.

Lemmon, Mark A., and Joseph Schlessinger. "Cell Signaling by Receptor Tyrosine Kinases." *Cell* 141.7 (2010): 1117-134.

Mackenzie, Kevin R., James H. Prestegard,, and Donald M. Engelman. "A Transmembrane Helix Dimer: Structure and Implications." *Science* 276.5309 (1997): 131-33.

Manni, Sandro, Konstantin S. Mineev, Dinara Usmanova, Ekaterina N. Lyukmanova, Mikhail A. Shulepko, Mikhail P. Kirpichnikov, Jonas Winter, Milos Matkovic, Xavier Deupi, Alexander S. Arseniev, and Kurt Ballmer-Hofer. "Structural and Functional Characterization of Alternative Transmembrane Domain Conformations in VEGF Receptor 2 Activation." *Structure* 22.8 (2014): 1077-089.

Mineev, K.S., N.F. Khabibullina, E.N. Lyukmanova, D.A. Dolgikh, M.P. Kirpichnikov, and A.S. Arseniev. "Spatial Structure and Dimer–monomer Equilibrium of the ErbB3 Transmembrane Domain in DPC Micelles." *Biochimica Et Biophysica Acta (BBA) - Biomembranes* 1808.8 (2011): 2081-088.

Mineev, Konstantin S., Eduard V. Bocharov, Yulia E. Pustovalova, Olga V. Bocharova, Vladimir V. Chupin, and Alexander S. Arseniev. "Spatial Structure of the Transmembrane Domain Heterodimer of ErbB1 and ErbB2 Receptor Tyrosine Kinases." *Journal of Molecular Biology* 400.2 (2010): 231-43.

Mottamal, Madhusoodanan, and Themis Lazaridis. "The Contribution of Cα −H···O Hydrogen Bonds to Membrane Protein Stability Depends on the Position of the Amide." *Biochemistry* 44.5 (2005): 1607-613.

Munter, Lisa-Marie, Philipp Voigt, Anja Harmeier, Daniela Kaden, Kay E. Gottschalk, Christoph Weise, Rüdiger Pipkorn, Michael Schaefer, Dieter Langosch, and Gerd Multhaup. "GxxxG Motifs within the Amyloid Precursor Protein Transmembrane Sequence Are Critical for the Etiology of Aβ42." *The EMBO Journal* 26.6 (2007): 1702-712.

Park, Hahnbeom, Jungki Yoon, and Chaok Seok. "Strength of Cα −H···O=C Hydrogen Bonds in Transmembrane Proteins." *The Journal of Physical Chemistry B* 112.3 (2008): 1041-048.

Pauling, Linus. *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry*. Ithaca, NY: Cornell UP, 1960. 459.

Popot, J. L., and D. M. Engelman. "Membrane Protein Folding and Oligomerization: The Two-stage Model." *Biochemistry* 29.17 (1990): 4031-037.

Russ, William P., and Donald M. Engelman. "The GxxxG Motif: A Framework for Transmembrane Helix-helix Association." *Journal of Molecular Biology* 296.3 (2000): 911-19.

Russ, W. P., and D. M. Engelman. "TOXCAT: A Measure of Transmembrane Helix Association in a Biological Membrane." *Proceedings of the National Academy of Sciences* 96.3 (1999): 863-68.

Scheiner, S., T. Kar, and Y. Gu. "Strength of the CαH···O Hydrogen Bond of Amino Acid Residues." *Journal of Biological Chemistry* 276.13 (2001): 9832-837.

Schneider, D., and D. M. Engelman. "GALLEX, a Measurement of Heterologous Association of Transmembrane Helices in a Biological Membrane." *Journal of Biological Chemistry* 278.5 (2002): 3105-111.

Senes, Alessandro, Mark Gerstein, and Donald M. Engelman. "Statistical Analysis of Amino Acid Patterns in Transmembrane Helices: The GxxxG Motif Occurs Frequently and in Association with β-branched Residues at Neighboring Positions." *Journal of*

Molecular Biology 296.3 (2000): 921-36.

Senes, A., I. Ubarretxena-Belandia, and D. M. Engelman. "The C-H···O Hydrogen Bond: A Determinant of Stability and Specificity in Transmembrane Helix Interactions." *Proceedings of the National Academy of Sciences* 98.16 (2001): 9056-061.

Sulistijo, Endah S., and Kevin R. Mackenzie. "Structural Basis for Dimerization of the BNIP3 Transmembrane Domain." *Biochemistry* 48.23 (2009): 5106-120.

Sulistijo, Endah S., and Kevin R. Mackenzie. "Sequence Dependence of BNIP3 Transmembrane Domain Dimerization Implicates Side-chain Hydrogen Bonding and a Tandem GxxxG Motif in Specific Helix–Helix Interactions." *Journal of Molecular Biology* 364.5 (2006): 974-90.

The UniProt Consortium. “UniProt: a hub for protein information.” *Nucleic Acids Res.* 43 (2015) D204-D212.

Ullrich, Axel, and Joseph Schlessinger. "Signal Transduction by Receptors with Tyrosine Kinase Activity." *Cell* 61.2 (1990): 203-12.

Vargas, Rubicelia, Jorge Garza, David A. Dixon, and Benjamin P. Hay. "How Strong Is the Cα −H···O=C Hydrogen Bond?" *Journal of the American Chemical Society* 122.19 (2000): 4750-755.

Wallin, Erik, and Gunnar Von Heijne. "Genome-wide Analysis of Integral Membrane Proteins from Eubacterial, Archaean, and Eukaryotic Organisms." *Protein Science* 7.4 (1998): 1029-038.

Walters, R. F. S., and W. F. Degrado. "Helix-packing Motifs in Membrane Proteins." *Proceedings of the National Academy of Sciences* 103.37 (2006): 13658-3663.

Weiss, Carol D. “HIV-1 gp41: Mediator of Fusion and Target for Inhibition.” *AIDS Rev* 5 (2003): 214-21.

Winograd-Katz, Sabina E., Reinhard Fässler, Benjamin Geiger, and Kyle R. Legate. "The Integrin Adhesome: From Genes and Proteins to Human Disease." *Nature Reviews Molecular Cell Biology* 15.4 (2014): 273-88.

Yin, H., J. S. Slusky, B. W. Berger, R. S. Walters, G. Vilaire, R. I. Litvinov, J. D. Lear, G. A. Caputo, J. S. Bennett, and W. F. Degrado. "Computational Design of Peptides That Target Transmembrane Helices." *Science* 315.5820 (2007): 1817-822.

Yohannan, Sarah, Salem Faham, Duan Yang, David Grosfeld, Aaron K. Chamberlain, and James U. Bowie. "A Cα −H···O Hydrogen Bond in a Membrane Protein Is Not

Stabilizing." *Journal of the American Chemical Society* 126.8 (2004): 2284-285.

You, Min, Edwin Li, William C. Wimley, and Kalina Hristova. "Förster Resonance Energy Transfer in Liposomes: Measurements of Transmembrane Helix Dimerization in the Native Bilayer Environment." *Analytical Biochemistry* 340.1 (2005): 154-64.

Zhang, Shao-Qing, Daniel W. Kulp, Chaim A. Schramm, Marco Mravic, Ilan Samish, and William F. Degrado. "The Membrane- and Soluble-Protein Helix-Helix Interactome: Similar Geometry via Different Interactions." *Structure* 23.3 (2015): 527-41.

Zhang, Han, Qilin Ma, Yun-Wu Zhang, and Huaxi Xu. "Proteolytic Processing of Alzheimer’s β-amyloid Precursor Protein." *Journal of Neurochemistry* 120 (2011): 9-21.

Chapter 2

A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα-H hydrogen bonds

This chapter was prepared for publication as:
Benjamin K. Mueller*, Sabareesh Subramaniam*, and Alessandro Senes “A frequent, GxxxG-mediated, transmembrane association motif is optimized for the formation of interhelical Cα-H hydrogen bonds” *Proc Natl Acad Sci USA* 2014 111(10) E888-95

*This work was completed with equal contributions from myself and Sabareesh Subramaniam

Abstract

Carbon hydrogen bonds between Cα–H donors and carbonyl acceptors are frequently observed between transmembrane helices. Networks of these interactions occur often at helix-helix interfaces mediated by GxxxG and similar patterns. Cα–H hydrogen bonds have been hypothesized to be important in membrane protein folding and association, but evidence that they are major determinants of helix association is still lacking. Here we present a comprehensive geometric analysis of homo-dimeric helices that demonstrates the existence of a single region in conformational space with high propensity for Cα–H···O=C hydrogen bond formation. This region corresponds to the most frequent motif for parallel dimers, GAS-right, whose best known example is glycophorin A. The finding suggests a causal link between the high frequency of occurrence of GAS-right and its propensity for carbon hydrogen bond formation. Investigation of the sequence dependency of the motif determined that Gly residues are required at specific positions where only Gly can act as a donor with its “side chain” Hα. Gly also reduces the steric barrier for non-Gly amino acids at other positions to act as Cα donors, promoting the formation of cooperative hydrogen bonding networks. These findings offer a structural rationale for the occurrence of GxxxG patterns at the GAS-right interface. The analysis identified the conformational space and the sequence requirement of Cα–H···O=C mediated motifs; we took advantage of these results to develop a structural prediction method. The resulting program, CATM, predicts *ab initio* the known high-resolution structures of homo-dimeric GAS-right motifs at near atomic level.

2.1 Introduction

The transmembrane (TM) domains of membrane proteins that span the bilayer with a single helix are commonly engaged in oligomeric interactions that are essential for the structure and function of these proteins [Moore, et al. (2008)]. The interaction between these TM helices are often mediated by recurrent structural motifs, which are characterized by specific geometries and display sequence signatures in the form of specific amino acid patterns [Walters, et al. (2006)]. In this work we present a geometric analysis of one of the most important structural motifs, and implement a method for its structural prediction. The primary feature of this motif is the presence of inter-helical carbon hydrogen bonds that occur across the helix-helix interface between Cα–H donors and backbone carbonyl oxygen acceptors (Cα–H···O=C bonds) [Senes, et al. (2001)]. The sequence "signature" is the occurrence of glycine and other small amino acids (Ala, Ser) at the helix-helix interaction interface, generally spaced at *i*, *i*+4 to form patterns such as GxxxG, AxxxG, GxxxA, etc. [Senes, et al. (2000)]. These small amino acids are important to reduce the steric barrier for bringing the backbones of the opposing helices in close proximity, allowing the Cα and carbonyl oxygen (two backbone atoms) to come in contact and form hydrogen bonds [Senes, et al. (2001)].

While Cα–H···O=C hydrogen bonds can be observed in right- and left-handed TM helical pairs and in both parallel and anti-parallel orientations, they are most frequently associated with parallel right-handed pairs with a crossing angle around -40° [Senes, et al. (2001)]. This structural motif has been named GAS_{right} by Walters and DeGrado, from its sequence signature (Gly, Ala, Ser) and its crossing angle [Walters, et al. (2006)]. GAS_{right} – the fold of the glycophorin A TM dimer – is the most frequent motif for pairs of parallel helices and it appears to be extremely frequent in 2-fold symmetrical homo-dimers of single-pass proteins. Indeed, out of approximately a dozen high-resolution TM homo-dimers solved to date as many as five are representatives of the GAS_{right} motif [Lomize, et al. (2006)]. However, whether the Cα hydrogen bonds indeed represent a major stabilizing force in GAS_{right} motifs has yet to be demonstrated.

Carbon hydrogen bonds are commonly observed in proteins and nucleic acids,

where they can contribute to protein structure, recognition or catalysis [Horowitz, et al. (2012)]. While carbons are generally weak donors, the Cα atom of all amino acids is activated by the electron withdrawing amide groups on both sides, and quantum mechanics calculations suggest that the energy of Cα–H hydrogen bonds may be as much as one third to half of that of canonical donors in vacuum [Vargas, et al. (2000); Scheiner, et al. (2001)]. Carbon hydrogen bonds have been proposed to be particularly important in membrane proteins, the membrane being a low dielectric environment that, in principle, should enhance their strength [Senes, et al. (2001)]. However, obtaining an experimental measurement of their contribution remains difficult. To date, two groups have addressed this question experimentally, with differing results. Arbely and Arkin calculated a favorable contribution of -0.88 kcal/mol for the carbon hydrogen bond formed by Gly 79 in glycophorin A, using isotope-edited IR spectroscopy [Arbely, et al. (2004)]. Conversely, Bowie and coworkers found that a Cα–H···O bond to the side chain hydroxyl group of Thr 24 was only marginally stabilizing or even slightly destabilizing in a folding study of bacteriorhodopsin variants [Yohannan, et al. (2004)]. Mottamal and Lazaridis were able to reconcile this discrepancy by analyzing the different hydrogen bonding geometries of the two systems [Mottamal, et al. (2005)]. Further quantum mechanical calculations performed on geometries from protein crystal structures also suggested that indeed the orientation of the groups can determine whether an interaction may be strongly favorable or unfavorable [Park, et al. (2008)].

More studies are certainly needed to fully understand the energetic contribution of Cα hydrogen bonds in membrane protein folding and interaction. However, their common occurrence as structural elements in membrane proteins postulates that they play an important role [Senes, et al. (2001); Senes, et al. (2004)]. To further investigate this issue, we present an analysis of the propensity for Cα hydrogen bond formation as a function of helical geometry in symmetric homo-dimers. Remarkably, the analysis reveals the existence of a single high-propensity conformation that corresponds to the common GAS_{right} motif. By defining a suitable frame of reference for the geometries, we were able to investigate the specific sequence requirements of each position at the

helix-helix interface. The results rationalize the occurrence of GxxxG patterns in GAS_{right}, and provide a physical explanation for the typical right-handed geometry of the motif based on steric interactions and optimization of hydrogen bonding. Overall, the analysis suggests a strong causal link between the high frequency of occurrence of GAS_{right} and its propensity for Cα hydrogen bond formation.

The analysis defines a map of the conformational space that allows the formation of networks of carbon hydrogen bonds between helical dimers. It also identified strict sequence dependencies at specific positions of each individual geometry. Based on this information, we have also created a rapid structural prediction method for the identification of Cα–H···O=C mediated homo-dimers, which we call CATM (Cα TransMembrane). We show that CATM can predict the known high-resolution structures of homo-dimeric GAS_{right} motifs at near atomic level. Interestingly and perhaps surprisingly, we found that a minimalistic set of energy functions composed of a hydrogen bonding and a van der Waals function, is sufficient for achieving a highly accurate level of prediction.

2.2 Results and Discussion

2.2.1 Geometric definition based on the unit cell of the helical lattice

The first step for our geometric analysis was to identify a practical frame of reference to express the relative orientation of the helices, as illustrated in Fig. 2.1a. Two parameters are straightforward: the inter-helical distance, *d*, and the crossing angle, θ. The other two parameters, the axial rotation, ω, and the position of the crossing point along the helical axis, *Z,* require a reference, such as a specific Cα. We found that it is most intuitive to define the geometry relative to a reference unit cell in the helical lattice (the parallelogram connecting four Cα atoms on the helical face illustrated in Fig. 2.1b, and, as a planar projection, in Fig. 2.1c). For completeness, we explored conformational space so that the position of the point of closest approach *P* (i.e. the crossing point) samples the entirety of the unit cell. This is done by expressing *Z* and ω relative to the helical screw, producing two transformed unit vectors, *Z'* and ω', that run parallel to the principal components of the unit cell (Fig. 2.1c and 2.9). For convenience, we defined a naming convention for the positions that is relative to the reference unit cell. The positions at the four corners were designated as N1, N2, C1 and C2, where “N” and “C” indicate the N- and C-terminal sides of the parallelogram. These four atoms are relatively spaced at *i*, *i*+1, *i*+4 and *i*+5. The above reference frame and convention greatly helps the analysis and the discussion of the results.

2.2.2 Carbon hydrogen bond analysis reveals a bias for right-handed structures

To investigate the precise geometric requirements for the formation of inter-helical carbon hydrogen bonds, we performed a systematic evaluation of all homo-dimer geometries beginning with poly-Gly. Gly is the only amino acid that doubles the opportunity for hydrogen bond formation by the virtue of having two alpha hydrogens oriented approximately perpendicular to each other (109°) as well as being the residue that permits the two helices to come into the closest proximity. Therefore, poly-Gly is the “best case” sequence for forming carbon hydrogen bond networks, from a geometric stand point.

The hydrogen bonding propensity for each individual geometry was estimated with a hydrogen bonding function borrowed from SCWRL 4 [Krivov, et al. (2009)] and re-parameterized to include Cα donors (see Methods). The results are presented as color-coded heat maps in Fig. 2.2a. Each graph shows total hydrogen bond energy as a function of axial rotation (ω', on the x-axis) and crossing angle (θ, y-axis) for a different slice in *Z'*. For simplicity the inter-helical distance *d* is not explicitly graphed; instead, for each [ω', θ, *Z'*] point we plot only the energy (E_{min}) at the optimal distance (d_{min}). A larger number of *Z'* stacks, as well as the corresponding d_{min} values for each point are plotted in Fig. 2.10.

A single major high-propensity region is observed in the lower half of the plot, for right-handed crossing angles in the -30° to -50° range. This minimum is situated mid-way between the Cα carbon atoms (C2 and C1) in the ω' dimension, between 40˚ to 60˚. The region persists, with some variation, across the entire range of *Z'*. Interestingly, the minimum corresponds to the important GAS_{right} structural motif [Walters, et al. (2006)], a right-handed dimer characterized by presence of GxxxG-like patterns at the helix-helix interface [Senes, et al. (2000)]. Structural examples of GAS_{right} homo-dimers are glycophorin A [MacKenzie, et al. (1997)] and BNIP3 [Sulistijo, et al. (2009)], and the motif is also common within the fold of polytopic membrane proteins [Walters, et al. (2006); Senes, et al. (2001)].

2.2.3 GAS_{right} homo-dimeric motifs require a Gly at position C1

To investigate the sequence requirements for carbon hydrogen bonding and to understand the role of GxxxG like patterns in GAS_{right} motifs, we expanded the geometric analysis to poly-Ala helices in which one or more Gly were inserted in the sequence at specific positions. The poly-Ala sequence has minimal propensity to form hydrogen bonds (Fig. 2.11a) but when a single Gly is placed at C1, a significant restoration of the energies is observed for *Z'* values between 1.5 to 4.5 Å, that is, for dimers that have the point of closest approach in the middle section of the parallelogram (Figs. 2.2d and in more detail in Fig. 2.12). Above and below these *Z'* values the backbones are separated by the Cβ methyl groups of either positions N1 or C5 (the

amino acid at *i*±4 with respect to C1).

When a single Gly is placed at any position other than C1 (N1, N2 or C2) the hydrogen bonding energy landscapes present only very shallow minima (Figs. 2.2b, 2.2c, and in further detail Fig. 2.11b, 2.11c and 2.11d). This is because the Cβ of C1-Ala invariably comes in contact with the opposing helix, preventing the two helices from being in sufficient proximity. Therefore we conclude that C1 is the position with the most stringent requirement for Gly.

2.2.4 GxxxG motifs are important on the right-hand side of the unit cell

If a second Gly is added at *i*-4 (N1) or *i*+4 (C5) with respect to C1 to form a GxxxG motif on the right-hand side of the unit cell, the hydrogen bonding propensity increases very significantly. If two Gly are placed at N1 and C1, significant restoration of the propensities is present for *Z′* values that bring the crossing point closest to N1 (Figures 2.5e and 2.13a). If two Gly are placed at C1 and C5, the increase is observed for low *Z′* values that have a crossing point closest to C1 (Fig. 2.13b). Finally, when N1, C1 and C5 are all Gly to form a Gly zipper motif GxxxGxxxG [Kim, et al. (2005)], the energy landscape looks very similar to the poly-Gly results (Figs. 2.5f and 2.13c). Again, addition of Gly residues on any of the left side positions (N2, C2 or C6, while keeping C1 as Gly) has a negligible effect on the hydrogen bonding energies (Fig. 2.14).

The marked distinction between the positions on the right side of the unit cell (N1, C1, C5) and those on the left side (N2, C2, C6) arises from the different orientation of the Cβ and Hα atoms with respect to the interface. This is schematically illustrated in Fig. 2.3a. The Cβ atom of C2 points away from the interface whereas the Cβ of C1 is oriented directly toward the opposing helix. For this reason, larger amino acids can be accommodated at C2, but Gly is required in C1 to allow the two backbones to come into close proximity. A similar argument applies to N1/N2 and C5/C6 as well.

2.2.5 GAS_{right} motifs are optimized for Cα hydrogen bond network formation

Gly performs a second important function as a donor when present at the right-

hand side positions. As illustrated in Fig. 2.3a, any amino acid can donate at C2 because the Hα atom is pointed toward the interface. However, that same hydrogen is oriented laterally and away from the interface at C1. As schematically illustrated in Fig. 2.3b, only Gly can donate from the right side positions (C1, N1, C5) because its “side chain” hydrogen is in the correct orientation. The same point is illustrated in structural terms in Fig. 2.3c.

It follows that both amino acids at C1 and C2 can simultaneously donate to the opposing helix only if C1 is a Gly. However, this requires a correct alignment with acceptors on the opposing helix. As illustrated in Fig. 2.4 using a superimposition of helical lattice projections, the crossing angle of GAS_{right} motifs is optimal for the this purpose. A -40° crossing angle aligns the two donors at C1 and C2 with two carbonyl oxygen atoms spaced at *i* and *i*+3 on the opposing helix. This is also shown in structural terms in Fig. 2.4c.

Overall, the analysis presents a compelling picture: the GAS_{right} coincides with the major hot-spot for carbon hydrogen bonding. From a steric stand point, the geometry appears ideal to allow backbone contacts as long as C1 and either N1 or C5 (or both) are Gly residues. The Gly residues at these same positions are also able to cooperatively extend the hydrogen bonding network by the virtue of having their second hydrogen oriented toward the interface. Finally, the -40° crossing angle is ideal for the simultaneous involvement of C1 and C2 (and, similarly, N1/N2 or C5/C6) in hydrogen bonding interactions. In our opinion, this finding suggests a strong causal link between the high frequency of the GAS_{right} motif in the structural database and its propensity to form networks of carbon hydrogen bonds, supporting the hypothesis that these interactions are important contributors to helix-helix association.

2.2.6 A high-throughput structural prediction method for GAS_{right} motif

The analysis presented above shows that only a small fraction of homo-dimer conformational space allows for the formation of Cα–H···O=C hydrogen bond networks. It also indicates that positions at the interface may have stringent sequence requirements for Gly or a limited set of amino acids. On these premises, we

hypothesized that it would be possible to create a rapid method to recognize sequence signatures compatible with the formation of GAS_{right} motifs.

To develop and implement the method, which we named CATM, we systematically subdivided the homo-dimer conformational space that allows formation of Cα–H···O=C bonds into a comprehensive “grid” of representative dimer conformations. We then established the specific sequence requirements of each conformation (sequence rules). In this implementation, we did not limit the space to the right-handed region, but allowed any dimer that displayed formation of at least two pairs of symmetrical hydrogen bonds.

When the primary sequence of a TM domain of interest is provided to CATM, the sequence is built in full atoms over each representative dimer that is compatible with the sequence rules. The two helices are placed at the inter-helical distance in which the two backbones still form a network of carbon hydrogen bonds (d_{out}, which is pre-calculated for each dimer). The helices are then moved closer followed by optimization of the side chains, until the energy reaches a minimum. At that point, the geometry of the dimer is locally optimized with a brief Monte Carlo procedure consisting of cycles in which all four inter-helical parameters changed randomly (d, Z, ω, θ).

At the end of each docking, the energy of the dimer is subtracted from the energy of the helices separated at a distance to obtain an interaction energy. Only the solutions with a negative interaction energy are preserved. Finally, all closely related solutions are clustered by similarity (RMSD < 2Å), and the lowest energy structure is reported as a representative model of its cluster. CATM is explained in full detail in the Methods, and is freely available for download with MSL, a C++ open source macromolecular modeling software library, at http://msl-libraries.org [Kulp, et al. (2012)].

2.2.7 A minimalistic set of energy functions predicts known structures with near atomic accuracy

We tested CATM against five known homo-dimeric GAS_{right} structures: glycophorin A [MacKenzie, et al. (1997)], BNIP3 [Sulistijo, et al. (2009); Bocharov, et al.

(2007)], and three members of the Tyrosine Receptor Kinase family, EphA1 [Bocharov, et al. (2008)], ErbB1 (EGFR) [Endres, et al. (2013)] and ErbB4 [Bocharov, et al. (2012)]. We began testing using a simple combination of hydrogen bonding (E_{hbond}) and van der Waals (E_{vdw}) to score the structural models. Perhaps surprisingly, we found that this minimalistic set of energy functions predicts the structures at near atomic precision, and in all but one case, the native structure corresponds to the lowest energy model. The finding validates our hypothesis that Cα hydrogen bonds can be an important guiding element for structure recognition, because they offer multiple anchor points between backbones, and because they are strongly dependent on good packing, given that the interactions can be easily disallowed by steric clashes [Senes (2011)]. All predicted models discussed below can be downloaded from http://seneslab.org/CATM/structures.

CATM returned 63 solutions for Glycophorin A, the first TM dimer solved by solution NMR [MacKenzie, et al. (1997)] and a major biophysical model systems for membrane protein association [MacKenzie, et al. (2008)]. The 63 solutions were clustered into 5 distinct models. The relationship between RMSD and energy for all 63 structures is plotted in Fig. 2.15. The lowest energy model predicted by CATM (Model 1) is a very close match of the NMR structure (Fig. 2.5). Measured over the entire TM helix (residues 73-95), the Cα RMSD is 1.31 ± 0.24 Å (average and standard deviation measured against the 20 NMR models). Measured over the segment that encompasses the interaction interface, discarding the contribution of the divergent ends, the RMSD reduces to 1.1 ± 0.21 Å (residues 75-87, marked in darker blue in Fig. 2.5a). A side by side comparison of the predicted model and the experimental structure shows the matching hydrogen bonding network and the conformation of the interfacial side chains (Fig. 2.5, panels b and c). A difference between the two structures is the conformation of Thr 87 which accepts an Cα hydrogen from Val 84 on the opposing helix in the NMR structure, while in the lowest energy CATM model the hydroxyl group of Thr 87 is involved in an inter-helical canonical hydrogen bond, which is consistent with a solid state NMR structure of the dimer [Smith, et al. (1994)]. Fig. 2.5 also shows the position of the point of closest approach in the unit cell at the interface of the dimer.

It should be noted that glycophorin A complies with the “Gly at C1” rule identified in our analysis, as all other structures analyzed in the following paragraphs. In fact, C1 is the only position that is invariably Gly across all the examples.

The second structural prediction is BNIP3, a very stable TM dimer [Sulistijo, et al. (2006)] characterized by a very short inter-helical distance (6.5 Å). The interface consists of a glycine zipper motif (A_{176}xxxG_{180}xxxG_{184}). As shown in Fig. 2.6a, the model is extremely similar to the NMR structure [Sulistijo et al. (2009); Bocharov, et al. (2007)]. The RMSD of the helical region of the entire TM domain is 1.10 ± 0.36 Å and only 0.56 ± 0.17 Å when it is computed only for the region that participates to the helix-helix interaction. The model replicates the network of carbon hydrogen bonds observed in BNIP3 and all interfacial side chains are predicted in the correct rotamer, as evident in the side-by-side comparison of panels b and c of Fig. 2.6. In addition, CATM accurately captures the inter-helical hydrogen bond between the side chain of His 173 (donor) and Ser 172 (acceptor), an important feature that contributes to the dimer's stability [Lawrie, et al. (2010)].

The third comparison is EphA1, which was solved by solution NMR in bicelles at two different pH conditions [Bocharov, et al. (2008)]. The dimer displays a conformational change induced by change in protonation state of a membrane embedded Glu residue (E547). CATM captures both conformations with good accuracy (Fig. 2.7). The low pH structure is predicted by Model 1 with a Cα RMSD of 1.26 Å. The higher pH structure is predicted by Model 4 with an RMSD of 1.48 Å. The structures are related by a shift of the crossing point of about 3 Å toward the C-terminus that brings the crossing point from the top half to the bottom half of the Glycine zipper motif (A_{550}xxxG_{554}xxxG_{558}), as schematically shown in Fig. 2.7c. Interestingly, the authors also report the presence of a minor component of some cross-peaks in the higher pH conditions, suggesting a second species (about 10%) was present in the sample [Bocharov, et al. (2008)]. While a structural model could not be calculated and was not reported for this minor species, the authors suggest that this competing state associates through the C-terminal GxxxG-like motif (A_{560}xxxG_{564}), and identify the amino acids

involved at the interface as Leu 557, Ala 560, Gly 564 and Val 567. This description is consistent with the interface of Model 2 produced by CATM.

The final two test cases are both members of the epidermal growth factor receptor family [Schlessinger (2000)]. As shown in Fig. 2.8a, the NMR structure of ErbB4 [Bocharov, et al. (2012)] is predicted well by CATM, with an RMSD of 0.81 Å across the interacting region. However, our prediction of ErbB1 (EGFR) is not in agreement with the experimental structure, the only case among the five structures tested. The experimental structure interacts through the N-terminal TxxxG motif [Endres NF et al. (2013)], and this structure is predicted by CATM's Model 3 with a Cα RMSD 0.77Å (Fig. 2.8b). Instead Model 1 is a well packed dimer that interacts through C-terminal side AxxxG motif, of the TM helix and is a likely candidate for a postulated inactive state of the receptor [Endres, et al. (2013); Fleishman, et al. (2002)]. As in the case of EphA1, this finding highlights the potential of offering alternative structural models that may reflect distinct functional states of the TM dimers.

The TM region of another member of the same family, ErbB2, has also been solved by NMR in dimeric form [Bocharov, et al. (2008)]. The NMR model has a crossing angle of -41° and an inter-helical distance of 7.6 Å, however, this structure is not mediated by Cα hydrogen bonds. Analysis of its geometry reveals that the ω' angle (12°) is incompatible with Cα hydrogen bond formation (Fig. 2.16). For this reason, the structure is outside the scope of conformational space explored by CATM, and thus it cannot be predicted by the program. Instead, CATM produced two unrelated GAS_{right} models, one mediated by the C-terminal GxxxG motif, the other mediated by the N-terminal SxxxG motif. Similarly to the previous cases, we note that it is possible that the CATM models may correspond to alternative physiological states of the dimer. The ErbB2 models are also available at http://seneslab.org/CATM/structures/.

2.3 Conclusions

We have presented an analysis of carbon hydrogen bonding as a function of helix orientation in TM homo-dimers. The analysis demonstrates that there is a single region of conformational space for homo-dimers with a high propensity for formation of hydrogen bond networks. Remarkably, this area corresponds to the GAS_{right} motif – the frequently occurring fold of glycophorin A – lending strong support to the hypothesis that optimization of carbon hydrogen bonding is a major driving factor in its assembly. The analysis also provides a rational structural interpretation of the occurrence of GxxxG motifs in GAS_{right} homo-dimers, indicating that the Gly residues are essential on a specific side of the helix interface for steric reasons and to act as hydrogen bonding donors.

Based on the analysis, we have created a rapid method for the structural prediction of GAS_{right} homo-dimers. We have shown that with a surprisingly simple set of energy functions (E_{hbond}+E_{vdw}), CATM predicts the known structures of GAS_{right} homo-dimers with near atomic precision. Future work is necessary to refine, verify and expand the scoring functions. For example, a membrane model such as a depth-depended potential [Senes, et al. (2007)] or an implicit solvent [Lazaridis (2003)], is likely to improve the predictions and any correlation between the computational score and the thermodynamic stability. Nevertheless, CATM appears to capture the essence of GAS_{right} motifs already in the current form, and therefore the method is already applicable to the rapid prediction of unknown structures.

2.4 Methods

2.4.1 Software

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

2.4.2 Creation of inter-helical geometries

Two helices, 31 residues in length, were created in idealized conformation, oriented with their axes aligned with the z-axis and the Cα atom at position 16 placed on the x-axis. Position 16 is the position designated as C2 in Fig. 2.1c. To create a dimer, the following transformations were performed in order: a rotation around the z-axis (determining the axial rotation ω), a translation along the z-axis (determining the position of the crossing point *Z* in the z-dimension), a rotation around the x-axis (determining the crossing angle θ), and a translation along the x-axis (determining the inter-helical distance *d*). One of the two helices was finally rotated around the z-axis by 180° to produce 2-fold symmetry.

The geometric analysis was performed so that the point of closest approach *P* would explore the entire unit cell defined by N1,N2,C1,C2 as in Fig. 2.1c. The transformations were performed using a redefined set of geometric parameters [d, θ, ω', *Z'*], where ω', *Z'* are unit vectors that go in the direction of the principal components of the unit cell of the helical lattice using the mathematical relationships defined in Fig. 2.9. The conformational space was explored at discrete intervals with the following step sizes: *d*: 0.1 Å; ω': 1°; *Z'*: 0.1 Å; θ: 1°. The crossing angle θ was constrained to be in the -55° to +55° range.

2.4.3 Energy functions and definitions

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 C++ libraries [Kulp, et al. (2012)]. Cα hydrogen bonds have been included as part of the energy functions of ROSETTA Membrane [Barth, et

al. (2007)]. We derived a similar adaptation for the SCWRL 4 function by adding the following parameters for Cα donors: B=60.278; D_0=2.3 Å; σ_d=1.202 Å; α_{max}=74.0°; β_{max}=98.0°. These parameters reduce the hydrogen bonding energy to approximately half that of canonical bonds, and adjust the optimal distance and the angular dependencies.

In the text below the energy of a model is computed as the difference between the the dimer energy minus the energy of the separated monomers (referred to as interaction energy), with the side chains optimized independently in the two states. All side chain optimization procedures were performed using the Energy-Based Conformer Library applied at the 95% level [Subramaniam, et al. (2012)] with a greedy trials algorithm [Xiang, et al. (2001)] as implemented in MSL.

2.4.4 Determination of Cα–H···O energy landscapes

The energy landscapes were determined for all [θ, ω', *Z'*] coordinates. Two helices were initially placed at d = 10 Å. The energies were evaluated and the helices were moved closer to each other in 0.1 Å steps until a lowest energy (E_{min}) conformation was identified at a distance d_{min}. Fig. 2.2 plots E_{min} as a function of [θ, ω', *Z'*]. A plot of the corresponding d_{min} values is provided for poly-Gly in Fig. 2.10.

2.4.5 Development of CATM

CATM is a structure prediction program that performs a systematic search in the subset of homo-dimer conformational space that allows formation of inter-helical Cα–H···O hydrogen bonds. The creation of CATM consisted of the definition of the search space and the derivation of a set of sequence exclusion rules. The execution phase of CATM (the actual structure prediction for a given sequence) is schematically illustrated in Fig. 2.17.

2.4.6 Definition of the search space

The definition of the search space was based on the geometric analysis of poly-Gly. We selected all conformations in [θ, ω', *Z'*] space that display at least four inter-helical Cα–H···O hydrogen bonds (two symmetrical pairs). This search yielded a set of approximately 90,000 structures which were then filtered by similarity using a 2.0 Å

RMSD criterion to create a representative set of 463 geometries. For each representative geometry we recorded the maximum inter-helical distance in which four hydrogen bonds still exist (d_{out}).

2.4.7 Definition of the sequence rules

Each representative geometry *G* was constructed as poly-Gly and was set at d_{out}. Every amino acid *X* type was built at every position *j* in every *G* and its conformation was optimized. If the interaction energy was unfavorable by more than 10 kcal/mol, a sequence rule was recorded stating that the *X* is not allowed at *j* in *G*. These rules allow for the exclusion of non-productive sequences from the expensive all atom modeling phase.

2.4.8 The CATM program

The input sequence is threaded into a set of different registers at each of the 463 representative geometries (Fig. 2.17). For each register, CATM checks if the sequence rules are met. If the rules are met, the sequence is built on the backbone in all atoms, and the helices are placed at d_{out}. The inter-helical distance is reduced in steps of 0.1 Å, and 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 10 Monte Carlo backbone perturbation cycles in which all inter-helical parameters (d, θ, ω, Z) are locally varied. If the final interaction energy is negative, the solution is accepted. The solutions are then clustered using an RMSD criterion (2 Å) to produce a series of distinct models, with all individual solutions provided as an NMR-style PDB file.

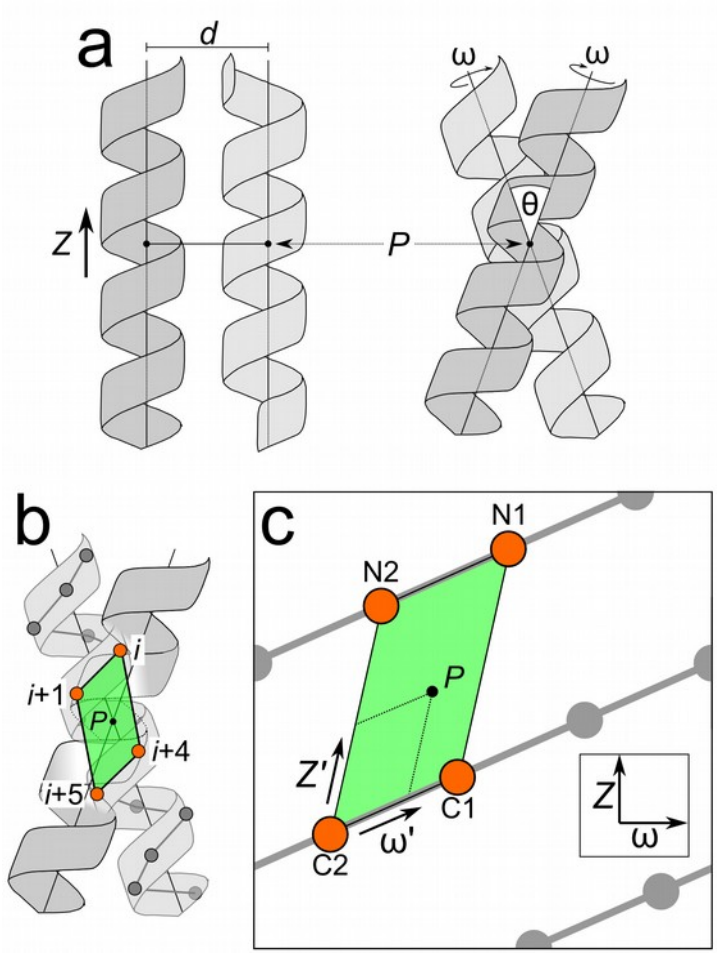


Fig. 2.1 Carbon hydrogen bond formation has preferential regions in inter-helical space. a) Definition of 4 parameters that define the geometry of a symmetrical dimer: the inter-helical distance *d*; the crossing angle θ; the rotation of the helix around its axis ω; and the vertical position *Z* of the point of closest approach between the two helical axes (the crossing point *P*). b) The coordinates can be redefined by expressing them as a function of the unit cell (green) on the helical lattice that contains the point of closest approach *P.* The four interfacial positions that surround the the point of closest approach are designated as N1 (relative position *i*), N2 (*i*+1), C1(*i*+4) and C2 (*i*+5). The

principal axes are the rotation along the helical screw (ω') and the vector between C2 and C2 (*Z'*). The mathematical relationship between (ω, *Z*) and (ω', *Z'*) is provided in Fig. 2.9.

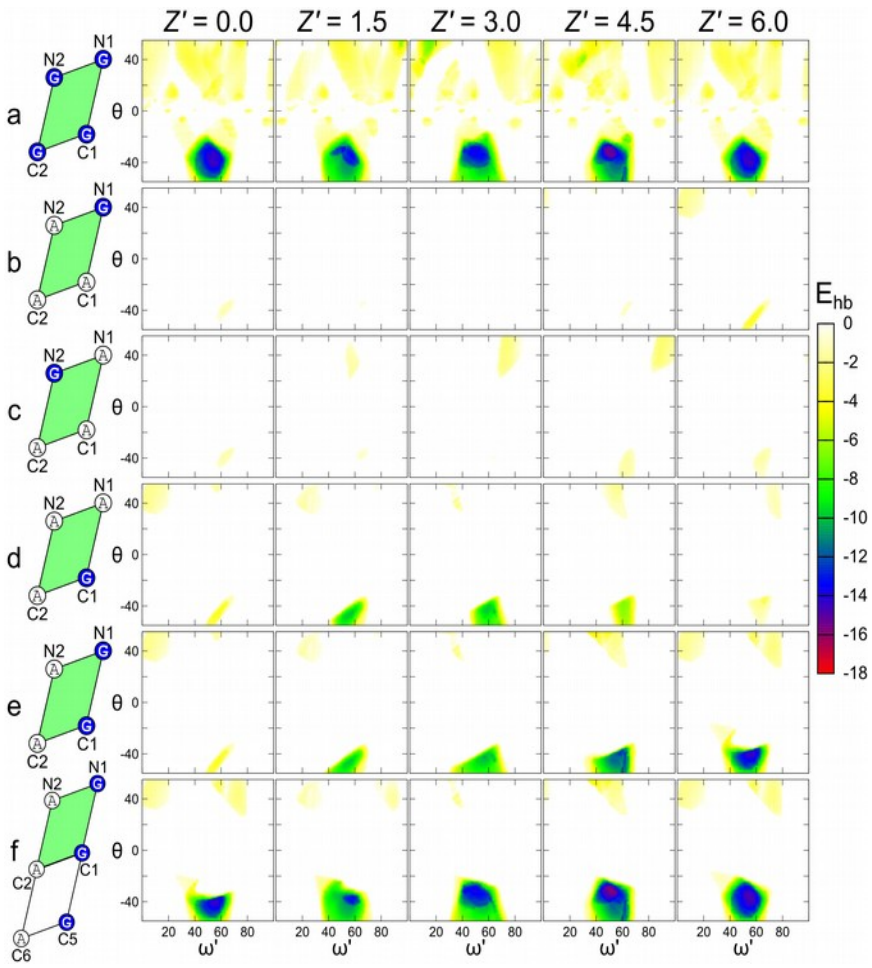


Fig. 2.2 Position C1 must be a Gly for carbon hydrogen bond formation. A map of the carbon hydrogen bonding energy (color bar) as a function of inter-helical geometry (ω': x-axis, θ: y-axis; *Z′*: panels). a) Analysis of poly-Gly: a single broad minimum is observed centered around a region with a right handed crossing angle θ of approximately -30° to -50°. The minimum persists with variation along the entire *Z′* stack. b, c and d) Poly-Ala sequences with a single Gly at specific positions as indicated on the left-hand side of the figure. The propensity to form hydrogen bonds is almost completely removed compared to poly-Gly unless the amino acid at position C1

is a Gly. (d) e) Introduction of a GxxxG motif at the positions N1 and C1 restores some of the low energy regions for higher *Z′* values. f) When a third Gly is added at C5 the propensity becomes very similar to poly-Gly. In each panel the lowest energy (*vdw* + *hbond*) across all inter-helical distances (E_{min} at d_{min}) is plotted for each point.

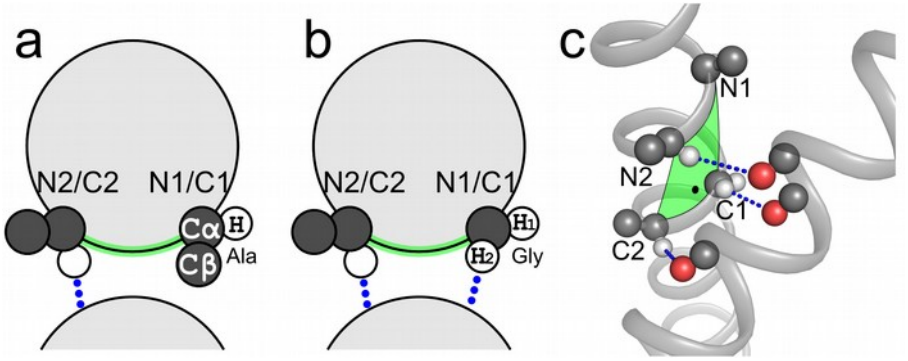


Fig. 2.3 Structural distinction between interfacial positions. a) The amino acids on the left side of the unit cell (N2 and C2) orient their α-hydrogen toward the interface while their the Cβ points laterally, and thus these position can accommodate larger amino acid types. The situation is reversed for positions N1 and C1: the α-hydrogen is oriented laterally and the side chain points directly toward the opposing helix. Larger amino acids in this position may not be accommodated. b) Gly is the only amino acid type that can form a hydrogen bond using the “side chain” hydrogen when present at positions N1 or C1. c) Structural example: in this case the crossing point is close to C1, and there is sufficient space to allow Ala at N1.

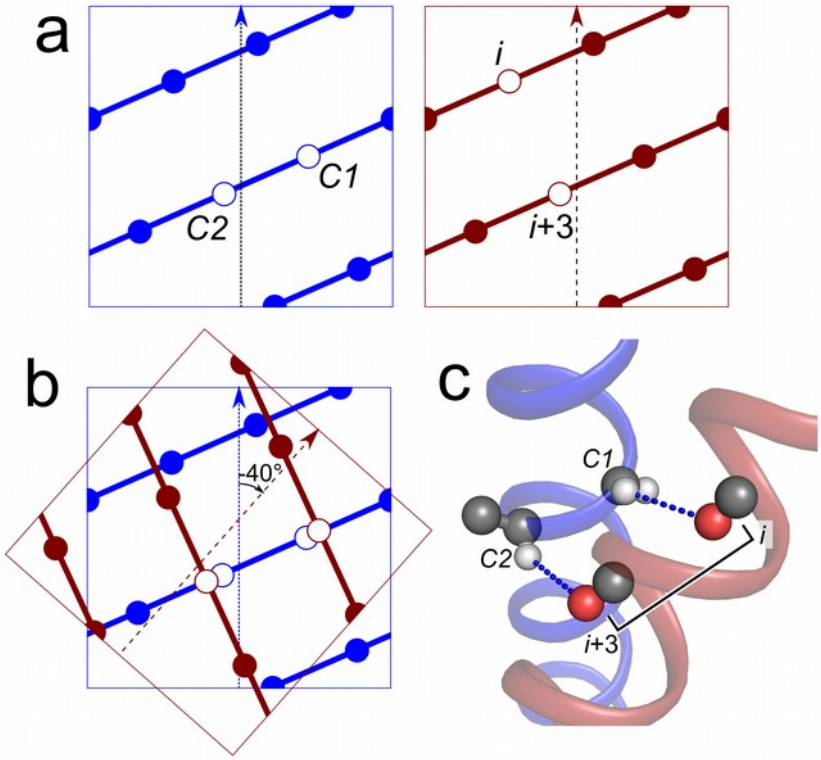


Fig. 2.4 In a GAS_{right} motif the C1 and C2 donors are aligned with carbonyl acceptors at *i*, *i*+3 on the opposing helix. a) Helical lattices highlighting the C1 and C2 donor positions (left, blue) and carbonyl acceptors at *i*, *i*+3 on the opposing helix (right, dark red). b) A superimposition of the two lattices followed by a -40° rotation aligns the donors and acceptors. c) Structural representation of the same alignment.

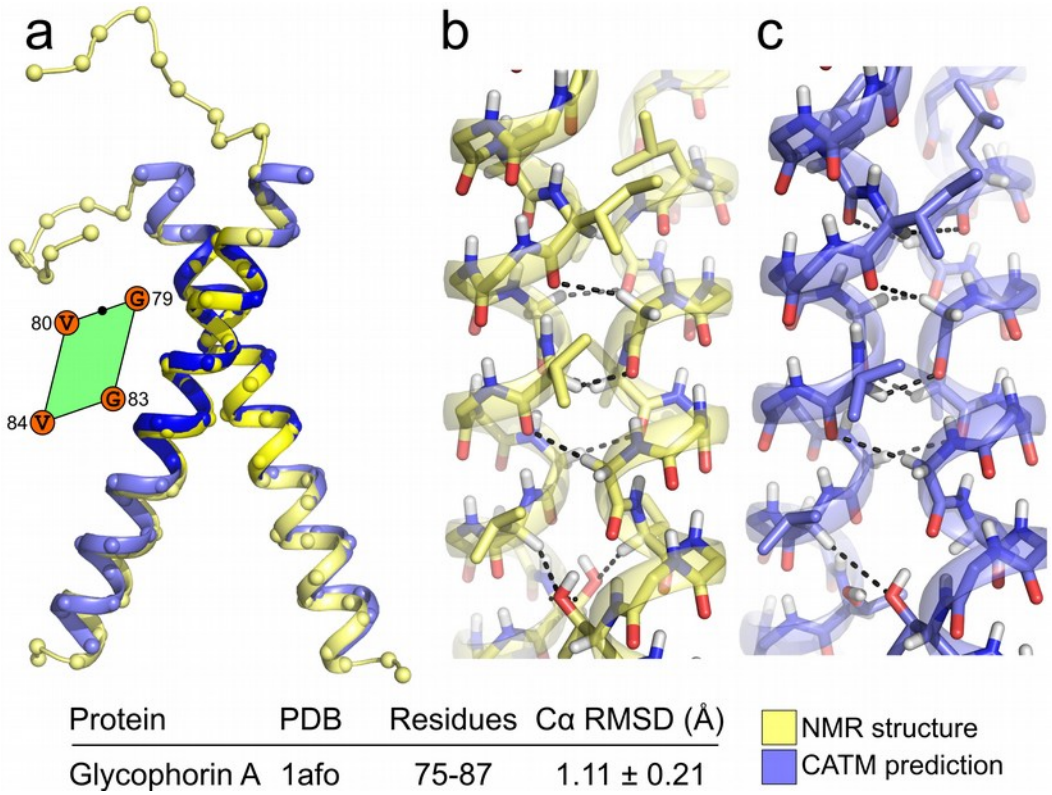


Fig. 2.5 CATM prediction of the TM domain of Glycophorin A. a) Backbone superimposition of the NMR structure (yellow) and the predicted model (blue). The Cα RMSD in the region that encompasses the interface is indicated and highlighted in darker blue and yellow in the ribbon. Panels b and c show the full-atom comparison between the experimental structure and the prediction. The CATM model is close to atomic level, with a similar network of carbon hydrogen bonds. The NMR structure and CATM model differ in the orientation of Thr 87, which hydrogen bonds to its own backbone, while CATM predicts the formation of an inter-helical canonical hydrogen bond.

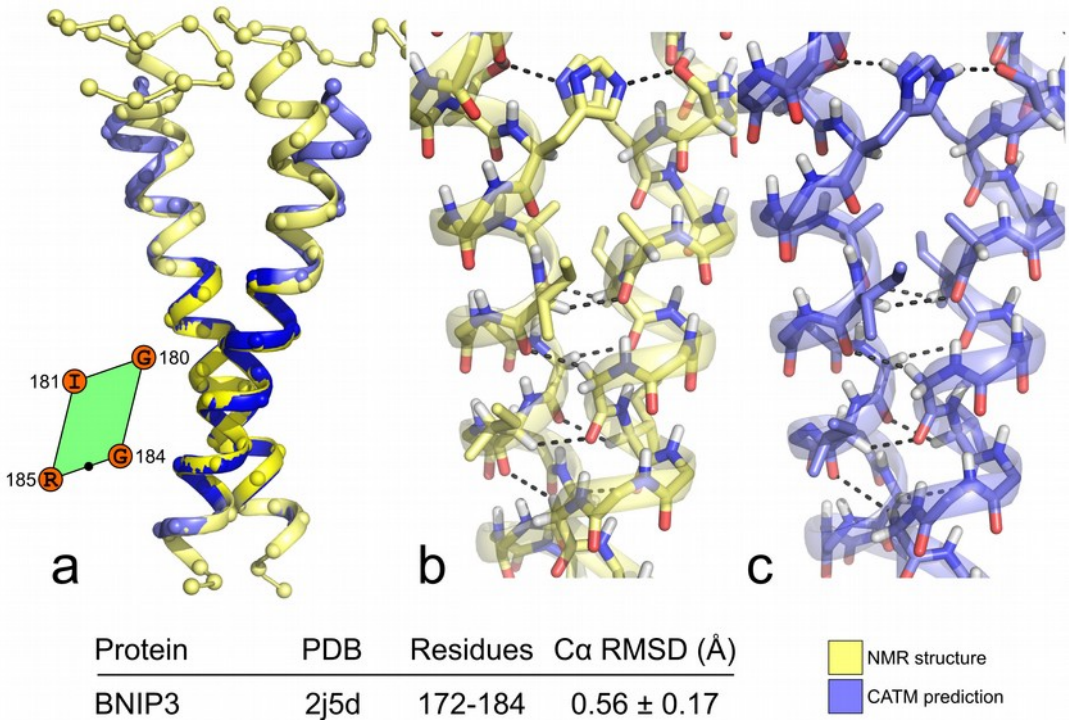
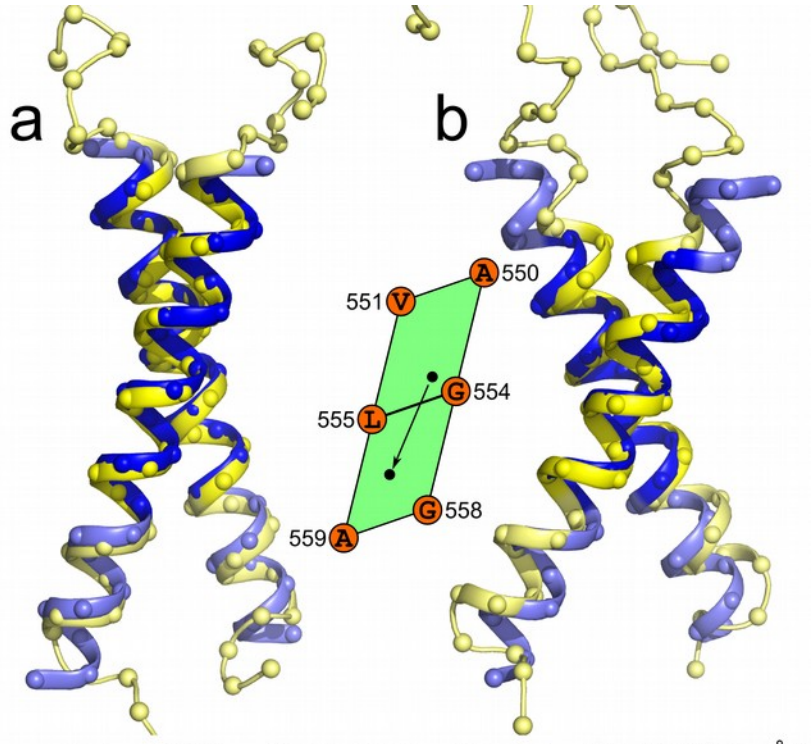
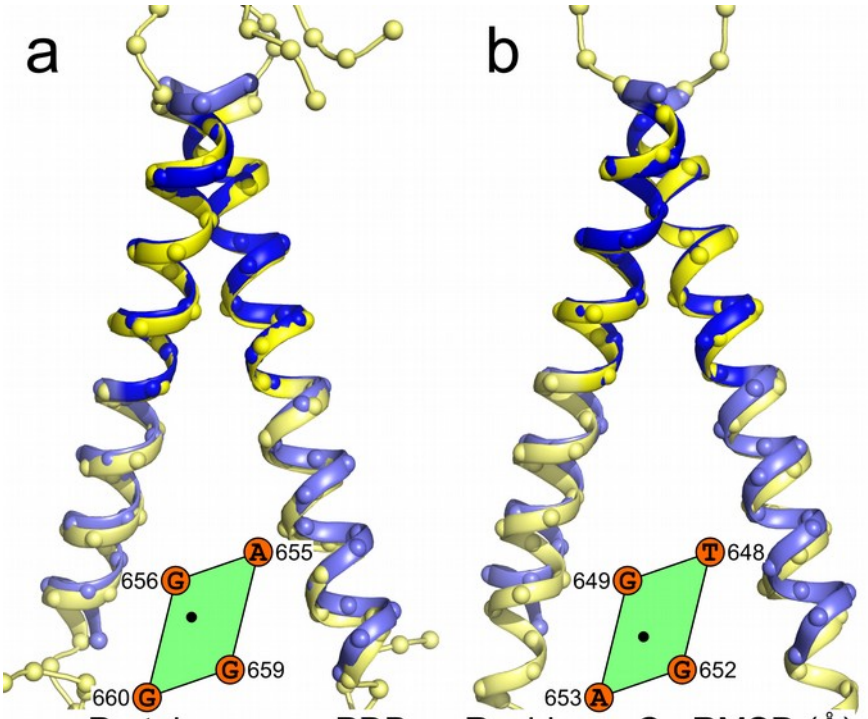


Fig. 2.6 Structural prediction of BNIP3. CATM produces a single model for BNIP3 that is extremely similar to the NMR structure. The Cα RMSD of the helical region of the entire TM domain is 1.10 ± 0.36 Å, which falls to 0.56 ± 0.17 Å when only the region in contact (darker blue and yellow) is considered. The side by side prediction (panels b and c) shows close similarity in the network of carbon hydrogen bonds and correct prediction of the orientation of all interfacial side chains. The model also accurately captures the canonical hydrogen bond between Ser 172 and His 173.



	pH	PDB	Residues	Model	Cα RMSD (Å)
a	4.3	2k1k	547-562	1	1.26 ± 0.01
b	6.3	2k1l	550-562	4	1.48 ± 0.01

Fig. 2.7 CATM predicts multiple states of the EphA1 Tyrosine Receptor Kinase. a) the structure of the TM domain EphA1 determined at a low pH is well predicted by CATM Model 1. b) the structure obtained at higher pH is matched by Model 4. The conformational shift between low and high pH is highlighted schematically in the unit cell representation. The interface remains centered on the Gly-zipper motif (AxxxGxxxG) but the crossing point shifts (arrow) toward the C-terminus in the adjacent unit cell. There is also an increase of the crossing angle. EphA1 has multiple GxxxG-like motifs and produces four models. Model 2 interacts through a C-terminal AxxxG motif. Model 3 is closely related to Model 1.



	Protein	PDB	Residues	Cα RMSD (Å)
a	ErbB4	2lcx	652-664	0.81 ± 0.05
b	ErbB1	2m20	645-667	0.78 ± 0.05

Fig. 2.8 Prediction of ErbB4 and ErbB1. a) ErbB4 is predicted by the top CATM model, while b) ErbB1 (EGFR), is predicted by the third model. Among the five structured tested, ErbB1 is the only structure that is not predicted by the lowest energy model.

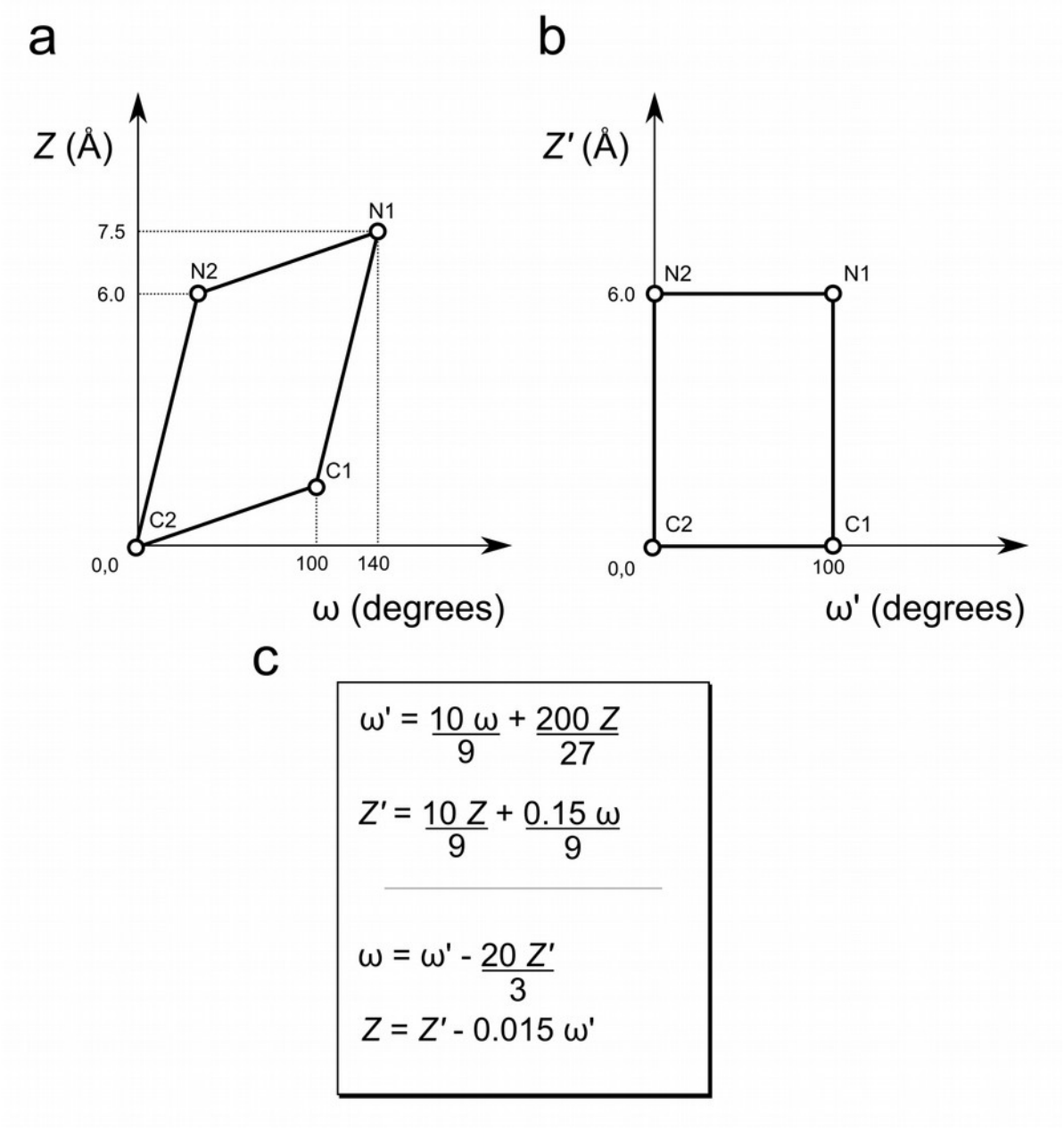


Fig. 2.9 Mathematical definition of *Z′* and ω' coordinates. a) Unit cell of the helical lattice as with C2 at the origin, as shown in Fig. 1. In the [ω, *Z*] set of coordinates C2 is at [0º, 0Å]; C1 is at [100º, 1.5Å]; N2 is at [40º, 6Å]; and N1 at [140º, 7.5Å]. b) The unit vectors ω' and *Z′* go in the direction of the principal components of the unit cell (C2-C1

and C2-N2, respectively). In the [ω', *Z'*] set of coordinates C2 is at [0º, 0Å]; C1 is at [100º, 0Å]; N2 is at [0º, 6Å]; and N1 at [100º, 6Å]. c) Mathematical equations for the transformation from one to the other set of coordinates.

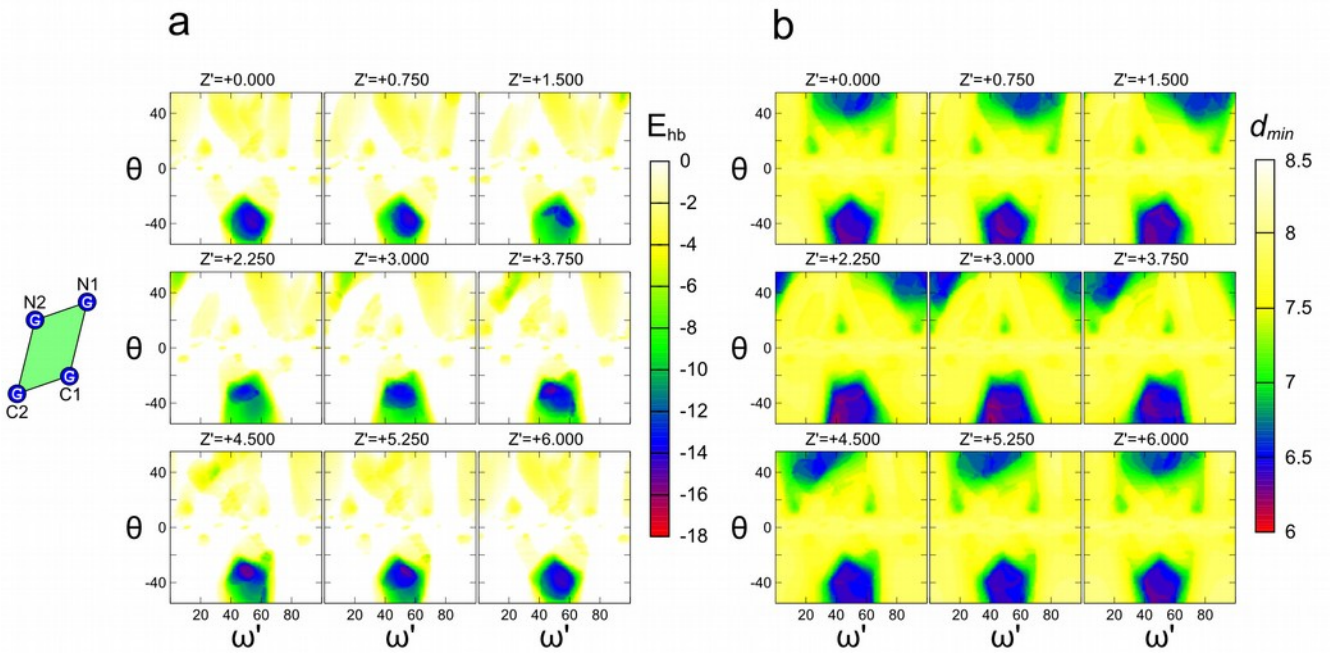


Fig. 2.10 Hydrogen bonding energies and d_{min} values of poly-Gly. a) Extended version of the hydrogen bonding energy maps for poly-Gly as a function of inter-helical geometry (ω': x-axis, θ: y-axis; *Z'*: panels) for poly-Gly, as in Fig. 2a. Only the minimum energy E_{min} across the *d* dimension is reported. b) Plot of the corresponding d_{min} distances at which the minimum energy was recorded. While the high-propensity region for Cα hydrogen bonding for right-handed structures display short d_{min} distances (<7.5Å), the plot demonstrates that other short distance conformations exist that do not lead to strong Cα hydrogen bond network formation.

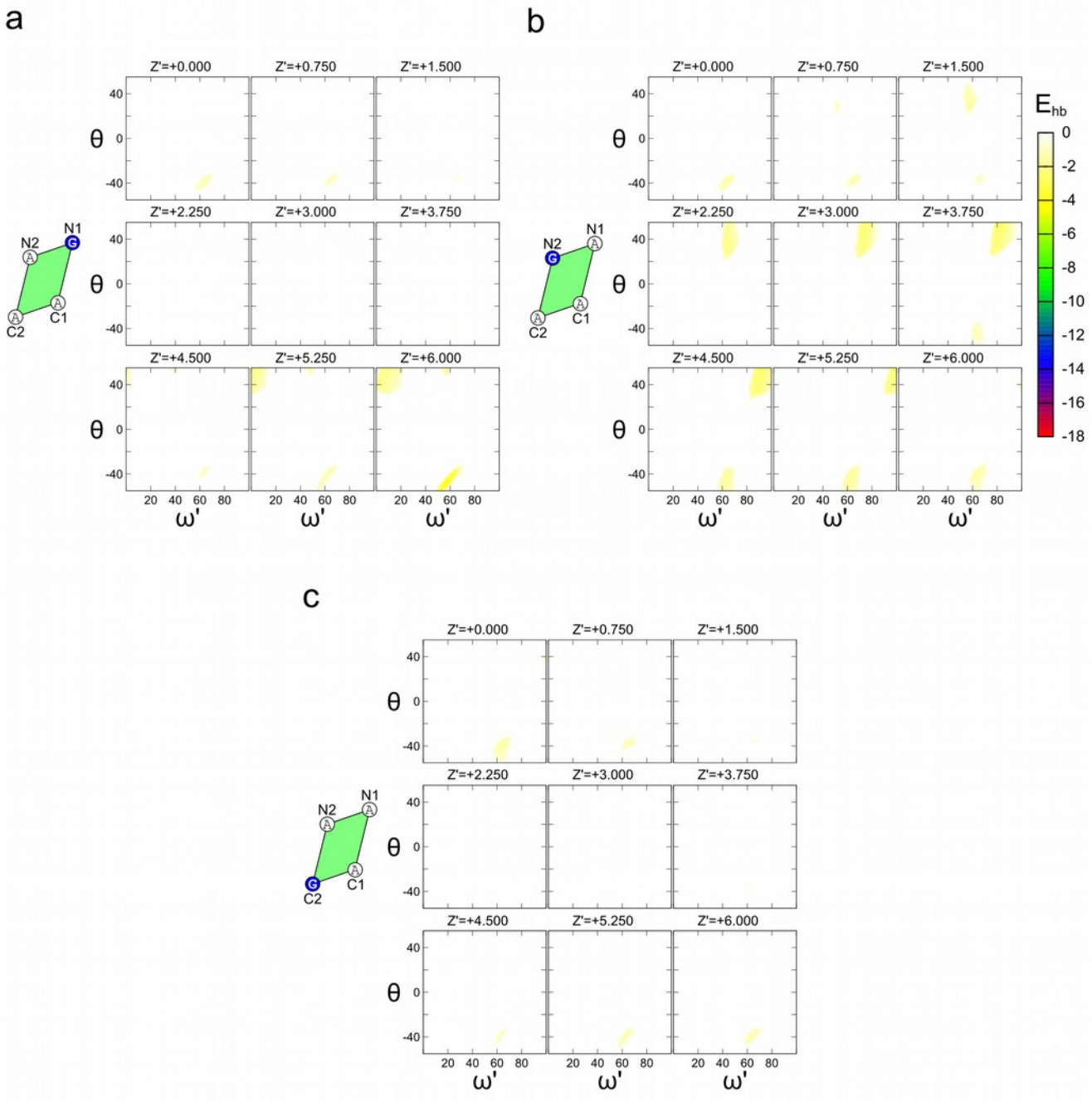


Fig. 2.11 Gly at N1, N2 and C2 in a poly-Ala background does not restore hydrogen bond propensity. Extended version of the hydrogen bonding energy maps for poly-Ala sequences with a single Gly at positions other than C1. a) Poly-Ala with no

Gly. b) Gly at N1 as in Fig. 2b. c) Gly at N2 as in Fig.2c. d) Gly at C2 (not shown in Fig. 2).

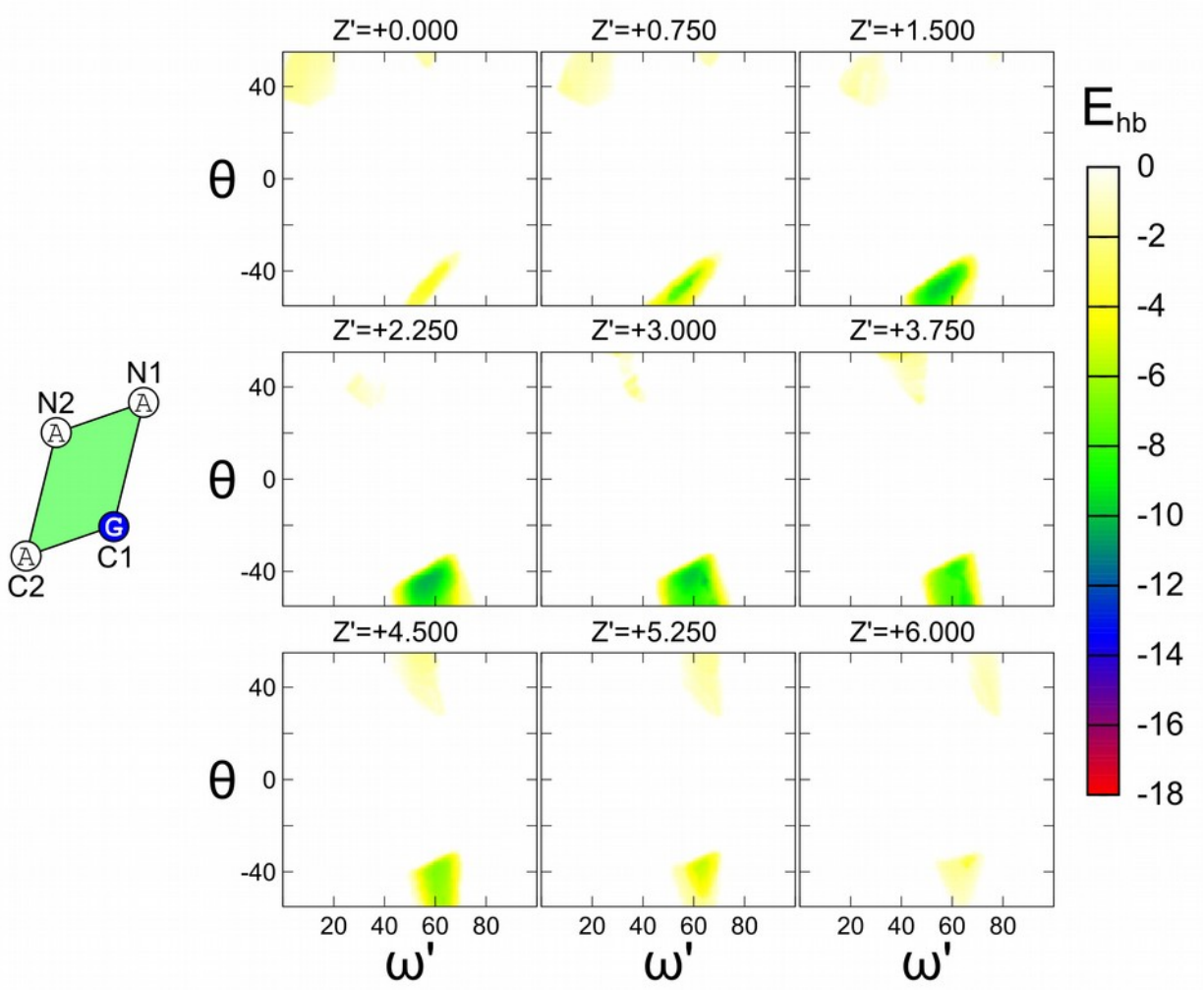


Fig. 2.12 Gly at C1 partially restores hydrogen bond propensity. Extended version of the hydrogen bonding energy maps for poly-Ala with a single Gly at positions C1 as in Fig. 2d.

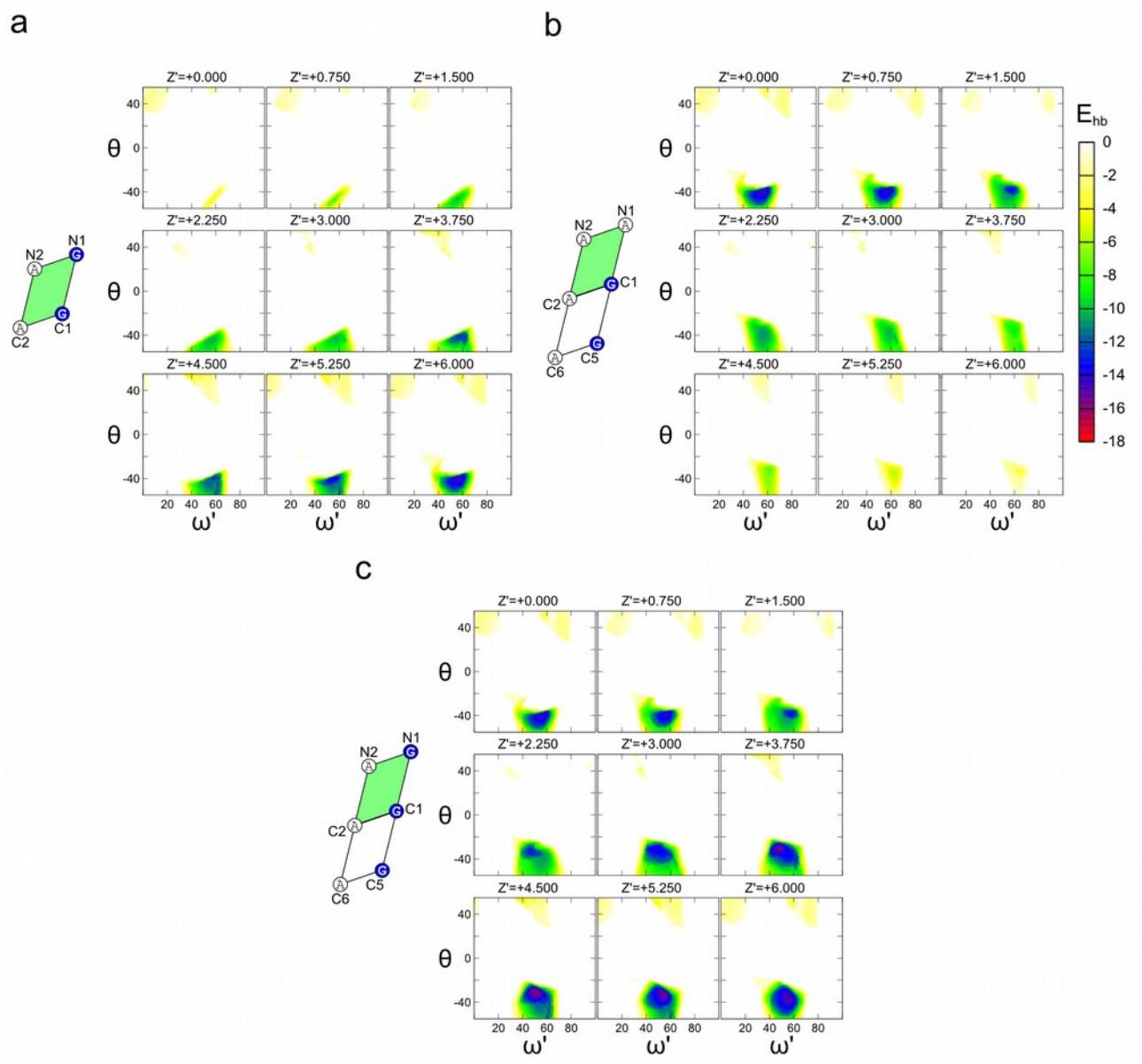


Fig. 2.13 Gly residues at N1 or C5 enhances hydrogen bonding in the presence of Gly at C1. Extended version of the hydrogen bonding energy maps for poly-Ala with two or three Gly residues on the right-hand side of the unit cell. a) Gly at N1 and C1 as in Fig. 2e. b) Gly at C1 and C5. c) Gly at N1, C1 and C5, as in Fig. 2f.

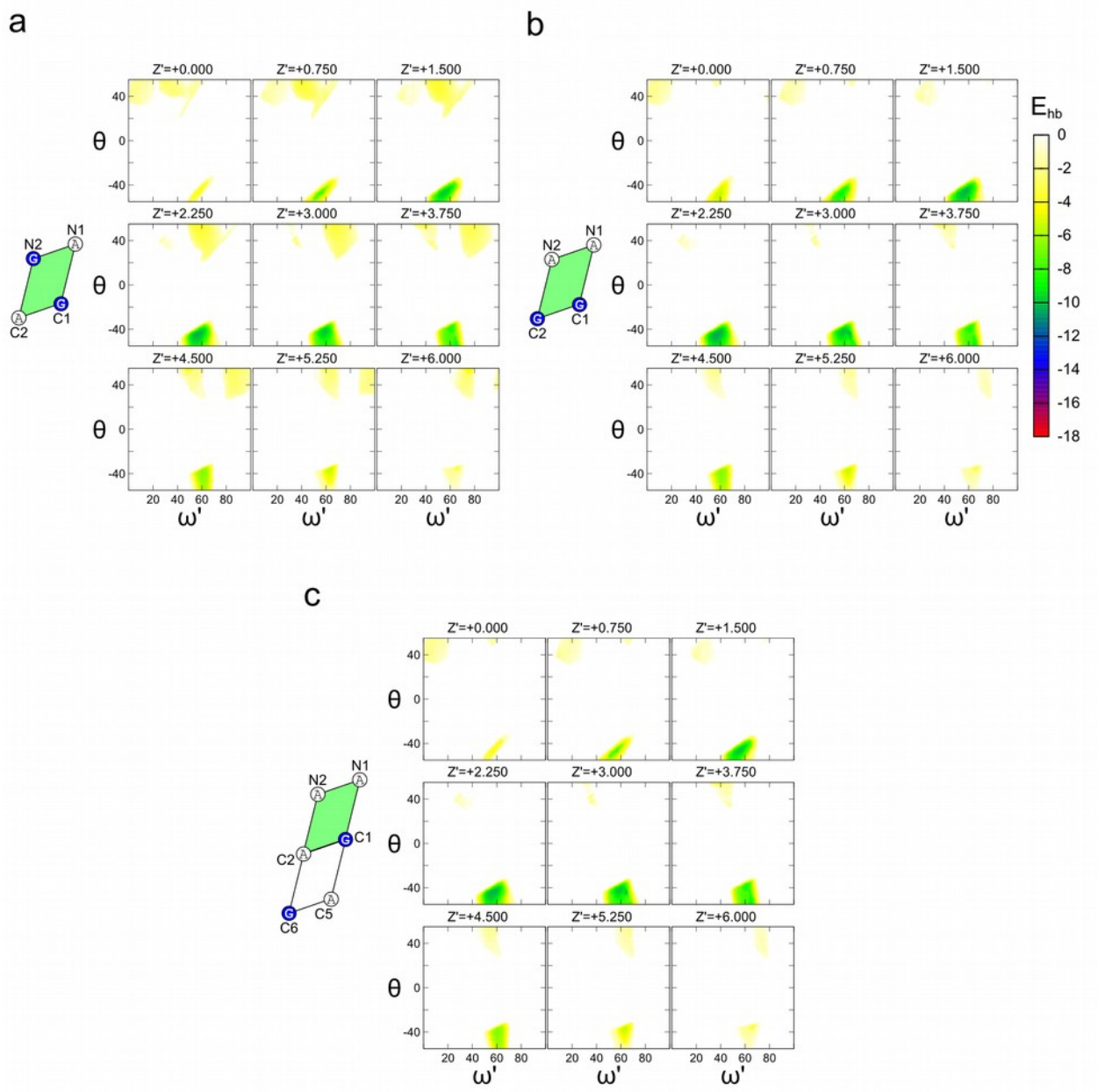


Fig. 2.14 A second Gly at N1, N2 or C6 does not restore hydrogen bond propensity. Hydrogen bonding energy maps for poly-Ala with two Gly residues a) Gly at N2/C1. b) Gly at C1/C2. c) Gly at C1/C6.

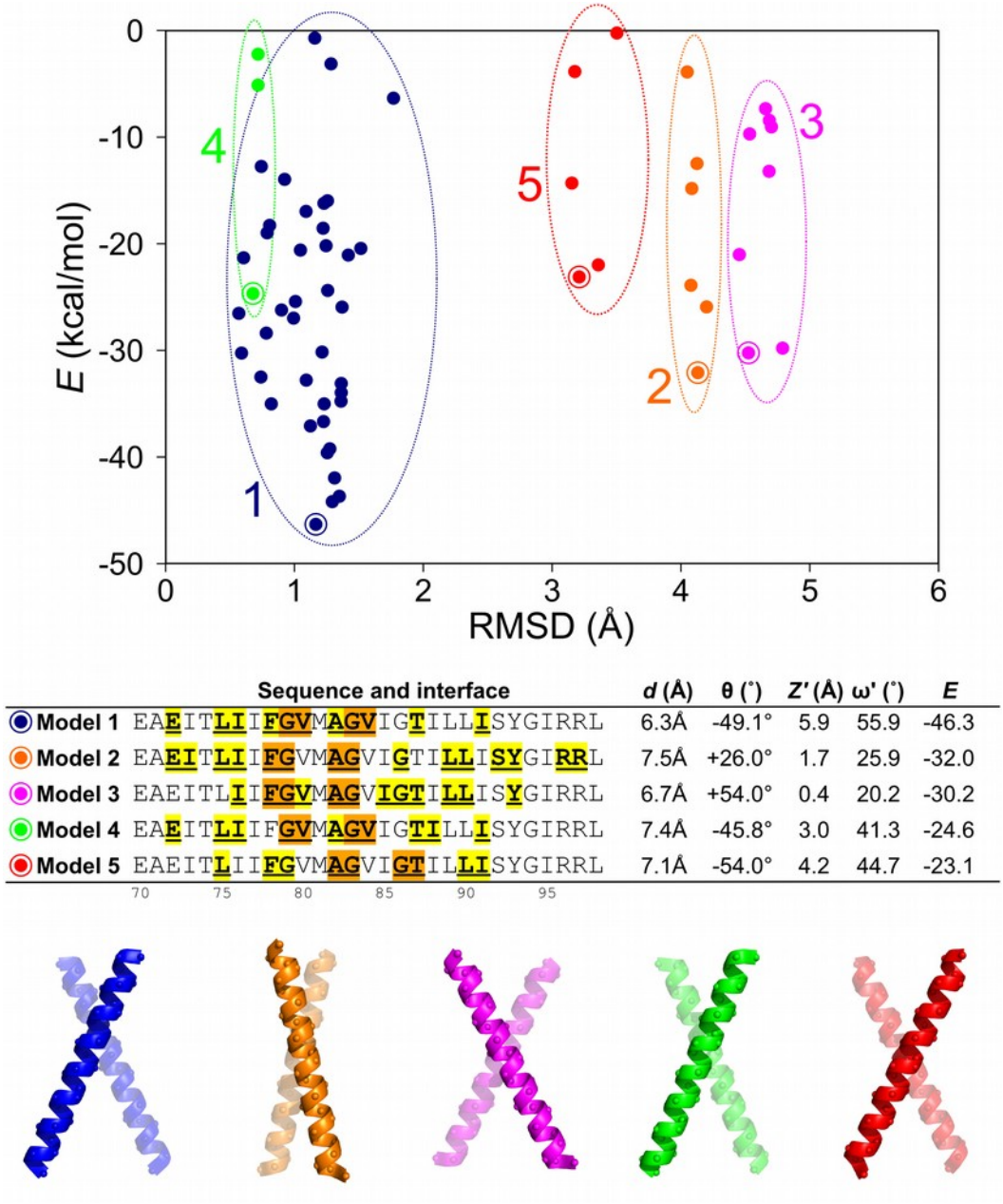


Fig. 2.15 RMSD from the NMR structure vs CATM energy for glycophorin A. CATM produces 63 structures for the transmembrane sequence of GpA, clustered into 5 representative models. The five clusters are color coded, and the lowest energy model highlighted by a circle. Model 1 and Model 4 are closely related neighboring clusters,

both right-handed dimers with a geometry similar to the experimental structure. As for all comparison, the RMSD were calculated in the range of amino acids that encompasses the dimer interfacial region (from L75 to T87) as in Fig. 5.

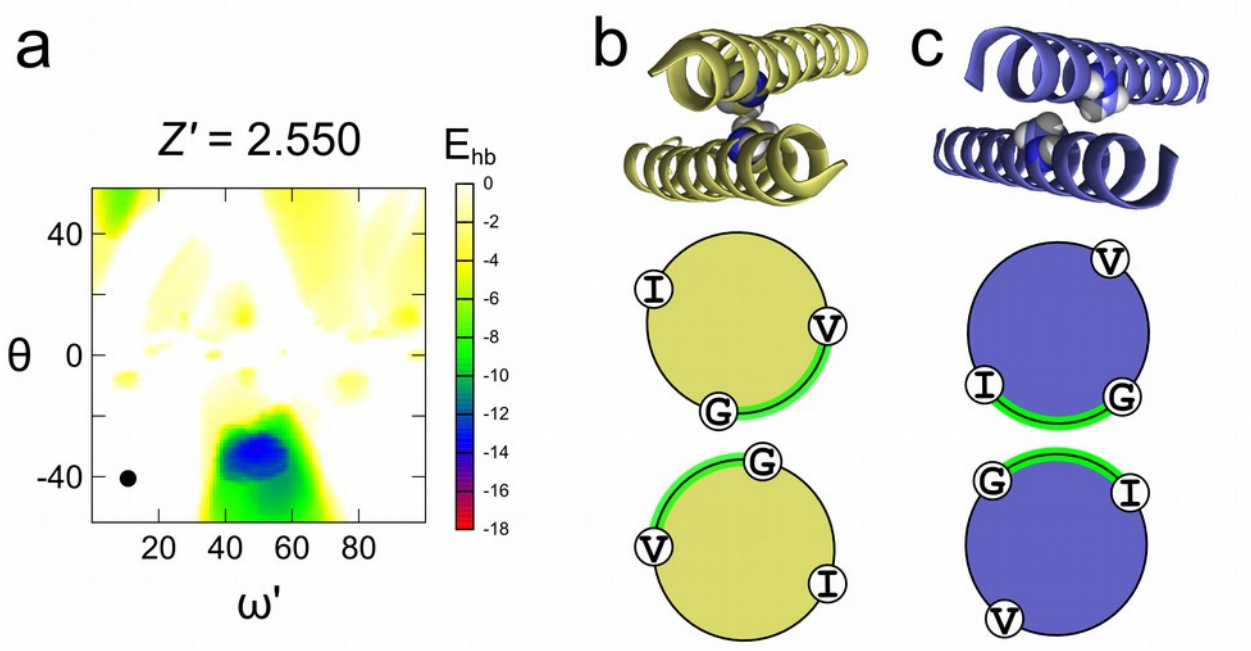


Fig. 2.16 Prediction of ErbB2 and comparison with the NMR structure. The NMR structure of ErbB2 has a crossing angle of -41° and an inter-helical distance of 7.6 Å, but it is not mediated by Cα hydrogen bonds. a) Its ω' angle (12°) falls in a region that is incompatible with Cα hydrogen bond formation (black dot). Therefore the structure is outside the scope of conformational space explored by CATM, and thus it cannot be predicted by the program. Comparison of the geometries of b) the NMR structure and c) the CATM model (Model 2) of ErbB2 shows a different orientation of the interface. The reference position Gly 660 is highlighted in both structures (above) and in the scheme (below).

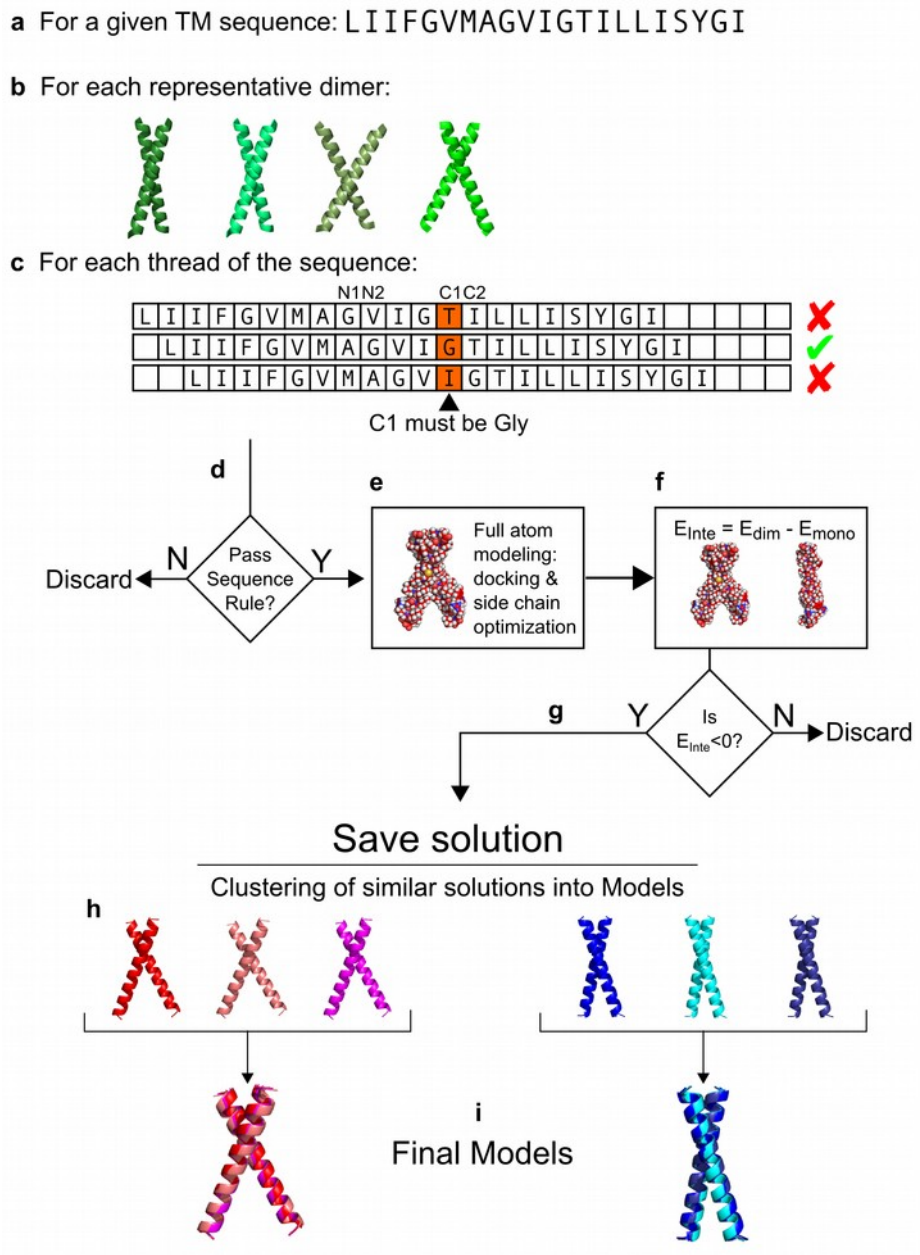


Fig. 2.17 Schematic illustration of CATM. Given a sequence (a), the sequence is threaded onto each of the 463 representative geometries (b) in all possible registries on the α-helices (c). For each thread the sequence rules are checked (d, in this example, we are only checking for a required Gly at C1). If the rules are met, the sequence is

built in all atoms and the structure is optimized (e), and an interaction energy is calculated (f). If the interaction energy is negative, the solution is accepted (g). The solutions are then clustered (h) to produce a series of final models, ranked by energy (i).

2.5 References

Arbely E, Arkin IT (2004) Experimental measurement of the strength of a Cα-H...O bond in a lipid bilayer. *J Am Chem Soc* 126:5362–5363.

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

Bocharov EV 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 EV et al. (2008) Spatial structure and pH-dependent conformational diversity of dimeric transmembrane domain of the receptor tyrosine kinase EphA1. *J Biol Chem* 283:29385–29395.

Bocharov EV et al. (2008) 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 EV, Mineev KS, Goncharuk MV, Arseniev AS (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.

Endres NF et al. (2013) Conformational coupling across the plasma membrane in activation of the EGF receptor. *Cell* 152:543–556.

Fleishman SJ, Schlessinger J, Ben-Tal N (2003) A putative molecular-activation switch in the transmembrane domain of erbB2. *Proc Natl Acad Sci USA* 99:15937–15940.

Horowitz S, Trievel RC (2012) Carbon-oxygen hydrogen bonding in biological structure and function. *J Biol Chem* 287:41576–41582.

Kim S et al. (2005) Transmembrane glycine zippers: physiological and pathological roles in membrane proteins. *Proc Natl Acad Sci USA* 102:14278–14283.

Krivov GG, Shapovalov MV, Dunbrack RL (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 77:778–795.

Kulp DW et al. (2012) Structural informatics, modeling, and design with an open-source Molecular Software Library (MSL). *J Comput Chem* 33:1645–1661.

Lawrie CM, Sulistijo ES, MacKenzie KR (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.

Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI (2006) OPM: orientations of proteins in membranes database. *Bioinformatics* 22:623–625.

MacKenzie KR, Fleming KG (2008) Association energetics of membrane spanning α-helices. *Curr Opin Struct Biol* 18:412–419.

MacKenzie KR, Prestegard JH, Engelman DM (1997) A transmembrane helix dimer: structure and implications. *Science* 276:131–133.

MacKerell et al. (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins†. *The Journal of Physical Chemistry B* 102:3586–3616.

Moore DT, Berger BW, DeGrado WF (2008) Protein-Protein Interactions in the Membrane: Sequence, Structural, and Biological Motifs. *Structure* 16:991–1001.

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

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

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

Schlessinger J (2000) Cell signaling by receptor tyrosine kinases. *Cell* 103:211–225.

Senes A (2011) Computational design of membrane proteins. *Curr Opin Struct Biol* 21:460–466.

Senes A et al. (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.

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

Senes A, Gerstein M, Engelman DM (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* 296:921–36.

Senes A, Ubarretxena-Belandia I, Engelman DM (2001) The Cα ---H...O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc Natl Acad Sci U S A* 98:9056–61.

Smith SO, Jonas R, Braiman M, Bormann BJ (1994) Structure and orientation of the transmembrane domain of glycophorin A in lipid bilayers. *Biochemistry* 33:6334–6341.

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

Sulistijo ES, MacKenzie KR (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 ES, Mackenzie KR (2009) Structural basis for dimerization of the BNIP3 transmembrane domain. *Biochemistry* 48:5106–5120.

Vargas R, Garza J, Dixon DA, Hay BP (2000) How Strong Is the Cα−H···OC Hydrogen Bond? *J Am Chem Soc* 122:4750–4755.

Walters RFS, DeGrado WF (2006) Helix-packing motifs in membrane proteins. *Proc Natl Acad Sci U S A* 103:13658–63.

Xiang Z, Honig B (2001) Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol* 311:421–430.

Yohannan S et al. (2004) A Cα-H...O hydrogen bond in a membrane protein is not stabilizing. *J Am Chem Soc* 126:2284–2285.

Chapter 3

A Gly-zipper motif mediates homo-dimerization of the transmembrane domain of the mitochondrial kinase ADCK3

This chapter was prepared for publication as:
Ambalika S. Khadria, Benjamin K. Mueller, Jonathan A. Stefely, Chin Huat Tan, David J. Pagliarini and Alessandro Senes "A Gly-zipper motif mediates homo-dimerization of the transmembrane domain of the mitochondrial kinase ADCK3" *J Am Chem Soc* 2014 136, 14068-77

*I was the primary computational author

Abstract

Interactions between α-helices within the hydrophobic environment of lipid bilayers are integral to the folding and function of transmembrane proteins; however, the major forces that mediate these interactions remain debated, and our ability to predict these interactions is still largely untested. We recently demonstrated that the frequent transmembrane association motif GAS_{right} – the GxxxG-containing fold of the glycophorin A dimer – is optimal for the formation of extended networks of Cα–H hydrogen bonds, supporting the hypothesis that these bonds are major contributors to association. We also found that optimization of Cα–H hydrogen bonding and inter-helical packing is sufficient to computationally predict the structure of known GAS_{right} dimers at near atomic level. Here, we demonstrate that this computational method can be used to characterize the structure of a protein not previously known to dimerize, by predicting and validating the transmembrane dimer of ADCK3, a mitochondrial kinase. ADCK3 is involved in the biosynthesis of the redox active lipid, ubiquinone, and human ADCK3 mutations cause a cerebellar ataxia associated with ubiquinone deficiency, but the biochemical functions of ADCK3 remain largely undefined. Our experimental analyses show that the transmembrane helix of ADCK3 dimerizes, with an interface based on an extended Gly-zipper motif, as predicted by our models. The data provide strong evidence for the hypothesis that optimization of Cα–H hydrogen bonding is an important factor in the association of transmembrane helices. This work also provides a structural foundation for investigating the role of transmembrane association in regulating the biological activity of ADCK3.

3.1 Introduction

A fundamental event in the folding and oligomerization of membrane proteins is the association of the transmembrane (TM) helices [Popot, et al. (1990); Engelman, et al. (2003)]. After the TM helices have been inserted in the membrane, helix-helix association is required to achieve the final fold and oligomeric state of the protein. A favorite system for investigating the rules that govern TM helix association are the single-span membrane proteins [Rath,et al. (2012); Senes, et al. (2004); Mackenzie (2006); MacKenzie, et al. (2008); Moore, et al. (2008)], primarily because a variety of methods are available for measuring their oligomerization (including FRET [Fisher, et al. (1999); Merzlyakov, et al. (2006); You, et al. (2005); Khadria, et al. (2013)], sedimentation equilibrium analytical ultra-centrifugation [Fleming (2002); Choma, et al. (2000)], *in vivo* assays in biological membranes [Russ, et al. (1999); Lindner, et al. (2006); Schneider, et al. (2003)], SDS-PAGE [Lemmon, et al. (1992); Rath, et al. (2009)], and steric trapping [Hong, et al. (2010); Hong, et al. (2013)]). Conversely, assessing the folding energetics of multi-span membrane proteins still represents a tremendous challenge [Bowie (2005), Chang, et al. (2014)].

In addition to being a tractable system, the single-span membrane proteins attract interest because of their biological importance. These proteins comprise the most numerous class of membrane proteins, constituting about half of the total [Wallin, et al. (1998); Arkin, et al. (1998); Hubert, et al. (2010)]. Rather than acting as mere membrane anchors for soluble domains, as it was once assumed, the oligomerization of the single TM domains actively plays roles in assembly, signal transduction, ion conduction and regulation in a wide variety of biological processes [Moore, et al. (2008)].

To investigate the basis of oligomerization in TM helices, our group and others have pursued a strategy based on the analysis of frequently occurring association motifs [Cunningham, et al. (2011); Walters, et al. (2006); Russ, et al. (2000); Senes, et al. (2000); Kim, et al. (2005); Doura, et al. (2004); Unterreitmeier, et al. (2007)]. One of

the most important motifs is GAS_{right} [Walters, et al. (2006)] (Fig. 3.1), which is best known as the fold of a widely studied model system for TM association, the glycophorin A TM dimer [MacKenzie, et al. (1997)]. GAS_{right} gets its name from its right-handed crossing angle (Fig. 3.1b), and from the characteristic small amino acids at its interface (GAS: Gly, Ala, Ser) [Walters, et al. (2006)], which are arranged to form GxxxG and GxxxG-like patterns (GxxxA, AxxxG, etc) [Russ, et al. (2000), Senes, et al. (2000)]. In many ways GAS_{right} parallels the important coiled coil, a frequently occurring interaction motif and model for folding and association for soluble proteins [Harbury, et al. (1993); Harbury, et al. (1995); Betz, et al. (1995)]. Like the coiled coil, GAS_{right} is characterized by a specific geometry (a short inter-helical distance and a crossing angle near −40°), it has a distinctive sequence signature (the GxxxG patterns), and is one of the most common – if not the most common – oligomerization motif [Walters, et al. (2006)].

In a recent computational analysis of transmembrane dimer geometry, we proposed that the primary role of GxxxG in GAS_{right} is to promote the formation of networks of stabilizing hydrogen bonds between Cα–H donors and carbonyl oxygen acceptors on opposed helices [Mueller, et al. (2014)] (Fig. 3.1b,c). More specifically, we proposed that the small amino acids perform two distinct functions: the first is to create permissive steric conditions, allowing the two helices to come in backbone contact, thus bringing the Cα–H donors and carbonyl acceptors in proximity. The second function, which is performed exclusively by Gly, is to increase the number of hydrogen bonds by donating with the second Hα, which corresponds to the side chain R-group in all other amino acids. To perform these functions, the small amino acids are required to be present at specific positions. The formation of this network of hydrogen bonds is also dependent on the specific crossing angle of GAS_{right} (−40°), which precisely aligns Cα–H donors spaced at *i*, *i*+1 on one helix against carbonyl acceptors spaced at *i*, *i*+3 on the opposing helix (see Fig. 4 in Mueller *et al*. [Mueller, et al. (2014)[j]). Overall, GAS_{right} appears geometrically optimized for inter-helical Cα–H formation [Mueller, et al. (2014)].

Cα–H hydrogen bonds are commonly observed in proteins [Horowitz, et al. (2012)]. Carbons are generally weak donors, but the Cα in proteins is activated by the

electron-withdrawing amide groups on both sides, and quantum calculations indicate that the energy of Cα–H hydrogen bonds may be as much as one third to half of that of canonical donors in vacuum [Vargas, et al. (2000); Scheiner, et al. (2001)]. 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. 3.1)[Senes, et al. (2001)]. An IR-based investigation of the CD_2 stretching mode of a Cα–H donor in the transmembrane domain of glycophorin A produced an estimated contribution of −0.88 kcal/mol for the hydrogen bond [Arbely, et al. (2004)]. Conversely, a folding study of the multi-span membrane protein bacteriorhodopsin in which a Cα–H···O side chain hydroxyl acceptor (Thr-24) was mutated indicated that this particular bond was not stabilizing [Yohannan, et al. (2004)]. Subsequent computational work suggested that the orientation of the groups can determine whether an interaction may be strongly favorable or unfavorable [Mottamal, et al. (2005); Park, et al. (2008)].

The exact contribution of hydrogen bonds – whether the donor is a Cα–H or a more “canonical” N–H or O–H group – to membrane protein folding and association is still unresolved [Bowie (2011)]. A governing assumption maintains that donors and acceptors buried in the membrane would not pay a significant desolvation penalty upon helix association, and therefore the formation of hydrogen bonds should contribute appreciably to the stability of membrane proteins. Indeed, polar residues can promote interaction of transmembrane helices [Choma, et al. (2000); Zhou, et al. (2000); Herrmann, et al. (2010)]. Yet, the limited number of experimental observations made to date seem to indicate that the contribution of hydrogen bonding in the membrane may be – surprisingly – of the same magnitude observed for water soluble proteins [Bowie (2011)].

Despite the scarce experimental evidence regarding the contribution of Cα–H hydrogen bonds to TM interactions, the hypothesis that they drive folding and oligomerization remains compelling. In particular, the fact that the prevalent GAS_{right} motif corresponds to the only inter-helical geometry that maximizes formation of Cα–

H···O=C networks, strongly suggests that these bonds are indeed a major contributor to association[Mueller, et al. (2014)]. Under these premises, we hypothesized that a computational structural search based on the simultaneous optimization of side chain packing and Cα–H hydrogen bonding may be able to predict the structure of GAS_{right} dimers. The resulting program, named CATM, was tested against the small database of known GAS_{right} homo-dimeric structures [Mueller, et al. (2014)]. We found that CATM predicts these known structures at near atomic precision. The finding provides further indirect support that Cα–H hydrogen bonding is likely to be a structural determinant of GAS_{right} dimers [Mueller, et al. (2014)]. The positive result also indicates that CATM may be a powerful tool for assisting the experimental investigation of GAS_{right} homo-dimers of unknown structure.

To test the ability of our methods to predict *ab initio* the structure of unknown GAS_{right} dimers, here we investigate ADCK3, a human mitochondrial protein that is a member of the highly conserved UbiB protein kinase-like family [Leonard, et al. (1998)]. UbiB family members account for approximately one-quarter of microbial PKL sequences [Kannan, et al. (2007)], are ubiquitous among eukaryotes [Kannan, et al. (2007)], and are strongly associated with lipid metabolism [Tan, et al. (2013); Martinis, et al. (2013); Lundquist, et al. (2013)]. Most organisms have a <u>Ubi</u>B family member that is required for the biosynthesis of coenzyme Q (CoQ, <u>ubi</u>quinone). Deletion of the *E. coli* gene *ubiB* [Poon, et al. (2000)] or the yeast gene *coq8* [Do, et al. (2001)] completely halts CoQ biosynthesis. Similarly, mutations to human ADCK3 are known to cause CoQ deficiency and cerebellar ataxia [Lagier-Tourenne, et al. (2008); Mollet, et al. (2008); Gerards, et al. (2010); Horvath, et al. (2012)], and mutations to human ADCK4 were recently shown to cause CoQ deficiency and a steroid-resistant nephrotic syndrome [Ashraf, et al. (2013)]. Our knowledge of the molecular mechanism by which UbiB family proteins enable CoQ biosynthesis is limited, primarily because the endogenous substrates of UbiB proteins have not yet been discovered. However, we do know that coq8p in yeast somehow stabilizes a complex of CoQ biosynthesis enzymes [He, et al. (2006)]. CoQ biosynthesis occurs within the context of cellular membranes — either the

plasma membrane of prokaryotes or the inner mitochondrial membrane of eukaryotes — and the responsible enzymes are either integral membrane proteins or peripherally associated membrane proteins [He, et al. (2006)]. ADCK3, which contains a predicted TM domain, is also likely to associate with membranes, but this hypothesis has not yet been tested. Biochemical characterization of the ADCK3 TM domain would provide an important foundation for understanding how it enables CoQ biosynthesis.

The potential functional importance of the ADCK3 TM region is underlined by the existence of a mutation at the putative edge of the TM domain (R213W) that disrupts CoQ biosynthesis and causes cerebellar ataxia in human patients [Mollet, et al. (2008)]. Furthermore, dimerization of single-span TM domains is known to be central to the regulation of some kinase families, such as the receptor tyrosine kinases [Lemmon, et al. (2014); Li, et al. (2006)]. However, it was unknown whether the predicted TM helix of ADCK3 could actually insert into biological membranes and whether the TM helix can self-associate to potentially drive dimerization of ADCK3.

Here, we demonstrate experimentally that the TM domain of ADCK3 inserts into membranes and self-associates. Using extensive mutagenesis, we also show that the interaction interface is consistent with the structural models predicted by CATM, which involves an extended Gly-zipper motif [Kim, et al. (2005)] (i.e. a series of Gly amino acids separated at *i*, *i*+4). The experimental and computational data also indicates that the Gly-zipper interface is potentially compatible with alternative conformations of the TM domain, opening the possibility that conformational changes of the TM dimer may be important for ADCK3 function.

3.2 Methods

3.2.1 Vectors and strains

All oligonucleotides were purchased in desalted form from Integrated DNA Technologies and used without purification. The expression vectors pccKAN, pccGpA-wt, and pccGpA-G83I, and *malE* deficient *Escherichia coli* strain MM39 were kindly provided by Dr. Donald M. Engelman [Russ, et al. (1999)]. The genes encoding the TM domain of ADCK3 (214-LANFGGLAVGLGFGALA-230) and ADCK4 (92-LANFGGLAVGLGLGVLA-108) were cloned into the NheI-BamHI restriction sites of the pccKAN vector. Site directed mutations to produce single amino acid variants in the TM domain of ADCK3 were introduced with the QuikChange kit (Stratagene).

3.2.2 Expression of Chimeric Proteins in MM39 cells

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. 30 µL of overnight cultures were inoculated into 3 mL of LB broth and grown to an OD_{420} of approximately 0.8-1.1 (OD_{600} of 0.4 to 0.6) at 37 ˚C. After recording the optical density, 1 mL of cells was spun down for 10 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 8 seconds 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.

3.2.3 MalE Complementation 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 hours [Russ, et al. (1999)].

3.2.4 Chloramphenicol Acetyltransferase (CAT) spectrophotometric assay

CAT activity was measured as described [Shaw (1975); LaPointe, et al. (2013)]. Briefly, 1 mL of buffer containing 0.1 mM acetyl CoA, 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 40 µL of cleared cell lysates and the absorbance at 412 nm was measured for two minutes 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 two minutes to determine CAT activity. The basal CAT activity was subtracted and the value was normalized by the cell density measured as OD_{420}. All measurements were determined at least in duplicate and the experiments were repeated at least twice.

3.2.5 Quantification of expression by immunoblotting

Protein expression was confirmed by immunoblotting. The cell lysates (10 µL) were loaded onto a NuPAGE 4-12% Bis-Tris SDS-PAGE gel (Invitrogen) and then transferred to PVDF 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 two hours at 4 ˚C, incubated with biotinylated anti-Maltose Binding Protein antibodies (Vector labs) overnight at 4 ˚C, followed by peroxidase-conjugated streptavidin (Jackson ImmunoResearch) for two hours at 4 ˚C. Blots were developed with the Pierce ECL Western Blotting Substrate Kit and chemiluminescence was measured using an ImageQuant LAS 4000 (GE Healthsciences).

3.2.6 Computational modeling

The structure of ADCK3-TM was predicted with CATM [Mueller, et al. (2014)], which is distributed with the open source MSL C++ library v. 1.2 [Kulp, et al. (2012)] at http://msl-libraries.org. The computational mutagenesis was performed on all ADCK3 models by applying the same point mutations measured experimentally in the context of a fixed backbone, followed by side chain optimization. Side chain mobility was modeled using the Energy-Based conformer library applied at the 95% level [Subramaniam, et al. (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; D_0=2.3 Å; σ_d=1.202 Å; α_{max}=74.0°;

β_{max}=98.0° [Mueller, et al. (2014)]. The relative energy of each mutant was calculated as

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

where $E_{WT,dimer}$ and $E_{mut,dimer}$ are the energies of the wild type and mutant sequence, respectively, in the dimeric state, and $E_{WT,monomer}$ and $E_{mut,monomer}$ are the energies of the wild type and mutant sequence, respectively, in a side chain optimized monomeric state with the same sequence. As reported previously [LaPointe, et al. (2013)], the effect of each mutation was classified in four categories (analogous to the experimental mutagenesis) using the following criterion: category 0, “WT-like”, $\Delta E_{mut} < 2$ kcal/mol; category 1, “Mild”, $2 \leq \Delta E_{mut} < 4$; category 2, “Severe”, $4 \leq \Delta E_{mut} < 8$; category 3, “Disruptive”, $\Delta E_{mut} \geq 8$. The numerical category values were averaged to calculate the average position-dependent disruption value.

3.3 Results and Discussion

3.3.1 ADCK3 is predicted to have a TM helix

The protein kinase-like domain of ADCK3 is preceded on the N-terminal side by a region of undefined function. A predicted TM helix within this region is annotated in UniProt for the close homolog ADCK4 [UniProt Consortium (2013)], providing a potential anchor for the protein at the inner mitochondrial membrane (Fig 3.2a). UniProt does not report a predicted TM domain for the corresponding region of ADCK3, but the sequence of the putative TM segment is highly conserved between the two proteins. The same general domain organization and function is also predicted for the yeast homolog Coq8p. Given that ADCK3 and ADCK4 are localized to the mitochondrial matrix [Rhee, et al. (2013)], the TM domain would position their catalytic kinase domains on the matrix face of the inner membrane, the same localization of the enzymes involved in the biosynthesis of coenzyme Q [He, et al. (2006); Rhee, et al. (2013)]. Therefore, it is important to verify the TM domain experimentally and to investigate its potential functional role.

The sequences of the putative TM domains of ADCK3 and ADCK4 are aligned in Fig. 3.2b. As summarized in Table 3.1, these sequences have low hydrophobicity and a relative short length (17 amino acids), and thus are not well recognized by prediction servers. The TM domain of ADCK3, which contains one polar amino acid (Asn 216), is recognized as a borderline TM sequence by most servers. Specifically, the segment is not recognized by *TMHMM* [Krogh, et al. (2001)] and *E(z)* [Senes, et al. (2007)] but the segment is predicted as transmembrane by *MemBrain* [Shen, et al. (2008); Yang, et al. (2013)] and *HMMTOP* [Tusnády, et al. (1998); Tusnády, et al. (2001)], and *Phobius* [Käll, et al. (2004); Käll, et al. (2007)] and *ΔG prediction* [Hessa, et al. (2007)] recognize it with low confidence. ADCK4 shares over 50% sequence identify with ADCK3 but their TM domains are almost identical, differing only at two positions (Fig. 3.2b). Because of these two substitutions (and primarily because of the A228V substitution), the hydrophobicity of the TM domain of ADCK4 is higher (as calculated with either the Wimley-White octanol scale [Wimley, et al. (1996)] or the “biological” scale [Hessa, et al.

(2007); Hessa, et al. (2005)]) and is sufficient to be predicted by most servers, except *E(z)*, with good confidence (Table 3.1).

In interpreting these prediction data, it is important to consider that TM prediction servers are trained against a majority of proteins that are inserted in the membrane via a translocon mediated mechanism. The sequence requirements for translocon mediated insertion in an eukaryotic system are well understood [Hessa, et al. (2007); Hessa, et al. (2005)]. A recent analysis in a bacterial system shows good overall correspondence to the mammalian system, but the hydrophobicity threshold appears to be distinctly lower [Ojemalm, et al. (2013)]. Much less is known about the requirements for membrane insertion of mitochondrial integral membrane proteins that are encoded in the nucleus, such as ADCK3. There is, however, good indication that the hydrophobicity threshold for these proteins should be even lower, to avoid mistargeting of these proteins to the endoplasmic reticulum and to facilitate their translocation to the mitochondrion [Tong, et al. (2011); Supekova, et al. (2010); Horie, et al. (2002); Daley, et al. (2002); Horie, et al. (2003); Daley, et al. (2005)]. Based on the above considerations, it is highly probable that the predicted TM segments of ADCK3 and ADCK4 are indeed *bona fide* TM domains.

3.3.2 The TM domain of ADCK3 has conserved GxxxG-like motifs

As shown in Fig. 3.2b, the predicted TM regions of ADCK3 and ADCK4 are very rich in small amino acids such as Gly, Ala and Ser (9 in each). The sequences contain a number of GxxxG and GxxxG-like (AxxxG) helix association patterns, which appear to be evolutionarily conserved (Fig. 3.2c). In particular, they contain an extended Gly-zipper motif [Kim, et al. (2005)], i.e. a series of small amino acids (215-AxxxGxxxGxxxG-227) spaced at *i*, *i*+4, highlighted in red in Fig. 3.2b. They also contain an additional AxxxG motif (magenta), which is off-frame by two positions with respect to the Gly-zipper. This spacing projects the two motifs on opposite helical faces.

3.3.3 CATM predicts that the TM domain of ADCK3 can form a GAS_{right} homo-dimer

GxxxG-like patterns can drive helix-helix association [Russ, et al. (2000)]. They occur with high frequency in TM helices [Senes, et al. (2000)], both in multi-span

proteins and in oligomerizing single-span membrane proteins [Senes, et al. (2001)], and are often important for biological function [Senes, et al. (2004)]. The presence of GxxxG-like motifs in the putative TM sequence of ADCK3 raised the question of whether this domain oligomerizes. To investigate this question, we analyzed the sequence with CATM [Mueller, et al. (2014)], a program for the structural prediction of GAS_{right} motifs, an important and common class of GxxxG-mediated dimers [Walters, et al. (2006)].

As shown in Fig. 3.3, CATM predicts five alternative models for the TM sequence of ADCK3. The figure schematically depicts the geometrical features of the dimers. The position of the crossing point between the two helices is marked (dot), and the interfacial positions that surround this crossing point are highlighted by a green parallelogram. All the positions that are involved in inter-monomer contacts at the interface are highlighted in either yellow, or in red if they belong to the Gly-zipper motif. The scores of the top models of ADCK3 in CATM (−59.8, −50.8 and −47.7 for Models 1, 2 and 3 respectively) are comparable to the scores obtained for the five known structures of GAS_{right} motifs (which range between −56 and −38) , which CATM is able to predict at near atomic precision [Mueller, et al. (2014)].

Notably, the extended Gly-zipper is involved at the helix-helix interface in all models (Fig. 3.3). Model 1 and 2 are related geometries whose crossing points fall in the quadrilateral defined by Gly 219, Leu 220, Gly 223 and Leu 224 (Axxx**GL**xx**GL**xxG). These two models differ by the position of the crossing point and, most importantly, by their crossing angle, which is near the canonical −40° of GAS_{right} motifs for Model 2, and narrower for Model 1 (−27.1°). The smaller crossing angle causes Model 1 to have a more extended interface, which is reflected also by the more extensive der Waals interaction of Model 1. Both models have twelve inter-helical Cα–H hydrogen bonds, although Model 2 has a better overall hydrogen bonding score.

The other three predicted models cross at different sections of the Gly-zipper. Model 3 and Model 5 are variations that cross within the N-terminal side of the zipper (**AN**xx**GL**xxGxxxG). Conversely, Model 4 crosses on the C-terminal side of the zipper (AxxxGxxx**GL**xx**GA**). CATM does not produce any model mediated by the off-frame

AxxxG motif of ADCK3 (magenta in Fig. 3.2b). The coordinates of all ADCK3 models are available as Supplementary Information and for download at http://seneslab.org/ADCK3_models.

3.3.4 ADCK3-TM self-associates strongly in *E. coli* membranes.

To investigate the structural predictions of CATM, we assessed the dimerization of ADCK3-TM and ADCK4-TM experimentally using TOXCAT, a widely used assay for TM association in biological membranes [Russ, et al. (1999)]. This assay involves the biological expression in the membrane of *Escherichia coli* of a chimeric construct that fuses the TM domain of interest with the ToxR transcriptional activator of *Vibrio cholera* (Fig. 3.4a). TM helix association leads to the dimerization of the ToxR domain, resulting in expression of the reporter gene chloramphenicol acetyltransferase (CAT). The expression level of CAT (measured by its enzymatic activity) is compared to that of a stable dimer, Glycophorin A (GpA), and to a monomeric GpA variant (GpA-G83I) as standards.

We first tested whether the constructs inserted correctly in the plasma membrane of *E. coli*, using a complementation test in the *malE* deficient strain MM39. The ADCK3-TM and ADCK4-TM TOXCAT constructs supported growth in minimal media with maltose as the sole carbon source (Fig. 3.4b), indicating that the fusion proteins are recognized as a TM domain and are expressed in the bacterial inner membrane in the correct orientation, with the MBP moiety positioned on the periplasmic side.

To examine whether ADCK3 oligomerizes in TOXCAT, we quantified the enzymatic activity of the reporter gene CAT, as an indirect measure of its expression. As shown in Fig. 3.4c, the CAT activities of the ADCK3-TM and ADCK4-TM constructs are higher than the activity of the GpA standard, which is a stable homo-dimer. These results indicate that the TM domain of ADCK3 and ADCK4 form strong homo-oligomers in TOXCAT.

3.3.5 Large scale mutagenesis demonstrates that the Gly zipper motif is important for association

To assess experimentally the interaction interface of the ADCK3-TM oligomer and validate the computational predictions, we performed large scale mutagenesis along the entire span of the TM segment, and measured their self-association in TOXCAT. Each position was individually changed to a variety of large and small hydrophobic amino acids. The expectation is that the changes at interfacial positions are more likely to perturb oligomerization than changes at lipid exposed positions, as commonly observed (for example [Lemmon, et al. (1992); Adams, et al. (1995); Fleming, et al. (2001); Jenei, et al. (2011); Li, et al. (2004)]). A total of 53 mutants were generated and analyzed in TOXCAT.

The TOXCAT data is shown in Fig. 3.8 and is schematically represented in Fig. 3.5a. To compute an overall position-dependent sensitivity to mutation, we applied a classification scheme for the variants' phenotypes using four categories (dashed lines in Fig. 3.8), labeled as “WT-like” (>80% of wild type CAT activity), “Mild” (50-80%), “Severe” (20-50%) and “Disruptive” (0-20%). These scores were then averaged to obtain a position specific “average disruption”. Position-based averaging reduces some of the natural variability of the biological assay and the method has been reliable in identifying the most sensitive positions at the helix-helix binding interface [LaPointe, et al. (2013); Adams, et al. (1995); Li, et al. (2004); Sulistijo, et al. (2003)]. The position-dependent “average disruption” is also plotted in numerical form in Fig. 3.5b.

A majority of the variants had CAT activity levels similar or higher compared to the wild type sequences (Fig. 3.5b). However, a number of variants showed dramatically reduced activity in a position specific fashion. In particular, all variants of the two C-terminal Gly residues of the Gly-zipper (G223 and G227) have the strongest disruptive phenotypes. Interestingly, the next most sensitive positions are L220 and L224, which are also predicted to be interfacial in almost all CATM models (Fig. 3.3). Conversely, the N-terminal positions of the zipper are either mildly affected by mutation (G219) or appear completely tolerant (A215).

The off-frame 221-AxxxG-225 (magenta) is also relatively insensitive to mutation. Substitution for a large Leu at these two positions has only a mild effect, and the Ile variants are completely tolerated. This is consistent with the CATM predictions, which do not identify any model in which this motif is at the interface.

A position of interest for self-association was Asn 216. Polar residues can drive TM helix oligomerization through the formation of hydrogen bonds, and have been found to be important for the association of model peptides [Choma, et al. (2000); Zhou, et al. (2000)] and of biological systems [Fleming, et al. (2001); Stanley, et al. (2007); Li, et al. (2006); Lawrie, et al. (2010)], including in the context of GAS_{right} motifs [Sulistijo, et al. (2003); Sulistijo, et al. (2009)]. In addition, some polarity of position 216 appears to be relatively conserved, as the main substitutions of N216 in a sequence alignment (Fig. 3.2c) are Gly, Ser, Gln and Glu. However, neither the computational nor the experimental analysis suggest that N216 is important for self-association. Asn 216 can be mutated to Ala, Leu or Phe in TOXCAT without reduction of self-association. CATM is in agreement with the experimental data, as it does not identify any potential strong polar interaction (i.e., N–H···O hydrogen bonds) involving the side chain of N216, although the side chain carbonyl oxygen (Oδ1) acts as a Cα–H bond acceptor in most models.

Overall the data indicates that the interface of the ADCK3-TM oligomer is mediated by Gly-zipper motif and, in particular, by the C-terminal side of this interaction motif.

3.3.6 Computational mutagenesis suggests potential alternative conformations for ADCK3-TM

In order to identify the structure most consistent with the TOXCAT data, we performed a mutational analysis of the five models generated by CATM. Using a protocol developed previously to analyze similar mutational data [LaPointe, et al. (2013)], we created *in silico* the same set of variants that were tested experimentally, and computed an analogous position-dependent “average disruption” index based on the interaction energies.

The experimental and theoretical disruption patterns are compared in Fig. 3.6. Given that all CATM structures interact through portions of the extended Gly-zipper, the computed patterns have similar periodicity across all models, with disruption peaking at position G219, G223 and/or G227. Models 1, 3 and 5 (Fig. 3.6a,c,e) show high sensitivity to mutation on the N-terminal side of the TM domain, in disagreement with the experimental observations. In these three models, mutations to G219 are completely disruptive, whereas the position is only mildly sensitive in TOXCAT. Models 1 and 5 are also very sensitive at positions A215 and N216, which are completely tolerant experimentally.

The structures of Models 2 and 4 are compared in Fig. 3.7. Model 2 (Fig. 3.6b) and Model 4 (Fig. 3.6d) are in better agreement with the experimental data and represent two possible structural solutions for the ADCK3 TM dimer. The disruption for both models peak at G223 and G227, which are also the two most disruptive positions in TOXCAT. However, Model 4 appears insensitive at position G219 (which is mildly sensitive experimentally) and it is extremely disruptive at position A228 (which is insensitive experimentally). In addition to being a better match, Model 2 also has lower energy, better packing and a larger number of hydrogen bonds. Therefore, Model 2 appears to be the best structural candidate for the ADCK3 TM dimer.

In a recent analysis of known GAS_{right} structures, we demonstrated that CATM is capable of capturing alternative conformations of biological importance [Mueller, et al. (2014)]. Therefore, an additional possibility is that the TM domain of ADCK3 may be in equilibrium between two or more structures. We observed that a linear combination of the mutagenesis profiles of the two models that best fit the data, 60% of Model 2 and 40% of Model 4, improves the fit with the TOXCAT data, producing an excellent correspondence between the two experiments (Fig. 3.6f). This interpolation is not necessarily quantitative, but it suggests that a conformational equilibrium would be compatible with the data. If such an equilibrium occurs in the biological context, it would postulate that the TM domain of ADCK3 may be a switchable element, a trait that could be important for regulation or signaling, as observed in a number of other single-span

TM proteins [Moore, et al. (2008)]. In this framework, the Gly-zipper would provide a dynamic interface for structural changes that can potentially affect either the distance of the helical termini or the relative rotation of the helices.

3.4 Conclusions

We have presented a computational and experimental analysis of the structural organization of the TM domain of the mitochondrial kinase ADCK3. While more experiments are necessary to fully test CATM, the work provides a first practical demonstration of the applicability of the program to the characterization of a TM dimer of unknown structure. It also confirms the ability of the algorithm (which is based on optimization of van der Waals and Cα–H hydrogen bonding) to correctly predict GAS_{right} motifs.

We have experimentally demonstrated that the TM domain of ADCK3 self-associates in *E. coli* membranes. While the specific oligomeric state could not determined by TOXCAT, the evidence suggests that ADCK3-TM is likely dimeric. Although Gly-zipper motifs can be involved in the formation of higher-oligomeric complexes [Kim, et al. (2005)], the good agreement between the experimental and computational mutagenesis supports the homo-dimeric hypothesis. Moreover, such oligomeric state is also consistent with a large body of structural evidence which shows that kinases frequently form dimeric complexes (for example [Cobb, et al. (2000); West, et al. (2001); Lemmon, et al. (2010)]), while higher-oligomers are rarely observed.

The analysis reveals a number of leads that may be biologically important. The helix-helix interaction interface was determined and the mutagenesis identified a number of disruptive interfacial mutations that will be useful for follow-up functional studies. The computational prediction of alternative models in which the helices adopt a different crossing point along Gly-zipper interface raises also the hypothesis that the TM domain of ADCK3 may possibly undergo conformational changes.

Indirectly, the work also provides important insight about ADCK4. All the amino acids that participate at the dimerization interface of ADCK3 are identical in ADCK4. The two positions that differ between the two sequences (F228L and A230V, Fig. 3.2) are insensitive to variation when they are mutated individually (Fig. 3.5). The computational predictions obtained for ADCK3 and ADCK4 are nearly identical and it is

thus expected that both TM domains dimerize with the same structure. Because the two interfaces are compatible with each other, it also possible that the TM domains could associate to drive formation of a hetero-dimeric complex between ADCK3 and ADCK4. These hypotheses need to be investigated in a biological context; the present analysis provides the theoretical foundation necessary for testing *in vivo* the role of these TM domains.

Table 3.1 Prediction of the transmembrane domain of the ADCK3 homologs

Name	Sequence	ΔG_{Oct}[1]	ΔG_{App}[2]	TMPRED[3]	Phobius[4]	TMHMM[5]	ΔG predictor[6]	MemBrain[7]	E(z)[8]
ADCK3	LANFGGLAVGLGFGALA	−0.28	+2.11	Yes	50%	No	Yes (+1.80)	Possible (70%)	No
ADCK4	LANFGGLAVGLGLGVLA	−0.78	+1.91	Yes	90%	40%	Yes (+1.69)	Yes (80%)	No

[1]Wimley-White octanol scale (kcal/mol)
[2]Biological hydrophobicity scale (kcal/mol)
[3]TMPRED at http://www.ch.embnet.org/software/TMPRED_form.html
[4]Phobius at http://phobius.sbc.su.se
[5]TMHMM at http://www.cbs.dtu.dk/services/TMHMM-2.0
[6]ΔG predictior at http://dgpred.cbr.su.se (in parethesis the ΔG_{App} for the predicted TM segment, kcal/mol)
[7]MemBrain at http://www.csbio.sjtu.edu.cn/bioinf/MemBrain
[8]E(z) potential at http://ez.degradolab.org/ez/original

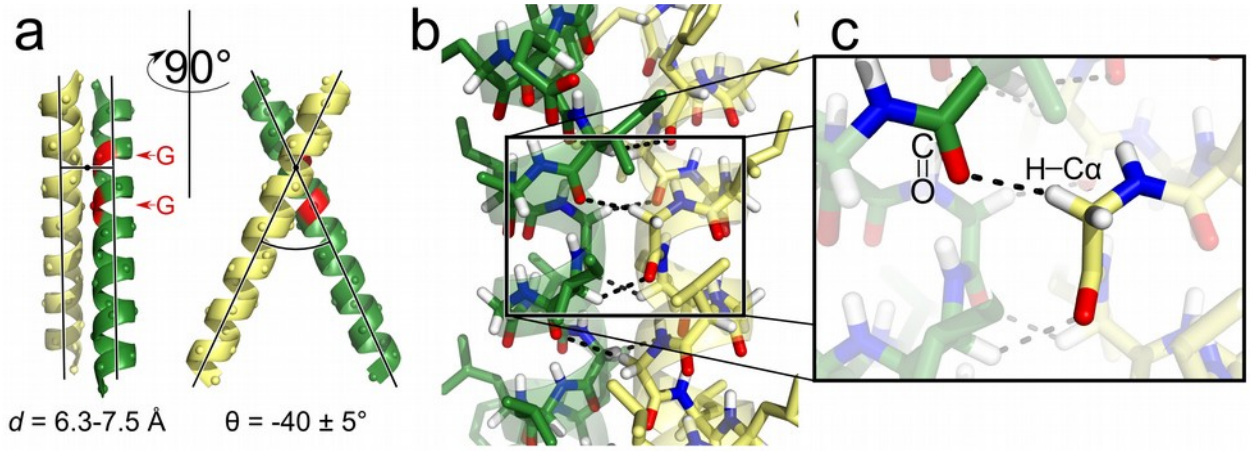


Fig. 3.1 Structural features of the GAS$_{right}$ TM association motif. a) The GAS$_{right}$ motif (which is best known as the fold of the TM region of glycophorin A) is a right-handed helical dimer with a short inter-helical distance *d* and a right-handed crossing angle θ of approximatively −40°. The GxxxG sequence pattern near the crossing point (marked in red in the green helix) allows the backbones to come into close contact. b) The contact enables the formation of networks of inter-helical hydrogen bonds between Cα–H donors and carbonyl oxygen acceptors (shown in detail in c).

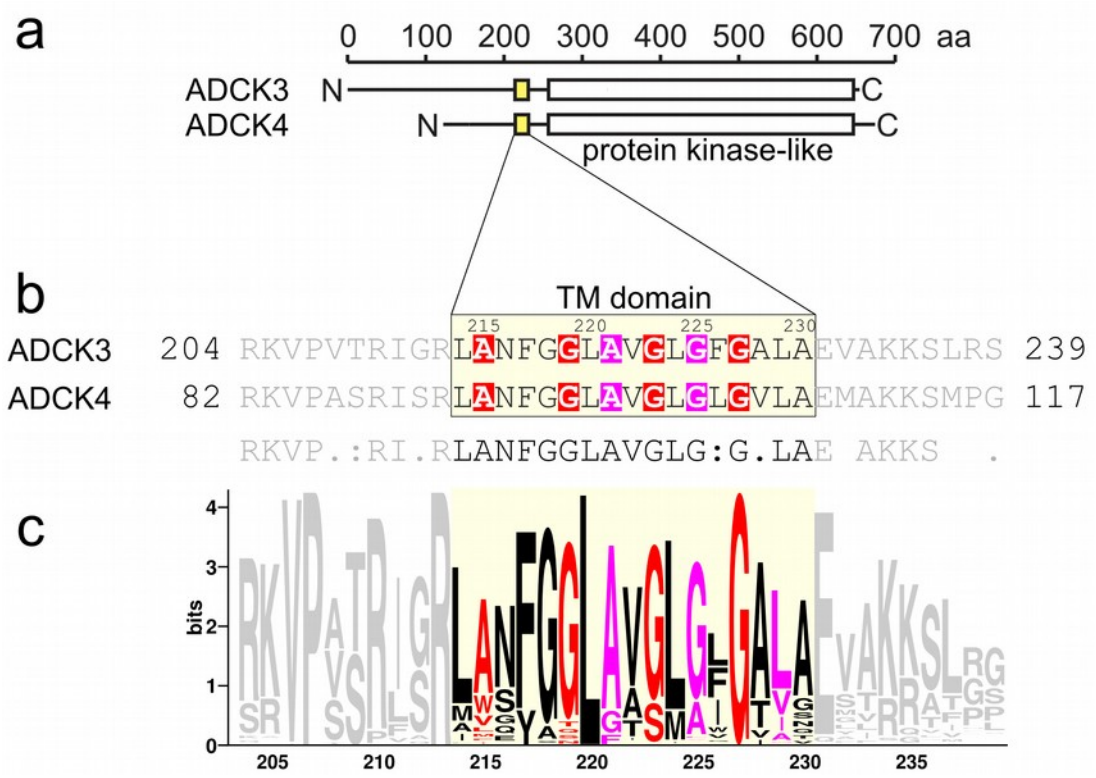
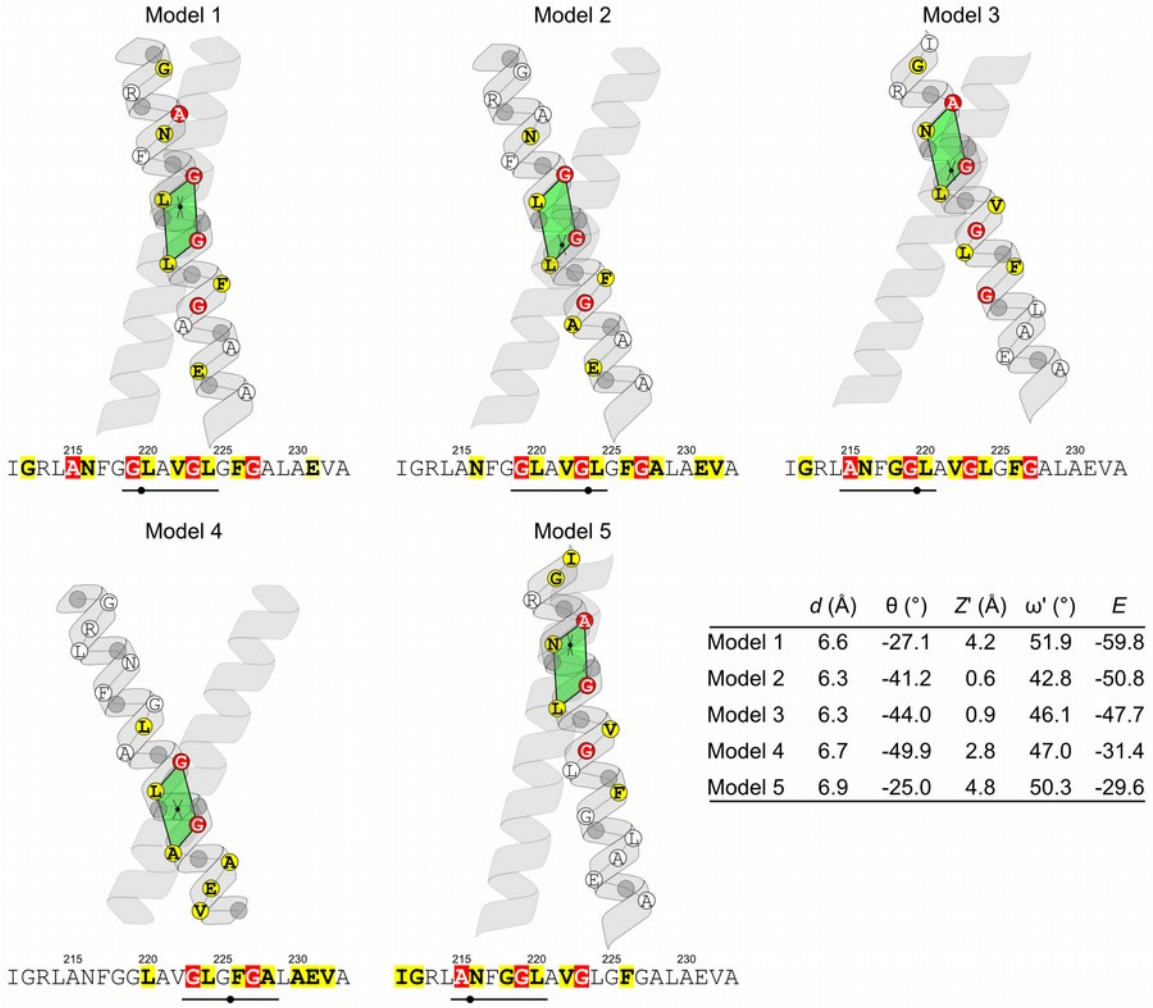


Fig. 3.2 The transmembrane domain of ADCK3 has a conserved Gly-zipper motif. a) Domain organization of ADCK3 homologs, which are proteins associated with the mitochondrial inner membrane. They are predicted to contain a TM domain (yellow) and a protein kinase-like domain (white). b) The sequence alignment of the TM domains of ADCK3, ADCK4 (yellow box). The TM domains of ADCK3 and ADKC4, which differ only at two positions, contain a number of GxxxG-like motifs, including an extended Gly-zipper motif (red) and a second AxxxG motif which is off-register by two

positions (magenta). c) Sequence logo of the alignment of 400 sequences homologous to ADCK3 from a broad range of eukaryotic species highlights conservation in the TM domain and in the N-terminal side of the juxta-membrane region. All Gly positions in the Gly-zipper (red) appear strongly conserved. The most conserved positions in the TM region are L220 and G227.



	d (Å)	θ (°)	*Z'* (Å)	ω' (°)	*E*
Model 1	6.6	-27.1	4.2	51.9	-59.8
Model 2	6.3	-41.2	0.6	42.8	-50.8
Model 3	6.3	-44.0	0.9	46.1	-47.7
Model 4	6.7	-49.9	2.8	47.0	-31.4
Model 5	6.9	-25.0	4.8	50.3	-29.6

Fig. 3.3 CATM predicts multiple modes of interaction along the Gly-zipper motif of ADCK3. Schematic representation of the five models of GAS_{right} homo-dimers generated by CATM for ADCK3-TM. The crossing point is marked by a black dot. The four positions that surround the crossing point are marked by a green parallelogram and are underlined in the sequence. The positions involved in inter-helical packing at the dimer interface are highlighted: in red are the interfacial positions that belong to the

extended Gly-zipper motif of ADCK3; all other interfacial positions are highlighted in yellow. The table summarizes the geometry of the five models: interhelical distance *d;* crossing angle θ; vertical (*Z*') and axial (ω') coordinates of the crossing point within the parallelogram of closest approach; and energy score *E*. For the geometric definitions see Fig. 3.9.

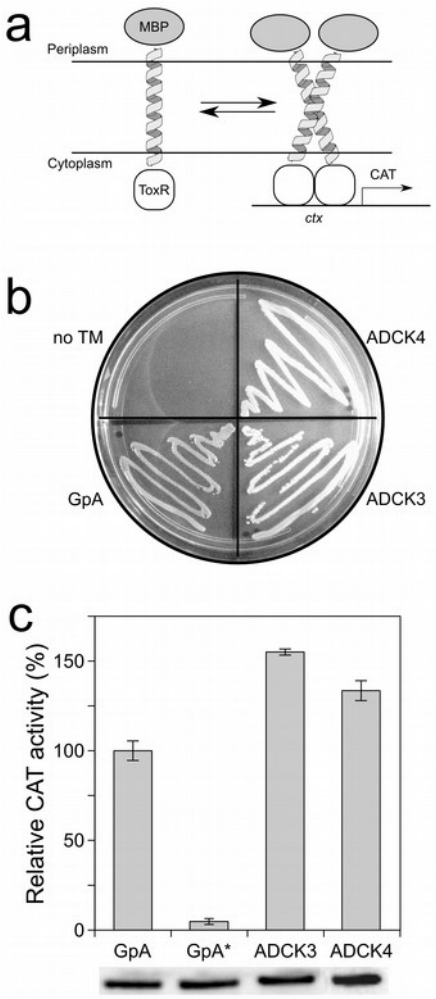


Fig. 3.4 ADCK3-TM and ADCK4-TM associate strongly in TOXCAT. a) TOXCAT is an in vivo assay based on a construct in which the transmembrane domain under investigation is fused to the ToxR transcriptional activator of *V. cholerae*. Transmembrane association results in the expression of a reporter gene in *E. coli* cells, which can be quantified. b) *malE* complementation assay. The TOXCAT construct containing the TM domain of ADCK3 and ADCK4 can use maltose as a carbon source, demonstrating correct insertion. GpA: Glycophorin A positive control; no TM: pcckan

plasmid without TM insert, negative control. c) TOXCAT assay of ADCK3 and ADCK4. ADCK3 shows approximately 150% of the CAT activity of the strong transmembrane dimer of Glycophorin A (GpA). The monomeric G83I mutant (GpA*) is used as a negative control. Data reported as average and standard deviation over four replicate experiments. Expression levels were controlled by immunoblotting.

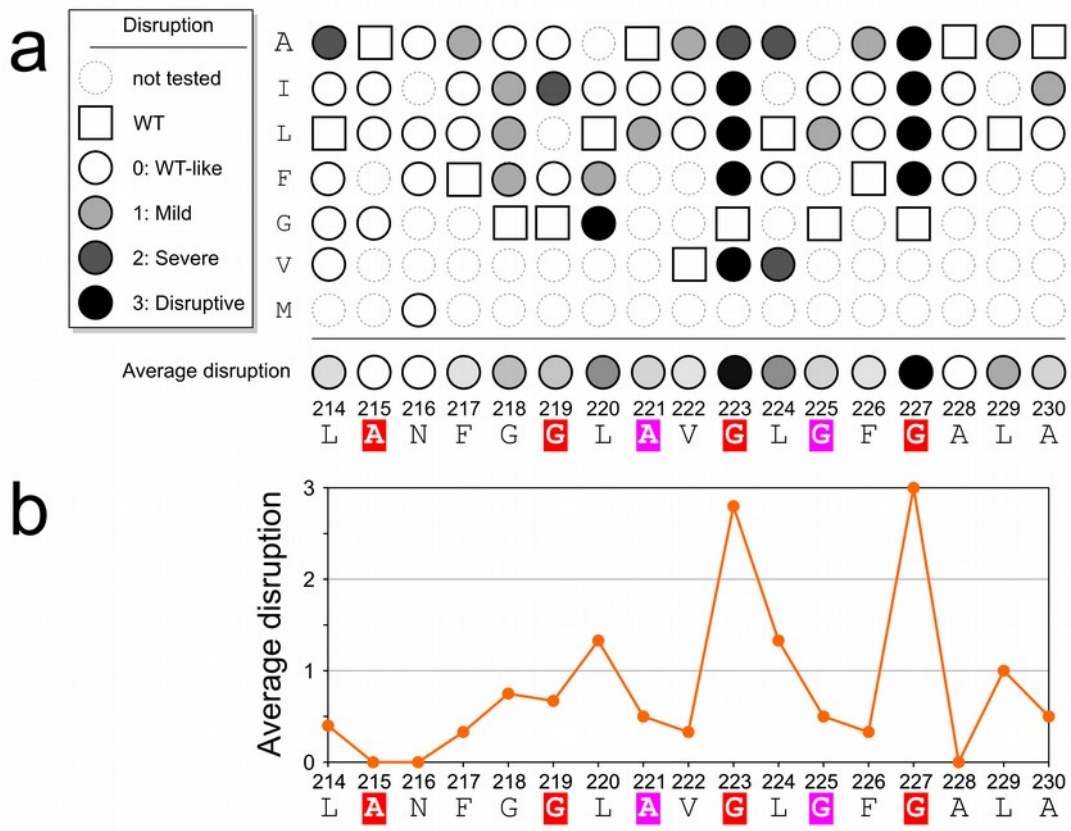


Fig. 3.5 Position specific "average disruption" suggests that the Gly-zipper is at the helical interface. a) "MacKenzie plot" summarizing the effect of all mutations of ADCK3-TM measured in TOXCAT. The color coding of the GxxxG motifs in the sequence corresponds to Fig. 1. The data has been subdivided in three categories as in the legend. The raw TOXCAT data is shown in supplementary Fig. 3.8. A calculated average disruption score for each position is displayed at the bottom of the scheme. b) The same average disruption plotted numerically (0 = as TW; 3 = disruptive). The mutagenesis reveals two positions that are essential for self-association, G223 and

G227, which are the last two position of ADCK3's Gly-zipper (red).

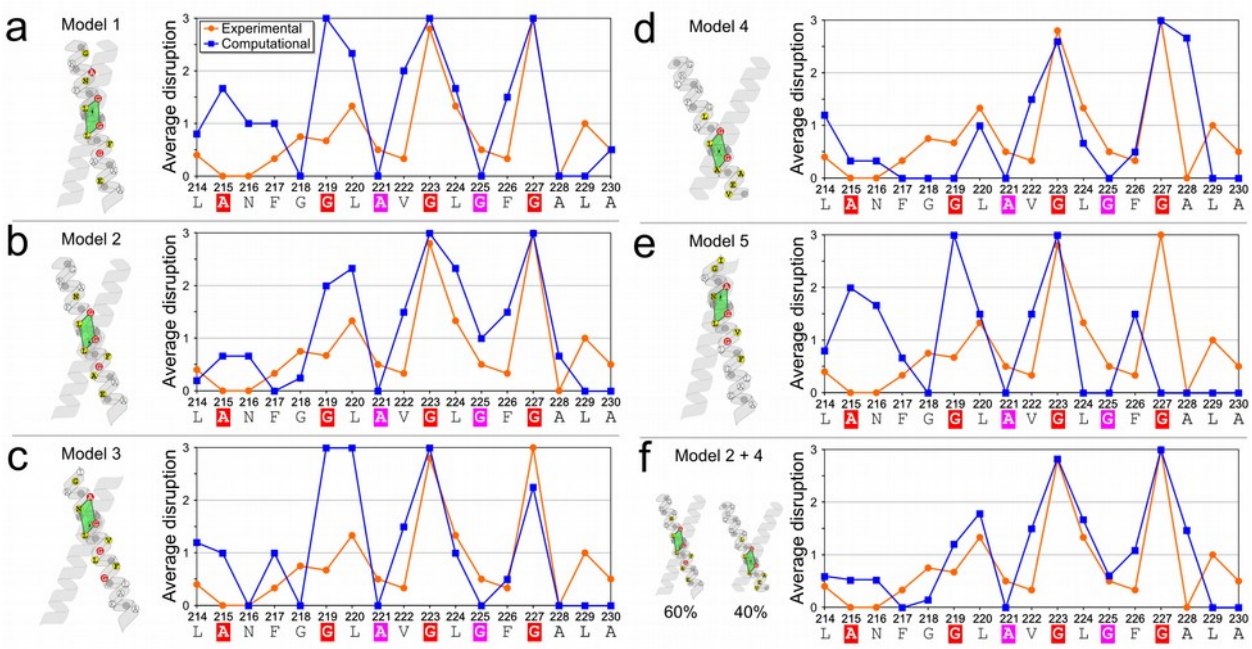


Fig. 3.6 Computational mutagenesis identifies compatible models. Comparison of the mutagenesis obtained in TOXCAT (same as Fig. 5b) with the computational mutagenesis performed on the five CATM models (panels a-e). The comparison suggest that Model 2 is the best fit to the experimental data, followed by Model 4. A linear combination of Model 2 (60%) and Model 4 (40%) produces an excellent fit to the data, suggesting that the TM of ADCK3 may be in equilibrium between at least two conformations in the TOXCAT system.

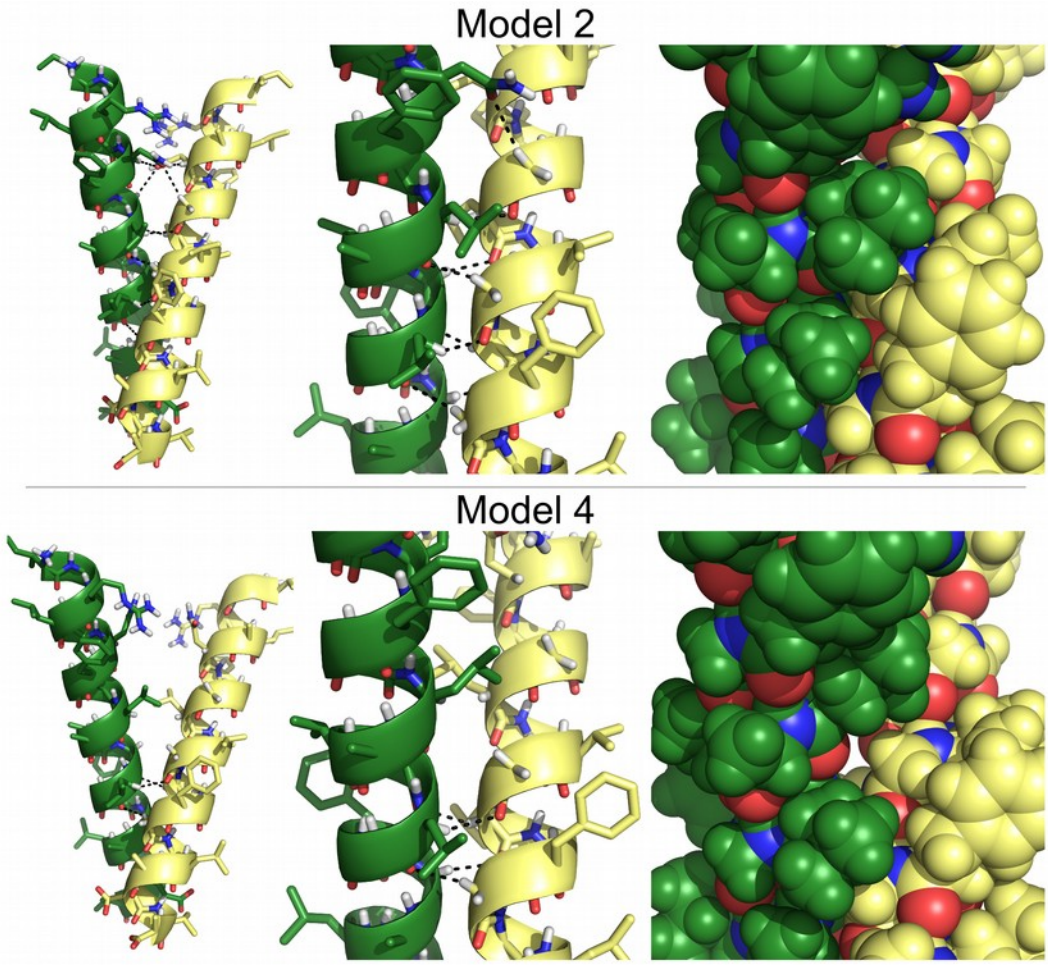


Fig. 3.7 Structural Models 2 and 4. Comparison of the structures of CATM Models 2 and 4 for ADCK3. From left to right, entire TM helix, detail of the interface, and same conformation in full atom spheres. Model 2 has lower energy, a larger number of hydrogen bonds (12 in Model 2 versus 4 in Model 4) and more extended and complementary packing.

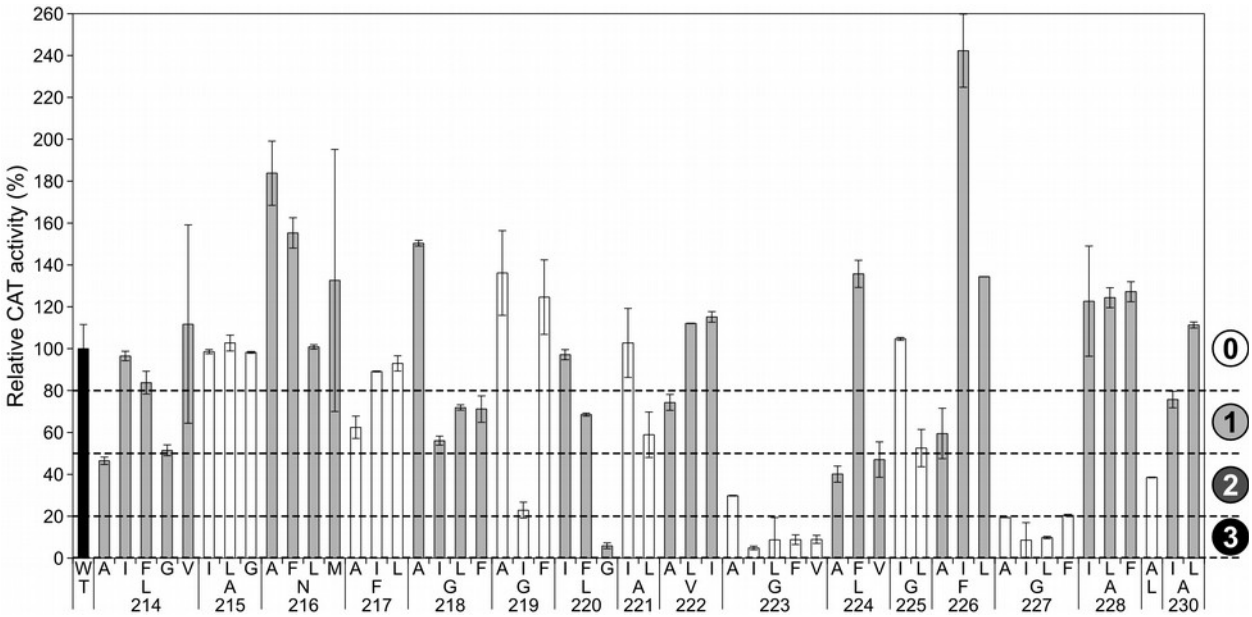


Fig. 3.8 Mutagenesis of the TM helix of ADCK3. The figure shows the TOXCAT result for each of the point mutants of the TM domain of ADCK3 schematically summarized in Fig. 5. The CAT activity (left axis) is normalized to that of the wild type sequence, shown in black. The mutations at each position are visually grouped by color. Each mutation has been categorized relative to the wild type activity (back bar) as “WT-like” (0: >80% of WT), “Mild” (1: 50-80%), “Severe” (2: 20-50%) or “Disruptive” (3: 0-20%), as indicated on the right axis and by the dashed lines.

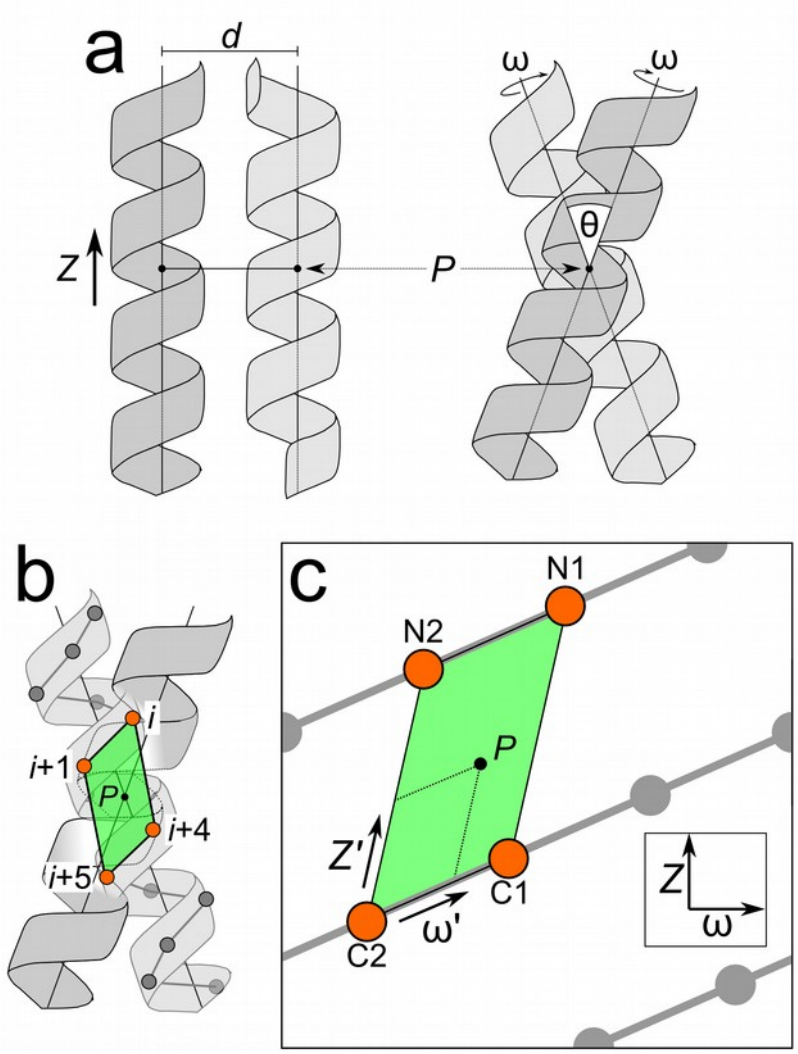


Fig. 3.9 Definition of 4 parameters that define the geometry of a symmetrical dimer. a) *d*: inter-helical distance; θ: crossing angle; ω: rotation of the helix around its axis; *Z*: vertical position of the point of closest approach between the two helical axes (the crossing point *P*). b) The coordinates can be redefined by expressing them as a function of the unit cell (green) on the helical lattice that contains the point of closest approach *P.* The four interfacial positions that surround the the point of closest

approach are designated as N1 (relative position *i*), N2 (*i*+1), C1(*i*+4) and C2 (*i*+5). The principal axes are the rotation along the helical screw (ω') and the vector between C2 and C2 (*Z'*).

3.5 References

Adams, P. D.; Arkin, I. T.; Engelman, D. M.; Brünger, A. T. *Nat. Struct. Biol.* 1995, 2, 154.

Arbely, E.; Arkin, I. T. *J. Am. Chem. Soc.* 2004, 126, 5362.

Arkin, I. T.; Brunger, A. T. *Biochim. Biophys. Acta* 1998, 1429, 113.

Ashraf, S.; Gee, H. Y.; Woerner, S.; Xie, L. X.; Vega-Warner, V.; Lovric, S.; Fang, H.; Song, X.; Cattran, D. C.; Avila-Casado, C.; Paterson, A. D.; Nitschké, P.; Bole-Feysot, C.; Cochat, P.; Esteve-Rudd, J.; Haberberger, B.; Allen, S. J.; Zhou, W.; Airik, R.; Otto, E. A.; Barua, M.; Al-Hamed, M. H.; Kari, J. A.; Evans, J.; Bierzynska, A.; Saleem, M. A.; Böckenhauer, D.; Kleta, R.; El Desoky, S.; Hacihamdioglu, D. O.; Gok, F.; Washburn, J.; Wiggins, R. C.; Choi, M.; Lifton, R. P.; Levy, S.; Han, Z.; Salviati, L.; Prokisch, H.; Williams, D. S.; Pollak, M.; Clarke, C. F.; Pei, Y.; Antignac, C.; Hildebrandt, F. J. *Clin. Invest.* 2013, 123, 5179.

Betz, S. F.; Bryson, J. W.; DeGrado, W. F. *Curr. Opin. Struct. Biol.* 1995, 5, 457.

Bowie, J. U. *Curr. Opin. Struct. Biol.* 2011, 21, 42.

Bowie, J. U. *Nature* 2005, 438, 581.

Chang, Y.-C.; Bowie, J. U. *Proc. Natl. Acad. Sci. U. S. A.* 2014, 111, 219.

Choma, C.; Gratkowski, H.; Lear, J. D.; DeGrado, W. F. *Nat. Struct. Biol.* 2000, 7, 161.

Cobb, M. H.; Goldsmith, E. J. *Trends Biochem. Sci.* 2000, 25, 7.

Cunningham, F.; Poulsen, B. E.; Ip, W.; Deber, C. M. *Biopolymers* 2011, 96, 340.

Daley, D. O.; Clifton, R.; Whelan, J. *Proc. Natl. Acad. Sci. U. S. A.* 2002, 99, 10510.

Daley, D. O.; Whelan, J. *Genome Biol.* 2005, 6, 110.

Do, T. Q.; Hsu, A. Y.; Jonassen, T.; Lee, P. T.; Clarke, C. F. J. Biol. Chem. 2001, 276, 18161.
Doura, A. K.; Kobus, F. J.; Dubrovsky, L.; Hibbard, E.; Fleming, K. G. *J. Mol. Biol.* 2004, 341, 991.

Engelman, D. M.; Chen, Y.; Chin, C.-N.; Curran, A. R.; Dixon, A. M.; Dupuy, A. D.; Lee, A. S.; Lehnert, U.; Matthews, E. E.; Reshetnyak, Y. K.; Senes, A.; Popot, J.-L. *FEBS Lett.* 2003, 555, 122.

Fisher, L. E.; Engelman, D. M.; Sturgis, J. N. *J. Mol. Biol.* 1999, 293, 639.

Fleming, K. G. *J. Mol. Biol.* 2002, 323, 563.

Fleming, K. G.; Engelman, D. M. *Proc. Natl. Acad. Sci. U. S. A.* 2001, 98, 14340.

Fleming, K. G.; Engelman, D. M. *Proteins* 2001, 45, 313.

Gerards, M.; van den Bosch, B.; Calis, C.; Schoonderwoerd, K.; van Engelen, K.; Tijssen, M.; de Coo, R.; van der Kooi, A.; Smeets, H. *Mitochondrion* 2010, 10, 510.

Harbury, P. B.; Tidor, B.; Kim, P. S. *Proc. Natl. Acad. Sci. U. S. A.* 1995, 92, 8408.

Harbury, P. B.; Zhang, T.; Kim, P. S.; Alber, T. *Science* 1993, 262, 1401.

He, C. H.; Xie, L. X.; Allan, C. M.; Tran, U. C.; Clarke, C. F. *Biochim. Biophys. Acta* 2014, 1841, 630.

Herrmann, J. R.; Fuchs, A.; Panitz, J. C.; Eckert, T.; Unterreitmeier, S.; Frishman, D.; Langosch, D. *J. Mol. Biol.* 2010, 396, 452.

Hessa, T.; Kim, H.; Bihlmaier, K.; Lundin, C.; Boekel, J.; Andersson, H.; Nilsson, I.; White, S. H.; von Heijne, G. *Nature* 2005, 433, 377.

Hessa, T.; Meindl-Beinker, N. M.; Bernsel, A.; Kim, H.; Sato, Y.; Lerch-Bader, M.; Nilsson, I.; White, S. H.; von Heijne, G. *Nature* 2007, 450, 1026.

Hong, H.; Blois, T. M.; Cao, Z.; Bowie, J. U. *Proc. Natl. Acad. Sci. U. S. A.* 2010, 107, 19802.

Hong, H.; Chang, Y.-C.; Bowie, J. U. *Methods Mol. Biol.* Clifton NJ 2013, 1063, 37.

Horie, C.; Suzuki, H.; Sakaguchi, M.; Mihara, K. *J. Biol. Chem.* 2003, 278, 41462.

Horie, C.; Suzuki, H.; Sakaguchi, M.; Mihara, K. *Mol. Biol. Cell* 2002, 13, 1615.

Horowitz, S.; Trievel, R. C. *J. Biol. Chem.* 2012, 287, 41576.

Horvath, R.; Czermin, B.; Gulati, S.; Demuth, S.; Houge, G.; Pyle, A.; Dineiger, C.; Blakely, E. L.; Hassani, A.; Foley, C.; Brodhun, M.; Storm, K.; Kirschner, J.; Gorman, G. S.; Lochmüller, H.; Holinski-Feder, E.; Taylor, R. W.; Chinnery, P. F. J. *Neurol. Neurosurg. Psychiatry* 2012, 83, 174.

Hubert, P.; Sawma, P.; Duneau, J.-P.; Khao, J.; Hénin, J.; Bagnard, D.; Sturgis, J. *Cell*

Adhes. Migr. 2010, 4, 313.

Jenei, Z. A.; Warren, G. Z. L.; Hasan, M.; Zammit, V. A.; Dixon, A. M. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* 2011, 25, 4522.

Käll, L.; Krogh, A.; Sonnhammer, E. L. L. *J. Mol. Biol.* 2004, 338, 1027.

Käll, L.; Krogh, A.; Sonnhammer, E. L. L. *Nucleic Acids Res.* 2007, 35, W429.

Kannan, N.; Taylor, S. S.; Zhai, Y.; Venter, J. C.; Manning, G. *PLoS Biol.* 2007, 5, e17.

Khadria, A.; Senes, A. *Methods Mol. Biol.* Clifton NJ 2013, 1063, 19.

Kim, S.; Jeon, T.-J.; Oberai, A.; Yang, D.; Schmidt, J. J.; Bowie, J. U. *Proc. Natl. Acad. Sci. U. S. A.* 2005, 102, 14278.

Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L. *Proteins* 2009, 77, 778.

Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L. *J. Mol. Biol.* 2001, 305, 567.

Kulp, D. W.; Subramaniam, S.; Donald, J. E.; Hannigan, B. T.; Mueller, B. K.; Grigoryan, G.; Senes, A. *J. Comput. Chem.* 2012, 33, 1645.

Lagier-Tourenne, C.; Tazir, M.; López, L. C.; Quinzii, C. M.; Assoum, M.; Drouot, N.; Busso, C.; Makri, S.; Ali-Pacha, L.; Benhassine, T.; Anheim, M.; Lynch, D. R.; Thibault, C.; Plewniak, F.; Bianchetti, L.; Tranchant, C.; Poch, O.; DiMauro, S.; Mandel, J.-L.; Barros, M. H.; Hirano, M.; Koenig, M. Am. *J. Hum. Genet.* 2008, 82, 661.

LaPointe, L. M.; Taylor, K. C.; Subramaniam, S.; Khadria, A.; Rayment, I.; Senes, A. *Biochemistry* 2013, 52, 2574.

Lawrie, C. M.; Sulistijo, E. S.; MacKenzie, K. R. *J. Mol. Biol.* 2010, 396, 924.

Lemmon, M. A.; Flanagan, J. M.; Treutlein, H. R.; Zhang, J.; Engelman, D. M. *Biochemistry* 1992, 31, 12719.

Lemmon, M. A.; Schlessinger, J. *Cell* 2010, 141, 1117.

Lemmon, M. A.; Schlessinger, J.; Ferguson, K. M. *Cold Spring Harb. Perspect. Biol.* 2014, 6, a020768.

Leonard, C. J.; Aravind, L.; Koonin, E. V. *Genome Res.* 1998, 8, 1038.

Li, E.; Hristova, K. *Biochemistry* 2006, 45, 6241.

Li, E.; You, M.; Hristova, K. *J. Mol. Biol.* 2006, 356, 600.

Li, R.; Gorelik, R.; Nanda, V.; Law, P. B.; Lear, J. D.; DeGrado, W. F.; Bennett, J. S. *J. Biol. Chem.* 2004, 279, 26666.

Lindner, E.; Langosch, D. *Proteins* 2006, 65, 803.

Lundquist, P. K.; Poliakov, A.; Giacomelli, L.; Friso, G.; Appel, M.; McQuinn, R. P.; Krasnoff, S. B.; Rowland, E.; Ponnala, L.; Sun, Q.; van Wijk, K. J. *Plant Cell* 2013, 25, 1818.

Mackenzie, K. R. *Chem. Rev.* 2006, 106, 1931.

MacKenzie, K. R.; Fleming, K. G. *Curr. Opin. Struct. Biol.* 2008, 18, 412.

MacKenzie, K. R.; Prestegard, J. H.; Engelman, D. M. *Science* 1997, 276, 131.

MacKerell; Bashford, D.; Bellott; Dunbrack; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* 1998, 102, 3586.

Martinis, J.; Glauser, G.; Valimareanu, S.; Kessler, F. *Plant Physiol.* 2013, 162, 652.

Merzlyakov, M.; You, M.; Li, E.; Hristova, K. *J. Mol. Biol.* 2006, 358, 1.

Mollet, J.; Delahodde, A.; Serre, V.; Chretien, D.; Schlemmer, D.; Lombes, A.; Boddaert, N.; Desguerre, I.; de Lonlay, P.; de Baulny, H. O.; Munnich, A.; Rötig, A. Am. *J. Hum. Genet.* 2008, 82, 623.

Moore, D. T.; Berger, B. W.; DeGrado, W. F. *Structure* 2008, 16, 991.

Mottamal, M.; Lazaridis, T. *Biochemistry* 2005, 44, 1607.

Mueller, B. K.; Subramaniam, S.; Senes, A. *Proc. Natl. Acad. Sci. U. S. A.* 2014, 111, E888.

Ojemalm, K.; Botelho, S. C.; Stüdle, C.; von Heijne, G. *J. Mol. Biol.* 2013, 425, 2813.

Park, H.; Yoon, J.; Seok, C. J. *Phys. Chem. B* 2008, 112, 1041.

Poon, W. W.; Davis, D. E.; Ha, H. T.; Jonassen, T.; Rather, P. N.; Clarke, C. F. *J.*

Bacteriol. 2000, 182, 5139.

Popot, J. L.; Engelman, D. M. *Biochemistry* 1990, 29, 4031.

Rath, A.; Deber, C. M. *Annu. Rev. Biophys.* 2012, 41, 135.

Rath, A.; Glibowicka, M.; Nadeau, V. G.; Chen, G.; Deber, C. M. *Proc. Natl. Acad. Sci. U. S. A.* 2009.

Rhee, H.-W.; Zou, P.; Udeshi, N. D.; Martell, J. D.; Mootha, V. K.; Carr, S. A.; Ting, A. Y. *Science* 2013, 339, 1328.

Russ, W. P.; Engelman, D. M. *J. Mol. Biol.* 2000, 296, 911.

Russ, W. P.; Engelman, D. M. *Proc. Natl. Acad. Sci.* 1999, 96, 863.

Scheiner, S.; Kar, T.; Gu, Y. *J. Biol. Chem.* 2001, 276, 9832.

Schneider, D.; Engelman, D. M. *J. Biol. Chem.* 2003, 278, 3105.

Senes, A.; Chadi, D. C.; Law, P. B.; Walters, R. F. S.; Nanda, V.; Degrado, W. F. *J. Mol. Biol.* 2007, 366, 436.

Senes, A.; Engel, D. E.; DeGrado, W. F. *Curr. Opin. Struct. Biol.* 2004, 14, 465.

Senes, A.; Gerstein, M.; Engelman, D. M. *J. Mol. Biol.* 2000, 296, 921.

Senes, A.; Ubarretxena-Belandia, I.; Engelman, D. M. *Proc. Natl. Acad. Sci. U. S. A.* 2001, 98, 9056.

Shaw, W. V. *Methods Enzymol.* 1975, 43, 737.

Shen, H.; Chou, J. J. *PloS One* 2008, 3, e2399.

Stanley, A. M.; Fleming, K. G. *J. Mol. Biol.* 2007, 370, 912.

Subramaniam, S.; Senes, A. *Proteins* 2012, 80, 2218.

Sulistijo, E. S.; Jaszewski, T. M.; MacKenzie, K. R. *J. Biol. Chem.* 2003, 278, 51950.

Sulistijo, E. S.; Mackenzie, K. R. *Biochemistry* 2009, 48, 5106.

Supekova, L.; Supek, F.; Greer, J. E.; Schultz, P. G. *Proc. Natl. Acad. Sci. U. S. A.* 2010, 107, 5047.

Tan, T.; Ozbalci, C.; Brügger, B.; Rapaport, D.; Dimmer, K. S. *J. Cell Sci.* 2013, 126, 3563.

Tong, J.; Dolezal, P.; Selkrig, J.; Crawford, S.; Simpson, A. G. B.; Noinaj, N.; Buchanan, S. K.; Gabriel, K.; Lithgow, T. *Mol. Biol. Evol.* 2011, 28, 1581.

Tusnády, G. E.; Simon, I. *Bioinforma. Oxf. Engl.* 2001, 17, 849.

Tusnády, G. E.; Simon, I. *J. Mol. Biol.* 1998, 283, 489.

UniProt Consortium. *Nucleic Acids Res.* 2013, 41, D43.

Unterreitmeier, S.; Fuchs, A.; Schäffler, T.; Heym, R. G.; Frishman, D.; Langosch, D. *J. Mol. Biol.* 2007, 374, 705.

Vargas, R.; Garza, J.; Dixon, D. A.; Hay, B. P. *J Am Chem Soc* 2000, 122, 4750.

Wallin, E.; von Heijne, G. *Protein Sci. Publ. Protein Soc.* 1998, 7, 1029.

Walters, R. F. S.; DeGrado, W. F. *Proc. Natl. Acad. Sci. U. S. A.* 2006, 103, 13658.

West, A. H.; Stock, A. M. *Trends Biochem. Sci.* 2001, 26, 369.

Wimley, W. C.; Creamer, T. P.; White, S. H. *Biochemistry* 1996, 35, 5109.

Yang, J.; Jang, R.; Zhang, Y.; Shen, H.-B. *Bioinforma. Oxf. Engl.* 2013, 29, 2579.

Yohannan, S.; Faham, S.; Yang, D.; Grosfeld, D.; Chamberlain, A. K.; Bowie, J. U. *J. Am. Chem. Soc.* 2004, 126, 2284.

You, M.; Li, E.; Wimley, W. C.; Hristova, K. *Anal. Biochem.* 2005, 340, 154.

Zhou, F. X.; Cocco, M. J.; Russ, W. P.; Brunger, A. T.; Engelman, D. M. *Nat. Struct. Biol.* 2000, 7, 154.

Chapter 4

Determination of the dimerization potential of human genome GAS_{right} mediated single-pass membrane proteins

This chapter is being finalized, I am the primary contributor, with additional work done by Samantha Anderson, Evan Lange, Sabareesh Subramaniam, and Alessandro Senes

4.1 Introduction

Single-pass transmembrane (TM) proteins are proteins that span the lipid bilayer as an α-helix only once. Single-pass TM proteins are quite common across most organisms [Krogh, et al. (2001)], and their contributions in humans to many important biological functions are well studied [Arnaout, et al. (2005); Ullrich, et al. (1990); Weiss (2003); Munter, et al. (2007)]. The TM domain in particular has been shown to be an important and necessary component for function [He, et al. (2012); Kondo, et al. (2010); Munter, et al. (2007); Yin, et al. (2007)]. The activity and function of a protein is determined by its composition and many studies have looked at the amino acids responsible for facilitating TM helical association [Senes, et al. (2000); Russ, et al. (2000); Zhou, et al. (2000); Choma, et al. (2000); Sal-Man, et al. (2004)]. Particular attention has been given to the role of commonly occurring sequence or structural motifs. One of the most studied TM protein motifs are two glycine residues spaced at *i* and *i+4*, the GxxxG motif, along with closely related motifs containing serine and alanine, such as AxxxG, SxxxG [Senes, et al. (2000)]. These small residue motifs G/A/SxxxG/A/S are the important component in the GAS_{right} structural motif [Walters, et al. (2006)]. The GAS_{right} fold is the association of two parallel transmembrane helices at a right-handed crossing angle of around -40 degrees, placing Glycine, Alanine, or Serine (GAS) at the interface (see Fig. 4.1).

In our previous work we showed that the GAS_{right} motif is optimized for the formation of interhelical Cα–H···O=C hydrogen bonds [Mueller, et al. (2014)]. These hydrogen bonds form between the α-carbon hydrogen on one helix to the carbonyl oxygen on the opposing helix. The presence of small residues, most importantly glycine, allow for the formation of backbone-to-backone, Cα–H···O=C hydrogen bonds. While the energetic contribution of these bonds is still a matter of debate [Arbely, et al. (2004); Yohannan, et al. (2004)], they are commonly seen in GAS_{right} structures and are predicted to have a favorable contribution to protein folding. By understanding the formation of these bonds we have shown that it is possible to predict *ab initio* the formation of GAS_{right} structures.

Hydrogen bonds form due to a hydrogen atom bonded to an electronegative atom, pulling the electron density away from the hydrogen. In biology, the most common example of a hydrogen bond is the hydrogen being bonded to either oxygen or nitrogen, while comparatively carbon is not as electronegative. However, in the context of a protein the α-carbon is surrounded by two electron withdrawing amide groups, enhancing the hydrogen bond strength. Many quantum calculations have calculated the strength of the C–H···O hydrogen bond in both idealized model systems and more native-like amino acids and, on average, have shown that C–H···O hydrogen bonds can be 1/3 to ½ the strength of a canonical hydrogen bond [Gu, et al. (1999)]. In the low-dielectric environment of the lipid bilayer, the Cα–H···O hydrogen bond should be a more important force, as there will be no water to satisfy the polar group. Unfortunately, *in vitro* measurements of Cα–H···O hydrogen bonds in membrane proteins has yielded relatively few results. One attempt probed the Cα-H···O bond from Cα-H of Ala51 to the O_γ of Thr24 in bacteriorhodopsin (bR) using an sodium dodecyl sulfate (SDS) unfolding assay. Mutations were performed to remove the backbone-to-sidechain Cα-H···O bond and the results found that the elimination of the bond was negligible or even slightly favorable [Yohannan, et al. (2004)]. Using a different technique Arbely and Arkin found a different result. They used fourier transform infrared spectroscopy (FTIR), and calculated a favorable ΔG of -0.88 kcals/mol for a carbon hydrogen bond interaction, by compared the CD_2 asymmetric stretching mode between dimeric glycophorin A (GpA) and a known monomeric mutant [Arbely, et al. (2004)]. The apparent contradiction was reconciled in a study done by Mottamal and Lazaridis; computationally analyzing the geometries and finding that the placement of the carbon and hydrogen in the bR protein were unfavorable for Cα–H···O hydrogen bond formation as opposed to the GpA protein [Mottamal, et al. (2005)]. These weak hydrogen bonds often appear in networks [Senes, et al. (2001)], and the individual bonds can vary in strength due to their geometry [Park, et al. (2008)]. Therefore it is likely that their contribution to GAS_{right} TM dimerization can exist over a range of values.

While the energetic contribution to folding and association of membrane proteins is still not well understood, in our previous work we have shown it is possible to predict the structure of homo-dimeric GAS_{right} proteins [Mueller, et al. (2014)]. The occurrence of the G/A/SxxxG/A/S is frequent in single-pass proteins, with over 60% of non-redundant TMs containing at least one motif. Even when the pattern is constrained to contain at least one glycine on the N or C side of the motif occurs in 42% of all TM domains [Senes et al. (2000)]. However the occurrence of a G/A/SxxxG/A/S motif does not in itself confer a strong dimerization potential.

The GxxxG motif has been extensively studied in the context of the GpA dimer. While maintaining the GxxxG motif, point mutations at other positions on the helix can vary the dimerization energy by -0.5 to +3.2 kcals/mol [Doura, et al. (2004)]. The G/A/SxxxG/A/S motif has also been studied in an array of human proteins. The Receptor Tyrosine Kinase family has been extensively studied, with many protein dimer structures solved via NMR [Endres, et al. (2013); Bocharov, et al. (2008); Mineev, et al. (2011); Bocharov, et al. (2012); Mineev, et al. (2010); Bocharov, et al. (2010); Bocharov, et al. (2008); Bocharov, et al. (2013); Manni, et al. (2014)], and the relative dimerization of all 58 human RTKs analyzed with the biological assay TOXCAT. Even though many of the proteins contain GxxxG or G/A/SxxxG/A/S motifs the level of dimerization is varied, with the weakest dimer having a tenth of the relative dimerization signal of the strongest dimer [Finger, et al. (2009)]. Mutations that cause changes in RTK TM dimerization have been shown to have phenotypic effects, resulting in disease [He, et al. (2012)]. A similar assay to TOXCAT, ToxR, was used to study representative sequences of all human single-pass TMs. Approximately 60% of the representative sequences contain a G/A/SxxxG/A/S motif, and the realtive dimerization range of motif containing structures vary from 40% to 140% of a known standard [Kirrbach, et al. (2013)]. To test for the G/A/SxxxG/A/S motif being present at the dimer interface, mutations were performed on the motif residues. Among proteins with mutation sensitive motif residues (likely to have the motif at the dimer interface), the range of relative dimerization was ~80% to 140%.

While the GxxxG motif and its associated G/A/SxxxG/A/S counterparts are common and seen in many strongly dimerizing TM regions, the context of the motif is important. The motif exists in a large dimerization range, and as shown by the GpA mutation study even when the GxxxG motif is maintained single point mutations have a large effect on the dimerization energy. A single glycine residue and to a greater extent the GxxxG motif are necessary for the formation of Cα–H···O hydrogen bonds in TM homo-dimers [Mueller, et al. (2014)], however the extent of formation of these bonds and their impact on the dimerization of the protein is dependent on the entire sequence space.

In this study we wanted to predict the structure of GAS_{right} proteins on a genome-wide level, understand the breath and importance of the GAS_{right} motif and further study the contribution of Cα–H···O hydrogen bonds in the GAS_{right} motif. We found that approximately half of the single-pass human genome proteins can potentially homo-dimerize via a GAS_{right} motif. We also found that our algorithm, CATM, can predict the relative strength of dimerization of TM domains that contain only hydrophobic and glycine residues – proteins that most likely dimerize via a GAS_{right} motif. CATM can also predict to a lesser degree TM domains containing serine, threonine and tyrosine residues as well. However, CATM cannot correctly predict proteins containing strong polar or charged residues, most likely due to the fact that the proteins do not dimerize via a GAS_{right} motif. Preliminary results also show that there may be a measurable difference in relative dimerization between proteins with high and low Cα–H···O=C hydrogen bond energy. Dimers with high Cα–H···O=C hydrogen bond energy dimerize to a greater extent that proteins with low Cα–H···O=C hydrogen bond energy.

4.2 Methods

4.2.1 Software

All calculations were implemented and performed using the MSL molecular modeling libraries v. 1.1 , an open source C++ library that is freely available.

4.2.2 CATM Algorithm to Predict Structure and Dimerization Energy

The CATM algorithm is described in our previous work [Mueller, et al. (2014)]. Briefly the sequence of interest is threaded into a set of different registers at each of the representative geometries. If the sequence rules are met, the sequence is built on the backbone in all atoms and the side chains are optimized. If the final interaction energy is negative, the solution is accepted. Previously Monte Carlo backbone perturbation cycles were performed at each step, now only the lowest energy structures undergo backbone perturbation.

Energies were previously 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)] with modified parameters for Cα donors [Mueller, et al. (2014)]. For a more accurate representation of the membrane environment the IMM1 implicit solvation potential was used in this work [Lazaridis, (2003)].

4.2.3 Vectors and Strains

The expression vectors pccKAN, pccGpA-wt, and pccGpA-G83I, and *malE* deficient *E. coli* strain MM39 were kindly provided by Dr. Donald M. Engelman [Russ, et al. (1999)]. Genes containing the transmembrane region of interest were cloned into the NheI-BamHI restriction sites of the pccKAN vector.

The TM encoding genes of interest were ordered 5 to 12 on a gblock purchased from Integrated DNA Technologies. TM genes (on the order of 54 bp) were placed between the restriction sites NheI and DpnII, genes and their respective cut sites were spaced by a few bp to allow for proper enzymatic digestion. Digestion was done sequentially, the product purified, and the multiple genes were cloned into the NheI-BamHI restriction

sites of the pccKAN vector.

4.2.4 Expression of Chimeric Proteins in MM39 Cells and MalE Complementation Assay

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. 30 µL of overnight cultures were inoculated into 3 mL of LB broth and grown to an OD600 of approximately 1.0 at 37 °C. After recording the optical density, 1 mL of cells was spun down for 10 min at 17000g 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 5-8 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 17000g, 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.

4.2.5 Chloramphenicol Acetyltransferase (CAT) Spectrophotometric Assay

CAT activity was measured as described. Briefly, 1 mL of buffer containing 0.1 mM acetyl CoA, 0.4 mg/mL 5,5′- dithiobis(2-nitrobenzoic acid) or Ellman’s reagent, and 0.1 M TrisHCl pH 7.8, were mixed with 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 OD600. All measurements were determined by at least two biological and two technical replicates.

4.2.6 Quantification of Expression by Immunoblotting

Protein expression was confirmed by immunoblotting. The cell lysates (10 µL) were loaded onto a NuPAGE 4-12% Bis-Tris SDS-PAGE gel and then transferred to PVDF membranes for 1 hour at 100 millivolts. Blots were blocked using 5% Bovine serum albumin in TBS-Tween buffer (50 mM Tris, 150 mM NaCl, 0.05% Tween 20) for two hours at 4 ˚C, incubated with biotinylated anti-Maltose Binding Protein antibodies overnight at 4 ˚C, followed by peroxidase-conjugated streptavidin for two hours at 4 ˚C. Blots were developed with the Pierce ECL Western Blotting Substrate Kit and chemiluminescence was measured..

4.3 Results & Discussion

4.3.1 Prediction of Human GAS_{right} Dimeric Proteins

Both the G/A/SxxxG/A/S sequence motif and the GAS_{right} structural motif are common in membrane proteins, especially single-pass membrane proteins. The CATM algorithm has shown its ability to predict the structure of known GAS_{right} proteins [Mueller, et al. (2014)] and well as predict the interface of a protein of unknown structure, ADCK3 [Khadria, et al. (2014)]. Therefore we wanted to predict the structure of GAS_{right} proteins on a genome-wide level, to be able to understand the breath and importance of the GAS_{right} motif. By predicting genome-wide dimerization we can determine the amount of single-pass membrane proteins that homo-dimerize, assess the extent of G/A/SxxxG/A/S motif usage by homo-dimeric proteins, and determine the energetic strength and distribution of GAS_{right} homo-dimers.

We chose to predict the human genome as a source of wild-type sequences. We used the entire Uniprot annotated set of wild-type single-pass human TM domain sequences, approximately 2200 [Uniprot Consortium (2014)]. The sequences were predicted by the CATM algorithm and approximately half of the proteins returned models that CATM estimated have favorable dimerization energy, indicating that there is a large number of potential GAS_{right} structure in the genome that could be identified (see Table 4.1, 4.2 and 4.3).

4.3.2 Human TM proteins in a leucine background show a wide range of relative dimerization activity

The TOXCAT assay is not quantitative, but due to the ease of creating and assaying multiple samples, it is frequently used for screening TM domain oligomerization (Fig 4.2)[Russ, et al. (1999); LaPointe et al. (2013); Khadria, et al. (2014); Finger, et al. (2009)]. To further reduce the qualitative nature of the TOXCAT assay and focus our study on the prediction of the interaction interface, only the predicted interface residues were used in the assay. Previous work has shown that the interface residues of glycophorin A in a leucine residue background behave similar to

the wt sequence [Russ, et al. (1999)]. The eight residues that were predicted by CATM to be at the interface were stitched into a leucine background, with a total TM length of 18 residues, as seen in figure 4.3. The stitching solved a number of problems: the varying length of the TM domain, the position in the membrane of the crossing point of the helix, and the variance of insertion (essentially the hydrophobicity) of the helix.

Our previous work showed that the GAS_{right} motif is optimized for the formation of Cα–H···O=C hydrogen bonds. Therefore it is to be expected that GAS_{right} dimers form Cα–H···O=C hydrogen bonds, and to form Cα–H···O=C hydrogen bonds a glycine residue must be present near the point of closest approach. Specifically, the glycine residue is required at position C1. Position C1 is the closest residue to the point of closest approach on the right, C-terminal side (see figure 4.4). CATM predicts GAS_{right} structures, therefore our focus is on proteins containing glycine in their TM domain sequence. In an effort to further standardize the TOXCAT assay the C1 position was placed at position 12 in the TM sequence (fig. 4.3). All TM sequences picked had glycine at the C1 position, however not all were predicted to dimerize. About 55 TM domains were cloned into the TOXCAT vector, representing a wide range of dimerization energy scores, including TM domains not predicted to interact. When assayed in TOXCAT, the TM domains showed a wide range of relative dimerization scores as well.

4.3.3 Validation of the Structure Predictions Using Mutagenesis

CATM predicts GAS_{right} motifs, however the TOXCAT assay merely reports if the TM domain being assayed associates, therefore a check needs to be made for proper GAS_{right} dimer formation. Optimally, a full mutagenesis would be performed on all 8 interface positions to determine their sensitivity to mutation of all positions. However with 55 samples, a full mutagenesis is unrealistic. Therefore, knowing the importance of the glycine at the C1 position, this position became the mutagenesis target. A glycine to isoleucine mutation was performed at this position, this mutation in glycophorin A is the universal negative dimerization control of the TOXCAT assay. The mutation results in

the loss of Cα–H···O=C hydrogen bonds as well as resulting in steric clashes at the interface.

Currently 19 G to I mutations at C1 have been completed *in vivo* (see Table 4.4). All mutations were first analyzed by predicting the structure in CATM. Of the 19 G to I mutations, 16 completely eliminated dimer formation. The three TM domain mutants that didn't completely abolish dimer formation *in silico* all have glycine-zipper motifs, GxxxGxxxG, which can place alternate glycine residues at the interface and still associate via a GAS_{right} motif [Khadria, et al. (2014)]. Of the G to I mutations assayed via TOXCAT 18 of the 19 dimerize less strongly than the original construct, and 10 of the 19 have a relative dimerization score less than 50% of the original. An additional 5 of the mutants that don't half the dimerization of the original had a low dimerization score to begin with. The one mutant that had a higher dimerization score than the original (AVC1B), may dimerize via a different motif, and will be discussed in a later section.

4.3.4 Correlation of TOXCAT Data to CATM Scores

All 55 of the wild-type TM domain relative dimerization signal in TOXCAT were compared to their CATM dimerization score. The correlation coefficient of these two datasets was -0.41 (the TOXCAT score is positive, and the CATM score is negative), showing little correlation between the two datasets. The inclusion of 18 additional mutants (mutations predicted to have small, but noticeable effects on the dimerization of the proteins (see Table 4.7 and 4.8)) increased the correlation to -0.51 (see fig. 4.5A). This still shows little correlation between the two datasets. However, as shown by the poor performance of the Gly to Ile AVC1B mutant, not all proteins may be dimerizing via a GAS_{right} motif. While the GAS_{right} dimer is a common structure [MacKenzie, et al. (1997); Bocharov, et al. (2007); Bocharov, et al. (2008); Bocharov, et al. (2012)], TM domains have been shown to dimerize by other means. Often association can be driven by polar residues [(Choma, et al. (2000); Sal-Man, et al. (2004)]. Therefore the TM domains were broken into types based on their composition. If a TM domain contained even one residue of the listed type it was labeled that type. The listing is

hierarchical, if the TM contains even one amino acid of that type it is labeled as such, with the groups listed in order of precedence: “Charged” - “glutamate, aspartate, arginine, lysine, histidine”, “Strong Polar” - “glutamine, asparagine”, “Tyrosine” - “tyrosine”, “Small Polar” - “serine, threonine”, “Other” - “leucine, valine, isoleucine, alanine, glycine, tryptophan, cystine, methionine, phenylalanine, proline”. For instance if a TM domain contains both a glutamate and a serine, it is labeled “Charged” (see fig. 4.6).

Considering the apolar environment of the bilayer it is unlikely that strongly polar residues would have unsatisfied hydrogen bonds, dimers may form to satisfy these bonds, irrespective of the GAS_{right} motif residues. With only the “Other” TM domains considered, a dataset of 13 residues, the correlation coefficient is -0.88 between TOXCAT and CATM, seen in figure 4.5b. With the “Other” and “Small Polar” residues, a dataset of 29 residues the correlation coefficient is -0.65, and is -0.72 when the additional mutants are added (13 additional TM domains). Addition of the “Polar Tyrosine” domains lowers the coefficient to -0.67 (52 total domains), see figure 4.5c. The correlation coefficient of the “Charged” and “Strong Polar” are -0.03 and 0.07 respectively (figure 4.5d). Preliminary results suggest that relative dimerization of GAS_{right} mediated structures, that is TM domains not containing strong polar or charged residues, can be predicted by CATM.

4.3.5 Mutation of Charged and Strong Polar residues to hydrophobic counterparts

In analyzing the datasets both the “Charged” and “Strong Polar” TM domains do not correlate well between the CATM and TOXCAT scores, perhaps due to either improper insertion or due to polar to polar hydrogen bonds. Therefore to test to see if this is the case charged/polar residues were mutated to their closest hydrophobic counterpart. Mutations were first done in TOXCAT and all mutants had similar energies and structures to their wild-type counterpart. Currently all “Charged”/ “Strong Polar” mutant genes have been ordered from IDT, 18 in total. Of these 18, 8 have been successfully cloned and tested in TOXCAT (see Table 4.5 and 4.6). Of these eight

samples the correlation coefficient between CATM and TOXCAT improves from 0.43 (a good correlation will be negative) to -0.8. Considering half of the samples still need to be tested, we cannot conclude much about the data, however preliminary results look promising, see figure 4.7.

The results suggest that CATM can correctly predict GAS_{right} mediated dimer structure and relative dimerization, but cannot predict structures mediated by other types of interactions. This makes sense as CATM was developed to optimize Cα–H···O=C hydrogen bonds.

4.3.6 Relationship of TMs with High and Low Cα–H···O=C bond scores

Currently there are very few *in vivo/vitro* studies that measure the strength of the Cα–H···O=C hydrogen bond or its contribution to the dimerization of TM domains [Yohannan, et al. (2004); Arbely, et al. (2004)]. The CATM algorithm, along with predicting the structure of TM dimers, predicts the number and strength of Cα–H···O=C hydrogen bonds formed. The strength of the hydrogen bond being based on the SCWRL4 hydrogen bond term [Krivov, et al. (2009)], and adapted for Cα–H···O=C hydrogen bonds [Mueller, et al. (2014)]. The analysis of the human genome protein interfaces reveals that GAS_{right} protein dimers can have varying numbers and strengths of hydrogen bonds. We intend to use TOXCAT to measure the relative dimerization score difference between proteins with strong Cα–H···O=C hydrogen bonding and low Cα–H···O=C hydrogen bonding. If Cα–H···O=C hydrogen bonds are an important force in dimerization, protein dimers with strong hydrogen bonds should have higher relative dimerization scores than similar protein dimers.

Proteins were analyzed in CATM for sets of structures with similar van der Waals (vdw) and solvation energy scores, but with high Cα–H···O=C hydrogen bonding and low to non-existent Cα–H···O=C hydrogen bonding. Sixteen such structures were found, divided into three groups. The three groups comprising high, medium and low vdw plus solvation scores, these three groups were subdivided into two further groups, structures

with high and low Cα–H···O=C hydrogen bonding. Currently about half of the structures have been cloned and assayed via TOXCAT. Results are too preliminary to analyze in depth, but in the high vdw plus solvation TM dimer there existed a distinct difference between the high and low Cα–H···O=C hydrogen bond dimer relative dimerization scores.

4.4 Conclusions

Single-pass membrane proteins are commonly occurring proteins and are frequently found with G/A/SxxxG/A/S sequence motifs. TM proteins with G/A/SxxxG/A/S sequence motifs have often been shown to exist in GAS_{right} protein dimers. However the formation of this motif does not automatically confer strong dimerization potential, like the GAS_{right} standard glycophorin A [Russ, et al. (1999)]. As we have seen here, and shown in previous studies [Finger, et al. (2009); Kirrbach, et al. (2013)], there are a wide range of dimerization energies within the GAS_{right} motif, and it is important to understand how TM protein dimers behave. Considering the difficulty in both assaying TM proteins and determining their structure CATM becomes an important tool to understand protein behavior from sequence alone.

Currently the TOXCAT assay has been simplified by removing the non-interfacial residues, to hopefully eliminate variation in the ΔG of insertion in the membrane by the non-interacting residues. However we have found that even in this simplified system the GAS_{right} motif dimerization energy is quite varied.

The CATM algorithm can currently predict TM dimerization energy provided the sequences do not contain strongly polar or charged residues. The difficulty in predicting structures of this kind is most likely due to non-GAS_{right} mediated contacts, such as side-chain to side-chain hydrogen bonding. Structures with small polar residues, serine and threonine, do not predict as well as the “other” group, however out of the 42 dimers predicted (other + small polars), just two proteins (THS7A and TNR12) account for the poor correlation coefficient. With those two proteins removed the correlation coefficient of CATM to TOXCAT improves to -0.82, showing the strong prediction power of the algorithm.

The predictive capabilities of CATM when determining GAS_{right} structures is further highlighted by the, currently, limited dataset of strong polar / charged mutated to non-charged residues. When these residue types are eliminated, and presumably the

GAS_{right} structure established, CATM more closely predicts the association energetics of the dimer.

Using the structural prediction capabilities of CATM, we analyzed our database of structures to find similarly packed structures, but with different Cα–H···O=C hydrogen bond energies. With our preliminary results we have seen that there may be a difference in dimerization energies when Cα–H···O=C hydrogen bonds are eliminated from structures. If these results hold, it is a key piece of evidence that Cα–H···O=C hydrogen bonds are important for the dimerization of membrane proteins.

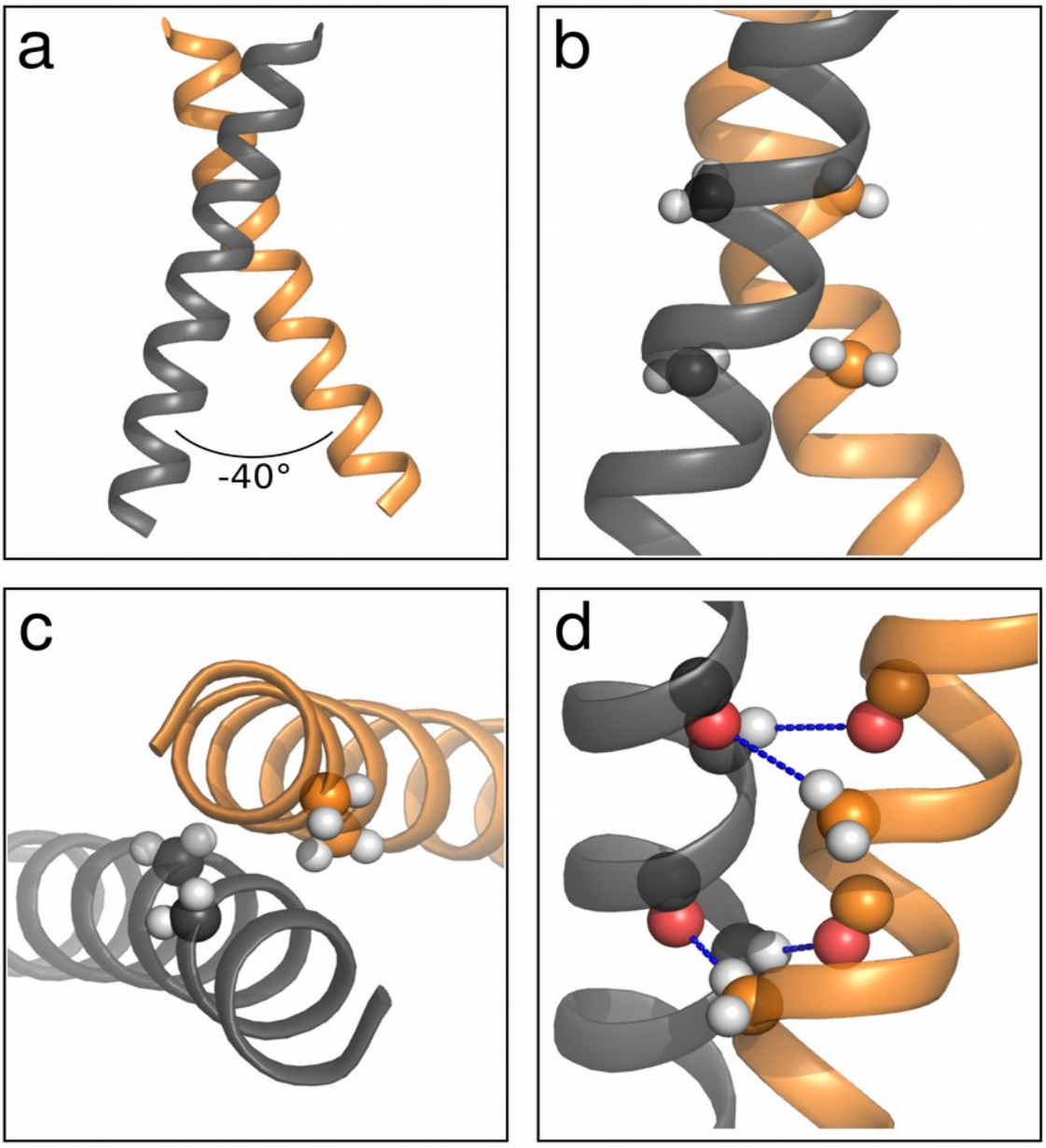


Fig. 4.1 The GAS_{right} motif. a) The GAS_{right} motif is the association of two parallel transmembrane helices at a right-handed crossing angle of around -40 degrees. b&c) The structural motif places Glycine, Alanine, or Serine (GAS) residues at the interface. In the figure the common GxxxG motif is placed at the interface, with the glycine residues placed at *i* and *i*+4, the two residues are at the same face of the helix. d) The

placement of the GxxxG motif in the GAS_{right} motif is optimized for the formation of interhelical Cα–H···O=C hydrogen bonds.

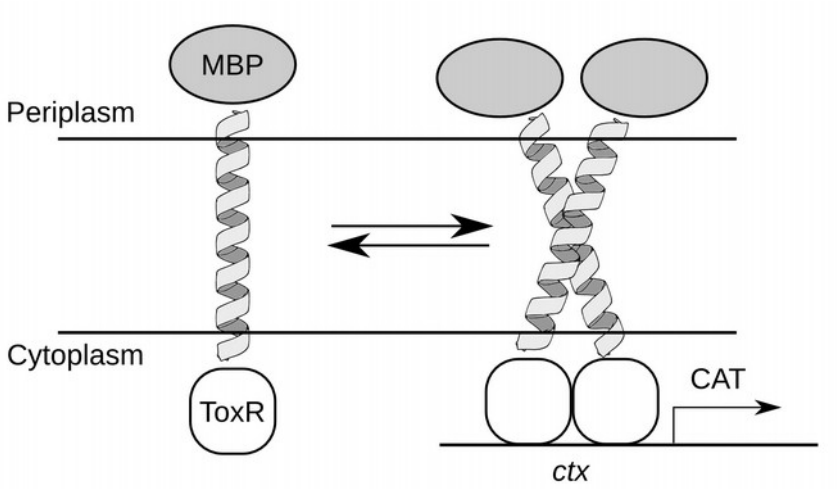


Fig. 4.2 TOXCAT Assay. TOXCAT is an in vivo assay based on a construct in which the transmembrane domain under investigation is fused to the ToxR transcriptional activator of V. cholerae. Transmembrane association results in the expression of a reporter gene in E. coli cells, which can be quantified.

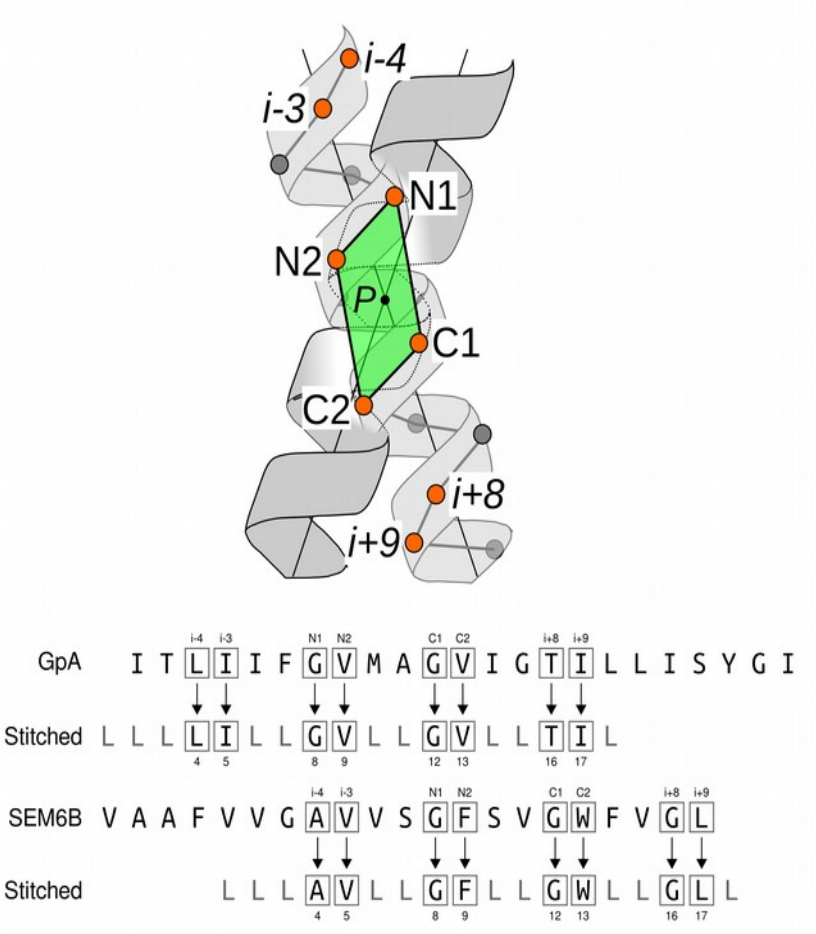


Fig. 4.3 Interface Residues Stitched into Leucine Residue Background. A schematic dimer is represented with the interface residues labeled. CATM predicts the point of closest approach between two helices, labeled as “P”, bounding this point are four residues in a parallelogram, labeled N1, N2, C1 and C2 (N residues on the N-terminus, and C residues on the C-terminus). The other residues at the interface are, relative to N1, *i-4, i-3, i+8,* and *i+9*. These interface residues are stitched into a leucine background of a standard length. For example, glycophorin A (GpA) is shown. Uniprot annotates the GpA TM domain as 23 residues, CATM places GVxxGV at the parallelogram positions. The interface is stitched into a TM domain composed of 18

positions, with the interface positions at 4, 5, 8, 9, 12, 13, 16 and 17 and leucine residues in the remaining positions. Regardless of the position of C1 in the wild-type sequence, it will always be placed at position 12 in the stitched sequence, to maintain a similar placement of the crossing point in the membrane. A further example shows the TM domain of SEM6B stitched into the leucine background. While the interface positions are more C-terminal than GpA, the C1 glycine residue is still placed at position 12.

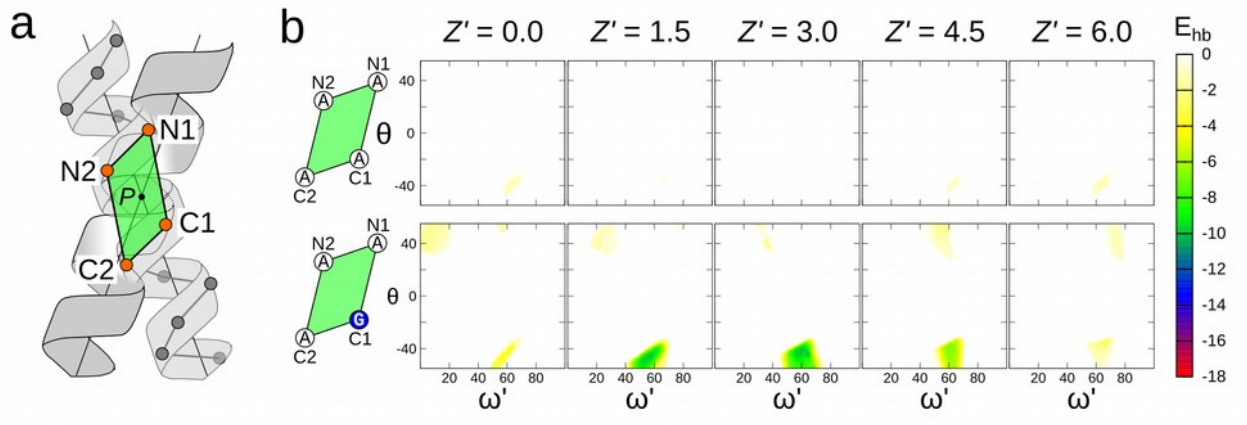


Fig. 4.4 Glycine must be present at Position C1. a) A helix dimer, with the point of closest approach labeled as “P”. Four residues bound this point of closest approach as a parallelogram, represented by their Cα atoms in the schematic. The two N-terminal positions are labeled as N1 and N2, the two C-terminal positions as C1 and C2. The C1 position is the critical residue for Cα–H···O=C hydrogen bond formation. b) A map of the carbon hydrogen bonding energy (E_{hb}, color bar) as a function of interhelical geometry (ω’: x axis, θ: y axis; Z’: panels). The amino acids at the interfacial positions (white circle A, Ala; blue circle G, Gly), are indicated in the unit cell schemes on the left. The top panels show a poly-Ala sequence, almost no hydrogen bond propensity is seen. In the bottom panels a glycine is introduced at the critical position, C1, into the poly-Ala helix. When the introduction of this single glycine a single broad minimum is observed centered around a region with a right-handed crossing angle θ of approximately −30° to −50°. In each panel, the hydrogen bond energy (E_{hb}) is plotted at the interhelical distance in which the overall energy is minimized. Further explanation can be found in *Mueller, et. al.* figures 2 & 3.

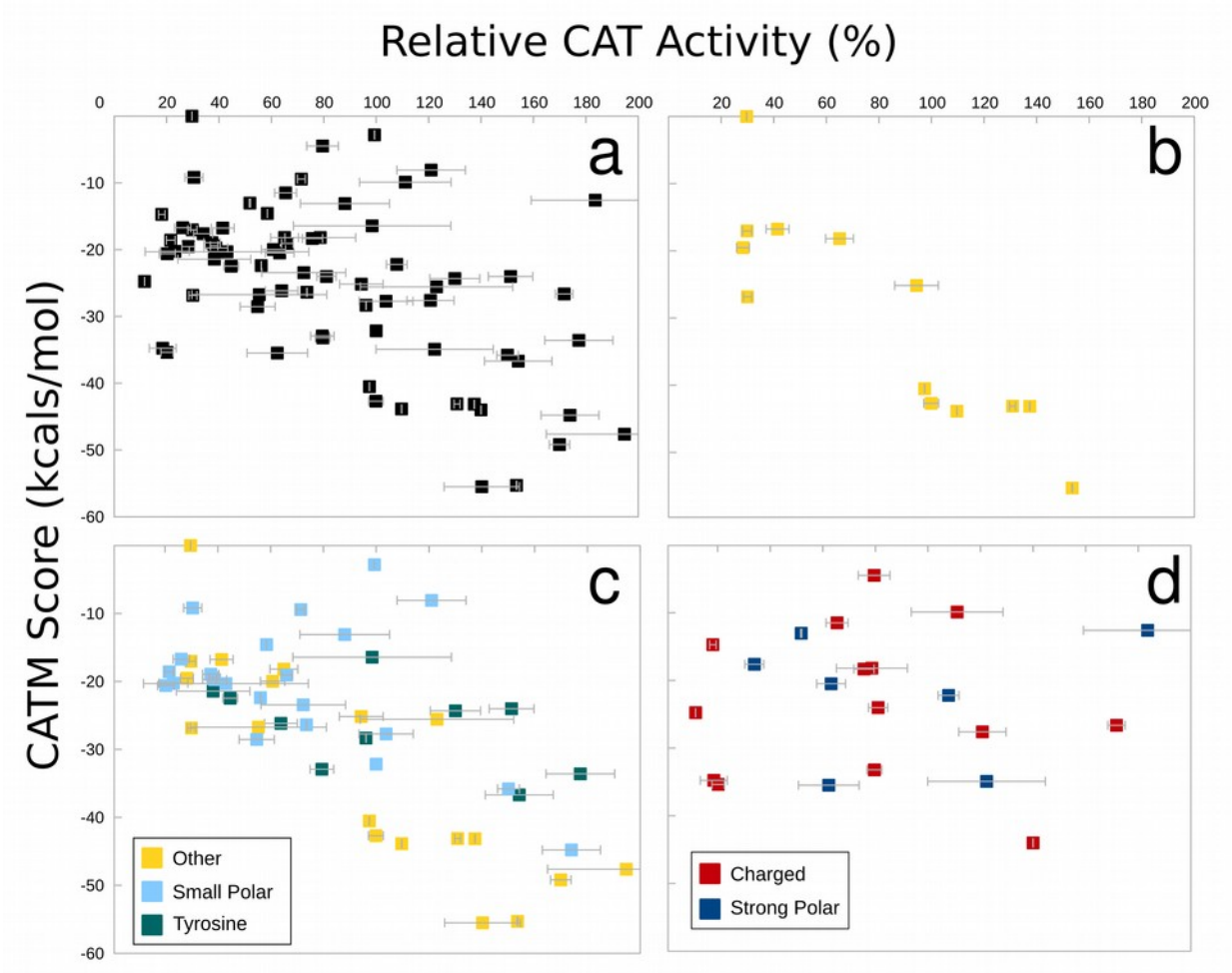


Fig. 4.5 Plots of Relative CAT Activity versus CATM Score. Relative CAT Activity as a percentage of GpA activity is plotted on the x-axis, with the CATM Score plotted on the y-axis. a) Plot of all 55 wt proteins, plus the additional 18 mutants, correlation coefficient of -0.51. b) Plot of the 13 “Other”-type wt proteins, correlation coefficient of -0.88. c) Plot of the “Other”-type, “Small Polar”-type, and “Tyrosine” proteins (all with wt + mutants). “Other” + “Small Polar” correlation coefficient of -0.72, “Other” + “Small Polar” + “Tyrosine” correlation coefficient of -0.67. d) Plot of “Charged” and “Strong Polar”-type proteins, correlation coefficient of -0.03 and 0.07 respectively.

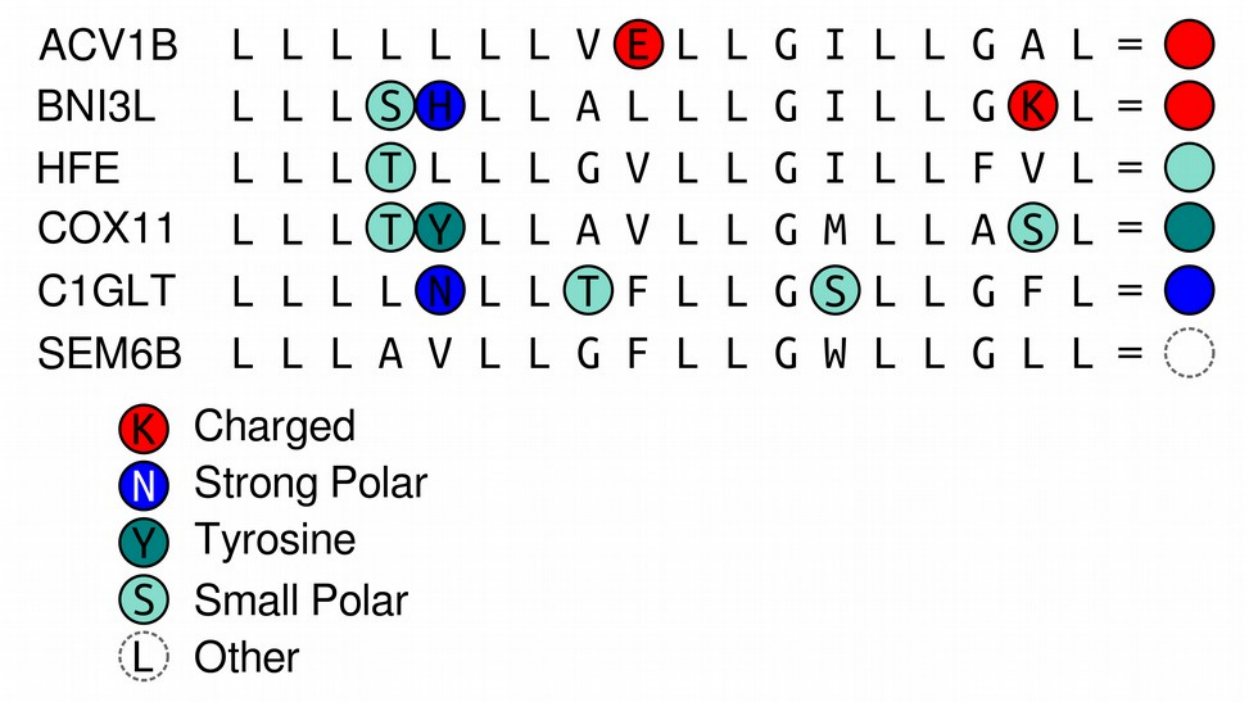


Fig. 4.6 TM Domain Sequence Label. Shown here are representative sequences, and the type of protein the sequence is labeled as. The key shows the order of the hierarchy, “Charged” with the most precedence, followed by “Strong Polar”, “Tyrosine”, “Small Polar” and “Other”. For example, ACV1B contains one charged residue, glutamate, and therefore is labeled as “Charged”. BNI3L contains serine, histidine and lysine, the charged residue lysine takes precedence and is labeled as “Charged”. SEM6B contains no charged, strong polar, tyrosine or small polar and therefore is labeled “Other”.

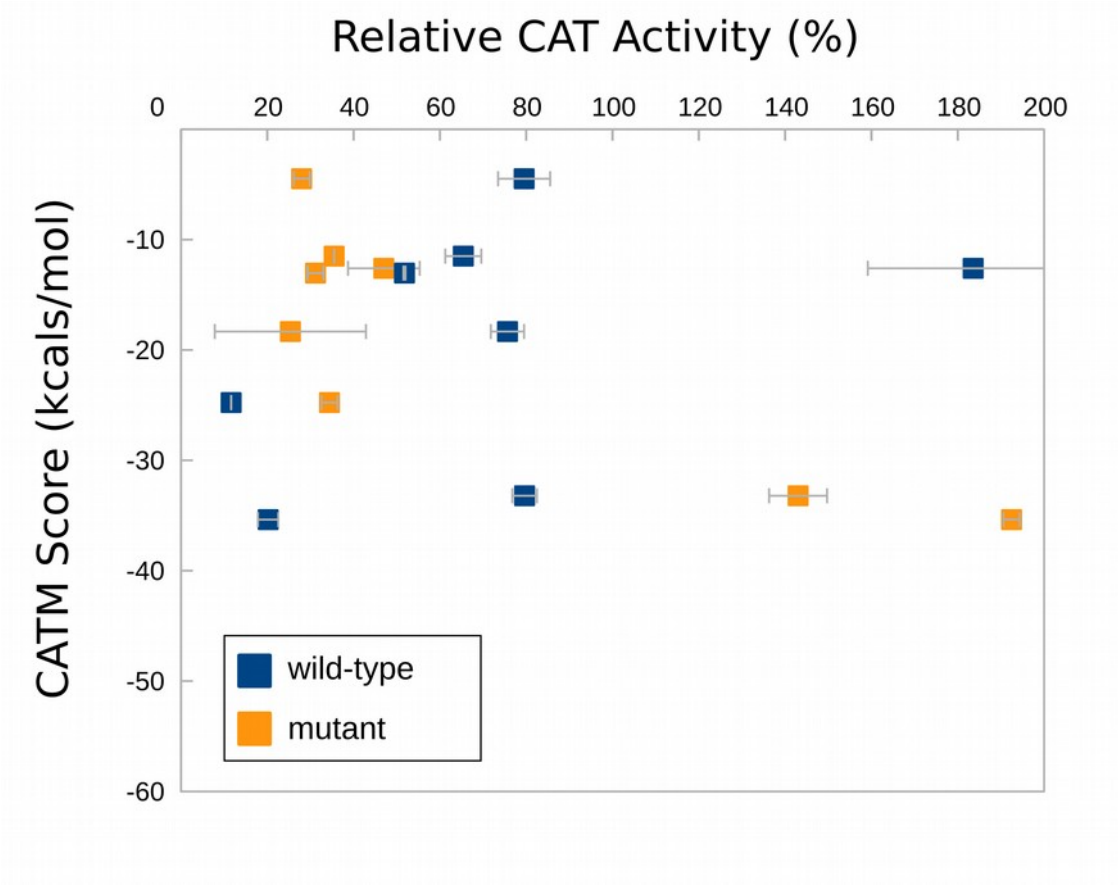


Fig. 4.7 Comparison of wt to Mutant Polar/Charged TM Domains. Mutation of "Charged" and "Strong Polar"-type proteins to their closest hydrophobic counterpart. Mutations are glutamate to methionine, arginine to methionine, aspartate to leucine, asparagine to leucine, and glutamine to methionine. Relative CAT Activity as a percentage of GpA activity is plotted on the x-axis, with the CATM Score plotted on the y-axis. The wild-type proteins have a correlation coefficient between CATM and TOXCAT of 0.43 (a good correlation will be negative). The mutations improve this correlation to -0.8.

Table 4.1 List of Cloned TM Domains.

List of Uniprot annoted TM domains, ranked by dimerization energy as calculated by the previous version of the CATM algorithm (see *Mueller, et. al.*). Both the Uniprot name and Accession Number are listed. Also listed is whether the TM domain was cloned into the TOXCAT vector. If the domain was not cloned, the reason is listed. As proline is not modeled by CATM, TM domains with proline residues at the interface were not cloned.

Past CATM Rank	Accession Number	Name	Cloned
1	Q96D53	ADCK4	Yes
2	P17342	ANPRC	Yes
3	Q9H3T3	SEM6B	Yes
4	Q9NUV7	SPTC3	Yes
5	Q6P4H8	F173B	Yes
6	P57679	EVC	did not clone properly
7	Q9NS00	C1GLT	Yes
8	Q9NP84	TNR12	Yes
9	Q15904	VAS1	Yes
10	Q9P2E5	CHPF2	PRO AT INTERFACE
11	P60602	ROMO1	Yes
12	Q5TGZ0	MOS1	did not clone properly
13	P56962	STX17	Yes
14	Q96EU7	C1GLC	did not clone properly
15	Q9NY15	STAB1	Yes
16	Q9BTM6	ARM10	Yes
17	Q9NYM9	BET1L	Yes
18	Q16651	PRSS8	PRO AT INTERFACE
19	Q86VU5	CMTD1	did not clone properly
20	B6SEH8	ERVV1	Yes
21	Q7L4S7	ARMX6	Yes
22	Q13740	CD166	Yes
23	Q6P7N7	TMM81	Yes
24	Q9Y5Y7	LYVE1	did not clone properly
25	Q6ZS62	COLC1	did not clone properly
26	Q9UKU0	ACSL6	Yes
27	Q86X52	CHSS1	Yes
28	Q9C091	GRB1L	did not clone properly
29	Q6UW88	EPGN	Yes
30	Q9P1Z9	CC180	did not clone properly
31	P10314	1A32	Yes
32	P30459	1A74	interface residues same as 1A32
33	O60238	BNI3L	Yes
34	O60313	OPA1	Yes
35	Q96I36	COX14	Yes
36	P15509	CSF2R	Yes
37	Q8N387	MUC15	Yes
38	A6NL88	SHSA7	Yes
39	Q12983	BNIP3	Yes
40	Q6UXN7	TO20L	Yes
41	Q9UHR8	MCL1	Yes
42	Q9BXM7	PINK1	did not clone properly
43	Q6UXE8	BTNL3	did not clone properly
44	O43557	TNF14	did not clone properly
45	O14788	TNF11	Yes
46	Q9NP80	PLPL8	did not clone properly

47	Q8N6P7	I22R1	Yes
48	O75354	ENTP6	Yes
49	Q6UX72	B3GN9	Yes
50	Q8N326	CJ111	PRO AT INTERFACE
51	Q9UH62	ARMX3	did not clone properly
52	Q9NVM1	EVA1B	Yes
53	Q8TAY3	FOLH1	did not clone properly
54	Q8N3T1	GLT15	Yes
55	Q3V5L5	MGT5B	Yes
56	Q9UJQ1	LAMP5	PRO AT INTERFACE
57	Q13586	STIM1	Yes
58	Q9Y6N1	COX11	Yes
59	Q6GPH6	IPIL1	Yes
60	Q86XE3	MICU3	did not clone properly
61	Q08ET2	SIG14	Yes
62	Q9TP01	DRB3	did not clone properly
63	Q99420	BT3A1	PRO AT INTERFACE
64	Q8NCR0	B3GL2	PRO AT INTERFACE
65	Q9UQU7	1A31	Yes
66	Q9BQD7	F173A	Yes
67	Q9H3N1	TMX1	Yes
…			
709	P20333	TNR1B	Yes
…			
735	O15197	EPHB6	Yes
…			
786	Q59FL7	NCAM1	Yes
…			
821	Q9H0Y8	PTPRT	Yes
…			
1088	Q9H3R2	MUC13	Yes
…			
1093	Q9UPZ6	THS7A	Yes
…			
1120	Q96KU8	HFE	Yes
…			
1208	Q8N967	LRTM2	Yes
…			
1236	P36896	ACV1B	Yes
…			
Unranked	Q6UX71	PXDC2	Yes

Table 4.2 List of TM Domain Assay Results.

List of TM domains, ranked by dimerization energy as calculated by the previous version of the CATM algorithm (see *Mueller, et. al.*). The cells containing the TM domain of interest can either grow, not grow, or have minimal growth on maltose; growth on maltose indicates proper insertion (see Methods). The TOX Avg, is the average TOXCAT dimerization signal as a percentage of the standard glycophorin A (see Methods), also listed is the standard deviation of the signal. Finally listed is the dimerization energy as calculated by the current version of the CATM algorithm.

Past CATM Rank	Accession Number	Name	Maltose Growth	TOX Avg	Stdev	Current CATM Energy
1	Q96D53	ADCK4	Min	107.8	4.0	-22.21
2	P17342	ANPRC	Yes	37.3	3.1	-18.95
3	Q9H3T3	SEM6B	Yes	109.7	0.0	-43.87
4	Q9NUV7	SPTC3	Yes	96.1	0.1	-28.34
5	Q6P4H8	F173B	Yes	34.0	3.6	-17.62
…						
7	Q9NS00	C1GLT	Min	183.6	24.4	-12.60
8	Q9NP84	TNR12	Yes	121.0	13.1	-8.07
9	Q15904	VAS1	Min	174.0	11.0	-44.80
…						
11	P60602	ROMO1	Yes	153.5	0.1	-55.30
…						
13	P56962	STX17	Min	140.0	0.0	-44.01
…						
15	Q9NY15	STAB1	Yes	97.4	0.0	-40.53
16	Q9BTM6	ARM10	Min	18.2	1.2	-14.72
17	Q9NYM9	BET1L	Min	20.2	2.4	-35.38
…						
20	B6SEH8	ERVV1	Yes	28.3	2.4	-19.53
21	Q7L4S7	ARMX6	Yes	79.5	6.0	-4.48
22	Q13740	CD166	Yes	18.5	5.1	-34.77
23	Q6P7N7	TMM81	Yes	137.4	0.1	-43.15
…						
26	Q9UKU0	ACSL6	Min	88.1	17.0	-13.10
27	Q86X52	CHSS1	Yes	150.1	4.1	-35.81
…						
29	Q6UW88	EPGN	No	51.2		
…						
31	P10314	1A32	Yes	94.3	8.3	-25.16
…						
33	O60238	BNI3L	Min	171.7	3.3	-26.63
34	O60313	OPA1	Yes	79.4	4.5	-32.92
35	Q96I36	COX14	Yes	20.3	3.1	-20.61
36	P15509	CSF2R	Min	79.6	2.9	-33.21
37	Q8N387	MUC15	Yes	99.9	2.8	-42.70
38	A6NL88	SHSA7	Yes	56.1	0.0	-22.39
39	Q12983	BNIP3	Yes	120.7	9.0	-27.60
40	Q6UXN7	TO20L	Yes	30.0	1.5	-26.85
41	Q9UHR8	MCL1	Yes	44.7	2.8	-22.47
…						
45	O14788	TNF11	Yes	62.3	11.5	-35.49
…						
47	Q8N6P7	I22R1	Yes	38.2	13.9	-21.42
48	O75354	ENTP6	Min	151.3	8.5	-24.02
49	Q6UX72	B3GN9	Min	65.4	4.2	-11.50
…						

52	Q9NVM1	EVA1B	Yes	154.2	12.9	-36.69
…						
54	Q8N3T1	GLT15	Yes	29.8	1.7	-17.03
55	Q3V5L5	MGT5B	Yes	66.1	2.2	-19.06
…						
57	Q13586	STIM1	Yes	23.2	5.5	-20.25
58	Q9Y6N1	COX11	Yes	177.3	13.0	-33.59
59	Q6GPH6	IPIL1	Yes	75.6	3.9	-18.33
…						
61	Q08ET2	SIG14	Yes	73.5	0.1	-26.37
…						
65	Q9UQU7	1A31	Yes	65.0	5.3	-18.18
66	Q9BQD7	F173A	Yes	78.7	13.5	-18.19
67	Q9H3N1	TMX1	Yes	30.5	3.4	-9.18
…						
674	P02724	GLPA	Yes	96.6	3.6	-32.17
…						
709	P20333	TNR1B	Yes	54.7	6.7	-28.54
…						
735	O15197	EPHB6	Yes	38.8	4.5	-19.73
…						
786	Q59FL7	NCAM1	Yes	130.8	1.1	-43.11
…						
821	Q9H0Y8	PTPRT	Yes	11.6	0.0	-24.77
…						
1088	Q9H3R2	MUC13	Yes	51.8	0.2	-13.03
…						
1093	Q9UPZ6	THS7A	Yes	99.4	0.0	-2.82
…						
1120	Q96KU8	HFE	Yes	72.4	16.0	-23.44
…						
1208	Q8N967	LRTM2	Yes	41.4	4.4	-16.76
…						
1236	P36896	ACV1B	Yes	111.1	17.5	-9.89
…						
Unranked	Q6UX71	PXDC2	Yes	29.7	0.0	0.00

Table 4.3 List of TM Domain Type and Sequence.

List of the cloned TM domains, ranked by dimerization energy as calculated by the previous version of the CATM algorithm (see *Mueller, et. al.*). Shown is the TM domain type (see Fig. 4.6) and the TM sequence cloned into the TOXCAT vector. Residues in bold are the wild-type interface residues (see Fig. 4.3).

Past CATM Rank	Name	Type	Sequence			4	5			8	9			12	13			16	17	
1	ADCK4	Strn P	L	L	L	I	S	L	L	A	N	L	L	G	L	L	L	G	L	L
2	ANPRC	Sm P	L	L	L	L	L	L	L	S	A	L	L	G	I	L	L	G	A	L
3	SEM6B	Other	L	L	L	A	V	L	L	G	F	L	L	G	W	L	L	G	L	L
4	SPTC3	Tyr	L	L	L	F	T	L	L	G	Y	L	L	G	T	L	L	G	Y	L
5	F173B	Strn P	L	L	L	L	N	L	L	F	L	L	L	G	L	L	L	G	T	L
…																				
7	C1GLT	Strn P	L	L	L	L	N	L	L	T	F	L	L	G	S	L	L	G	F	L
8	TNR12	Sm P	L	L	L	A	L	L	L	T	F	L	L	G	L	L	L	G	F	L
9	VAS1	Sm P	L	L	L	F	F	L	L	G	I	L	L	G	L	L	L	S	L	L
…																				
11	ROMO1	Other	L	L	L	G	F	L	L	G	C	L	L	G	M	L	L	G	A	L
…																				
13	STX17	Chrg	L	L	L	A	A	L	L	G	G	L	L	G	F	L	L	G	K	L
…																				
15	STAB1	Other	L	L	L	V	L	L	L	G	A	L	L	G	L	L	L	G	A	L
16	ARM10	Chrg	L	L	L	R	G	L	L	W	V	L	L	G	L	L	L	G	A	L
17	BET1L	Chrg	L	L	L	K	L	L	L	G	M	L	L	G	L	L	L	A	F	L
…																				
20	ERVV1	Other	L	L	L	L	G	L	L	L	A	L	L	G	A	L	L	G	M	L
21	ARMX6	Chrg	L	L	L	R	E	L	L	W	M	L	L	G	L	L	L	G	A	L
22	CD166	Chrg	L	L	L	K	L	L	L	G	I	L	L	G	L	L	L	A	A	L
23	TMM81	Other	L	L	L	A	L	L	L	G	I	L	L	G	V	L	L	G	V	L
…																				
26	ACSL6	Sm P	L	L	L	S	A	L	L	L	V	L	L	G	A	L	L	A	I	L
27	CHSS1	Sm P	L	L	L	S	V	L	L	G	L	L	L	G	F	L	L	A	S	L
…																				
29	EPGN	Tyr	L	L	L	L	Y	L	L	Y	I	L	L	G	I	L	L	G	L	L
…																				
31	1A32	Other	L	L	L	V	L	L	L	A	M	L	L	G	A	L	L	A	A	L
…																				
33	BNI3L	Chrg	L	L	L	S	H	L	L	A	L	L	L	G	I	L	L	G	K	L
34	OPA1	Tyr	L	L	L	Y	L	L	L	G	S	L	L	G	G	L	L	T	A	L
35	COX14	Sm P	L	L	L	S	M	L	L	L	T	L	L	G	G	L	L	C	S	L
36	CSF2R	Chrg	L	L	L	T	L	L	L	G	I	L	L	G	F	L	L	K	R	L
37	MUC15	Other	L	L	L	G	I	L	L	G	A	L	L	G	A	L	L	G	V	L
38	SHSA7	Sm P	L	L	L	S	F	L	L	A	V	L	L	G	A	L	L	A	F	L
39	BNIP3	Chrg	L	L	L	S	H	L	L	A	I	L	L	G	I	L	L	G	R	L
40	TO20L	Other	L	L	L	L	L	L	L	A	A	L	L	G	A	L	L	F	L	L
41	MCL1	Tyr	L	L	L	A	F	L	L	V	A	L	L	G	A	L	L	A	Y	L
…																				
45	TNF11	Strn P	L	L	L	V	A	L	L	G	L	L	L	G	Q	L	L	C	S	L
…																				
47	I22R1	Tyr	L	L	L	L	T	L	L	Y	S	L	L	G	A	L	L	F	S	L
48	ENTP6	Tyr	L	L	L	V	A	L	L	A	Y	L	L	G	L	L	L	G	V	L
49	B3GN9	Chrg	L	L	L	D	A	L	L	T	L	L	L	G	A	L	L	G	L	L

…			
52	EVA1B	Tyr	L L L **Y** **F** L L **G** **V** L L **G** **L** L L **T** **L** L
…			
54	GLT15	Other	L L L **F** **L** L L **L** **L** L L **G** **C** L L **M** **M** L
55	MGT5B	Sm P	L L L **L** **C** L L **M** **T** L L **G** **G** L L **S** **A** L
…			
57	STIM1	Sm P	L L L **V** **S** L L **I** **G** L L **G** **C** L L **A** **Y** L
58	COX11	Tyr	L L L **T** **Y** L L **A** **V** L L **G** **M** L L **A** **S** L
59	IPIL1	Chrg	L L L **L** **L** L L **E** **G** L L **G** **W** L L **G** **N** L
…			
61	SIG14	Sm P	L L L **T** **L** L L **G** **A** L L **G** **A** L L **L** **L** L
…			
65	1A31	Other	L L L **V** **L** L L **A** **V** L L **G** **A** L L **A** **A** L
66	F173A	Chrg	L L L **L** **E** L L **Q** **A** L L **G** **S** L L **A** **A** L
67	TMX1	Sm P	L L L **F** **A** L L **T** **L** L L **G** **L** L L **G** **L** L
…			
674	GLPA	Sm P	L L L **L** **I** L L **G** **V** L L **G** **V** L L **T** **I** L
…			
709	TNR1B	Sm P	L L L **T** **G** L L **A** **L** L L **G** **L** L L **G** **V** L
…			
735	EPHB6	Sm P	L L L **L** **L** L L **S** **L** L L **G** **S** L L **G** **A** L
…			
786	NCAM1	Other	L L L **G** **L** L L **G** **A** L L **G** **I** L L **V** **I** L
…			
821	PTPRT	Chrg	L L L **V** **K** L L **G** **V** L L **G** **L** L L **F** **I** L
…			
1088	MUC13	Strn P	L L L **F** **Q** L L **L** **T** L L **G** **T** L L **G** **I** L
…			
1093	THS7A	Sm P	L L L **L** **L** L L **T** **W** L L **G** **V** L L **G** **A** L
…			
1120	HFE	Sm P	L L L **T** **L** L L **G** **V** L L **G** **I** L L **F** **V** L
…			
1208	LRTM2	Other	L L L **V** **I** L L **G** **V** L L **G** **V** L L **I** **M** L
…			
1236	ACV1B	Chrg	L L L **L** **L** L L **V** **E** L L **G** **I** L L **G** **A** L
…			
Unrnk	PXDC2	Other	L L L **L** **L** L L **G** **L** L L **G** **I** L L **L** **V** L

Table 4.4 List of Gly to Ile Mutations.

List of all the cloned glycine to isoleucine mutations at position C1 (see Fig. 4.3, 4.4 and Section 4.3.3). The average and standard deviation of the TOXCAT dimerization signal is shown as a percentage of the standard glycophorin A (see Methods). The dimerization energy is calculated by the current version of the CATM algorithm, for most TM domains the score will be zero, as the G to I mutation will destroy the dimer.

Past CATM Rank	G C1 I Mut Name	TOX Avg	Stdev	Current CATM Energy
…				
2	ANPRC	27.3	1.6	0.0
…				
9	VAS1	29.2	4.8	0.0
…				
11	ROMO1	69.5	60.9	-24.3
…				
13	STX17	14.8	8.2	-21.2
…				
21	ARMX6	48.9	4.3	0.0
…				
26	ACSL6	27.5	4.3	0.0
…				
27	CHSS1	18.2	5.0	0.0
…				
31	1A32	21.9	5.1	0.0
…				
36	CSF2R	9.9	2.9	0.0
…				
41	MCL1	20.4	4.1	0.0
…				
45	TNF11	36.5	3.8	0.0
…				
48	ENTP6	78.8	2.4	0.0
49	B3GN9	39.4	2.9	0.0
…				
674	GLPA	26.6	9.7	0.0
…				
709	TNR1B	30.5	3.0	0.0
…				
786	NCAM1	38.2	1.8	-26.7
…				
1088	MUC13	50.4	15.3	0.0
…				
1120	HFE	60.3	5.2	0.0
…				
1236	ACV1B	123.2	1.8	0.0

Table 4.5 List of Strong Polar / Charged Mutant Assay Results.

Lists all TM domains labeled as “Charged” or “Strong Polar” (see Fig. 4.6) and the mutations to their hydrophobic counterparts (see Section 4.3.5). The cells containing the TM domain of interest can either grow, not grow, or have minimal growth on maltose; growth on maltose indicates proper insertion (see Methods). The average and standard deviation of the TOXCAT dimerization signal is shown a percentage of the standard glycophorin A (see Methods). Finally listed is the dimerization energy of the wild-type and mutated version of the TM domain as calculated by the current version of the CATM algorithm.

Past CATM Rank	Name	Mutation	Cloned	Maltose Growth	TOX Avg	Stdev	CATM Energy	CATM Mut Energy
7	C1GLT	N5L	yes	yes	47.0	8.3	-12.6	-13.3
17	BET1L	K4M	yes	yes	192.5	1.9	-35.4	-46.3
21	ARMX6	R4M/E5M	yes	yes	28.0	1.9	-4.5	-24.6
36	CSF2R	K16M/R17M	yes	yes	143.0	6.7	-33.2	-40.5
49	B3GN9	D4L	yes	yes	35.5	0.0	-11.5	-14.0
59	IPILI	E8M/N17L	yes	yes	25.3	17.5	-18.4	-22.6
821	PTPRT	K5M	yes	yes	34.4	1.9	-24.8	-21.9
1088	MUC13	Q5M	yes	yes	31.2	1.9	-13.0	-16.1
1	ADCK4	N9L	yes	no	23.7	1.9	-22.2	-23.3
22	CD166	K4M	yes	no	178.5	7.4	-34.8	-45.6
33	BNI3L	K17M	yes	no	287.6	103.8	-26.6	-22.0
39	BNIP3	R17M	yes	no	257.1	65.6	-27.6	-23.6
5	F173B	N5L	no				-17.6	-13.4
13	STX17	K17M	no				-44.0	-42.1
16	ARM10	R4M	no				-14.7	-24.0
45	TNF11	Q13M	no				-35.5	-35.0
66	F173A	E5M / Q8M	no				-18.2	-22.9
1236	ACV1B	E9M	no				-9.9	-11.7

Table 4.6 List of Strong Polar / Charged Mutant Sequences.

Lists all TM domains labeled as “Charged” or “Strong Polar” (see Fig. 4.6) and the mutations to their hydrophobic counterparts (see Section 4.3.5). Shown is the TM domain cloned into the TOXCAT vector, highlighted are the mutated residues.

Past CATM Rank	Name	Mutation	Sequence																	
						4	5			8	9			12	13			16	17	
1	ADCK4	N9L	L	L	L	I	S	L	L	A	L	L	L	G	L	L	L	G	L	L
5	F173B	N5L	L	L	L	L	L	L	L	F	L	L	L	G	L	L	L	G	T	L
7	C1GLT	N5L	L	L	L	L	L	L	L	T	F	L	L	G	S	L	L	G	F	L
13	STX17	K17M	L	L	L	A	A	L	L	G	G	L	L	G	F	L	L	G	M	L
16	ARM10	R4M	L	L	L	M	G	L	L	W	V	L	L	G	L	L	L	G	A	L
17	BET1L	K4M	L	L	L	M	L	L	L	G	M	L	L	G	L	L	L	A	F	L
21	ARMX6	R4M/E5M	L	L	L	M	M	L	L	W	M	L	L	G	L	L	L	G	A	L
22	CD166	K4M	L	L	L	M	L	L	L	G	I	L	L	G	L	L	L	A	A	L
33	BNI3L	K17M	L	L	L	S	H	L	L	A	L	L	L	G	I	L	L	G	M	L
36	CSF2R	K16M/R17M	L	L	L	T	L	L	L	G	I	L	L	G	F	L	L	M	M	L
39	BNIP3	R17M	L	L	L	S	H	L	L	A	I	L	L	G	I	L	L	G	M	L
45	TNF11	Q13M	L	L	L	V	A	L	L	G	L	L	L	G	M	L	L	C	S	L
49	B3GN9	D4L	L	L	L	L	A	L	L	T	L	L	L	G	A	L	L	G	L	L
59	IPILI	E8M/N17L	L	L	L	L	L	L	L	M	G	L	L	G	W	L	L	G	L	L
66	F173A	E5M/Q8M	L	L	L	L	M	L	L	M	A	L	L	G	S	L	L	A	A	L
821	PTPRT	K5M	L	L	L	V	M	L	L	G	V	L	L	G	L	L	L	F	I	L
1088	MUC13	Q5M	L	L	L	F	M	L	L	L	T	L	L	G	T	L	L	G	I	L
1236	ACV1B	E9M	L	L	L	L	L	L	L	V	M	L	L	G	I	L	L	G	A	L

Table 4.7 List of Additional Mutations Assay Results.

Lists all of the additional mutations performed to the wild-type sequence, ranked by the wild-type dimerization energy as calculated by the previous version of the CATM algorithm (see *Mueller, et. al.*). The average and standard deviation of the TOXCAT dimerization signal is shown a percentage of the standard glycophorin A (see Methods). Finally listed is the dimerization energy of the mutated version of the TM domain as calculated by the current version of the CATM algorithm.

Past CATM Rank	Name	Mutation	TOX Avg	Stdev	CATM Energy
2	ANPRC	A9S	21.6	1.7	-18.6
3	SEM6B	A4V	60.8	4.6	-20.0
4	SPTC3	G8S	98.5	30.0	-16.4
7	C1GLT	T9A	63.2	5.5	-20.5
9	VAS1	S16I	55.4	25.7	-26.7
11	ROMO1	A17L	140.3	14.4	-55.5
15	STAB1	G8S	26.2	3.1	-16.7
26	ACSL6	L8F	38.1	1.4	-19.4
29	EPGN	Y8A	63.9	6.2	-26.2
31	1A32	V4L	123.0	29.1	-25.6
37	MUC15	A9L/A13L	169.9	3.9	-49.2
38	SHSA7	A8S	58.5	0.0	-14.6
45	TNF11	S17L	122.3	22.4	-34.9
48	ENTP6	V19A	130.0	9.5	-24.3
786	NCAM1	A9L	194.7	29.8	-47.6
1093	THS7A	L5V	71.4	1.5	-9.4
“	“	V13A	43.1	31.2	-20.3
1208	LRTM2	I16T	103.8	10.3	-27.7
1236	ACV1B	I13L	106.0	6.7	-12.9
“	“	L4A/V8A	81.0	3.7	-24.0

Table 4.8 List of Additional Mutation Sequences.

Lists all the additional TM domains mutations. Shown is the TM domain cloned into the TOXCAT vector, highlighted are the mutated residues.

Past CATM Rank	Name	Mutation	Sequence																	
						4	5			8	9			12	13			16	17	
2	ANPRC	A9S	L	L	L	L	L	L	L	S	S	L	L	G	I	L	L	G	A	L
3	SEM6B	A4V	L	L	L	V	V	L	L	G	F	L	L	G	W	L	L	G	L	L
4	SPTC3	G8S	L	L	L	F	T	L	L	S	Y	L	L	G	T	L	L	G	Y	L
7	C1GLT	T9A	L	L	L	L	N	L	L	T	A	L	L	G	S	L	L	G	F	L
9	VAS1	S16I	L	L	L	F	F	L	L	G	I	L	L	G	L	L	L	I	L	L
11	ROMO1	A17L	L	L	L	G	F	L	L	G	C	L	L	G	M	L	L	G	L	L
15	STAB1	G8S	L	L	L	V	L	L	L	S	A	L	L	G	L	L	L	G	A	L
26	ACSL6	L8F	L	L	L	S	A	L	L	F	V	L	L	G	A	L	L	A	I	L
29	EPGN	Y8A	L	L	L	L	Y	L	L	A	I	L	L	G	I	L	L	G	L	L
31	1A32	V4L	L	L	L	L	L	L	L	A	M	L	L	G	A	L	L	A	A	L
37	MUC15	A9L/A13L	L	L	L	G	I	L	L	G	L	L	L	G	L	L	L	G	V	L
38	SHSA7	A8S	L	L	L	S	F	L	L	S	V	L	L	G	A	L	L	A	F	L
45	TNF11	S17L	L	L	L	V	A	L	L	G	L	L	L	G	Q	L	L	C	L	L
48	ENTP6	V19A	L	L	L	V	A	L	L	A	Y	L	L	G	L	L	L	G	A	L
786	NCAM1	A9L	L	L	L	G	L	L	L	G	L	L	L	G	I	L	L	V	I	L
1093	THS7A	L5V	L	L	L	L	V	L	L	T	W	L	L	G	V	L	L	G	A	L
“	“	V13A	L	L	L	L	L	L	L	T	W	L	L	G	A	L	L	G	A	L
1208	LRTM2	I16T	L	L	L	V	I	L	L	G	V	L	L	G	V	L	L	T	M	L
1236	ACV1B	I13L	L	L	L	L	L	L	L	V	E	L	L	G	L	L	L	G	A	L
“	“	L4A/V8A	L	L	L	A	L	L	L	A	E	L	L	G	I	L	L	G	A	L

4.5 References

Arbely, Eyal, and Isaiah T. Arkin. "Experimental Measurement of the Strength of a Cα−H···O Bond in a Lipid Bilayer." *Journal of the American Chemical Society* 126.17 (2004): 5362-363.

Arnaout, M.A., B. Mahalingam, and J.-P. Xiong. "Integrin Structure, Allostery, And Bidirectional Signaling." *Annual Review of Cell and Developmental Biology* 21.1 (2005): 381-410.

Bocharov, E. V., Y. E. Pustovalova, K. V. Pavlov, P. E. Volynsky, M. V. Goncharuk, Y. S. Ermolyuk, D. V. Karpunin, A. A. Schulga, M. P. Kirpichnikov, R. G. Efremov, I. V. Maslennikov, and A. S. Arseniev. "Unique Dimeric Structure of BNip3 Transmembrane Domain Suggests Membrane Permeabilization as a Cell Death Trigger." *Journal of Biological Chemistry* 282.22 (2007): 16256-6266.

Bocharov, Eduard V., Konstantin S. Mineev, Marina V. Goncharuk, and Alexander S. Arseniev. "Structural and Thermodynamic Insight into the Process of “weak” Dimerization of the ErbB4 Transmembrane Domain by Solution NMR." *Biochimica Et Biophysica Acta (BBA) - Biomembranes* 1818.9 (2012): 2158-170.

Bocharov, Eduard V., Maxim L. Mayzel, Pavel E. Volynsky, Konstantin S. Mineev, Elena N. Tkach, Yaroslav S. Ermolyuk, Alexey A. Schulga, Roman G. Efremov, and Alexander S. Arseniev. "Left-Handed Dimer of EphA2 Transmembrane Domain: Helix Packing Diversity among Receptor Tyrosine Kinases." *Biophysical Journal* 98.5 (2010): 881-89.

Bocharov, E. V., M. L. Mayzel, P. E. Volynsky, M. V. Goncharuk, Y. S. Ermolyuk, A. A. Schulga, E. O. Artemenko, R. G. Efremov, and A. S. Arseniev. "Spatial Structure and pH-dependent Conformational Diversity of Dimeric Transmembrane Domain of the Receptor Tyrosine Kinase EphA1." *Journal of Biological Chemistry* 283.43 (2008): 29385-9395.

Bocharov, E. V., K. S. Mineev, P. E. Volynsky, Y. S. Ermolyuk, E. N. Tkach, A. G. Sobol, V. V. Chupin, M. P. Kirpichnikov, R. G. Efremov, and A. S. Arseniev. "Spatial Structure of the Dimeric Transmembrane Domain of the Growth Factor Receptor ErbB2 Presumably Corresponding to the Receptor Active State." *Journal of Biological Chemistry* 283.11 (2008): 6950-956.

Bocharov, Eduard V., Dmitry M. Lesovoy, Sergey A. Goncharuk, Marina V. Goncharuk, Kalina Hristova, and Alexander S. Arseniev. "Structure of FGFR3 Transmembrane Domain Dimer: Implications for Signaling and Human Pathologies." *Structure* 21.11 (2013): 2087-093.

Choma, Christin, Holly Gratkowski, James D. Lear, and William F. DeGrado

“Asparagine-mediated self-association of a model transmembrane helix.” *Nature Structural Biology* 7.2 (2000): 161-166.

Doura, Abigail K., Felix J. Kobus, Leonid Dubrovsky, Ellen Hibbard, and Karen G. Fleming. "Sequence Context Modulates the Stability of a GxxxG-mediated Transmembrane Helix–Helix Dimer." *Journal of Molecular Biology* 341.4 (2004): 991-98.

Endres, Nicholas F., Rahul Das, Adam W. Smith, Anton Arkhipov, Erika Kovacs, Yongjian Huang, Jeffrey G. Pelton, Yibing Shan, David E. Shaw, David E. Wemmer, Jay T. Groves, and John Kuriyan. "Conformational Coupling across the Plasma Membrane in Activation of the EGF Receptor." *Cell* 152.3 (2013): 543-56.

Finger, C., C. Escher, and D. Schneider. "The Single Transmembrane Domains of Human Receptor Tyrosine Kinases Encode Self-Interactions." Science Signaling 2.89 (2009).

Gu, Yanliang, Tapas Kar, and Steve Scheiner. "Fundamental Properties of the CH···O Interaction: Is It a True Hydrogen Bond?" *Journal of the American Chemical Society* 121.40 (1999): 9411-422.

He, Lijuan, and Kalina Hristova. "Physical–chemical Principles Underlying RTK Activation, and Their Implications for Human Disease." *Biochimica Et Biophysica Acta (BBA) - Biomembranes* 1818.4 (2012): 995-1005.

Khadria, Ambalika S., Benjamin K. Mueller, Jonathan A. Stefely, Chin Huat Tan, David J. Pagliarini, and Alessandro Senes. "A Gly-Zipper Motif Mediates Homodimerization of the Transmembrane Domain of the Mitochondrial Kinase ADCK3." *Journal of the American Chemical Society* 136.40 (2014): 14068-4077.

Kirrbach, J., M. Krugliak, C. L. Ried, P. Pagel, I. T. Arkin, and D. Langosch. "Self-interaction of Transmembrane Helices Representing Pre-clusters from the Human Single-span Membrane Proteins." *Bioinformatics* 29.13 (2013): 1623-630.

Kondo, N., K. Miyauchi, F. Meng, A. Iwamoto, and Z. Matsuda. "Conformational Changes of the HIV-1 Envelope Protein during Membrane Fusion Are Inhibited by the Replacement of Its Membrane-spanning Domain." *Journal of Biological Chemistry* 285.19 (2010): 14681-4688.

Krivov, Georgii G., Maxim V. Shapovalov, and Roland L. Dunbrack. "Improved Prediction of Protein Side-chain Conformations with SCWRL4." *Proteins* 77.4 (2009): 778-95.

Krogh, Anders, Björn Larsson, Gunnar Von Heijne, and Erik L.l Sonnhammer. "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application

to Complete Genomes." *Journal of Molecular Biology* 305.3 (2001): 567-80.

Lapointe, L.M., K.C. Taylor, S. Subramaniam, A. Khadria, I. Rayment, and A. Senes. "Structural Organization of FtsB, a Transmembrane Protein of the Bacterial Divisome." *Biochemistry* 52 (2013): 2574-85.

Lazaridis, Themis. "Effective Energy Function for Proteins in Lipid Membranes." Proteins 52.2 (2003): 176-92

Mackenzie, Kevin R., James H. Prestegard,, and Donald M. Engelman. "A Transmembrane Helix Dimer: Structure and Implications." *Science* 276.5309 (1997): 131-33.

Mackerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-Mccarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins." *The Journal of Physical Chemistry B* 102.18 (1998): 3586-616.

Manni, Sandro, Konstantin S. Mineev, Dinara Usmanova, Ekaterina N. Lyukmanova, Mikhail A. Shulepko, Mikhail P. Kirpichnikov, Jonas Winter, Milos Matkovic, Xavier Deupi, Alexander S. Arseniev, and Kurt Ballmer-Hofer. "Structural and Functional Characterization of Alternative Transmembrane Domain Conformations in VEGF Receptor 2 Activation." *Structure* 22.8 (2014): 1077-089.

Mineev, K.S., N.F. Khabibullina, E.N. Lyukmanova, D.A. Dolgikh, M.P. Kirpichnikov, and A.S. Arseniev. "Spatial Structure and Dimer–monomer Equilibrium of the ErbB3 Transmembrane Domain in DPC Micelles." *Biochimica Et Biophysica Acta (BBA) - Biomembranes* 1808.8 (2011): 2081-088.

Mineev, Konstantin S., Eduard V. Bocharov, Yulia E. Pustovalova, Olga V. Bocharova, Vladimir V. Chupin, and Alexander S. Arseniev. "Spatial Structure of the Transmembrane Domain Heterodimer of ErbB1 and ErbB2 Receptor Tyrosine Kinases." *Journal of Molecular Biology* 400.2 (2010): 231-43.

Mottamal, Madhusoodanan, and Themis Lazaridis. "The Contribution of Cα −H···O Hydrogen Bonds to Membrane Protein Stability Depends on the Position of the Amide." *Biochemistry* 44.5 (2005): 1607-613.

Mueller, B. K., S. Subramaniam, and A. Senes. "A Frequent, GxxxG-mediated, Transmembrane Association Motif Is Optimized for the Formation of Interhelical C -H Hydrogen Bonds." *Proceedings of the National Academy of Sciences* 111.10 (2014).

Munter, Lisa-Marie, Philipp Voigt, Anja Harmeier, Daniela Kaden, Kay E. Gottschalk, Christoph Weise, Rüdiger Pipkorn, Michael Schaefer, Dieter Langosch, and Gerd Multhaup. "GxxxG Motifs within the Amyloid Precursor Protein Transmembrane Sequence Are Critical for the Etiology of Aβ42." *The EMBO Journal* 26.6 (2007): 1702-712.

Park, Hahnbeom, Jungki Yoon, and Chaok Seok. "Strength of Cα −H···O=C Hydrogen Bonds in Transmembrane Proteins." *The Journal of Physical Chemistry B* 112.3 (2008): 1041-048.

Russ, William P., and Donald M. Engelman. "The GxxxG Motif: A Framework for Transmembrane Helix-helix Association." *Journal of Molecular Biology* 296.3 (2000): 911-19.

Russ, W. P., and D. M. Engelman. "TOXCAT: A Measure of Transmembrane Helix Association in a Biological Membrane." *Proceedings of the National Academy of Sciences* 96.3 (1999): 863-68.

Sal-Man, Neta, Doron Gerber, and Yechiel Shai. "The Composition Rather than Position of Polar Residues (QxxS) Drives Aspartate Receptor Transmembrane Domain Dimerization *in Vivo*." *Biochemistry* 43.8 (2004): 2309-313.

Senes, Alessandro, Mark Gerstein, and Donald M. Engelman. "Statistical Analysis of Amino Acid Patterns in Transmembrane Helices: The GxxxG Motif Occurs Frequently and in Association with β-branched Residues at Neighboring Positions." *Journal of Molecular Biology* 296.3 (2000): 921-36.

Senes, A., I. Ubarretxena-Belandia, and D. M. Engelman. "The C-H···O Hydrogen Bond: A Determinant of Stability and Specificity in Transmembrane Helix Interactions." *Proceedings of the National Academy of Sciences* 98.16 (2001): 9056-061.

The UniProt Consortium. “UniProt: a hub for protein information.” *Nucleic Acids Res.* 43 (2015) D204-D212.

Ullrich, Axel, and Joseph Schlessinger. "Signal Transduction by Receptors with Tyrosine Kinase Activity." *Cell* 61.2 (1990): 203-12.

Walters, R. F. S., and W. F. Degrado. "Helix-packing Motifs in Membrane Proteins." *Proceedings of the National Academy of Sciences* 103.37 (2006): 13658-3663.

Weiss, Carol D. “HIV-1 gp41: Mediator of Fusion and Target for Inhibition.” *AIDS Rev* 5 (2003): 214-21.

Yin, H., J. S. Slusky, B. W. Berger, R. S. Walters, G. Vilaire, R. I. Litvinov, J. D. Lear, G. A. Caputo, J. S. Bennett, and W. F. Degrado. "Computational Design of Peptides That Target Transmembrane Helices." *Science* 315.5820 (2007): 1817-822.

Yohannan, Sarah, Salem Faham, Duan Yang, David Grosfeld, Aaron K. Chamberlain, and James U. Bowie. "A Cα −H···O Hydrogen Bond in a Membrane Protein Is Not Stabilizing." *Journal of the American Chemical Society* 126.8 (2004): 2284-285.

Zhou, Fang Xiao, Melanie J. Cocco, William P. Russ, Axel T. Brunger, and Donald M. Engelman. "Interhelical Hydrogen Bonding Drives Strong Interactions in Membrane Proteins." *Nature Structural Biology* 7.2 (2000): 154-60.

Chapter 5

Future Directions and Continuing Work

5.1 Introduction

My graduate work focused on the determination and prediction of Cα-H bonds and the GAS_{right} motif. From this work, four main conclusions were drawn. First, the GAS_{right} motif geometry is integral to the formation of Cα-H bonds and for the dimerization of single-pass transmembrane (TM) proteins. Second, the knowledge of the geometries required for Cα-H bond formation can successfully predict, *ab initio*, the structure of single-pass transmembrane dimers mediated by the GAS_{right} motif (the CATM algorithm). Third, CATM can correctly predict the structure of a protein with unknown structure (ADCK3). Finally, CATM can predict the dimerization potential of homo-dimers assayed by the biological assay TOXCAT [Russ, et al. (1999)]. These conclusions lend themselves to future study.

Continuing and future work breaks down into three main categories: further analysis of the allowed geometries for hydrogen bonding propensity, expanding CATM to model additional GAS_{right} mediated interactions, and refining the energy function to more accurately model TM dimer interactions. These improvements will allow for future studies to more accurately understand the role of single-pass transmembrane oligomerization in the cell, and for in-depth studies examining interactions in protein families.

5.2 Analysis of Serine Residues at the GAS_{right} Motif Interface

The focus of Chapter 2 was on the importance of glycine at transmembrane dimer

interfaces given its propensity to form Cα-H bonds. In contrast to this, it was found that alanine hinders Cα-H bond formation. What Chapter 2 did not focus on was the third amino acid of the GAS_{right} motif, serine. A serine residue represents a unique amino acid at the interface, like glycine and alanine, it is one of the smallest amino acids. In addition to its ability to form non-canonical Cα-H bonds, it can also form hydrogen bonds using its hydroxyl group. In unpublished work, the addition of serine at the interface maintains the singular energy well for Cα-H bond formation. However the interfacial serine residue allowed for stronger hydrogen bond potential overall due to hydroxyl hydrogen bond formation. The full impact of side-chain hydrogen bond formation at the close interface was not explored, and no *in vivo* or *in vitro* work was performed. Therefore, the *in vitro* contributions of interfacial serine residues should be explored to determine whether the addition of a hydroxyl group at the interface increases the strength of dimerization as predicted.

5.3 Modeling Anti-parallel Structures

Currently, CATM can only predict structures mediated by the GAS_{right} motif; however transmembrane helices can also dimerize via an anti-parallel motif, and are quite common [Walters, et al. (2006); Zhang, et al. (2015)]. Therefore, the algorithm should be extended to cover this common interaction type. This will involve determining where the propensity to form Cα-H bonds occur (see section 2.2.2 and figure 2.2), and updating the algorithm to sample the anti-parallel configuration (see section 2.2.6). However, there are currently no solved structures of anti-parallel transmembrane

dimers, making validation of this extension difficult.

5.4 Modeling Hetero-dimeric Structures

The next extension of CATM involves the sampling of hetero-dimeric structures. TM dimers are not limited to self-interactions, and some TM proteins can exist in both homo- and hetero-dimers (e.g. Erbb1 and Erbb2) [Endres, et al. (2013); Bocharov, et al. (2008); Mineev, et al. (2010)]. The CATM algorithm currently does not support this feature; however, the implementation is relatively straightforward.

Homo-dimeric space is defined by four degrees of freedom (see section 2.2.1). In the hetero-dimeric space, two additional degrees of freedom are necessary to define the search space (both helices have independent axial rotations and crossing points along their helical axes). The addition of search space increases the runtime of the algorithm from N^4 to N^6. Creative heuristics will need to be employed to counter this gain in runtime.

In section 4.3.1, the entire set of approximately 2,200 single-pass transmembrane proteins (as annotated by Uniprot [Uniprot Consortium (2014)]) were analyzed. However, once extended to hetero-dimeric interactions, a more limited approach must be taken. A full sample of all possible human TM-to-TM pairs cannot be tractably performed, as this would be ~4.8 million pairs (for comparison sampling 2,200 structures takes around 1 week of compute time). From a biological standpoint, not all

TM domains can pair, as proteins are expressed at different levels given the cell type and are segregated by membrane type (mitochondrial, nuclear, etc). This limitation is not without an upside, as more focused studies can be performed on a protein family, analyzing the homo- and hetero-dimeric interaction network that is formed.

5.5 Modeling Higher-order Oligomer Structures

The GxxxG motif, apart from being integral to GAS_{right} dimer formation, has been observed at the interface of higher-order TM interactions (e.g. trimers, tetramers) [Kim, et al. (2005)]. The current CATM algorithm only determines dimeric structures of TM proteins; any higher-order oligomeric interactions mediated by a GxxxG motif would be modeled as a dimer, or predicted to remain in monomeric form. Modeling these structures would involve sampling higher-order symmetry, (C3, C4 rotational symmetry, as opposed to C2), but, in principle, limited work would need to be performed.

5.6 Improving Predictions With Additional Energy Terms

As seen in Chapter 4, there is strong correlation between the CATM energy score and the relative dimerization measured by TOXCAT; however, there are outliers which are not properly predicted. One way to combat this issue is to modify the energy score. The measure of helix dimerization energy is currently represented by three energy terms: a van der Waals packing term, a hydrogen bonding term (for both canonical and non-canonical bonds), and a solvation term to mimic the lipid bilayer. However, these terms are not the full extent of energy terms that have been outlined in the literature. For

instance, the CATM van der Waals term is from CHARMM, which includes energy terms for bonded atomic distance, bond angle, dihedrals, and other geometric energy terms, as well as an electrostatic term [MacKerell, et al. (1998)]. Rosetta, one of the most widely used protein structure prediction software packages, has a large array of energy terms to best capture the correct structure. Some terms mimic closely ones used in CHARMM; however, other terms score the preference for a given rotatmer, the preference for an amino acid given a set backbone, and a term that scores the geometry of the backbone [Kuhlman, et al. (2000)]. The energy score is a linear combination of the individual terms, with weighting factors associated with each. These weighting factors allow for refinements to the total energy score, which can lead to more accurate models [Barth, et al. (2007)]. CATM is currently being developed to add further energy terms, and work is being done to refine the term weights. These improvements should lead to better correlation between the energy score and the relative TOXCAT dimerization.

5.7 Refining CATM Energy Scoring with *in vitro* Assays

In Chapter 4, a comparison of CATM energy scores and relative dimerization in the TOXCAT assay was performed. The results showed good correlation between the two systems. However, TOXCAT is not a quantitative assay, and was chosen over more quantitative *in vitro* systems due to the relative speed of construct preparation and assay length. While work has shown that there is some correlation between TOXCAT and sedimentation equilibrium analytical ultra-centrifugation [Duong, et al. (2007)], true

free energy of dimerization cannot be measured. In order for CATM to better represent this energy, work must be performed in *in vitro* quantitative systems.

5.8 Addition of Backbone Flexibility at Proline Residues

The current CATM algorithm (see section 2.2.6) models TM domains as idealized helices: 3.6 residues per turn, and a rise per residue of 1.5 angstroms. However, helical parameters of solved structures show deviation from the ideal. To maintain this ideality, proline, which is known to cause kinks in helices, was not modeled. Instead, all occurrences of proline were modeled as alaine residues. In order to properly model helix deviations, backbone flexibility must be modeled in CATM. This could be accomplished by allowing Monte Carlo moves of the phi/psi backbone torsional angles, or by finding a representative set of alpha-helical kinks in the Protein Data Bank and sampling this set. Any addition of backbone flexibility will increase the runtime of the algorithm; however, further precision would be gained in atomic-level modeling.

5.9 Conclusions

Continuing work on the CATM project involves refining the groundwork laid in order to ask more biologically relevant questions. Updates to the algorithm will allow for the modeling of more structure types: anti-parallel dimers, hetero-dimers and higher-order oligomers. These updates will allow for comparisons to be made between CATM generated models to make better predictions on the oligomeric state of a protein complex. The ability to predict hetero-dimeric structures will allow for more complex

studies of protein families, for instance, the ability to create interaction maps, detailing the strength of association between all protein partners. The other main updates to CATM are focused around refining the computational model, either by adding and/or re-weighting energy terms, using more assays to better quantify interaction energies, and by adding more backbone flexibility to more accurately model a real protein backbone. These improvements will help to eliminate false positives from the data set and better describe the interaction between transmembrane protein dimers. Overall the future of CATM lies in the transition from method design to applications of the method, using the knowledge of single-pass transmembrane oligomerization to better understand their role in cellular function.

5.10 References

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

Bocharov, E. V., K. S. Mineev, P. E. Volynsky, Y. S. Ermolyuk, E. N. Tkach, A. G. Sobol, V. V. Chupin, M. P. Kirpichnikov, R. G. Efremov, and A. S. Arseniev. "Spatial Structure of the Dimeric Transmembrane Domain of the Growth Factor Receptor ErbB2 Presumably Corresponding to the Receptor Active State." *Journal of Biological Chemistry* 283.11 (2008): 6950-956.

Duong MT, Jaszewski TM, Fleming KB, MacKenzie KR. “Changes in Apparent Free Energy of Helix–Helix Dimerization in a Biological Membrane Due to Point Mutations” *Journal of Molecular Biology* 371 (2007): 422-434.

Endres, Nicholas F., Rahul Das, Adam W. Smith, Anton Arkhipov, Erika Kovacs, Yongjian Huang, Jeffrey G. Pelton, Yibing Shan, David E. Shaw, David E. Wemmer, Jay T. Groves, and John Kuriyan. "Conformational Coupling across the Plasma Membrane in Activation of the EGF Receptor." *Cell* 152.3 (2013): 543-56.

Kim S, Jeon TJ, Oberai A, Yang D, Schmidt JJ, Bowie JU. (2005) “Transmembrane glycine zippers: physiological and pathological roles in membrane proteins.” *Proc Natl Acad Sci USA* 102:14278–14283.

Kuhlman B, Baker D. “Native protein sequences are close to optimal for their structures.” *Proc Natl Acad Sci USA* 97 (2000): 10383-10388.

Mackerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-Mccarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins." *The Journal of Physical Chemistry B* 102.18 (1998): 3586-616.

Mineev, Konstantin S., Eduard V. Bocharov, Yulia E. Pustovalova, Olga V. Bocharova, Vladimir V. Chupin, and Alexander S. Arseniev. "Spatial Structure of the Transmembrane Domain Heterodimer of ErbB1 and ErbB2 Receptor Tyrosine Kinases." *Journal of Molecular Biology* 400.2 (2010): 231-43.

Russ, W. P., and D. M. Engelman. "TOXCAT: A Measure of Transmembrane Helix Association in a Biological Membrane." *Proceedings of the National Academy of Sciences* 96.3 (1999): 863-68.

The UniProt Consortium. “UniProt: a hub for protein information.” *Nucleic Acids Res.* 43 (2015) D204-D212.

Walters, R. F. S., and W. F. Degrado. "Helix-packing Motifs in Membrane Proteins." *Proceedings of the National Academy of Sciences* 103.37 (2006): 13658-3663.

Zhang, Shao-Qing, Daniel W. Kulp, Chaim A. Schramm, Marco Mravic, Ilan Samish, and William F. Degrado. "The Membrane- and Soluble-Protein Helix-Helix Interactome: Similar Geometry via Different Interactions." *Structure* 23.3 (2015): 527-41.