ABSTRACT

Title of Document: THE DESIGN OF TWO PROTEINS THAT HAVE 100% SEQUENCE IDENTITY BUT ENCODE DIFFERENT FOLDS

Dana Motabar, Master of Science, 2015

Directed By: Dr. Philip Bryan
Institute for Bioscience and Biotechnology Research

It is well-established that proteins adopt specific three-dimensional structures. However, examples of proteins that can adopt more than one folded state have become increasingly more common. The objective of this thesis is to determine how three common, small folds are connected in sequence space. The folds this work focuses on are a 3-α-helix bundle, an α/β plait, and a 4β+α fold. Topological alignment and site-directed mutagenesis were used to develop engineered variants of the 3-α-helix bundle and the α/β plait folds that maintain their highly distinct native folds even though their sequences are 100% identical. CD and NMR data suggest that both proteins were stable and folded. This engineered fold switch demonstrates that the fold preference of a sequence is dependent upon stabilizing interactions within the context of the protein. These fold switching proteins have important implications in areas such as protein design, human disease, and structural biology.

THE DESIGN OF TWO PROTEINS THAT HAVE 100% SEQUENCE IDENTITY BUT ENCODE DIFFERENT FOLDS

By

Dana Motabar

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2015

Advisory Committee:
Dr. Philip Bryan, Chair
Dr. John Orban
Dr. Edward Eisenstein

Acknowledgements

I would first like to thank my advisor, Dr. Philip Bryan, for giving me the opportunity to be a part of his lab and work on this exciting and challenging project. I am extremely grateful for his invaluable mentorship and continuing support during the course of this research. I will surely benefit from all that I have learned from him throughout the rest of my life.

I would like to express my appreciation to the members of my committee, Dr. Philip Bryan, Dr. John Orban and Dr. Edward Eisenstein, for accepting to evaluate my thesis and dedicating their time to my defense.

A warm thank you to Dr. Lauren Porter, Dr. Biao Ruan, Dr. Yingwei Chen, and Dr. Richard Simmerman for their friendship, advice, and guidance throughout our time working together. I am extremely grateful for all their encouraging support and their helpful suggestions during my research.

I thank all members of the Orban lab, especially Dr. John Orban, Yihong Chen, and Tsegsa Solomon, for obtaining the NMR spectra shown in this thesis.

I would also like to express my gratitude to Dr. Tracy Chung for her advice and her continuous support throughout my undergraduate and graduate studies.

Lastly, I would like to thank my family for standing by me throughout all my endeavors. I owe everything that I accomplish to them.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Motivation and Objectives

Proteins spontaneously collapse into well-defined tertiary structures, known as folds. Although it is well-established that a protein's amino acid sequence determines its fold, it remains unclear how seemingly simple one-dimensional amino acid sequences can encode complex three-dimensional structures (Dill, Ozkan, Shell, & Weikl, 2008). It is well-understood that proteins can shift their equilibrium between a native, folded state and a disordered, unfolded state (Dobson, 2003). More recently, it has been demonstrated that proteins can also be engineered to switch into new, alternative folded states and that the sequence space separating the two distinct folds can be small (P. A. Alexander, He, Chen, Orban, & Bryan, 2007, 2009; Porter, He, Chen, Orban, & Bryan, 2015). Learning how proteins acquire these alternative folds and how to identify switchable sequences will lead to a better understanding of both protein folding and the evolution of new folds and functions. Therefore, the motivation behind this work is both to determine how proteins are connected in sequence space and to understand how that connectivity can be predicted.

The objective of this thesis is to determine how three common, small folds are connected in a sequence space by engineering two proteins with high sequence identity that encode different folds. The three folds that this work will focus on are a 3-α-helix bundle fold, an α/β plait fold, and a 4β+α fold. Alexander et al. showed that two different fold topologies, the 3-α-helix bundle fold and 4β+α fold, nevertheless have extremely high sequence identities (up to 98%) while retaining their respective folds (P. A. Alexander et al., 2007, 2009). This thesis will demonstrate that the engineered variants of the 3-α-helix bundle and the α/β plait folds maintain their highly distinct native folds even though their sequences are 100% identical. The key to this successful fold switch was that the stabilizing interactions within the protein allowed the 3-α-helix bundle fold to change conformations to the alternative α/β plait fold topology.

1.2 Background

This section will establish the context for the thesis. It will explain both the folds that are being studied and the thermodynamics of how proteins fold in general. Furthermore, it will discuss both naturally occurring and previously engineered fold switches.

1.2.1 Protein Structures

1.2.1.1 Overview

Proteins have evolved to perform various functions. These functional properties are determined by the protein's three-dimensional fold, which is encoded by a specific sequence of amino acids (Creighton, 1993). The three-dimensional protein structure is composed of secondary structural elements. There are two major types of secondary structure (Rossmann & Argos, 1981). The most common type of secondary structure in proteins is the alpha (α-) helix (Doig & Baldwin, 1995). One characteristic of the α-helix is that all of its residues have backbone dihedral angles approximately equal to -60 and -40 degrees, respectively. The angles fall within the lower left quadrant of the Ramachandran plot (Ramakrishnan & Ramachandran, 1965). The α- helix has 3.6 residues per turn and contains hydrogen bonds which connect all -NH and- CO groups except for the first -NH groups and the last -CO groups at the end of the helix (Branden & Tooze, 1999). The second major structural element in proteins is the beta (β-) sheet. The β-sheet is comprised of β-strands that interact to form either parallel, antiparallel, or mixed pleated sheets (Rossmann & Argos, 1981). β-strands are approximately 5 to 10 residues long and are in a fully extended conformation with backbone dihedral angles in the upper left quadrant of the Ramachandran plot (Branden & Tooze, 1999; Ramakrishnan & Ramachandran, 1965). Secondary structure elements then connect to form domains, which are units of function (Rossmann & Argos, 1981).

1.2.1.1 The three proteins discussed in this thesis

This thesis will focus on three proteins: G_A, G_B, and S6. Two of the proteins, G_A and G_B, are domains within Protein G. Protein G is a multi-domain, cell surface receptor protein of *Streptococcus*, Lancefield group G (Gallagher, Alexander, Bryan, & Gilliland, 1994). Protein G contains tandem repeats of two types of domains, G_A and G_B, which bind to serum proteins in blood (P. A. Alexander et al., 2007). This ability to bind serum proteins helps the organism to evade the host defenses by creating a coat of host proteins (Gallagher et al., 1994; Kraulis et al., 1996). The G_A domain (PDB: 2FS1) is comprised of 56 structured amino acids that take a 3-α-helix bundle fold and that bind to human serum albumin (HSA) (Falkenberg, Bjoerck, & Aakerstroem, 1992). The G_B domain (PDB: 1PGB) is comprised of 56 structured amino acids that take a 4β+α fold and that bind to the constant (FC) region of immunoglobulin (IgG) (He, Chen, Alexander, Bryan, & Orban, 2012; Myhre & Kronvall, 1977). Because of their low levels of sequence identity, the wild-type forms of these domains do not appear to be evolutionarily related. The other starting protein, which has an α/β plait fold (PDB: 1RIS), is the ribosomal protein S6 from the small ribosomal subunit of *Thermus thermophilus* (Lindahl, 1994). We chose to study these three common folds because they are small, exhibit two state behavior, and fold without any intermediates (Bryan & Orban, 2010).

1.2.1.2 G_A domain: 3-α-Helix Bundle Fold

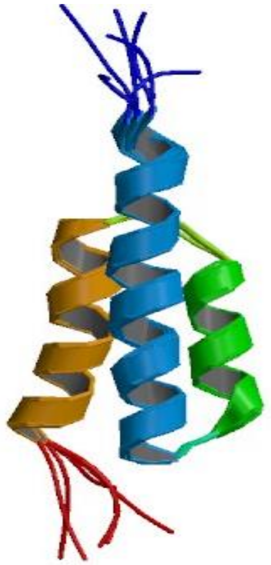


Figure 1. Solution strcture of the G_A domain (He et al., 2006)

3-α-helix bundle proteins are some of the smallest and fastest cooperatively folding structural domains (Wickstrom et al., 2006). The surface of an α-helix can be described by a row of adjacent side chains that form ridges which are separated by grooves. These ridges and grooves are created by side chains that are approximately three to four residues apart. α-helix bundles are then packed by fitting the ridges of one helix into the grooves of another helix (Chothia, Levitt, & Richardson, 1977). Helix stability is a result of hydrogen bonding, tight main chain packing, and the release of bound water when the chain folds into a helix (Aurora, Creamer, Srinivasan, & Rose, 1997).

1.2.1.3 G_B domain: 4β+α Fold

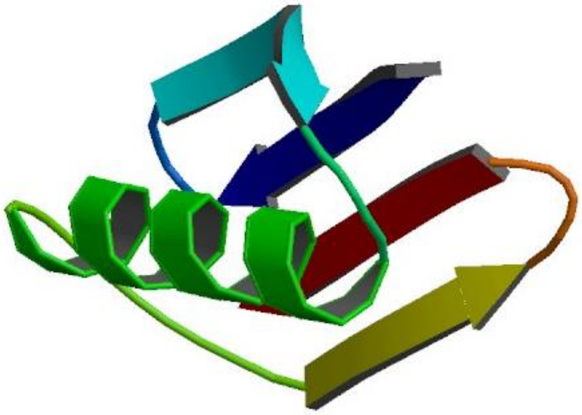


Figure 2. The structure of the G_B domain (Gallagher et al., 1994)

α/β structures are the most frequent and regular of all protein structures (Branden & Tooze, 1999). The folding pattern of the 4β+ α fold (also called the α/β grasp) is a four-stranded, anti-parallel β-sheet that forms stabilizing hydrophobic contacts with an α-helix (P. Alexander, Orban, & Bryan, 1992). The connectivity scheme is β1-β2-α1-β3-β4. β1 and β2 are adjacent and parallel to each other while β3 and β4 are anti-parallel (Burroughs, Balaji, Iyer, & Aravind, 2007). The side of the sheet that does not interact with the α-helix is solvent-exposed. The 4β+ α fold of G_B is stable without disulfide bonds or tight ligand binding even though it is small in size (P. Alexander et al., 1992).

1.2.1.4 S6 domain: α/β Plait Fold

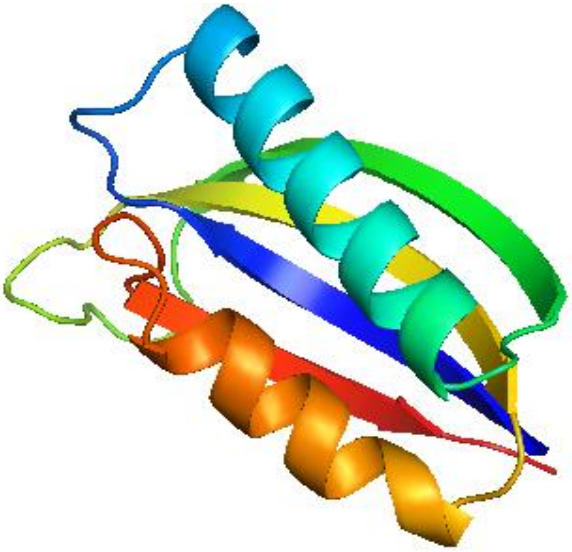


Figure 3. The structure of the S6 domain

The α/β plait is the third most populated fold after TIM barrels and Rossmann folds (Grant, Lee, & Orengo, 2004). Proteins with this fold perform many diverse functions, and their thermodynamic and kinetic properties vary widely. The α/β plait has an anti-parallel, α+β topology that consists of two helices packed against a four-stranded β-sheet (Mirny & Shakhnovich, 1999). The connectivity scheme is β1-α1-β2-β3-α2-β4. The amino acids in the inner β-strands (β1 and β3) are mostly hydrophobic whereas the rest of the protein has a substantial amount of charged and polar residues. Furthermore, there is an extended loop region between β strands 2 and 3 which forms a hook shape that partly folds over the β-sheet, giving the protein a concave nature (Lindahl, 1994).

1.2.2 Thermodynamics of Protein Folding

The process by which a polypeptide chain acquires its biologically active native state is called protein folding (Branden & Tooze, 1999). Generally, proteins in their native state are unstable as small changes in temperature and pH can cause them to unfold. Unfolded proteins populate the denatured state. Under physiological conditions, the energy difference between the native and denatured states is quite small, about 5-15 kcal/mol (Dill et al., 2008). It is important that the energy difference remain small in order for cells to degrade and synthesize proteins and for proteins to retain their structural flexibility. The protein folding reaction of the three proteins discussed in this thesis can be described in terms of a two-state reaction:

$$N \underset{K_F}{\overset{K_U}{\longleftrightarrow}} U$$

where K_U and K_F are first-order rate constants for unfolding and folding, respectively. The folded state (N) has a free energy of unfolding between 5 and 15 kcal/mol. The unfolded state (U) can be approximated by a random coil (Bryan & Orban, 2010).

The thermodynamics of protein folding is defined by the Gibbs free energy (ΔG) equation that is shown below:

$$\Delta G = \Delta H - T\Delta S \text{ where } T = \text{temperature}$$

There are two major contributors to the energy difference between the folded and unfolded states, enthalpy and entropy. Enthalpy changes (ΔH) arise from the non-covalent interactions of the polypeptide chain including the formation or breaking of hydrophobic interactions, hydrogen bonds, and ionic bonds. These interactions are stronger in the native state, which is packed more tightly than the denatured state (Dill, 1990). Entropy changes (ΔS) describe conformational heterogeneity. In the denatured state, proteins are highly disordered and thus adopt many different conformations. In the native state, proteins take a highly ordered conformation that fluctuates slightly (e.g. its loops can have different structures in solution). Therefore, proteins in the disordered state have more energetically favorable entropies than native proteins (Brady & Sharp, 1997).

One of the major driving forces of protein folding is the large free energy change that occurs by bringing hydrophobic side chains out of contact with water and into contact with each other in the interior of the protein. This phenomenon is often referred to as the hydrophobic effect (Pace, 1992). This effect significantly restricts the number of conformations the protein can take thus allowing the protein to fold in only a matter of seconds (Dill, 1990). When the hydrophobic side chains are buried, this also causes their polar backbone, -NH and –CO groups, to be buried as well; this is unfavorable for the polar backbone groups as they are unable to form hydrogen bonds with water. In order to counteract this energetically unfavorable situation, the –NH and –CO groups of the main chain form hydrogen bonds with each other which results in the creation of secondary structure elements.

There are many obstacles that need to be overcome for a protein to fold. These obstacles include formation of incorrect disulfide bonds, isomerization of proline residues, and aggregation of intermediates through exposed hydrophobic groups (Branden & Tooze, 1999). Cells contain various other types of proteins to overcome these energy barriers to folding.

1.2.3 Proteins that Switch Folds

1.2.3.1 Overview

Under physiological conditions, a protein will adopt a specific, three-dimensional fold. A protein changing folds due to a single point mutation is very rare and unlikely; more likely options would be that the protein would either retain its fold or it would unfold and lose function (Elber, 2015). It is generally understood that proteins with as little as 30% sequence identity typically adopt the same fold (Porter et al., 2015). However, examples of 'metamorphic' proteins, or proteins that can adopt more than one folded state, have become increasingly more common (He et al., 2006).

1.2.3.2 Naturally Occurring Fold Switches

Naturally occurring fold switches occur due to a change in environmental factors such as the presence of a ligand, temperature, salt concentration, or a redox state. The three common features of switchable folds are: (1) The structural transitions need states with diminished stability in order to allow large scale changes; (2) Flexible regions in the protein are needed to allow the transition from one conformer to another; (3) The generation of a new binding surface occurs which stabilizes the alternative fold and expands function (Bryan & Orban, 2010; He et al., 2012). Currently, there are eight identified naturally occurring fold switches (Chang et al., 2015; Murzin, 2008; Porter et al., 2015).

One example of a naturally occurring fold switch is lymphotactin. This protein acquires a new function by reversibly switching folds. Under physiological conditions, lymphotactin adopts two distinct but approximately equimolar conformations: a monomeric chemokine fold (Ltn10) and a dimeric β-sandwich fold (Ltn40) (Tuinstra et al., 2008). Varying salt concentration and temperature can shift the conformational equilibrium to favor one fold over the other. The major structural transition from the Ltn10 conformer to the Ltn40 conformer is the creation of a hydrophobic dimer interface in the Ltn40 subunit. The dimer interface of the Ltn40 form binds glycosaminoglycans whereas the Ltn10 form acts as an agonist of the G-protein coupled XCR1 receptor (Bryan & Orban, 2010).

The mitotic arrest deficiency 2 (Mad2) protein is another example of a protein that, upon switching folds, performs a new function. Mad2 proteins are used to ensure the correct binding of microtubules to kinetochores (Luo et al., 2004). This protein switches between an inactive open state (O-Mad2) and an active closed state (C-Mad2) by undergoing major C-terminal conformational changes. This conformational switch is important as it allows for the exposure of a latent cdc20-binding site (Bryan & Orban, 2010).

1.2.3.3 Previously Engineered Conformational Switches

In addition to naturally occurring fold switches, high-identity proteins that are capable of switching folds have been developed through various protein engineering and design methods. Recently, a pair of proteins with two different fold topologies, but highly identical sequences (up to 98%) was designed (P. A. Alexander et al., 2007, 2009). The two proteins were the G_A and G_B domains from the *Streptococcus* cell surface protein G discussed previously (Section 1.2.2). In this study, latent binding sites were created by engineering HSA and IgG binding epitopes into the domains; these binding sites were exposed when the protein switched folds. Identity was increased via stepwise mutation in the binary sequence space (choice of either the G_A or G_B amino acid at positions of non-identity) of the G_A and G_B sequences using phage display and site-directed mutagenesis. This resulted in NMR structures of the G_A-G_B pairs with 88% , 95% , and 98% identities (He et al., 2012). The 98% identical fold pair switched conformations via a single amino acid substitution. This work demonstrates that a protein can switch both its fold and its function via a short mutational path.

1.2.3.4 Significance

Proteins that can switch fold topologies are important in many areas such as protein design and evolution, human disease, and structural biology (Bryan & Orban, 2013). The amino acid sequences of fold switching proteins contain an extensive amount of information as both a stable native state and the propensity for an alternative state are encoded in their sequences. This information could be used to improve protein structure prediction algorithms and can lead to a greater understanding of how a protein’s amino acid sequence specifies its structure. Moreover, fold switching seems to be more likely to occur between some folds than others (P. A. Alexander et al., 2009). Analyzing the structural aspects of fold switching proteins will allow us to understand what makes these folds more amenable to conformational changes and can improve the prediction of other fold switching proteins. Furthermore, the development of fold switching proteins can contribute to the field of protein evolution. Fold switching proteins suggest that many diverse protein folds may have evolved from an one or several existing folds rather than evolving independently (Bryan & Orban, 2010). Finally, the design of protein conformational switches could potentially be used to develop new therapeutics, such as multifunctional proteins, to better treat diseases (Bryan & Orban, 2013).

Chapter 2: 3-α-Helix Bundle and α/β Plait Fold Switch

2.1 Design of Variants

2.1.1 Topological Alignment

Threading takes the amino acid sequence of a protein and evaluates how well it fits into a known three-dimensional protein structure (Rost, Schneider, & Sander, 1997). Threading methods are based upon two facts: (1) That many protein structures in the PDB are similar and (2) That there must be a limited number of unique protein folds found in nature. Therefore, due to the limited number of folds available to a protein sequence, there is an increased probability of solving the problem of structure prediction using threading. There are two components to most threading methods: the actual threading of a sequence into a specific structure and the evaluation of whether that alignment corresponds to a correct sequence-structure match (Lemer, Rooman, & Wodak, 1995).

The most basic threading method only uses protein sequence alignment. This method determines the optimal alignment between the new (target) sequence and the sequence of a known fold based on the alignment with the highest pairwise identity (or similarity) between the two sequences. This information is then used to infer the structure of the target sequence by evaluating the optimal sequence alignment. The most important limitation of this method is that it does not use any structural information from the known protein to determine the optimal alignment. This presents problems for determining the optimal alignment for fold-switching proteins;

highly similar sequences can adopt different folds, but it is not clear in advance whether two sequences with reasonably high levels of sequence identity can switch between one fold and another.

More advanced methods of threading use algorithms to fit a target sequence to a known structure in a library of folds. The alignment between the spatial positions of the three-dimensional structure and the protein sequence is evaluated using a specific scoring function that calculates the energy of that alignment. These methods rely heavily on the parameters of the programming algorithms used. A variety of different studies have been performed to test different threading algorithms. Progress of these computational protein structure prediction methods is assessed in the biannual Critical Assessment of Protein Structure Prediction (CASP) experiments (Moult, Fidelis, Kryshtafovych, Rost, & Tramontano, 2009). In CASP experiments, research groups test their prediction methods on sequences for which there is an unknown native structure. These community-wide experiments provide a way to assess and monitor progress in the field (Floudas, 2007).

Topological alignment was the method used for the design of G_A and $S6_{GA}$ variants. Topological alignment uses secondary structure alignment in order to find the best possible alignment of sequences. For this threading method, as S6 is a larger protein than G_A, the most main chain matches over an extended region of the fold topology gives the initial anchor point. Structures were aligned by inspection using PYMOL software (DeLano, 2002). Various topological alignments were evaluated; for this specific pair of proteins, the best match occurred when the largest helices of both structures were aligned. This is illustrated in Figure 4 which shows the close alignment of the G_A and $S6_{GA}$ helices via the overlapping cyan and green colors. Using this alignment, the optimal register was determined to be residue 1 of G_A and residue 11 of S6, which is a residue in the first turn that follows the first β-strand. There is 16.1% sequence identity between the wild-type G_A sequence and the 56 amino acid subsequence (residues 11-66) in wild-type S6 where threading occurred. Once the register was defined, custom mutations were made to the sequence using PYMOL in order to increase sequence identity. Both models were then assessed with each sequence change to ensure that there were no unresolvable steric clashes and that the hydrophobic core is conserved. If a clash between side chains was created, a change at another residue position was evaluated to see if it could alleviate the problem. The designs were submitted to the RosettaDesign server to calculate their energies.

Before threading, the N- and C- termini of the wild-type S6 sequence were altered. The residues that were removed from the sequence were for an RNA binding site; this deletion occurred in order to assist with protein purification as it ensured that the target protein and RNA molecules were not co-purified. It is important to note that the topological alignment method is imperfect mainly because the computation cannot predict small changes in main chain structure that could result from the mutations of interest.

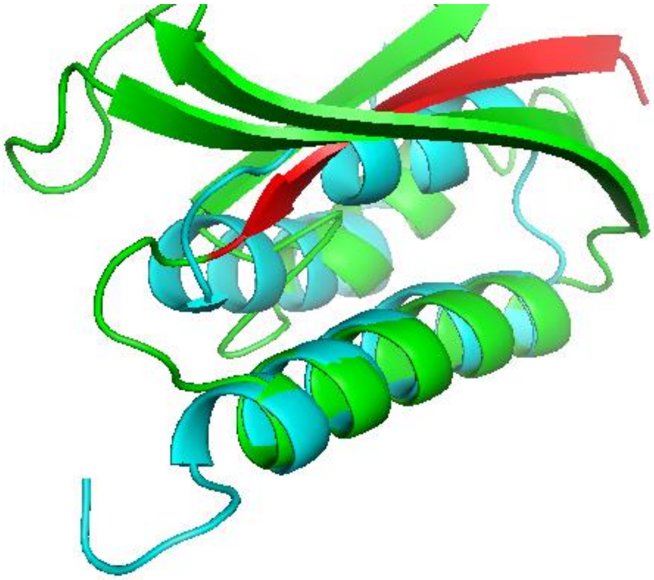


Figure 4. Topological alignment of the largest α-helices of the S6 domain and the G_A domain using PYMOL. The bottom helix shows the alignment of the helices (overlapping *cyan* and *green* helices). *(Cyan)* the G_A domain; *(Green)* the S6 domain; *(Red)* Residues 1-10 of the S6 domain. The G_A sequence was threaded through the S6 domain at residue 11.

2.1.2 RosettaDesign

Protein design software has been used for a variety of purposes such as stabilizing naturally occurring proteins, altering protein binding specificity, and designing novel protein structures. For this thesis, the RosettaDesign server was used in order to find the lowest energy sequences for each fold structure. The RosettaDesign server uses the design module of the Rosetta program to perform fixed backbone protein design simulations. RosettaDesign consists of two components: an energy function to determine the favorability of the sequence and an optimization procedure for searching sequence space. The energy function consists of many different components such as the Lennard- Jones potential, the Lazaridis-Karplus implicit solvation model, a hydrogen bonding term, torsion potentials, reference values for all amino acids, and an electrostatic interaction term for charged residues (Liu &

Kuhlman, 2006). Side chains can populate a number of low energy conformations called rotamers. RosettaDesign has a library of permissible rotamers for each amino acid in the protein. To determine the lowest energy sequences, RosettaDesign uses Monte Carlo optimization with simulated annealing (Liu & Kuhlman, 2006). The efficiency of the search and the accuracy of the energy function are the biggest challenges and the most distinguishing features of each design program (Lazar & Handel, 1998). RosettaDesign has been used successfully to redesign nine naturally occurring proteins (Dantas, Kuhlman, Callender, Wong, & Baker, 2003). It has also been used to redesign smaller regions of proteins in order to increase protein stabilities or binding affinities (Eletr, Huang, Duda, Schulman, & Kuhlman, 2005; Nauli, Kuhlman, & Baker, 2001).

2.1.3 Designed Mutants

After topological alignment and threading, protein design principles were used to increase identity between the G_A and $S6_{GA}$ sequences. The two sequences were first aligned and positions of non-identity were identified. Given the binary sequence space (choice of either the G_A or $S6_{GA}$ amino acid at positions of non-identity), several different approaches were used to evaluate sites that could be mutated while still maintaining stability:

1) Creating salt bridges by identifying amino acids with a lone charge (such as Asp, Glu, Lys, Arg) and attempting to engineer a partner for that charge. Surface salt bridges increase protein stability (Makhatadze, Loladze, Ermolenko, Chen, & Thomas, 2003). We sought to design optimal salt bridges by optimizing their spatial

orientation and considering the interactions of individual side chains forming the salt bridge with the rest of the protein.

2) Repacking the hydrophobic core by identifying hydrophobic cavities that could be filled with other, larger hydrophobic amino acids. Core-packing plays a critical role in protein stability (Lazar & Handel, 1998). Replacing large amino acids in the hydrophobic core with smaller, non-polar side chains destabilize protein structures considerably (Cordes, Davidson, & Sauer, 1996). By replacing smaller hydrophobic residues with larger ones, a sizeable increase in stability may be gained, however it is generally more challenging to engineer due to the difficulty in creating a tightly packed core.

3) Removing glycines from certain positions may decrease the conformational entropy of the sequence thus increasing the stability of the protein. Glycine has the greatest backbone conformational entropy of all amino acids. Thus, it requires more free energy during the folding process in order to restrict the conformation of glycine (Matthews, Nicholson, & Becktel, 1987).

4) Similar to the logic in approach 3, inserting prolines at specific positions can also decrease the chain entropy. The pyrolidine ring of proline restricts the residue to fewer conformations that are available to other amino acids (Matthews et al., 1987). A proline residue in the chain restricts not only the (Φ, Ψ) values of the proline residue but also hinders the (Φ, Ψ) values of the preceding residue (Matthews et al., 1987; Schimmel & Flory, 1968). This mutation is best utilized at positions in tight turns (Fu, Grimsley, Razvi, Scholtz, & Pace, 2009).

5) Engineering an intramolecular disulfide bond to form between two cysteines can increase the stability of the folded state. Engineered intramolecular disulfide bonds have stabilized many proteins including T4 lysozyme (Perry & Wetzel, 1986) and subtilisin BPN' (Mitchinson & Wells, 1989). The cross-linking of the disulfide bond restricts the degrees of freedom of the unfolded chain and thus stabilizes the folded state.

6) Mutating polar side chains that are buried in the protein core to hydrophobic side chains. It is energetically favorable for polar side chains (such as Ser, Thr, Cys, Asn, Gln, Tyr) to be solvent-exposed (Kohn, 1998).

7) Mutating hydrophobic groups that are on the protein surface. It is energetically favorable for hydrophobic groups (such as Val, Leu, Ile, Met, Pro) to be buried in the interior of the protein, shielding them from solvent (Kohn, 1998).

8) Creating helix caps can help stabilize α-helices. Helix-capping motifs are specific patterns of hydrogen bonding and hydrophobic interactions found at or near the ends of helices in proteins. In an α-helix, the first four -NH groups and last four -CO groups lack intrahelical hydrogen bonds; these groups are often capped by alternative hydrogen bond partners (referred to as N-cap and C-cap) (Aurora & Rose, 1998). Capping motifs can stabilize α-helices by fulfilling hydrogen-bonding potential and burying hydrophobic surfaces (Cordes et al., 1996). Helix preferences differ between amino acids at interior positions and the helix termini (Doig & Baldwin, 1995). It was found that amino acids whose side chains can accept hydrogen bonds from otherwise free backbone -NH groups such as Asn, Asp, Ser, Thr, and Cys, generally have the

highest preference for the N-cap whereas Gly is highly preferred in the C-cap position. (Doig & Baldwin, 1995; Richardson & Richardson, 1988).

Furthermore, 19 of the 20 amino acids bias the protein backbone to adopt either the α-helix, the β-sheet, or the reverse turn conformation (Levitt, 1978). The preferences for a particular secondary structure are dependent on the chemical structure and stereochemistry of the amino acid. Amino acids with bulky side chains such as those branched at the β-carbon (like Val, Ile, Thr) and those with large, aromatic rings (like Phe, Tyr, Trp) bias the backbone to adopt the β-sheet conformation. β-strand preferences are enhanced by the ability of a side chain to shield the β-sheet hydrogen-bonding networks from solvent (Cordes et al., 1996). Amino acids with short polar side chains (like Ser, Asp, Asn) or with side chains that either expand or restrict the conformational space accessible to the protein backbone (i.e. Gly and Pro, respectively) bias the protein main chain to adopt reverse turns. The remaining amino acids, except for Arg which has no preference, bias the protein backbone to adopt α-helices (Levitt, 1978). The propensity for these amino acids to be found in α-helices is dependent in part on the loss of side chain entropy through interactions with side chain atoms in the preceding turn. The effects of secondary structure preferences of the amino acids are modest, with the average substitution (excluding proline and glycine) changing stability by less than 0.5 kcal/mol. Although the influence of single substitutions may be small and variable, the net effect of secondary structure preferences on stability becomes substantial when summed over the entire protein sequence (Cordes et al., 1996).

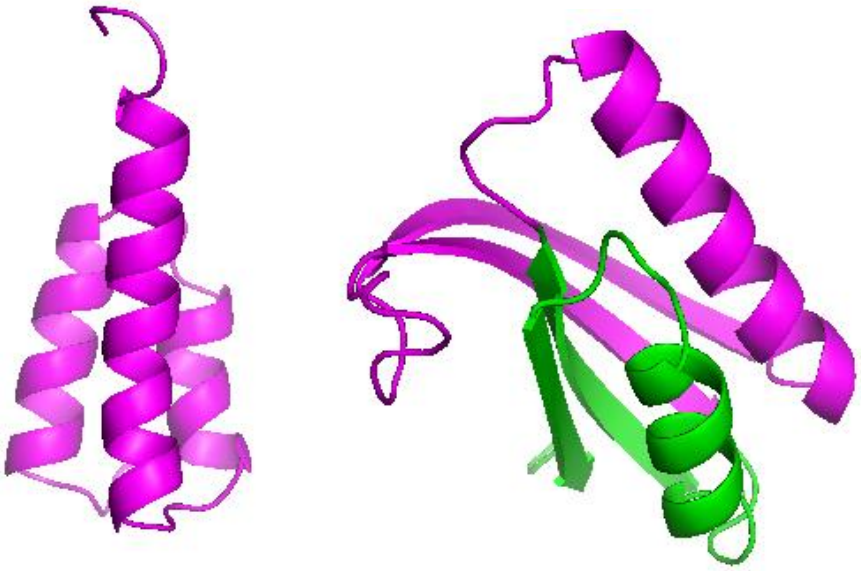


Figure 5. Engineered protein varaiants for G_A and $S6_{GA}$ illustrated in ribbon style. (*Left*) The G_A variant (*Right*) The $S6_{GA}$variant(s). (*Magenta)* Residues that are included in the G_A subsequence. (*Green)* Residues that are not included in the G_A subsequence.

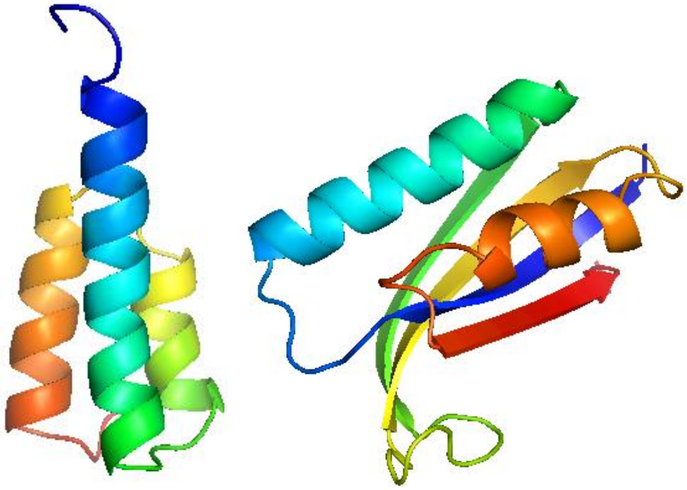
The engineered proteins with different folds, the 3-α-helix bundle (Fig.5A) and the α/β plait (Fig.5B), are both pictured in figure 5. The magenta portion of the α/β plait fold (Fig.5B) denotes the residues that correspond to the G_A subsequence.

As shown in Table 1, there were sequence changes for all S6 domain and G_A domain variants. Wild-type G_A and ${}^{100}G_A$ have 60.7% sequence identity whereas wild-type S6 and ${}^{100}S6_{GA}$ have 51.5% sequence identity. For the ${}^{100}G_A$ and ${}^{100}S6_{GA}$ variants, all positions between the G_A sequence and the G_A subsequence within the S6 domain are identical. To obtain the 100% identity between G_A and $S6_{GA}$, an I26A mutation was made in ${}^{98}S6_{GA}$ variant. For the S6 protein, the residues that flank the G_A subsequence were required in order assist with the folding of the sequence into an α/β plait topology that is not highly favored in isolation. Rainbow coloring was used in Figure 6 in order to indicate the amino acids which correspond to the secondary structure elements of both fold topologies.

Table 1. High identity protein variants for G_A and $S6_{GA}$

Variant	**Sequence**[1,2,3]
G_A(2FS1)	MEAVDANSLAQAKEAAIKELKQYGIGDYYIKLINNAKTVEGVESLKNEILKALPTE*
$^{100}G_A$	NPNLDQKQLAQAKELAIKALKQYGIGVEKIKLIGNAKTVEAVEKLKQGILLVYQIE*
S6(1RIS)	MRRYEVNIVLNPNLDQSQLALEKEIIQRALENYGARVEKVEELGLRRLAYPIAKDPQGYFLWYQVEMPEDRVNDLARELRIRDNVRRVMVVKSQEPFLANA*
$^{98}S6_{GA}$	SKTFEVNIVLNPNLDQKQLAQAKELIIKALKQYGIGVEKIKLIGNAKTVEAVEKLKQGILLVYQIEAPADRVNDLARELRILDAVRRVEVTYAAD*
$^{100}S6_{GA}$	SKTFEVNIVLNPNLDQKQLAQAKELAIKALKQYGIGVEKIKLIGNAKTVEAVEKLKQGILLVYQIEAPADRVNDLARELRILDAVRRVEVTYAAD*

[1] Colored backgrounds indicated where a new mutation was made. The color was changed in subsequent sequences.
[2] The red font indicates the amino acids that are a part of the G_A sequence/subsequence.
[3] The green colored background corresponds to amino acids that were removed before threading.



$^{100}G_A$ sequence:

NPNLDQKQLAQAKELAIKALKQYGIGVEKIKLIGNAKTVEAVEKLKQGILLVYQIE

$^{100}S6_{GA}$ sequence:

SKTFEVNIVLNPNLDQKQLAQAKELIIKALKQYGIGVEKIKLIGNAKTVEAVEKLKQGILL
VYQIEAPADRVNDLARELRILDAVRRVEVTYAAD

Figure 6. The rainbow coloring of the structures and the sequences illustrates the secondary structure elements that the amino acids correspond to for both $^{100}G_A$ (*left*) and $^{100}S6_{GA}$ (*right*).

2.2 Methods

2.2.1 Cloning

The G_A and S6 gene constructs were ordered from GeneArt™ Gene Synthesis (ThermoFisher Scientific, Grand Island, NY). The genes were amplified using iProof PCR. The primers used to amplify these genes contained the EcoRI and HindIII restriction sites. The amplified PCR product was purified with QIAquick® PCR Purification Kit (Qiagen, Valencia, CA) before and after digestion with EcoRI and HindIII restriction enzymes in order to be ligated into the PPAL8 vector which encodes a His-tagged subtilisin prosequence at the N-terminus of the fusion protein. The PPAL8 vector allows for the rapid purification of the protein in a one-step reaction by subtilisin (Ruan, Fisher, Alexander, Doroshko, & Bryan, 2004). Point mutations for the subsequent variants were made using Q5 mutagenesis (New England BioLabs, Ipswich, MA). The plasmids were then transformed into XL1-Blue Supercompetent Cells (Agilent, Santa Clara, CA) or 5α Competent Cells (New England BioLabs, Ipswich, MA), plated onto LB agar containing 100 μg/mL carbenicillin, and incubated overnight at 37℃. Single colonies were chosen and grown overnight at 37℃ in LB media with 100 μg/mL carbenicillin. Plasmid DNA was extracted from the cell cultures using Wizard® Plus SV Minipreps (Promega Corporation, Madison, WI) according to the manufacturer's instructions. Concentrations were determined by ultra violet (UV) absorption at 260nm. DNA sequencing was performed by MacrogenUSA (Rockville, MD).

2.2.2 Protein Expression and Purification

Fusion protein variants were expressed in BL-DE3 cells by autoinduction at 37℃. Cells were then pelleted by centrifugation at 10,000 g for 10 minutes. Cells were lysed by sonication in 0.1M phosphate buffer pH 6.8, 0.1mM EDTA, and one Complete Mini Protease pill (Roche, Basal, Switzerland). Samples were then fractionated by high-speed centrifugation at 40,000 g for 45 minutes. Soluble cell extract of prodomain fusion protein was loaded onto a 1mL pT2197 column (S189) at 1 mL/min to allow binding and then washed with 0.1M Kpi + 300mM sodium chloride to remove impurities. The protein was then eluted with 6mL of 100mM imidazole (pH 6.8) at 0.1 mL/min. Its purity was confirmed by SDS-PAGE. The purified protein was then dialyzed into 0.1M phosphate buffer (pH 6.8). Protein concentration was then determined by UV absorption at 280nm.

2.2.3 CD and NMR

CD measurements were performed with a Chirascan™ CD spectrometer (Applied Photophysics, Surrey, UK) using quartz cells with a path length of 1-mm. Protein concentrations of 20 μM were used for all variants. The ellipticity results were expressed as mean residue ellipticity, [θ] (, degrees per cm^2 /dmol), with extinction coefficents estimated by EXPASY (Gasteiger et al., 2005). Temperature-induced unfolding was performed in the temperature range between 20°C and 95°C in a 1-mm cuvette. Ellipticities at 222nm were continuously monitored at a scanning rate of 1° per minute. (P. A. Alexander et al., 2007). Reversibility of the denaturation was confirmed by comparing the CD spectra at 25°C before melting and after heating to 95°C and cooling to 25°C. NMR spectra were obtained using an Avance III 600 MHz spectrometer at 25°C (Bruker, Billerica, MA). Minimal media was used for ^{15}N labelling. The culture was incubated overnight at 25°C. The cultures were then spun down and sonicated in lysis buffer. Samples were then fractionated by high-speed centrifugation at 40,000 g for 45 minutes. Isotope- labelled proteins were purified similarly except on a 5mL pT2197 column.

2.3 Experimental Results

2.3.1 CD

We used CD to determine the secondary structure content of the $^{100}G_{A}$, $^{98}S6_{GA}$, and $^{100}S6_{GA}$ variants. Their stabilities were then gauged through thermal denaturation, monitoring changes in their CD spectra at 222 nm. The CD spectra are shown in figures 7, 8, and 9.

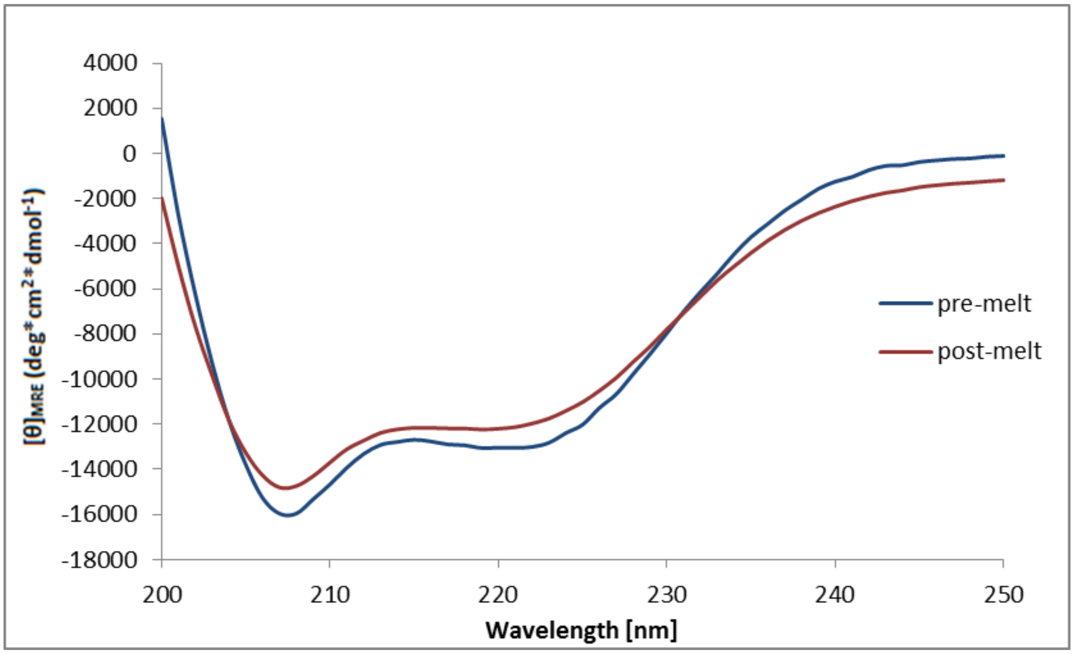


Figure 7. CD spectra of $^{100}G_{A}$ suggests a folded structure. G_{A} has a 3-α -helix bundle fold topology. The pre- and post- melt spectra both have similar shapes, which indicates reversible folding. All spectra were measured in 100mM potassium phosphate buffer pH 6.8.

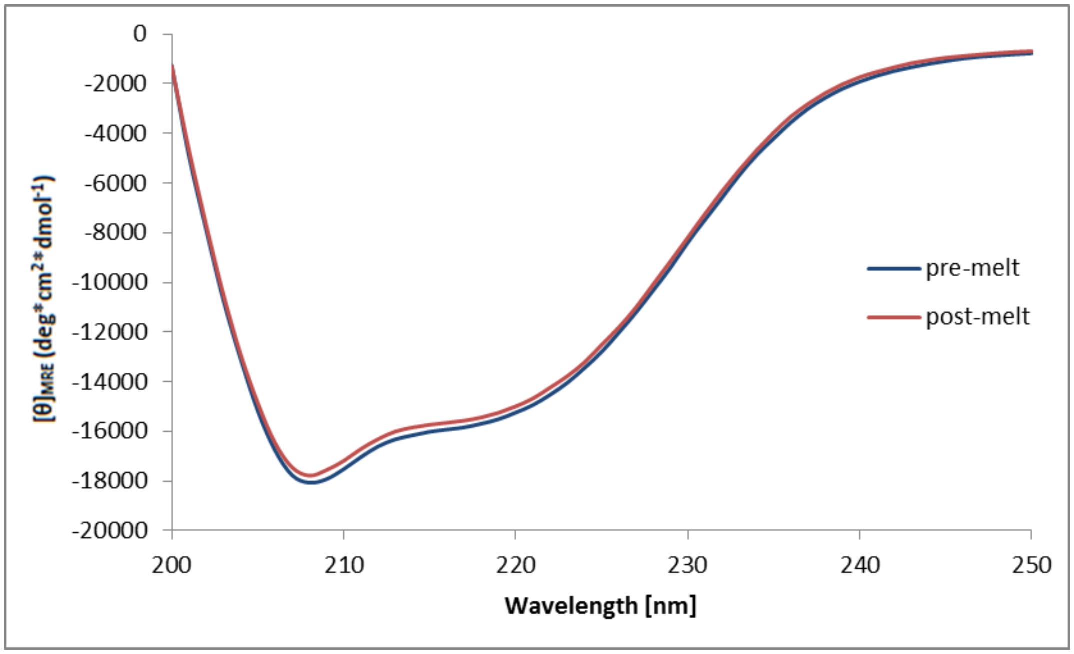


Figure 8. CD spectra of $^{98}S6_{GA}$ suggests a folded structure. All spectra were measured in 100mM potassium phosphate buffer pH 6.8.

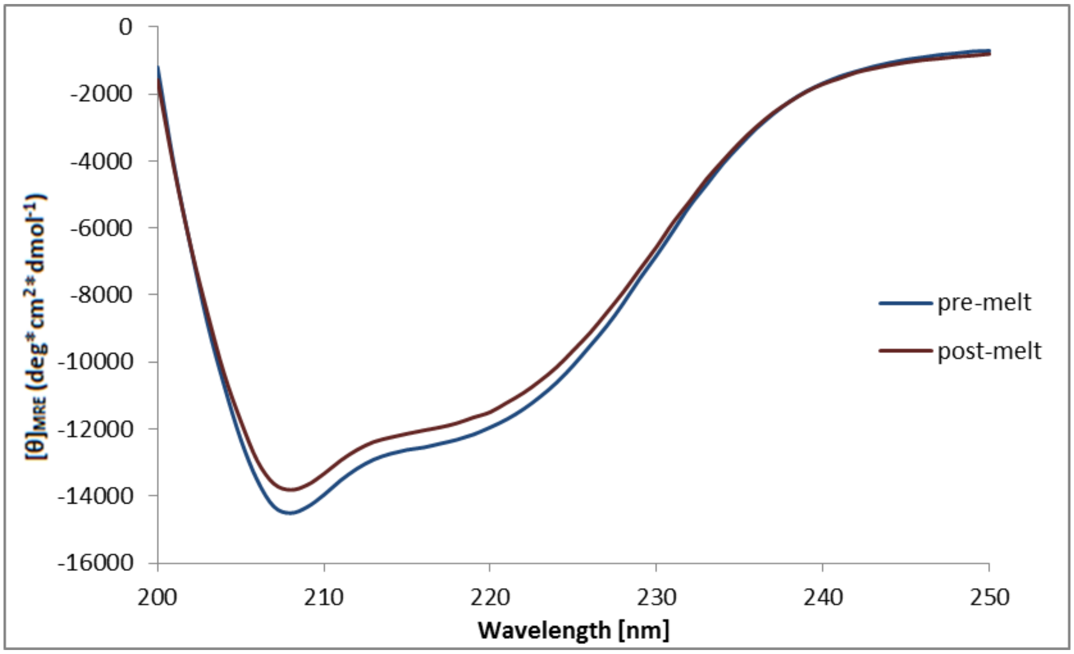


Figure 9. CD spectra of $^{100}S6_{GA}$ suggests a folded structure. The pre- and post- melt spectra both have similar shapes, which indicates reversible folding. All spectra were measured in 100mM potassium phosphate buffer pH 6.8.

The CD spectra of $^{100}G_A$ was very similar to that of its parent fold (Porter et al., 2015). This suggests that $^{100}G_A$ retains its 3-α-helix bundle fold topology. As can be seen in figure 9, the CD spectra of $^{100}S6_{GA}$ indicates that it is folded as well and has a mixed α+β fold topology. The CD spectra of $^{98}S6_{GA}$ was very similar to that of $^{100}S6_{GA}$. The ellipticity of post- melt spectra for all variants decreased as compared to the pre-melt ellipticity which is expected due to aggregation that can occur after thermal denaturation.

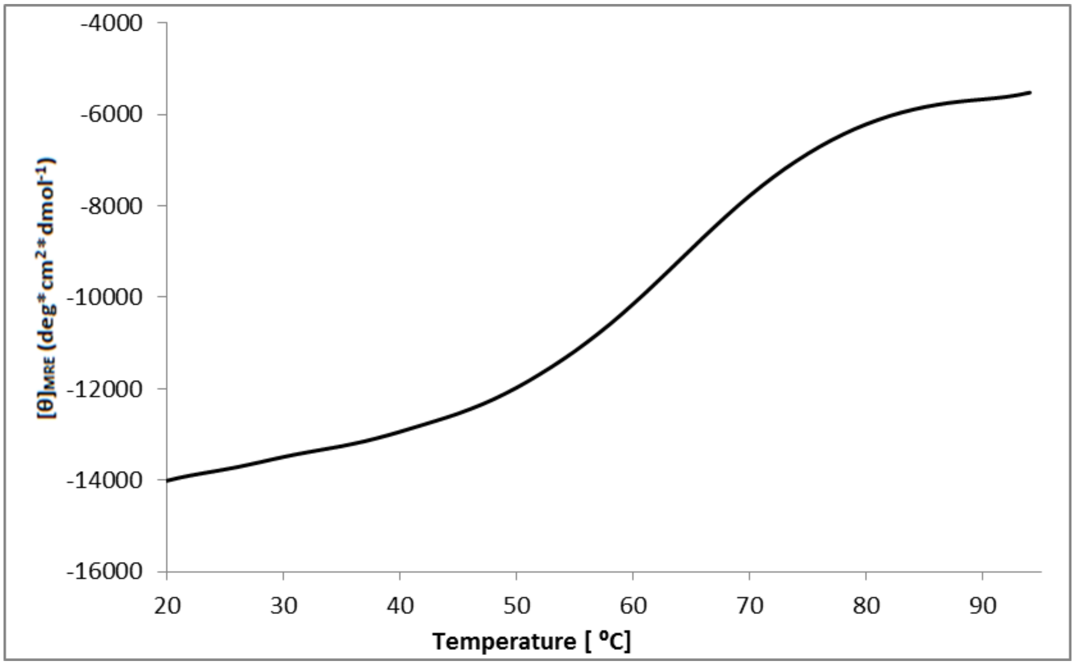


Figure 10. The thermal denaturation profile of $^{100}G_A$ indicates cooperative unfolding of the initially folded structure. The denaturation midpoint gives a melting temperature of 64˚C. All measurements were taken in 100mM potassium phosphate buffer pH 6.8.

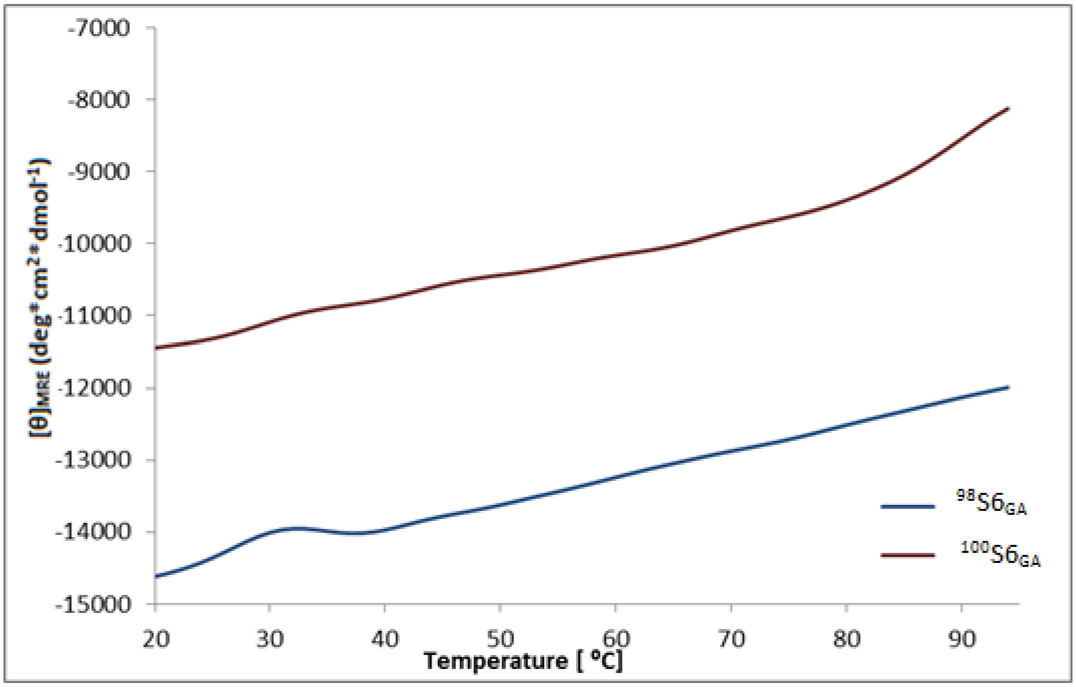


Figure 11. Thermal denaturation profiles of both S6$_{GA}$variants (^{98}S6$_{GA}$, blue; ^{100}S6$_{GA}$, red). Profiles indicate that S6$_{GA}$ variants do not unfold cooperatively in response to heat. ^{100}S6$_{GA}$ may begin unfolding around 85 ˚C, suggesting that the structure is less stable than ^{98}S6$_{GA}$.

Thermal denaturation data for both variants is shown in figures 10 and 11. As shown in figure 10, ^{100}G$_{A}$ exhibits a cooperative unfolding transition. The thermal denaturation midpoint gives the melting temperature of the variant. The derivative of the thermal denaturation profile of ^{100}G$_{A}$ indicated a melting temperature of 64 ^{o}C. As shown in figure 11, ^{98}S6$_{GA}$ and ^{100}S6$_{GA}$ variants were hyper-stable, which prevented them from fully unfolding in the given temperature range (20 ^{o}C to 95 ^{o}C). In order to see if a full unfolding transition could be observed, a CD spectra and a thermal denaturation profile were taken of ^{98}S6$_{GA}$ at pH 12. Even at a pH 12, the variant did not undergo a full unfolding transition. Comparing the thermal denaturation profiles shown in figure 11 of ^{98}S6$_{GA}$ and ^{100}S6$_{GA\ ,}$ it can be seen that ^{100}S6$_{GA}$ begins to exhibit a melting transition around 85 ˚C whereas ^{98}S6$_{GA}$ does not show any indication of a melting transition. This suggests that the I26A mutation destabilized ^{100}S6$_{GA}$ relative to ^{98}S6$_{GA}$.

2.3.2 NMR

Although CD spectra and thermal denaturation profiles can indicate well-defined tertiary structure, they do not prove conclusively that structure has been retained (P. A. Alexander et al., 2007; Blanco, Angrand, & Serrano, 1999). To further characterize folded structure, NMR spectra were obtained. Proteins were isotopically labelled with ^{15}N. The spectra of ^{98}G$_A$ (which has the same amino acid sequence as ^{100}G$_A$) and ^{98}S6$_{GA}$ are shown in figures 12 and 13. HSQC spectra of both proteins have well-dispersed main-chain amide signals.

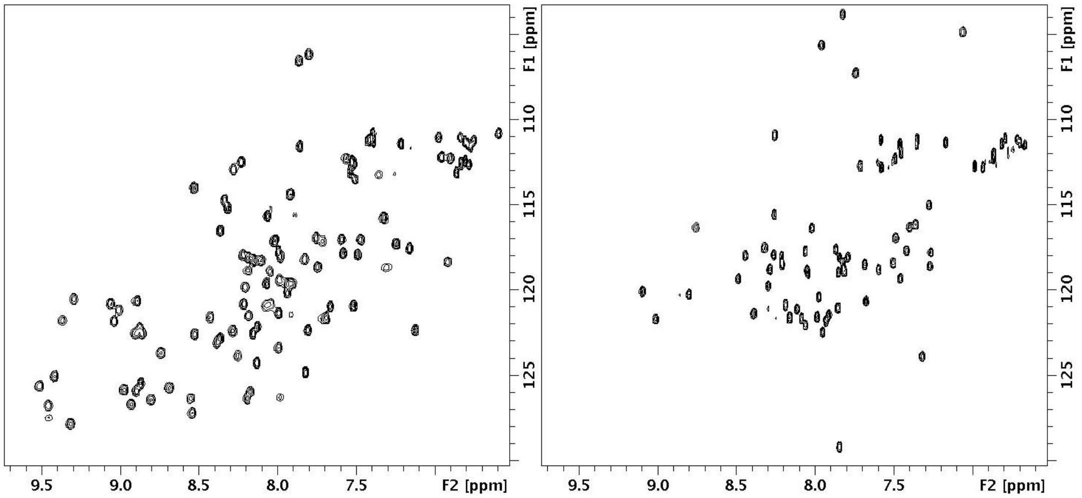


Figure 12. Two-dimensional ^{15}N HSQC spectra of ^{98}S6$_{GA}$ (*left*) and ^{98}G$_A$ (*right*) at 25˚C (us600). Both spectra indicate that the proteins are folded.

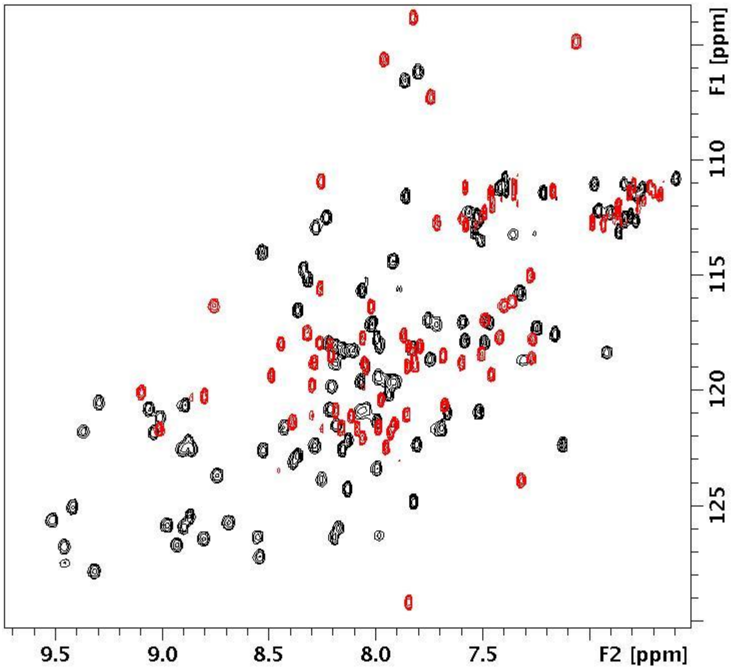


Figure 13. Overlay of the two-dimensional ^{15}N HSQC spectra of ^{98}S6$_{GA}$ (*black*) and ^{98}G$_{A}$ (*red*) at 25˚C (us600) which illustrates the folded nature of both variants.

Figure 13 shows the overlay of the HSQC spectra of ^{98}S6$_{GA}$ and ^{98}G$_{A}$. It is evident that the proteins have two different fold topologies based upon the major differences in their HSQC spectra. The HSQC spectra for ^{98}S6$_{GA}$ has greater dispersion than that of ^{98}G$_{A}$, which suggests that the structure contains β strands. Sequence assignments and high-resolution structures are currently being completed as well and will be published elsewhere.

Figure 14 shows the overlay of the HSQC spectra for ^{98}S6$_{GA}$ and ^{100}S6$_{GA}$. Both spectra suggest that the S6$_{GA}$ domains are folded. However, as indicated by the different locations of the chemical shifts, ^{100}S6$_{GA}$ has undergone changes in structure due to the probable repacking of the hydrophobic core. The I26A mutation is the only change in the sequence between ^{98}S6$_{GA}$ and ^{100}S6$_{GA}$. I26 stabilizes the core of the α/β plait fold. Thus, replacing it with a smaller amino acid (Ala in this case) is expected to be a destabilizing mutation based upon the structure of S6$_{GA}$.

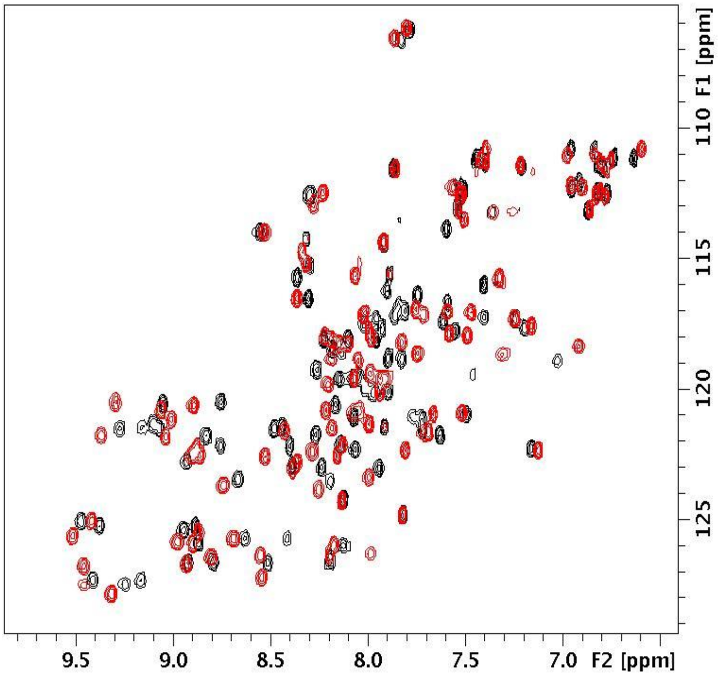


Figure 14. Overlay of the two-dimensional ^{15}N HSQC spectra of ^{100}S6$_{GA}$ (*black*) and ^{98}S6$_{GA}$ (*red*) at 25˚C (us600). As evident by the change in position of the chemical shifts, ^{100}S6$_{GA}$ has undergone changes in structure as a result of the I26A mutation.

2.4 Discussion

The pair of proteins designed, ^{100}G$_A$ and ^{100}S6$_{GA}$, was used to explore the sequence space that connects two distinct folds, the 3-α-helix bundle and the α/β plait. Approximately, 22 amino acids (~39%) in the wild-type G$_A$ sequence and 30 amino acids (~53%) in the wild-type S6 sequence were mutated to obtain 100% identity. Mutations that were made to increase identity were complementary to both sequences in order to retain fold structure. Deciding what mutations to make was not a trivial task as any mutation that is made in the sequence can change the context of how the other amino acids interact with one another (P. A. Alexander et al., 2009)

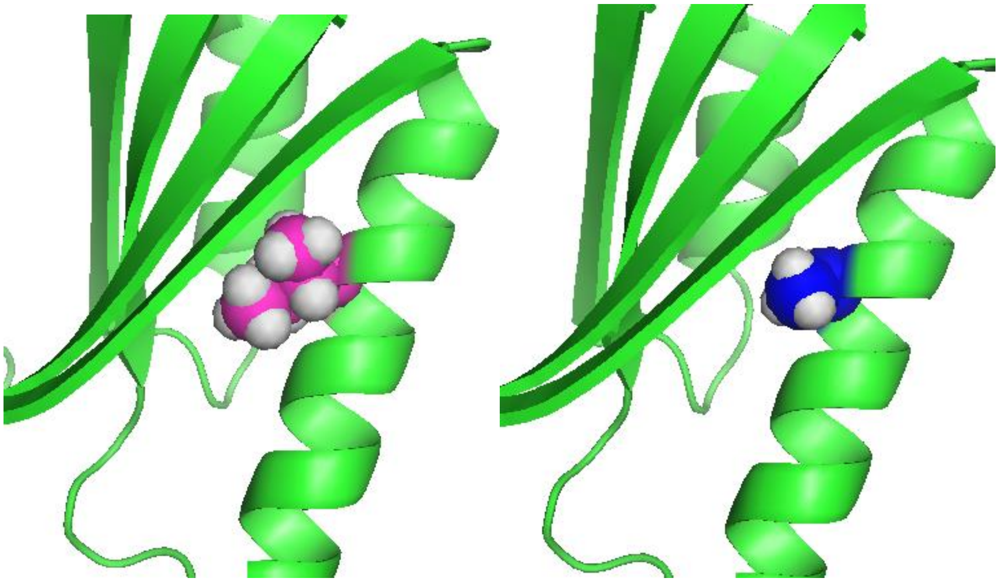


Figure 15. Depiction of residue 26 which is an Ile (*pink* residues) in $^{98}S6_{GA}$ (*left*) and an Ala (*blue* residues) in $^{100}S6_{GA}$ (*right*). The size difference between the amino acids likely causes the re-packing of the hydrophobic core when the I26A mutation occurs.

The final mutation in order to obtain 100% identity between the variants was the I26A change in the $^{98}S6_{GA}$ variant. As shown in figure 14, this single amino acid mutation caused significant shifts in the HSQC spectra, suggesting a restructuring or change in the α/β plait fold. The I26 residue is found on the first α-helix of the structure, facing inward towards the hydrophobic core of the protein. Although both Ala and Ile are hydrophobic amino acids, Ile has a branched hydrocarbon side chain that it larger than the methyl side chain of Ala. Figure 15 depicts the difference in size between the Ile side chain and the Ala side chain in the $S6_{GA}$ structure and also shows that the I26 side chain is fully buried. Hydrophobic interactions are a major contributor to the stability of the native structure of proteins (Prevost, Wodak, Tidor, & Karplus, 1991). Previous studies have used site-directed mutagenesis to change hydrophobic core residues with other apolar residues that are slightly smaller or that have different steric properties. These studies demonstrated that modest changes in packing resulting from these mutations will destabilize the protein, but the overall structure of the protein will change very little (Lim, Farruggio, & Sauer, 1992). For $^{100}S6_{GA}$, the change from I26 to A26 likely led to the creation of a cavity that caused the amino acids in the hydrophobic core to repack extensively. This probable repacking would destabilize the folded structure while leaving its overall topology intact.

It has been previously shown in work by Alexander et al. that the structure of a monomeric protein is context-dependent (P. A. Alexander et al., 2007, 2009). The development of the G_A and $S6_{GA}$ fold pair further reinforces the idea that fold preference of a protein is dependent upon stabilizing interactions within the context of the protein. There have been examples of monomeric proteins that have an alternative fold topology in a multimeric protein, such as the previously discussed lymphotactin (Bryan & Orban, 2010). For the G_A and $S6_{GA}$ pair, the switch occurs between one monomeric fold to another monomeric fold. The energetic driving force behind this switch could potentially be the more extensive hydrophobic core that is created upon the switch to the α/β plait fold topology. The amino acid residues that flanked the G_A subsequence, residues 1-10 and 67-95, were able to provide the stabilizing contacts that were needed to switch to the α/β fold topology. Many identity- increasing mutations destabilized the folded structure. Thus, the interactions from the long flanking sequences compensated for the loss of stability due to the increased identity. Without these stabilizing interactions, the alternative 3-α-helix bundle conformation was taken. Moreover, the G_A and $S6_{GA}$ fold pair is in agreement with the computational model for protein space as a network, which is based on explicit modeling of the kinetics of evolution (Meyerguz, Kleinberg, & Elber, 2007; Porter et al., 2015). Taking into account previous fold pairs that have been developed with G_A variants (P. A. Alexander et al., 2009; Porter et al., 2015), it is likely that the 3-α-helix bundle is a hub fold that is connected to other alternative folds as well.

$^{98}S6_{GA}$ and $^{100}S6_{GA}$ are hyper-stable proteins that could not be thermally denatured by CD. The S6 domain may possibly be resistant to denaturation by heat because it is from a thermophilic bacterium, *Thermus thermophilus* (Lindahl, 1994). The use of a chemical denaturant, such as urea or guanidinium chloride, is another method that could be used to obtain quantitative stability data for both variants. Chemical denaturants are frequently used to unfold proteins and to characterize mechanisms and transition states of protein folding reactions (Möglich, Krieger, & Kiefhaber, 2005). The molar concentration of the chemical denaturant can be increased until the protein fold is destabilized and the unfolding transition can be measured using NMR. Using this method, the general stability of the protein can be found. Furthermore, hydrogen deuterium (H-D) exchange experiments can be used to measure exchange rates for main-chain amide protons (Orban, Alexander, Bryan, & Khare, 1995). An HSQC spectrum can be obtained at a series of time points while the hydrogen is exchanging with the deuterium. Using an exponential fit, an exchange constant can be found and from that the $\Delta G_{unfolding}$ can be calculated for both proteins. These experiments will be able to provide stability data that can be used to better understand this fold switch.

The CD and NMR data obtained thus far is very promising as it suggests that both G_A and $S6_{GA}$ variants are folded and likely retain the structure of their parent folds. Moreover, the most conclusive way to know if the folds of these proteins are retained is by solving their structures. NMR studies to determine their structures are in progress.

Chapter 3: 4β+α and α/β plait Fold Pair

3.1 Design of Variants

3.1.1 Threading and Designed Mutants

As discussed in the previous chapter, there are various threading methods that can be used to design proteins. Unlike the G_A and $S6_{GA}$ variants, topological alignment was not used for the G_B and $S6_{GB}$ variants. For this pair of proteins, a simple threading method was used. The G_B sequence was inserted into the S6 sequence at each register in the sequence. All designs were then submitted to RosettaDesign to calculate energies (Liu & Kuhlman, 2006). The design with the lowest energy, renamed pG1016, was used.

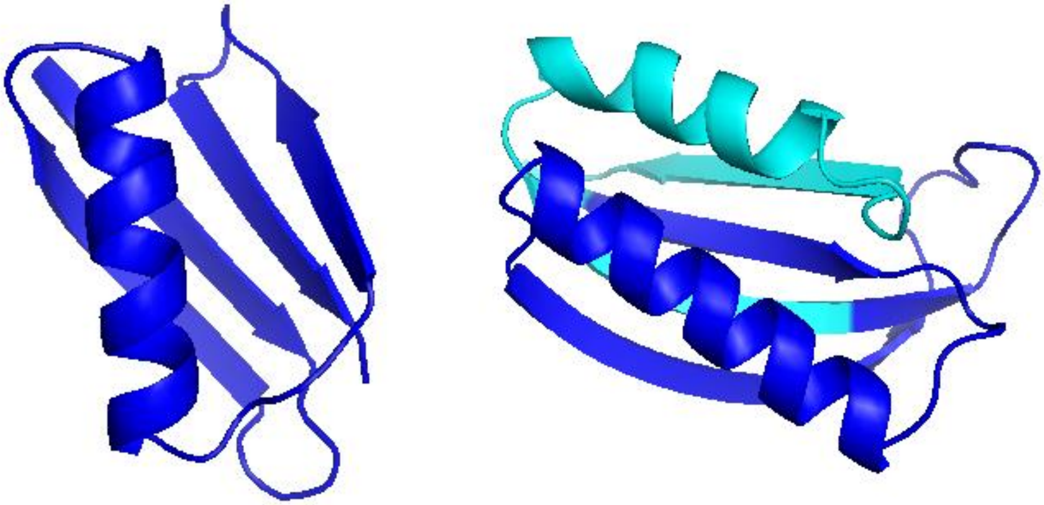


Figure 16. Engineered protein variants for G_B and $S6_{GB}$. (*Left*) The G_B variant(s) (*Right*) The S6 variant (s). (*Blue)* Residues that are included in the G_B subsequence. (*Cyan)* Residues that are not included in the G_B subsequence.

Figure 16 illustrates the structure of the G_B and $S6_{GB}$ variants. The binary sequence space was methodically examined in order to determine what positions of non-identity could be mutated to increase identity without changing the folds of these variants. The protein design principles previously discussed for the G_A and $S6_{GA}$ variants were used. Mutations in the hydrophobic core were generally avoided. Mutation tolerant sites were first chosen by inspection of the structure in PYMOL using the wizard mutation tool (DeLano, 2002) and then those mutations were further examined using the RosettaDesign server. If the mutation caused an irreconcilable clash, that mutation was avoided. Reverting to wild-type amino acids at destabilized sites assisted in partially regaining lost stability (Johnson, Gintner, Park, & Snow, 2015). Mutations to all variants are shown in Table 2. The standard nomenclature was not used for these variants as there was no folded, high-identity pair developed.

Table 2. High- identity protein variants for G_B and $S6_{GB}$

Variant	Sequence[1,2,3]
G_B (1PGB)	MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE*
pG1015	ATFKLVLNGKTLKGETTTEAVDAATALKNFGAYAQDVGVDGAWTYDDATKTFTVGE*
pG1017	ATFKLVLNGKTLKGETTTEAVDAATALKNFGAYAQDVGVDGAWTYDDATKTFTVGY*
pG1022	YTFKIVLNGKTNKGETTTEAVDAATALKNFGAYAQDVGVDGAWTYDDATKTFTVGY*
pG1023	YTFKIVLNGKTNKGETTTEAVDAATALKNFGAYAQDVGVDGAWTYSDPTKTFTVGY*
S6(1RIS)	MRRYEVNIVLNPNLDQSQLALEKEIIQRALENYGARVEKVEELGLRRLAYPIAKDPQGYFL WYQVEMPEDRVNDLARELRIRDNVRRVMVVKSQEPFLANA*
pG1016	GIATFKLVLNGKTLKGETTTEAVDAATALKNFGAYAQDVGVDGAWTYDDATKTFTVGERL IFKVKMPEDRMNDLARQLRQRDNVSRVEVTRYK*
pG1018	GIATFKLVLNGKTNKGELTTEAVDAATALKNFGAKAQDVGVDGAWTYDDATKTFTVGYRL IFKVEMPEDRMNDLARQLRQRDNVSRVEVTRYK*
pG1020	GIYTVKIVLNPKTNKGELTTEAVDAATALKNFGAKAQDVGVDGAWTYSDPTKTFPVGYRL IFKVEMPEDRVNDLARQLRQRDNVSRVEVTRYK*

[1] Colored backgrounds indicated where a new mutation was made. The color was changed in subsequent sequences.
[2] The red font indicates the amino acids that are a part of the G_B sequence/subsequence.
[3] The green colored background corresponds to amino acids that were removed.

3.1.2 Purification Tags

The $S6_{GB}$ variants had difficulty with both soluble protein expression and protein purification due to the low stability of these engineered variants. As a result of these issues with the $S6_{GB}$ variants, we sought to optimize the amount of soluble protein expressed by experimenting with several purification tags. The PPAL8 tag with an N-terminal Histidine tag was initially used for all $S6_{GB}$ variants. As stated previously, the PPAL8 tag allows for cleavage via a subtilisin column (Ruan et al., 2004). The pG1020 variant, which was the third redesign of $S6_{GB}$, had difficulty with soluble expression at both 37℃ and 25°C via autoinduction and IPTG induction. In order to attempt to optimize protein expression, the pG1020 sequence was then ligated into two vectors with different purification tags: the TK-pro tag and the 1RIS tag. These vectors were used in an attempt to increase pG1020's stability. Tk-pro is the prodomain tag for Tk-subtilisin, which is from the hyperthermophilic archaeon *Thermococcus kodakaraensis* (Pulido, Koga, Takano, & Kanaya, 2007). The 1RIS tag is the S6 domain that takes an α/β plait fold. Both the Tk-pro and 1RIS tags strongly bind to subtilisin therefore the subtilisin column with an imidazole activator (pT2197) was used for purification.

Soluble protein expression of pG1020 significantly increased with both the Tk-pro and 1RIS tags, respectively, likely due to the increase in stability. However, even though soluble expression was increased, pG1020 continued to degrade on the pT2197 column. To avoid this purification-induced degradation, pG1020 was then ligated into a new vector, pG5. After ligation, a 6-histadine tag was added onto pG1020's N- and C- termini in two separate constructs, using primers and Q5 PCR. This method avoided using another protein domain as a purification tag, as those domains may interfere with the folding of the $S6_{GB}$ protein. Moreover, the small 6-histadine tag was needed to purify the protein. This method also increased soluble protein expression as compared to the pG1020 in the PPAL8 vector. However, as with the other vectors, a significant amount of purified protein was unable to be obtained via column purification. Another method of purification could be used to circumvent the technical problems posed by column purification. One possible approach could be to cleave the variants in solution with subtilisin to see if that would improve protein yield (Schwyter, Phillips, & Reisler, 1989).

3.2 Methods and Results

3.2.1 Methods

The methods for cloning, expressing and purifying proteins, and obtaining structural information used for the $S6_{GB}$ and G_B fold pair were similar to the methods used for the $S6_{GA}$ and G_A fold pair. However, the methods differ slightly in a few ways.
For protein purification of pG1015 and pG1017, two columns were used: a 1mL nickel column and a Profinity Exact column (Bio-Rad, Hercules, CA) (Ruan et al., 2004) that cleaves the protein by an azide activator. The nickel column allows for the protein to be purified initially before cleavage on the Profinity Exact column. The purified protein was then dialyzed into 0.1M phosphate buffer (pH 6.8). Protein concentrations were determined by UV absorption at 280nm. CD measurements were performed for the pG105 and pG1017 variants with a spectropolarimeter (model J-720; JASCO, Easton, MD) using quartz cells with a pathlength of 1-cm. All measurements were taken in 100mM potassium phosphate buffer pH 6.8. The ellipticity results were expressed as mean residue ellipticity, [θ] (, degrees per cm^2 /dmol), with extinction coefficents estimated by EXPASY (Gasteiger et al., 2005). Temperature-induced unfolding was performed in the temperature range between 25^oC and 95^oC. NMR was not performed for any of the variants as no high-identity fold pairs were developed.

3.2.2 Results

CD spectra were obtained for two of the G_B variants, pG1015 and pG1017. The CD spectra for both variants suggests that they were folded and stable; therefore, their thermal stability was also assessed. The thermal denaturation profiles are shown in figures 17 and 18. The thermal profiles for pG1015 and pG1017 show the cooperative unfolding of the protein. The thermal denaturation midpoints give the melting temperature of the protein. The melting temperatures for pG1015 and pG1017 were both 56 ^{o}C. This indicates that the protein did not lose any stability due to the single point mutation of E56Y in pG1017.

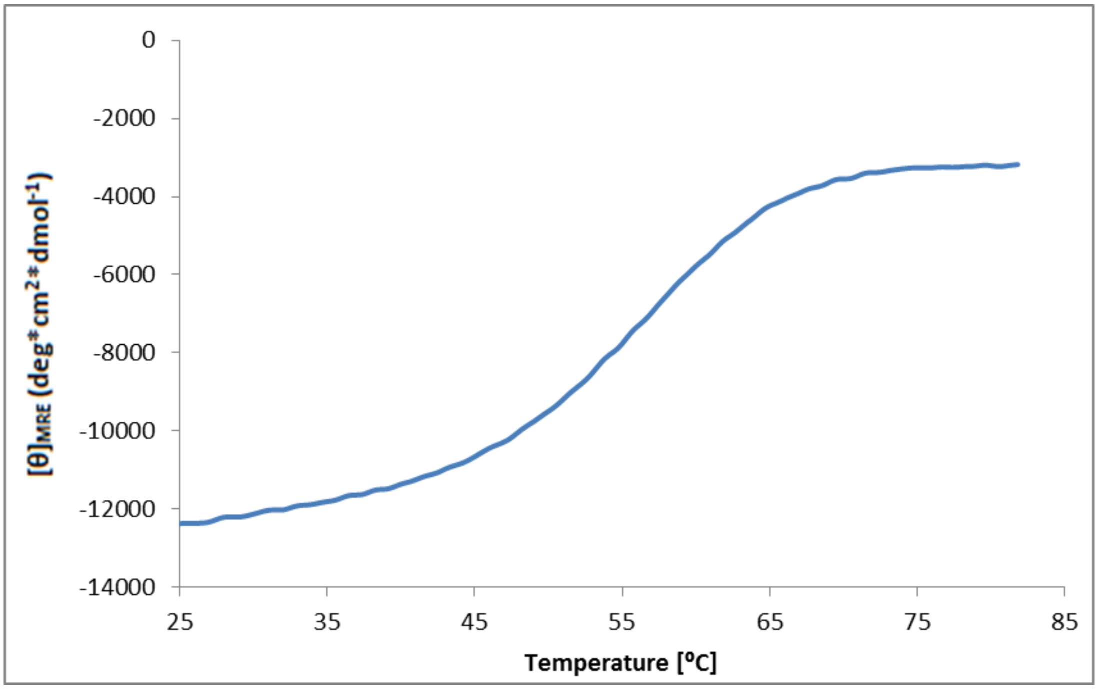


Figure 17. Thermal denaturation profile of pG1015 which suggests cooperative unfolding of the protein. The denaturation midpoint gives a melting temperature of 56˚C. All measurements were taken in 100mM potassium phosphate buffer pH 6.8.

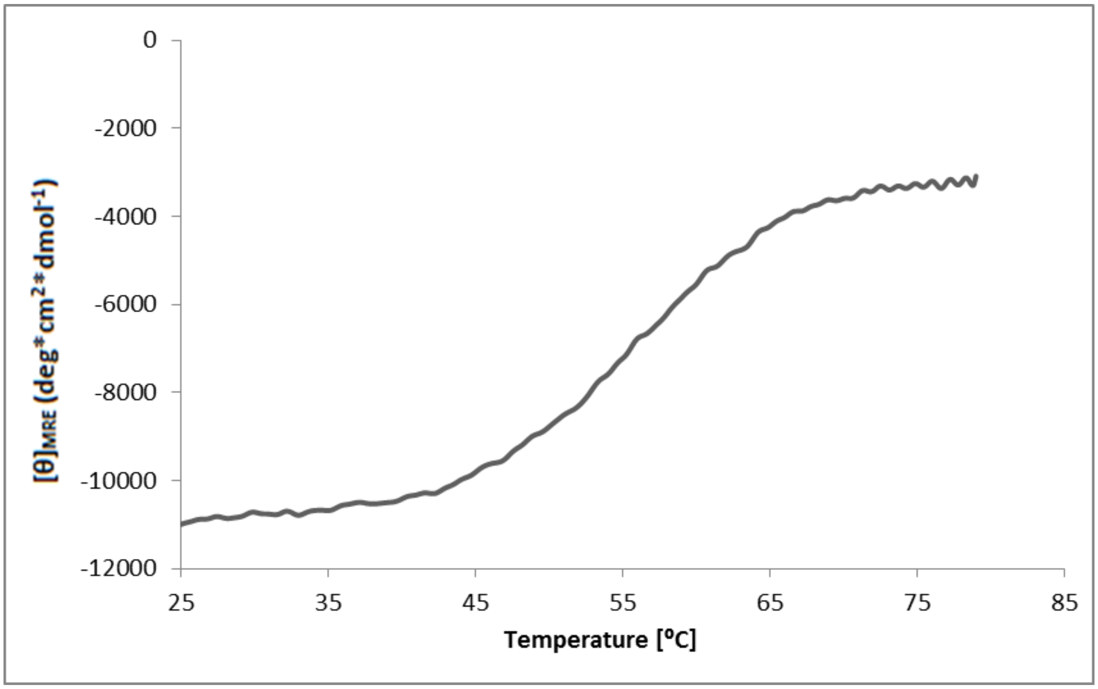


Figure 186. Thermal denaturation profile of pG1017 which suggests cooperative unfolding of the protein. The denaturation midpoint gives a melting temperature of 56˚C. All measurements were taken in 100mM potassium phosphate buffer pH 6.8.

Since CD data for both pG1015 and pG1017 suggests that they were stable proteins, the mutations made in the pG1022 and pG1023 variants were only made to increase sequence identity between G_B and $S6_{GB}$. The three new mutations included in pG1022 were: A1Y, L5I, L12N and the additional two mutations in pG1023 were: D46S and A48P. CD data for the pG1022 and pG1023 variants was not able to be obtained as the variants were unstable and could not be purified via column purification.

pG1016, which was the first design of $S6_{GB}$, did not express well solubly. New designs of the $S6_{GB}$ protein, pG1018 and pG1020, reverted mutations that were possibly destabilizing back to their wild-type amino acids (Johnson et al., 2015). Table 3 shows the percent of identity between all designed variants. It is evident that sequence identity decreased as variant designs progressed; $S6_{GB}$ variants were not sufficiently stable therefore mutated residues were changed back to wild-type, resulting in a loss of identity.

Table 3. Percent identity between G_B and $S6_{GB}$ variants

		$S6_{GB}$ variants			
	% Identity	**WT S6**	**pG1016**	**pG1018**	**pG1020**
G_B Variants	**pG1015**	30.3%	100%	92.9%	80.4%
	pG1017	30.3%	98.2%	94.6%	82.1%
	pG1022	32.4%	94.4%	92.9%	87.5%
	pG1023	28.6%	89.3%	89.3%	91.1%

3.3 Discussion

There are many possible reasons as to why a successful, high-identity fold pair was not able to be developed for the G_B and $S6_{GB}$ protein pair as compared to the G_A and $S6_{GA}$ pair. One factor may have been the different threading method used for the G_A and $S6_{GA}$ pair. The topological alignment approach may produce better results for all variants as it preserves the most native structure possible. Designing a well-threaded sequence is extremely important as it is the foundation for all subsequent variants. The residues that flank the G_B subsequence, residues 1-2 and 59- 93, may not be able to provide the contact points necessary for G_B sequence to take an alternative fold.

Another factor that may have hindered the development of this fold pair as compared to the G_A and $S6_{GA}$ fold pair was that the wild-type structure of G_A may be more amenable to sequence changes than the wild-type structure of $G_{B.}$ Amino acids 1-7 and 55-56 are disordered in G_A while it's other 47 amino acids which are well-ordered. On the other hand, all 56 amino acids of the G_B domain are all well-ordered in the secondary structure elements (P. A. Alexander et al., 2007). The unstructured N- and C- termini of the G_A domain may cause the protein retain its structure more easily in response to destabilizing mutations as compared to the G_B domain. Furthermore, the core of G_A contains fewer stabilizing hydrophobic contacts than the core of G_B. G_A has seven critical core residues (A12, A16, I33, A36, V42, K46, and I49), which are preserved in all $S6_{GA}$ and G_A variants. The G_B domain has a slightly more extensive core that includes nine critical residues (P. A. Alexander et al., 2009).

Lastly, another possibility is that the large number of mutations made to $S6_{GB}$ variants caused the protein to have diminished stability below that of most natural proteins, $\Delta G_{unfolding} \geq 5$ kcal/mol. All mutations should be reassessed in order to determine which specific mutations potentially destabilize the fold. It is possible that reverting to wild-type amino acids at destabilized sites would allow for regained stability (Johnson et al., 2015) .

Chapter 4: Conclusion and Future Work

4.1 Conclusion

For this thesis, a fold pair was designed with two protein domains, G_A and S6, which have 100% sequence identity but encode two very different fold topologies. The G_A domain retains a 3-α-helix bundle and the S6 domain retains an α/β plait fold. The G_A sequence was threaded into the S6 sequence by inspection using topological alignment. Mutations were made using PYMOL in the binary sequence space of the G_A and $S6_{GA}$ variants in order to increase sequence identity. RosettaDesign was used to evaluate all mutations and identify the lowest energy designs. The designed $^{98}S6_{GA}$ and $^{98}G_A$ fold pair was then experimentally screened and appeared folded. A single residue mutation, I26A, was then made to the $^{98}S6_{GA}$ variant to obtain a fold pair, $^{100}S6_{GA}$ and $^{100}G_A$, with 100% identity. CD and NMR data suggest that both proteins of the fold pair were stable. In isolation, the G_A sequence will adopt a 3-α-helix bundle fold. However, in the context of the $S6_{GA}$ sequence, the G_A subsequence will take an α/β plait fold topology. The amino acid residues that flanked the G_A subsequence, residues 1-10 and 67-95, were able to provide the stabilizing contacts that were needed to adopt the α/β plait fold topology. Further design work that can be done for this fold pair is discussed in section 4.2.

We attempted to engineer another fold pair with high levels of sequence identity using the G_B and S6 domains. This protein pair was optimized using different expression methods, purification methods, and purification tags. Although CD data suggests that a stable G_B variant was created, we were unable to create a S6 variant that maintained the α/β plait fold topology. Thus, the maximum level of sequence identity that was reached was 30.3%. There are various factors that may have prevented our designs from being more successful. This includes the threading method used, the well-ordered structure of G_B, and the number of mutations that were made.

4.2 Future Work

Fold switching is more likely to occur between some folds than others. Alexander et al. showed that latent binding function can be linked to alternative folding propensity. If latent binding function exists in an alternative fold, a fold switch is probable (P. A. Alexander et al., 2009). Based on this previous work, a ligand induced fold switch can be designed for the G_A and $S6_{GA}$ fold pair. Both the $^{100}G_A$ and $^{100}S6_{GA}$ variants contain the seven residues (G26, L32, A36, T38, E40, L45, and I49) that compose the HSA binding epitope of G_A (He, Chen, Rozak, Bryan, & Orban, 2007). Figure 19 highlights the seven residues of the binding epitope in both the G_A and $S6_{GA}$ folds.

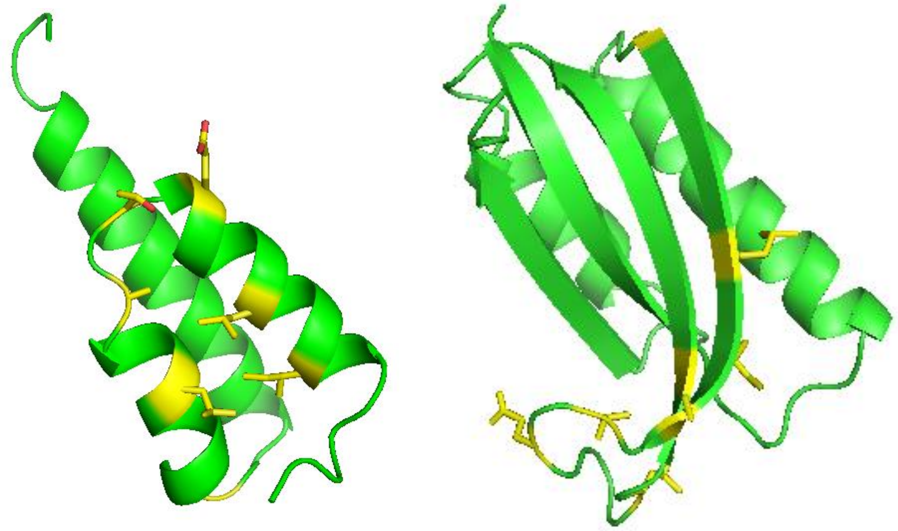


Figure 19. Ribbon representation of the G_A (*left*) and the $S6_{GA}$ (*right*) structures that highlight the residues (in stick form; both, yellow) that are included in the HSA binding epitope.

A fold switch can be developed from $^{100}S6_{GA}$ to $^{100}G_A$ in order to expose a cryptic HSA binding epitope. In order for a ligand-induced switch to occur, the ^{100} $S6_{GA}$ fold needs to first be destabilized as the fold is highly stable. If the stability of the α/β plait fold topology is diminished, alternative folds become more accessible.

Another area that can be further investigated is the development of a G_B and $S6_{GB}$ fold pair that can successfully switch conformations. The G_B sequence can be inserted into the same register as the G_A subsequence was in the $S6_{GA}$ sequence. It is evident that the flanking residues, residues 1-10 and 67-95, of the $S6_{GA}$ sequence are efficient in generating interactions to stabilize a small fold. Therefore, it is likely that these flanking residues may be able to provide enough stabilizing contacts to create an $S6_{GB}$ variant that retains the α/β plait fold topology.

Nomenclature and Glossary

Standard nomenclature used to distinguish between variants:

$^{\%\ \text{identiity}}\text{variant}_{\text{reference sequence}}$

All reference sequences for both pairs are the highest identity versions unless otherwise noted.

Amino acids abbreviations:

Ala: Alanine

Asp: Aspartic Acid

Asn: Asparagine

Arg: Arginine

Cys: Cysteine

Glu: Glutamic Acid

Gln: Glutamine

Gly: Glycine

His: Histidine

Ile: Isoleucine

Leu: Leucine

Lys: Lysine

Met: Methionine

Phe: Phenylalanine

Pro: Proline

Ser: Serine

Thr: Threonine

Trp: Tryptophan

Tyr: Tyrosine

Val: Valine

CD: Circular Dichroism

NMR: Nuclear Magnetic Resonance

PDB: Protein Data Bank

WT: Wild-type

Bibliography

Alexander, P. A., He, Y., Chen, Y., Orban, J., & Bryan, P. N. (2007). The Design and Characterization of Two Proteins with 88% Sequence Identity but Different Structure and Function. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(29), 11963–11968. doi:10.2307/25436232

Alexander, P. A., He, Y., Chen, Y., Orban, J., & Bryan, P. N. (2009). A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21149–21154. doi:10.1073/pnas.0906408106

Alexander, P., Orban, J., & Bryan, P. (1992). Kinetic analysis of folding and unfolding the 56 amino acid IgG-binding domain of streptococcal protein G. *Biochemistry, 31*(32), 7243–7248. doi:10.1021/bi00147a006

Aurora, R., Creamer, T. P., Srinivasan, R., & Rose, G. D. (1997). Local Interactions in Protein Folding: Lessons from the α-Helix. *Journal of Biological Chemistry* , *272* (3), 1413–1416. doi:10.1074/jbc.272.3.1413

Aurora, R., & Rosee, G. D. (1998). Helix capping. *Protein Science, 7*(1), 21–38.

Blanco, F. J., Angrand, I., & Serrano, L. (1999). Exploring the conformational properties of the sequence space between two proteins with different folds: an experimental study. *Journal of Molecular Biology*, *285*(2), 741–753.

Brady, G. P., & Sharp, K. A. (1997). Entropy in protein folding and in protein—protein interactions. *Current Opinion in Structural Biology*, *7*(2), 215–221. doi:10.1016/S0959-440X(97)80028-0

Branden, C., & Tooze, J. (1999). *Introduction to Protein Structure* (2nd ed.). New York: Garland Publishing.

Bryan, P. N., & Orban, J. (2010). Proteins that switch folds. *Current Opinion in Structural Biology*, *20*(4), 482–8. doi:10.1016/j.sbi.2010.06.002

Bryan, P. N., & Orban, J. (2013). Implications of protein fold switching. *Current Opinion in Structural Biology*, *23*(2), 314–6. doi:10.1016/j.sbi.2013.03.001

Burroughs, A. M., Balaji, S., Iyer, L. M., & Aravind, L. (2007). Small but versatile: the extraordinary functional and structural diversity of the beta-grasp fold. *Biol Direct*, *2*(3), 18.

Chang, Y.-G., Cohen, S. E., Phong, C., Myers, W. K., Kim, Y.-I., Tseng, R., … Lee,

Y. (2015). A protein fold switch joins the circadian oscillator to clock output in cyanobacteria. *Science*, *349*(6245), 324–328.

Chothia, C., Levitt, M., & Richardson, D. (1977). Structure of proteins: packing of alpha-helices and pleated sheets. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(10), 4130–4134. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC431889/

Cordes, M. H., Davidson, A. R., & Sauer, R. T. (1996). Sequence space, folding and protein design. *Current Opinion in Structural Biology*, *6*(1), 3–10. doi:10.1016/S0959-440X(96)80088-1

Creighton, T. E. (1993). *Proteins: structures and molecular properties*. Macmillan.

Dantas, G., Kuhlman, B., Callender, D., Wong, M., & Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *Journal of Molecular Biology*, *332*(2), 449–460.

DeLano, W. L. (2002). The PyMOL molecular graphics system.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, *29*(31), 7133–7155. doi:10.1021/bi00483a001

Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The Protein Folding Problem. *Annual Review of Biophysics*, *37*, 289–316. doi:10.1146/annurev.biophys.37.092707.153558

Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, *426*(6968), 884–890. Retrieved from http://dx.doi.org/10.1038/nature02261

Doig, A. J., & Baldwin, R. L. (1995). N- and C-capping preferences for all 20 amino acids in α-helical peptides. *Protein Science*, *4*(7), 1325–1336. doi:10.1002/pro.5560040708

Elber, R. (2015). Two Is a Pair, Three Is a Network. *Biophysical Journal*, *108*(1), 22.

Eletr, Z. M., Huang, D. T., Duda, D. M., Schulman, B. A., & Kuhlman, B. (2005). E2 conjugating enzymes must disengage from their E1 enzymes before E3-dependent ubiquitin and ubiquitin-like transfer. *Nature Structural & Molecular Biology*, *12*(10), 933–934.

Falkenberg, C., Bjoerck, L., & Aakerstroem, B. (1992). Localization of the binding site for streptococcal protein G on human serum albumin. Identification of a 5.5-kilodalton protein G binding albumin fragment. *Biochemistry*, *31*(5), 1451–1457.

Floudas, C. A. (2007). Computational methods in protein structure prediction.

Biotechnology and Bioengineering, *97*(2), 207–213. doi:10.1002/bit.21411

Fu, H., Grimsley, G. R., Razvi, A., Scholtz, J. M., & Pace, C. N. (2009). Increasing protein stability by improving beta-turns. *Proteins: Structure, Function, and Bioinformatics*, *77*(3), 491–498.

Gallagher, T., Alexander, P., Bryan, P., & Gilliland, G. L. (1994). Two Crystal Structures of the B1 Immunoglobulin-Binding Domain of Streptococcal Protein G and Comparison with NMR. *Biochemistry*, *33*(15), 4721–4729. doi:10.1021/bi00181a032

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). *Protein identification and analysis tools on the ExPASy server*. Springer.

Grant, A., Lee, D., & Orengo, C. (2004). Progress towards mapping the universe of protein folds. *Genome Biology*, *5*(5), 107. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC416458/

He, Y., Chen, Y., Alexander, P. A., Bryan, P. N., & Orban, J. (2012). Mutational tipping points for switching protein folds and functions. *Structure (London, England : 1993)*, *20*(2), 283–91. doi:10.1016/j.str.2011.11.018

He, Y., Chen, Y., Rozak, D. A., Bryan, P. N., & Orban, J. (2007). An artificially evolved albumin binding module facilitates chemical shift epitope mapping of GA domain interactions with phylogenetically diverse albumins. *Protein Science*, *16*(7), 1490–1494.

He, Y., Rozak, D. A., Sari, N., Chen, Y., Bryan, P., & Orban, J. (2006). Structure, Dynamics, and Stability Variation in Bacterial Albumin Binding Modules: Implications for Species Specificity†,‡. *Biochemistry*, *45*(33), 10102–10109. doi:10.1021/bi060409m

Johnson, L. B., Gintner, L. P., Park, S., & Snow, C. D. (2015). Discriminating between stabilizing and destabilizing protein design mutations via recombination and simulation. *Protein Engineering Design and Selection* , *28* (8), 259–267. doi:10.1093/protein/gzv030

Kohn, W. (1998). De novo design of α-helical coiled coils and bundles: models for the development of protein-design principles. *Trends in Biotechnology*, *16*(9), 379–389. doi:10.1016/S0167-7799(98)01212-8

Kraulis, P. J., Jonasson, P., Nygren, P.-Å., Uhlén, M., Jendeberg, L., Nilsson, B., & Kördel, J. (1996). The serum albumin-binding domain of streptococcal protein G is a three-helical bundle: a heteronuclear NMR study. *FEBS Letters*, *378*(2), 190–194. doi:10.1016/0014-5793(95)01452-7

Lazar, G. A., & Handel, T. M. (1998). Hydrophobic core packing and protein design. *Current Opinion in Chemical Biology*, *2*(6), 675–679. doi:10.1016/S1367-5931(98)80102-6

Lemer, C. M.-R., Rooman, M. J., & Wodak, S. J. (1995). Protein structure prediction by threading methods: Evaluation of current techniques. *Proteins: Structure, Function, and Bioinformatics*, *23*(3), 337–355. doi:10.1002/prot.340230308

Levitt, M. (1978). Conformational preferences of amino acids in globular proteins. *Biochemistry*, *17*(20), 4277–4285. doi:10.1021/bi00613a026

Lim, W. A., Farruggio, D. C., & Sauer, R. T. (1992). Structural and energetic consequences of disruptive mutations in a protein core. *Biochemistry*, *31*(17), 4324–4333. doi:10.1021/bi00132a025

Lindahl, M. . S. L. A. . L. A. . S. S. E. . E. I. A. . F. N. P. . N. N. . N. S. V. . G. M. B. . M. T. A. . R. A. I. . A. R. (1994). Crystal structure of the ribosomal protein S6 from Thermus thermophilus. *EMBO J.*, *13*, 1249–1254. doi:8137808

Liu, Y., & Kuhlman, B. (2006). RosettaDesign server for protein design. *Nucleic Acids Research* , *34* (suppl 2), W235–W238. doi:10.1093/nar/gkl163

Luo, X., Tang, Z., Xia, G., Wassmann, K., Matsumoto, T., Rizo, J., & Yu, H. (2004). The Mad2 spindle checkpoint protein has two distinct natively folded states. *Nature Structural & Molecular Biology*, *11*(4), 338–345.

Makhatadze, G. I., Loladze, V. V., Ermolenko, D. N., Chen, X., & Thomas, S. T. (2003). Contribution of Surface Salt Bridges to Protein Stability: Guidelines for Protein Engineering. *Journal of Molecular Biology*, *327*(5), 1135–1148. doi:10.1016/S0022-2836(03)00233-X

Matthews, B. W., Nicholson, H., & Becktel, W. J. (1987). Enhanced Protein Thermostability from Site-Directed Mutations that Decrease the Entropy of Unfolding. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(19), 6663–6667. doi:10.2307/30639

Meyerguz, L., Kleinberg, J., & Elber, R. (2007). The network of sequence flow between protein structures. *Proceedings of the National Academy of Sciences*, *104*(28), 11627–11632.

Mirny, L. A., & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *Journal of Molecular Biology*, *291*(1), 177–96. doi:10.1006/jmbi.1999.2911

Mitchinson, C., & Wells, J. A. (1989). Protein engineering of disulfide bonds in subtilisin BPN'. *Biochemistry*, *28*(11), 4807–4815. doi:10.1021/bi00437a043

Möglich, A., Krieger, F., & Kiefhaber, T. (2005). Molecular Basis for the Effect of Urea and Guanidinium Chloride on the Dynamics of Unfolded Polypeptide Chains. *Journal of Molecular Biology*, *345*(1), 153–162. doi:10.1016/j.jmb.2004.10.036

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., & Tramontano, A. (2009). Critical assessment of methods of protein structure prediction—Round VIII. *Proteins: Structure, Function, and Bioinformatics*, *77*(S9), 1–4.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., & Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP)—round x. *Proteins: Structure, Function, and Bioinformatics*, *82*(S2), 1–6.

Moult, J., Fidelis, K., Kryshtafovych, A., & Tramontano, A. (2011). Critical assessment of methods of protein structure prediction (CASP)—round IX. *Proteins: Structure, Function, and Bioinformatics*, *79*(S10), 1–5.

Murzin, A. G. (2008). Metamorphic proteins. *SCIENCE-NEW YORK THEN WASHINGTON-*, *320*(5884), 1725.

Myhre, E. B., & Kronvall, G. (1977). Heterogeneity of nonimmune immunoglobulin Fc reactivity among gram-positive cocci: description of three major types of receptors for human immunoglobulin G. *Infection and Immunity*, *17*(3), 475–482.

Nauli, S., Kuhlman, B., & Baker, D. (2001). Computer-based redesign of a protein folding pathway. *Nature Structural & Molecular Biology*, *8*(7), 602–605.

Orban, J., Alexander, P., Bryan, P., & Khare, D. (1995). Assessment of stability differences in the protein G B1 and B2 domains from hydrogen-deuterium exchange: Comparison with calorimetric data. *Biochemistry*, *34*(46), 15291–15300.

Pace, C. N. (1992). Contribution of the hydrophobic effect to globular protein stability. *Journal of Molecular Biology*, *226*(1), 29–35. doi:10.1016/0022-2836(92)90121-Y

Perry, L. J., & Wetzel, R. (1986). Unpaired cysteine-54 interferes with the ability of an engineered disulfide to stabilize T4 lysozyme. *Biochemistry*, *25*(3), 733–739.

Porter, L. L., He, Y., Chen, Y., Orban, J., & Bryan, P. N. (2015). Subdomain interactions foster the design of two protein pairs with ~80% sequence identity but different folds. *Biophysical Journal*, *108*(1), 154–62. doi:10.1016/j.bpj.2014.10.073

Prevost, M., Wodak, S. J., Tidor, B., & Karplus, M. (1991). Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the Ile-96----Ala mutation in barnase. *Proceedings of the National Academy of Sciences* , *88* (23), 10880–10884. Retrieved from http://www.pnas.org/content/88/23/10880.abstract

Pulido, M. A., Koga, Y., Takano, K., & Kanaya, S. (2007). Directed evolution of Tk-subtilisin from a hyperthermophilic archaeon: identification of a single amino acid substitution responsible for low-temperature adaptation. *Protein Engineering Design and Selection* , *20* (3), 143–153. doi:10.1093/protein/gzm006

Ramakrishnan, C., & Ramachandran, G. N. (1965). Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophysical Journal*, *5*(6), 909–933.

Richardson, J. S., & Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. *Science*, *240*(4859), 1648–1652.

Rossmann, M. G., & Argos, P. (1981). Protein folding. *Annual Review of Biochemistry*, *50*(1), 497–532.

Rost, B., Schneider, R., & Sander, C. (1997). Protein fold recognition by prediction-based threading. *Journal of Molecular Biology*, *270*(3), 471–480.

Ruan, B., Fisher, K. E., Alexander, P. A., Doroshko, V., & Bryan, P. N. (2004). Engineering Subtilisin into a Fluoride-Triggered Processing Protease Useful for One-Step Protein Purification†. *Biochemistry*, *43*(46), 14539–14546. doi:10.1021/bi048177j

Schimmel, P. R., & Flory, P. J. (1968). Conformational energies and configurational statistics of copolypeptides containing L-proline. *Journal of Molecular Biology*, *34*(1), 105–120.

Schwyter, D., Phillips, M., & Reisler, E. (1989). Subtilisin-cleaved actin: polymerization and interaction with myosin subfragment 1. *Biochemistry*, *28*(14), 5889–5895. doi:10.1021/bi00440a027

Tuinstra, R. L., Peterson, F. C., Kutlesa, S., Elgin, E. S., Kron, M. A., & Volkman, B. F. (2008). Interconversion between two unrelated protein folds in the lymphotactin native state. *Proceedings of the National Academy of Sciences*, *105*(13), 5057–5062.

Wickstrom, L., Okur, A., Song, K., Hornak, V., Raleigh, D. P., & Simmerling, C. L. (2006). The Unfolded State of the Villin Headpiece Helical Subdomain: Computational Studies of the Role of Locally Stabilized Structure. *Journal of Molecular Biology*, *360*(5), 1094–1107. doi:10.1016/j.jmb.2006.04.070