Spring 2021

Computational and Experimental Investigation into the Determinants of Protein Structure, Folding, and Stability in the β-Grasp Superfamily

John T. Bedford II
Old Dominion University, jbedford@odu.edu

COMPUTATIONAL AND EXPERIMENTAL INVESTIGATION INTO THE DETERMINANTS OF PROTEIN STRUCTURE, FOLDING, AND STABILITY IN THE β-GRASP SUPERFAMILY

by

John T. Bedford II
B.S. December 2012, Old Dominion University
B.S. December 2012, Old Dominion University
M.S. August 2016, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

CHEMISTRY

OLD DOMINION UNIVERSITY
May 2021

Approved by:

Lesley H. Greene (Director)

Yaohang Li (Member)

Chris Osgood (Member)

Steven Pascal (Member)

Jennifer Poutsma (Member)

ABSTRACT

COMPUTATIONAL AND EXPERIMENTAL INVESTIGATION INTO THE DETERMINANTS OF PROTEIN STRUCTURE, FOLDING, AND STABILITY IN THE β-GRASP SUPERFAMILY

John T. Bedford II
Old Dominion University, 2021
Director: Dr. Lesley H. Greene

Elucidating the mechanisms of protein folding and unfolding is one of the greatest scientific challenges in basic science. The overarching goal is to predict three-dimensional structures from their amino acid sequences. Understanding the determinants of protein folding and stability can be facilitated through the study of evolutionarily related but diverse proteins. Insights can also be gained through the study of proteins from extremophiles that may more closely resemble the primordial proteins. In this doctoral research, three aims were accomplished to characterize the structure, folding and unfolding behavior within the β-grasp superfamily. We propose that the determinants of structure, stability, and folding are conserved as sequence and interaction patterns in the β-grasp fold. To elucidate key residues, bioinformatics studies were conducted and identified nine structurally conserved amino acids in the core of the B1 domain of protein G (GB1). A network analysis of all long-range interactions in the structure of GB1 revealed the relative significance of each conserved amino acid. Within the β-grasp superfamily, two proteins, GB1 and the small archaeal modifier protein 1 (SAMP1), were investigated to elucidate the key determinants of structural stability at the level of individual interactions. They were subjected to high temperature molecular dynamics simulations and the detailed behavior of each long-range interaction was characterized. The results revealed that in GB1 the most stable region was the C-terminal hairpin and in SAMP1 it was the opposite, the N-terminal hairpin. The

folding behavior of SAMP1 was investigated due to its nature as a divergent superfamily member and extremophile. The results revealed that SAMP1 at high ionic strength folds more rapidly than in low ionic strength. These findings clearly indicate that adaption at high salt produces rapid and less-frustrated folding. The results of these research aims provide insight into determinants of the β-grasp fold and the folding and unfolding behavior of two key members. Perhaps the most surprising finding is the presence of a significant number of non-native long-range interactions during unfolding which has largely gone unnoticed in the scientific community and appears to be pivotal.

This dissertation is dedicated to my parents John Jeffrey and Charlene Renee Cole Bedford for their constant love, support, and encouragement throughout life, especially during my time in the doctoral program. I would not have made it this far without them.

ACKNOWLEDGMENTS

First and foremost, I thank my Lord and Savior Jesus Christ for blessing me with this opportunity and providing me all I needed to fulfill it. I'd like to thank my family for their prayers and motivation and for pushing me to continue toward this goal when all I wanted to do was give up. Thank you for attending all of my presentations; it was a relief to have familiar faces in the audience cheering me on. I also thank my friends for being willing to listen and for supporting me through the highs and lows of graduate school.

I would like to sincerely thank my advisor and mentor Dr. Lesley H. Greene for all of her support and guidance. Thank you for giving me the opportunity to do undergraduate research with your group. It proved to be a major factor in my decision to continue into graduate school. The constant love and patience you exhibit are a real blessing to everyone. Thank you for always having time to listen and for all the pep talks, lunches, and gifts for every occasion. You helped to humanize the graduate school experience.

I would like to thank the members of my dissertation committee. Dr. Jennifer Poutsma for teaching me molecular dynamics, without which part of this dissertation would not have been possible. Dr. Steven Pascal, Dr. Christopher J. Osgood, and Dr. Yaohang Li for their contributions in guiding me through this process and for helping to shape me into a better scientist. I thank my colleagues in the Greene group, both past and present, for their helpful recommendations and support. I would like to thank my collaborators Dr. Heinrich Roder, Dr. Takuya Mizukami, Dr. ShanHui Liao, and Dr. Norou Diawara for their expert help in completing my research aims. I also thank the members of the ODU ITS department for their priceless assistance.

The dissertation research of J.T.B.II is supported in part by funding from two Old Dominion University CIBA Fellowships, two Virginia Space Grant Consortium Research Fellowships, the Old Dominion University Van Norman Award, the Old Dominion University Student Engagement and Enrollment Services Student Travel Award, and the International Symposium on Bioinformatics Research and Applications Travel Fellowship.

NOMENCLATURE

3D	Three-Dimensional
BC	Betweenness Centrality
CATH	Class, Architecture, Topology, Homology
CD	Circular Dichroism
CHARMM	Chemistry at HARvard Macromolecular Mechanics
CI2	Chymotrypsin Inhibitor 2
CO	Contact Order
DNA	Deoxyribonucleic Acid
GA	Albumin-binding Domain of Protein G
GB1	Immunoglobulin-binding Domain of Protein G
IPTG	Isopropyl-β-D-1-thiogalactopyranoside
LB	Luria Bertani
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PID	Percent Identity
RMSD	Root Mean Square Deviation
SAMP1	Small Archaeal Modifier Protein 1
SCOP	Structural Classification of Proteins database
SIAS	Sequence Identity and Structure
TS	Transition State

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

OVERVIEW OF PROTEINS

Protein Structure

Protein folding is an area of research that has caught the attention of many researchers from different disciplines in the scientific community. The field was most notably brought into the public spotlight when researchers sequenced the human genome [1-4]. Proteins play a crucial role in the onset and sustainability of all life in both mesophilic and extremophilic conditions. Proteins are one of four main classes of biological molecules, the others being carbohydrates, lipids, and deoxyribonucleic acid (DNA). They are polymers composed of monomeric units called amino acids. These polymers in the context of proteins are called polypeptide chains. There are 20 naturally occurring amino acids found in nature (Appendix A). As a protein is synthesized by the ribosome, it progresses through several structural stages in a hierarchical fashion to reach its final functional form (Figure 1).

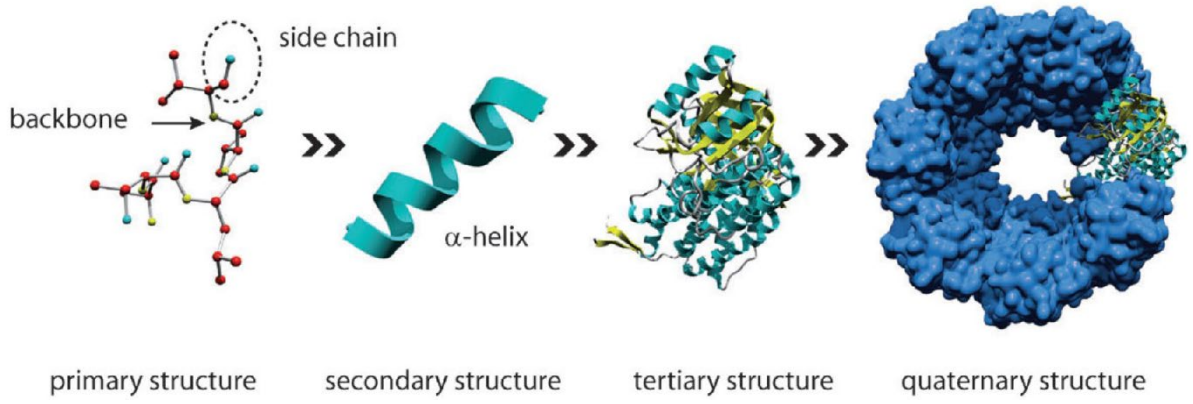


Figure 1. Hierarchical levels of protein structure. Protein structures were visualized using Pymol (version 2.1.1). Figure adapted from [5] and used with permission.

The primary structure is its linear chain of amino acids connected via peptide bonds. A peptide bond is formed via a condensation reaction when the amide nitrogen of one amino acid is deprotonated and the carbonyl carbon of another amino acid is dehydroxylated. The primary structure then arranges itself into secondary structural elements including α-helices, β-sheets, and β-turns. The tertiary structure is defined as the coalescence of secondary elements into a protein's overall structure in three-dimensional (3D) space. These structures are classified into three main groups: all α-helical, all β-sheet, and mixed α/β (Figure 2). It has been proposed that the mixed *α/β* class of proteins may be in general the older of the three protein classes [6]. These three classes of proteins however have a vast amount of variability in nature. Most proteins adopt one of ten main superfolds found in nature (Figure 3).

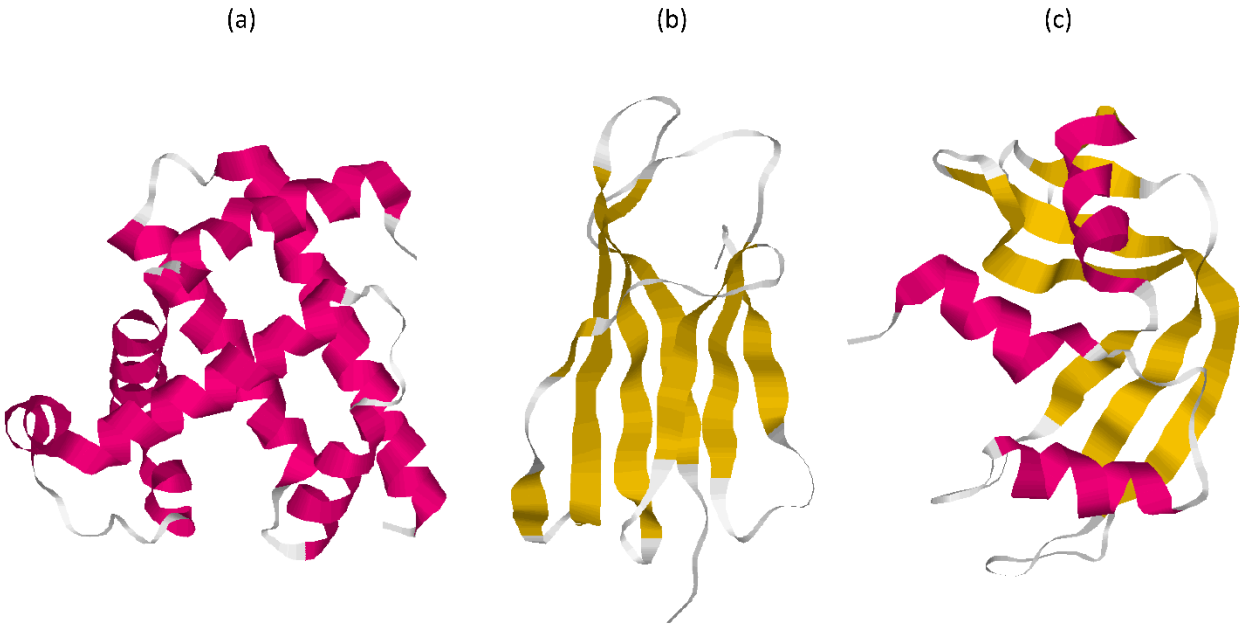


Figure 2. Protein classes. (a) all-α, (b) all-β, (c) mixed α/β. α-helices, β-strands and loops are shown in magenta, yellow, and white, respectively. Structures visualized using RasMol Ver. 2.7.2.1.1.

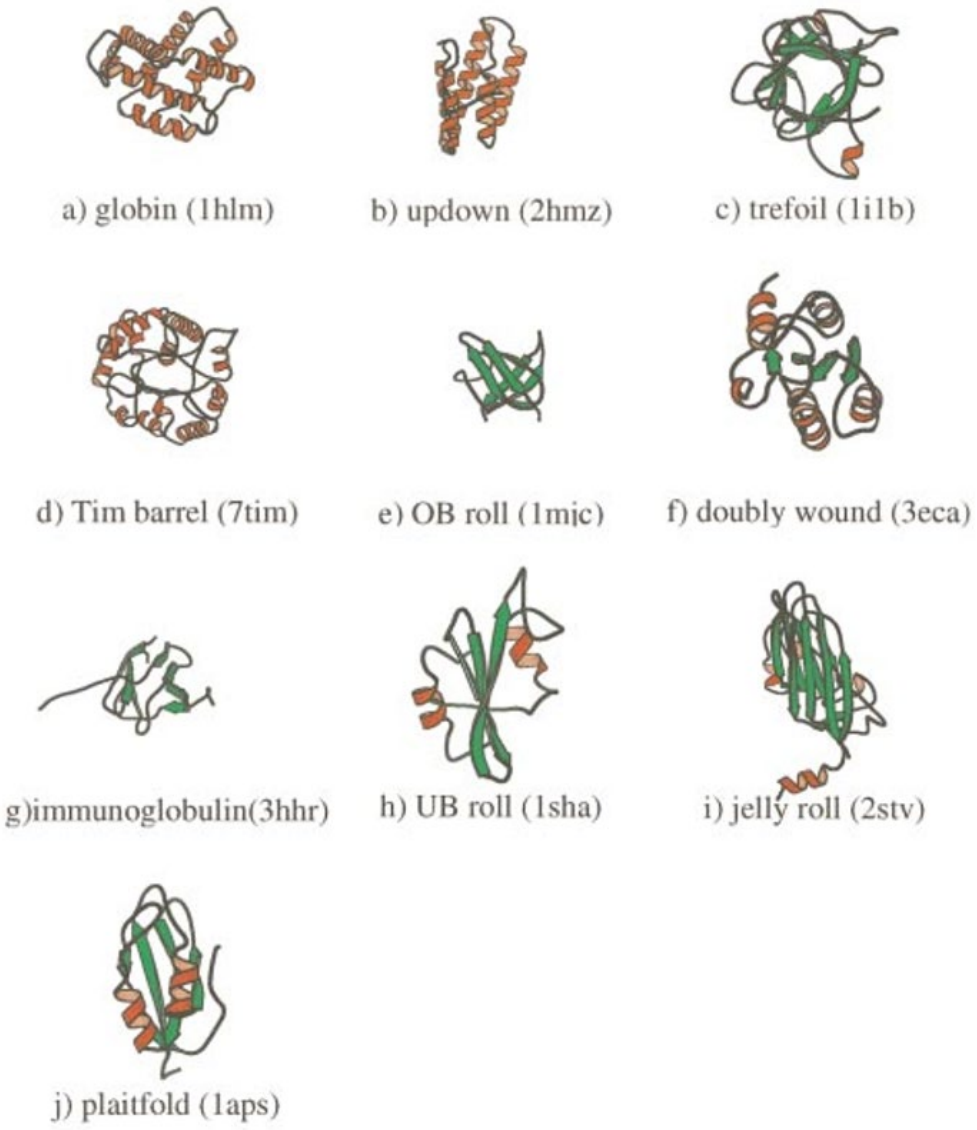


Figure 3. Representations of the ten superfolds. Figure reproduced from [7] and used with permission.

Although monomeric proteins can be functional in their in tertiary form, some proteins need to form complexes comprised of multiple polypeptide chains to be functional. This is

defined as the quaternary structure. At the time of this publication, the Protein Data Bank (PDB), a repository for elucidated protein structures, contains over 170,970 biological macromolecular structures. Yet, this vast array of structures share common folds, motifs, and topologies.

Protein Interactions

Protein structures are stabilized by various types of interactions. Secondary structural elements are formed through local, short-range interactions and are primarily stabilized through hydrogen bonding. Local or short-range interactions are those between residues that are close in sequence and 3D space. A protein's tertiary structure is stabilized through an assortment of non-local, long-range interactions including but not limited to hydrophobic interactions, salt bridges, disulfide bonds, hydrogen bonds, and van der Waals interactions (Figure 4). Non-local or long-range interactions are those between residues that are distant in sequence but still close in 3D space. Long-range interactions are more important for structuring of the native state and its overall stability [8-15].

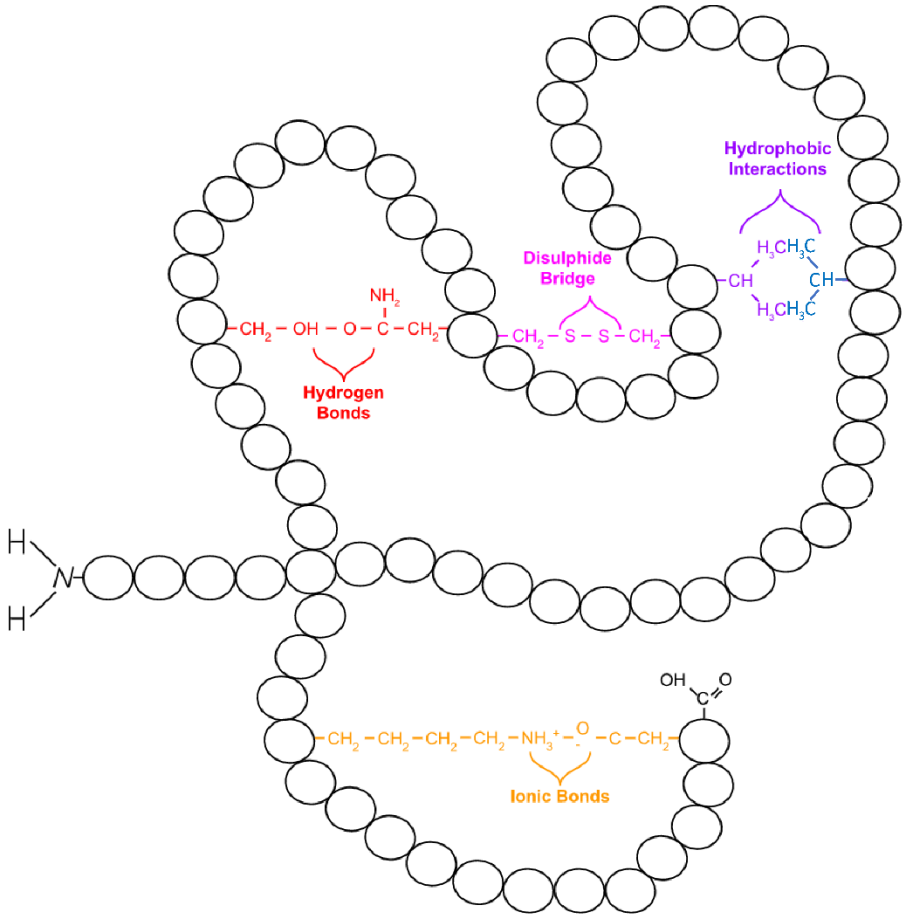


Figure 4. Stabilizing interactions of tertiary structure.

These interactions are also important for determining a protein's folding rate. The number of contacts and their location also play a key role. Proteins containing more local, short-range contacts will generally fold faster than those with more non-local, long-range contacts [16]. The importance of long- and short-range interactions in a protein's native structure and how they affect the protein's folding rate can be assessed using contact order (CO) [16-20]. CO is the average sequence separation between interacting residues normalized by the total sequence length. CO can be calculated using equation 1, where N is the total number of contacts, $\Delta S_{i,j}$ is

the sequence separation between interacting amino acids *i* and *j*, and *L* is the total number of amino acids [16].

$$CO = \frac{1}{L*N} \sum_{i=0}^{N} \Delta S_{i,j} \quad (1)$$

Proteins that have slower folding rates with ordered transition states and larger non-local interaction networks have higher CO values [16-20]. Early long-range interaction formation could allow a more stable native structure to be formed by slowing down the folding rate.

Protein Folding

Cyrus Levinthal proposed that a 100 amino acid protein, sampling one possible conformation every 10^{-13} seconds, would take 10^{27} years to find the correct native fold [21, 22]. Therefore, he concluded that the process of protein folding must be ordered and not random. Later work done by Anfinsen, for which he won the Nobel Prize, indicated that amino acid interactions are the sole determinant of protein structure [23, 24]. A profound development in the field of proteomics was the use of nuclear magnetic resonance spectroscopy (NMR) and X-ray crystallography to solve the structure of proteins with atomic resolution. These solved structures are stored in the aforementioned PDB so that they are publicly accessible. Once proteins are visualized, they can be classified based upon their topology into families and superfamilies. Two main databases are predominately used: CATH (Class, Architecture, Topology, and Homology) and SCOP (Structural Classification of Proteins) [2, 3]. Five mechanisms have been proposed to describe the protein folding process. In the hydrophobic collapse model, non-polar amino acids form a hydrophobic core followed by the formation of secondary elements around the core to form the native structure (Figure 5). In the framework model, secondary elements are formed first and are then assembled into the native conformation (Figure 5). In the nucleation-

condensation model, secondary elements collapse to form a folding nucleus, which the remaining polypeptide orients around resulting in the native state (Figure 5). In the jigsaw model, there are many different pathways which an unfolded protein can take to reach its native conformation (Figure 5).

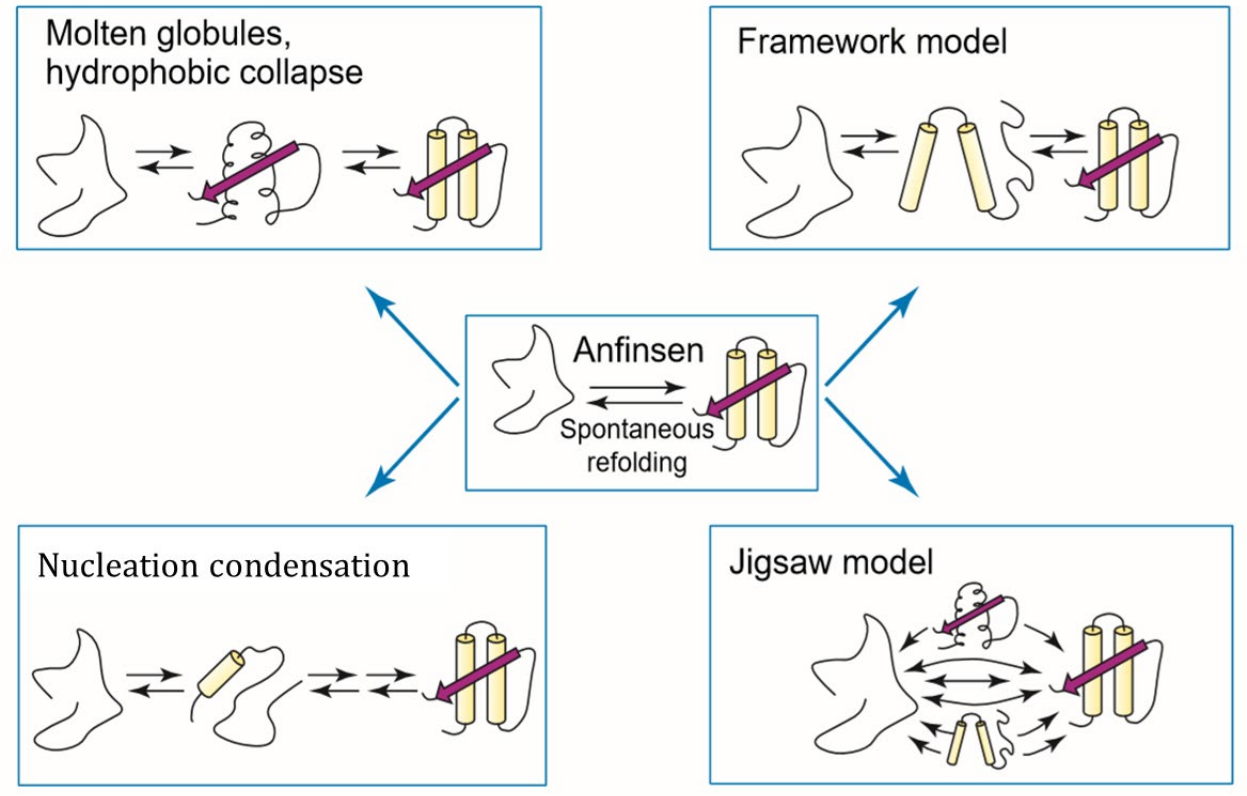


Figure 5. Proposed protein folding models: hydrophobic collapse model, framework model, nucleation-condensation model, and jigsaw model. Figure adapted from [25] and used with permission.

First explored by Alan Fersht and co-workers at University of Cambridge the nucleation-condensation mechanism has been the focus of many experimental studies [26, 27]. Their work with the chymotrypsin inhibitor 2 (CI2) helped to support this mechanism. They showed the development of nucleation site in the transition state (TS) during the folding process. Using Φ-value analysis, the nucleus was determined to be composed of an α-helix stabilized by long-range interactions to the remaining protein structure. This analysis eliminates or reduces amino acid interactions by reduction of the side chain. The mutant protein's interactions are then reassessed during the folding and unfolding process using kinetic and equilibrium techniques [26, 27]. Φ-value is the ratio of changes in the folding free energy of activation ($\Delta\Delta G_{‡\text{-}D}$) and the folding equilibrium free energy ($\Delta\Delta G_{N\text{-}D}$), as seen in equation 2 [28].

$$\Phi_F = \frac{\Delta\Delta G_{‡-D}}{\Delta\Delta G_{N-D}} \quad (2)$$

Φ-values range from 0 to 1, where a Φ-value of 0 is indicative of a mutation in which the TS is not affected and thus the interaction does not form in the TS. A Φ-value of 1 is indicative of a mutation where the TS is affected and thus the interaction is present in the TS (Figure 6).

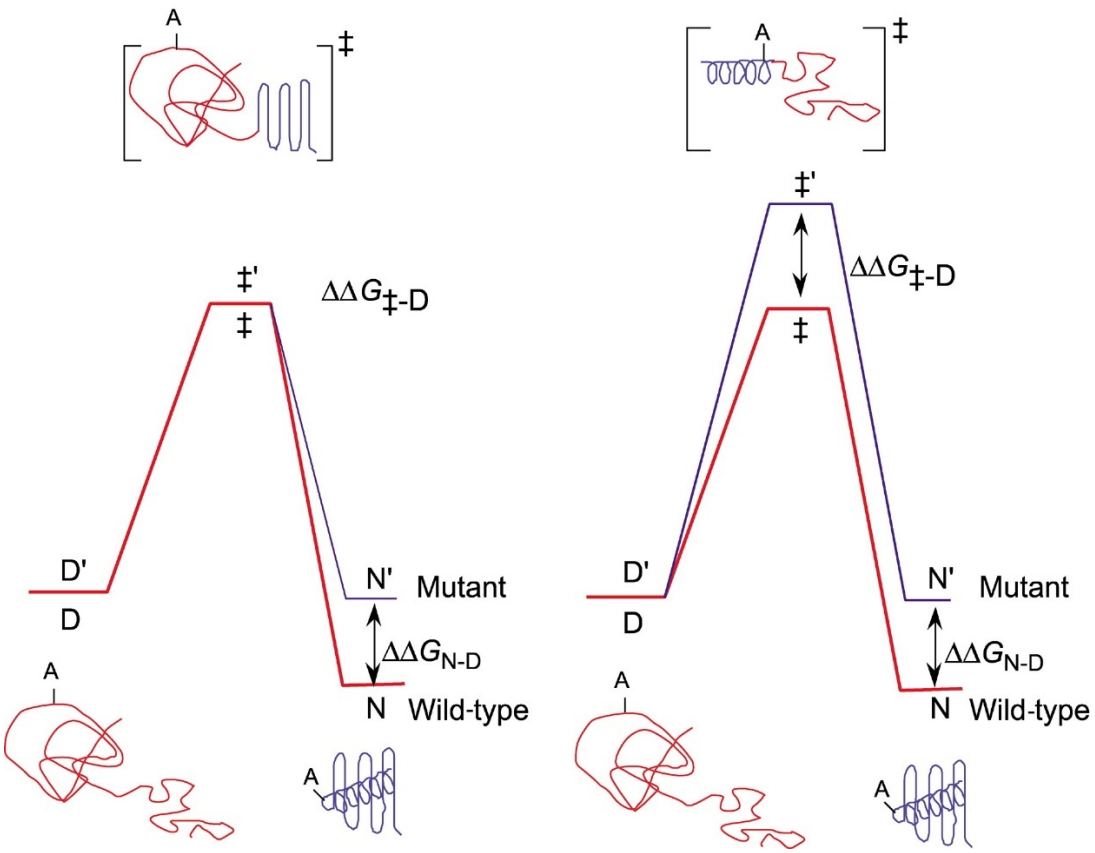


Figure 6. Φ-value diagrams. The left diagram shows a Φ-value of 0 and the diagram on the right shows a Φ-value of 1. Figure reproduced from [29] and used with permission.

The protein folding funnel model starts with numerous unfolded peptide conformations with few native interactions in a high-energy, high-entropy state. As the protein proceeds down the funnel-shaped energy landscape, conformational space is restricted causing an increase in the number of contacts, the result of which is a low-energy, low-entropy native state (Figure 7) [30, 31].

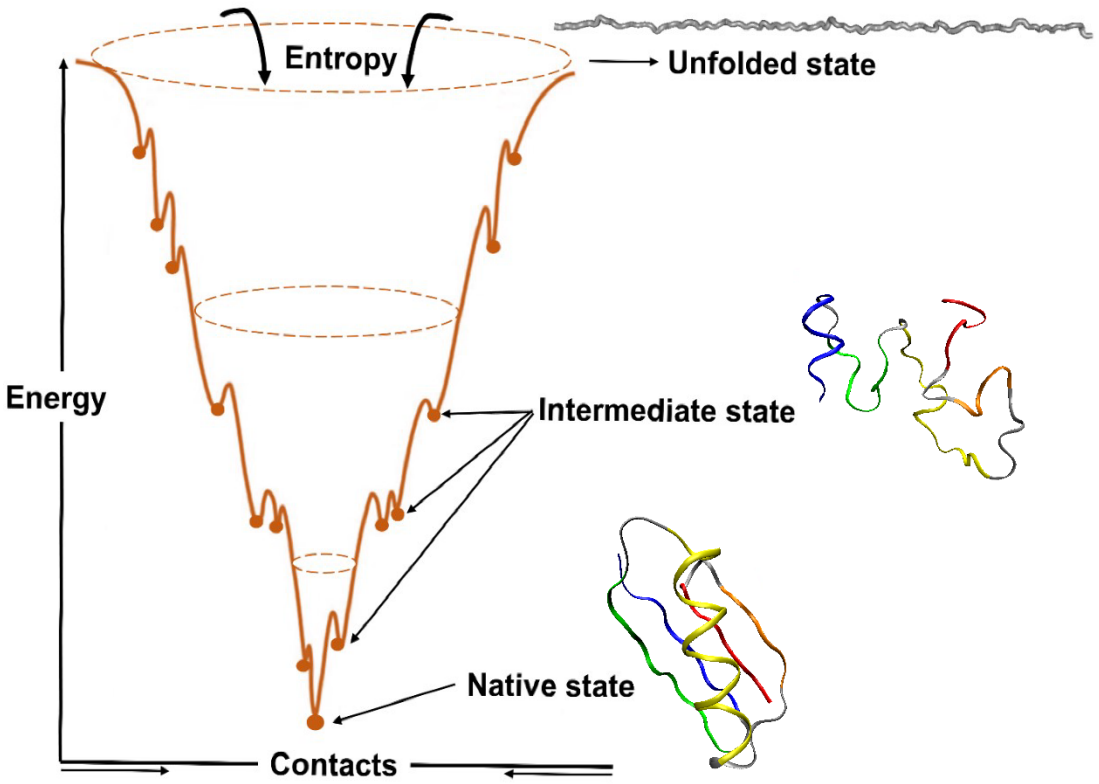


Figure 7. The schematic of the folding energy landscape funnel. The folding of GB1 (PDB code: 1PGB) is shown here. Figure adapted from [32] and used with permission. Copyright (1998) National Academy of Sciences, U.S.A.

Proteins fold along an energy landscape by forming in large part specific non-covalent short- and long-range interactions in an ordered process which results in the proper organization of structural components into a native conformation. This transition is not a smooth one; the funnel consists of many energy wells in which the protein adopts a misfolded conformation. There are many factors that influence the folding behavior of proteins, including size, shape, and stability [26, 33-51].

In Vivo Protein Folding and the Role of the Ribosome and Chaperones

In most cells, protein synthesis occurs on the ribosome. The ribosome is comprised of two subunits, the small subunit is responsible for reading the incoming mRNA and the large subunit is responsible for the elongation of the polypeptide chain [52]. Protein folding can initiate while still inside the ribosome [53]. As a protein's polypeptide chain is elongated it begins to exit the ribosome. If the protein is small enough it can fold inside the ribosome as it exits, however if the protein is too large folding will occur once the polypeptide chain exits the ribosome without assistance (Figure 8(a)). The process of folding outside the ribosome supports the *in vitro* folding of isolated proteins [53-58].

If a protein misfolds during this process the result may be a loss of function or the formation of certain disease states such as Parkinson's and Alzheimer's. Hsp70 or trigger factor may aid in the folding process as a means of prevention (Figure 8(b)). Another group of proteins that aid in the folding process are chaperones. They function by binding partially folded or misfolded proteins to help them reach their correct native fold (Figure 8(c)) [53]. Archetypal examples of a chaperonin systems are GroEL/GroES in prokaryotes and TRiC/CCT in eukaryotes. The structure of this system consists of two rings with a central cavity and a cap, where the partially folded or misfolded protein is unfolded and correctly folded [59].

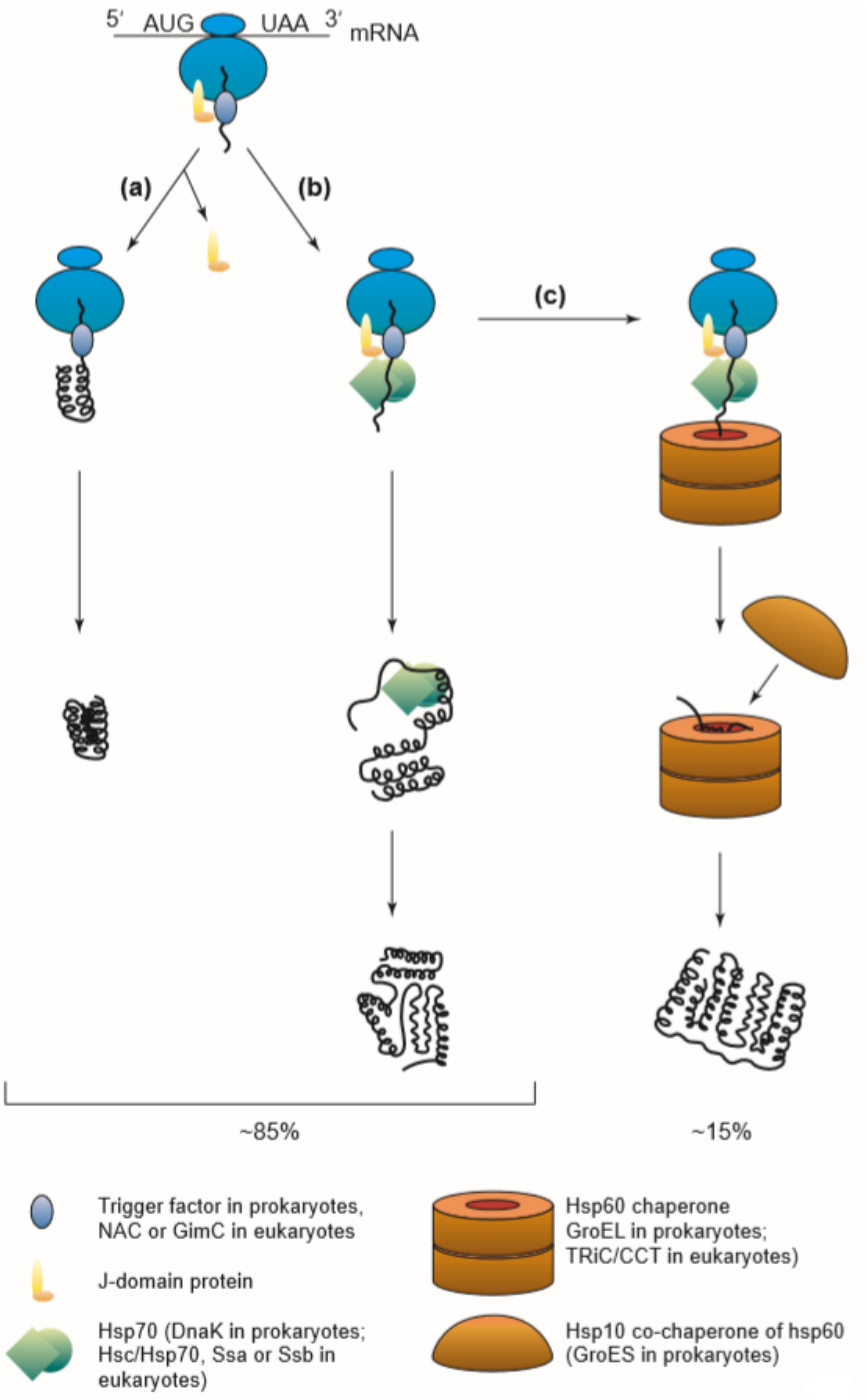


Figure 8. Cytosolic *de novo* folding in prokaryotes and eukaryotes. (a) folding that is independent of hsp60 and hsp70. (b) folding assisted by hsp70 or trigger factor. (c) folding that is assisted by either hsp70 or trigger factor and hsp60 chaperonin. Figure reproduced and figure legend adapted from [25] and used with permission.

The β-grasp superfamily

The aim of this doctoral research is to elucidate determinants of structure, folding, and stability using the β-grasp superfamily as a model system. The β-grasp superfamily encompasses a vast array of proteins that occupy seven distinct branches of the evolutionary tree [60]. These proteins have diverse functions which is largely attributed to the β-sheet. Despite this diversity, the proteins belonging to this superfamily share a common fold, termed the β-grasp fold, because the β-sheet appears to grasp the α-helix (Figure 9). The research presented in this dissertation will focus on two members of the β-grasp superfamily, the immunoglobulin-binding domain of protein G (GB1) from *Streptococcus sp*. and the small archaeal modifier protein 1 (SAMP1) from *Haloferax volcanii*.

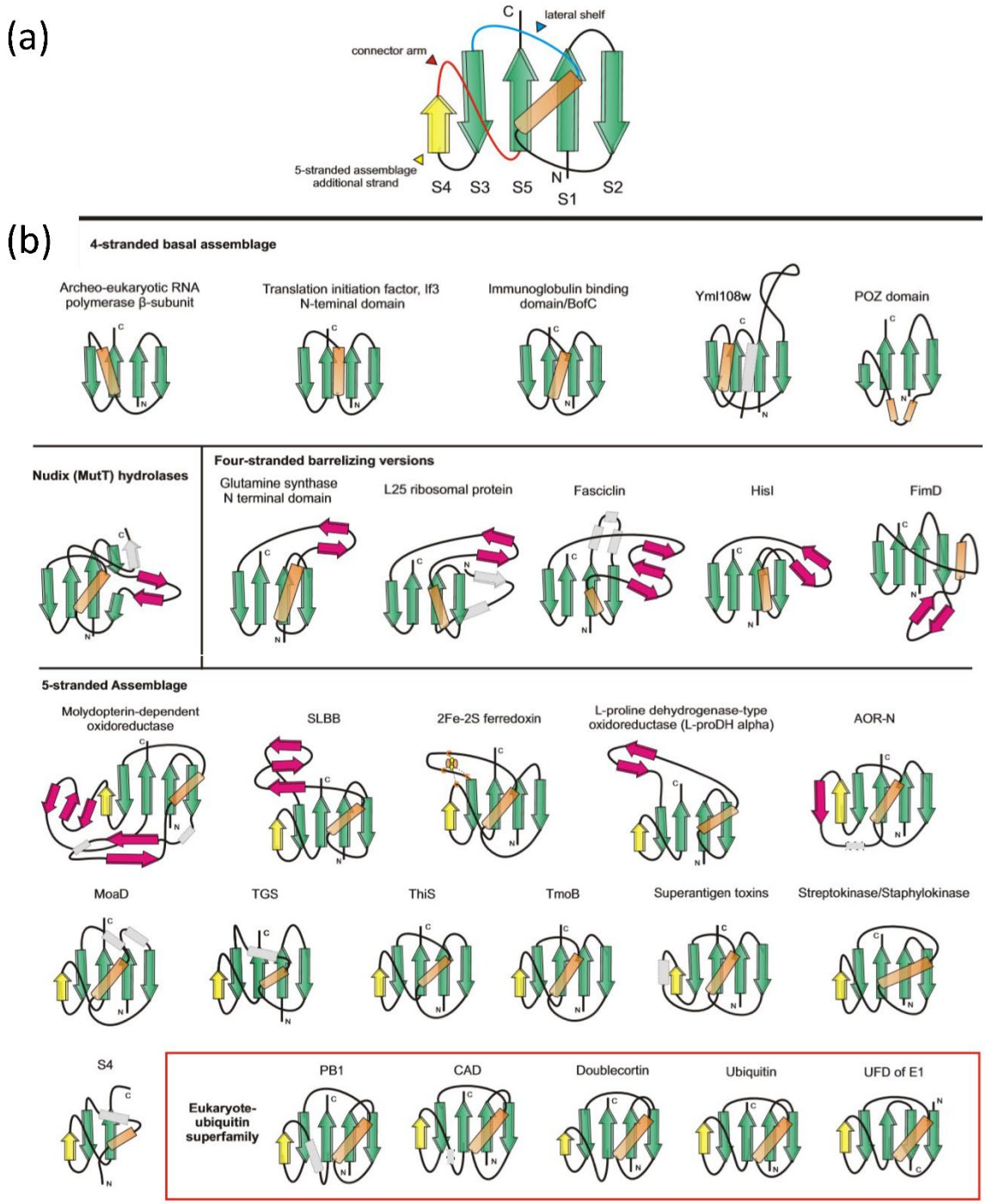


Figure 9. Topology diagrams of select β-grasp superfamily members. (a) The four-stranded β-sheet and core α-helix that are conserved among all members are shown in green and orange respectively. (b) Embellishments to the core structure are shown in yellow, magenta, and grey. Figure reproduced from [60] and used with permission.

GB1

GB1 was selected as a model protein for this research because of its intrinsic characteristics and its history of being well studied both computationally [61-72] and experimentally [73-84]. It is a small, 56 residue protein containing a four-stranded β-sheet packed against an α-helix (Figure 10). The two hairpins and helix form a symmetrical fold that is rarely seen amongst proteins. It is isolated from *Streptococcus sp*., a mesophilic organism.

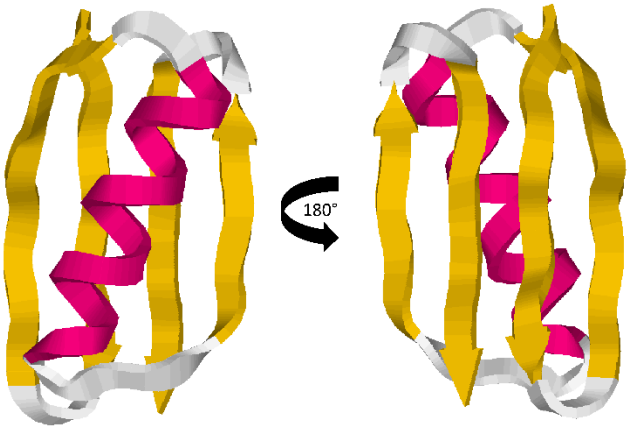


Figure 10. The immunoglobulin-binding domain of protein G (PDB code: 1PGB). The α-helix and β-sheet are shown in magenta and yellow, respectively. Structures visualized using RasMol Ver. 2.7.2.1.1.

SAMP1

SAMP1 was chosen as a model protein because of its topological similarity to GB1 and because it contains structural embellishments, in part, due to its belonging to *Haloferax volcanii*, a halophilic organism isolated from the silt sands of the Dead Sea in Israel. It is an 87-residue

protein containing a four-stranded β-sheet and three α-helices (Figure 11). Two of the helical elements are embellishments of the common β-grasp fold.

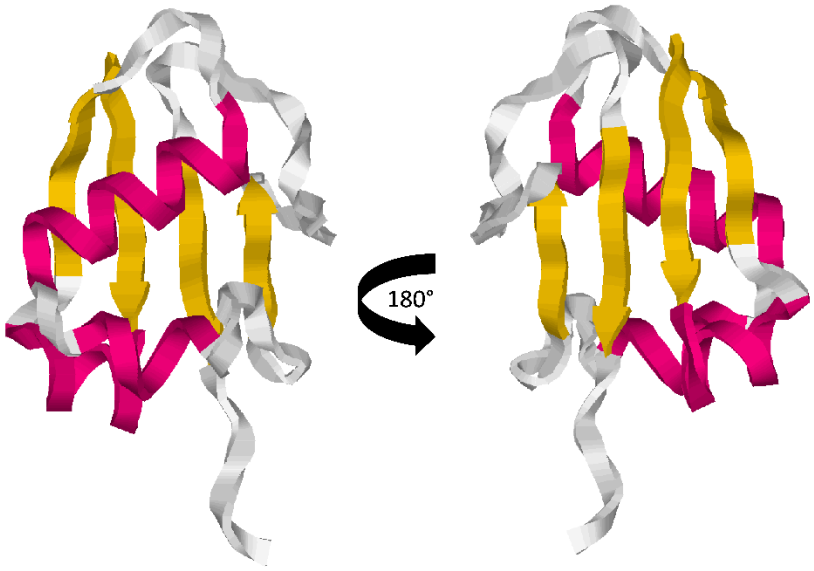


Figure 11. The small archaeal modifier protein 1 (PDB code: 3PO0). The α-helix and β-sheet are shown in magenta and yellow, respectively. Structures visualized using RasMol Ver. 2.7.2.1.1.

RESEARCH AIMS

The aim of this dissertation was to conduct a thorough investigation into the determinants of structure, folding, and stability among model members of the β-grasp superfamily. This is necessary because in over five decades significant advancements have been made in our understanding but the fundamental questions of how proteins fold, how are they stabilized, and

how is the structure predicted remain unresolved. In aim one, a bioinformatics analysis was completed on select proteins in the β-grasp superfamily in order to identify the nature of conserved residues proposed as critical determinants for folding and stability. While numerous computational and experimental studies have been performed to analyze these determinants [82, 85-94], our studies are focused on identifying and characterizing the role of conserved residues using bioinformatics. Conserved residues have been analyzed in a number of different proteins from different superfamilies [95-106]. Once identified, we are able to characterize them using network principles which provides a depth of understanding [107-110]. This approach is more unique in the protein folding field and has more recently become a very valuable way to analyze proteins and investigate determinants of folding. The idea is that what is conserved in proteins that differ in function and sequence identity but are related by a common ancestor and share the same overall topology is a key determinant of the folding and structure. More specifically, we examine long-range interactions using network approaches, which we propose are central to encoding the native 3D structure [13, 18, 111-117]. In aim two, molecular dynamics was used to unfold GB1 and SAMP1 to elucidate the determinants of stability and map the unfolding process. These studies involve an analysis of every long-range interaction which is computationally intensive and rigorous and is largely missing in studies of GB1 and other members of the superfamily. Thus, our research will provide the most comprehensive view of the unfolding process at atomic level resolution. The analysis consists of calculating the average persistence of each long-range interaction over the course of multiple unfolding simulations. Long-range interactions that are found to be among the most persistent are proposed to be important for the formation of the β-grasp fold. The last aim was to conduct a biophysical analysis of SAMP1 using kinetic and equilibrium techniques to characterize the folding kinetics

and establish a divergent member as a future model. SAMP1 is a halophilic protein and to the best of our knowledge no protein stability, folding and unfolding studies have been performed. Our experiments will characterize the stability and kinetic behavior of SAMP1 at different sodium chloride concentrations. This provides the first insight into an extremophile in the β-grasp fold and establishes the foundation for future experimental studies to elucidate the role of conserved residues and long-range interactions which we have identified computationally in aims one and two.

COMPUTATIONAL AND EXPERIMENTAL METHODOLOGY

Bioinformatics

Bioinformatics is the union between computers, biology, and chemistry. It allows large quantities of sequence information to be analyzed to find patterns and determinants of biological processes such as genomic sequencing of tumors to identify causative mutations that direct chemotherapy treatments. It is also used to identify adaptive changes in organisms during evolution or from environmental challenges. It is used to track viruses and mutations that can lead to enhanced virulence and infectivity, and to study protein sequences and structures to gain insight into their functional and structural behavior. The most common bioinformatics tools are the position-specific iterated basic local alignment search tool which utilizes algorithms to identify protein families and superfamilies [118], DaliLite which utilizes structural information found in the PDB to search for other proteins that contain a similar structure [119], and MUSCLE, a sequence based alignment program [120]. Visualization programs such as RasMol, PyMOL, and VMD are also used to analyze structural details. All of these programs allow one to interrogate in great detail, macromolecular sequence and structure information.

Network Science

A network is a system of interconnected nodes. Interconnected systems like social networks, businesses, and systems even as basic as power distribution grids have been the subject of network analysis to understand their development, robustness, and dynamics [107, 108, 121, 122]. This concept can also be applied to proteins. A protein is a network of amino acids (nodes) interconnected through various types of interactions [108]. An upward trend can be seen of network science helping to answer questions about protein structure, stability, and folding [123]. Network science is suited for such a task [124-130].

The way to begin to analyze a protein as a network system is to calculate interactions between amino acids. Here, an amino acid is a node and the interaction, a link. These interactions can be short- or long-range and consist of hydrogen bonds, van der Walls forces, hydrophobic forces, and salt bridges.

One very powerful approach to analyzing protein structure networks is to apply the concept of betweenness centrality (BC). BC is a measure of the total number of shortest paths between all pairs of nodes that pass through a specific node. Nodes with a high BC value play a critical role in network connectivity. This value concept can be applied to many systems including proteins [109, 131].

Molecular Dynamics

As detailed and amazing as protein structures are to observe, viewing them in motion is even more so. This can be achieved using molecular dynamics (MD). MD utilizes computer-generated force fields to simulate the *in vivo* movement of proteins and molecules on a picosecond time scale by assigning random velocities appropriate for a given temperature to each

atom in a simulation. These atoms then move in response to forces acting on them, which are determined by Newton's equations of motion [132]. Some popular force fields include CHARMM, AMBER, and GROMOS [133-139].

The dynamic equation for motion used for MD simulation is derived from the following. The position of atoms can be propagated forward using equation 3 given the atoms initial positions, $x_i(t_0)$, and their respective velocities, $v_i(t_0)$ at time t_0.

$$x_i(t_1) = x_i(t_0) + v_i(t_0)\Delta t \tag{3}$$

New velocities can then be calculated using equation 4.

$$v_i(t_1) = v_i(t_0) + \Delta v_i(t_0) \tag{4}$$

Using equations 5 and 6 and Newton's equation (F = ma or F = mdV/dt) the change in velocity can be calculated.

$$\Delta v_i(t_0) = \frac{F_{i(t_0)}}{m_i}\Delta t \tag{5}$$

$$v_i(t_1) = v_i(t_0) + \frac{F_{i(t_0)}}{m_i}\Delta t \tag{6}$$

where F_i is the sum of the forces acting on the i^{th} particle, Thus,

$$F(r) = -\nabla U(r) \tag{7}$$

$$U(r) = \Sigma U_{bonded}(r) + \Sigma U_{nonbonded}(r) \tag{8}$$

$$U_{bonded} = U_{bond} + U_{angle} + U_{dihedral} + U_{improper} \tag{9}$$

$$U_{nonbonded} = U_{LJ} + U_{elec} \tag{10}$$

$$U(r) = \Sigma_{bonds} K_b(b - b_0)^2 + \Sigma_{angles} K_\Theta(\Theta + \Theta_0)^2 + \tag{11}$$

$$\Sigma_{dihedral} K_\Phi[1 + \cos(n\Phi - \delta)] + \Sigma_{impropers} K_\omega(\omega - \omega_0)^2 +$$

$$\Sigma_{Urey-Bradly} K_\mu(\mu + \mu_0)^2 + \Sigma_{nonbonded}\epsilon[(\frac{R_{min,i,j}}{r_{rj}})^{12} - \left(\frac{R_{min,jj}}{r_{rj}}\right)^6] + \Sigma_{nonbonded}\frac{q_i q_j}{\epsilon r_{ij}}$$

In the first term of equation 11, K_b is the bond force constant and $b\text{-}b_0$ is the distance from equilibrium for a given bond. In second term K_θ is the angle force constant and $\theta + \theta_0$ is the degrees from equilibrium for a given angle. In the third term, K_Φ is the dihedral force constant, n is the multiplicity, Φ is the dihedral angle, and δ is the phase shift. In the fourth term, K_ω is the force constant and $\omega - \omega_0$ is the out of plane angle. In the fifth term, K_μ is the force constant and $\mu + \mu_0$ is the distance from equilibrium of the 1,3-nonbonded interactions. In the sixth term, ε is the electric permittivity constant, r_{ij} is the distance between two nonbonded atoms in the configuration and $R_{in,ij}$ is the constant distance at which the potential is zero. In the last term, q_i and q_j are partial charges of atoms i and j. Figure 12 describes some of these terms.

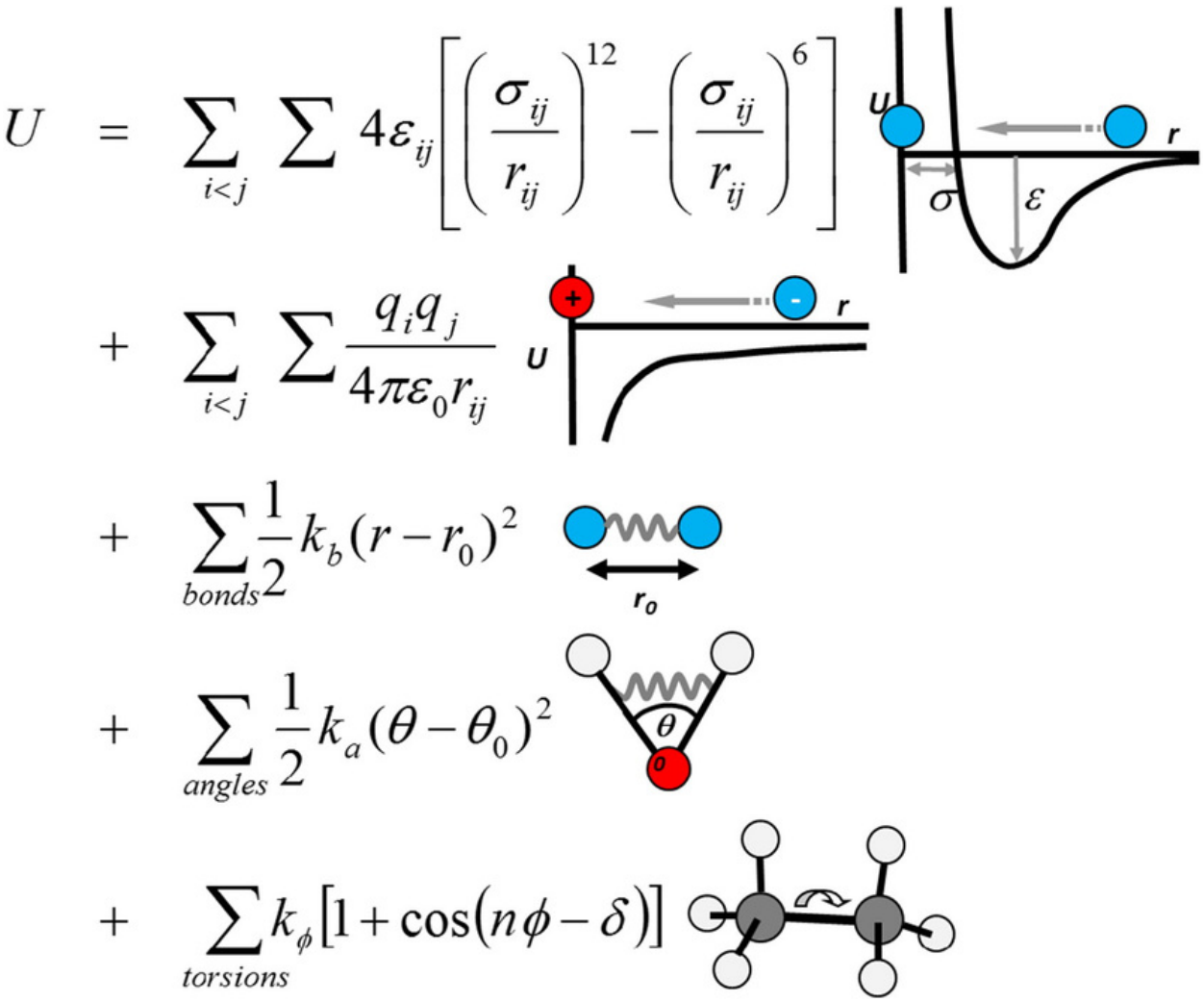


Figure 12. Schematic of select terms describing potential energy. The molecular mechanics potential energy function comprising the van der Waals (term 1, Lennard-Jones) and coulombic (term 2) interactions, and the three valence terms, bond, angle bending, and dihedral energy. The summations for van der Waals and coulombic terms indicate all pairwise interactions between atoms that are not either bonding or linked via a bond angle. The Lennard Jones parameters εij and σij, partial charges qi and qj, and the force constants kb, ka, and kϕ are all atom-specific parameters that comprise the force field and are inputs to the simulation. Figure and figure legend reproduced from [140] and used with permission.

One of the major limitations of using MD to simulate the atomistic movement of proteins is computation time. Modeling the smallest proteins could take upwards of a month to simulate 100 nanoseconds of protein movement. This limitation can be overcome using parallel computing and specially designed algorithms. One such computer is Anton, which can simulate proteins on the order of millions of atoms for timescales in the millisecond range [141, 142].

The ability to view an MD protein trajectory provides invaluable insight into how folding and unfolding occur. To view a trajectory, one must first load a minimized crystal structure into VMD. The minimized structure has been neutralized and solvated into a box of water. After loading the minimized structure, the trajectory file is loaded. The trajectory file contains a merged list of PDB codes that were generated at selected intervals during the simulation.

Fluorescence

It is essential to have experimental studies to complement computational analyses. Many techniques that are utilized to study protein folding rely on intrinsic fluorescence of aromatic residues, namely, phenylalanine, tyrosine, and tryptophan (Appendix A) [143-145]. These residues absorb wavelengths of light at 260nm, 280nm, and 285nm respectively [146]. When these residues are buried in an environment excluding solvent, representative of the native state, fluorescence intensity increases [146]. As the protein unfolds and the residues are exposed to solvent, the fluorescence is quenched. This unfolding can be due to heat, chemical denaturant, pH, or pressure [143]. While these three amino acids exhibit fluorescence, tryptophan is most often used as a probe for folding and unfolding experiments due to its high fluorescence intensity and large molar extinction coefficient [146].

Circular Dichroism

Circular dichroism (CD) is a method used for ascertaining the secondary and tertiary structural information of proteins from their native environments. This technique measures the differential absorption of left- or right-handed circularly polarized UV light by chiral molecules (Figure 13(a)) [147, 148].

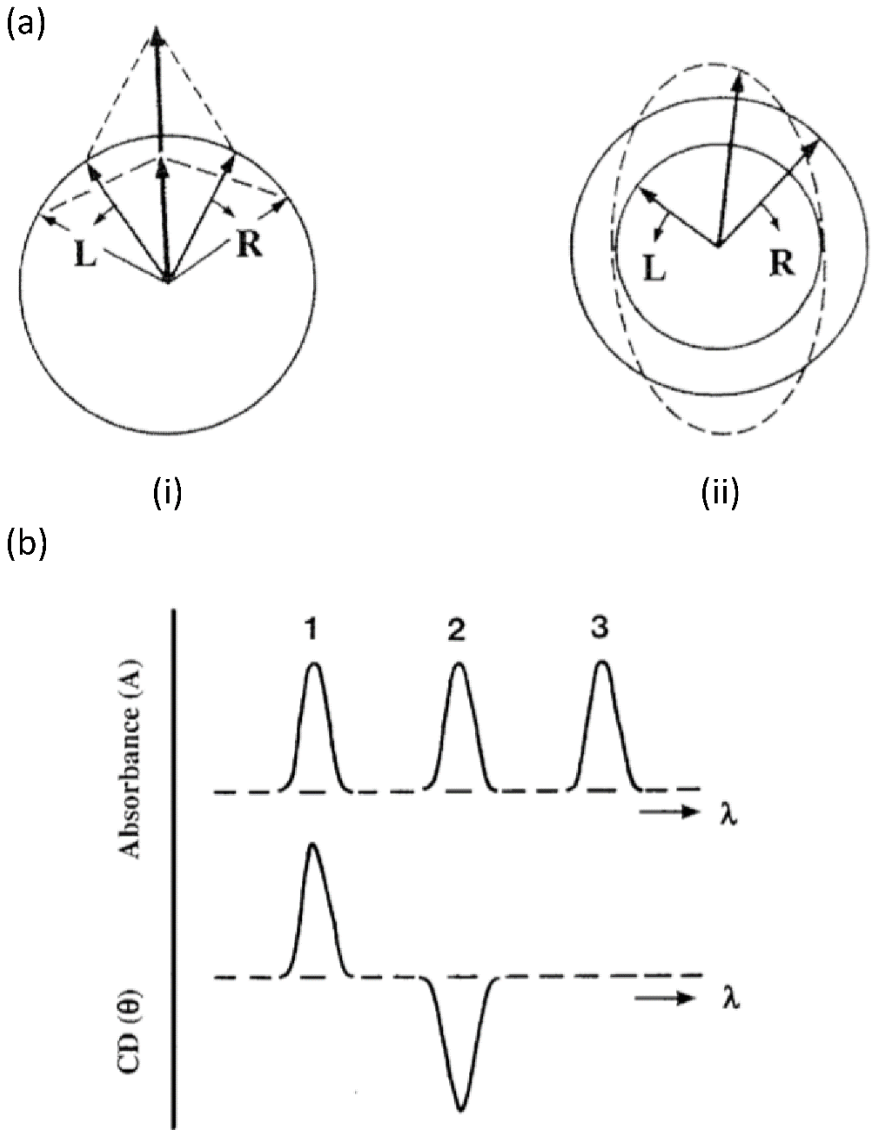


Figure 13. CD effect origin. (a) The left (L) and right (R) circularly polarized components of plane polarized radiation: (i) the two components have the same amplitude and when combined generate plane polarized radiation; (ii) the components are of different magnitude and the resultant (dashed line) is elliptically polarized. (b) Absorption versus CD spectra. Band 1 has a positive CD spectrum with L absorbed more than R; band 2 has a negative CD spectrum with R absorbed more than L; band 3 is due to an achiral chromophore. Figure and figure legend reproduced from [147] and used with permission.

This differential absorption results in a CD spectrum which can contain both positive and negative peaks (Figure 13(b)). There are two types of CD, far and near. In far-UV CD, the protein backbone, and thus the secondary structure, can be monitored due to its preferential absorption of UV light 240nm and below [147-149]. In near-UV CD (260-320nm), tertiary structure can be monitored due to the number, mobility, and environment of aromatic amino acids. Phenylalanine, tyrosine, and tryptophan are observed in the 255-270nm, 275-282nm, and 290-305nm regions respectively [147, 148].

Continuous and Stopped-Flow

The folding kinetics of a protein is commonly studied using stopped-flow spectroscopy. It is rapid-mixing technique which can monitor the folding or unfolding of a protein as a measure of intrinsic fluorescence. When conducting a folding experiment, protein that has been denatured with concentrated denaturant is diluted with a refolding buffer (Figure 14).

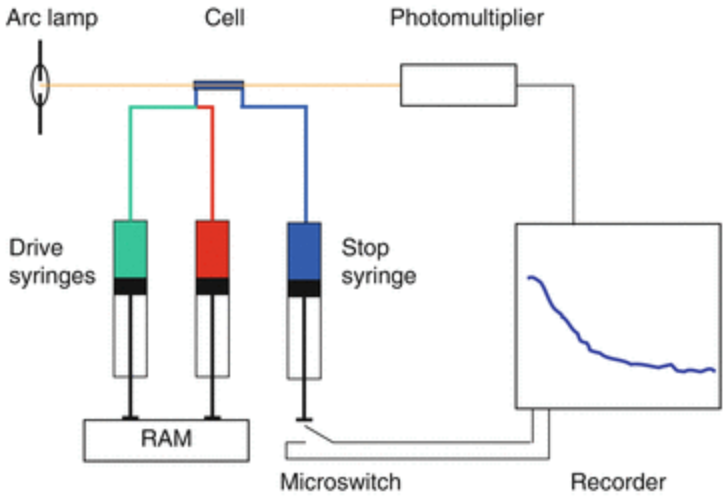


Figure 14. Schematic of a stopped-flow spectrophotometer. Protein (red) and buffer (green) are mixed and passed through the cell. Mixed samples (blue) are collected in a stop syringe. Figure and figure legend reproduced from [150] and used with permission.

As the concentration of denaturant is decreased, the protein folds and intrinsic fluorescence increases. Conversely, an unfolding experiment mixes protein in its native state with concentrated denaturant. As the concentration of denaturant is increased, the protein unfolds and intrinsic fluorescence decreases. Other methods exist for denaturing proteins, such as altering the pH or temperature, and detection of folding or unfolding, such as near- and far-UV CD, Fourier transform infrared spectroscopy, X-ray scattering, and real time NMR [146, 147, 151-156]. A major limitation of the stopped-flow technique is dead-time. Dead-time is defined as the amount of time it takes a protein sample to move from the mixer to the observation window. This time can be on the order of milliseconds. In the event of a rapidly folding protein, many of the major folding events could occur in this dead-time and thus would not be detected [87, 157-

159]. Researchers are developing better systems with dead-times in the microsecond range [160-162]. One such system is a continuous flow spectrophotometer. In continuous flow the protein and buffer solutions are mixed as they enter the observation cell, and the reaction occurs as the sample flows through the cell. Fluorescence is measured for the entire length of the cell (Figure 15).

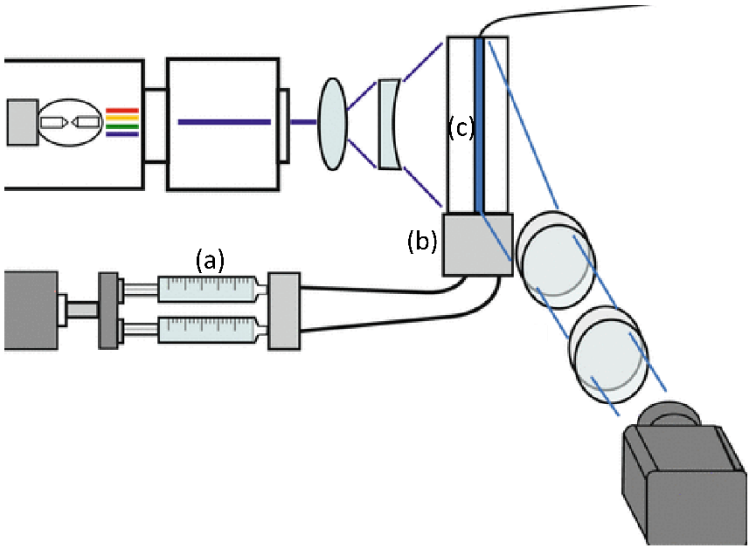


Figure 15. Schematic of continuous flow spectrophotometer. (a) protein and buffer are passed through (b) a mixer just prior to entering (c) the observation cell. Figure adapted from [163] and used with permission.

CHAPTER II

ELUCIDATING THE KEY DETERMINANTS OF STRUCTURE, FOLDING, AND STABILITY OF GB1 USING BIOINFORMATICS APPROACHES

OVERVIEW

Folding along a funnel-shaped energy landscape from an ensemble of denatured states occurs through the restriction of conformational space. One of the key determinants hypothesized to restrict shape space is the formation of a native-like topology dictated by long-range contacts between evolutionarily conserved residues. In this view, select amino acids are conserved in a superfamily of proteins, in part because they make critical interactions that are more important in forming and maintaining the common fold than biological function. These critical interactions are proposed to work by structuring a hydrophobic "fold-determining core" to stabilize the initial native-like topology [97, 164]. The role of conserved amino acids has been the subject of a number of computational and experimental studies which seek to investigate a link between conserved amino acids and how they might or might not facilitate rapid and correct folding of a protein into its native state [97, 100-105, 109, 165-168].

Long-range interactions are the focus of this research because they are the key determinants of tertiary structure and can be classified as interactions between amino acids that are greater than seven residues from each other in the primary structure but within 5Å in the tertiary structure [109, 169, 170]. Using bioinformatics approaches we can identify and assess which amino acids are conserved for the fold of a protein.

Content in this chapter is reprinted with permissions from "Collins J, Bedford JT, Greene LH. Elucidating the Key Determinants of Structure, Folding, and Stability for the (4β + α) Conformation of the B1 Domain of Protein G Using Bioinformatics Approaches. IEEE Transactions On Nanobioscience. 2016; 15:140-147."

The application of network science has also become important in the study of protein structure and folding [97, 107, 114, 171-173]. Network measures have most recently been applied to protein transition-states and native folds, which further our understanding of the underlying determinants of protein structure [97, 114].

A protein superfamily is similar to a family lineage tree (Figure 16) [174, 175]. As a protein diverges from a common ancestral sequence, there will be some degree of evolutionary drift and some features will be retained. In terms of sequence similarity, it has been determined that proteins that contain >40% identity are generally conserved in function [176]. Whereas sequences conserved for the fold can contain <25% identity and have a significant degree of functional diversity. Thus, the construction of a divergent superfamily provides a method of searching and identifying a conserved sequence and structural signature that are hypothesized to be critical in determining the fold.

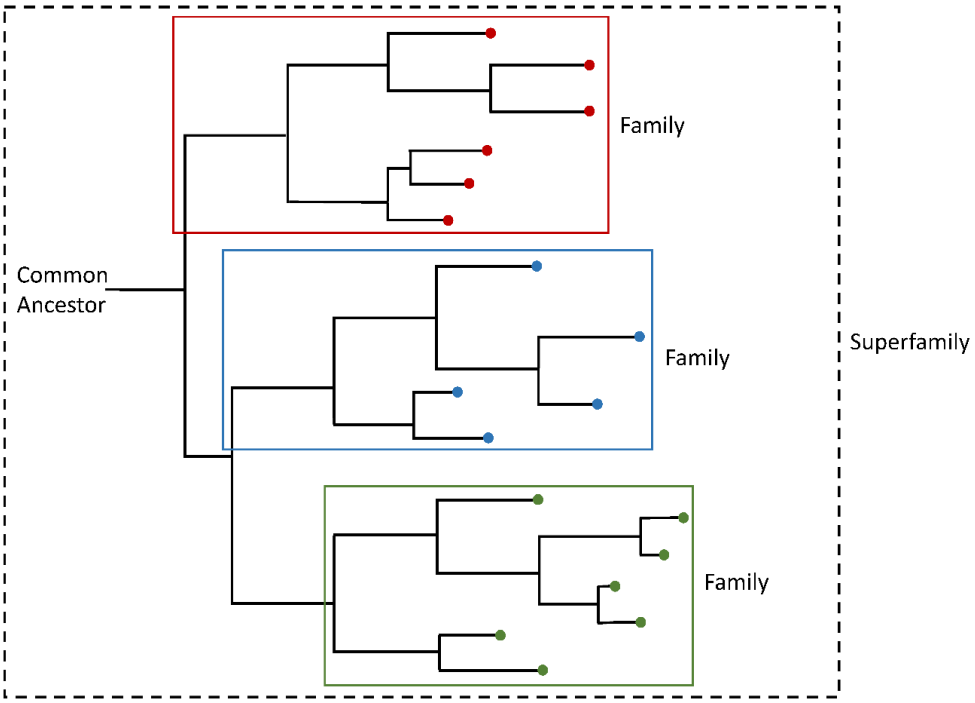


Figure 16. Hypothetical representative schematic of a superfamily.

Our model system is GB1. In this chapter, I present the use of bioinformatics methods in concert with a network analysis to elucidate the conserved amino acids and their relative importance in the fold of GB1. This establishes a foundation for experimental and computational studies to test the proposed role of conserved residues in structure, folding, and stability.

MATERIALS AND METHODS

Structural Alignment and Percent Identity

A structural alignment was constructed for GB1 using the DaliLite Server (v.3) [119]. This was used to identify and select proteins with a similar superimposable structure that also shared low sequence identity and varied in function. The DaliLite server is a comprehensive

search method that surveys the protein data bank and does a sum-of-pairs comparison of superimposable structures. This method produces a measure of similarity by comparing intramolecular distances and calculating a similarity measure called the Dali Z-score. Structures that are significantly similar have a Z-score above 2, and usually have similar folds [119]. The final alignment was constructed by removing all gap regions in the Dali generated alignment to give a contiguous GB1 sequence.

The sequence identity of the structural alignment was obtained by importing the alignment into the sequence identity and similarity (SIAS) server which uses the following equation to calculate the percent identity (PID) of each alignment:

$$PID = 100\left(\frac{Identical\ Positions}{Length\ of\ the\ Alignment}\right) \quad (12)$$

Our aim in using this program was to evaluate the multiple sequence alignment of selected structures to ensure that they were significantly divergent. This ideally consists of pairwise identities ≤ 25%, although a few pairs were between 25–35% identity. We also sought to have a broad range of functional diversity so that similarities obtained would be related to structure. Structural modifications were made by hand based on the visual comparison of the side-chain orientation in each selected structure using RasMol (Ver. 2.7.2.1.1) and Insight II (Ver. 2005, Accelrys). This manual analysis is required to ensure that the obtained structural alignment is further refined as DaliLite only considers the α-carbon backbone in the 3D superposition.

Conservation Analysis and Hydropathy

The completed and verified structural alignment was analyzed for position specific residue type and residue character conservation. The number of each residue type at each

position in the structure-based sequence alignment is calculated by summation of each type of amino acid using a computer program written to calculate these values. With the number of amino acids of each type at each position, entropy can be calculated by equation 13 in SigmaPlot (Ver. 13.0, Systat Software):

$$S(i) = \sum_{j=1}^{m} - \{Pj(i) \ln[Pj(i)]\} \quad (13)$$

In the equation *Pj(i)* is the fractional occurrence of each amino acid type *j* at each residue position *i* and *m* is the number of amino acid types or groups possible in the particular analysis [177]. Positional entropy tells us about the amino acid variability at each position. High entropy indicates high variability and thus infers low conservation and vice versa. Thus, to calculate conservation the results from the application of (13) is used as follows:

$$C(i) = 1 - \frac{S(i)}{\ln(m)} \quad (14)$$

The conservation parameter C*(i)*, ranges from 0 to 1. At maximum entropy, where all amino acids types are equally represented, a conservation of 0 is obtained. Whereas, at minimal entropy, when only one amino acid type is represented, a conservation of 1 is obtained. From the analysis, residue positions whose conservation is ≥ 0.45 are considered highly conserved whereas conservation between 0.45 and 0.30 are considered moderately conserved. Any conservation values <0.30 are considered to be less conserved. Positions containing one or more gaps with respect to GB1 are given a value of zero and considered non-conserved. The analysis of conservation of character involved dividing the amino acids into four groups (polar, nonpolar, acidic, and basic) for (13) and (14). To calculate residue specific hydropathy as it relates to persistence within the structural superfamily, the average hydrophobicity of all the amino acids at the selected position are assigned a hydropathy value. We then applied equation 15 at each position in the alignment.

$$Average\ Hydrophobicity = \frac{(sum\ of\ the\ number\ of\ each\ amino\ acid\ type * hydrophobicity\ of\ that\ amino\ acid)}{14} \quad (15)$$

The hydrophobicity values used were adapted from a commonly used amino acid hydrophobicity index [178]. The data from both conservation and hydropathy analyses were analyzed and plotted using SigmaPlot.

Network Analysis

Using the PDB structure of GB1 (1 PGB) we calculated all of the long-range amino acid interactions. This was accomplished using the program Contact which calculated every interacting atom between pairs of residues within 5Å in the tertiary structure [179]. The output file was further analyzed using a program we coded in C, called DegLR which identified pairs of contacting residues that were seven or more residues apart in the primary structure. This data was converted to a Pajek input file and a network of all the long-range interactions was constructed using Pajek-XXL (64 bit) [180]. Betweenness centrality was calculated within the resulting GB1 network using Pajek and plotted with SigmaPlot. The betweenness centrality measure is based on (16).

$$\sigma(m) \equiv \sum_{i \neq j} \frac{B(i,m,j)}{B(i,j)} \quad (16)$$

B(i, j) is the total number of shortest paths between vertices *i* and *j*. *B(i,m, j)* is the total number of shortest paths between vertices *i* and *j* that pass through vertex *m*. The ratio *B(i,m, j)/B(i, j)* produces a measure of importance (0 = low importance and 1 = high importance) of vertex m in traversing the network from vertices *i* to *j.* The betweenness measure, *σ(m)* of the vertex *m*, is the sum over all pairs of *i* and *j* vertices which have at least one path [181]. Thus, *B(i, j)* > 0. Betweenness centrality facilitates identifying the central importance of each node.

The methods described above regarding the structural alignment and percent identity, conservation analysis and hydropathy, and network analysis were done in collaboration with Dr. Jason Collins.

RESULTS AND DISCUSSION

To determine the evolutionary conservation of amino acids in the sequence of GB1 we developed a structural superfamily using the DaliLite server [119]. We input the PDB file of GB1 into the server and obtained a list of proteins whose structures are superimposable with GB1. Ideally, we would have selected 20 or more proteins to allow for the greatest amino acid variability. However, we were limited in the numbers of available structures that fit our criteria for inclusion. From the server we selected 13 proteins whose fold matched GB1 but varied significantly in sequence identity (Figure 17) and were functionally diverse (Table 1). This ensured that the structure and structure-based sequence alignments would provide information on which amino acids and side-chain interactions were important in dictating the fold and not biological function. In addition, we assessed the sequence identity to ensure to a significant degree that each value was below 25% for the majority of selected proteins. This percent identity is considered in the “twilight zone” [182]. The “twilight zone” is a threshold of percent identity in which you cannot be sure or guarantee the proteins have the same 3D structure. Thus, we work in this region to enhance sequence variability but use known 3D structures for accuracy of the analysis.

1PGB	100%													
2PTL	14.54%	100%												
1RLF	12.96%	12.96%	100%											
3PO0	7.69%	13.46%	7.69%	100%										
1ENF	11.11%	9.25%	9.25%	3.84%	100%									
1FMA	7.84%	11.76%	3.92%	23.52%	5.88%	100%								
2K8H	12.00%	8.00%	6.00%	10.00%	14.00%	8.00%	100%							
1F2R	6.12%	6.12%	6.12%	12.24%	8.16%	10.20%	12.24%	100%						
1EUV	10.20%	8.16%	8.16%	8.16%	6.12%	12.24%	30.61%	10.20%	100%					
1WM2	10.20%	12.24%	8.16%	10.20%	14.28%	10.20%	28.57%	10.20%	32.65%	100%				
3A4R	4.00%	8.00%	10.00%	6.00%	14.00%	12.00%	12.00%	2.04%	22.44%	34.69%	100%			
1C4P	8.92%	5.45%	11.11%	3.84%	11.11%	3.92%	12.00%	12.24%	6.12%	6.12%	8.00%	100%		
2BS2	5.55%	1.85%	7.40%	7.69%	9.25%	7.84%	8.00%	6.12%	6.12%	8.16%	10.00%	3.70%	100%	
1WSP	8.33%	2.08%	6.25%	10.41%	8.33%	6.25%	14.58%	6.25%	10.41%	8.33%	4.16%	14.58%	4.16%	100%
	1PGB	**2PTL**	**1RLF**	**3PO0**	**1ENF**	**1FMA**	**2K8H**	**1F2R**	**1EUV**	**1WM2**	**3A4R**	**1C4P**	**2BS2**	**1WSP**

Figure 17. Percent sequence identity for proteins in the structure-based multiple sequence alignment in Figure 18. Identities were calculated using equation (12) within the SIAS server. http://imed.med.ucm.es/Tools/sias.html

Table 1 General functional classification of structurally aligned proteins.

PDB Code	**Species**	**Protein Length**	**Functional Classification** *(Based on RCSB PDB)*
1pgb	Bacteria (*Streptococcus sp.* Group G)	56	Immunoglobulin-Binding Protein
2ptl	Bacteria (*Peptostreptococcus magnus*)	78	Protein-Binding (Immunoglobulin L Chain)
1rlf	Mouse (*Mus musculus*)	90	Signal Transduction Protein
3po0	Halophile (*Haloferax volcanii*)	89	Protein-Binding
1enf	Bacteria (*Staphylococcus aureus*)	212	Toxin
1fma	Bacteria (*Escherichia coli*)	81	Transferase
2k8h	Human African Trypanosomiasis (*Trypanosoma brucei*)	110	Signaling Protein
1f2r	Mouse (*Mus musculus*)	87	DNA-Binding Protein
1euv	Baker's Yeast (*Saccharomyces cerevisiae*)	221	Hydrolase
1wm2	Human (*Homo sapiens*)	78	Protein Transport
3a4r	Mouse (*Mus musculus*)	79	Transcription
1c4p	β-hemolytic Bacteria (*Streptococcus equisimilis*)	137	Blood Clotting
2bs2	Proteobacteria (*Wolinella succinogenes*)	660	Oxidoreductase
1wsp	Rat (*Rattus norvegicus*)	84	Signaling Protein

Although the structural alignment provided by the DaliLite server is quite advanced, it may not be perfect and only provides an optimal sequence alignment based on α-carbon superposition. To verify the sequence alignment, the general side-chain orientation of each

amino acid aligned with GB1 was assessed by manually inspecting the superposition using the 3D structure visualization programs, RasMol and Insight II [183]. Once each side-chain orientation was verified the finalized structure-based sequence alignment was completed (Figure 18).

```
        1         10        20        30        40        50
        |         |         |         |         |         |
[1pgb] MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
[2ptl] VTIKANLIfagstQTAEFKgTFe-KATSEAYAYADTLkgeYTVDVAdkgYTLNIKF
[1rlf] RIIRVQMElgdgsVYKSILVT--dkAPSVISRVlkknnseFELVQldashDFLLRQ
[3po0] GSMEWKLF-ADlarTVRVDVDtvGDALDALvgahlesrv-iNVLRn--gdELALFP
[1enf] RVIGANVWVdiqkETELIRTvtlqELDIKIRKILSDky--GLIEFDMkisHIDVNL
[1fma] -MIKVLFFrelvtdATEVAad--fptVEALRQHMAAalallLAAVn--gdEVAFFP
[2k8h] VAVKVVNA---dgaEMFFRIKs-rtALKKLIDTYCkkqnsVRFLFd--ddVIDAMV
[1f2r] KCVKLRALh--sacKFGVAArsCQELLRKGCVRFq-----sRLCLfpglaELLLLT
[1euv] INLKVSDg----ssEIFFKIKkttpL-RRLMEAFKRqgkeLRFLYd--ndIIEAHR
[1wm2] INLKVAGQ---dgsVVQFKItplSKLMKAYCERqgl--rqIRFRFd--edTIDVFQ
[3a4r] LRLRVQGk--ekhqMLEISLSplkVLMSHYEeamgl--hkLSFFfd--gdLIEVWG
[1c4p] VEYTVQFTpfrpglKDTKLLitsqELLAQAqsilnkpgytYeRSsivtliSEKYYV
[2bs2] RMLTIRVFkYphfqEYKIEeap-smtIFIVLNmirepdlnmMIn-lfedGVITLLP
[1wsp] IVVAYYFcg--epiPYRTLVravGQFKE-LL---tkkg-sYRYYFkk-kiIGKVEK
```

Figure 18. Structure-based sequence alignment. The sequence alignment was generated from a structural alignment of 14 superimposed proteins, some of which are domains within larger proteins. Side chains that are not in a similar orientation are shown as lowercase letters in accordance with DaliLite. Gaps are delineated by dashes. All positions were verified upon visual inspection of the aligned structures and adjusted accordingly. In brackets on the left are the PDB codes for the structures selected for inclusion in the alignment. The numbering system corresponds to the GB1 structure (1PGB).

As expected, most variability in side-chain orientation is found in the loop regions which have higher variation in structure and length. Also of note, is that the third β-strand of GB1 appears to only have one position in which there was total side-chain orientation alignment. This could indicate that formation of β-strand 3 is not as evolutionarily conserved for this fold as the other β-strands and could also suggest that stabilization of this strand is formed post initial collapse of the structure during folding. In order to know which amino acids would be most significant to select for experimental and computational study, an analysis is performed to determine position specific conservation over the superfamily (Figure 19). Using a modified Shannon's entropy equation, amino acid conservation is determined based on the number of amino acid types at each position. This analysis calculates the entropy of a given residue based on a position specific variability over the superfamily, where high entropy indicates high amino acid variability while low entropy indicates low variability. From the conservation analysis we found that there were twelve residue positions that were considered evolutionarily conserved. In GB1 these correspond to: Tyr3, Lys4, Leu5, Thr18, Ala20, Ala26, Phe30, Glu42, Asp46, Lys50, Phe52, and Val54. There are eleven positions that were considered moderately conserved (>0.30) and one position, residue Ala26, considered highly conserved (≥ 0.45). It is interesting to note that there is at least one conserved amino acid found in each major secondary structure component.

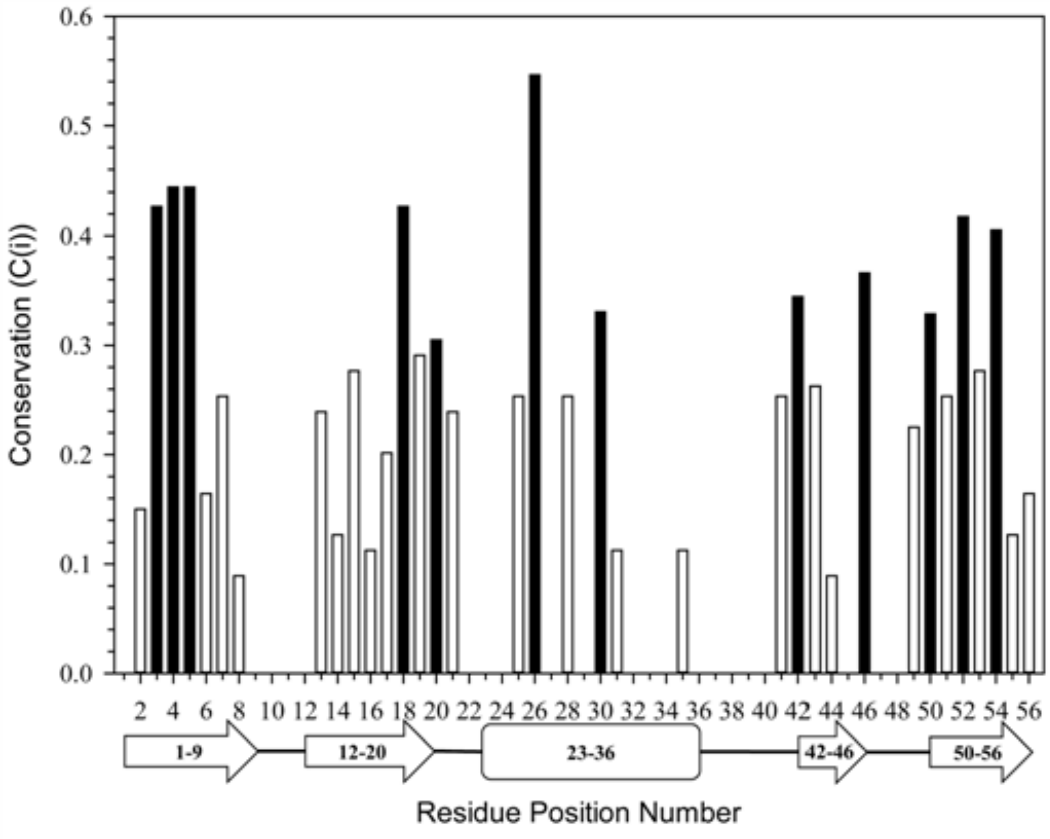


Figure 19. Amino acid conservation analysis. Positions colored in black are positions considered conserved. Positions ≥ 0.45 are considered highly conserved and 0.45 > positions ≥ 0.30 are considered moderately conserved. Arrows indicate β-strands, the rounded rectangle indicates an α-helix. Data plotted using SigmaPlot 12.5.

The initial analysis determined conservation based strictly on identity, thus we also wanted to get a sense of conserved residue positions with respect to amino acid character. A second conservation analysis was conducted by counting similar character types rather than the same specific amino acid and the results were plotted similarly (Figure 20). The data indicated that eleven positions were conserved in amino acid character. In GB1 these are: Tyr3, Leu5,

Leu7, Thr18, Ala20, Ala26, Phe30, Gly41, Trp43, Phe52, and Val54. Between the two conservation analyses (Figures 19 and 20) there are four positions (Lys4, Glu42, Asp46, and Lys50) that were considered conserved based on identity but are not similarly conserved in the character analysis. This indicates from an evolutionary perspective that these positions may be more dependent on the particular amino acid chemical structure. However, when these positions are modified it does not favor a particular amino acid character type which could mean that its role in forming the overall shared conformation across the superfamily could be secondary. In GB1, experimental data indicates that Asp46 is structured in the transition-state and necessary for early formation of the second β-turn [81]. Interestingly, the following eight positions are conserved in both amino acid position and character: Tyr3, Leu5, Thr18, Ala20, Ala26, Phe30, Phe52, and Val54. In addition, three positions (Leu7, Gly41, and Trp43) that were not conserved in amino acid identity are now considered conserved with respect to amino acid character which indicates that these positions may be important in the fold of GB1 because of the character of the amino acid. This suggests that during evolution when these positions were varied, they did not favor any specific amino acid in particular but required that the character of the position be maintained.

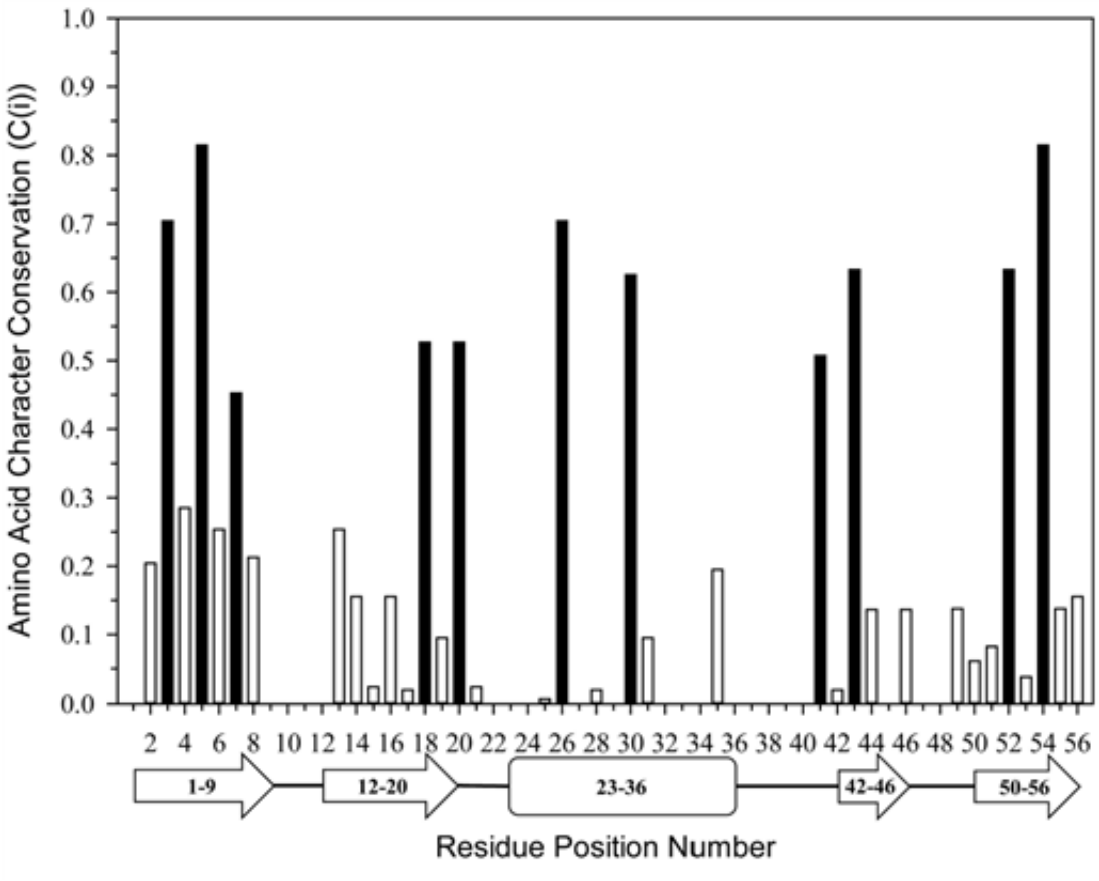


Figure 20. Amino acid character conservation analysis. Positions colored in black are positions considered conserved. Positions ≥ 0.45 are considered highly conserved and 0.45 > positions ≥ 0.30 are considered moderately conserved. Arrows indicate β-strands, the rounded rectangle indicates an α-helix. Data plotted using SigmaPlot 12.5.

To identify hydrophobic positions versus hydrophilic positions a hydropathy analysis was done (Figure 21). From the hydropathy analysis we see that of the fifteen positions conserved by amino acid type or character, eleven were hydrophobic while four were hydrophilic in nature. This makes sense as the four positions considered hydrophilic are either acidic or basic in character and would be expected to be found on the surface of the protein exposed to the solvent.

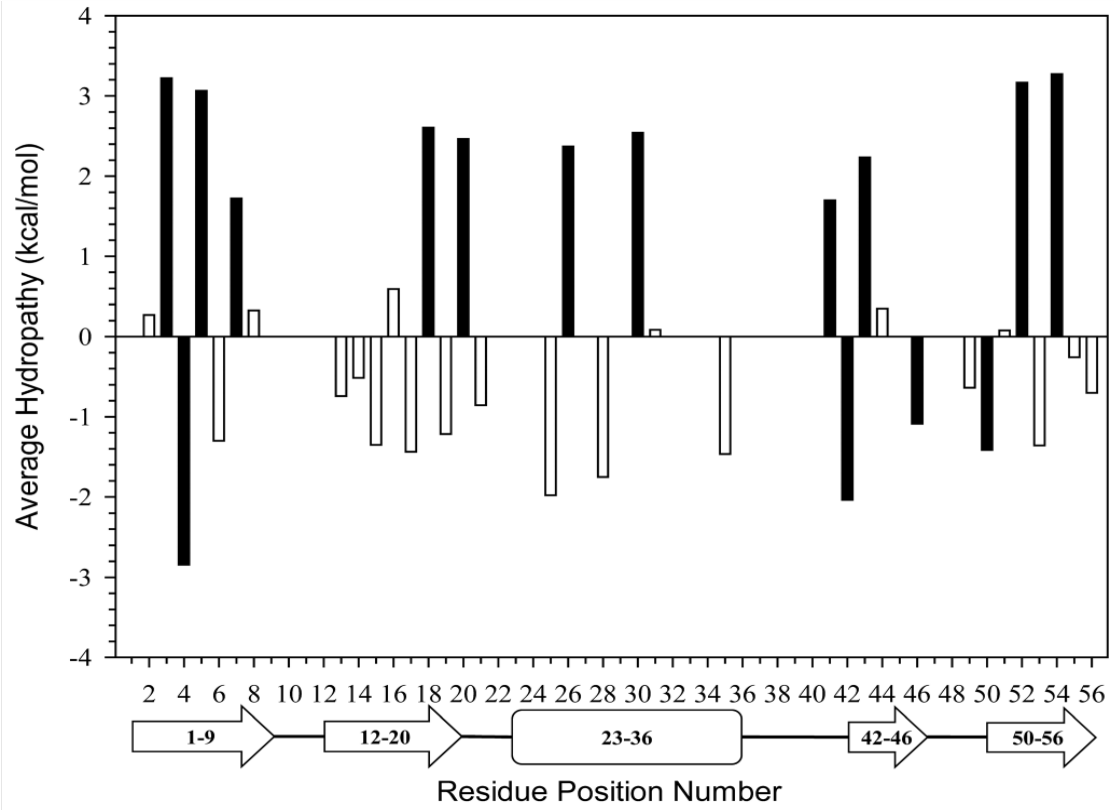


Figure 21. Position specific hydropathy analysis. Positions colored in black are positions considered conserved either by position or character conservation. Positive values indicate hydrophobicity and negative values indicate hydrophilicity. Arrows indicate *β*-strands, the rounded rectangle indicates an *α*-helix. Data plotted using SigmaPlot 12.5.

Table 2 Summary of Structure Alignment Analysis

Amino Acid	Secondary Structure	Amino Acid Conservation	Character Conservation	Hydropathy
Tyr 3	β-Strand 1	Moderately	Highly	Hydrophobic
Lys 4	β-Strand 1	Moderately	Less	Hydrophilic
Leu 5	β-Strand 1	Moderately	Highly	Hydrophobic
Leu 7	β-Strand 1	Less	Highly	Hydrophobic
Thr 18	β-Strand 2	Moderately	Highly	Hydrophobic
Ala 20	β-Strand 2	Moderately	Highly	Hydrophobic
Ala 26	α-Helix	Highly	Highly	Hydrophobic
Phe 30	α-Helix	Moderately	Highly	Hydrophobic
Gly 41	Loop 3	Less	Highly	Hydrophobic
Glu 42	β-Strand 3	Moderately	Less	Hydrophilic
Trp 43	β-Strand 3	Less	Highly	Hydrophobic
Asp 46	β-Strand 3	Moderately	Less	Hydrophilic
Lys 50	β-Strand 4	Moderately	Less	Hydrophilic
Phe 52	β-Strand 4	Moderately	Highly	Hydrophobic
Val 54	β-Strand 4	Moderately	Highly	Hydrophobic

Based on all the bioinformatics data gathered there are fifteen positions that were revealed to have conservation by at least one measure. However, when compared to the structure-based sequence alignment only nine of the fifteen positions can be considered reliably conserved when we take into account side-chain orientation. A comprehensive summary of all the conserved positions and the results of each analysis can be found in Table 2 and Figure 18. Experimental results show in part some correspondence with our conservation data [81].

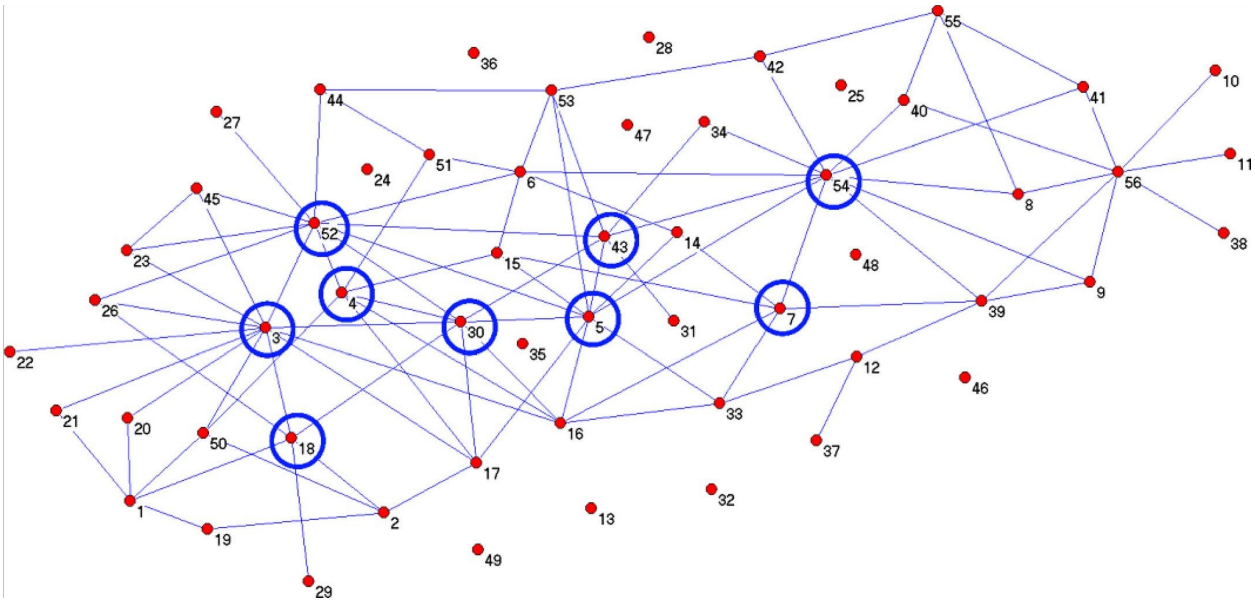


Figure 22. Network of long-range interactions in the structure of GB1. Individual amino acids are the filled circles connected by long-range interactions shown as lines. The nine conserved amino acids based on Table 2 and Figure 18 (considering only positions with similar side-chain orientation) are shown by open circles. Data plotted using Pajek64-XXL 4.08.

A network analysis provides insight into the nature of each amino acid position within the structure of GB1. We can initially assess the relative importance at each position in the structure by the number of contacts made with other amino acids. Using this approach, we calculated all the long-range interactions found in the structure of GB1 and modeled an interaction network (Figure 22). From the network we see that the nine conserved residues are highly interconnected. Of the nine conserved positions, eight form what appears to be a predominantly hydrophobic core of GB1 (Figure 23(a)). A network model of the long-range interactions overlaying the 3D structure of GB1 shows a group of amino acids linked in the core. It is interesting that six of the nine conserved residues are found in the N- and C-termini β-strands.

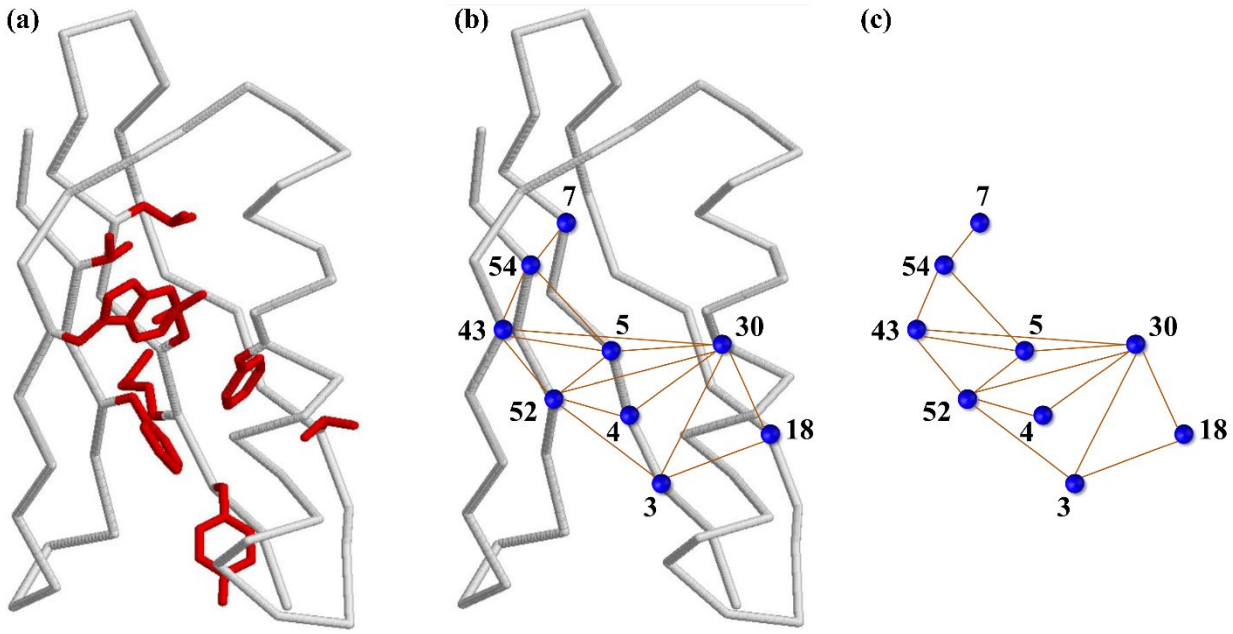


Figure 23. Conserved amino acid network overlay. Backbone structure of GB1 in light gray with (a) conserved amino acid side chains shown and (b) network overlay with amino acid nodes located on the *Cα* in filled circles and long-range interaction links as lines. (c) Core network of conserved amino acids. Structures visualized using RasMol Ver. 2.7.2.1.1.

It appears as if the formation of connections between the N- and C-termini may be important to forming this fold with five of the fifteen core long-range interactions found between *β*-strands 1 and 4 (Figure 23 (b)). There are also three conserved interactions potentially important in forming the *β*-hairpins and six in stacking the *α*-helix onto the *β*-sheet. This could be necessary in bringing the two anti-parallel *β*-hairpins together in 3D space during the folding process.

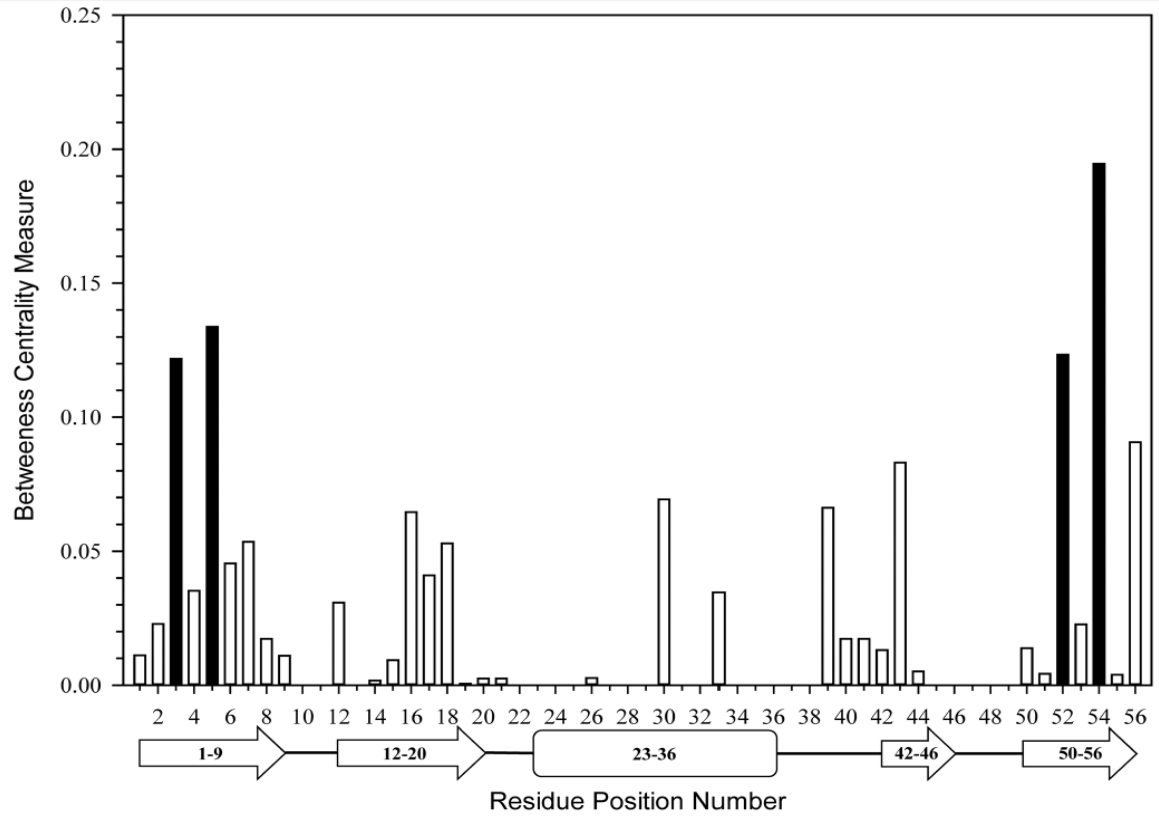


Figure 24. Betweenness centrality analysis of the GB1 network. Positions with high betweenness centrality are colored in black. These are: 3, 5, 52, and 54. Arrows indicate β-strands, the rounded rectangle indicates an α-helix. Data plotted using SigmaPlot 12.5.

To further investigate the importance of the core conserved amino acids, an analysis using the betweenness centrality (BC) measure on the GB1 long-range interaction network was conducted (Figure 24). A betweenness measure indicates centrality of an amino acid node in the network. It calculates the importance of a node in traversing the network [184, 185]. The BC analysis revealed four nodes (Tyr3, Leu5, Phe52, and Val54) with high betweenness. Interestingly, the four amino acids are found on the N- and C-termini β-strands and could be

important in bringing the two hairpins together. Further, these four amino acids interact in the GB1 network (Figure 23(b-c)). This result indicates that these positions appear to be more centrally important to the network and may be of higher importance, perhaps fixing the topology in the folding process. However, a focused investigation is necessary to determine if the hypothesized role of the conserved features in the formation of the β-grasp fold is supported by existing and future computational and experimental results.

SUMMARY

GB1 has served as the model system in a number of significant studies that encompass both computational and experimental approaches. Orban and co-workers engineered GB1 and another protein, the three-helical bundle called protein A to maintain their distinct folds and functions yet share up to 98% sequence identity [83]. The folding behavior of these artificially designed proteins was further studied by Giri and co-workers to shed light on this fascinating discovery [186]. GB1 was also a structure used in successful *de novo* folding simulation studies by Shaw and co-workers [187]. Additionally, while a number of computational [85, 86] and experimental studies [82, 88-94, 188] have been conducted to characterize the structure, stability, and folding behavior of GB1, our work is directed at elucidating and characterizing the role of conserved residues from a bioinformatics perspective. Thus, the results of our present study provide an important avenue of investigation for experimental research as well as future theoretical and simulation studies which in combination with existing results published in the literature could help lead to a more comprehensive understanding of the folding process of GB1.

CHAPTER III

THE NATURE OF PERSISTENT INTERACTIONS IN TWO MODEL β-GRASP PROTEINS REVEALS THE ADVANTAGE OF SYMMETRY IN STABILITY

OVERVIEW

Computational approaches have significantly advanced our understanding of the determinants of protein structure, folding, and dynamics. In particular, molecular dynamics (MD) simulations have allowed us to peer into protein forms and interrogate the nature of the interactions in both the native state, transition-state (TS), and during the folding and unfolding process. More recently, significant advancements in the application of MD simulations were made through the success in folding a select group of small proteins with the use of a purpose-built supercomputer named Anton [187]. Other advances have also come from further development of Monte Carlo simulations [189, 190].

In this chapter the results of *in silico* unfolding studies conducted using molecular dynamics is presented. The two model systems are β-grasp proteins, GB1 (Figure 25(a)) and SAMP1 (Figure 25(b)). GB1 contains 95 calculated long-range interactions, where contacts that are seven or more residues apart in the primary sequence but closer than 5Å in the tertiary structure are defined as long-range interactions.

Content in this chapter is reprinted with permissions from "Bedford JT, Diawara N, Poutsma J, Greene LH. The nature of persistent interactions in two model β-grasp proteins reveals the advantage of symmetry in stability. Journal of Computational Chemistry. 2021; 42:600-607."

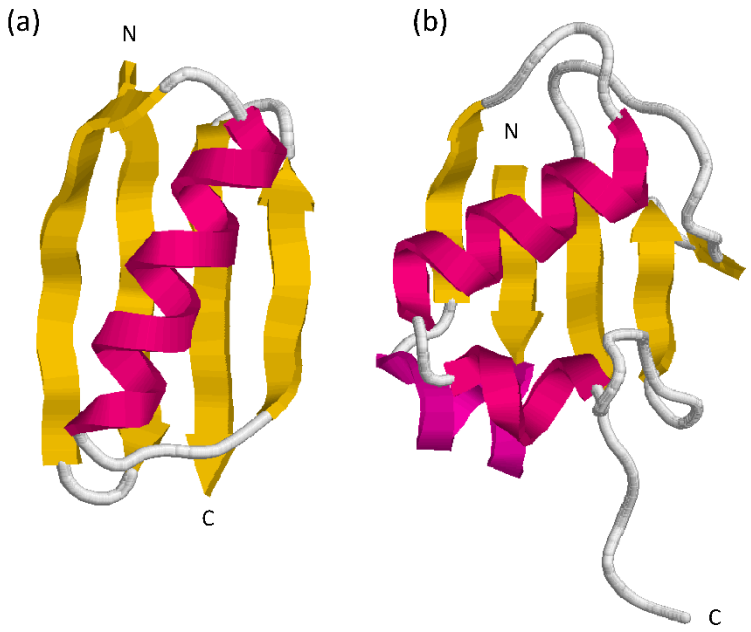


Figure 25. X-ray crystal structures of (a) GB1 and (b) SAMP1. α-helices and β-strands are shown in magenta and yellow, respectively. Structures visualized using RasMol Ver. 2.7.2.1.1.

The understanding of the folding process emerging thus far from kinetic studies suggests that GB1 is a rapid two-state folder [81, 84]. This conclusion remains controversial with the identification of an on-pathway intermediate based on the techniques and conditions of select kinetic studies [61, 76, 77]. In the TS, based on Φ-value analysis, the second hairpin is more structured. In agreement, the application of Ψ-value analysis suggests the four-stranded β-sheet is partially formed through select interactions with the strongest located in the second hairpin [81, 94].

A number of research studies on GB1 focusing on the unfolding of the individual β-hairpins in isolation have been published [191, 192]. The work of Pande and Rokhsar supports our research findings. They found that the first step during unfolding is the total loss of

secondary structure, where intra-backbone hydrogen bonding is lost. This event however does not disrupt the hydrophobic cluster in the C-terminal hairpin comprised of Trp43, Tyr45, Phe52, and Val54. This cluster remains intact until the water penetrates and disrupts the hydrophobic core [191]. The work of Lee and Shin also support our research findings. They concluded that the hydrophobic core consisting of Trp43, Tyr45, Phe52, and Val54 is located in the middle of the two strands comprising the C-terminal hairpin and that the formation of this core is responsible for the initial folding and stability of the C-terminal hairpin [192].

Unfolding studies by MD simulation of the full-length protein have been conducted by previous research groups. Scott and Daggett found during the unfolding of a GB1 variant with three mutations in the C-terminal hairpin (GB1 variant G311), the hydrophobic core is opened and repacked followed by the dissociation of the N-terminal hairpin from the main protein structure [193]. Next the C-terminal hairpin moves away from the helix and the helix rotates to a near parallel position to the C-terminal hairpin. Simultaneously to this event the hydrophobic core comprising Tyr3, Val5, Phe30, Ile34, Trp43, Tyr45, Phe52, and Val54 is rearranged however the long-range interactions between these residues are maintained but are non-native. This series of unfolding events is similar to what we observe in our simulations however in our study we monitor all native long-range interactions. They further conclude that considering the unfolding simulations in reverse, the earliest interactions are between strands three and four which form the C-terminal hairpin and that the β-turn of this hairpin then interacts with the helix. These two events fix the topology of GB1 early in the folding pathway [193]. Morrone *et al.* conducted five thermal unfolding MD simulations on GB1 and described their results from the perspective of folding [76]. They found that in the first transition state, contacts between strands in the C-terminal hairpin were almost fully formed. This is in agreement with our unfolding

simulations which found some of these contacts to be the most persistent during unfolding. Their findings also suggest the formation of an extended nucleus which not only incorporates residues from the C-terminal hairpin and α-helix but also the N-terminal hairpin [76]. Our conclusions are similar in that we find residues in the first strand of the N-terminal hairpin help to stabilize GB1.

In the present study, GB1 is subjected to high temperature all-atom MD simulations in order to identify key long-range interactions governing native-state thermodynamic stability. Long-range interactions were studied because they are the dominant determinants of the 3D protein structure and provide the chemical forces between secondary elements. Of the 95 long-range interactions, 9 are most persistent and located in the C-terminal hairpin. Comparisons to experimental studies at the residue-level are drawn to present a picture of the determinants of structural stability.

For a deeper look into the β-grasp superfamily we compare the MD simulations of GB1 to those conducted with a distant homologue, SAMP1 (Figure 25(b)) which shares 2% identity with GB1. This selection is based on needing a member which was very divergent to best investigate the common determinants in the fold which can be obscured if the proteins are too similar. SAMP1 is a ubiquitin-like protein found in the halophilic organism *Haloferax volcanii*. SAMP1 is 87 residues in length compared to GB1's 56 and contains many charged residues due to its highly saline environment. SAMP1, like GB1, has four β-strands and one central α-helix [194]. It also has two small inserted α-helical segments, α1 and α3, and much longer loops. A total of 20 persistent long-range interactions out of 155 were identified in SAMP1. Comparative studies between GB1 and SAMP1 were conducted and reveal the flexibility in stability for simple symmetrical proteins.

MATERIALS AND METHODS

The X-ray crystal structures for GB1 and SAMP1 were obtained from the Protein Data Bank (PDB codes: 1PGB and 3PO0). Using CHARMM v.39 the structures were minimized and solvated into truncated octahedron water boxes containing 6525 and 6369 explicit water molecules after molecules overlapping with the GB1 and SAMP1 structures respectively, were removed. To neutralize the structure of GB1, four sodium ions were randomly added to its water box. To the box containing SAMP1, 13 chloride ions and 25 potassium ions were randomly added. In this case, the structure is still neutralized however extra ions were added to increase the ionic strength of the simulation. Equilibration was run at 450K and 475K for 280ps for GB1 and SAMP1, respectively. Start temperature was 110K which reached a final temperature of 450K or 475K. Once equilibrated, MD simulations were performed employing CHARMM39 with a CHARMM27 force field and using an isothermal-isobaric ensemble. Four separate simulations were run for GB1 while only three simulations were run for SAMP1. The dynamics simulations were 120ns each with time steps of 2fs. Ewald was utilized to treat long-range electrostatics and van der Waals interactions employed a cutoff of 12Å. The SHAKE algorithm was used to freeze all covalent bonds involving hydrogen. The simulations were visualized using VMD.

The RMSD of each simulation was calculated as a function of time to ensure that the proteins unfolded. The contact distance between residue pairs was measured every 2ps over the course of the trajectory. In each 4ns window, the number of times the amino acids of a long-range interaction were within 10Å of each other were counted. A value of 2000 indicates that the contact was present during the entire window. These counts are referred to as persistence values in the context of this research study. The values for the four simulations were scaled to values ranging from zero to one so that data could be fitted using a logistic regression model and was

subsequently rescaled to the range of the original data. Fitted data was then averaged over the different simulations and plotted as persistence over time.

An analysis to elucidate the common long-range interactions between GB1 and SAMP1 was conducted. For the long-range interactions only two heavy atoms, one on each residue, need to be closer than 5Å in the tertiary structure. They are identified using Contact (CCP4) and DegLR, a program coded in the Greene Lab.

The hydrophobic core of a protein is a region of high density containing non-polar residues. The program Naccess (http://wolf.bms.umist.ac.uk/naccess/) was used to calculate solvent accessibility in both GB1 and SAMP1. The long-range interactions monitored in the more detailed analysis of the MD trajectories are confined to those that have a percent burial of 60% or higher. Any relevant short-range interactions thought to play a role were also analyzed. The orientation of each amino acid side chain was visually verified, and our analysis focused on interactions involving side chains.

RESULTS AND DISCUSSION

Persistence in GB1

The RMSDs of the four GB1 unfolding simulations were plotted as a function of time. As seen in Figure 26(a), these simulations, despite having the same dynamics parameters, have similar yet distinct RMSD values over the course of the simulations. This is due to each atom being assigned a random velocity vector at the beginning of the simulation. It is interesting to note that the RMSDs are fluctuating indicating that GB1 is undergoing many unfolding and folding events but the overall RMSDs are trending upward indicating that the protein is unfolding over the course of the simulations. Also, two of the four GB1 simulations diverge to higher RMSD values earlier in the simulation. The persistence of the 95 long-range interactions

in GB1 was measured over the course of each simulation and averaged. The averaged persistence values were plotted versus time and are shown in Figure 27. The most persistent long-range interactions clearly separate themselves from the total population and are mostly red. Specifically, 9 of 95 or 9.5% of the long-range interactions within GB1 are persistent. Furthermore, these nine interactions are located in the C-terminal hairpin (Figure 28).

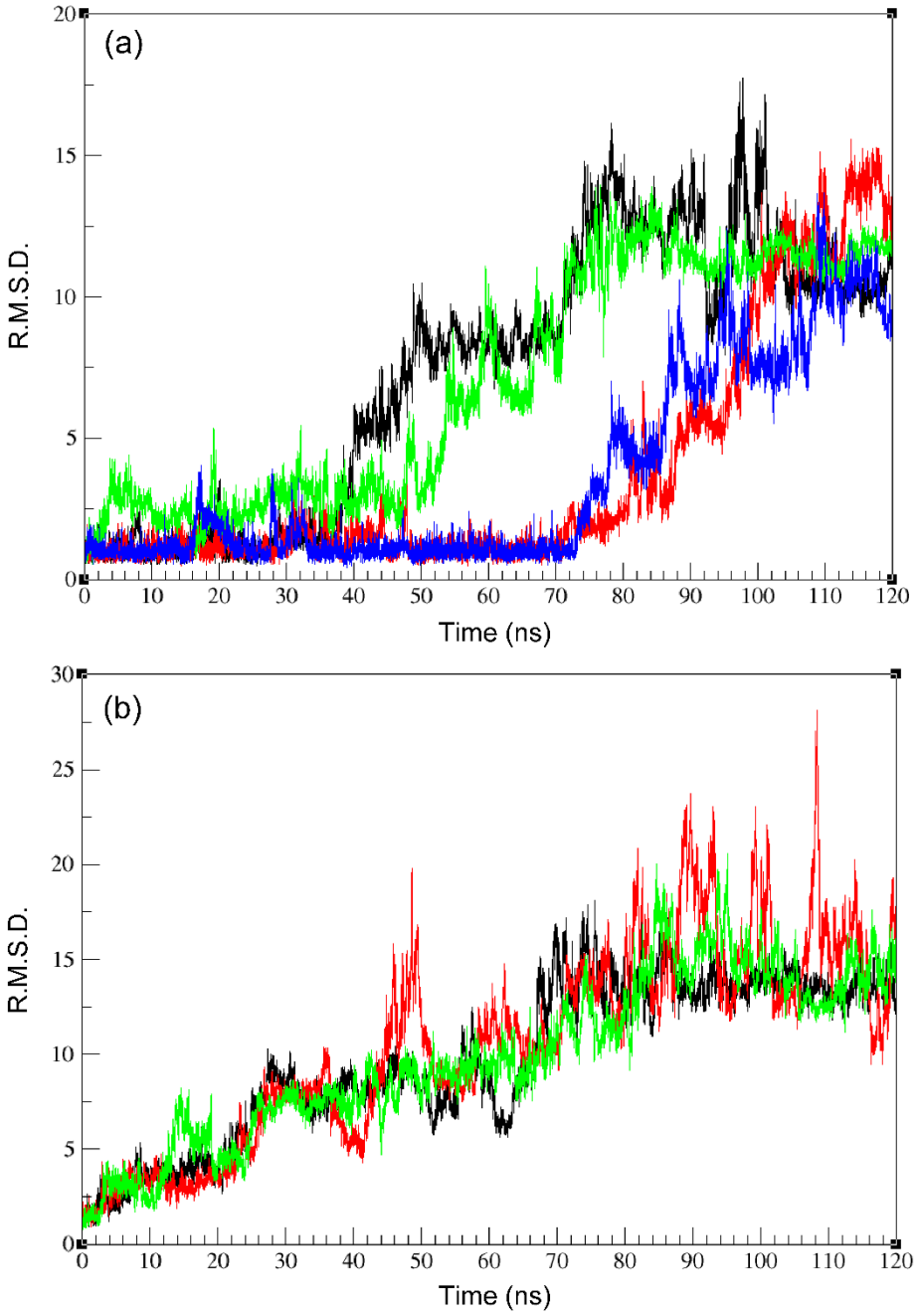


Figure 26. RMSD of MD simulations. (a) GB1 and (b) SAMP1. Simulations 1, 2, 3, and 4 are shown in black, red, green, and blue, respectively. Data plotted using XMGR

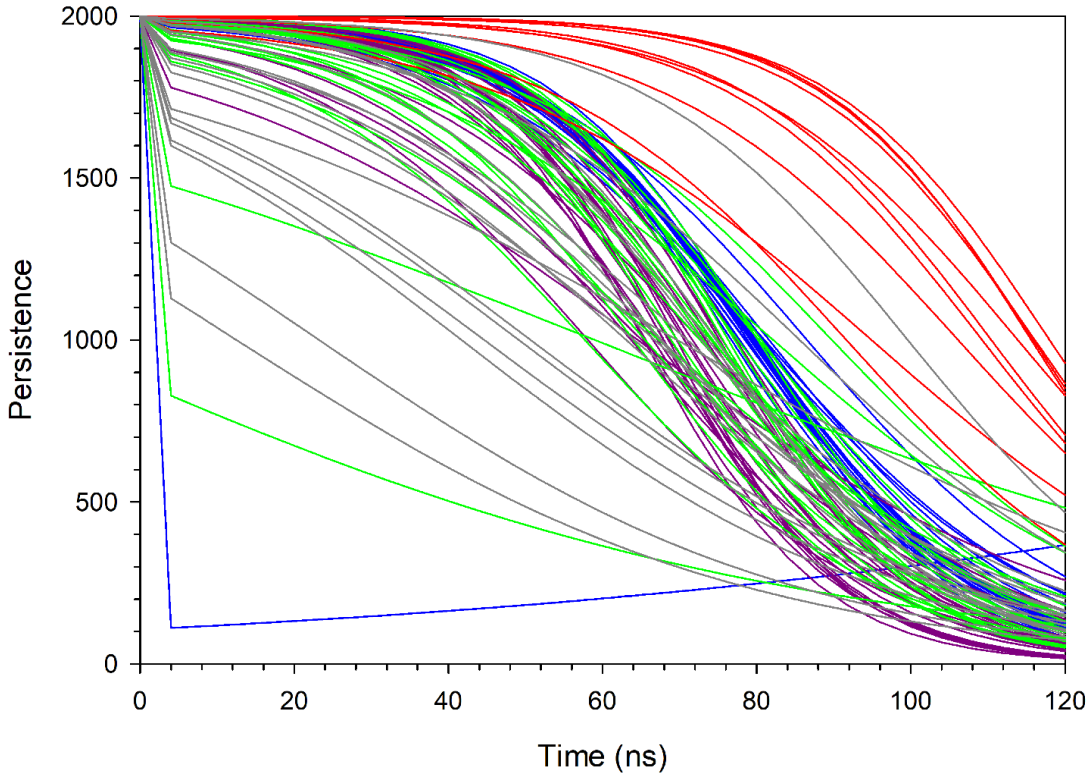


Figure 27. Persistence of long-range interactions in GB1. Interactions within the N- and C-terminal hairpins are shown in blue and red, respectively. Interactions between the α-helix and either hairpin are shown in green. Interactions between the N- and C-terminal hairpins are shown in purple. Interactions involving loop regions are shown in gray. Data plotted using SigmaPlot 12.5.

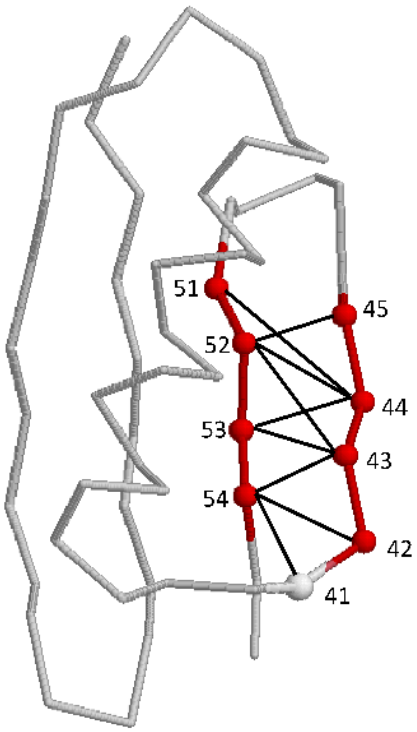


Figure 28. Persistent long-range interactions in GB1. Residues involved in persistent long-range interactions are shown as spheres. Residues are located in the C-terminal hairpin (red) and loop region (gray). Structure visualized using RasMol Ver. 2.7.2.1.1.

To further explore the structural nature of the unfolding transition, snapshots of GB1 were taken over the course of the simulation to visually assess the unfolding transitions. These snapshots are shown in Figure 29. The assessment of the long-range interactions during the simulations reveal that the two β-hairpins separate first at 52ns followed by movement of the α-helix away from the main structure at 56ns. The N-terminal hairpin is next to unfold at 78ns followed by partial unfolding of the C-terminal hairpin at 107ns.

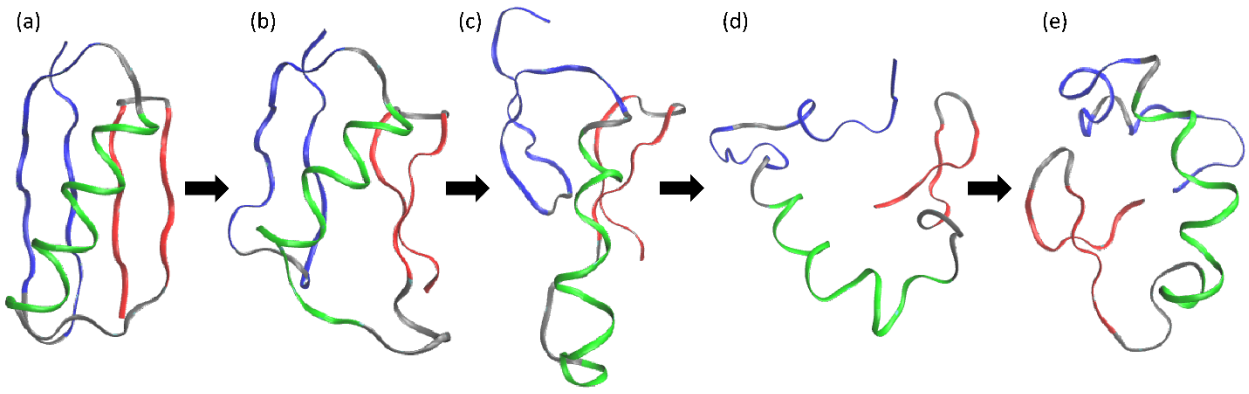


Figure 29. GB1 unfolding simulation snapshots. The N- and C-terminal hairpins are shown in blue and red, respectively. The central α-helix is shown in green. Loops and termini are in gray. The structures in (a-e) represent configurations from the third trajectory at time points: 0ns, 52ns, 56ns, 78ns and 107 ns, respectively. Structures visualized using VMD 1.9.1.

Previous computational studies revealed the preferential formation of the C-terminal hairpin during early folding in comparison to the N-terminal hairpin of GB1 [64, 70] suggesting that during an unfolding event the C-terminal hairpin would be the most persistent. It is also likely that the C-terminal hairpin acts as a structurally stable element for helix docking [67]. The formation of the C-terminal hairpin is characterized by the long-range hydrophobic interaction between Trp43 and Phe52 [68], one of the nine persistent long-range interactions in the present study. Further evidence using MD analysis suggests the conformation of the C-terminal hairpin determines whether the final structure will be properly folded [65]. Previous molecular dynamics simulations involving ten folding and unfolding events by Shaw *et al*. found that either hairpin could fold first for a redesigned variant of GB1 [187].

Computational work, however advanced, ideally requires comparison to experimental research whenever possible. The computationally derived persistence data was then compared to experimental work done by Baker and coworkers [81]. These researchers experimentally determined ΔΔG values based on site-directed mutagenesis.

ΔΔG is a measure of the stability of the mutated protein against the wild-type. Residues that resulted in ΔΔG values greater than 0.3 kcal mol^{-1} were found to be important for the stability of GB1. The three residues with the highest ΔΔG values within the C-terminal hairpin, Tyr45, Phe52, and Val54, are involved in two-thirds (6 of 9) of the persistent long-range interactions we identified in the unfolding simulations of GB1. Thus, there seems to be a correlation between ΔΔG, and residues involved in persistent long-range interactions (Figure 28) indicating that they are important for protein stability. Additional experimental work by Bu *et al.* revealed that three residues (Phe30, Tyr45, and Phe52) were key to stability [73]. Two residues are located in the C-terminal hairpin and one residue is located in the central helix. We also identified these residues and detailed their stabilizing interactions. Idiyatullin *et al.* show experimentally, the locations of residues with the highest internal motion activation energy, which equates to stability [195]. They are located in strands 1, 3, and 4 as well as part of the helix, which in our present work are also the more stable elements. However, we are uniquely able to monitor stability at the level of individual long-range interactions during the entire unfolding process.

Our results were further compared to the experimental work of Orban and co-workers [83, 84, 196, 197]. This research involved mutating two proteins, the albumin binding domain of protein G (GA) and GB1, to near identical sequences (95%) while maintaining their distinct 3D folds. In three rounds of mutations, residues were mutated from the GB1 sequence to GA and

vice versa. The total number of mutations made to the GB1 sequence was 20. When compared to our data only one of the 20 mutated residues (Glu42) was found to be involved in persistent long-range interactions (Figure 30(a)). This research ultimately led to the seminal discovery that three critical residues were responsible for fold-switching, Ala20, Phe30, and Tyr45. One of the three critical residues (Tyr45) was found to be involved in persistent long-range interactions (Figure 30(b)). These findings further support our proposal that the residues comprising persistent long-range interactions in GB1 are important for structural stability of the protein.

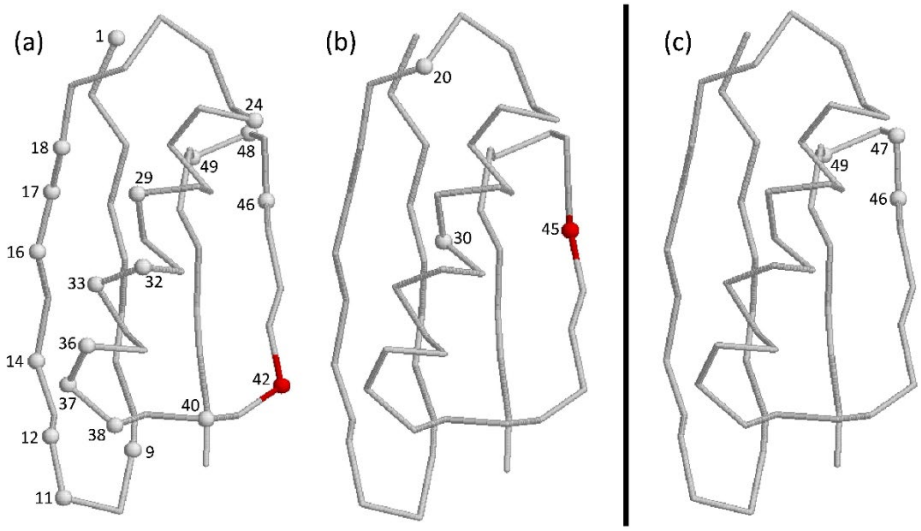


Figure 30. Select experimental studies of GB1. (a) Residues mutated by Orban and co-workers. The mutated residues are shown with numbered spheres. Residue 42 located in the C-terminal hairpin is shown in red and is persistent. (b) Fold switching residues in GB1. Residue 45 located in the C-terminal hairpin is shown in red and is persistent [83, 196, 197]. (c) Residues with high Φ-values (Asp46, Asp47, Thr49) are shown as numbered spheres [81]. Structures visualized using RasMol Ver. 2.7.2.1.1.

Lastly, our data was compared to Baker and co-workers Φ-value analysis [81]. Φ-values can range from zero to one with values closer to one indicating the residue is structured in the transition state. The three residues with the highest Φ-values make no long-range interactions within the protein, only short-range interactions and are in the turn of the second hairpin (Figure 30(c)). This suggests that residues involved in persistent long-range interactions are not the most important for the TS during folding as other residues within GB1. However, two residues with moderate Φ-values (Tyr45 and Thr51) are involved in persistent long-range interactions.

SAMP1 Persistence

The RMSDs of the three SAMP1 unfolding simulations were plotted as a function of time (Figure 26(b)). As in the case of GB1, these simulations used the same dynamics parameters, yet distinct RMSD values over the course of the simulations were observed. Further, the RMSDs are fluctuating indicating that SAMP1 is undergoing many unfolding and folding events but the overall RMSDs are trending upward indicating that the protein is unfolding over the course of the simulations.

The persistence of the 155 long-range interactions in SAMP1 was measured over the course of each simulation and averaged. The averaged persistence values were plotted versus time and are shown in Figure 31. The most persistent long-range interactions clearly separate themselves from the total population and are mostly blue. Specifically, 20 of 155 or 13% of the long-range interactions within SAMP1 are persistent. Unlike GB1 however, these 20 interactions are located in the N-terminal hairpin (Figure 32).

To further explore the structural nature of the unfolding transition, snapshots of SAMP1 were taken over the course of the third simulation to visually assess the unfolding transitions.

These snapshots are shown in Figure 33. The assessment of long-range interactions during the simulations reveal that the extra secondary elements, not present in GB1, move away from the main structure at 24ns followed by movement of the α-helix away from the main structure at 42ns. The C-terminal hairpin is next to unfold at 44ns followed by the separation of the hairpins at 48ns. Lastly, the N-terminal hairpin partially unfolds at 80ns.

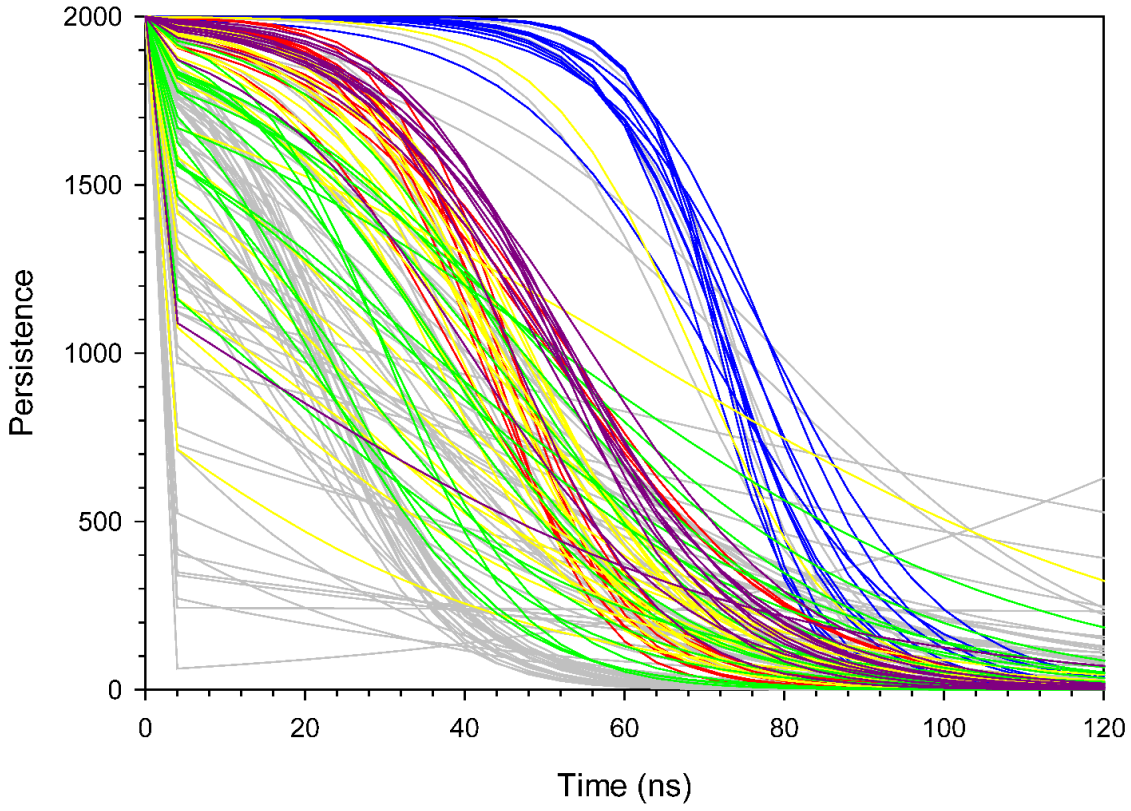


Figure 31. Persistence of long-range interactions in SAMP1. Interactions within the N- and C-terminal hairpins are shown in blue and red, respectively. Interactions between the central α-helix and either hairpin are shown in green. Interactions between the hairpins are shown in purple. Interactions involving loop regions and extra secondary elements are shown in gray and yellow, respectively. Data plotted using SigmaPlot 12.5.

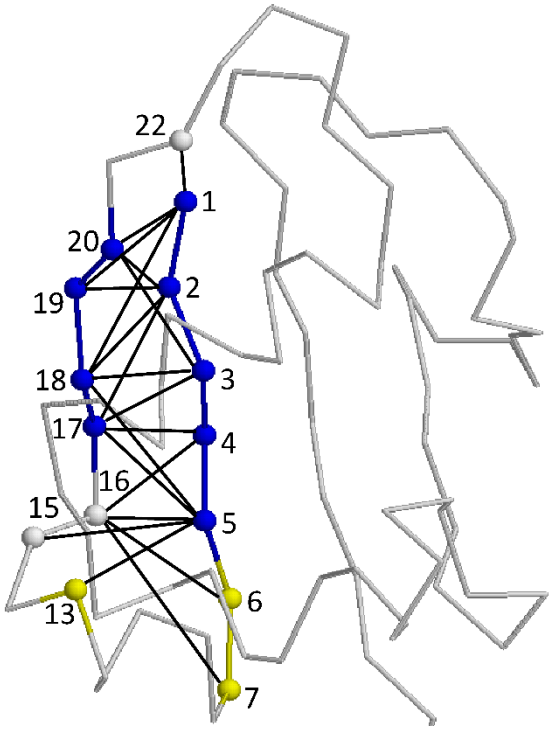


Figure 32. Persistent long-range interactions in SAMP1. Residues involved in persistent long-range interactions are shown as spheres. gray, yellow, and blue residues are located in loops, an α-helix, and the N-terminal hairpin, respectively. Residues 84-87 removed for visual clarity. Structure visualized using RasMol Ver. 2.7.2.1.1.

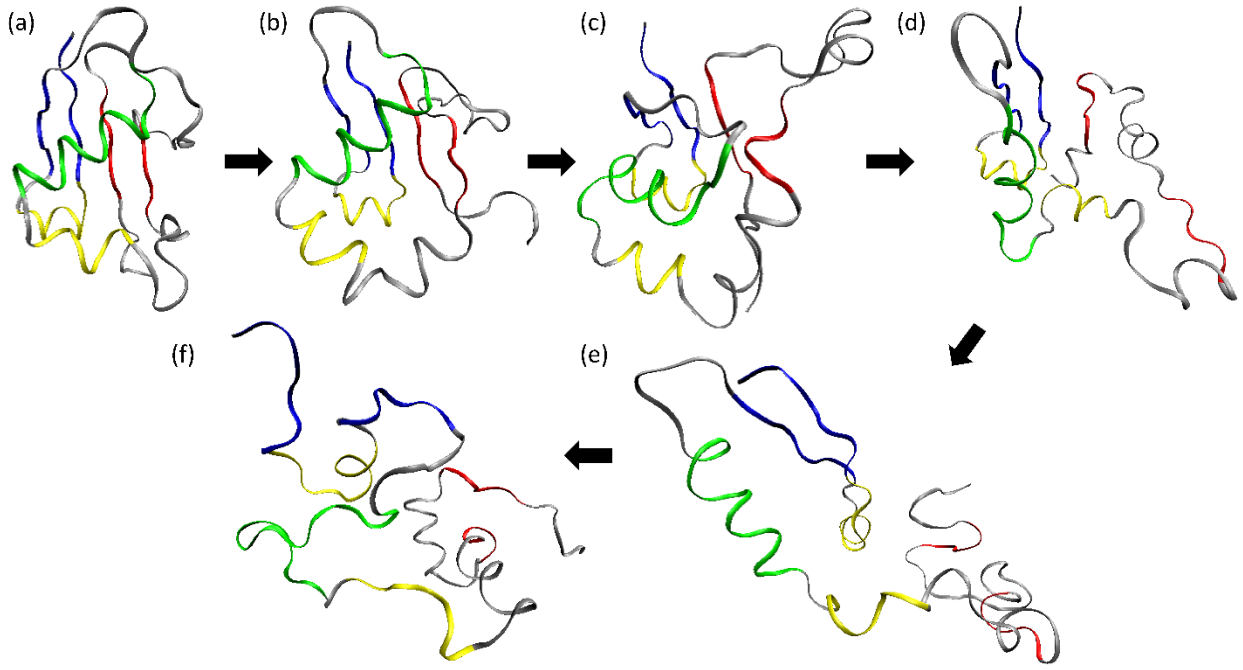


Figure 33. SAMP1 unfolding simulation snapshots. The N- and C-terminal hairpins are shown in blue and red, respectively. The central α-helix is shown in green with the two additional helical segments in yellow. Loops and termini are in gray. The structures in (a-f) represent configurations from the third trajectory at time points: 0ns, 24ns, 42ns, 44ns, 48ns, and 80ns, respectively. Structures visualized using VMD 1.9.1.

A comparative analysis of the structural nature and stability of these proteins was performed by creating a long-range consensus network of GB1 and SAMP1 and plotting these on a contact map (Figure 34). The analysis revealed 57 long-range interactions in common between the two structures. These interactions are predominately located within and between each of the hairpins. These results seem to indicate that β-hairpin and β-sheet formation is critical in maintaining the overall shape and stability of this topology.

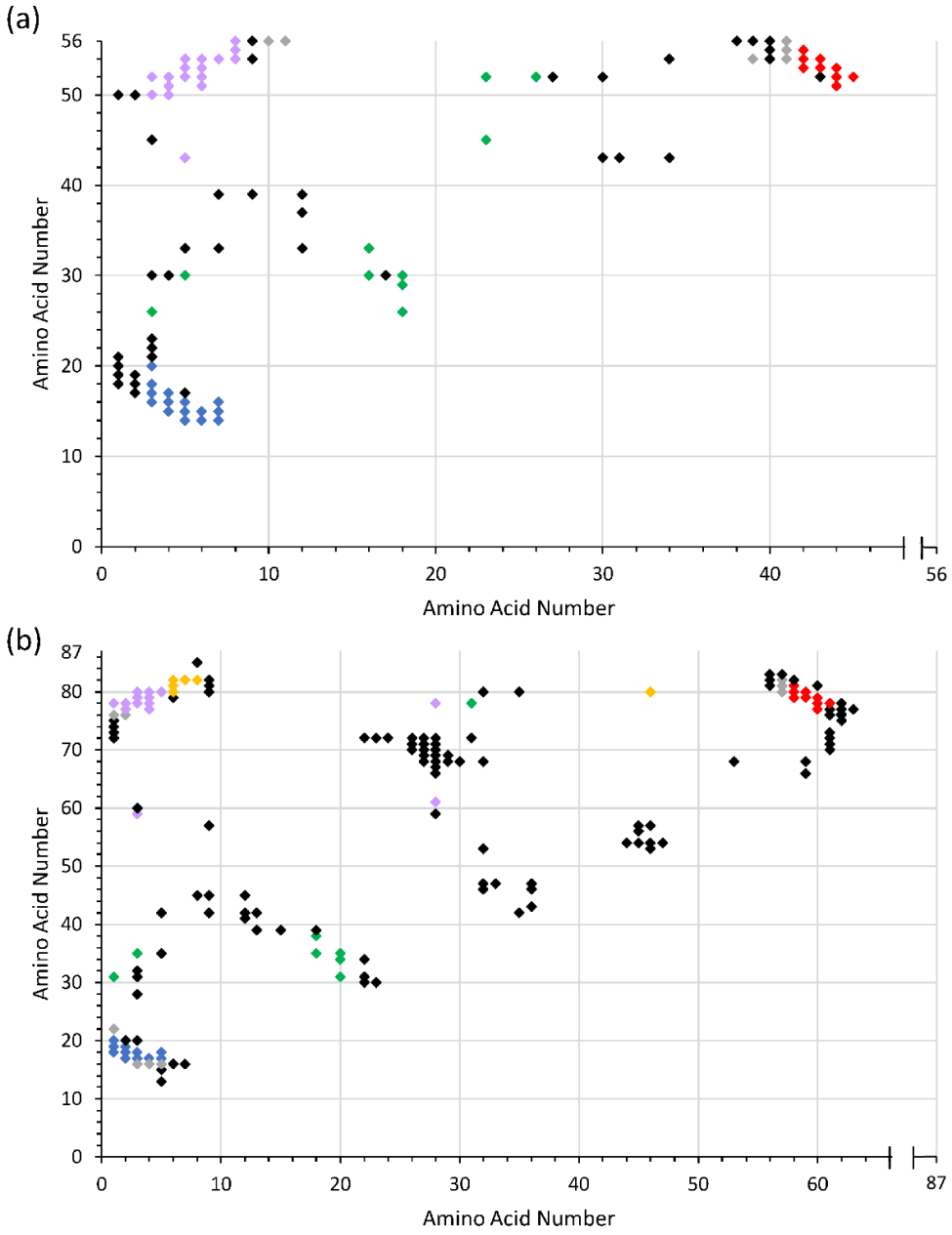


Figure 34. Long-range interaction contact maps of (a) GB1 and (b) SAMP1. The use of color (other than black) indicates a common interaction in both proteins. Interactions within the N- and C-terminal hairpins and between them are shown in blue, red, and purple, respectively. Green indicates interactions between the central α-helix and either hairpin. Interactions in extra secondary elements and loops are shown in yellow and gray, respectively. Data plotted using Microsoft Excel 365.

To address the nature of the forces that shift the local stability profile and kinetics from one hairpin to another we analyzed short- and long-range hydrophobic interactions involving residues that are in the core (90% buried) or peripheral to it (60-90% buried). Throughout the simulations we find that the bias in the native patterning of hydrophobicity in the core and periphery plays a central role. Here, destabilization of the core in layers facilitates unfolding of the protein. We also propose that transient non-native interactions (salt bridges and hydrophobic interactions) play a fundamental role in unfolding, which has received little recognition in the protein folding field (data not shown). In GB1, we propose the C-terminal hairpin is more stable than its N-terminal hairpin for several reasons. Within the C-terminal hairpin, there are three long-range hydrophobic interactions (Figure 35(a)) that have an average persistence of 88% in the MD simulations (Figure 35(b)). Within the N-terminal hairpin of GB1, one long-range and one short-range hydrophobic interaction (Figure 35(a)) have an average persistence of 26% and 34%, respectively (Figure 35(b)). This outcome is due to the fact that the central region of the second strand in the N-terminal contains no hydrophobic residues. However, the ends do have hydrophobic residues: Leu and Ala. Interestingly, there are weaker long-range interactions between the N-terminal hairpin and the loops with a persistence of 17% (Figure 35(c)). Thus, the C-terminal hairpin has stronger and more central interactions. Therefore, location of the hydrophobic residues is also key to conferring structural stability.

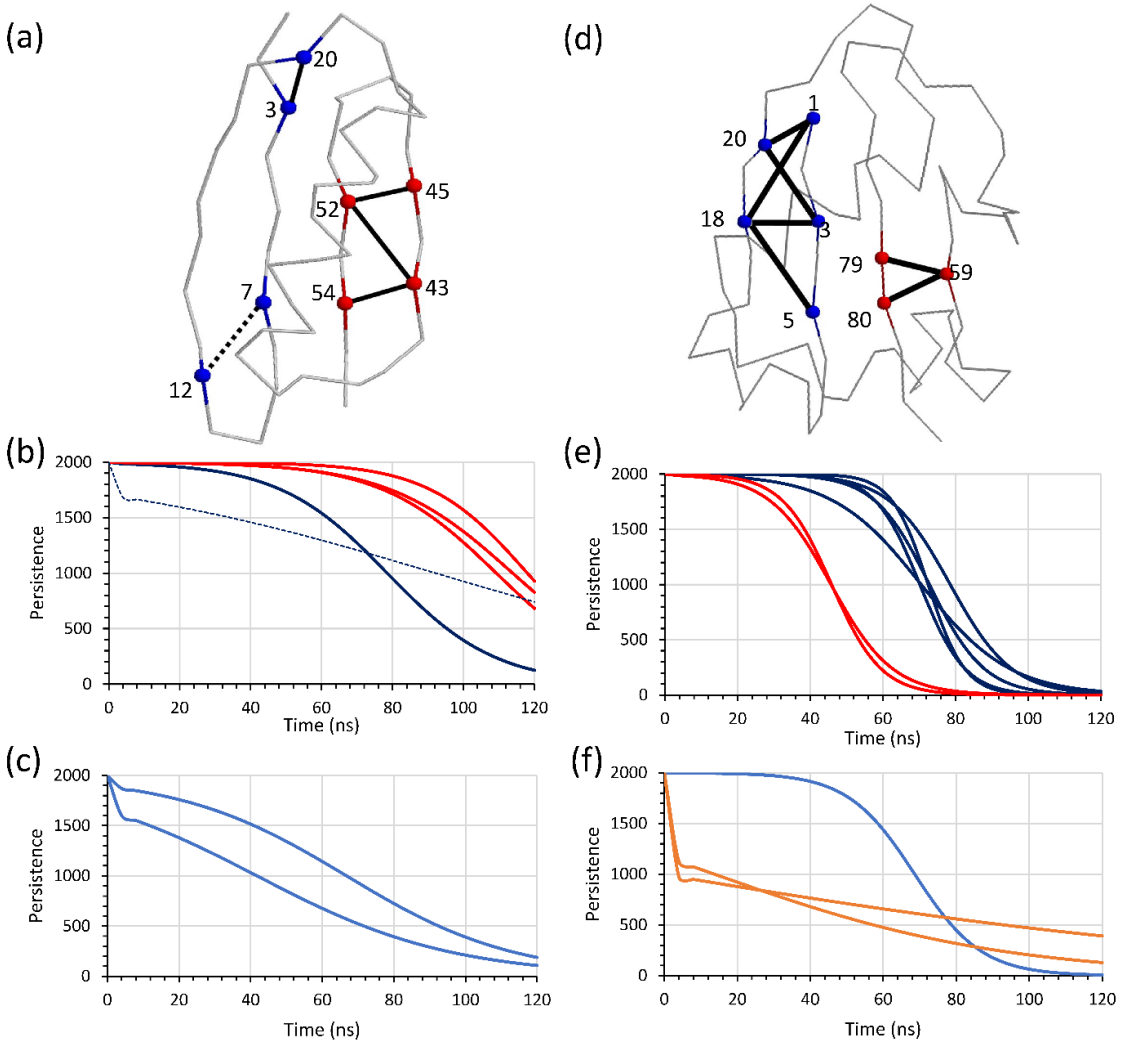


Figure 35. Persistence in the hydrophobic core and peripheral core of (a-c) GB1 and (d-f) SAMP1. (a, d) long-range (black solid lines) and short-range (black dashed lines) hydrophobic interactions in the N-terminal (blue) and C-terminal (red) hairpins. Residues 84-87 of SAMP1 removed for visual clarity. (b, e) Persistence of long-range (solid lines) and short-range (dashed lines) hydrophobic interactions in the N-terminal (blue) and C-terminal (red) hairpins. (c, f) persistence of long-range (solid lines) hydrophobic interactions between a residue in a hairpin and a residue in a loop. N-terminal and C-terminal hairpins are shown in light blue and orange, respectively. Structures visualized and data plotted using RasMol Ver. 2.7.2.1.1 and Microsoft Excel 365, respectively.

The hairpins are also further stabilized through interactions with the central helix, which is an integral part of the hydrophobic core (Figure 36(a)). In GB1, both hairpins are interacting with the central helix. In the N-terminal hairpin, the four long-range interactions with the helix involve residues in the first strand and have an average persistence of 25%. The one short-range interaction, containing a residue in strand two, has a persistence of 58%. Both strands of the C-terminal hairpin have interactions with the helix There are seven long-range interactions, and they have an average persistence of 29%. Thus, fewer persistent contacts and the fact that only one strand of the hairpin interacts with the helix indicates that N-terminal hairpin has weaker interactions with the hydrophobic core, making it susceptible to unfolding first.

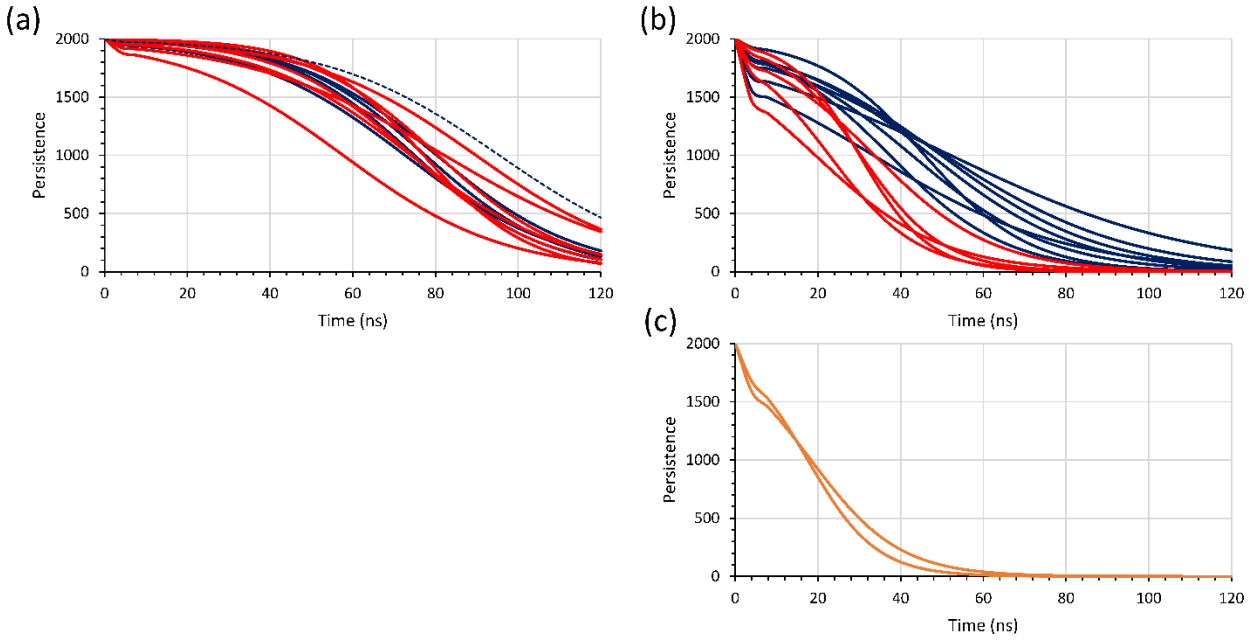


Figure 36. Persistence between the hairpins and central α-helix of (a) GB1 and (b) SAMP1. (a, b) Persistence of long-range (solid lines) and short-range (dashed lines) hydrophobic interactions in the N-terminal (blue) and C-terminal (red). (c) persistence of long-range (solid lines) hydrophobic interactions between residues in the loop of the C-terminal hairpin and the central α-helix. Data plotted using Microsoft Excel 365.

Within the N-terminal hairpin of SAMP1 are five hydrophobic interactions (Figure 35(d)) that have an average persistence of 59% in the MD simulations (Figure 35(e)). One long-range interaction exists between the N-terminal hairpin and a loop with a persistence of 56% (Figure 35(f)). Within the C-terminal hairpin of SAMP1 are two hydrophobic interactions (Figure 35(d)) that have an average persistence of 37% (Figure 35(e)) in the MD simulations. Two weak long-range interactions exist between the C-terminal hairpin and a loop with an average persistence of 30% (Figure 35(f)).

As in GB1, the central helix has a stabilizing effect on the two hairpins in SAMP1 (Figure 36(b)). The N-terminal hairpin contains eight hydrophobic interactions with the central α-helix that have an average persistence of 37%. The C-terminal hairpin contains seven long-range hydrophobic interactions with the central α-helix. Four involve residues in strand four, one involves a residue in strand three, and two residues are in a loop (Figure 36(c)). The average persistence values are 21%, 28%, and 15%, respectively. The stabilizing hydrophobic interactions with the N- and C-terminal hairpins of both proteins are listed in Table 3.

Table 3 Hydrophobic core and peripheral core interactions in GB1 and SAMP1. Italics indicate short-range interactions

GB1	SAMP1
Tyr3 - Ala20	Met1 - Val18
Leu7 - Leu12	Met1 - Val20
Trp43 - Phe52	Trp3 - Val18
Trp43 - Val54	Trp3 - Val20
Tyr45 - Phe52	Leu5 - Val18
	Val59 - Ala79
	Val59 - Leu80

As we find with GB1's strand two, SAMP1's strand three is weakly associated with the hydrophobic core. Thus, this is in large part the reason for the swapping of stability in the symmetrical proteins.

Comparing our results to a homologue of GB1, protein L, which has been studied experimentally using Φ-value analysis, indicates the N-terminal hairpin of protein L forms early [198]. Interestingly, in this protein, unlike GB1, the most stable region is the N-terminal hairpin.

This finding is analogous to what we find in the computational studies of SAMP1. However, ψ-value studies support the β-sheet comprised of both the N- and C-terminal hairpins forming early. Interestingly, the computational MD work of Cheng and co-workers have simulated transitions state structures that accommodate both of these experimental results [64]. Additionally, several other research investigations reveal that the N-terminal hairpin is the earlier hairpin to become structured [198, 199].

SUMMARY

The results of the comparative study with GB1 and SAMP1 reveal that either of the two β-hairpins can be the most stable. This is in large part due to the polarization of the hydrophobic cores and location of key long-range interactions. GB1 contains a hydrophobic core comprised of select residues with low solvent accessibility that is polarized toward the C-terminal hairpin. Conversely, SAMP1 contains a hydrophobic core comprised of select residues with low solvent accessibility that is polarized toward the N-terminal hairpin. Thus, the location of the hydrophobic core and select long-range interactions and the hydrophobic forces therein appears to lend itself to the preferential stability of one hairpin over the other as seen in the native state. The unfolding simulations uniquely allowed an in-depth investigation into the unfolding process and governing forces. When coupled to the results from previous studies of protein L, it is clear, that symmetrical stability is a central feature of the β-grasp fold.

CHAPTER IV

EFFECTS OF IONIC STRENGTH ON FOLDING AND STABILITY OF A HALOPHILIC PROTEIN

OVERVIEW

The research presented in this chapter focuses on a halophilic protein, the small archaeal modifier protein 1 (SAMP1), within the β-grasp fold superfamily. In the field of protein biochemistry, far less is known about the thermodynamic and kinetic behavior of halophilic proteins in comparison to mesophilic proteins. Thus, a deeper understanding can provide a clearer view of the determinants of folding and stability [200-205]. Halophilic proteins have a number of unique features such as a larger number of acidic residues which is consistent with analysis of SAMP1 (Table 4) [206]. Aspartic acid and glutamic acid residues are also considered two of the prebiotic residues [207]. Halophilic proteins are attractive models for study as a high salt environment is considered one of the potential primordial conditions in which proteins first evolved [207]. Halophiles are also thought to be particularly adaptive to changing temperature and pH conditions which would have been a favorable feature in an evolving world [206]. Interestingly, a study conducted towards investigating the evolution of halophiles to mesophiles revealed that incorporating an aromatic residue, which is not considered a primordial amino acid, in the core of a designed primitive protein, converted the folding behavior from halophilic to mesophilic conditions [207]. The formation of the essential peptide bond is also considered more favorable under high salt [207-209].

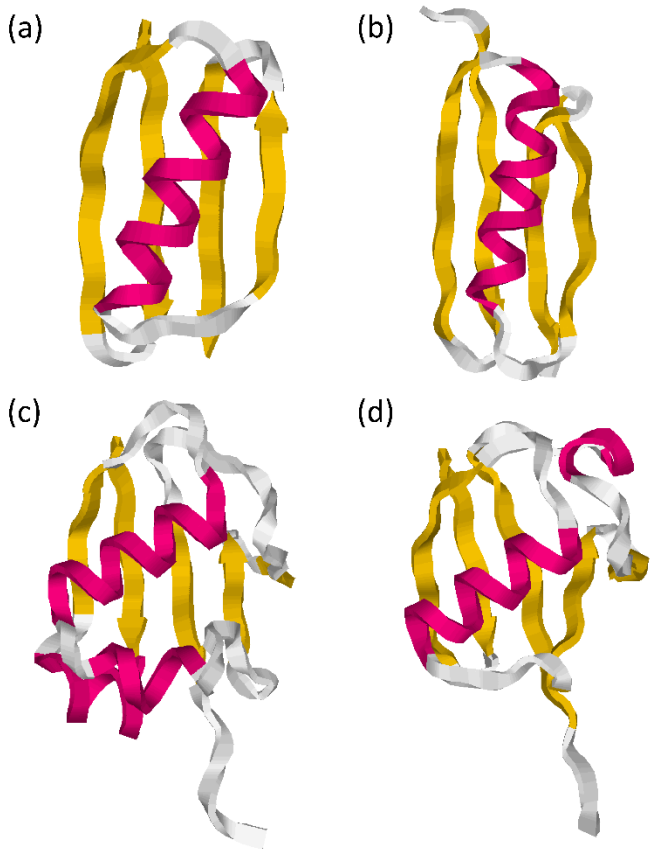


Figure 37. Structures of select β-grasp superfamily members. (a) GB1, (b) protein L with residues 1-14 removed, (c) SAMP1, and (d) ubiquitin. PDB codes are 1PGB, 2PTL, 3PO0, and 1UBQ, respectively. α-helices and β-strands are shown in magenta and yellow, respectively. Structures visualized using RasMol Ver. 2.7.2.1.1.

Here I present the results of experimental folding and unfolding studies conducted using SAMP1 from *Haloferax volcanii* [194]. While sharing the same fold as GB1, ubiquitin, and protein L, it differs in sequence length and contains additional helical secondary structure (Figure 37). SAMP1 is more closely related in structure and sequence to ubiquitin (Table 4). SAMP1 has very interesting features in contrast to other β-grasp superfamily members studied thus far. As a

protein expressed by a halophile, it evolved to fold and maintain its stability at higher salt concentrations. It is therefore more stable and highly structured in comparison to GB1, protein L, and ubiquitin, which is expected of halophilic proteins [206].

Table 4 Characteristics of β-grasp superfamily members.

	GB1	Protein L	Ubiquitin	SAMP1
Sequence Length	56	78	76	87
Number of Secondary Elements	5	5	7	7
Number of Positive Residues	6	8 (7)[a]	11	5
Number of Negative Residues	10	17 (8)[a]	11	17
Relative Contact Order	0.17	0.18	0.15	0.14
Absolute Contact Order	9.70	11.13	11.47	12.47
$\ln(k_f)$	6.00[c]	4.10[b]	5.90[d]	3.41 (1.0 M NaCl) 0.202 (0.3 M NaCl) -0.675 (0.1 M NaCl)
Percent Identity and *RMSD*	GB1	Protein L	Ubiquitin	SAMP1
GB1	100 *0*	10 *2.5*	7 *3.1*	2 *3.1*
Protein L	10 *2.5*	100 *0*	7 *4*	6 *4.1*
Ubiquitin	7 *3.1*	7 *4*	100 *0*	8 *2.5*
SAMP1	2 *3.1*	6 *4.1*	8 *2.5*	100 *0*

[a] The number shown in parenthesis is lower when the long intrinsically disordered tail is not included (residues 1-17).
[b] Folding rate based on a protein length of 63 residues [210, 211].
[c] $\ln(k_f)$ obtained from [212].
[d] $\ln(k_f)$ obtained from [212, 213].

The results of folding studies revealed that SAMP1 folds faster at high versus low ionic strength. With little information on the folding of halophilic proteins this study also provided the opportunity to examine the folding behavior near the solubility point of NaCl at 25°C. A comparison of SAMP1 to the folding kinetics mechanisms of GB1, ubiquitin, and protein L provides greater insight into the underlying nature of the β-grasp protein fold.

MATERIALS AND METHODS

Materials and Equipment

The buffer for the protein study was composed of a mixture of mono- and di-basic sodium phosphate and sodium chloride from Fisher Scientific (Pittsburgh, PA). Ultrapure urea for protein unfolding was obtained from MP Biochemicals, Inc. (Solon, OH). The studies with *Escherichia coli* BL21(DE3) to produce the protein included Luria-Bertani media (LB) and isopropyl β-D-thiogalactoside (IPTG). The protein was purified using a Ni-NTA column (GE Healthcare). Equilibrium fluorescence was conducted using a PTI QM-2000 spectrofluorometer (Photon Technology International, Inc., South Brunswick, NJ). Folding and ultrafast folding kinetics were conducted using a SX-20 stopped-flow instrument (Applied Photophysics, Ltd., Leatherhead, U.K.) and an in-lab built continuous flow mixer [214], respectively. Circular dichroism was conducted using a JASCO J-815 spectropolarimeter.

Protein Expression and Purification

The recombinant plasmid (pET-22b(+)) containing the SAMP1 gene cloned in the Nde I/Xho I sites was used for expression of the SAMP1 protein [215]. It was transformed into *E. coli* BL21(DE3) and the cells grown in LB at 37 °C to an OD_{600} of 0.8. Expression of the

recombinant protein was induced at 16 °C for 20 hours using 0.5 mM IPTG. The cells were harvested by centrifugation at 4 °C, 4000 g for 10 mins, and were resuspended with lysis buffer (20 mM Tris, 400 mM NaCl, 5 mM imidazole, pH 7.8). Cells were lysed by sonication and the lysate was cleared by centrifugation at 4 °C, 15000 g for 30 mins. The supernatant was collected and passed through a Ni-NTA column and the column was washed with 30 ml wash buffer (20 mM Tris, 400 mM NaCl, 70 mM imidazole, pH 7.8). The recombinant protein, including an N-terminal Leu-Gln-His_6 sequence, was eluted with elution buffer (20 mM Tris, 400 mM NaCl, 70 mM imidazole, pH 7.8) and dialyzed in buffer containing 10 mM Tris, 50 mM NaCl pH 7.5. These studies were performed by my collaborator Dr. ShanHui Liao at the University of Science and Technology of China.

Equilibrium Unfolding Monitored by Fluorescence

Urea-induced equilibrium fluorescence of SAMP1 was measured in 50 mM sodium phosphate, pH 7.0, with varying amounts of NaCl: 100 mM, 300 mM, and 1.0 M. Trp fluorescence was monitored using a PTI QM-2000 spectrofluorometer. All experiments were conducted with 2 μM protein at 20°C and three wavelength scans were collected for each concentration. Excitation and emission wavelengths were 290 nm and 300-450 nm, respectively. Slit widths were 1 nm for excitation and 6 nm for emission. The complete 2D data set (fluorescence vs. urea concentration and wavelength) was fitted to a 2-state model using a global fitting procedure in Igor Pro, ver. 6.37 (WaveMetrics, Inc), as described in Latypov and in Maki [216, 217]. These studies were completed at the Fox Chase Cancer Center in Philadelphia, PA in collaboration with Drs. Heinrich Roder and Takuya Mizukami.

Folding Kinetics Monitored by Fluorescence

All data were acquired at 20 ºC in 0.1 M sodium phosphate buffer (pH 7.0) containing varying concentrations of sodium chloride. An in-lab built continuous flow mixing instrument comprising a microfluidic mixer with a mixing time of ~10 μs and a 266 nm DPSS laser (Mizukami et al., in preparation) was used to monitor fluorescence changes associated with the kinetics of folding and unfolding of SAMP1 on the sub-millisecond time scale. Tryptophan fluorescence emission was measured using a 310 nm cut-on filter. The observation channel had a depth of 0.2 mm and a variable width from 0.15-0.6 mm. For folding/unfolding measurements on a time scale of ~1 ms and longer we used an Applied Photophysics SX-20 stopped flow instrument equipped with a 1 mm cuvette. Tryptophan fluorescence was excited at 290 nm and emission was monitored using a 310 nm cut-on filter. The final protein concentration in stopped-flow and continuous flow experiments was 1.1 and 2.0 μM, respectively. The folding data from the different urea and salt concentrations were fitted to a 2-state and 3-state model following a global fitting procedure in Igor Pro, ver 6.37 (WaveMetrics, Inc). These studies were completed at the Fox Chase Cancer Center in Philadelphia, PA in collaboration with Drs. Heinrich Roder and Takuya Mizukami. Kinetics figures were constructed from data acquired with Dr. Takuya Mizukami.

Circular Dichroism

Near- and far-UV circular dichroism spectra were obtained for SAMP1 using a Jasco J-815 spectropolarimeter under native and denatured conditions, 50 mM sodium phosphate and 50 mM sodium phosphate/6.0 M guanidinium chloride, respectively, for both low salt, 62.5 mM,

and high salt, 962.5 mM conditions at 20.0 °C. A 1 cm and 0.1 cm cuvette were used for near- and far-UV CD, respectively. Relevant IBC protocol numbers are 16-005 and 17-010.

Computational Studies

Native protein structures were analyzed and visualized using Chimera for electrostatics and VMD (v. 1.9.1) for salt bridges using a 4.0 Å cutoff. SAMP1 unfolding simulations were conducted as described in Bedford *et al*. [218]. Salt bridges in the MD trajectories were calculated using VMD (v.1.9.1) using a 4.0 Å cutoff. To calculate relative and absolute contact order, the website developed by Plaxco and Baker was used. https://depts.washington.edu/bakerpg/contact_order/contact_order.cgi

RESULTS AND DISCUSSION

Circular Dichroism

Figure 38 presents an equilibrium analysis of SAMP1 monitored by near- and far-UV CD. The results suggest that at high ionic strength SAMP1 is more structured than at low ionic strength as evidenced by the difference in molar ellipticity of the near-UV CD spectra, particularly around the Phe residues (wavelengths 255-270 nm) [147].

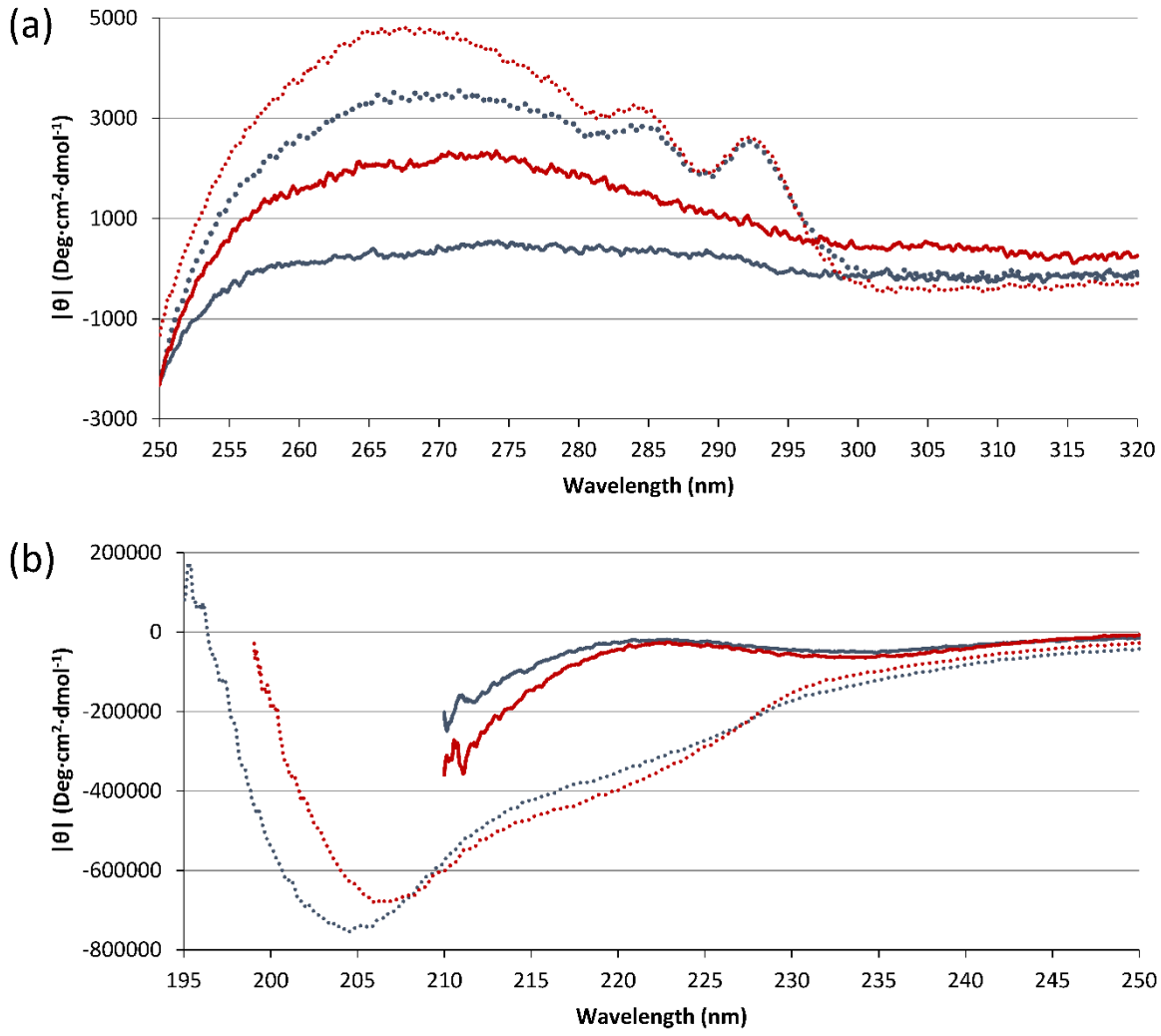


Figure 38. Circular dichroism spectra of SAMP1. SAMP1 at 62 mM NaCl (blue) and 965 mM NaCl (red) under native (dotted-line) and denatured (solid-line) conditions. Panels (a) and (b) are near- and far-UV CD, respectively.

Analysis of electrostatic interactions

We also analyzed the electrostatic interactions of SAMP1 at both high and low ionic strength. The surface potential of SAMP1 at high-ionic strength was calculated using the

Coulombic model in Chimera (Figure 39). Most of the surface is negative with only a couple small patches of net positive charge. A salt bridge analysis utilizing VMD with a cutoff distance of 4.0 Å [219] on the native crystal structure revealed two valid results, Glu70-Arg61, Asp76-Arg61. Upon visual analysis it appears there may also be a third salt bridge present, Glu2-Arg19. In comparison, both GB1 and ubiquitin contain three salt bridges in their native states while protein L contains none.

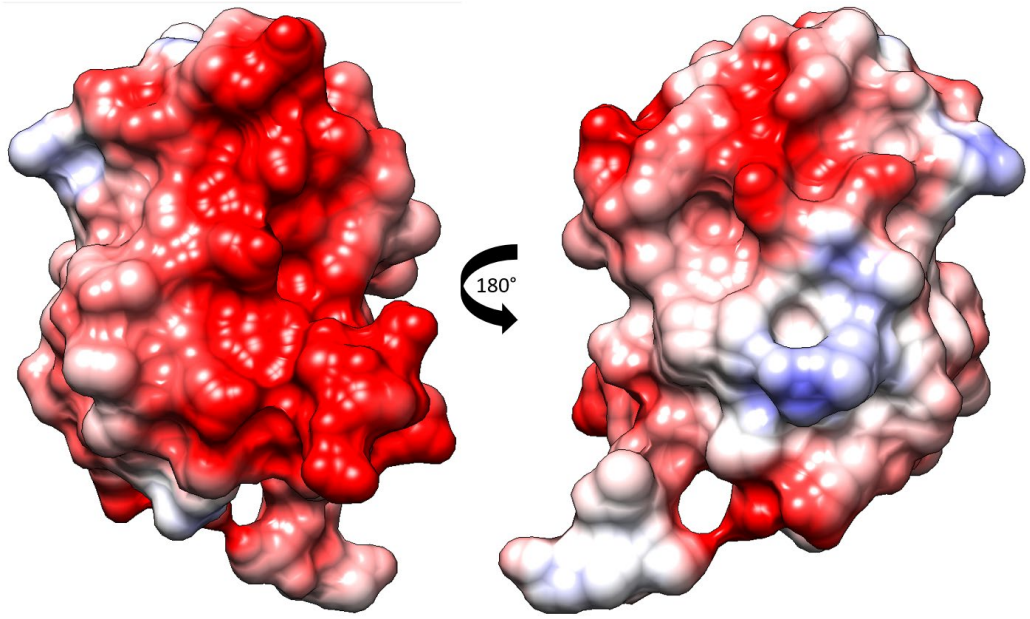


Figure 39. Surface potential of the high-ionic strength form of SAMP 1 (PDB code: 3PO0). Red, white, and blue indicate negative, neutral, and positive surface potential, respectively. Surface potential was calculated using the coulombic model in Chimera 1.14.

In another study, monitoring the salt bridges during the unfolding of SAMP1 by molecular dynamics simulations at high temperature reveals that Glu70-Arg61 and Asp76-Arg61 are lost early in the 120 ns simulation at approximately 20 ns and 35 ns, respectively. The third possible salt bridge, Glu2-Arg19, is maintained for approximately 70 ns and then abruptly breaks. Interestingly, between the three simulations there are 58 transient salt bridges formed during the unfolding process (Table 5). Details of the unfolding simulations using molecular dynamics can be found in Bedford *et al.* [218].

Table 5 Transient salt bridges formed during all three SAMP1 unfolding simulations.

Salt Bridges				
Asp-Arg16	Asp25-Lys4	Asp50-Arg61	Glu2-Lys4	Glu64-Arg61
Asp8-Arg45	Asp30-Arg16	Asp55-Arg16	Glu11-Arg16	Glu64-Lys4
Asp8-Lys4	Asp30-Arg19	Asp55-Arg19	Glu11-Arg19	Glu70-Arg16
Asp21-Arg16	Asp30-Arg61	Asp55-Arg45	Glu11-Arg45	Glu70-Arg19
Asp21-Arg19	Asp33-Arg16	Asp76-Arg16	Glu11-Lys4	Glu70-Arg61
Asp21-Arg61	Asp33-Arg19	Asp76-Arg19	Glu43-Arg16	Glu70-Lys4
Asp23-Arg16	Asp33-Arg61	Asp76-Arg61	Glu43-Arg45	Glu77-Arg16
Asp23-Arg19	Asp33-Lys4	Asp76-Lys4	Glu43-Lys4	Glu77-Arg19
Asp23-Arg61	Asp49-Arg19	Glu2-Arg16	Glu52-Arg16	Glu77-Arg61
Asp25-Arg16	Asp49-Arg45	Glu2-Arg19	Glu52-Arg45	Glu77-Lys4
Asp25-Arg19	Asp49-Lys4	Glu2-Arg45	Glu52-Lys4	
Asp25-Arg61	Asp50-Arg45	Glu2-Arg61	Glu64-Arg19	

Folding and Unfolding Studies

The fluorescence changes at a representative wavelength associated with urea-induced unfolding and refolding of SAMP1 at NaCl concentrations of 0.1, 0.3, and 1.0 M are shown in Figure 40. Folded and unfolded populations that were obtained by a global fit of a two-state

model at each salt concentration are shown and parameters characterizing these transitions are listed in Table 6.

Table 6 *m*-values at corresponding sodium chloride concentrations.

[NaCl]/M	*m*(kcal/mol ·M)		
	No constraints	Global baselines	Global *m*-value
0.1	0.645	1.737	
0.3	1.370	0.938	0.859
1.0	0.937	0.740	

The observed increase in the mid-point concentration of the unfolding transitions, C_m, with increasing salt concentration indicates that the folded state is strongly favored at higher ionic strength (Figure 40). *m*-values are calculated in order to understand the change in accessible surface area of the transition state (Table 6). If the low-salt state, M, is less compact than the high-salt state, N, then we should have m(low-salt) < m(high-salt). However, we if fact observed that m(0.1 M NaCl) < m(1.0 M NaCl) < m(0.3 M NaCl).

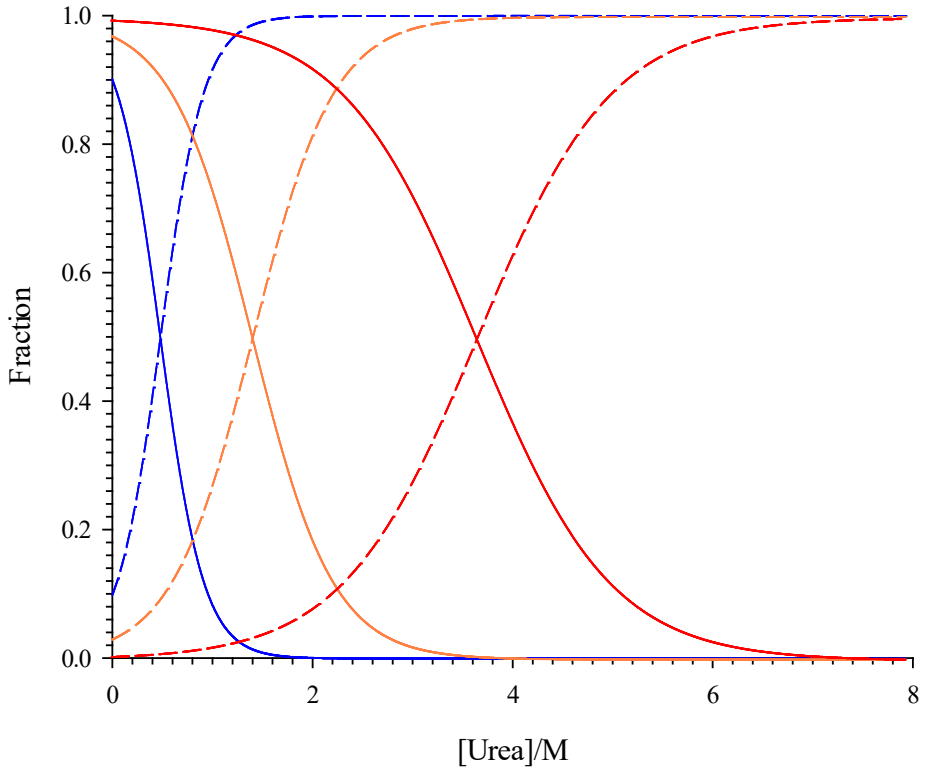


Figure 40. Equilibrium population at 100 mM (blue), 300 mM (orange) and 1.0 M (red) NaCl estimated by fluorescence spectroscopy. The solid and dashed lines show the population of native and unfolded states, respectively. Data plotted using SigmaPlot 12.5.

This is also evident in Figure 41, which shows that as salt concentration increases the denaturation midpoints of the chevron plots also increase. Chevron plots are constructed by combining two rate constants as shown in equation 17. When combined they form a V-shaped curve [220].

$$\ln k_{obs} = \ln \left(k_f^{H_2O} \exp\left(-m_{k_f}[denaturant]\right) + k_u^{H_2O} \exp\left(m_{k_u}[denaturant]\right)\right) \quad (17)$$

The data reveals that the structure at high ionic strength (1.0 M) folds more rapidly.

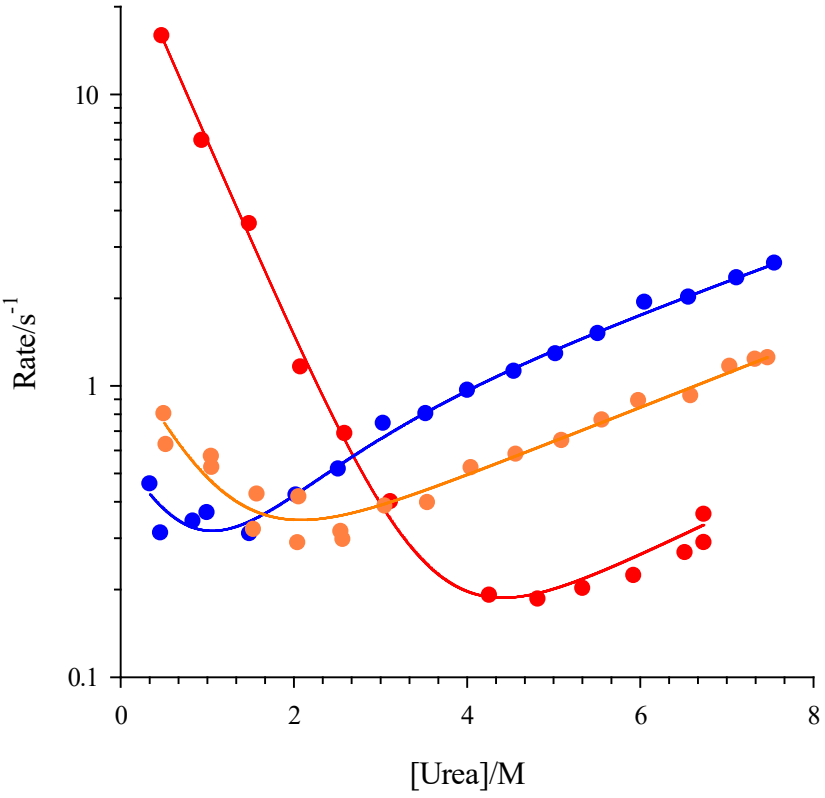


Figure 41. Chevron plots of stopped-flow experiments done at medium speeds at [NaCl] of 100 mM (blue), 300 mM (orange), 1.0 M (red). The lines show fitting curves to a two-state model. Data plotted using SigmaPlot 12.5.

The salt dependence on the free energy landscape was calculated and plotted in Figure 42(a). The results show that as sodium chloride concentration increases, protein stability increases. The Tanford β value was then applied to better understand the rates for denaturation unfolding. β_T is a measure of the degree of exposure of the transition state relative to the native and unfolded states and is therefore a good indicator of the compactness of the transition state [220]. Figure 42(b) shows that the protein structure becomes more compact as it folds.

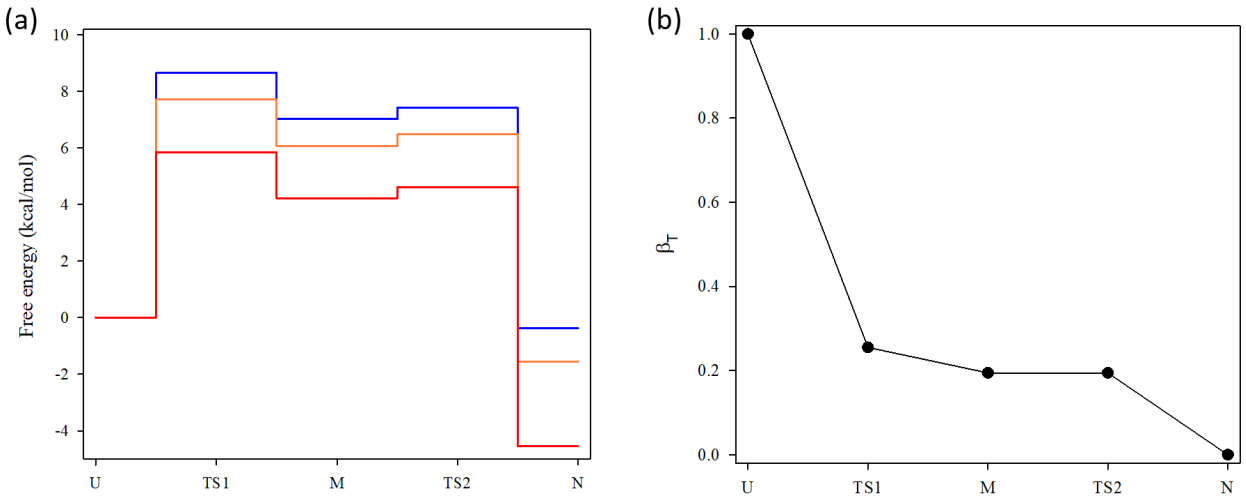


Figure 42. Salt dependence of the free energy landscape and Tanford β value. The three colors represent different NaCl concentrations: 100 mM (blue), 300 mM (orange) and 1.0 M (red) NaCl. Panel (a) shows the free energy calculations for each state (U = unfolded state, TS1 and TS2 are the two transition-states, M is the near native state and N = native state. Panel (b) graphs the β_T-value against the U, TS1, M, TS2 and N-states. Data plotted using SigmaPlot 12.5.

A salt concentration dependent study examining folding rates was conducted by monitoring the change in fluorescence intensity over time by using continuous-flow and stopped-flow fluorescence (Figure 43). The reaction was initiated by the salt-jump to the target NaCl concentration and by the dilution of urea concentration from 4.0 M to 0.36 M. The salt induced folding is a triphasic reaction. The kinetic traces show a minor increasing phase in the sub-millisecond time window followed by a fast major decreasing phase and a slow minor decreasing phase. The rate constant of the major phase, which is within the same order of magnitude, well matches the data obtained from stopped-flow shown in Figure 41. The results indicate that as the concentration of salt increases, the folding rate of the dominant phase increases up to 2.0 M NaCl

followed by a small decrease, indicating that the reaction reaches the upper rate limit around 300 s^{-1} in the presence of high salt (Figure 43 and Table 7).

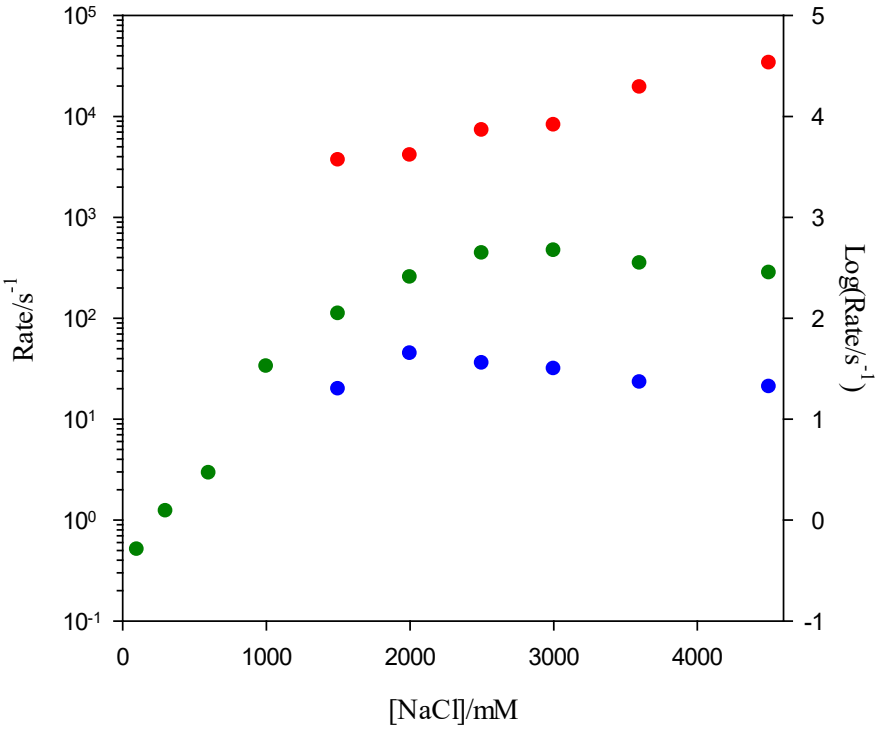


Figure 43. Salt-dependence on folding rates. The salt-induced folding kinetics is initiated by salt-jump and/or urea-jump by the continuous flow and stopped-flow fluorescence. The three rate constants measured are plotted as a function of salt concentration. The rate constant of the fastest rising phase observed in the continuous flow experiments is shown as red circles. The second fastest rate constant of the major decreasing phase observed in the stopped flow experiments is shown in green, while the slowest minor phase in blue. The green circles at 0.1, 0.3 and 1.0 M NaCl are obtained from the chevron plots shown in Figure 41. Data plotted using SigmaPlot 12.5.

Table 7 Salt-dependence on folding rates.

[NaCl]/M	k_1/s^{-1}	k_2/s^{-1}		k_3/s^{-1}
		Observed	Extrapolated to 0M urea	
0.1	-	0.46[a]	0.51[b]	-
0.3	-	0.81[a]	1.22[b]	-
0.6	-	2.92	-	-
1.0	-	15.9[a]	30.3[b]	-
1.5	0.37 X 10^4	110	-	20
2.0	0.41 X 10^4	254	-	45
2.5	0.73 X 10^4	440	-	36
3.0	0.82 X 10^4	469	-	31
3.6	1.94 X 10^4	350	-	23
4.5	3.38 X 10^4	282	-	21

[a] The rate constants are obtained at the lowest urea concentration (0.36 M) of the chevron plot experiments.
[b] The rate constants are obtained by extrapolating the fitting results of chevron plots to the 0.0 M urea.

There is limited knowledge of how extremophiles, particularly halophiles, fold and stabilize their native structure. The equilibrium unfolding data showed that with increasing salt, the protein becomes more structured and more stable, thus the more urea needed for unfolding. A limitation was reached where the studies could not go above 10.0 M urea due to solubility issues. In terms of Debye-Huckel screening of the large number of unfavorable interactions among acidic side chains, stability increases strongly with increasing salt concentration. For example, the FynSH3 domain, contains a negatively charged cluster on its surface, and its mutant are approximately 1-2 kcal/mol stable at low salt concentrations due to unfavorable electrostatic

repulsion. The FynSH3 domain is stabilized to approximately 4 kcal/mol at high salt concentrations due to the screening effect of unfavorable interactions [221]. Similarly, the stability of SAMP1 increases from 0.06 kcal/mol at 100 mM NaCl to 3.14 kcal/mol at 1.0 M NaCl, indicating that changes in stability due to salt concentration is a common property of highly charged proteins.

In addition to equilibrium studies, kinetic behavior can be directly understood through the use of stopped-flow. Continuous flow experiments are performed only at higher sodium chloride concentrations because the kinetic amplitude of the fastest phase becomes smaller at lower concentrations of sodium chloride. Stopped-flow data was obtained by refolding SAMP1 at sodium chloride concentrations of 0.1 M, 0.3 M, and 1.0 M sodium chloride. The kinetic data was fit to a chevron plot that enables a direct evaluation of rate versus urea concentration under the different salt concentrations. Folding branches of the chevron plots for 0.1 M and 0.3 M NaCl are short compared to 1.0 M NaCl. The denaturation midpoints for 0.1 M and 0.3 M NaCl occur between 0.0 M and 2.0 M urea while the midpoint for 1.0 M NaCl occurs at approximately 4.0 M urea. This indicates that SAMP1 is more stable at higher salt concentrations. Low-salt conditions (0.1 M NaCl) show a roll-over in the unfolding branch of the chevrons. This may be present at higher salt concentrations however data can only be obtained within the solubility range of urea.

To gain insight into the relative changes in solvent-accessible surface area, β_T was calculated for the rate-limiting transition state and the intermediate by normalizing the cumulative kinetic *m*-values with respect to the equilibrium *m*-value. In the absence of denaturant, the crossing of the first transition state (TS1) forms the intermediate (M), this is the rate-limiting step during refolding. The steep change in the β_T value from the unfolded state to

TS1 indicates that early folding steps include chain compaction followed by structural optimization (fine-tuning).

Relative and absolute contact order were calculated and can be found in Table 4. Both are correlated with folding rates, however relative contact order is normalized to the protein's sequence length [16, 220]. It is expected that proteins with small contact orders will fold quicker due to increased local over non-local interactions [16]. SAMP1 has the smallest relative contact order and the highest absolute contact order, which is not normalized by sequence length. However, it has the slowest folding rate in comparison to GB1, protein L, and ubiquitin (Table 4). Through this study, we can also ascertain the optimal rate of folding for this family of proteins determined by chain topology and contact order.

Structural studies using several proteins from halophilic organisms have been studied at both high- and low-ionic strength [201-205]. These studies are in agreement with our observations that SAMP1 exhibits increased structural stability at high- versus low-ionic strength.

More specifically, Muller-Santos *et al*. found, using CD, that an esterase from *Haloarcula marismortui* was completely unfolded in a salt-free medium. Using pH end point titration, they also determined that the enzyme had no activity. Upon increasing the NaCl concentration to 2.0 M they observed an increase in helical structure and an increase in specific activity, indicative of a folded protein structure [201]. Miyashita *et al.* found using CD that dihydrofolate reductase from *Haloarcula japonica* is only partially structured in the absence of salt but increasing the concentration to 0.5 M induced significant structural formation [202]. Additionally, this protein was stabilized for thermal and urea-induced unfolding. Ishibashi *et al*. found that nucleoside diphosphate kinase from *Halobacterium salinarum* contained more

secondary structure in 3.8 M salt versus 0.2 M and that increasing the salt concentration from 0.2 M to 3.8 M progressively stabilizes the protein [203]. Additionally, the melting temperature of the protein is reduced by 30 degrees at 0.2 M salt vs 3.8 M. You *et al*. found using CD that RNase H1 from *Halobacterium sp*. NRC-1 requires at least 2.0 M salt for folding and that at low-salt concentrations the protein is only partially folded [204]. Additionally, they found that increasing the salt concentration from 0.0 M to 3.0 M raises the fraction of protein in the native state from 0 to 100 percent. Pundak and Eisenberg found using CD that malate dehydrogenase from a halophilic bacterium found in the Dead Sea begin to lose ellipticity at NaCl concentrations less than 1.0 M and below 0.5 M complete distortion of ellipticity occurred [205]. They also measured enzyme activity and found that once the NaCl concentration is below 0.5 M all activity is lost. The results of these structural studies with these proteins are in agreement with our findings which suggest that SAMP1 is more structured at 965 mM NaCl than it is at 62.5 mM.

Bandyopadhyay and Krishnamoorthy studied the kinetics of the salt-dependent unfolding of the 2Fe-2S ferredoxin from *Halobacterium salinarum* using stopped-flow [222]. They concluded high salt confers stability of the native state against urea denaturation. They also concluded that unfolding in low salt appears to be a two-phase process with an intermediate. In our studies at 100 mM NaCl we see evidence of an intermediate at low salt where the protein is unstable (Figure 42(a)). With respect to folding kinetics, it appears there is only one comprehensive folding study on a halophilic protein, dihydrofolate reductase from *Haloferax volcanii.* Gloss *et al*. used manual mixing kinetics, stopped-flow, and 8-anilinonaphthalene-1-sulfonic acid fluorescence to characterize the behavior of this protein [223]. They found that dihydrofolate reductase folding proceeds through three kinetic phases as monitored by Trp

fluorescence: a burst-phase and a fast phase detectable by stopped-flow and a slow phase requiring manual mixing. The results for SAMP1 also show three kinetic phases (Figure 43).

SUMMARY

The results of this research investigation provide an opportunity to examine the nature of the folding behavior of proteins from halophiles and supports the notion that proteins adapted and evolved to fold rapidly and correctly in a high saline environment. Thus, the observations revealing that the folding rates increase in high salt are reasonable. One finding of particular interest is based on an analysis of simulated unfolding trajectories using molecular dynamics, which revealed that 58 salt bridges are transiently present during the unfolding process of all three SAMP1 simulations whereas only four are present in the native state (Table 5). Therefore, salt bridges may play a more important role in protein dynamics than previously understood.

CHAPTER V

CONCLUSIONS AND FUTURE WORK

CONCLUSIONS

In the first aim, using bioinformatics approaches, we investigated which residues may be key determinants of this fold. We identified nine conserved amino acids based on the analysis of a structure-based sequence alignment. The conservation analysis considered amino acid identity, character, and side chain orientation. We propose that these conserved residues are important for forming and stabilizing the fold. The nine conserved residues form a predominantly hydrophobic nucleus within the core of GB1. A network analysis of all the long-range interactions in the structure of GB1 in concert with a BC analysis revealed the relative significance of each conserved amino acid residue based on the number and location of the interactions. Interestingly, the four residues which exhibited the greatest BC are conserved. This therefore shows correlation between the proposal that residues with high BC govern the network and conserved residues govern the formation of the network. These conserved residues with high BC are located on the central two strands of the four-stranded β-sheet and act as topological buttresses for the overall structure. This bioinformatics analysis provides an important foundation for the design and interpretation of both computational and experimental work for proteins in the β-grasp superfamily which may be helpful in solving the protein folding problem.

In the second aim, two proteins within the β-grasp superfamily, GB1 and SAMP1, were investigated to elucidate the key determinants of structural stability at the level of individual long-range interactions. This type of interaction is the focus of the study because it is fundamental to tertiary structure and the least understood. What we find most interesting about the β-grasp fold is that it is symmetrical. The core structure is composed of two β-hairpins which

form a β-sheet flanked by a central α-helix. The proteins were subjected to high temperature molecular dynamics simulations and the detailed behavior of each native long-range interaction was characterized. The results revealed that in GB1 the most stable region was the C-terminal hairpin and in SAMP1 it was the opposite, the N-terminal hairpin. Experimental results for GB1 support this finding. It appears that the difference in the location and number of hydrophobic interactions dictate the differential stability which is accommodated due to the structural symmetry of the β-grasp fold. Thus, the hairpins are interchangeable and in nature this lends itself to adaptability and flexibility when selective pressures occur.

In the third aim, the folding behavior of SAMP1, which is a halophile found in *Haloferax volcanii*, from the Dead Sea was investigated. To gain insight into the effects of salt at low and high concentration near the saturation point, experimental protein folding studies were conducted. The results revealed that SAMP1 folds more rapidly at high- versus low-ionic strength. Further, studies conducted at high ionic strength provided insight into the folding behavior near the solubility limit of salt at 25 °C. Thus, these results clearly indicate that adaption at high salt produces rapid and less-frustrated folding. The results of these studies help to experimentally establish the folding and unfolding behavior of SAMP1 and help lay the foundation of future, more detailed, experimental studies.

The results of these research aims provide insight into determinants of the highly populated β-grasp fold and folding and unfolding behavior of two key members. Perhaps the most surprising finding is the presence of a significant number of non-native long-range interactions during unfolding which has largely gone unnoticed in the scientific community since the study of protein kinetics and thermodynamics at atomic resolution began. These findings

together provide a solid foundation for advancement of the protein folding question and structure prediction.

FUTURE WORK

To ascertain a more complete picture of the underlying mechanisms and forces guiding the folding/unfolding process, folding simulations of GB1 should be analyzed in an identical manner as the unfolding simulations discussed in chapter three. Three simulations were donated by the Shaw research group to the Greene research group and were performed using the Anton supercomputer. They will complement the unfolding simulations to give a unified picture of the folding and unfolding behavior of GB1 as well as allow the role of the conserved residues in all-atom folding simulations to be characterized.

The transient salt bridges found in our unfolding simulations should be analyzed in greater detail in future work. They should also be analyzed in the Shaw folding simulations of GB1 to determine what role they may play in the folding process. These mercurial salt bridges were observed for both GB1 and SAMP1 in their respective unfolding simulations. The persistence of transient salt bridges during the unfolding of GB1 is shown in Figure 44.

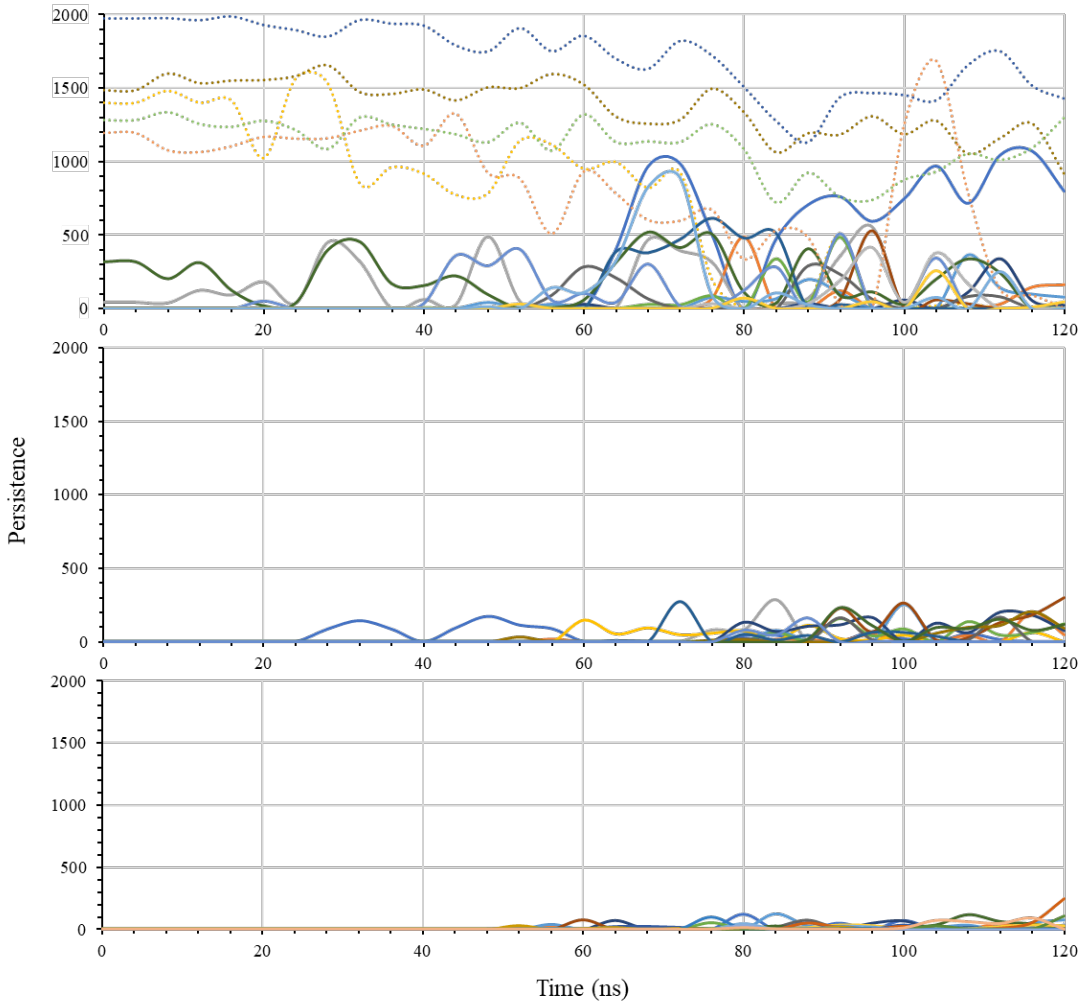


Figure 44. Persistence of transient salt bridges in GB1. The top, middle, and bottom graphs seem to be the most to least persistent as grouped by visual analysis. Dotted lines in the top graph indicate native salt bridges and those found in loops. The colors are arbitrary and were used for distinguishability. Data plotted using Microsoft Excel 365.

Specifically, in GB1, there exists no native salt bridge between Lys13 and Glu56 (Figure 45(a)). During the unfolding simulation these residues remain at a distance (Figure 45(b)) until the end of the simulation, where they form a salt bridge (Figure 45(c)). In SAMP1, no native salt

bridge is present between Arg19 and Glu77 (Figure 45(d)). During the unfolding simulation they form a temporary salt bridge (Figure 45(e)) and then move apart (Figure 45(f)) as the protein continues to unfold. Investigation into the persistence of transient salt bridges may provide key insights into protein folding that have largely gone unnoticed.

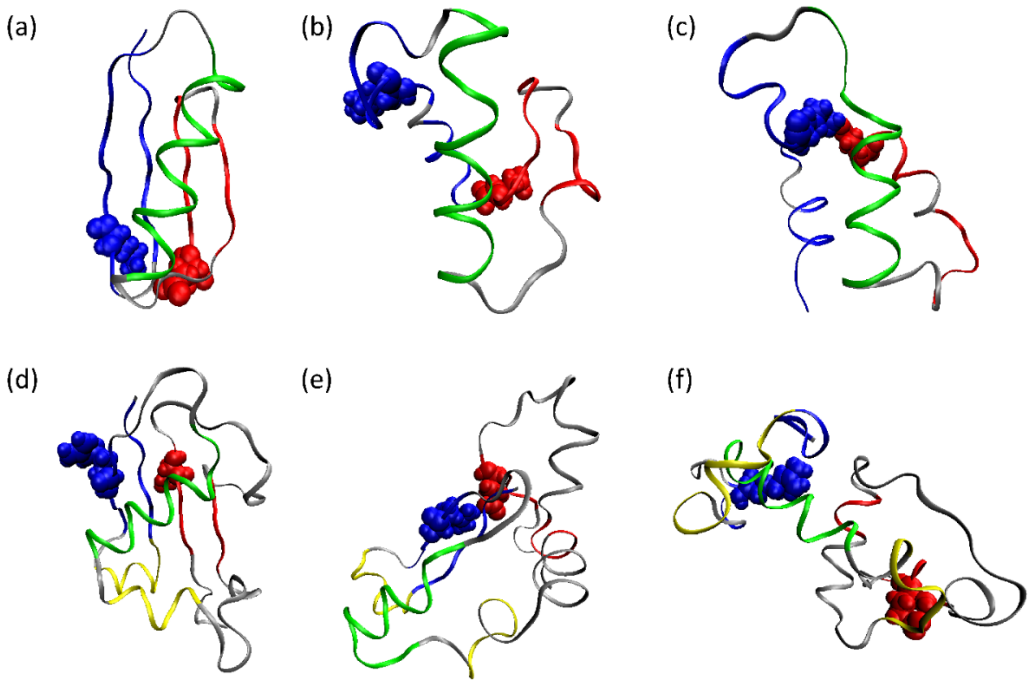


Figure 45. Transient salt bridges in (a-c) GB1 and (d-f) SAMP1. The transient salt bridge between Lys13 (blue spheres) and Glu56 (red spheres) in GB1 at (a) 0 ns, (b) 68 ns, and (c) 112 ns during the first unfolding simulation. The transient salt bridge between Arg19 (blue spheres) and Glu77 (red spheres) in SAMP1 at (d) 0 ns, (e) 58 ns, and (f) 100 ns during the first unfolding simulation. Residues comprising the N- and C-terminal hairpins, central α-helix, and loops are shown in blue, red, green, and gray, respectively. Secondary structure embellishments are shown in yellow. Structures visualized using VMD 1.9.1.

Transient hydrophobic interactions were also observed during the unfolding simulations for both GB1 (Figure 46(a-c)) and SAMP1 (Figure 46(d-e)) in our unfolding simulations. These temporary interactions may play a role in the folding of the protein and thus future work should include the development of a program that is able to calculate and track all transient hydrophobic interactions over the entire protein structure during the course of an unfolding or folding simulation.

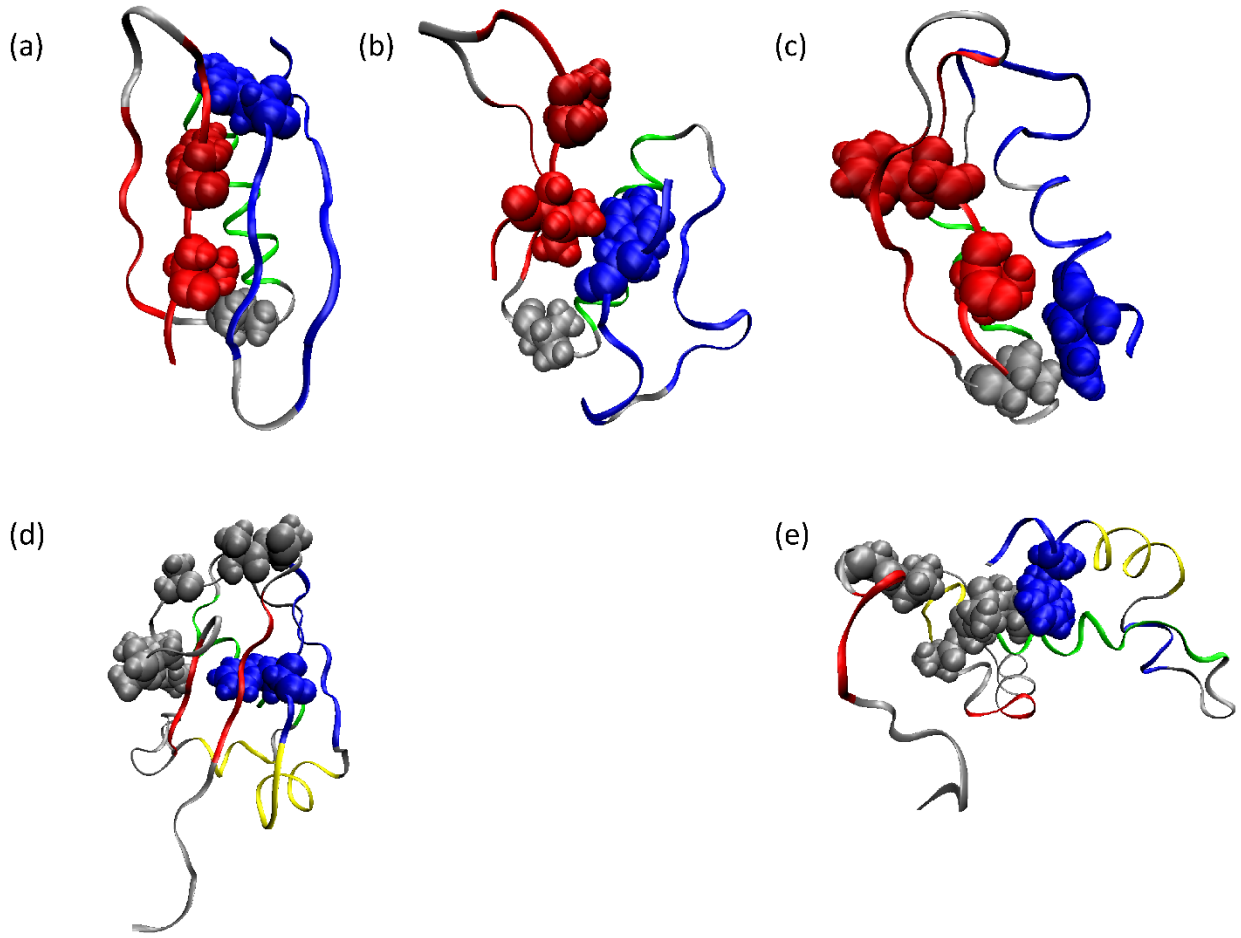


Figure 46. Residues comprising transient hydrophobic interactions in (a-c) GB1 and (d-e) SAMP1. In GB1, residues involved in hydrophobic interactions are Tyr3 (blue spheres), Phe52 (dark red spheres), Val54 (red spheres), and Val39 (gray spheres) and are shown at (a) 0 ns, (b) 45 ns, and (c) 55 ns during the first unfolding simulation. In SAMP1 residues involved in hydrophobic interactions are Trp3 (blue spheres) and cluster of hydrophobic loop residues comprised of Ala65, Ala66, Ala67, Leu 68, Ala71, Ala73, and Ala74 (gray spheres) and are shown at (d) 0 ns and (e) 85 ns during the first unfolding simulation. Residues comprising the N- and C-terminal hairpins, central α-helix, and loops are shown in blue, red, green, and gray, respectively. Secondary structure embellishments are shown in yellow. Structures visualized using VMD 1.9.1.

To capture some of these intermediate states and characterize the transient interactions in the future, freeze folding in combination with solid state NMR can be used. In the case of SAMP1, folding via jumping the salt concentration is possible due to its increased structural stability in high ionic concentrations. Recognition of these non-native interactions by the scientific community is less well described.

Work on SAMP1, both experimental and computational, is minimal compared to other proteins in the β-grasp superfamily due to its relevant recent discovery. Future work with SAMP1 should include folding simulations on the Anton supercomputer. These simulations would prove priceless in helping to ascertain the determinants of folding in SAMP1, but they would also allow for the comparison to the folding simulations by the Shaw research group and to our MD unfolding simulations and experimental work. An extensive mutagenesis study and subsequent Φ-value analysis of SAMP1 would help to elucidate key residues responsible for the folding of SAMP1 and give insight into its structure in the transition state.

REFERENCES

1. Gregory, S.G., Barlow, K.F., McLay, K.E*., et al.* (2006). The DNA sequence and biological annotation of human chromosome 1. *Nature*. **441**, 315-321.
2. Collins, F.S., Lander, E S, Rogers, J, Waterston, R H, Conso, IHGS. (2004). Finishing the euchromatic sequence of the human genome. *Nature*. **431**, 931-945.
3. Schmutz, J., Wheeler, J., Grimwood, J*., et al.* (2004). Quality assessment of the human genome sequence. *Nature*. **429**, 365-368.
4. Dunham, I., Hunt, A.R., Collins, J.E*., et al.* (1999). The DNA sequence of human chromosome 22. *Nature*. **402**, 489-495.
5. Heim, M., Römer, L. and Scheibel, T. (2010). Hierarchical structures made of proteins. The complex architecture of spider webs and their constituent silk proteins. *Chem Soc Rev*. **39**, 156-164.
6. Abeln, S. and Deane, C.M. (2005). Fold usage on genomes and protein fold evolution. *Proteins: Struct Funct Bioinf*. **60**, 690-700.
7. Salem, G.M., Hutchinson, E.G., Orengo, C.A*., et al.* (1999). Correlation of observed fold frequency with the occurrence of local structural motifs. *J Mol Biol*. **287**, 969-981.
8. Makhatadze, G. (2017). Linking computation and experiments to study the role of charge–charge interactions in protein folding and stability. *Phys Biol*. **14**, 013002.
9. Durell, S.R. and Ben-Naim, A. (2017). Hydrophobic-hydrophilic forces in protein folding. *Biopolymers*. **107**, e23020.
10. Ben-Naim, A. (2011). The rise and fall of the hydrophobic effect in protein folding and protein-protein association, and molecular recognition. *Open J Biophys*. **1**, 1-7.
11. Nick Pace, C., Scholtz, J.M. and Grimsley, G.R. (2014). Forces stabilizing proteins. *FEBS Lett*. **588**, 2177-2184.
12. Samuel, S. and Gromiha, M. (2003). Role of hydrophobic clusters and long-range contact networks in the folding of (α/β)8 barrel proteins. *Biophys J*. **84**, 1919-1925.
13. Gromiha, M.M. and Selvaraj, S. (1999). Importance of long-range interactions in protein folding. *Biophys Chem*. **77**, 49-68.
14. Sengupta, D. and Kundu, S. (2012). Role of long- and short-range hydrophobic, hydrophilic and charged residues contact network in protein's structural organization. *BMC Bioinformatics*. **13**, 142.
15. Cregut, D., Civera, C., Macias, M.J*., et al.* (1999). A tale of two secondary structure elements: when a β-hairpin becomes an α-helix. *J Mol Biol*. **292**, 389-401.
16. Plaxco, K.W., Simons, K.T. and Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*. **277**, 985-994.
17. Ouyang, Z. and Liang, J. (2008). Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Sci*. **17**, 1256-1263.
18. Gromiha, M.M. and Selvaraj, S. (2001). Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J Mol Biol*. **310**, 27-32.
19. Istomin, A.Y., Jacobs, D.J. and Livesay, D.R. (2007). On the role of structural class of a protein with two-state folding kinetics in determining correlations between its size, topology, and folding rate. *Protein Sci*. **16**, 2564-2569.
20. Harihar, B. and Selvaraj, S. (2009). Refinement of the long-range order parameter in predicting folding rates of two-state proteins. *Biopolymers*. **91**, 928-935.

21. Zwanzig, R., Szabo, A. and Bagchi, B. (1992). Levinthal's paradox. *Proc Natl Acad Sci*. **89**, 20-22.
22. Daggett, V. and Fersht, A.R. (2003). Is there a unifying mechanism for protein folding? *Trends Biochem Sci*. **28**, 18-25.
23. Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science*. **181**, 223-230.
24. Anfinsen, C.B. and Scheraga, H.A. (1975). Experimental and theoretical aspects of protein folding. *Adv Protein Chem*. **29**, 205-300.
25. Radford, S.E. (2000). Protein folding: progress made and promises ahead. *Trends Biochem Sci*. **25**, 611-618.
26. Fersht, A.R. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc Natl Acad Sci U S A*. **92**, 10869-10873.
27. Itzhaki, L.S., Otzen, D.E. and Fersht, A.R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J Mol Biol*. **254**, 260-288.
28. Fersht, A.R., Matouschek, A. and Serrano, L. (1992). The folding of an enzyme: I. Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol*. **224**, 771-782.
29. Fersht, A.R. and Daggett, V. (2002). Protein folding and unfolding at atomic resolution. *Cell*. **108**, 573-582.
30. Shoemaker, B.A. and Wolynes, P.G. (1999). Exploring structures in protein folding funnels with free energy functionals: the denatured ensemble. *J Mol Biol*. **287**, 657-674.
31. Shoemaker, B.A., Wang, J. and Wolynes, P.G. (1999). Exploring structures in protein folding funnels with free energy functionals: the transition state ensemble. *J Mol Biol*. **287**, 675-694.
32. Brooks, C.L., Gruebele, M., Onuchic, J.N.*, et al.* (1998). Chemical physics of protein folding. *Proc Natl Acad Sci*. **95**, 11037-11038.
33. Karplus, M. and Weaver, D.L. (1994). Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci*. **3**, 650-668.
34. Bryngelson, J.D., Onuchic, J.N., Socci, N.D.*, et al.* (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*. **21**, 167-195.
35. Wolynes, P.G. (1997). Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc Natl Acad Sci*. **94**, 6170-6175.
36. Finkelstein, A.V. and Badretdinov, A.Y. (1997). Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Fold Des*. **2**, 115-121.
37. Finkelstein, A.V. (1991). Rate of β-structure formation in polypeptides. *Proteins*. **9**, 23-27.
38. Wolynes, P.G. (1996). Symmetry and the energy landscapes of biomolecules. *Proc Natl Acad Sci*. **93**, 14249-14255.
39. Klimov, D.K. and Thirumalai, D. (1996). Factors governing the foldability of proteins. *Proteins*. **26**, 411-441.
40. Abkevich, V.I., Gutin, A.M. and Shakhnovich, E.I. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J Mol Biol*. **252**, 460-471.
41. Gutin, A.M., Abkevich, V.I. and Shakhnovich, E.I. (1996). Chain length scaling of protein folding time. *Phys Rev Lett*. **77**, 5433-5436.
42. S˘ali, A., Shakhnovich, E. and Karplus, M. (1994). How does a protein fold? *Nature*. **369**, 248-251.
43. Onuchic, J.N., Wolynes, P.G., Luthey-Schulten, Z.*, et al.* (1995). Toward an outline of the topography of a realistic protein-folding funnel. *Proc Natl Acad Sci*. **92**, 3626-3630.

44. Pande, V.S., Grosberg, A.Y. and Tanaka, T. (1997). On the theory of folding kinetics for short proteins. *Fold Des*. **2**, 109-114.
45. Doyle, R., Simons, K., Qian, H.*, et al.* (1997). Local interactions and the optimization of protein folding. *Proteins: Struct Funct Bioinf*. **29**, 282-291.
46. Gross, M. (1996). Linguistic analysis of protein folding. *FEBS Lett*. **390**, 249-252.
47. Unger, R. and Moult, J. (1996). Local interactions dominate folding in a simple protein model. *J Mol Biol*. **259**, 988-994.
48. Fersht, A.R., Dobson, C.M. and Fersht, A.R. (1995). Mapping the structures of transition states and intermediates in folding: delineation of pathways at high resolution. *Philos Trans R Soc Lond B Biol Sci*. **348**, 11-15.
49. Govindarajan, S. and Goldstein, R.A. (1995). Optimal local propensities for model proteins. *Proteins: Struct Funct Bioinf*. **22**, 413-418.
50. Orengo, C.A., Jones, D.T. and Thornton, J.M. (1994). Protein superfamilles and domain superfolds. *Nature*. **372**, 631-634.
51. Dill, K.A., Fiebig, K.M. and Chan, H.S. (1993). Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci*. **90**, 1942-1946.
52. Nelson, D.L., Lehninger, A. L., Cox, M. M. (2008). Lehninger principles of biochemistry. Macmillan.
53. Thommen, M., Holtkamp, W. and Rodnina, M.V. (2017). Co-translational protein folding: progress and methods. *Curr Opin Struct Biol*. **42**, 83-89.
54. Guinn, E.J., Tian, P., Shin, M.*, et al.* (2018). A small single-domain protein folds through the same pathway on and off the ribosome. *Proc Natl Acad Sci*. **115**, 12206-12211.
55. Tian, P., Steward, A., Kudva, R.*, et al.* (2018). Folding pathway of an Ig domain is conserved on and off the ribosome. *Proc Natl Acad Sci*. **115**, E11284-E11293.
56. Nilsson, O.B., Hedman, R., Marino, J.*, et al.* (2015). Cotranslational protein folding inside the ribosome exit tunnel. *Cell reports*. **12**, 1533-1540.
57. Holtkamp, W., Kokic, G., Jäger, M.*, et al.* (2015). Cotranslational protein folding on the ribosome monitored in real time. *Science*. **350**, 1104-1107.
58. Waudby, C.A., Dobson, C.M. and Christodoulou, J. (2019). Nature and regulation of protein folding on the ribosome. *Trends Biochem Sci*. **44**, 914-926.
59. Motojima, F. (2015). How do chaperonins fold protein? *Biophysics (Nagoya-shi, Japan)*. **11**, 93-102.
60. Burroughs, A.M., Balaji, S., Iyer, L.M.*, et al.* (2007). Small but versatile: the extraordinary functional and structural diversity of the β-grasp fold. *Biol direct*. **2**, 18.
61. Berkovich, R., Mondal, J., Paster, I.*, et al.* (2017). Simulated force quench dynamics shows GB1 protein is not a two state folder. *J Phys Chem B*. **121**, 5162-5173.
62. Best, R.B. and Mittal, J. (2011). Free-energy landscape of the GB1 hairpin in all-atom explicit solvent simulations with different force fields: Similarities and differences. *Proteins*. **79**, 1318-1328.
63. Collins, J.C., Bedford, J.T. and Greene, L.H. (2016). Elucidating the key determinants of structure, folding, and stability for the (4β+α) conformation of the B1 domain of protein G using bioinformatics approaches. *IEEE Trans Nanobioscience*. **15**, 140-147.
64. Cheng, Q., Joung, I., Lee, J.*, et al.* (2019). Exploring the folding mechanism of small proteins GB1 and LB1. *J Chem Theory Comput*. **15**, 3432-3449.
65. Bonomi, M., Branduardi, D., Gervasio, F.L.*, et al.* (2008). The unfolded ensemble and folding mechanism of the C-terminal GB1 β-hairpin. *J Am Chem Soc*. **130**, 13938-13944.

66. De Sancho, D., Mittal, J. and Best, R.B. (2013). Folding kinetics and unfolded state dynamics of the GB1 hairpin from molecular simulation. *J Chem Theory Comput*. **9**, 1743-1753.
67. Sheinerman, F.B. and Brooks, C.L., 3rd. (1998). Molecular picture of folding of a small α/β protein. *Proc Natl Acad Sci U S A*. **95**, 1562-1567.
68. Zagrovic, B., Sorin, E.J. and Pande, V. (2001). β-hairpin folding simulations in atomistic detail using an implicit solvent model. *J Mol Biol*. **313**, 151-169.
69. Kmiecik, S. and Kolinski, A. (2011). Simulation of chaperonin effect on protein folding: a shift from nucleation-condensation to framework mechanism. *J Am Chem Soc*. **133**, 10283-10289.
70. Kmiecik, S. and Kolinski, A. (2008). Folding pathway of the B1 domain of protein G explored by multiscale modeling. *Biophys J*. **94**, 726-736.
71. Ahalawat, N. and Mondal, J. (2018). Assessment and optimization of collective variables for protein conformational landscape: GB1 β-hairpin as a case study. *J Chem Phys*. **149**, 094101.
72. Brooks, C.L. (2002). Protein and peptide folding explored with molecular simulations. *Acc Chem Res*. **35**, 447-454.
73. Bu, T., Wang, H.C. and Li, H. (2012). Single molecule force spectroscopy reveals critical roles of hydrophobic core packing in determining the mechanical stability of protein GB1. *Langmuir*. **28**, 12319-12325.
74. Lapidus, L.J., Acharya, S., Schwantes, C.R.*, et al.* (2014). Complex pathways in folding of protein G explored by simulation and experiment. *Biophys J*. **107**, 947-955.
75. Lapidus, L.J., Yao, S., McGarrity, K.S.*, et al.* (2007). Protein hydrophobic collapse and early folding steps observed in a microfluidic mixer. *Biophys J*. **93**, 218-224.
76. Morrone, A., Giri, R., Toofanny, R.D.*, et al.* (2011). GB1 is not a two-state folder: identification and characterization of an on-pathway intermediate. *Biophys J*. **101**, 2053-2060.
77. Park, S.-H., Shastry, M.C.R. and Roder, H. (1999). Folding dynamics of the B1 domain of protein G explored by ultrarapid mixing. *Nat Struct Biol*. **6**, 943-947.
78. Shen, T., Cao, Y., Zhuang, S.*, et al.* (2012). Engineered bi-histidine metal chelation sites map the structure of the mechanical unfolding transition state of an elastomeric protein domain GB1. *Biophys J*. **103**, 807-816.
79. Campos-Olivas, R., Aziz, R., Helms, G.L.*, et al.* (2002). Placement of 19F into the center of GB1: effects on structure and stability. *FEBS Lett*. **517**, 55-60.
80. Ding, K., Louis, J.M. and Gronenborn, A.M. (2004). Insights into conformation and dynamics of protein GB1 during folding and unfolding by NMR. *J Mol Biol*. **335**, 1299-1307.
81. McCallister, E.L., Alm, E. and Baker, D. (2000). Critical role of β-hairpin formation in protein G folding. *Nat Struct Biol*. **7**, 669-673.
82. Nauli, S., Kuhlman, B., Le Trong, I.*, et al.* (2002). Crystal structures and increased stabilization of the protein G variants with switched folding pathways NuG1 and NuG2. *Protein Sci*. **11**, 2924-2931.
83. Alexander, P.A., He, Y., Chen, Y.*, et al.* (2009). A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci U S A*. **106**, 21149-21154.
84. Alexander, P., Orban, J. and Bryan, P. (1992). Kinetic analysis of folding and unfolding the 56 amino acid IgG-binding domain of streptococcal protein G. *Biochemistry*. **31**, 7243-7248.
85. Adhikari, A.N., Freed, K.F. and Sosnick, T.R. (2013). Simplified protein models: predicting folding pathways and structure using amino acid sequences. *Phys Rev Lett*. **111**, 028103.
86. Kouza, M. and Hansmann, U.H.E. (2012). Folding simulations of the A and B domains of protein G. *J Phys Chem B*. **116**, 6645-6653.

87. Roder, H. and Colón, W. (1997). Kinetic role of early intermediates in protein folding. *Curr Opin Struct Biol*. **7**, 15-28.
88. Kuszewski, J., Clore, G.M. and Gronenborn, A.M. (1994). Fast folding of a prototypic polypeptide: The immunoglobulin binding domain of streptococcal protein G. *Protein Sci*. **3**, 1945-1952.
89. Khare, D., Alexander, P. and Orban, J. (1999). Hydrogen bonding and equilibrium protium−deuterium fractionation factors in the immunoglobulin G binding domain of protein G. *Biochemistry*. **38**, 3918-3925.
90. Nauli, S., Kuhlman, B. and Baker, D. (2001). Computer-based redesign of a protein folding pathway. *Nat Struct Biol*. **8**, 602-605.
91. Frank, M.K., Dyda, F., Dobrodumov, A.*, et al.* (2002). Core mutations switch monomeric protein GB1 into an intertwined tetramer. *Nat Struct Biol*. **9**, 877-885.
92. Jee, J., Byeon, I.J., Louis, J.M.*, et al.* (2008). The point mutation A34F causes dimerization of GB1. *Proteins*. **71**, 1420-1431.
93. Bauer, M.C., Xue, W.F. and Linse, S. (2009). Protein GB1 folding and assembly from structural elements. *Int J Mol Sci*. **10**, 1552-1566.
94. Baxa, M.C., Yu, W., Adhikari, A.N.*, et al.* (2015). Even with nonnative interactions, the updated folding transition states of the homologs Proteins G & L are extensive and similar. *Proc Natl Acad Sci U S A*. **112**, 8302-8307.
95. Campion, S.R. (2016). Conserved aromatic residues as determinants in the folding and assembly of immunoglobulin variable domains. *Mol Immunol*. **70**, 63-71.
96. Engman, K.C., Sandberg, A., Leckner, J.*, et al.* (2004). Probing the influence on folding behavior of structurally conserved core residues in P. aeruginosa apo-azurin. *Protein Sci*. **13**, 2706-2715.
97. Greene, L.H., Hamada, D., Eyles, S.J.*, et al.* (2003). Conserved signature proposed for folding in the lipocalin superfamily. *FEBS Lett*. **553**, 39-44.
98. Gunasekaran, K., Eyles, S.J., Hagler, A.T.*, et al.* (2001). Keeping it in the family: folding studies of related proteins. *Curr Opin Struct Biol*. **11**, 83-93.
99. Haliloglu, T., Keskin, O., Ma, B.*, et al.* (2005). How similar are protein folding and protein binding nuclei? Examination of vibrational motions of energy hot spots and conserved residues. *Biophys J*. **88**, 1552-1559.
100. Hamill, S.J., Steward, A. and Clarke, J. (2000). The folding of an immunoglobulin-like greek key protein is defined by a common-core nucleus and regions constrained by topology. *J Mol Biol*. **297**, 165-178.
101. Heidary, D.K. and Jennings, P.A. (2002). Three topologically equivalent core residues affect the transition state ensemble in a protein folding reaction. *J Mol Biol*. **316**, 789-798.
102. Kragelund, B.B., Osmark, P., Neergaard, T.B.*, et al.* (1999). The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nat Struct Biol*. **6**, 594-601.
103. Larson, S.M., Ruczinski, I., Davidson, A.R.*, et al.* (2002). Residues participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. *J Mol Biol*. **316**, 225-233.
104. Li, H., Wojtaszek, J.L. and Greene, L.H. (2009). Analysis of conservation in the Fas-associated death domain protein and the importance of conserved tryptophans in structure, stability and folding. *Biochim Biophys Acta Proteins Proteom*. **1794**, 583-593.

105. Tseng, Y.Y. and Liang, J. (2004). Are residues in a protein folding nucleus evolutionarily conserved? *J Mol Biol*. **335**, 869-880.
106. Zarrine-Afsar, A., Larson, S.M. and Davidson, A.R. (2005). The family feud: do proteins with similar structures fold via the same pathway? *Curr Opin Struct Biol*. **15**, 42-49.
107. Greene, L.H. (2012). Protein structure networks. *Brief Funct Genomics*. **11**, 469-478.
108. Greene, L.H. and Higman, V.A. (2003). Uncovering network systems within protein structures. *J Mol Biol*. **334**, 781-791.
109. Higman, V.A. and Greene, L.H. (2006). Elucidation of conserved long-range interaction networks in proteins and their significance in determining protein topology. *Physica A*. **368**, 595-606.
110. Haratipour, Z., Aldabagh, H., Li, Y.*, et al.* (2019). Network connectivity, centrality and fragmentation in the greek-key protein topology. *Protein J*. **38**, 497-505.
111. Cao, A. (2020). The last secret of protein folding: The real relationship between long-range interactions and local structures. *Protein J*. **39**, 422-433.
112. Chen, J., Liu, X. and Chen, J. (2018). Atomistic peptide folding simulations reveal interplay of entropy and long-range interactions in folding cooperativity. *Sci Rep*. **8**, 13668.
113. Go, N. and Taketomi, H. (1978). Respective roles of short- and long-range interactions in protein folding. *Proc Natl Acad Sci U S A*. **75**, 559-563.
114. Greene, L.H. and Grant, T.M. (2012). Protein folding by 'levels of separation': A hypothesis. *FEBS Letters*. **586**, 962-966.
115. Harihar, B. and Selvaraj, S. (2011). Analysis of rate-limiting long-range contacts in the folding rate of three-state and two-state proteins. *Protein Pept Lett*. **18**, 1042-1052.
116. Melkikh, A.V. and Meijer, D.K.F. (2018). On a generalized Levinthal's paradox: The role of long- and short range interactions in complex bio-molecular reactions, including protein and DNA folding. *Prog Biophys Mol Biol*. **132**, 57-79.
117. Miyazawa, S. and Jernigan, R.L. (2003). Long- and short-range interactions in native protein structures are consistent/minimally frustrated in sequence space. *Proteins*. **50**, 35-43.
118. Altschul, S.F., Madden, T.L., Schäffer, A.A.*, et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**, 3389-3402.
119. Holm, L. and Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics (Oxford, England)*. **16**, 566-567.
120. Madeira, F., Park, Y.M., Lee, J.*, et al.* (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res*. **47**, W636-W641.
121. Barzel, B. and Barabási, A.-L. (2013). Universality in network dynamics. *Nat Phys*. **9**, 673-681.
122. Lü, L., Pan, L., Zhou, T.*, et al.* (2015). Toward link predictability of complex networks. *Proc Natl Acad Sci*. **112**, 2325-2330.
123. Barabási, A.-L. (2016). Network Science. Cambridge University Press.
124. Di Paola, L. and Giuliani, A. (2015). Protein contact network topology: a natural language for allostery. *Curr Opin Struct Biol*. **31**, 43-48.
125. Vishveshwara, S., Ghosh, A. and Hansia, P. (2009). Intra- and inter-molecular communications through protein structure network. *Curr Protein Pept Sci*. **10**, 146-160.
126. Costanzi, S. (2016). Topological analyses of protein-ligand binding: a network approach. *Curr Protein Pept Sci*. **17**, 37-40.
127. Bhattacharyya, M., Ghosh, S. and Vishveshwara, S. (2016). Protein structure and function: Looking through the network of side-chain interactions. *Curr Protein Pept Sci*. **17**, 4-25.

128. Gromiha, M.M. (2009). Multiple contact network is a key determinant to protein folding rates. *J Chem Inf Model*. **49**, 1130-1135.
129. Dokholyan, N.V., Li, L., Ding, F.*, et al.* (2002). Topological determinants of protein folding. *Proc Natl Acad Sci*. **99**, 8637-8641.
130. Böde, C., Kovács, I.A., Szalay, M.S.*, et al.* (2007). Network analysis of protein dynamics. *FEBS Lett*. **581**, 2776-2782.
131. Vendruscolo, M., Dokholyan, N.V., Paci, E.*, et al.* (2002). Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E*. **65**, 061910.
132. Hospital, A., Goñi, J.R., Orozco, M.*, et al.* (2015). Molecular dynamics simulations: advances and applications. *Adv Appl Bioinform Chem*. **8**, 37-47.
133. Lindorff-Larsen, K., Piana, S., Palmo, K.*, et al.* (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*. **78**, 1950-1958.
134. Best, R.B. and Hummer, G. (2009). Optimized molecular dynamics force fields applied to the helix−coil transition of polypeptides. *J Phys Chem B*. **113**, 9004-9015.
135. Best, R.B., Zhu, X., Shim, J.*, et al.* (2012). Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *J Chem Theory Comput*. **8**, 3257-3273.
136. Li, D.-W. and Brüschweiler, R. (2010). NMR-based protein potentials. *Angew Chem Int Ed*. **49**, 6778-6780.
137. Piana, S., Lindorff-Larsen, K. and Shaw, D.E. (2011). How robust are protein folding simulations with respect to force field parameterization? *Biophys J*. **100**, L47-L49.
138. Braun, A.R., Sachs, J.N. and Nagle, J.F. (2013). Comparing simulations of lipid bilayers to scattering data: The GROMOS 43A1-S3 force field. *J Phys Chem B*. **117**, 5065-5072.
139. Mayne, C.G., Saam, J., Schulten, K.*, et al.* (2013). Rapid parameterization of small molecules using the force field toolkit. *J Comput Chem*. **34**, 2757-2770.
140. Anwar, J. and Zahn, D. (2017). Polymorphic phase transitions: Macroscopic theory and molecular simulation. *Adv Drug Delivery Rev*. **117**, 47-70.
141. Shaw, D.E., Deneroff, M.M., Dror, R.O.*, et al.* (2008). Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM*. **51**, 91–97.
142. Shaw, D.E., Grossman, J.P., Bank, J.A.*, et al.* (2014). Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 41-53.
143. Royer, C.A., Mann, C.J. and Matthews, C.R. (1993). Resolution of the fluorescence equilibrium unfolding profile of trp aporepressor using single tryptophan mutants. *Protein Sci*. **2**, 1844-1852.
144. Ladokhin, A.S., Jayasinghe, S. and White, S.H. (2000). How to measure and analyze tryptophan fluorescence in membranes properly, and why bother? *Anal Biochem*. **285**, 235-245.
145. Vivian, J.T. and Callis, P.R. (2001). Mechanisms of tryptophan fluorescence shifts in proteins. *Biophys J*. **80**, 2093-2109.
146. Lakowicz, J.R. (2006). Principles of fluorescence spectroscopy. Springer.
147. Kelly, S.M., Jess, T.J. and Price, N.C. (2005). How to study proteins by circular dichroism. *Biochem Biophys Acta Proteins Proteom*. **1751**, 119-139.
148. Whitmore, L. and Wallace, B.A. (2008). Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers*. **89**, 392-400.

149. Berova, N., Nakanishi, K. J., Woody, R. (2000). Circular Dichroism: principles and applications. Wiley-VCH.
150. Bagshaw, C.R. (2013). Stopped-Flow Techniques. Springer Berlin Heidelberg.
151. Kelly, S.M. and Price, N.C. (1997). The application of circular dichroism to studies of protein folding and unfolding. *Biochim Biophys Acta*. **1338**, 161-185.
152. Kelly, S.M. and Price, N.C. (2006). Circular dichroism to study protein interactions. *Curr Protoc Protein Sci*. **46**, 1-18.
153. Zeeb, M. and Balbach, J. (2004). Protein folding studied by real-time NMR spectroscopy. *Methods (San Diego, Calif.)*. **34**, 65-74.
154. Zeeb, M. and Balbach, J. (2005). Millisecond protein folding studied by NMR spectroscopy. *Protein Pept Lett*. **12**, 139-146.
155. Fabian, H. and Naumann, D. (2004). Methods to study protein folding by stopped-flow FT-IR. *Methods (San Diego, Calif.)*. **34**, 28-40.
156. Kihara, H., Semisotnov, G.V. and Kotova, N.V. (1996). Kinetic study on protein folding studied by stopped-flow X-ray scattering. Proceedings of the 10 International conference on small-angle scattering; Workshop on synchrotron radiation and neutron SAS: instrumentation and industrial applications Abstracts. 295.
157. Ptitsyn, O.B. (1995). Molten Globule and Protein Folding. Academic Press.
158. Kuwajima, K. (1989). The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. *Proteins*. **6**, 87-103.
159. Marqusee, S., Robbins, V.H. and Baldwin, R.L. (1989). Unusually stable helix formation in short alanine-based peptides. *Proc Natl Acad Sci*. **86**, 5286-5290.
160. Shastry, M.C.R., Luck, S.D. and Roder, H. (1998). A continuous-flow capillary mixing method to monitor reactions on the microsecond time scale. *Biophys J*. **74**, 2714-2721.
161. Shastry, M.C.R. and Roder, H. (1998). Evidence for barrier-limited protein folding kinetics on the microsecond time scale. *Nat Struct Biol*. **5**, 385-392.
162. Xu, M., Beresneva, O., Rosario, R.*, et al.* (2012). Microsecond folding dynamics of apomyoglobin at acidic pH. *J Phys Chem B*. **116**, 7014-7025.
163. Gierusz, L.A. and Pinheiro, T.J.T. (2013). Continuous Flow. Springer Berlin Heidelberg.
164. Schueler-Furman, O. and Baker, D. (2003). Conserved residue clustering and protein structure prediction. *Proteins*. **52**, 225-235.
165. Samatova, E.N., Melnik, B.S., Balobanov, V.A.*, et al.* (2010). Folding intermediate and folding nucleus for I-->N and U-->I-->N transitions in apomyoglobin: contributions by conserved and nonconserved residues. *Biophys J*. **98**, 1694-1702.
166. Wilson, C.J. and Wittung-Stafshede, P. (2005). Role of structural determinants in folding of the sandwich-like protein Pseudomonas aeruginosa azurin. *Proc Natl Acad Sci U S A*. **102**, 3984-3987.
167. Gromiha, M.M., Pujadas, G., Magyar, C.*, et al.* (2004). Locating the stabilizing residues in (α/β)8 barrel proteins based on hydrophobicity, long-range interactions, and sequence conservation. *Proteins*. **55**, 316-329.
168. Mirny, L.A., Abkevich, V.I. and Shakhnovich, E.I. (1998). How evolution makes proteins fold quickly. *Proc Natl Acad Sci*. **95**, 4976-4981.
169. Shi, X., Bisaria, N., Benz-Moy, T.L.*, et al.* (2014). Roles of long-range tertiary interactions in limiting dynamics of the Tetrahymena group I ribozyme. *J Am Chem Soc*. **136**, 6643-6648.
170. Greene, L.H., Higman, V. A. (2005). Conserved networks and the determinants of protein topology. *FEBS Lett*. **272**, 87.

171. Rao, F. and Caflisch, A. (2004). The protein folding network. *J Mol Biol*. **342**, 299-306.
172. Lei, H., Su, Y., Jin, L.*, et al.* (2010). Folding network of villin headpiece subdomain. *Biophys J*. **99**, 3374-3384.
173. Estrada, E. (2010). Universality in protein residue networks. *Biophys J*. **98**, 890-900.
174. Kinch, L.N. and Grishin, N.V. (2002). Evolution of protein structures and functions. *Curr Opin Struct Biol*. **12**, 400-408.
175. Alexander, R.P. and Zhulin, I.B. (2007). Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors. *Proc Natl Acad Sci*. **104**, 2885-2890.
176. Todd, A.E., Orengo, C.A. and Thornton, J.M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*. **307**, 1113-1143.
177. Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*. **9**, 56-68.
178. Kyte, J. and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. **157**, 105-132.
179. Winn, M.D., Ballard, C.C., Cowtan, K.D.*, et al.* (2011). Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr*. **67**, 235-242.
180. De Nooy, W.M., A.; Batagelj, V. (2011). Exploratory social network analysis with Pajek. Cambridge Univ. Press.
181. Dorogovtsev, S.N.M., J.F.F. (2013). Evolution of networks: from biological nets to the internet and WWW. Oxford Univ. Press.
182. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng Des Sel*. **12**, 85-94.
183. Sayle, R.A. and Milner-White, E.J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem Sci*. **20**, 374.
184. Barthelemy, M. (2004). Betweenness centrality in large complex networks. *Eur Phys J B*. **38**, 163-168.
185. Newman, M.E. (2003). A measure of betweenness centrality based on random walks. *Soc Networks*. **27**, 39-54.
186. Giri, R., Morrone, A., Travaglini-Allocatelli, C.*, et al.* (2012). Folding pathways of proteins with increasing degree of sequence identities but different structure and function. *Proc Natl Acad Sci U S A*. **109**, 17772-17776.
187. Lindorff-Larsen, K., Piana, S., Dror, R.O.*, et al.* (2011). How fast-folding proteins fold. *Science*. **334**, 517-520.
188. Park, S.H., O'Neil, K.T. and Roder, H. (1997). An early intermediate in the folding reaction of the B1 domain of protein G contains a native-like core. *Biochemistry*. **36**, 14277-14283.
189. Shimada, J. and Shakhnovich, E.I. (2002). The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc Natl Acad Sci*. **99**, 11175-11180.
190. Kolinski, A., Klein, P., Romiszowski, P.*, et al.* (2003). Unfolding of globular proteins: Monte Carlo dynamics of a realistic reduced model. *Biophys J*. **85**, 3271-3278.
191. Pande, V.S. and Rokhsar, D.S. (1999). Molecular dynamics simulations of unfolding and refolding of a β-hairpin fragment of protein G. *Proc Natl Acad Sci*. **96**, 9062-9067.
192. Lee, J. and Shin, S. (2001). Understanding β-hairpin formation by molecular dynamics simulations of unfolding. *Biophys J*. **81**, 2507-2516.
193. Scott, K.A. and Daggett, V. (2007). Folding mechanisms of proteins with high sequence identity but different folds. *Biochemistry*. **46**, 1545-1556.

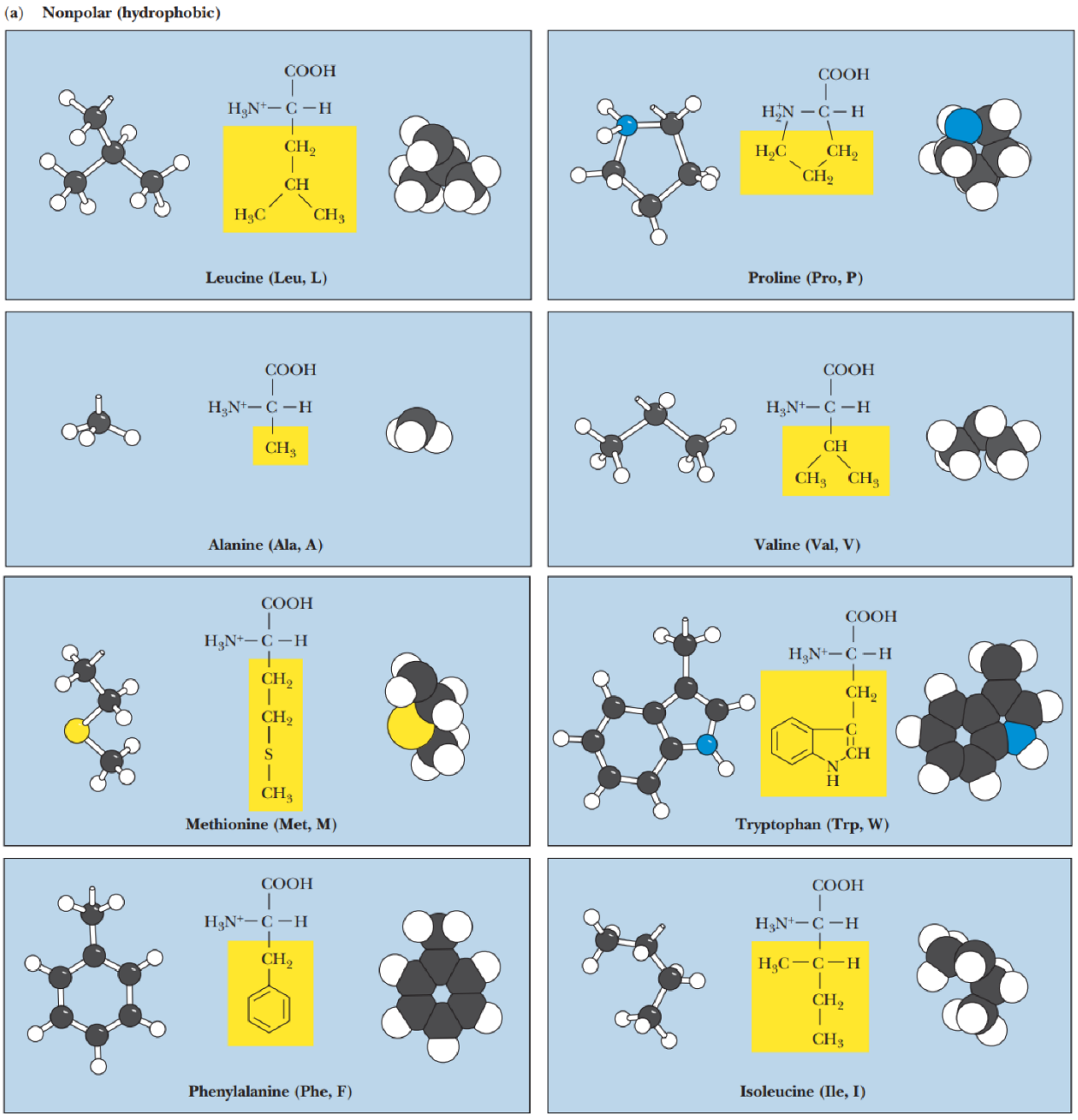
194. Jeong, Y.J., Jeong, B.C. and Song, H.K. (2011). Crystal structure of ubiquitin-like small archaeal modifier protein 1 (SAMP1) from Haloferax volcanii. *Biochem Biophys Res Commun*. **405**, 112-117.
195. Idiyatullin, D., Nesmelova, I., Daragan, V.A.*, et al.* (2003). Heat capacities and a snapshot of the energy landscape in protein GB1 from the pre-denaturation temperature dependence of backbone NH nanosecond fluctuations. *J Mol Biol*. **325**, 149-162.
196. Alexander, P.A., He, Y., Chen, Y.*, et al.* (2007). The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci U S A*. **104**, 11963-11968.
197. He, Y., Chen, Y., Alexander, P.*, et al.* (2008). NMR structures of two designed proteins with high sequence identity but different fold and function. *Proc Natl Acad Sci U S A*. **105**, 14412-14417.
198. Kim, D.E., Fisher, C. and Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein L. *J Mol Biol*. **298**, 971-984.
199. Karanicolas, J. and Brooks, C.L., 3rd. (2002). The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci*. **11**, 2351-2361.
200. Mao, Y.-J., Sheng, X.-R. and Pan, X.-M. (2007). The effects of NaCl concentration and pH on the stability of hyperthermophilic protein Ssh10b. *BMC Biochemistry*. **8**, 28.
201. Müller-Santos, M., de Souza, E.M., Pedrosa, F.d.O.*, et al.* (2009). First evidence for the salt-dependent folding and activity of an esterase from the halophilic archaea Haloarcula marismortui. *Biochim Biophys Acta Mol Cell Biol Lipids*. **1791**, 719-729.
202. Miyashita, Y., Ohmae, E., Nakasone, K.*, et al.* (2015). Effects of salt on the structure, stability, and function of a halophilic dihydrofolate reductase from a hyperhalophilic archaeon, Haloarcula japonica strain TR-1. *Extremophiles*. **19**, 479-493.
203. Ishibashi, M., Arakawa, T., Philo, J.S.*, et al.* (2002). Secondary and quaternary structural transition of the halophilic archaeon nucleoside diphosphate kinase under high- and low-salt conditions. *FEMS Microbiol Lett*. **216**, 235-241.
204. You, D.-J., Jongruja, N., Tannous, E.*, et al.* (2014). Structural basis for salt-dependent folding of ribonuclease H1 from halophilic archaeon Halobacterium sp. NRC-1. *J Struct Biol*. **187**, 119-128.
205. Pundak, S., Aloni, H. and Eisenberg, H. (1981). Structure and activity of malate dehydrogenase from the extreme halophilic bacteria of the Dead Sea. *Eur J Biochem*. **118**, 471-477.
206. Brininger, C., Spradlin, S., Cobani, L.*, et al.* (2018). The more adaptive to change, the more likely you are to survive: Protein adaptation in extremophiles. *Semin Cell Dev Biol*. **84**, 158-169.
207. Longo, L.M., Tenorio, C.A., Kumru, O.S.*, et al.* (2015). A single aromatic core mutation converts a designed "primitive" protein from halophile to mesophile folding. *Protein Sci*. **24**, 27-37.
208. Rode, B.M. (1999). Peptides and the origin of life. *Peptides*. **20**, 773-786.
209. Schwendinger, M.G. and Rode, B.M. (1989). Possible role of copper and sodium chloride in prebiotic evolution of peptides. *Anal Sci*. **5**, 411-414.
210. Scalley, M.L. and Baker, D. (1997). Protein folding kinetics exhibit an Arrhenius temperature dependence when corrected for the temperature dependence of protein stability. *Proc Natl Acad Sci U S A*. **94**, 10636-10640.
211. Scalley, M.L., Yi, Q., Gu, H.*, et al.* (1997). Kinetics of folding of the IgG binding domain of peptostreptococcal protein L. *Biochemistry*. **36**, 3373-3382.

212. Ivankov, D.N., Garbuzynskiy, S.O., Alm, E.*, et al.* (2003). Contact order revisited: influence of protein size on the folding rate. *Protein Sci*. **12**, 2057-2062.
213. Khorasanizadeh, S., Peters, I.D. and Roder, H. (1996). Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. *Nat Struct Biol*. **3**, 193-205.
214. Mizukami, T., Xu, M., Fazlieva, R.*, et al.* (2018). Complex folding landscape of apomyoglobin at acidic pH revealed by ultrafast kinetic analysis of core mutants. *J Phys Chem B*. **122**, 11228-11239.
215. Ye, K., Liao, S., Zhang, W.*, et al.* (2013). Ionic strength-dependent conformations of a ubiquitin-like small archaeal modifier protein (SAMP1) from Haloferax volcanii. *Protein Sci*. **22**, 1174-1182.
216. Latypov, R.F., Cheng, H., Roder, N.A.*, et al.* (2006). Structural characterization of an equilibrium unfolding intermediate in cytochrome c. *J Mol Biol*. **357**, 1009-1025.
217. Maki, K., Cheng, H., Dolgikh, D.A.*, et al.* (2004). Early events during folding of wild-type Staphylococcal nuclease and a single-tryptophan variant studied by ultrarapid mixing. *J Mol Biol*. **338**, 383-400.
218. Bedford, J.T., Poutsma, J., Diawara, N.*, et al.* (2021). The nature of persistent interactions in two model β-grasp proteins reveals the advantage of symmetry in stability. *J Comp Chem*. **42**, 600-607.
219. Kumar, S. and Nussinov, R. (2002). Close-range electrostatic interactions in proteins. *Chembiochem*. **3**, 604-617.
220. Fersht, A. (1998). Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. W. H. Freeman and Company.
221. de Los Rios, M.A. and Plaxco, K.W. (2005). Apparent Debye-Huckel electrostatic effects in the folding of a simple, single domain protein. *Biochemistry*. **44**, 1243-1250.
222. Bandyopadhyay, A.K., Krishnamoorthy, G., Padhy, L.C.*, et al.* (2007). Kinetics of salt-dependent unfolding of [2Fe-2S] ferredoxin of Halobacterium salinarum. *Extremophiles*. **11**, 615-625.
223. Gloss, L.M., Topping, T.B., Binder, A.K.*, et al.* (2008). Kinetic folding of Haloferax volcanii and Escherichia coli dihydrofolate reductases: haloadaptation by unfolded state destabilization at high ionic strength. *J Mol Biol*. **376**, 1451-1462.

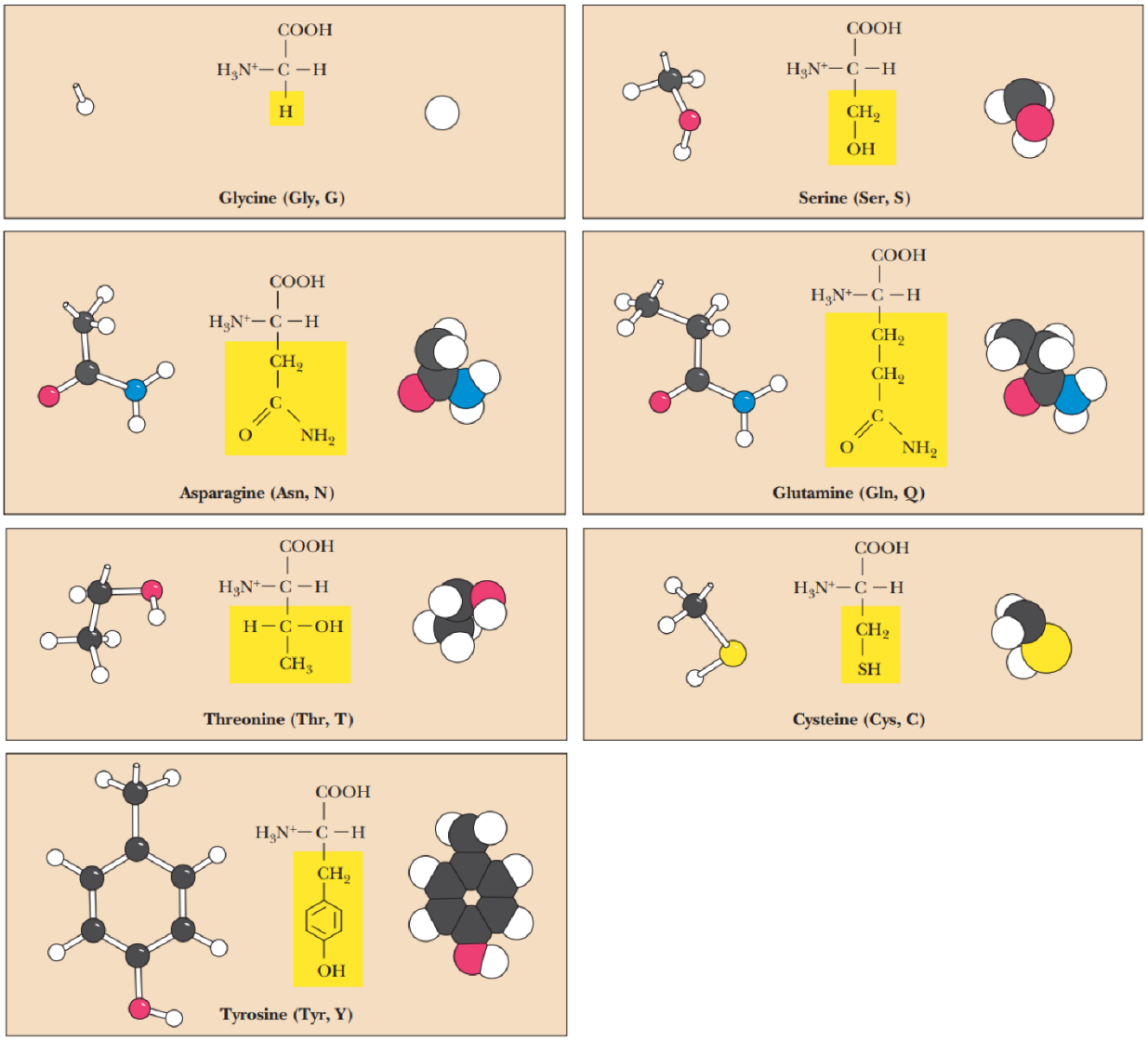
APPENDIX A

STRUCTURES OF THE 20 COMMON AMINO ACIDS

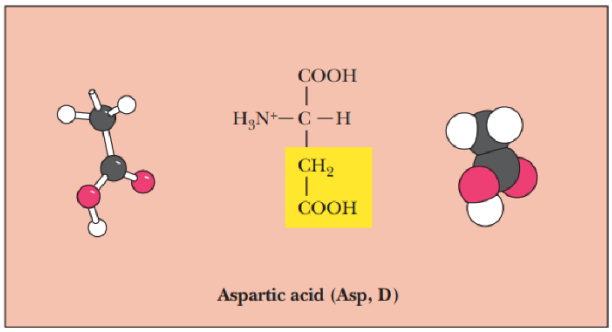
(Figure reproduced from Garrett, R.H. and Grisham, C.M. (2010). Biochemistry Brooks/Cole and used with permission)



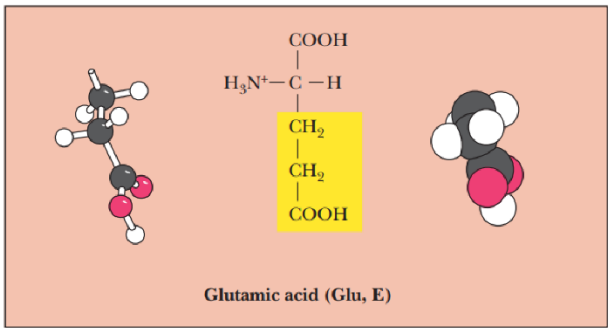
(b) **Polar, uncharged**



(c) **Acidic**

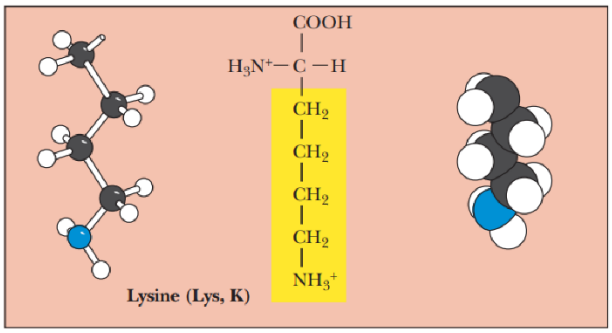


Aspartic acid (Asp, D)

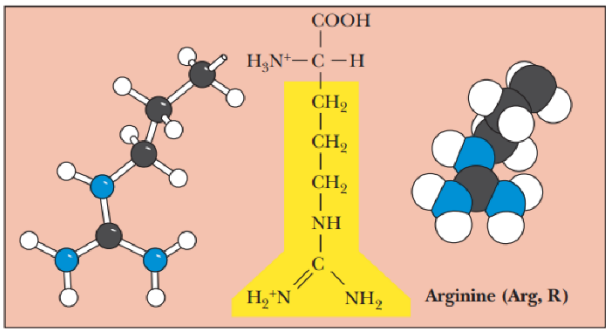


Glutamic acid (Glu, E)

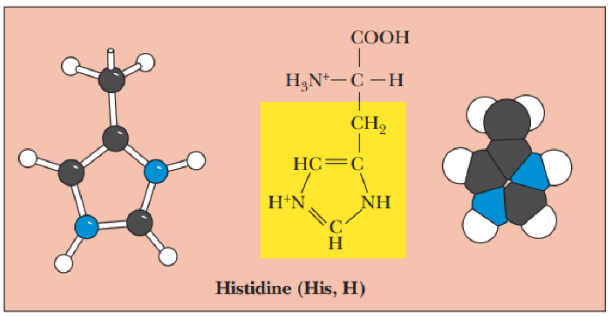
(d) **Basic**



Lysine (Lys, K)



Arginine (Arg, R)



Histidine (His, H)

APPENDIX B

COMPLETE β-GRASP SUPERFAMILY ALIGNMENT

```
1pgbA   ---------------------------------------------
2ptlA   ---------------------------------------------
1rlfA   ---------------------------------------------
3po0A   ---------------------------------------------
1enfA 2 dlhdkseltdlalanaygqynhpfikeniksdeisgekdlifrnqg 47
1fmaD   ---------------------------------------------
2k8hA   ---------------------------------------------
1f2rC   ---------------------------------------------
1euvB   ---------------------------------------------
1wm2A   ---------------------------------------------
3a4rA   ---------------------------------------------
1c4pD   ---------------------------------------------
1qlaB   ---------------------------------------------
1wspC   ---------------------------------------------
```

```
1pgbA    ----------------------------------------
2ptlA    ----------------------------------------
1rlfA    ----------------------------------------
3po0A    ----------------------------------------
1enfA 48 dsgndlrvkfatadlaqkfknknvdiygasfyykcekisenis 90
1fmaD    ----------------------------------------
2k8hA  1 ---------------------------------msnnggeps 9
1f2rC    ----------------------------------------
1euvB    ----------------------------------------
1wm2A    ----------------------------------------
3a4rA    ----------------------------------------
1c4pD    ----------------------------------------
1qlaB    ----------------------------------------
1wspC    ----------------------------------------
```

```
1pgbA   1 ----------------MTYKLILN---G------K--------TL- 12
2ptlA   1 enkeetpetpetdseeeVTIKANLI---f------a-------ngs- 30
1rlfA 646------------gssdcRIIRVQME---l------g-------edg-663
3po0A  -1 ----------------GSMEWKLF----------A--------Dl- 9
1enfA  91 eclyggttlnseklaqeRVIGANVW---V------d-------giq-120
1fmaD   1 ----------------MIKVLFFaqvr------e--------lvg 15
2k8hA  10 nnggegaegtckeetalVAVKVVNA-------------------d- 35
1f2rC   1 ---------mcavlrqpKCVKLRAL---h----------------s- 18
1euvB  20 ------------pethINLKVSDg---------------------- 31
1wm2A  12 -----------tenndhINLKVAGQ-------------------d- 26
3a4rA 339----------gplgsqeLRLRVQGk-----------------ek-350
1c4pD 149------kpiqnqaksvdVEYTVQFT---plnpdddf--------rp-177
1qlaB   1 ---------------mgRMLTIRVF---k------Ydpqsavskph- 22
1wspC 749------------pcdsIVVAYYFc---g----------------e-762
```

```
1pgbA  13 K-----GETTTEA--------V---DA---ATAEKVFKQYA 34
2ptlA  31 t-----QTAEFKg--------T---Fe----KATSEAYAYA 51
1rlfA 664 s-----VYKSILV--------T------sqdkAPSVISRVl 685
3po0A  10 aevagsrTVRVDV--------Dgdatv---GDALDALvgah 39
1enfA 121 k-----ETELIRT----nkknv---tl---qELDIKIRKIL 146
1fmaD  16 t-----dATEVAa--------d--------fptVEALRQHM 35
2k8hA  36 g-----aEMFFRI--------K---s----rtALKKLIDTY 56
1f2rC  19 a-----cKFGVAA--------r---sC---QELLRKGCVRF 40
1euvB  32 s-----sEIFFKI--------K---kt---tpL-RRLMEAF 53
1wm2A  27 g-----sVVQFKI-----krht---pl---SKLMKAYCERq 51
3a4rA 351 h-----qMLEISL--------Spdspl---kVLMSHYEeam 375
1c4pD 178 g-----lKDTKLLktlaigdti---ts---qELLAQAqsil 207
1qlaB  23 f-----qEYKIEe--------a---p----smtIFIVLNmi 43
1wspC 763 p-----iPYRTLV--------r-grav-tlGQFKE-LL--- 784
```

```
1pgbA 35--ND-----NG------VD--------------GEW-TY---------45
2ptlA 52--DT-----Lk---kdnge--------------YTV-DV---------65
1rlfA686--kk-----nnrdsavase--------------FEL-VQ---------702
3po0A 40pale-----sr------v-----fgddgelydhiNV-LR---------61
1enfA147--SD-----ky----------kiyykdseiskGLI-EF---------166
1fmaD 36--AAqsdrwal------al----------edgklLA-AV---------55
2k8hA 57--Ck-----kq--gisrns--------------VRF-LF---------71
1f2rC 41--q------------------------lpmpgsRL-CLyedgtevtd60
1euvB 54A-KR-----qg------ke----------mdsLRF-LY---------67
1wm2A 52--gl-----------smrq--------------IRF-RF---------62
3a4rA376--gl-----------sghk--------------LSF-Ff---------386
1c4pD208--nk---thpg------yt--------------iYeRdSs---------222
1qlaB 44--re--tydpd------lnfdfvcragicgscgmMI-n---grpslac77
1wspC785--tk-----kg-------s--------------YRY-YF---------794
```

```
1pgbA  46 --D---D-----------------A------------------ 48
2ptlA  66 --A---d-----------------k------------------ 68
1rlfA 703 --llpgdreltiphsanvfyamdga------------------ 725
3po0A  62 --n---------------------------------------- 62
1enfA 167 --D---M-----------------k---------tprdysfdiy 179
1fmaD  56 --n---------------------------------------- 56
2k8hA  72 --d---------------------------------------- 72
1f2rC  61 dcf---p-----------------g------------------ 65
1euvB  68 --d---------------------------------------- 68
1wm2A  63 --d---------------------------------------- 63
3a4rA 387 --d---------------------------------------- 387
1c4pD 223 --i---v-----------------thdndifrtilpmdqeftyh 244
1qlaB  78 rtltkdf-----------------e------------------ 85
1wspC 795 --k---k-----------------------------------vs 798
```

```
1pgbA  49 ------------------------T---K-TFTVTE 56
2ptlA  69 ------------------------g---Y-TLNIKF 76
1rlfA 726 ------------------------s---h-DFLLRQ 733
3po0A  63 ------------geaaalgeataag---d-ELALFP 82
1enfA 180 dlkgendyeidkiyednktlksddi---s-HIDVNL 211
1fmaD  57 ------------qtlvsfdhpltdg---d-EVAFFP 76
2k8hA  73 -------gtpidetktpeelgmedd---d-VIDAMV 97
1f2rC  66 ------------------------lpnda-ELLLLT 76
1euvB  69 -------giriqadqtpedldmedn---d-IIEAHR 93
1wm2A  64 -------gqpinetdtpaqlemede---d-TIDVFQ 88
3a4rA 388 -------gtklsgkelpadlglesg---d-LIEVWG 412
1c4pD 245 vknreqayeinkksglneeinntdl---i-SEKYYV 276
1qlaB  86 ------------------------d---G-VITLLP 93
1wspC 799 defdcgvvfeevredeailpvfeek---i-IGKVEK 830
```

```
1pgbA    ----------------------------------------
2ptlA    ----------------------------------------
1rlfA    ----------------------------------------
3po0A    ----------------------------------------
1enfA    ----------------------------------------
1fmaD    ----------------------------------------
2k8hA    ----------------------------------------
1f2rC    ----------------------------------------
1euvB    ----------------------------------------
1wm2A    ----------------------------------------
3a4rA    ----------------------------------------
1c4pD    ----------------------------------------
1qlaB 94 lpafklikdlsvdtgnwfngmsqrveswihaqkehdiskleerie 138
1wspC    ----------------------------------------
```

```
1pgbA     ---------------------------------
2ptlA     ---------------------------------
1rlfA     ---------------------------------
3po0A     ---------------------------------
1enfA     ---------------------------------
1fmaD     ---------------------------------
2k8hA     ---------------------------------
1f2rC     ---------------------------------
1euvB     ---------------------------------
1wm2A     ---------------------------------
3a4rA     ---------------------------------
1c4pD     ---------------------------------
1qlaB 139 pevaqevfeldrciecgcciaacgtkimredfvg 172
1wspC     ---------------------------------
```

```
1pgbA     ----------------------------------------
2ptlA     ----------------------------------------
1rlfA     ----------------------------------------
3po0A     ----------------------------------------
1enfA     ----------------------------------------
1fmaD     ----------------------------------------
2k8hA     ----------------------------------------
1f2rC     ----------------------------------------
1euvB     ----------------------------------------
1wm2A     ----------------------------------------
3a4rA     ----------------------------------------
1c4pD     ----------------------------------------
1qlaB 173 aaglnrvvrfmidphdertdedyyeligdddgvfgcmtlla 213
1wspC     ----------------------------------------
```

```
1pgbA     -------------------------
2ptlA  77 ------------------------ag 78
1rlfA 734 ------------------------rr 735
3po0A  83 --------------------pvsgg 87
1enfA 212 ------------------------yt 213
1fmaD  77 --------------------pvtgg 81
2k8hA  98 --------------------eqtgg 102
1f2rC  77 ---------------agetwhgyvsd 87
1euvB  94 --------------------eqigg 98
1wm2A  89 -------------------------q 89
3a4rA     -------------------------
1c4pD 277 ----------------------lkkg 280
1qlaB 214 chdvcpknlplqskiaylrrkmvsvn 239
1wspC 831 ------------------------vd 832
```

APPENDIX C

ELUCIDATING DETERMINANTS OF PROTEIN STABILITY AND FOLDING IN EXTREME ENVIRONMENTS – GB1 INVESTIGATION FOR THE VIRGINIA SPACE GRANT CONSORTIUM 2016-2017

ABSTRACT

On Earth there exist organisms that thrive in extreme conditions. Proteins are one of the most critical components of biological life, essential for cellular function and environmental adaptability. This robustness is achieved by varying amino acid content and the number and types of chemical contacts while maintaining the three-dimensional structure. The immunoglobulin-binding domain of protein G is a protein that is ideal to study due to its size and fundamental topology. To assess the role of amino acid type and amino acid interactions a protein alignment was generated, and long-range interaction networks calculated for wild-type and mutated GB1 *in silico* to analyze changes in the number of contacts. Mutated GB1 was also made *in vitro*. The variant protein was then expressed and purified in preparation for circular dichroism (CD) and fluorescence studies in pH 7.0, pH 2.0, and 3.0 M NaCl buffers. While the network analysis suggests that the variant protein will be more stable, the far- and near-UV CD and fluorescence spectroscopy data reveal that the variant protein is less stable compared to the wild-type. This may be due to overcrowding in the protein core.

INTRODUCTION

The vast unknown reaches of the universe have peaked mankind's curiosity for centuries. Questions such as what celestial masses lie beyond our own blue spheroid and does life exist elsewhere in the cosmos are timeless. There is no doubt that in the search for discovery explorers have found numerous extreme environments in outer space. From the intense temperatures on Venus [C1] to the methane lakes on Titan [C2], conditions are certainly not ideal for life.

On Earth, there are many places where extreme environments similar to those in space are found. The intense temperatures of hydrothermal vents, the high salinity of the Dead Sea, and the crushing pressure of the Marianas Trench are just a few examples. One might be tempted to think that life could not survive in such places, even on a planet thriving with life, but there are actually many diverse organisms that have found a way to flourish. One of these ways is through the increased robustness of their proteins.

Proteins are one of the four main types of macromolecules, which also include, nucleic acids, lipids, and carbohydrates, that are essential to life. They are responsible for most of the cell's vital functions and are composed of amino acids that are linked via peptide bonds. This amino acid chain is organized into secondary elements as α-helices and β-sheets that are further organized into a three-dimensional (3D) form. In many instances this tertiary structure is not functional alone and associates with other tertiary structures to become a functional quaternary structure. The secondary elements are held together with hydrogen bonds while the tertiary form of the protein is held together by a variety of interactions, including but not limited to, hydrogen bonds, hydrophobic interactions, disulfide bonds, and salt bridges. These interactions are critical to protein stability.

Proteins that are found in extremophilic organisms exhibit characteristics that allow for their enhanced stability. Some contain cysteine residues that when close enough in 3D space and under the right conditions form very stable, covalent disulfide bonds. Others contain additional noncovalent interactions which collectively increase overall stability. The length and orientation of the amino acid's side chain accounts for the number of contacts it displays. Upon mutation of an amino acid, which results in lengthening of its side chain, there is an expectation for a greater number of contacts.

We are interested in understanding the determinates of stability for protein folds and propose that increasing the number of contacts within the protein core will enhance its thermodynamic stability thus making it more amenable to extreme conditions on this planet and others. In order to undertake this investigation, we designed a combined approach involving both computational and experimental techniques using the immunoglobulin-binding domain of *Streptococcal* protein G (GB1) as a model system (Figure C1) [C3]. This is a small 56 residue protein that is 6.2 kDa in size. It contains one α-helix and one, four stranded, β-sheet. Its topology is that of a ubiquitin-like β-grasp fold. It is an ideal protein to study due to its small size and fundamental topology.

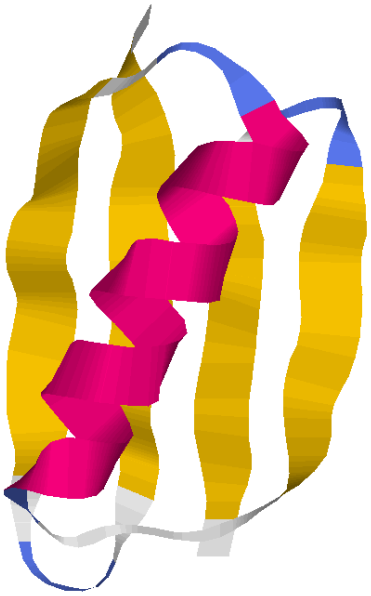


Figure C1. Immunoglobulin-binding domain of *Streptococcal* protein G. The α-helix (magenta), β-sheet (yellow), β-hairpins (blue), and loops (white) are colored accordingly. Structures visualized using RasMol Ver. 2.7.2.1.1.

Bioinformatics is a computational field of study which uses computer algorithms for the purpose of gathering and analyzing biological data [C4]. In the specific case of proteins, it is used to ascertain information about sequence, structure, and function. Bioinformatics is intrinsically linked with the concept of network analysis. This applies concepts from the field of network science to model protein structures as network systems thus allowing us to rigorously interrogate the nature of long-range interactions [C5]. Thus, a combination of bioinformatics and network science approaches will be applied to the present research investigation.

On the experimental side, we are able to test the role of interactions in the stability of the protein by making a mutation, expressing and purifying the variant protein, and conducting structural and stability studies. The latter is achieved with circular dichroism which is a

biophysical method used to study the structural components of proteins [C6]. There are two types of circular dichroism, far-UV (190-250nm), which monitors a protein's secondary structure and near-UV (250-320nm), which monitors a protein's tertiary structure. Circular dichroism is the differential absorption of left and right polarized light causing elliptical polarization. This work will be complimented with fluorescence spectroscopy to monitor the stability of the core.

In the course of our computational studies, we proposed specifically that changing alanine in the 26th position to phenylalanine would enhance thermodynamic stability without affecting the overall structure, however, based on experimental results we discovered that GB1 was destabilized and the protein structure altered. While unexpected, this is in fact very interesting, because it suggests that the β-grasp fold may be limited in its ability to accommodate and increase in hydrophobicity. Therefore, this protein form may not exist in extreme environments beyond those known to contain this fold on Earth and may not be amenable to other planets with more extreme environments than our own. This also makes us think about the nature of our present protein structure universe, how it evolved, and what constitutes an allowable fold.

MATERIALS

Bioinformatics studies were conducted using DALI [C7] and the CATH [C8] database for identifying related sequences and structures. Proteins were visualized using RasMol [C9]. A mutation *in silico* was made using Insight II (Accelrys). Contact distances were calculated using the Contact program (CCP4) [C10]. Networks were generated on a SUN Workstation running Linux using a DegLR program written in the laboratory. The networks were visualized using Pajek [C11]. Luria broth (LB) media, supplemented with 100 µg/ml carbenicillin (Teknova), was used for bacterial cultures and agar plates. Mutagenesis reactions were performed using the Q5

site-directed mutagenesis kit from New England Biolabs (NEB). A Strataprep mini plasmid prep kit (Agilent Technologies) was used for the purification of plasmid DNA. Plasmid DNA samples were sent to the Molecular Core Facility at Eastern Virginia Medical School for sequencing. Protein was expressed using BL21 (DE3) competent *E. coli* cells (NEB) via induction with 0.4 mM IPTG (IBI Scientific). An AKTA purification system (GE) was used for the purification of protein samples. Q Sepharose fast flow resin (GE) was used for anion exchange and Sephacryl S-100 (GE) was used for gel filtration. 1.0 kDa MWCO dialysis tubing (Spectrum Labs) was used for dialysis. Protein purity was verified using 4-12% Bis-Tris gels (Invitrogen). Relevant IBC protocol numbers are 16-005 and 17-010.

METHODS

The present research investigation is delineated into three major aims. In Aim 1, a search was done using the DALI database and the PDB FASTA sequence for GB1 as the query. The results generated were from many different organisms. Multiple chains of the same protein were also displayed as results. The results were filtered to obtain a list of proteins that were exclusively from extremophilic organisms. Only one chain from each protein was chosen for evaluation. The selected protein sequences were then structurally aligned using DALI. The alignment was then verified by evaluating each amino acid position with reference to GB1 using the RasMol visualization program and making corrections as necessary. The CATH database was then used to assign domains to the various proteins. Only the domain that aligned with GB1 was included in the alignment, all others were removed. Upon gathering the alignments, each amino acid position was visually verified as being aligned with GB1 by analyzing the 3D structures in RasMol. Corrections were made as necessary to produce the final alignment. The

networks were generated with the DegLR program which used a contact file as input and visualized in Pajek. The cutoffs used in the network were 5 Å contact distance between atoms and seven residue separation between pairs of interacting amino acids [C12].

In Aim 2, site-directed mutagenesis and transformation were performed using the protocol from NEB. Polymerase chain reaction running conditions were as follows: initial denaturation at 98 °C for 30 s, 25 cycles of denaturation at 98 °C for 10 s, annealing at 57 °C for 30 s, and extension at 72 °C for 3 min and 30 s, finally a final extension at 72 °C for 2 min. The transformed cells were plated and incubated at 37 °C overnight. A single colony was then obtained from the selective plate and cultured in 50 ml of selective LB media and incubated at 37 °C overnight with shaking at 250 rpm. Plasmid DNA was extracted and purified. Samples were the sent for sequencing. Upon confirmation of the mutated cDNA sequence, the plasmid was transformed into BL21 (DE3) competent *E. coli* cells using the protocol provided. The transformed cells were again plated on selective LB agar and incubated at 37 °C overnight. A single colony was then obtained from the selective plate and cultured in 50 ml of selective LB media and incubated at 37 °C overnight with shaking at 250 rpm. 6.0 L of selective LB media was inoculated with 2 ml of starter culture and incubated at 37 °C with shaking at 250 rpm until the OD_{600} was between 0.6 and 0.8 representing the mid-log phase of growth. Cultures were then inoculated with more carbenicillin bringing the final concentration to 200 μg/ml and IPTG at a final concentration of 0.4 mM. Cultures continued shaking incubation at 37 °C for 4 hours. After incubation cultures were centrifuged into a single pellet at 8000 rpm for 30 min and covered with 50 ml of buffer containing 20 mM Tris base, pH 8.5 and stored at -20 °C overnight. The bacterial pellet was thawed and solubilized in buffer covering the pellet and sonicated on ice at 25% amplitude for 4 hours with 10 s pulses per minute using an ultrasonic processor. Sonicated lysate

was then heated at 80 °C for 15 min to precipitate out proteins that are not thermostable. Lysate was centrifuged at 16000 rpm for 20 min to pellet bacterial debris and decanted into a sterile falcon tube. Lysate was loaded onto a well equilibrated XK26 anion exchange column containing 38 ml of Q sepharose fast flow resin. Sample was loaded and washed at a flow rate of 0.5 ml/min using a 20 mM Tris buffer, pH 8.5 and was eluted at a flow rate of 1.0 ml/min for 600 min using a 20 mM Tris, 500 mM NaCl buffer, pH 8.5. Peaks were ran using gel electrophoresis to check for purity. Peaks containing the variant protein were pooled and dialyzed into double deionized water using 1.0 kDa MWCO dialysis tubing, frozen using liquid nitrogen, and lyophilized on a freezer drier. The variant protein was solubilized at a concentration of 40 mg/ml and loaded onto a XK16 size exclusion column containing 120 ml of Sephacryl S-100 resin with a running buffer containing 50 mM Tris and 200 mM NaCl, pH 6.5. Column was run at 0.5 ml/min for 4 hours. Gel electrophoresis was again used to check for purity. Peaks containing the variant protein were pooled and dialyzed into double deionized water using 1.0 kDa MWCO dialysis tubing, frozen using liquid nitrogen, and lyophilized on a freezer drier.

The structural and stability studies were conducted in Aim 3. This involved using far- and near-UV CD (Jasco J-815) in the native condition (pH 7.0) and two extreme conditions (3.0 M NaCl and pH 2.0). Comparative stability studies were more clearly analyzed using thermal unfolding by fluorescence spectroscopy on a Cary Eclipse spectrophotometer. GB1 has an intrinsic tryptophan in the core (Trp 43) and was monitored using 295 nm excitation and 350 nm emission wavelengths. The slit widths were 5 and 10, respectively.

RESULTS AND DISCUSSION

The final corrected structural alignment of extremophilic proteins in reference to GB1 is shown in Figure C2. It includes the query sequence (1pgb) and sequences found in a cryophile (2pw9), halophile (3po0), a protein found in oil wells (2l52), and a thermoacidophile (2g1e).

```
          1         10                20          30
          |         |                 |           |
[1pgb] ---MTYKLILNGKTLK-----GETTTE-AV---DAATAEKVFKQYA---NDNG--V------
[2pw9] --eTPYAIALN---dR-----ViGSSM-VL---p-VDLEEFGAGFL---FGQG--Yikkae-
[3po0] ---GSMEWKLF-ADlAevagsRTVRVD-VDgdaTVGDALDALVGAH-palesr--v------
[2l52] ghmAEVKVKLF-AnlReaagtPELPLS-GE---KVIDVLLSLTDK----YPALkyvifekgd
[2g1e] ----MVTVRYYATlr-pitkkKEETFNgIS---KISELLERLKVEYgseftkq---------
                40                        50
                |                         |
[1pgb] ----------DGEWTYDD----------------ATKTFTVTE-----
[2pw9] ---------eiREILVCP----------------QGRISVYA-----
[3po0] -fgddgelydHINVLRn------geaaalgeataagDELALFPpvsgg
[2l52] ekseililcg--SINILIngnnirhlegletllkdSDEIGILPpvsgg
[2g1e] -mydgnnlfkn--VIILVngnnitsmkgldteikdDDKIDLFppvagg
```

Figure C2. Corrected structure-based alignment. Side chains that are not aligned are shown as lowercase letters. Gaps are delineated by dashes. PDB codes are in brackets on the left. Positions 26 and 52 are shown in bold.

It is interesting to note where the gap sequences are in GB1. These are areas in which amino acids have been added in some of the other proteins. These extra amino acids may be important in facilitating enhanced stability in their respective extremophilic proteins. It is also worthy to note that position 26 of the alignment is occupied by amino acids whose side chains are more moderate in size. In GB1 the amino acid is located in the α-helix and the side chain is

pointed toward the interior of the protein. It was for these reasons that the alanine residue in position 26 of GB1 was selected for mutation to a phenylalanine. This mutation is shown in Figure C3.

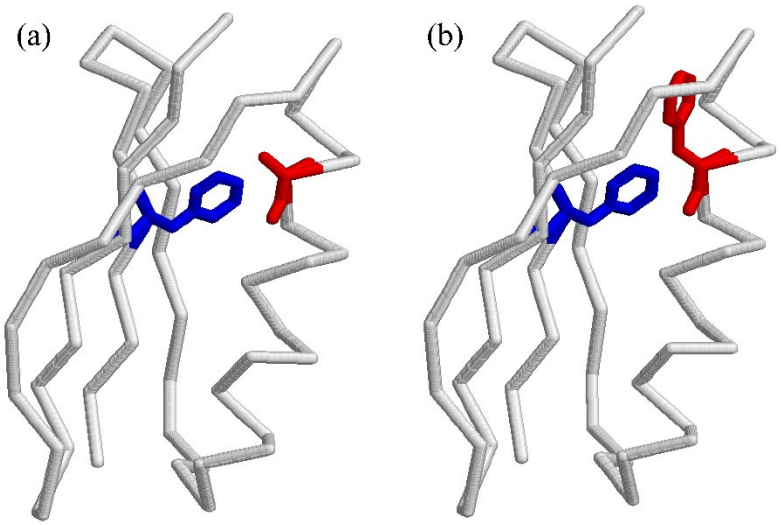


Figure C3. Backbone representation of GB1. In red, (a) alanine 26 is shown mutated to (b) phenylalanine. Phenylalanine in position 52 is shown in blue. Structures visualized using RasMol Ver. 2.7.2.1.1.

To evaluate how the mutation will affect the protein, networks were constructed for the wild-type and mutated forms of GB1. As seen in Figure C4, when a phenylalanine replaces the alanine in position 26, new long-range interactions are generated with residues 1, 2, and 19. Residues 1 and 2 are located in the first β-strand of the protein and residue 19 is located in the second β-strand. Since more contacts often lead to a more stable structure, it would stand to reason that this variant protein should exhibit increased stability.

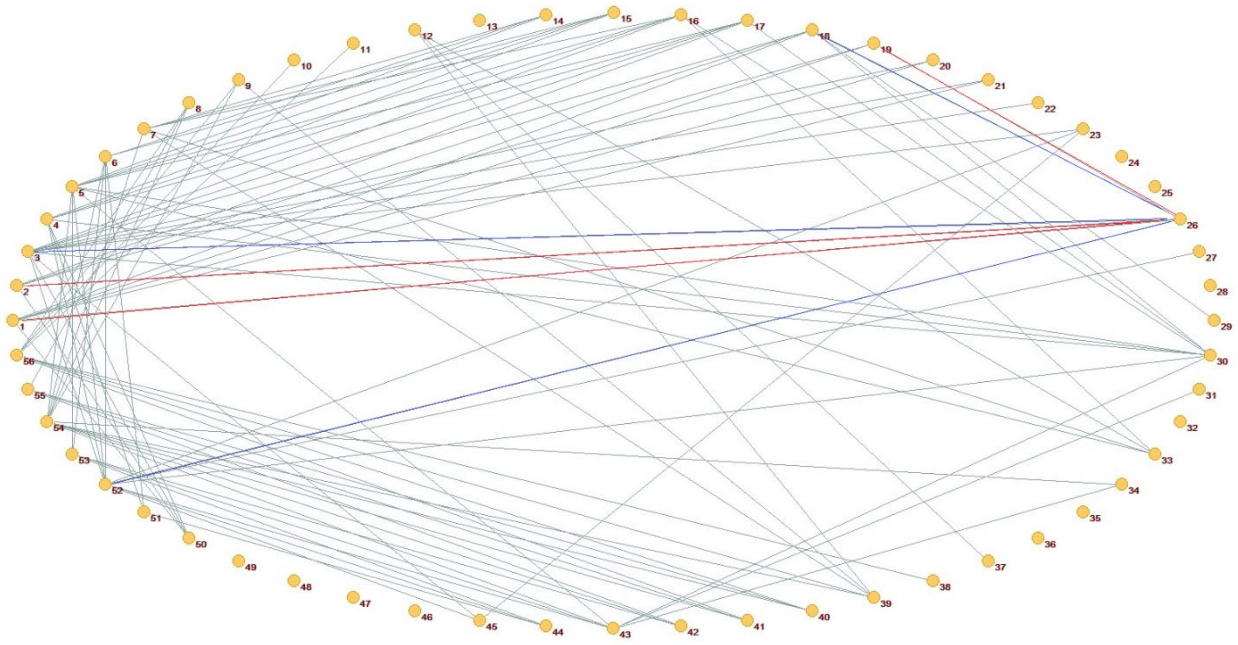


Figure C4. Long-range interaction network of GB1. Amino acids are filled circles connected by long-range interactions shown as lines. Long-range interactions involving residue 26 are shown for wild-type GB1 (dark blue). Upon mutation to phenylalanine, the formation of additional long-range interactions occurs (red). Data plotted using Pajek64-XXL 4.08.

To determine the effect on protein structure and stability experimentally, far- and near-UV CD studies were performed. Figure C5 shows the results from the far-UV CD analysis. Both wild-type and variant GB1 are the least stable in an acidic environment. In wild-type GB1 the presence or absence of NaCl causes little variation in protein stability however in the variant protein the addition of salt results in increased stability. The overall results show that the variant protein, independent of the buffer, exhibits a decrease in secondary structure when compared to the wild-type and is therefore less stable.

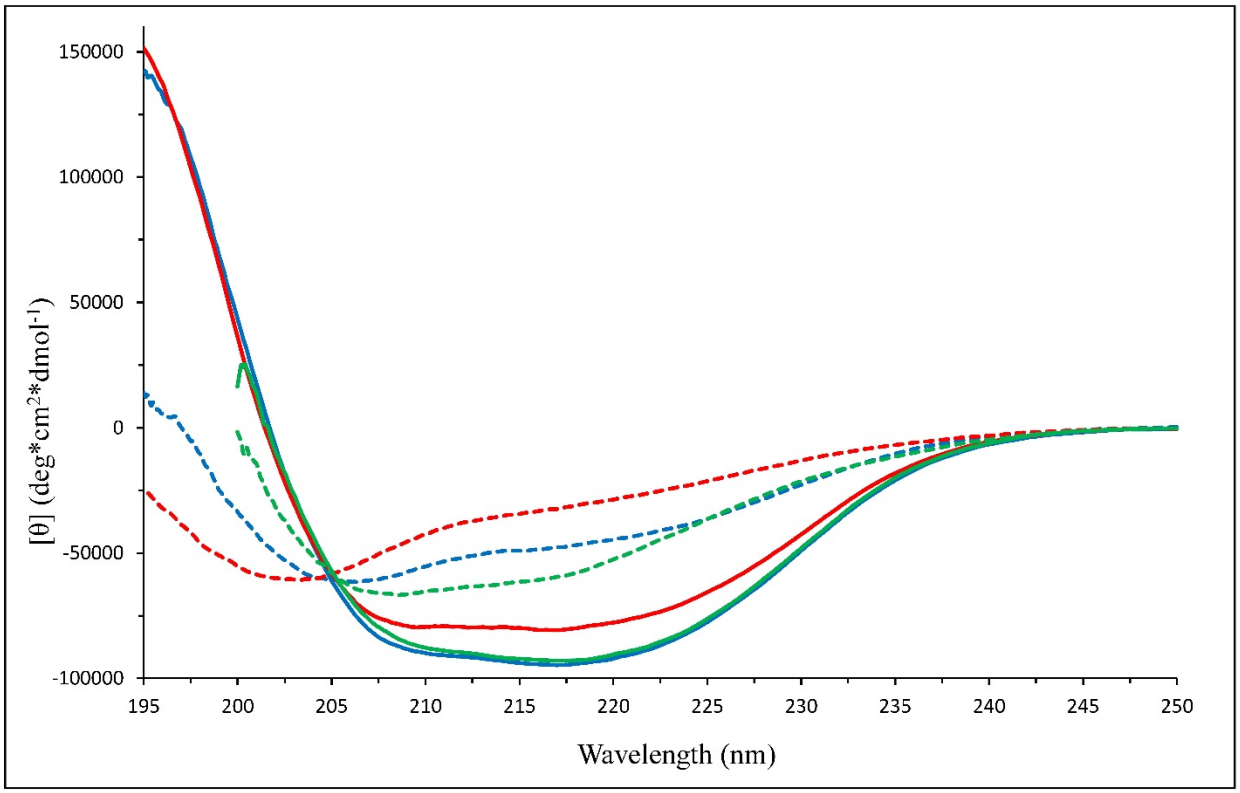


Figure C5. Far-UV CD of GB1. Wild-type and variant GB1 are shown as solid and dashed lines respectively. Samples were run at pH 7.0 (blue), pH 2.0 (red), and 3.0 M NaCl (green). Data plotted using Microsoft Excel 365.

Figure C6 shows the results from the near-UV CD analysis during thermal unfolding. Wild-type GB1 is the least stable in acidic conditions, losing its tertiary structure between 20 °C and 55 °C while variant GB1 retains little tertiary structure under all three conditions at 20 °C. The overall results show that the variant protein displays a decrease in tertiary structure under all conditions when compared to the wild-type. Thermal unfolding was also monitored using fluorescence spectroscopy.

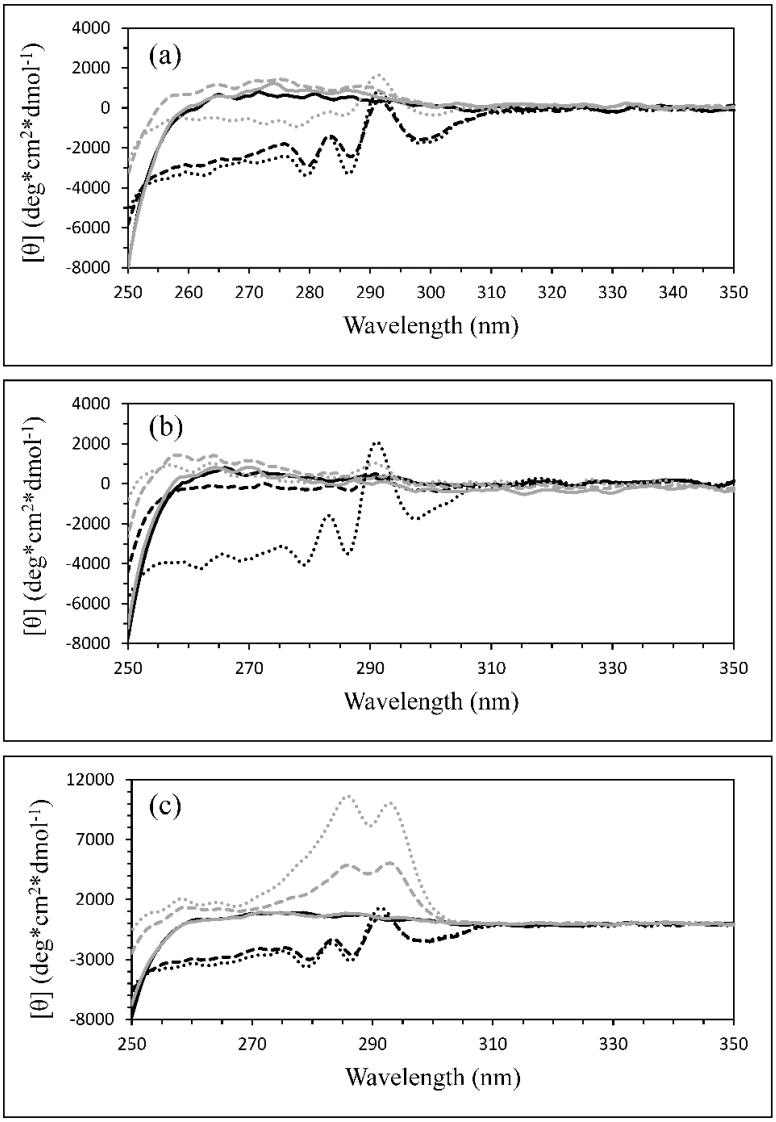


Figure C6. Near-UV CD of GB1 during thermal unfolding. Wild-type and variant GB1 are shown in black and grey respectively. Thermal unfolding was performed at (a) pH 7.0, (b) pH 2.0, and in (c) 3.0 M NaCl. Temperature data is shown for 20 °C (dotted lines), 55 °C (dashed lines), and 95 °C (solid lines). Data plotted using Microsoft Excel 365.

As shown in Figure C7, the variant protein is less stable than the wild-type. The profiles for the two proteins match in that they are more stable when salt is present and less stable in an

acidic environment. These results parallel those obtained for far-UV CD and also help to determine whether the presence of salt makes the wild-type protein more stable. The results of this study show that while theoretically, through a computational study, a mutation that produces a greater number of contacts should increase the stability of the protein, experimentally this is not always the case.

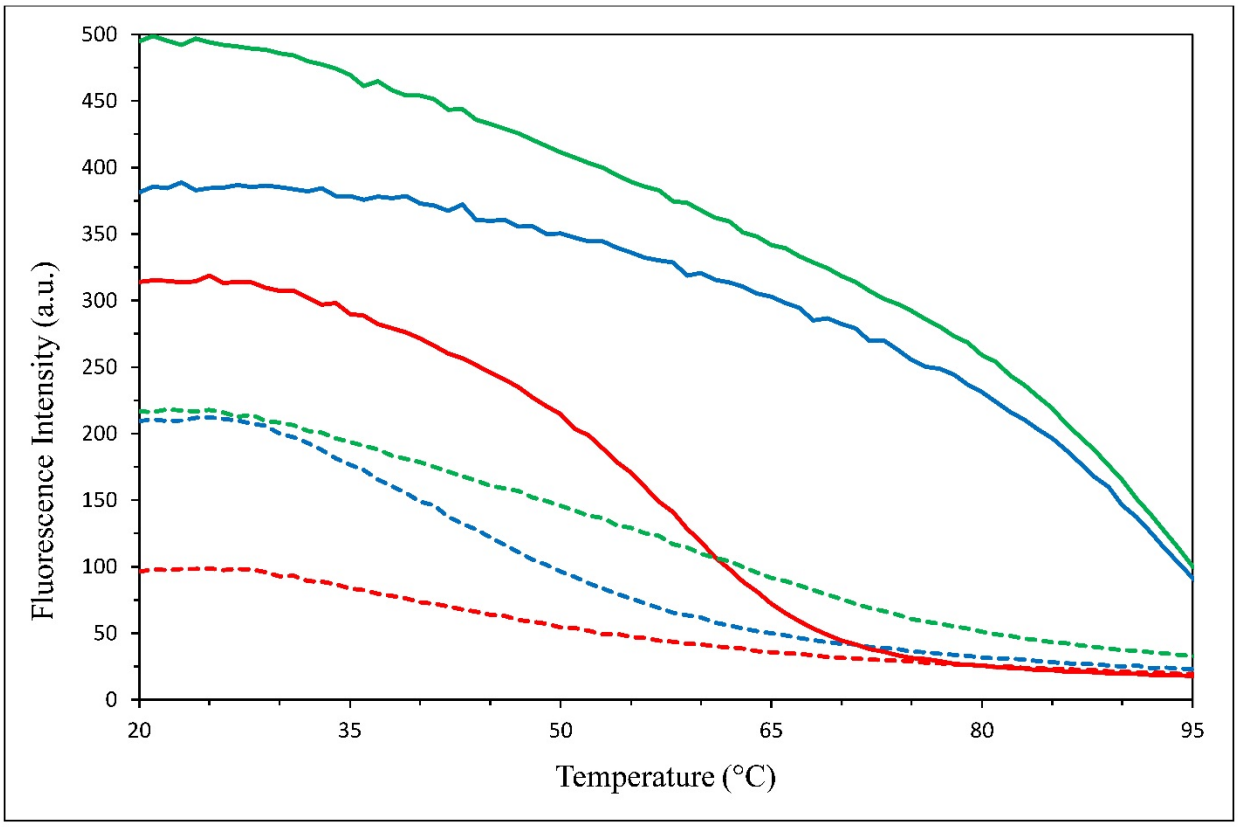


Figure C7. Fluorescence spectroscopy during thermal unfolding. Wild-type and variant GB1 are shown as solid and dashed lines respectively. Samples were run at pH 7.0 (blue), pH 2.0 (red), and 3.0 M NaCl (green). Data plotted using Microsoft Excel 365.

The axin dix domain (ADD) found in the Norwegian rat, PDB code 1WSP, when structurally aligned with GB1, contains a phenylalanine in position 26 (Figure C8). Upon analysis of the two protein structures, it was seen that opposite the amino acid side chain in ADD there is a glycine residue while in GB1 there is another phenylalanine. This means that while a phenylalanine residue is allowed to exist in this location there must be a compensatory mutation that shortens the side chain of the amino acid opposite that of the phenylalanine. If this mutation does not occur, overcrowding within the interior of the protein core results and the stability is decreased.

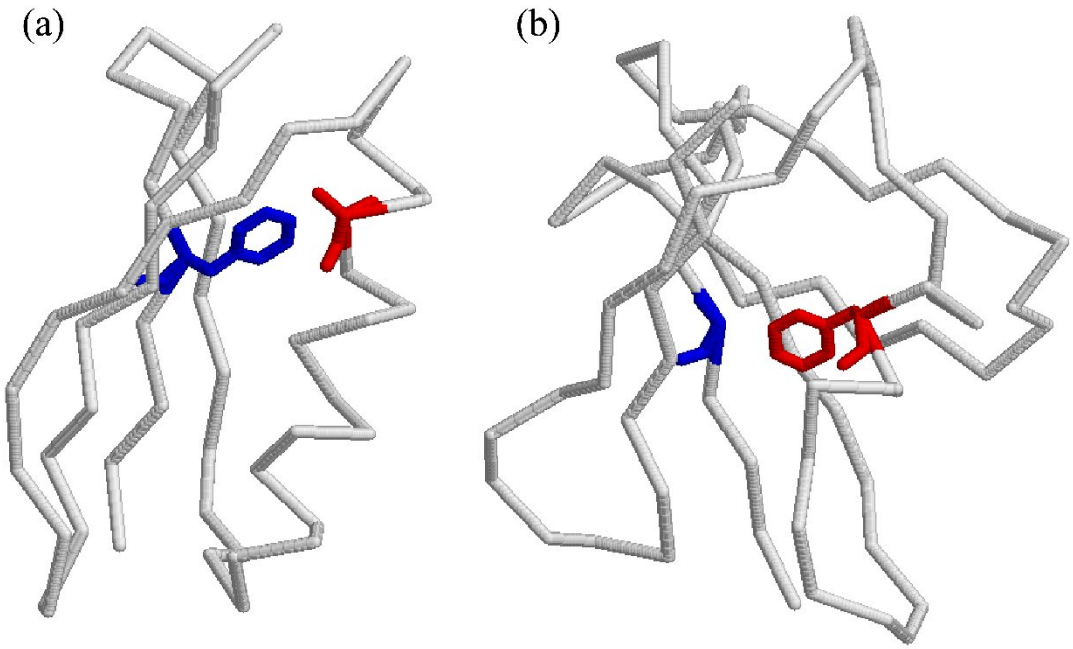


Figure C8. GB1 compared to the axin dix domain. In (a) GB1 alanine 26 and phenylalanine 52 are shown in red and blue, respectively. In (b) the axin dix domain phenylalanine and glycine are shown in red and blue, respectively. Structures visualized using RasMol Ver. 2.7.2.1.1.

These findings have implications for further understanding how protein folds evolved on this planet and how the density of protein cores may limit further adaption on other planets that have conditions more extreme than our own.

ACKNOWLEDGEMENTS

Funding from the Virginia Space Grant Consortium Graduate Fellowship.

REFERENCES

C1. Hashimoto G.L., Roos-Serote M., Sugita S., *et. al.* (2008) Felsic highland crust on Venus suggested by Galileo Near-Infrared Mapping Spectrometer data. *J Geophys. Res.* **113**, E00B24.
C2. Stofan E.R., Elachi C., Lunine J.I., *et. al.* (2007) The lakes of Titan. *Nat. Lett.* **445**, 61-64.
C3. Gallagher T., Alexander P., Bryan P., *et. al.* (1994) Two crystal structures of the B1 immunoglobulin-binding domain of *Streptococcal* protein G and comparison with NMR. *Biochemistry*. **33**, 4721-4729.
C4. Pevsner J. (2015) Bioinformatics and functional genomics. Wiley and Blackwell, 3rd edition, 3-4.
C5. Greene L.H., Higman V.A. (2003) Uncovering network systems within protein structures. *J Mol. Biol.* **334**, 781-791.
C6. Kelly S.M., Jess T.J., and Price N.C. (2005) How to study proteins by circular dichroism. *Biochim. Biophys. Acta.* **1751**, 119-139.
C7. Holm L. and Rosenström P. (2010) Dali Server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, 545-549.
C8. Sillitoe I., Lewis T.E., Cuff A., *et. al.* (2015) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* **43**, 376-381.
C9. Sayle R.A., Milner-White E.J. (1995) Rasmol – biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374-376.
C10. Winn M.D., Ballard C.C., Cowtan K.D., *et. al.* (2011) Overview of the CCP4 suite and current developments. *Acta Cryst.* **67**, 235-242.
C11. Mrvar A. and Batagelj V. (2016) Analysis and visualization of large networks with program package Pajek. *Complex Adapt. Syst. Model.* **4**, 6.
C12. Collins J.C., Bedford J.T. and Greene L.H. (2016) Elucidating the key determinants of structure, folding, and stability for the (4β + α) conformation of the B1 domain of protein G using bioinformatics approaches. IEEE *Trans. on Nanobioscience.* **15**, 140-147.

APPENDIX D

FURTHER INVESTIGATION INTO ELUCIDATING DETERMINANTS OF PROTEIN STABILITY AND FOLDING IN EXTREME ENVIRONMENTS – GB1 INVESTIGATION FOR THE VIRGINIA SPACE GRANT CONSORTIUM 2017-2018

ABSTRACT

Computational bioinformatics studies were conducted using the immunoglobulin-binding domain of protein G (GB1) to assess the role of amino acid type and amino acid interactions in dictating structure and stability. This directed experimental studies and a double mutant, GB1-Ala26Phe-Phe52Ala, which was synthesized. The variant protein was then expressed. Following biophysical studies in physiological and extreme conditions, (high salt and high temperature), it was determined that the variant was highly unstable in physiological conditions however interestingly in 3.0 M NaCl the structure and stability of the variant protein increased. This suggests that extreme conditions are not necessarily deleterious to the building blocks of cells and can facilitate adaptability.

INTRODUCTION

This work is a continuation of the research presented in Appendix C. As a result of the previous Virginia Space Grant Consortium computational studies, we discovered that changing alanine in the 26th position to phenylalanine in GB1 resulted in destabilization and altered the protein structure despite phenylalanine being moderately expected in a helix compared to other amino acids [D1]. We hypothesized this may have been due to overcrowding in the protein core and proposed that mutating a phenylalanine in the 52nd position to alanine would cause some

stability and structure to be regained. The characterization of this double variant of GB1 is the focus of one of our present research aims This new compensatory mutation was selected due to the presence of a glycine residue in a protein found in the Norwegian rat that has a structure similar to GB1. However, this was not the case and the resultant protein was further destabilized. While unexpected, this is in fact very interesting, because it suggests that compensatory mutations are limited by the 3D space around the residue.

MATERIALS

Bioinformatics studies were conducted using the DALI [D2] and BLAST [D3] databases for identifying related sequences and calculating their percent identity and RMSD. Protein sequences were aligned using MUSCLE [D4]. Proteins were visualized using RasMol [D5]. A mutation *in silico* was made using Insight II (Accelrys). Contact distances were calculated using the Contact program (CCP4) [D6]. Networks were generated on a SUN Workstation running Linux using a DegLR program written in the laboratory. The networks were visualized using Pajek [D7]. Luria broth (LB) media, supplemented with 100 μg/ml carbenicillin (Teknova), was used for bacterial cultures and agar plates. Mutagenesis reactions were performed using the Q5 site-directed mutagenesis kit from New England Biolabs (NEB). A Strataprep mini plasmid prep kit (Agilent Technologies) was used for the purification of plasmid DNA. Plasmid DNA samples were sent to the Molecular Core Facility at Eastern Virginia Medical School for sequencing. Protein was expressed using BL21 (DE3) competent *E. coli* cells (NEB) via induction with 0.4 mM IPTG (IBI Scientific). An AKTA purification system (GE) was used for the purification of protein samples. Q Sepharose fast flow resin (GE) was used for anion exchange and Sephacryl S-100 (GE) was used for gel filtration. 3.0 kDa MWCO concentrators

(Sartorius) were used for protein concentration and buffer exchange. Protein purity was verified using 4-12% Bis-Tris gels (Invitrogen). Relevant IBC protocol numbers are 16-005 and 17-010.

METHODS

The present research investigation is delineated into three major aims. In Aim 1, a search was done using the BLAST database and the PDB FASTA sequence for SAMP1 as the query. SAMP1 was used as the query because it has the same topology as GB1 and is found in a halophilic organism. The results generated were from many different organisms. The results were filtered to obtain a list of proteins that were exclusively from extremophilic organisms. The selected sequences were then aligned using MUSCLE. In Aim 2, we sought to expand upon the work previously carried out in the VSGC project. A mutation from alanine in position 26 to phenylalanine was made. This mutation destabilized the protein and was hypothesized that this was due to overcrowding in the core. To test this theory another mutation was made; phenylalanine in position 52 to alanine. The new network was generated with the DegLR program which used a contact file as input and visualized in Pajek. The cutoffs used in the network were 5 Å contact distance between atoms and seven residue separation between pairs of interacting amino acids [D8]. In Aim 3, site-directed mutagenesis and transformation were performed using the protocol from NEB followed by structural and stability studies. Polymerase chain reaction running conditions were as follows: initial denaturation at 98 °C for 30 s, 25 cycles of denaturation at 98 °C for 10 s, annealing at 62 °C for 30 s, and extension at 72 °C for 3 min and 30 s, finally a final extension at 72 °C for 2 min. The transformed cells were plated and incubated at 37 °C overnight. A single colony was then obtained from the selective plate and cultured in 50 ml of selective LB media and incubated at 37 °C overnight with shaking at 250

rpm. Plasmid DNA was extracted, purified and sent for sequencing. Upon confirmation of the mutated cDNA sequence, the plasmid was transformed into BL21 (DE3) competent *E. coli* cells using the protocol provided. The transformed cells were again plated on selective LB agar and incubated at 37 °C overnight. A single colony was then obtained from the selective plate and cultured in 50 ml of selective LB media and incubated at 37 °C overnight with shaking at 250 rpm. 5.0 L of selective LB media was inoculated with 500 µL of starter culture and incubated at 37 °C with shaking at 250 rpm until the OD_{600} was between 0.6 and 0.8 representing the mid-log phase of growth. Cultures were then inoculated with more carbenicillin bringing the final concentration to 200 µg/ml and IPTG at a final concentration of 0.4 mM. Cultures continued shaking incubation at 37 °C for 4 hours. After incubation cultures were centrifuged into a single pellet at 8000 rpm for 30 min and stored at -20 °C overnight. The bacterial pellet was thawed and solubilized in buffer containing 20 mM Tris base, pH 8.5 and sonicated on ice at 40% amplitude for 2 hours with 10 s pulses per minute using an ultrasonic processor. Sonicated lysate was then heated at 80 °C for 15 min to precipitate out proteins that are not thermostable. Lysate was centrifuged at 16000 rpm for 20 min to pellet bacterial debris and decanted into a sterile falcon tube. Lysate was filtered through a 0.45 µm membrane and loaded onto a XK26 anion exchange column containing 38 ml of Q sepharose fast flow resin. Sample was loaded and washed at a flow rate of 0.5 ml/min using a 20 mM Tris buffer, pH 8.5 and was eluted at a flow rate of 0.5 ml/min for 230 min using a 20 mM Tris, 500 mM NaCl buffer, pH 8.5. Peaks were ran using gel electrophoresis to check for purity. Peaks containing the variant protein were pooled and concentrated using 3.0 kDa MWCO concentrator. The variant protein was loaded onto a XK16 size exclusion column containing 120 ml of Sephacryl S-100 resin with a running buffer containing 50 mM Tris and 200 mM NaCl, pH 7.5. Column was run at 0.5 ml/min for 4 hours.

Gel electrophoresis was again used to check for purity. Peaks containing the variant protein were pooled, concentrated, and buffer exchanged into double deionized water using 3.0 kDa MWCO concentrator.

The structural and stability studies involved using CD (Jasco J-815). Near-UV CD was performed in the native condition (pH 7.0) and far-UV CD was performed in the native condition (pH 7.0) and at an extreme condition (3.0 M NaCl). Comparative stability studies were more clearly analyzed using thermal unfolding by fluorescence spectroscopy on a Cary Eclipse spectrophotometer. GB1 has an intrinsic tryptophan in the core (Trp 43) and was monitored using 295 nm excitation and 350 nm emission wavelengths at a concentration of 0.05 mg/ml. The slit widths were 5 and 10, respectively.

RESULTS AND DISCUSSION

The final corrected structural alignment of extremophilic proteins in reference to SAMP1 is shown in Figure D1. It includes the halophilic query sequence (3PO0) and sequences found in a halophile (WP_020445345.1), haloacidophile (WP_021780493.1), haloalkaliphile (WP_011324211.1), alkaliphile (WP_093047229.1), thermophile (WP_005588799.1), alkalithermophile (WP_007506270.1), and two hyperthermophiles (WP_010879121.1 ; WP_048091823.1).

```
                    1         10        20          30        40
                    |         |         |           |         |
[3PO0]              GSMEWKLFADLAEVAGSRTVRVD-VDGDA-TVGDALDALVGAHPALESRVFGD
[WP_020445345.1]    --MHWKLFADLAEVAGSDAVDVD-AGADAETVGEALDALLADRPDLATRVLDE
[WP_021780493.1]    --MEWKLFADLAEVAGGDRIAVS-VPGDA-TVGDALDALLDDRPALRERVLDE
[WP_011324211.1]    --MEWRLFANLAEAAGTKRVEVDAAPGD--TFGDAFEQLLAAHPDLEAEVLDE
[WP_093047229.1]    --MIIKVFATFREICGGKTIQVDYEDNS--RIGDLLDEVISQFPAMEEEIFTT
[WP_005588799.1]    --MKVKVFANFREICQSPVVEVK-ATGD--RVIDILEALTRQYPALETEIFDA
[WP_007506270.1]    --MQIKVFANFREICGGKSVTLDLTPPQ--TVASVLDALIDQYPPMKEELFTE
[WP_010879121.1]    -MVRVKLFANFREAAGVKEVEVE-AG----TVGEVLQELVRRFPKLES-LFYE
[WP_048091823.1]    -MIKVKFPTVFVQITGKRRIEVDGVT----TVSQLLEKLFDEYPQLKERLI-K

                    50        60             70        80
                    |         |              |         |
[3PO0]              DGELYDHINVLRNGE-----AAALGEATAAGDELALFPPVSGG
[WP_020445345.1]    DGEVYDHVNVLKNGSDVTATDAGLETAVEDGDELALFPPVSGG
[WP_021780493.1]    TGSLYDHVNLLRNGR-----DAALDEELEAGDELALFPPVSGG
[WP_011324211.1]    DGELRDHIRVLRNDRNPFVSDDGFETTLEEGDELALFPPVSGG
[WP_093047229.1]    DRQVKQHVHVFINGRNIIH-LQGLETIVKPDDQLALFPPVAGG
[WP_005588799.1]    ERKLKPFVHVFINGRNAIH-EQGLETKVKASDQFALFPPVAGG
[WP_007506270.1]    QKTLKPLVHVFVNGQNIIH-LDGLKTKVESKDEIALFPPVAGG
[WP_010879121.1]    EGRLRDYVNIMVNGRNV---RGDLNYPLSHTDEVAIFPPVSGG
[WP_048091823.1]    DGDLSPFINIFVNGEDIRF-LNGLDTEIKDGDEVAFIPAISGG
```

Figure D1. MUSCLE alignment. Gaps are delineated by dashes. Accession numbers are in brackets on the left. Numbering is with respect to 3PO0.

It is interesting to note that halophilic sequences in the alignment share the same amino acid at positions 3, 9, 31, and 46 with respect to 3PO0 whose side chains point toward the protein core and positions 60 and 75 with respect to 3PO0 whose side chains are solvent exposed. These amino acids may be important in facilitating enhanced stability in a high salinity environment. We propose that mutating a phenylalanine residue in position 52 to alanine will alleviate some of the overcrowding (Figure D2). To test our hypothesis, a network was constructed for the mutated form of GB1.

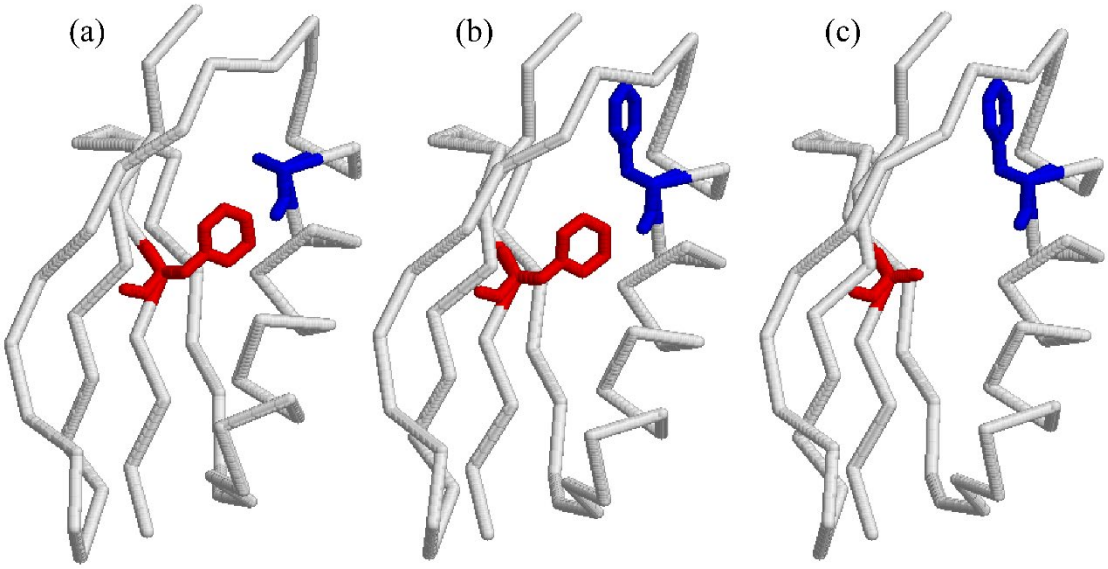


Figure D2. Backbone representation of GB1. Position 26 is in blue and position 52 is in red. (a) Wild-type, (b) A26F mutant, and (c) A26F and F52A mutant. Structures visualized using RasMol Ver. 2.7.2.1.1.

As seen in Figure D3, when an alanine replaces the phenylalanine in position 26, new long-range interactions are generated with residues 1 and 2 in the first β-strand and residue 19 in the second β-strand. However, upon mutation of the phenylalanine in position 52 to alanine, long-range interactions are lost to residue 3 in the first β-strand and residues 23, 26, and 27 in the α-helix.

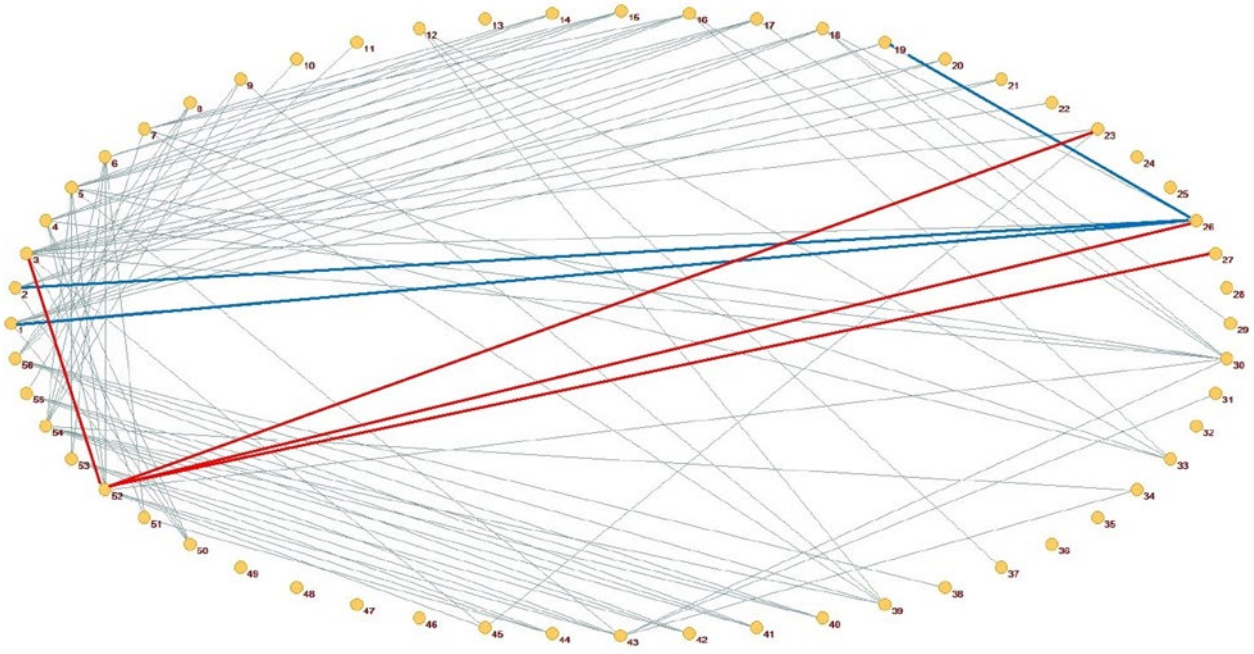


Figure D3. Long-range interaction network of GB1. Amino acids are filled circles connected by long-range interactions shown as lines. Long-range interactions gained by A26F mutation (dark blue) and lost by F52A mutation (red) are highlighted. Data plotted using Pajek64-XXL 4.08.

To determine the effect on protein structure and stability experimentally, far- and near-UV CD studies were performed. Figure D4 shows the results from the far-UV CD analysis. Both wild-type and variant GB1 show little variation in protein stability in the presence or absence of NaCl. The overall results show that the variant protein, independent of the buffer, exhibits a decrease in secondary structure when compared to the wild-type and is therefore less stable.

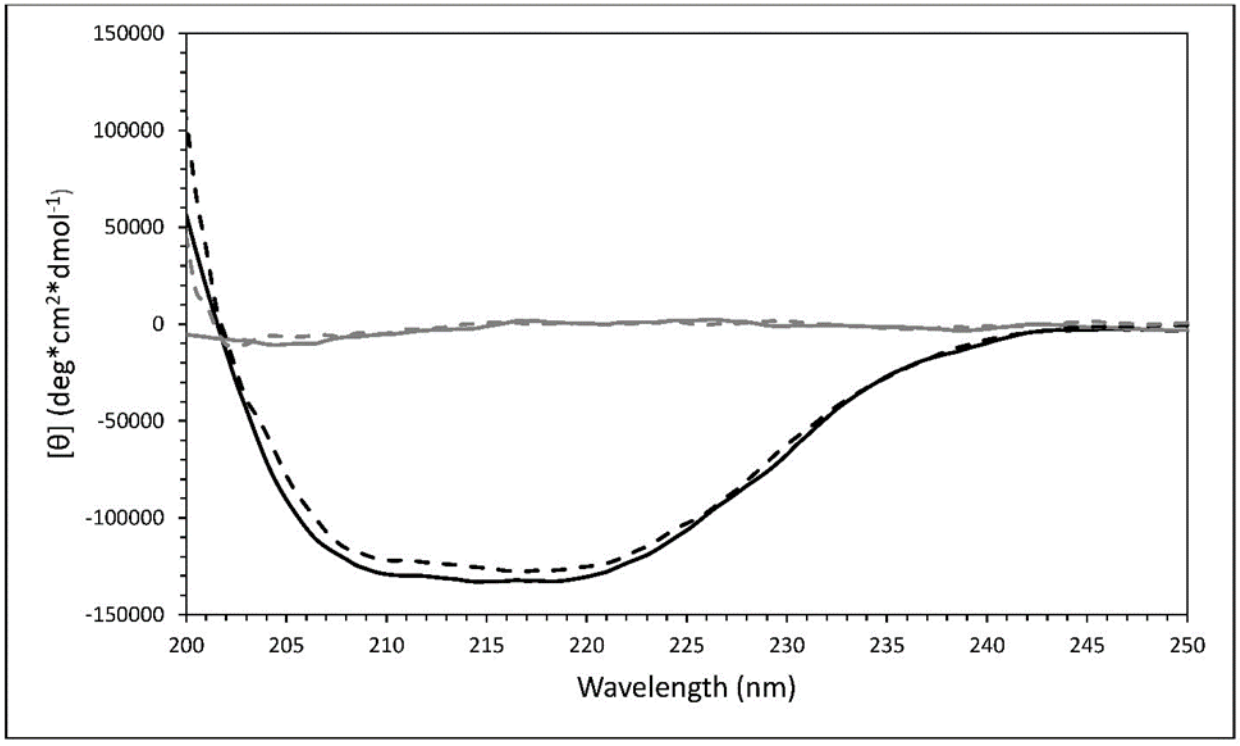


Figure D4. Far-UV CD of GB1. Wild-type and variant GB1 are shown in black and grey respectively. Samples were run at pH 7.0 (solid lines) and 3.0 M NaCl (dashed lines). Data plotted using Microsoft Excel 365.

Figure D5 shows the results from the near-UV CD analysis during thermal unfolding. Wild-type GB1 is stable at lower temperatures and loses its tertiary structure between 55 °C and 95 °C while variant GB1 adopts a disordered tertiary structure during thermal unfolding. The overall results show that the variant protein displays a decrease in tertiary structure during thermal unfolding when compared to the wild-type. Thermal unfolding was also monitored using fluorescence spectroscopy. As shown in Figure D6, the variant protein is less stable than the wild-type. These results parallel those obtained for far-UV CD and also help to determine whether the presence of salt makes the wild-type protein more stable.

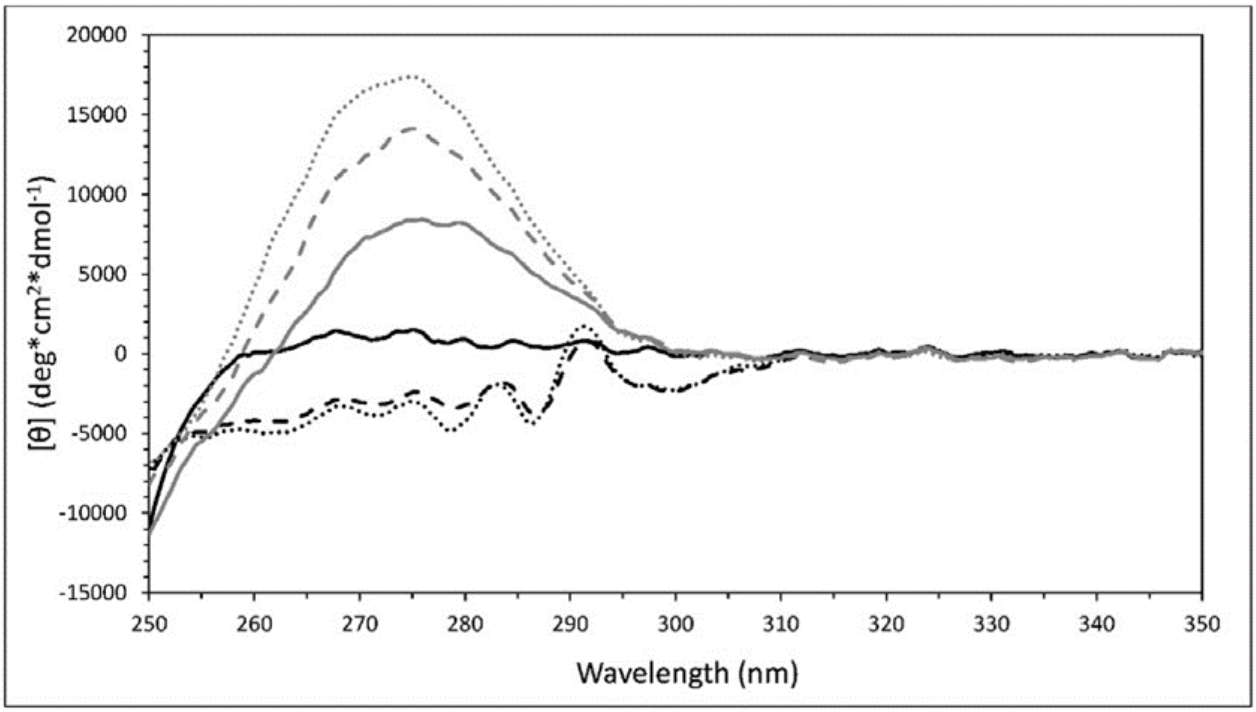


Figure D5. Near-UV CD of GB1 during thermal unfolding. Wild-type and variant GB1 are shown in black and grey respectively. Thermal unfolding was performed at pH 7.0. Temperature data is shown for 20 °C (dotted lines), 55 °C (dashed lines), and 95 °C (solid lines). Data plotted using Microsoft Excel 365.

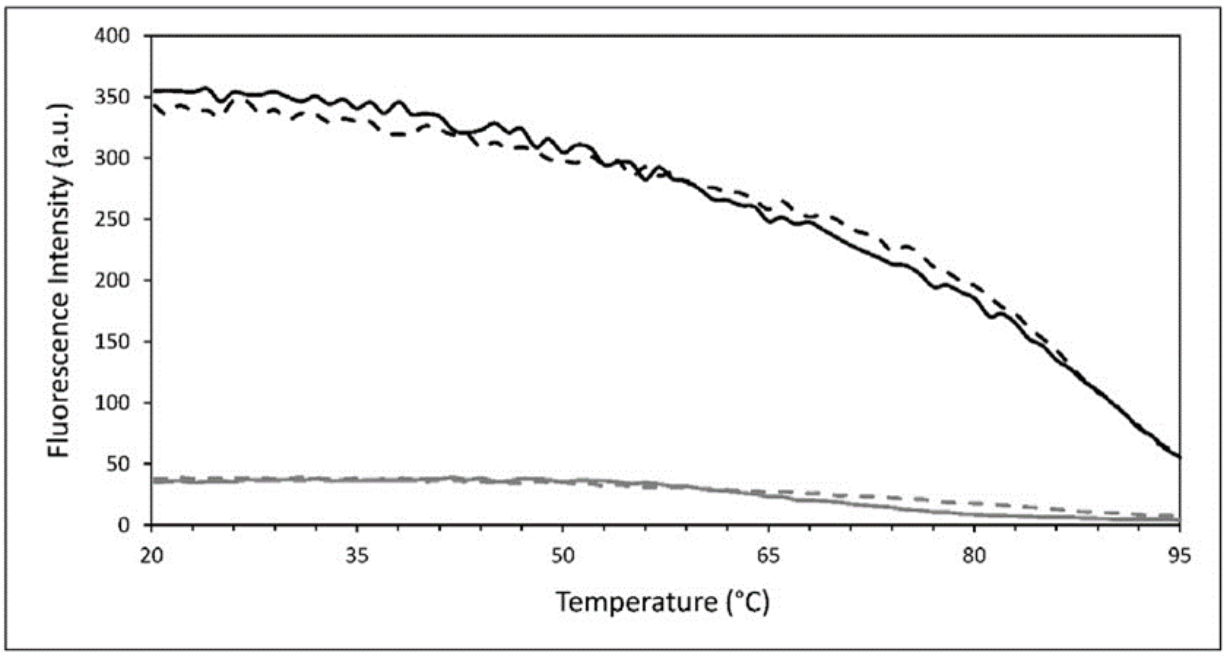


Figure D6. Fluorescence spectroscopy during thermal unfolding. Wild-type and variant GB1 are shown as black and grey lines respectively. Samples were run at pH 7.0 (solid lines) and 3.0 M NaCl (dashed lines). A protein concentration of 0.05 mg/ml was used with excitation and emission wavelengths of 295 nm and 350 nm respectively. Data plotted using Microsoft Excel 365.

The results of this study show that while theoretically a compensatory mutation would allow the protein to be more stable than a single mutation, experimentally this is not always the case. Many factors play a role in what mutations are allowed by nature. In light of our new findings, analysis of the axin dix domain (ADD), PDB code 1WSP, when structurally aligned with GB1, the phenylalanine in position 26 is located at the top of a helix. This perhaps helps accommodate the residue side chain. This means that while a phenylalanine residue is allowed to exist in this location there must room in the surrounding 3D environment.

ACKNOWLEDGEMENTS

Funding from the Virginia Space Grant Consortium Graduate Fellowship.

REFERENCES

D1. Pace C, and Scholtz J. (1998) A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. Journal*. **75**, 422-427.
D2. Holm L. and Rosenström P. (2010) Dali Server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, 545-549.
D3. Altschul S., Gish W., Miller W., *et. al.* (1990) Basic Local Alignment Search Tool. *J Mol. Biol.* **215**, 403-410.
D4. Edgar R. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797.Sayle R.A., Milner-White E.J. (1995) Rasmol – biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374-376.
D5. Sayle R.A., Milner-White E.J. (1995) Rasmol – biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374-376.
D6. Winn M.D., Ballard C.C., Cowtan K.D., *et. al.* (2011) Overview of the CCP4 suite and current developments. *Acta Cryst.* **67**, 235-242.
D7. Mrvar A. and Batagelj V. (2016) Analysis and visualization of large networks with program package Pajek. *Complex Adapt. Syst. Model.* **4**, 6.
D8. Collins J.C., Bedford J.T. and Greene L.H. (2016) Elucidating the key determinants of structure, folding, and stability for the (4β + α) conformation of the B1 domain of protein G using bioinformatics approaches. IEEE *Trans. on Nanobioscience.* **15**, 140-147.

VITA

JOHN T. BEDFORD II
jbedf001@odu.edu
Department of Chemistry and Biochemistry
Old Dominion University
Norfolk, VA 23529

EDUCATION

Old Dominion University	**Norfolk, VA**
PhD, Chemistry	May 2021
Old Dominion University	**Norfolk, VA**
Master of Science, Chemistry	August 2016
Old Dominion University	**Norfolk, VA**
Bachelor of Science, Chemistry	December 2012
Old Dominion University	**Norfolk, VA**
Bachelor of Science, Biochemistry	December 2012

PUBLICATIONS AND ORAL PRESENTATIONS

J.C. Collins, **J.T. Bedford**, L.H. Greene, *Elucidating the key determinants of structure, folding and stability of the (4β+α) conformation of the B1 domain of protein G using bioinformatics approaches*. IEEE Transactions on Nanobioscience, 2016, 15, 140-147.

J.T. Bedford, N. Diawara, J. Poutsma, L.H. Greene, *The nature of persistent interactions in two model β-grasp proteins reveals the advantage of symmetry in stability*. Journal of Computational Chemistry, 2021, 42, 600-607.

J.T Bedford, T. Mizukami, S. Liao, L.H. Greene, H. Roder, *Effects of ionic strength on the folding and stability of a halophilic protein, SAMP1.* (to be submitted)

D.S.K. Gamage, **J.T. Bedford**, N. Diawara, L.H. Greene, *Detecting dynamic contacts of temporal patterns with generalized Dirichlet processes.* (submitted)

A. Munyanyi, **J.T. Bedford**, J.C. Collins, N. Sori, and L.H. Greene, *Origins of the synucleins: orphans or superfamily members.* (in progress)

J.T. Bedford, J. Poutsma, L.H. Greene, *Persistence of key long-range interactions during GB1 unfolding simulations*, 96th VAS Meeting, Farmville, VA, May 23-25, 2018. (Presentation)

J.T. Bedford, L.H. Greene, *Further investigation into elucidating determinants of protein stability and folding in extreme environments*, Virginia Space Grant Consortium 2018 Student Research Conference, Norfolk, VA, April 11, 2018. (Presentation)

J.T. Bedford, L.H. Greene, J. Poutsma, *Computational exploration of the β-grasp fold for structure, function, and stability*, 255th ACS National Meeting, New Orleans, LA, March 18-22, 2018. (Presentation)

AWARDS

Virginia Space Grant Consortium Research Fellowship 2017 ($6000) and 2016 ($6000)
Old Dominion University Van Norman Award 2017 ($450)
Old Dominion University CIBA Fellowship 2016 ($4500) and 2015 ($4000)
ISBRA Travel Fellowship 2015 ($450)
Old Dominion University SEES Student Travel Award 2014 ($500)