Title
Combining in vivo and in silico screening for protein stability

Permalink
https://escholarship.org/uc/item/8df2j276

Author
Barakat, Nora Hisham

Publication Date
2007-01-01

Peer reviewed|Thesis/dissertation
Combining *in vivo* and *in silico* screening for protein stability

A Dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Chemistry

by

Nora Hisham Barakat

Committee in charge:

University of California, San Diego

Professor Randolph Hampton
Professor Katja Lindenberg
Professor Xuong Nguyen-Huu

San Diego State University

Professor John Love, Chair
Professor Anca Segall
Professor William Stumph

2007
The Dissertation of Nora Hisham Barakat is approved, and it is acceptable in quality and form for publication on microfilm:

University of California, San Diego

San Diego State University

2007
DEDICATION

I dedicate this dissertation to the following people:

The memory of my father, Hisham Barakat, who emphasized the importance of education.

My sweet mother, Souad Barakat, who has been my role-model of hard work.

My lovely sisters, Dr. Nermeen, Nesreen, and Neveen Barakat, who have been my anchors through my graduate school and my entire life.

and my dear husband, Hany El-Saidany, who has been proud and supportive of my work.
TABLE OF CONTENTS

Signature Page ................................................................. iii
Dedication ................................................................. iv
Table of Contents ............................................................ v
List of Figures ................................................................. vii
List of Tables ................................................................. ix
Acknowledgements ........................................................... x
Vita ................................................................. xii
Abstract of the Dissertation .................................................. xiv

I. Introduction to protein stability ........................................... 1
   1.1. An integrated view of protein structure, dynamics, and function .... 1
   1.2. Relationship between protein stability and protein function ....... 4
   1.3. Relationship of protein flexibility to thermostability ......... 6
   1.4. Industrial Enzyme Design ................................................. 8
   1.5. Protein misfolding and disease ......................................... 10
   1.6. Protein – based drug design ............................................. 12

II. Introduction to rational design and directed evolution ............... 17
   2.1. Abstract ................................................................. 17
   2.2. Rational design and protein stability .................................. 18
   2.3. Directed evolution and protein stability ............................. 21
   2.4. Emerging computational and in vivo approaches for engineering protein libraries ............................................................. 37
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Chimeric Construct</td>
<td>61</td>
</tr>
<tr>
<td>3.2</td>
<td>Chimeric Vector and Construct</td>
<td>62</td>
</tr>
<tr>
<td>3.3</td>
<td>Gβ1-WT Structure and Positions of Specific Residues</td>
<td>65</td>
</tr>
<tr>
<td>3.4</td>
<td>Correlation between protein stability and proteolysis resistance</td>
<td>70</td>
</tr>
<tr>
<td>3.5</td>
<td>Gβ1-WT structure and Mutant variants</td>
<td>76</td>
</tr>
<tr>
<td>3.6</td>
<td>Far UV Circular Dichroism Scans for Gβ1-WT and Mutant Variants...</td>
<td>78</td>
</tr>
<tr>
<td>3.7</td>
<td>Circular Dichroism Thermal Unfolding Curves for Gβ1-WT and Mutant Variants...</td>
<td>79</td>
</tr>
<tr>
<td>3.8</td>
<td>Chimeric Vector and Construct for Stable Test Variant...</td>
<td>81</td>
</tr>
<tr>
<td>3.9</td>
<td>Chimeric Vector and Construct for Unstable Test Variant...</td>
<td>82</td>
</tr>
<tr>
<td>3.10</td>
<td>Reporter Gene Expression for Gβ1 Variants...</td>
<td>85</td>
</tr>
<tr>
<td>3.11</td>
<td>Correlation between Bacterial Growth and Protein Thermal Stability...</td>
<td>86</td>
</tr>
<tr>
<td>3.12</td>
<td>Representative [1H, 15N] HSQC spectra...</td>
<td>90</td>
</tr>
<tr>
<td>3.13</td>
<td>1D Projections of the 2D HSQC Spectra...</td>
<td>91</td>
</tr>
<tr>
<td>3.14</td>
<td>Screening using chromatographic assays...</td>
<td>94</td>
</tr>
<tr>
<td>3.15</td>
<td>Correlation between β-galactosidase activity and protein thermal Stability...</td>
<td>96</td>
</tr>
<tr>
<td>3.16</td>
<td>Model of the Chimeric Construct...</td>
<td>103</td>
</tr>
<tr>
<td>4.1</td>
<td>Correlation between Bacterial Growth and Protein Thermal Stability...</td>
<td>120</td>
</tr>
<tr>
<td>4.2</td>
<td>In vivo screening scheme to identify stable MonB variants...</td>
<td>121</td>
</tr>
<tr>
<td>4.3</td>
<td>Far UV Circular Dichroism Scans for Gβ1-WT and Variants...</td>
<td>126</td>
</tr>
</tbody>
</table>
Figure 4.4. Circular Dichroism Thermal Unfolding Curves for Gβ1-WT and Mutant Variants

Figure 4.5. Correlation Between protein thermal stability and ORBIT energy calculation

Figure 4.6. Melting temperatures of the selected variants were plotted against their GMEC found in DEE search
# LIST OF TABLES

Table 1.1. The biophysical properties of proteins that can be optimized to obtain desired therapeutic outcomes .................................................. 14

Table 3.1. DNA oligomer sequences .................................................. 58

Table 3.2. Amino Acid Sequence for Gβ1-WT and the ORBIT Selected Sequences for MonA and MonB ................................................. 66

Table 3.3. Biophysical properties of the Gβ1 variants ................................. 67

Table 3.4. Amino Acid Sequences for Gβ1-WT and mutant variants ............... 73

Table 4.1. Oligonucleotides corresponding to the MonB Library gene ............. 114

Table 4.2. Amino acids at positions 23, 27, and 45 for Gβ1-WT and its variants 128

Table 4.3. Amino acids sequence, the global minimum energy solution, and subcategories of energies, for MonB mutant variants ...................... 132
Chapter 2, in part is a reprint of the materials as it appears in Nora H. Barakat, John J. Love, “Molecular Diversity in Engineered Protein Libraries”. Submitted to and accepted by *Current Opinion in Chemical Biology*.


Chapter 4, in part is a reprint of the materials as it appears in Nora H. Barakat, Nesreen H. Barakat, John J. Love, “Characterization of protein stability using experimental and computational methods”. Manuscript in final stage of revision.

I would like to thank the person who has had the most positive and significant effect on my graduate career, my advisor Professor John J. Love. I am very grateful for his time and effort to teach me and guide me through my scientific development process. I would like to thank Dr. Stumph for his guidance; and Kathy McNamara, for scientific insight. I would like to thank my colleagues Jennifer Reichert, Ushma Shukla, Mina Stemm, Shadi Saedinia, Youly Ly, Nesreen Barakat, Lisa Carmody, Sarah Hamilton, Mackenzi Patterson, Aditi Apte, and Mario Navarro for their great help and contributions. I would like to thank my Mother who taught me the importance of education and believed in me. I would like to thank my sisters Dr. Nermeen, Nesreen and Neveen for their endless support and for making me proud to have such great sisters. I wish success and the best of luck to them in pursuing their
dreams. Last but not least, I would like to thank my great husband, Hany, who supported me and motivated me throughout the entire process.
VITA

Education

1997       B.S., Biomedical Science, University of Qatar

2000       Certificate of Achievement, Computer Science and Information Systems
            Grossmont Community College

2006       M.A., Chemistry, San Diego State University

2007       Ph.D., Chemistry, 
            University of California, San Diego and San Diego State University (Joint 
            Doctoral Program)

Publications

Elements of Transcriptional Machinery to Enhance Protein Stability”. ”. Journal of 

experimental and computational methods”. Manuscript in final stage of revision.

Nora H. Barakat, John J. Love, “Molecular Diversity in Engineered Protein Libraries”. 
Submitted to and accepted by Current Opinion in Chemical Biology.

Patent Applications

Stability”. A provisional patent application was submitted by the SDSU Technology 
Transfer Office to the US Patent Office in the Spring of 2006
Awards

2006 recipient of an Arne N. Wick Pre-doctoral Research Fellowship from the California Metabolic Research Foundation

2005 recipient of a Don Eden Award for Excellence in Student Research in field of Biophysics and Physical Biochemistry, California State University Program for Education and Research in Biotechnology Symposium (CSUPERB)
ABSTRACT OF THE DISSERTATION

Combining in vivo and in silico screening for protein stability

by

Nora Hisham Barakat

Doctor of Philosophy in Chemistry

University of California, San Diego, 2007

San Diego State University, 2007

Professor John J. Love, Chair

In any protein design endeavor, thermal stability is an important issue. Increasing protein stability is currently one of the main goals of the field of protein design as well as the biomedical industry. It is important for developing stable protein-based therapeutics and ultimately enhances our understanding of the principles of protein folding. The overall goal of my project is to develop and combine in vivo and in silico screening and analysis to select stable proteins from a large combinatorial library of proteins. Our in vivo screening method is based solely on the stability of the test proteins, and is independent of protein function. In addition, the mutant variants
obtained from the *in vivo* screen were further analyzed and screened *in silico* with the ORBIT protein design package.

In the *in vivo* ‘chimeric’ screen a target protein is expressed as a fusion protein with a DNA binding domain fused at the N-terminus and a transactivation domain at the C-terminus. This three-component fusion protein acts as a functional transcription factor for a reporter gene. If the central target protein is relatively unstable it is more flexible and thereby allows the DNA binding domain and the transactivation domain to achieve an optimal orientation for transcriptional activation. Conversely, if the central target protein is highly stable it is too rigid to allow the DNA binding domain and the transactivation domain to achieve the proper orientation for transcriptional activation and thus is a less efficient transcription factor.

To supplement the results obtained from the *in vivo* screen we also use the ORBIT suite of protein design algorithms to computationally generate stability variants. These algorithms enable one to mutate specific residues while keeping the remainder wild-type. Discrete sets of side chain rotamers are considered in the calculation and an empirical potential energy function is used to calculate the energies between rotamer pairs. The energies are subjected to a very fast side chain selection algorithm, based on the Dead-End Elimination theorem, which ultimately returns the global minimum energy conformation. We successfully demonstrated the ability of the *in vivo* chimeric stability screen and *in silico* ORBIT screen to generate novel protein variants with improved thermal stability.
I. INTRODUCTION TO PROTEIN STABILITY

1.1 An integrated view of protein structure, dynamics, and function

Living cells utilize proteins to perform a variety of functions. For example, enzymes are responsible for catalyzing reactions in many important biological processes. Enzymes are highly efficient catalysts which accelerate reactions by many orders of magnitude (Radzicka and Wolfenden 1995; Neet 1998). For more than a century it has been known that the activity of enzymes is directly correlated to structure and that a “lock and key” and “induced fit” mechanism was initially hypothesized (Kraut 1988). The hypothesis suggested that the three-dimensional shape of the substrates acts like a key and fits exactly into the active site of its target enzyme and that interaction between substrate and enzyme are based primarily on physical shape interactions. Such a view is incomplete as it does not fully explain the detailed mechanism of enzyme catalysis. More recent studies provide evidence indicating that the dynamic motion of enzymes plays a major role in the catalytic function of enzymes (Agarwal 2005; Agarwal 2006). Protein motions are implicated in events such as binding of substrate or cofactor, and product release. These events are often mediated by conformational fluctuations of active site residues and/or loops in the vicinity of the active site.

Currently, there are many studies investigating the possible correlation between protein structure, dynamics, stability and function of several enzymes (e.g. Cyclophilin
A (Cyp A), dihydrofolate reductase (DHFR), Liver alcohol dehydrogenase (LADH)) (Agarwal 2005; Agarwal 2006). Changes in protein dynamics in response to ligand binding that potentially induce conformational changes; have been explored using experimental and computational approaches. Techniques including X-ray crystallography, X-ray scattering, fluorescence, nuclear magnetic resonance (NMR), and hydrogen-deuterium exchange continue to provide interesting details about the dynamic movement of different regions of protein enzymes. Theoretical and computational methods such as molecular dynamics simulations, and hybrid quantum/classical dynamics methods are complementing the knowledge gained from experimental studies.

To explore the potential link between protein dynamics and enzymatic catalysis the enzyme Cyclophilin A (CypA) has been investigated using both biochemical and computational methods. CypA is a single polypeptide chain with 165 amino acids. Its molecular architecture consists of an eight-stranded anti-parallel β-barrel with hydrophobic residues forming a core at the center and the active-site located on one face of the molecule. X-ray crystallographic studies revealed several flexible loop regions on the surface of the protein. CypA catalyzes the isomerization of peptidyl-prolyl amide bonds that are N-terminal to proline residues in a wide variety of protein substrates, such as capsid protein (CA) from human immunodeficiency virus type 1 (HIV-1). NMR studies (i.e. $^{15}$N relaxation and two-dimensional $^1$H-$^{15}$N heteronuclear exchange) of CypA have suggested a link between internal protein dynamics and the substrate isomerization step. Conformational fluctuations within the active-site of CypA were
detected and the frequency of which was on the time-scale of the actual reaction (hundreds of microseconds).

Computational modeling of CypA has revealed a variety of internal dynamics that may be linked to catalytic activity and range from femtosecond to microsecond and longer time scales. On one side of the range there are fast motions (vibrations), which occur on the femtosecond to nanosecond time-scale and consist primarily of bonds and angle movements. These fast motions presumably impact the hydrophobic and hydrophilic interactions that occur between the enzyme and substrate. On the other side of this range there are conformational fluctuations occurring on the microsecond and longer time scale. These slower motions are called breathing motions and span a large portion of the protein. The detailed characterization of protein motions indicates that protein dynamics plays a major role in protein functions.

In addition to protein dynamics there is increasing experimental evidence (primarily driven by high-field heteronuclear NMR) that overall structural stability, or lack there of, may also play an important role in protein function. The overall dynamic regime of protein backbone and side chain atoms correlate to intrinsic stability and thus it reasons that protein flexibility (or plasticity) may also correlate to function. Therefore the ability to tailor a designed protein’s intrinsic stability/flexibility to a desired function may ultimately enhance protein design endeavors. Furthermore, an in vivo screen for protein stability also enables the rapid analysis of large portions of a particular fold’s sequence space.
1.2 Relationship between protein stability and protein function

Functional proteins are normally well folded because of two reasons: 1) proteins fold to minimize free energy and 2) they fold into specific structures to be able to recognize and bind either other macromolecules or ligands and/or catalyze reactions. In general, protein residues that contribute to catalysis or binding are not optimal for protein stability. Minimization of folding free energy normally leads to a well-packed hydrophobic core surrounded by a hydrophilic surface. Maximizing function leads sometimes to active-sites clefts with buried hydrophilic amino acids, where hydrophobic amino acids are exposed to water. The balance between stability and function are important for maintaining the structural integrity of the protein as well as for catalysis and ligand binding. Several studies have focused on the relationship and balance between stability and function. In one approach, Shoichet et al. reported that protein residues that contribute to the function of the protein (e.g. catalysis or ligand binding) are not necessarily optimal for protein stability (Shoichet, 1995). The relationship between stability and function was investigated by substituting key active-site residues and measuring the changes in stability and activity. Using T4 lysozyme, an enzyme that is well-characterized structurally and mechanistically, five residues that have catalytic and substrate binding roles were mutated. Thermodynamic stability and kinetic activity measurements revealed that enzymatic activity was reduced while thermal stability was increased. In a similar approach, catalytic residues of the enzyme β-lactamase were mutated and the activity of the enzyme decreased by $10^3$–$10^5$-fold compared to wild-type (Beadle, 2002). Concomitantly, many of these substitutions increase the stability of the enzyme significantly (i.e., by 4.7 kcal/mol). These results
suggest that preorganization of functional residues in the active site has come at a cost to enzyme stability. In proteins of unknown function, the presence of such destabilized regions may be indicative of the presence of an enzyme binding site. These results further demonstrate the potential correlation between protein stability and function.
1.3 Relationship of protein flexibility to thermostability

Protein thermostability arises from the combined effect of many forces, such as electrostatic interactions, hydrophobic interactions, and hydrogen bonds, which invariably lead to overall decrease in protein flexibility. In general, the flexibility of proteins is reduced when thermostability is increased. Protein flexibility is essential for the regulation and function of several proteins. For example, substrate binding in yeast hexokinase requires conformational changes (Bennett, 1978). The yeast hexokinase crystallized as a complex with glucose has a conformation that is dramatically different from yeast hexokinase in the absence of glucose. Comparison of the high-resolution structures shows that one lobe of the molecule is rotated by 12 degrees relative to the other lobe, resulting in movements of as much as 8 Å in the polypeptide backbone and closing the cleft between the lobes into which glucose is bound. The conformational change is produced by the binding of glucose and is essential for catalysis. In some cases stability can be important on test protein through the introduction of a properly designed disulfide bond. Perry et al. introduced an intramolecular disulfide bond into T4 lysozyme and demonstrated that the protein is significantly more stable than the wild-type protein and correspondingly reduced its flexibility (Perry, 1986).

Závodszky et al. examined the question of whether the low activity of certain thermophilic enzymes at room temperature may be caused by restricted conformational movements of the protein (Zavodszyk, 1998). In other words, is the activity of such an enzyme related to its flexibility? As a model system they chose the enzyme, 3-isopropylmalate dehydrogenases (IPMDH), which has two variants with significantly different thermal stabilities. One originates from a mesophilic organism and the other
from a thermophilic organism. Hydrogen/Deuterium exchange (H/D exchange), followed by the FT-IR, was used to compare enzyme flexibilities and explore the potential correlation with thermodynamic stability (Zavodszky, 1998). Analysis of the resulting data clearly showed that the thermophilic IPMDH is significantly more rigid at room temperature than its mesophilic counterpart. However, the two enzyme variants showed very similar flexibility at temperatures near each of their activity optima. Their findings demonstrated that enzyme activity, conformational flexibility, and thermostability are closely correlated. The balance between these factors is important for maintaining structural integrity while allowing the protein to functional properly at the appropriate temperature.
1.4 Industrial Enzyme Design

Enzymes have been used since ancient times in the making of food products, such as cheese, wine, and vinegar. Additionally, they are important in manufacturing products such as leather. All these processes depend on enzymes produced from natural sources (i.e., microorganisms). In addition, the re-design of enzymes has received significant attention from the detergent industry. Recombinant technology and improvements in protein expression and purification has further advanced the enzyme manufacturing processes that lead to large scale production of enzymes. Moreover, the latest developments in protein engineering, rational design and directed evolution further revolutionized the development of the enzyme industry. These modern biotechnologies enabled the development of tailor-made enzymes with novel functions, adapted to harsh industrial processes.

Worldwide use of industrial enzymes has increased in recent years (Godfrey T 1996; M 2000). The enzyme industry is dominated by detergents, starch, fuels, leather and dairy products. These industries require enzymes with special properties, such as enhanced stability, which improves an enzyme’s ability to function under a wide range of temperatures and pH. Enzyme stability is an important parameter for industrial applications as it determines the economic feasibility of applying an enzyme in an industrial process. Additionally, high stability is considered an economic advantage because it reduces enzyme turnover and allows for their use in high temperatures, which also usually enhances reaction rates. For example, Cherry et al. described the evolution of a peroxidase from the ink cap mushroom Coprinus cinereus (CiP) to allow its use as a detergent additive (Cherry, Lamsa et al. 1999). Under mild conditions, this peroxidase
is able to catalyze the oxidation of dyes that leach out of colored clothing during the wash cycle, rendering them colorless and effectively preventing transfer of the dye to other clothes. While CiP is unstable in bleach-containing detergents, the authors have successfully engineered the protein such that it is able to withstand the highly alkaline (pH 8.5–10) and oxidative (H$_2$O$_2$=5 mM) detergent milieu.

The increased demand for enzymes in food, chemical industries has spurred the search for proteins with improved intrinsic stability. Implementation of directed evolution and rational design has improved the specific properties (e.g. stability) of a wide variety of enzymes, offering a robust means to transfer enzymes out of the laboratory and into industrial processes.
1.5 Protein misfolding and disease

Protein molecules have a finite tendency to either misfold or to fail to maintain their correctly folded states, under some circumstances. A “new view” of protein folding now incorporates ideas on protein misfolding and disease (Smith 2003). It is obvious that the structure of a protein and its ability to carry out its correct function are tightly linked and that small structural defects can lead to a number of protein folding diseases. Examples of such genetic diseases include cystic fibrosis and sickle cell anemia, which are caused by a single residue mutation, rendering the mutated protein, incapable of its normal function. More recently a number of diseases have been linked to protein misfolding and lead to the build up of insoluble protein plaques in the brain or other organs. These diseases include prion diseases such as bovine spongiform encephalopathy (BSE) and its human equivalent Creutzfeld-Jakob disease (CJD). Other examples include Alzheimer's disease, Parkinson's disease and type II (non-insulin dependent) diabetes. In all of these cases the protein isolated from the insoluble plaques is found to be coded for by the host genome although the natural functional roles are often not known.

Protein folding, structure, stability and the exact information in the sequence of the protein that is necessary to fold a specific protein to its biologically active state are still not fully understood. However, at the present time, protein folding is an extremely active field of research encompassing aspects of biology, chemistry, biochemistry, computer science, and physics. The complexity of the protein folding problem requires the ability to search large volume of sequence space. Two powerful approaches to help
understand protein folding and stability are combinatorial library selection (directed evolution) and computational protein design (rational design).
1.6 Protein – based drug design

Traditional drug discovery efforts normally focus on small organic molecules such as compounds that are obtained from either natural sources or produced using synthetic chemistry methods. In recent years, the use of protein-based drugs as mainstream therapeutic agents has risen significantly. Examples of new technologies that have led to advances in the field of protein-based drugs include, for example, recombinant DNA technology, molecular biology methods, techniques for protein purification, amino acid sequencing, and DNA sequencing. In particular, site-directed mutagenesis has been used to make specific changes in protein enzymes and protein hormones which has rendered these drugs more useful in clinical applications. In addition, the decoding of the human genomic and proteomic maps has lead to discoveries in molecular medicine that have contributed to the development of protein-based drugs.

The list of protein-based therapeutics has grown rapidly since 1982 when the FDA approved the first recombinant drug, insulin, for human use (Ho, 2003; Schluter, 1982; Lotz, 1982). Peptide and protein based drugs are continuing to develop and thus many protein based drugs are under pre-clinical and clinical development in the pharmaceutical industry. Several newly engineered protein and peptide therapeutics have successfully completed clinical trials as for example, Genentech’s Somavert (pegvisomant), an antagonistic variant of human growth hormone (hGH); Trimeris’s Fuzeon (enfuvirtide), an inhibitor of HIV fusion derived from the viral protein gp41; Amgen’s Aranesp (darbepoetin alfa), a hyperglycosylated variant of erythropoietin
(EPO); and Abbott’s Humira (adalimumab), a fully human monoclonal anti-TNFα antibody (Marshall, Lazar et al. 2003).

There has been rapid growth in the protein-based drug market this century. In 2003, sales for protein based drugs were nearly $40 billion dollars, and the annual growth rate for these drugs was 12.2%. This market is expected to reach $71 billion in 2008. Replacement proteins (e.g., recombinant insulin and growth hormone) constitute the majority of the protein-based drug market, and are expected to rise at an average annual growth rate of 9.8% from $31.4 billion in 2003 to $50 billion in 2008. Monoclonal antibodies and fusion proteins are expected to rise from $8.5 billion in 2003 to $20.1 billion in 2008 (Cyran 2004).

In spite of the increasing number of engineered protein therapeutics currently being developed, tested in clinical trials and marketed for use, the pharmaceutical industry is facing several challenges. The majority of which are due to the biophysical and chemical properties of protein therapeutics such as: stability, solubility, receptor binding affinity and specificity, oligomerization state, chemical modifications, post-translational modifications, sequence diversity and conformational states. Table 1.1 is a summary of the biophysical properties of proteins that can be optimized to obtain desired therapeutic outcomes. The physicochemical properties of protein therapeutics greatly affects their performance during development, manufacturing and clinical use (Marshall, Lazar et al. 2003).
Table 1.1 The biophysical properties of proteins that can be optimized to obtain desired therapeutic outcomes. Recreated from (Marshall, Lazar et al. 2003).

<table>
<thead>
<tr>
<th>Property</th>
<th>Enable discovery</th>
<th>Mechanism of action</th>
<th>Pharmacokinetics</th>
<th>Immuno-</th>
<th>Route of administration</th>
<th>Cost of goods</th>
<th>Shelf life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stability</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Solubility</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Receptor binding affinity</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Oligomerization state</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical modifications</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Posttranslational modifications</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Sequence diversity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Conformational state</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>
A major biophysical property that has been explored for my dissertation is protein stability. In general, protein therapeutics require as high a level of stability as possible (without compromising function). Stability is important to retain the activity of the protein drug during purification, formulation, storage, administration and \textit{in vivo} function. For example, Willuda et al. demonstrated the important role that protein folding and stability play when attempting to improve the biophysical properties of antibody fragments of potential clinical benefit (Willuda, Honegger et al. 1999). An unstable and poorly expressed anti-epithelial glycoprotein-2 single-chain Fv (scFv) was converted to a well-expressed and very stable humanized antibody fragment with the same antigen specificity. The epithelial glycoprotein-2 is abundantly expressed on many solid tumors and is a suitable target for antibody-based therapy. Despite its high affinity ($K_D = 3.9 \times 10^{-9} \text{ M}$), the original antibody fragment failed to significantly accumulate at the lung tumor in mice, and it was hypothesized that the failure was due to insufficient thermal stability. To overcome this limitation, the antigen-binding residues of the scFv fragment were grafted onto the framework of the highly stable and well-folded single-chain Fv. These modifications resulted in increased serum stability at $37^\circ \text{ C}$ and also significantly improved expression behavior while retaining the antigen specificity and affinity of the parent scFv. In mice a scFv which was radiolabeled exhibited efficient accumulation at lung tumor. This example demonstrates the important role that thermal stability plays in determining the total therapeutic efficacy of protein drugs \textit{in vivo}.

In addition, protein solubility also plays a major role in the drug development, cost and drug feasibility. There is a good correlation between protein stability and solubility. Stable and well-folded proteins are generally more soluble than unstable,
unfolded proteins which are more highly prone to aggregation. Therefore, protein stability also correlates to the pharmaco-kinetics and immunogenicity parameters of protein drugs. Due to the important role stability plays in protein function, and therefore therapeutics, much effort has been made to find effective means to increase protein stability. Thus far, successful attempts have been made through the utilization of a number of techniques such as rational protein design and directed evolution. A main goal of my dissertation research was to add to this growing body of techniques by developing novel screening methods designed to increase the stability of proteins
II. INTRODUCTION TO RATIONAL DESIGN & DIRECTED EVOLUTION

2.1 Abstract

The increased use of enzymes and other proteins in the chemical and pharmaceutical industries has generated significant interest in designing proteins with improved stabilities. The protein engineer's toolbox is expanding and the number of successful examples of engineered protein stability is increasing. Still, the design and selection of thermo-stable protein is not a standard process. Engineering increased stability into different proteins is complicated due to the large variety of structural parameters and variables that affect stability. However, two different but complementary technologies have been applied to accomplish this goal, 1) rational design and 2) directed evolution. Both approaches have met with success yet each has caveats and limitations. Studies using either one or a combination of both these distinct approaches are reviewed in this section.
2.2 Rational design and protein stability

Rational design relies on structural and mechanistic knowledge of protein folding combined with human input and expertise. It is information-intensive and requires both the availability of the structure of the protein and knowledge of the relationship between sequence, structure, stability, and function. The progress in solving protein structures by NMR spectroscopy and X-ray crystallography has significantly increased the number of structures stored in the Protein Data Base (PDB). Furthermore, if there is no structural data available in the PDB for a protein of interest, the structure of a homologous protein can be used as a model.

The conventional approach used for rational design is based on precise changes in amino acid sequences which are invariably introduced using site-directed mutagenesis. The amino acid mutations are based on a detailed knowledge of the protein’s structure, function and mechanism. This approach is useful for improving enzyme characteristics such as altering an enzyme’s mechanism, substrate specificity, or co-factor specificity. This traditional approach requires the mutated enzymes to be expressed and purified and the novel functional properties properly assessed after mutagenesis (Chen 2001). Such an approach can be tedious and expensive and successful examples of improved protein stability using this method are fairly rare (Zhang, Liu et al. 2005).

Another novel approach for the rational design of proteins is achieved through computational design and engineering. Over the past ten years there has been tremendous success in the area of computational design (Korkegian, 2005; Ashworth, 2006). A power of computational design comes from the ability to surpass the
combinatorial and physical limitations inherent to laboratory-based high-throughput or trial-and-error methods. Computational design techniques use protein design algorithms to perform *in silico* assessment of mutant amino acid sequences. By taking advantage of the speed of computers, these methods allow a large number of sequences to be assessed with a molecular mechanics force field. Searching vast areas of sequence space increases the possibility of finding proteins with novel functions (e.g. binding) or improved properties (e.g. higher thermal stability). There are two elements required for computational design. One element is accurate scoring functions to rank sequences as a function of calculated energy and the other is high-speed optimization methods designed to rapidly find the best sequences from the enormous combinatorial search space (Dahiyat 1999). Protein design software has thus far been used successfully to design stable proteins (Malakauskas and Mayo 1998), redesign substrate binding pockets (Schrag, Vernet et al. 1995), creating a novel protein fold (Socolich, Lockless et al. 2005), and design catalytic activity into a bacterial receptor (Dwyer, 2004).

Several computational design strategies have shown great improvements in the overall stability of target proteins by optimizing important intramolecular interactions. For example, computational design has been used to optimize packing interactions and hydrophobic burial in the protein core (Butterfoss and Kuhlman 2006). Optimizing secondary structure propensity, hydrogen bonds and electrostatic interactions can also improve protein stability substantially (Malakauskas and Mayo 1998). Several studies have applied these principles to clinically relevant proteins as for example human growth hormone (hGH) (Filikov, Hayes et al. 2002). Recombinant
hGH is used worldwide for the treatment of pediatric hypopituitary dwarfism and for children suffering from low levels of hGH. It has limited stability in solution, and because of poor oral absorption is administered by injection and typically several times a week. Development has therefore focused on more stable or sustained-release formulations and alternatives to injectable delivery that would increase bioavailability and make it easier for patients to use. Filikov, et al. attempted to redesign hGH computationally to improve its thermostability. A Protein Design Automation (PDA) method (Dahiyat and Mayo 1996), which incorporates the dead-end elimination (DEE) algorithm (Goldstein 1994; Desmet, De Maeyer et al. 1997), was used for the above purposes. Using a rotamer descriptions of the side-chains, an optimal sequence for a fixed backbone structure can be found by screening all possible sequences of rotamers, in which each backbone position can be occupied by each amino acid in all possible rotameric states. The resulting variants contained six to 10 mutations, exhibited enhancement in thermal stability of up to 16°C, have improved pharmacokinetics, better storage properties and are more amenable to use in alternative delivery systems and formulations, thus providing added convenience and improved patient compliance. These results demonstrate the utility of using an optimized energy function in automated protein design. Computational design technology has been applied to numerous other systems including important pharmaceutical and industrial proteins and has a demonstrated record of success in protein optimization.
2.3 Directed evolution and protein stability

Evolution has given rise to a fantastic array of different proteins. In the field of protein design there is great interest in generating proteins with novel functions and thus there is a strong desire to establish a system capable of obtaining such molecules. Directed evolution has been used to develop a system that mimics the course of natural protein evolution. Such a system can be used to rapidly perform rounds of selection and diversification while applying selective pressure for function, solubility, stability or protease resistance. Directed evolution approaches contrast with more conventional (rational) methods involving iterative computer design and site-directed mutagenesis that are sometimes rendered ineffective because of the absence of the detailed structural and mechanistic information required.

Directed evolution (also called evaluative biotechnology or molecular evolution) involves random mutagenesis of the gene encoding the protein of interest, or recombination of gene fragments (also called DNA shuffling), to generate libraries of genes that code for mutant variants of the target protein. The libraries are then screened using high-throughput technologies for identification of improved variants, both approaches (random mutagenesis or DNA shuffling) are well established and have been used throughout the molecular biology and biochemistry fields. Directed evolution has become a valuable tool for both industrial and pure research applications. For industrial applications, enzymes are engineered in order to produce suitable biocatalysts with high catalytic activity and stability in industrial environments. For research applications, methods have been established to quickly engineer new enzymes for certain catalytic steps and with improved properties (e.g.
Combinatorial methods (directed evolution) used to select for stable proteins are based on three main elements (Magliery and Regan 2004). The first element is the construction of a library of protein variants. This library can either be expressed in cells (usually bacteria), displayed on the surface of filamentous phage, displayed on stalled ribosomes or covalently linked to mRNA through the action of the compound puromycin. Puromycin is an antibiotic that inhibits translation by causing premature translational termination. It does this by binding to the A site of the ribosome in place of tyrosyl-tRNA during transpeptidation. The second element of directed evolution is the selection or screening of the library for proteins with desired properties. Cells which express proteins of interest are distinguished either by cellular survival (selection) or a phenotypic trait (screen). Many selection and screening methods for protein stability are based on genetic screens. For example, tryptophan synthase function is required for survival on tryptophan-free medium and the lac repressor prevents transcription of β-galactosidase which can assayed by hydrolysis of a chromogenic galactosidase. The third element is identification of selected variants by isolation of DNA from cells or phage, or RT-PCR of RNA linked to the selected protein.

The following are examples of designed proteins with improved thermal stability using different screening and selection strategies.

Selection of stable proteins with phage display coupled with in vitro proteolysis: In general, most proteins that are to be purified must be sufficiently
resistant to proteolysis. A number of researchers have shown that proteolytic resistance can be used directly as a marker of foldedness and stability.

Phage display is a selection tool that has been successfully applied to investigate protein stability. In adapting phage-display from the more common selection for binding affinity, investigators coupled phage-display with proteolysis to select for proteins with increased stability. Phage display coupled with proteolysis selects for stably folded proteins and is thus independent of binding ability. It links the increased protease resistance of stabilized protein variants with the infectivity of the phage. It was first demonstrated in 1998 by two research groups (Kristensen and Winter 1998; Sieber, Pluckthun et al. 1998). Their systems are based on a number of important features. First, proteolytic resistance can be used directly as a marker of foldedness (Parsell and Sauer 1989). Second, stably folded and well structured proteins are normally more resistant to protease digestion than those that are less stable and poorly folded. Third, M13 and fd phages are resistant to cleavage by many proteases. And finally, the surface g3p proteins of phage are needed for bacteria infection, which allows the coupling between the cleavage of inserted guest proteins and the loss of phage infection (Bai and Feng 2004).

Martin and colleagues (Martin, Sieber et al. 2001) used phage display with proteolysis to convert the mesophilic cold shock protein, Bs-CspB, from *Bacillus subtilis* to a hyperthermophilic protein. A large combinatorial library of mutated genes was created by randomizing the codons for six positions on the surface of Bs-CspB protein. The library was then subjected to phage display coupled with proteolysis which efficiently enabled the selection of stable variants. The conditions of
selection were based on two different strategies for stabilization of the protein surface. First the selection were performed in an ionic denaturant and thus non-polar surface interactions were optimized. Second, the selection were carried out at elevated temperature and resulted in the selection of variants with improved electrostatics. For the most stable selected variant, the midpoint of thermal unfolding increased by 28.2°C and the Gibbs free energy of unfolding (ΔG) by 19 kJ/mol relative to wild-type protein.

In a similar approach, Pedersen and co-workers attempted to engineer a stable mutant of ribonuclease barnase from *Bacillus amyloliquefaciens* (Pedersen, Otzen et al. 2002). The constructed library of different barnase variants was guided by the natural variability between two closely related ribonucleases (binase from *Bacillus intermedius*). When comparing the amino acid sequences for these two related ribonucleases there are 17 positions that differ in amino acid identity. A degenerate gene was designed so that the bases encoding the 17 amino acid residues were randomized to allow residues from both barnase and binase, while the rest of the gene remained constant. The library was then challenged with proteases and among the 20 clones selected 10 were studied for stability. The fact that none of the selected mutants were found to be more stable than wild type barnase may be attributed to possible local unfolding in barnase in the vicinity of the mutations.

The results of using phage-display combined with proteolysis to select stable folded protein structures clearly indicate that this method is a powerful tool for protein design. However, some studies failed to select proteins with higher stability. One possible reason for this is that there may be protease sites in loop regions that, upon
cleavage, may cause local unfolding. Further perfection of the method should help to provide insights into the forces that stabilize proteins and could also possibly lead to the design of proteins with new folds.

Selection of stable proteins using ribosome display coupled with in vitro proteolysis: Ribosome display is a powerful tool for selecting proteins based on their ability to bind ligands. Jermutus and co-workers tailored such an in vitro evolution system to perform selection based on the folding properties of proteins (Jermutus, Honegger et al. 2001). Ribosome display is based on transcribing a library of DNA to mRNA, translating it in vitro with stoichiometric amounts of ribosomes such that at the end of the reaction, the protein (as for example peptidyl-tRNA), and the mRNA are still associated with the ribosome. A stable ternary complex is formed because of the absence of a stop codon and a buffer designed to stabilize the complex. Folding of the protein is allowed by its fusion to a C-terminal unstructured tether, which can occupy the ribosomal tunnel. One of the attractions of the method is that, because all steps are carried out in vitro, ribosome display selection can easily be combined with in vitro mutagenesis and recombination during cycles of protein evolution.

Matsuura & Pluckthun used ribosome display to select for proteins with particular folding properties. They used two different schemes for selection: One selection scheme is based on protein hydrophobicity where ternary complexes are incubated with beads possessing hydrophobic groups in the presence of high concentrations of salt. Those with greater hydrophobicity bind the beads and those which remain unbound are recovered using the specific interactions between the N-terminal his-tag, anti-his-tag antibody and immobilized protein-L. Another scheme is
based on protease resistance and is similar to phage display coupled with proteolysis. The ternary complexes are incubated with protease and those that are sensitive to proteases are degraded, whereas those having a compact, stable structure are not. Ternary complexes displaying undigested proteins are recovered via the his-tag (Matsuura and Pluckthun 2003; Matsuura and Pluckthun 2004). Ribosome display is able to mimic a part of natural protein evolution by applying different selection pressures for function, solubility and proteases resistance.

In vivo monitoring of protein stability using fluorescent energy transfer (FRET): Fluorescence resonance energy transfer (FRET) was used to establish a novel in vivo screening system that allows rapid detection of protein folding and protein variants with increased thermodynamic stability in the cytoplasm of living cells (Philipps, Hennecke et al. 2003). The system is based on detection of protein folding in vivo and is independent of protein function. Furthermore, proteins are selected for based on resistance to E. coli proteases and monitored by FRET between blue fluorescent protein (BFP) and green fluorescent protein (GFP) fused to the N and C termini of the protein of interest. Efficient FRET between BFP and GFP in the ternary fusion protein is observed in vivo only when the protein of interest is folded and thus brings BFP and GFP into close proximity. One the other hand FRET is lost when BFP and GFP are far apart due to unfolding or intracellular degradation of the protein of interest. With this method an antibody VL domain was selected for improved stability. However, this method is probably limited to misfolded proteins which do not form inclusion bodies in the cytoplasm of E.coli.
Another screening approach for the selection of folded, stable proteins is the use of functional properties such as ligand-binding or catalytic activity. Protein functional properties arise from the precise balance of stability, structural specificity, solubility and resistance to proteolysis. With few exceptions, functional proteins are necessarily structured proteins. Thus, in general, proteins that bind ligands or catalyze reactions can be expected to be well-folded and native-like in structure (Shoichet, Baase et al. 1995). This approach is valuable if one wants functional proteins to differ only structurally and therefore one must not mutate residues that directly affect the function of the protein as some residues will have both functional and structural roles. Furthermore, most of the mutations should be made in the hydrophobic core and away from functional residues which are located at the active site or binding site. There are several screening methods to test for functionally stable proteins.

**In vitro binding selection using phage display**: As described above phage display is a selection technique that has been successfully applied to investigate protein stability. In adapting phage display from the more common selection for binding affinity, investigators have focused on mutating residues affecting protein stability, but not directly involved in ligand binding. Proteins are selected that retain binding capacity, with the implicit assumption that a properly folded protein is required for an intact binding interface (O'Neil, Hoess et al. 1995). As a directed evolution strategy, phage-display offers a number of important advantages. One advantage is the technology for generating large libraries has been well developed, permitting the simultaneous characterization of a relatively large number of mutants. Another advantage is the high *in vitro* stability of the phage particle permits the use of
a wide range of selection conditions (i.e. high temperature and denaturants). Varying
the stringency of selection conditions by these methods allows greater flexibility in
experimental design and is particularly relevant to questions of protein stability
(Sieber, Pluckthun et al. 1998). One limitation of this approach is the requirement for a
known binding partner with a binding interface that is unaffected by the mutations
introduced.

Cochran et al. have focused on investigating the effect of mutations on the
stability of the β1 domain of Streptococcal protein G (Gβ1) by displaying variants of
the β1 domain of the IgG-binding protein G on filamentous phage. This work relies
upon the binding of properly folded Gβ1 to the immunoglobulin Fc fragment to
separate the few functional variants from a large number of unfolded proteins (Kotz,
Bond et al. 2004).

A conceptually similar approach was employed by Bond et al. in the design of
a camelid heavy chain antibody scaffold for use in constructing naive antibody
libraries. In this work the association of the variable heavy chain (V\textsubscript{H}) with protein A
was used as a surrogate for direct stability measurements (Bond, Marsters et al. 2003).

These studies have employed phage display to identify stable clones from a
large pool of unfolded proteins. In both cases stable clones are identified from the
selection, and stability was confirmed when the individual proteins were
characterized.

\textit{In vitro binding selection by ribosome display and mRNA-display:}
Ribosome display and mRNA-display are an alternative approach for ligand binding
selection which has been developed by both Pluckthun and Szostak’s groups respectively (Hanes and Pluckthun 1997); (Roberts and Szostak 1997).

Ribosome display is a powerful tool for selecting protein function through ligand-binding. In this approach, an \textit{in vitro} translated protein is bound to immobilized ligand under conditions that leave the protein and its corresponding mRNA attached to the ribosome. After washing away unbound mRNA-ribosome-protein complexes, the remaining bound complexes are dissociated, and the eluted mRNA is amplified by RT-PCR. In an attempt to select for increased stability of a hag-peptide-binding scFv by ribosome display, the \textit{in vitro} translation steps between successive binding and amplification rounds were performed with increasing DTT concentration, 0.5 to 10 mM. Normally, transcription and translation are carried out separately during ribosome display, and translation is performed under oxidizing conditions. Reducing conditions prevent disulfide bond formation, such that the \textit{in vitro} translated proteins are kept in a reduced state. The disulfide bond is very critical for scFv stability, and only intrinsically stable fragments should be able to fold into the correct native structure in the presence of high concentrations of DTT. Indeed, some of the mutants selected by this procedure, excluding the original scFv, were shown to refold completely under reducing conditions \textit{in vitro}. In addition, Gdn HCl-denaturation curves indicate the selection of molecules which have increased stability with and without disulfide bonds (Jermutus, Honegger et al. 2001).

Szostak and co-workers have developed an alternative route to the \textit{in vitro} selection and directed evolution of proteins by generating covalent RNA-peptide fusions. Covalent fusions between an mRNA and the peptide or protein that it encodes
can be generated by \textit{in vitro} translation of synthetic mRNAs that carry puromycin, a peptidyl acceptor antibiotic, at their 3’ end. The stable linkage between the informational (nucleic acid) and functional (peptide) domains of the resulting joint molecules allows a specific mRNA to be enriched from a complex mixture of mRNAs based on the properties of its encoded peptide. Fusions between a synthetic mRNA and its encoded myc epitope peptide have been enriched from a pool of random sequence mRNA-peptide fusions by immunoprecipitation (Roberts and Szostak 1997). Keefe and Szostak isolated four new ATP-binding protein aptamers from fully randomized 80-mers that appear to be unrelated to each other or to anything found in the current databases of biological proteins (Keefe and Szostak 2001). Ribosome display and mRNA-display results show that selection for stability is possible during \textit{in vitro} ligand-binding.

\textit{In vivo binding to DNA (repressor) or RNA (rop)}: Lim and Sauer carried out combinatorial experiments in protein structure based on the binding of N-terminal variants of the $\lambda$ repressor to lytic $\lambda$ phage DNA, conferring resistance to phage infection and lysis to those cells with functional repressors. Using lytic phages of differing virulence, the ‘activity and stability’ of a repressor variant could be estimated (Lim and Sauer 1991).

Magliery and Regan have developed a novel, robust cell based screen for function and stability of the four-helix bundle protein Rop (Magliery and Regan 2004). Rop regulates the copy number of ColE1 plasmids. It facilitates the binding of an inhibitory RNA to the RNA that primes plasmid replication (by binding to hairpin loops in both of those RNAs). By expressing green fluorescent protein from a ColE1
plasmid, cellular fluorescence reports the copy number of the plasmid and therefore Rop functionality. This screen makes it possible to rapidly and effectively interrogate different libraries of hydrophobic core variants of Rop effectively for proper protein folding and stability. These tools that are based on binding activity provide the ability to sort native-like, stable protein molecules from unstructured or unstable variants.

**Selection of functional proteins (Catalytic Activity):** One can also understand native-like protein properties based on the catalytic activity of different enzyme variants. However, library design is usually even more complex than in the case of ligand binding. For example, Uchiyama et al. applied a directed evolution system to improve the thermostability of Prolyl endopeptidase (Uchiyama, Inaoka et al. 2000). Prolyl endopeptidase is the only endopeptidase that specifically cleaves peptides at proline residues. Although this unique specificity is advantageous for application in protein chemistry, the enzyme is less stable than commonly used peptidases such as subtilisin and trypsin. Therefore, the scientists desired to improve its thermostability with directed evolution. First, an efficient expression system for the enzyme in *E. coli* was established using the prolyl endopeptidase gene from *Flavobacterium meningosepticum*. Then, a method for screening thermostable variants was developed by combining heat treatment with active staining on membrane filters. Random mutagenesis by error-prone PCR and screening was repeated three times, and as a result the thermostability of the enzyme was increased step-by-step as the amino acid substitutions accumulated. The most thermostable mutant obtained after the third cycle showed a half-life of 42 min at 60° C, which was 60 times longer than that of the wild-type enzyme.
In a similar approach, Oh et al. reported the simultaneous improvement of both the oxidative stability and thermostability of N-carbamyl-D-amino-acid amidohydrolase from *Agrobacterium tumefaciens* (Oh, Nam et al. 2002; Oh, Nam et al. 2002). This enzyme is currently used in the industrial production of D-amino acids in combination with a D-hydantoinase. A mutant library was generated by DNA shuffling, and positive clones with improved oxidative and thermal stability were screened on the basis of the activity staining method on a solid agar plate containing pH indicator (phenol red) and substrate (N-carbamyl-D-p-hydroxyphenylglycine). Two rounds of directed evolution resulted in identification of an optimal mutant that had both improved thermal and oxidative stability compared to the wild-type. Sequencing of this variant revealed six mutations; four that were subsequently shown to enhance both oxidative stability and thermostability, and two mutations that gave rise only to an increase in oxidative stability.

Directed evolution coupled with selection for native-like protein properties is a very successful approach for designing thermostable protein variants with improved binding affinity and catalytic activity. However, this method is limited to proteins which have known functional activity such as binding or catalytic activity. In addition, it requires the knowledge of which residues are required for their particular functions.

*In vivo* protein solubility screen using a fusion reporter tag: There is generally a high correlation between protein stability and solubility. Stable and well-folded proteins are usually more soluble than unstable, unfolded proteins which are prone to aggregation. Additionally, globular proteins must remain soluble to perform
their functions. Recently, several protein solubility screens have been developed that do not require structural or functional information about the target protein.

These protein solubility screens and selection methods are based on a test protein fused to a reporter protein that has an easily detected function or biological activity. Information about the folding and solubility of the test protein is transduced to a screenable or selectable activity by the fused reporter domain. The reporter domain can be a selectable marker protein such as a metabolic enzyme or antibiotic resistance protein or a screenable marker such as fluorescent protein.

Waldo et al. have developed a method that is capable of screening soluble proteins in vivo using the screenable marker green fluorescent protein (GFP) as a fusion reporter tag (Waldo, Standish et al. 1999; Waldo 2003; Waldo 2003). GFP is a spontaneously fluorescent protein isolated from Pacific jellyfish, *Aequoria Victoria*. Formation of the chromophore of GFP depends on the correct folding of the protein. The green fluorescent protein folding reporter method is based on the idea that *E. coli* fluoresces when soluble proteins are fused to GFP, whereas aggregating proteins prevent GFP folding and thus block fluorescence. This method has been applied to improve the expression properties of hyperthermophilic proteins, which are expressed predominantly as inclusion bodies in *E. coli* (Pedelacq, Piltch et al. 2002). The resulting soluble proteins were used for structural studies. Hecht and co-workers also used the GFP folding reporter to discover soluble variants of the Alzheimer α/β precursor protein (Wurth, Guimard et al. 2002). They found mutations that had previously been described, as well as several new solubilizing mutations.
Waldo and coworkers have continued the development of the GFP solubility reporter by using intensive protein engineering (i.e., directed evolution) to generate a split GFP system in which the protein being tested is fused to a 15 amino acid GFP fragment (Cabantous, Terwilliger et al. 2005; Cabantous and Waldo 2006). When this chimeric fragment is co-expressed with a fragment consisting of the remainder of GFP they spontaneously reassemble and form fluorescent GFP if the attached test protein remains soluble. Cabantous et al, used a two tiered approach, in which fully intact GFP was used in combination with the split GFP system, to evolve higher solubility into proteins from *Mycobacterium tuberculosis* known to be insoluble and thus recalcitrant to protein expression, purification and ultimately crystallization (Cabantous, Pedelacq et al. 2005).

Additional production of mutant libraries of GFP, and selection based on enhanced fluorescence, has resulted in a ‘superfolder’ variant that folds robustly even when fused to poorly folded test proteins (Pedelacq, Cabantous et al. 2006). Pédelacq et al, started with the amino acid sequence of a well-folded GFP variant derived from a previous directed evolution cycle, which resulted in the mutations F99S, M153T, V163A (Crameri, Whitehorn et al. 1996), and two ‘enhanced GFP’ mutations F64L and S65T (Patterson, Knobel et al. 1997). After four rounds of DNA shuffling a highly fluorescent ‘superfolder’ GFP variant was isolated that contained the following six new mutations: S30R, Y39N, N105T, Y145F, I171V and A206V. Upon solving the crystal structure of this variant (PDB accession code: 2B3P) it was theorized that its faster folding kinetics and greater stability are likely due to the S30R mutation (Pedelacq, Cabantous et al. 2006). This mutation results in an intramolecular ionic
network, across four adjacent β-strands, which involves a sequence of five alternating acidic and basic residues. This form of surface exposed ionic network would likely be difficult to accurately predict computationally and thus reflects the value and continued importance of engineered protein libraries and associated screens.

Another example of a selectable marker was demonstrated by Davidson and co-workers in the survival of *E. coli* cells on media containing the antibiotic chloramphenicol. A positive correlation was demonstrated between the solubility of test proteins expressed as N-terminal fusions with chloramphenicol acetyltransferase (CAT) protein (Maxwell, Mittermaier et al. 1999). Growing the cells on media containing progressively higher levels of chloramphenicol provided selective pressure. This method was used to discriminate a minority population of known soluble variants from a pool of cells containing an excess of less soluble wild-type clones.

The fusion reporter method essentially screens or selects for proteins with reduced tendency to aggregate by probing intracellular concentrations of soluble, fluorescent GFP fusion proteins, or assessing the survival of *E. coli* cells on media containing an appropriate antibiotic. However, the fusion reporter method has not been shown to have the potential of generating proteins with increased thermodynamic stabilities.

In an excellent contribution, Barberis and co-workers have developed a fusion reporter system named ‘Quality Control’ using a selectable/screenable genetic marker for the selection/screen of stable and soluble single chain antibodies (scFv) in yeast independently of their antigen-binding specificity (Auf der Maur, Escher et al. 2001). The scFv is fused to a selectable marker protein which consists of a transcription
activation domain (AD) of Gal4 and a peptide fragment derived from Gal11P, which specifically interacts with the region of the DNA-binding Gal4. If the scFV is stable and soluble, the fused AD-Gal11P domain can associate with the DNA-bound Gal4 fragment, thereby activating transcription of HIS3 (select) and lacZ (screen) reporter genes. In the appropriate strain background and on selective media, HIS3 expression allows survival of the host cell (selection), whereas lacZ allows blue/white screening. Thus, cells expressing soluble scFV will survive, while those expressing unstable or insoluble scFV will die because of the lack of expression of the selectable marker HIS3. The authors validated the ‘Quality Control’ system by testing various scFvs that have previously been characterized for stability, solubility, and in vivo performance. Furthermore, this approach allowed efficient screens of scFv libraries to isolate ‘super-stable’ frameworks suitable for intracellular applications (Auf der Maur, Tissot et al. 2004).

Understanding the determinants of protein stability is critical for protein design. Rather than characterizing individual proteins with single mutations, or defined combinations of mutations, researchers have increasingly been using selection and screening methods to investigate protein stability. In comparison to the labor-intensive process of generating and characterizing individual mutant proteins, these combinatorial approaches offer the important advantage of simultaneously generating libraries of protein variants, thus allowing a much larger number of mutations to be investigated. However, interpreting the results from combinatorial experiments is not as straightforward as characterizing individual proteins. Consequently, results must be carefully assessed in light of the library design and selection pressure applied.
2.4 Emerging computational and *in vivo* approaches for engineering protein libraries

Engineered protein libraries, defined here as a collection of different mutant variants of a single specific protein, are intentionally designed to be rich in molecular diversity and can span a range from as little as 400 different variants to greater than 10^12 members per library. The goal of engineering libraries is to generate new protein variants, identified upon screening, that possess desired novel properties. Exploitation of the natural organization of the genetic code has led to ‘focused’ libraries that are lower in overall complexity yet biased towards variants with realistic biophysical properties. An emerging trend, in which computational algorithms are blended with *in vivo* screens, is also leading towards greater and more rapid success in the field of protein design.

Two key aspects of evolution are the creation (and maintenance) of molecular diversity and the selection of traits associated with specific fitness criteria (i.e., fitness conferred by particular sets of amino acids in different protein variants). Emulation of these two aspects in the laboratory, as in *in vivo*, *in vitro* or genetic screens, has led to significant advances in the field of protein design (Chapter 2, Section 3). The molecular diversity inherent to engineered protein libraries, and subsequent screening, has led to designed protein variants with novel enzymatic properties, variants with the ability to bind specifically to macromolecular targets, and variants with increased solubility and enhanced stability. In addition, the use of protein libraries and associated screening has provided unique insights into the biophysical properties of
proteins associated with certain human diseases (e.g., the Aβ42 peptide associated with Alzheimer’s disease). This section focuses on recent advances in the use of engineered protein libraries and the development of focused libraries designed to reduce sequence space to that which is most likely to produce variants with the desired properties. In addition, the use of computational algorithms to assist in the engineering of focused libraries is also described.

**Focused Protein Libraries for the de Novo Design of a Four-Helix Bundle:**

The power of focused libraries is nicely exemplified by the seminal work of Hecht and coworkers in which binary patterning of polar and non-polar amino acids was used to generate combinatorial libraries of proteins designed to fold into a specific target structure (Bradley, Thumfort et al. 2006). To generate a variant of a single specific conformation a focused library was engineered based on 1) a four-helix bundle scaffold, 2) the periodicity of the desired secondary structure elements and 3) exploitation of the natural binary organization of the genetic code (Kamtekar, Schiffer et al. 1993). To generate a focused library, the degenerate codon NTN (which encodes five non-polar amino acids; N represents any of the four DNA bases) was used in the synthetic genes for positions buried at the interface of the four helices. Conversely the degenerate codon NAN (which encodes six polar amino acids) was used for helical positions exposed to solvent. The initial library coded for a molecular diversity of $4.7 \times 10^{41}$. Forty eight variants were arbitrarily chosen for biophysical characterization and in a manner similar to that of the original design (i.e., a rational design attempt that did not utilize a combinatorial protein library (Hecht, Richardson et al. 1990)) most satisfied the design criteria except they exhibited evidence of molten globule-like
cores (Kamtekar, Schiffer et al. 1993). To address this issue the target scaffold was modified to elongate the four helices and four out of five variants arbitrarily chosen from this newly designed library exhibited thermodynamic and spectroscopic properties of stable, well-folded proteins (Wei, Liu et al. 2003). The structure of one variant, S-824, was solved to high-resolution using multidimensional NMR and revealed that all design criteria were successfully achieved (Wei, Kim et al. 2003). This finding definitively demonstrated that, even in the complete absence of a selection step, a judiciously designed focused library can be used to produce proteins that fold into the target structure and possess the desired biophysical properties.

An even more interesting characteristic of the above library is that, in addition to generating a high percentage of folded variants, the library also produced variants that possessed unplanned enzymatic capabilities. A major goal of protein design is to engineer proteins with catalytic properties. A structural prerequisite for enzymatic catalysis is that specific amino acid side chains must be properly juxtapositioned in the protein’s active site and thus protein enzymes must necessarily be well-folded. Proceeding on this premise, Wei and Hecht interrogated focused libraries of well-folded four-helix bundle proteins for enzymatic esterase activity (Wei and Hecht 2004). Surprisingly, six variants tested demonstrated esterase activity significantly above background and the S-824 variant displayed a rate enhancement \((k_{cat}/k_{uncat})\) of 8700 (Wei and Hecht 2004). The authors state that the measured activity is similar to, or better than, that observed for several esterases designed previously using rational design or automated computational methods. These results are especially impressive in light of the fact that the libraries interrogated are considered ‘naive’ as there was no
selection step for substrate binding, transition state analog binding or a screen for catalytic activity.

**Probing the Amyloidogenicity of Aβ42 with Focused Protein Libraries and GFP Reporter Fusions:** To explore the biophysical properties of peptide associates with Alzheimer’s disease Wurth et al utilized GFP fusions with fluorescence-based screening of engineered peptide libraries (Wurth, Guimard et al. 2002; Wurth, Kim et al. 2006, Chapter 2, section 3). Expanding on their previous work on the *de novo* design of β-sheet proteins, which exhibited high propensities to form amyloid-like fibril structures (Wang and Hecht 2002; Wurth, Kim et al. 2006), they employed combinatorial protein libraries to probe the amyloidogenic properties of the Aβ42 peptide known to be a major molecular component of the amyloid plaques associated with Alzheimer’s disease (Wurth, Guimard et al. 2002; Kim and Hecht 2005; Wurth, Kim et al. 2006). When fused to GFP as an N-terminal fusion the high propensity for Aβ42 to aggregate blocks proper GFP folding and thus bacteria harboring this chimera do not emit fluorescence. To explore the role that particular residues play in amyloidogenicity, mutant libraries of Aβ42 variants were generated using different methods for introducing mutations (Wurth, Guimard et al. 2002). The resulting libraries were probed in the context of the GFP screen and 36 mutant variants of Aβ42, which are more soluble relative to the wild-type peptide, were isolated. The bulk of the mutations cluster into three hydrophobic regions and, although most agree with previous model studies, there are a number of conservative mutations that simple models based solely on sequence hydrophobicity would not have predicted.
Aβ42, and the shorter variant Aβ40 (identical to Aβ42 but two residues shorter), are produced at relatively equal amounts in vivo, yet the senile plaques in diseased brains are composed primarily of Aβ42, which also more readily forms fibrils in vitro (Wurth, Guimard et al. 2002; Kim and Hecht 2005). To analyze the potential role these two terminal residues play in amyloidogenicity a library was generated by randomizing the codons for positions 41 and 42. In addition, focused libraries were engineered in which the degenerate codon NTN (which encodes five non-polar amino acids) was incorporated at both positions for one library and NAN (which encodes six polar amino acids) incorporated at both positions for another. The resulting screen indicated that hydrophobicity, as well as β-sheet propensity, greatly influence overall solubility. Interestingly, only one colony from the hydrophilic library displayed a white phenotype and subsequent DNA sequencing revealed two arginines at positions 41 and 42. Analysis of a structural model of an Aβ42 fibril, based on solid state NMR data (Petkova, Ishii et al. 2002), revealed that an arginine at the terminal positions would be in close spatial proximity to a glutamic acid at position 11 and that a putative salt bridge between these residues may function to stabilize the β-sheet structure of the Aβ42 fibril.

Kim and Hecht returned to the "binary code" of protein structure (which specifies the pattern of polar and hydrophobic residues in protein structure(Bradley, Thumfort et al. 2006; Bradley, Wei et al. 2007)) to engineer protein libraries for the purpose of probing another key issue associated with the amyloidogenicity of Aβ42 (Kim and Hecht 2006). Recently determined structures of model amyloidogenic peptides reveal the presence of highly ordered "steric zippers" comprising well packed
structures with specific side-chain interactions (Nelson, Sawaya et al. 2005; Kim and Hecht 2006). To ascertain if these specific side-chain interactions could be substituted for generic hydrophobic amino acids, a focused library of Aβ42 variants was engineered in which the degenerate codon NTN was used to code for five non-polar amino acids at 12 different positions. The findings demonstrate that generic hydrophobic amino acids are sufficient to drive Aβ42 into amyloid fibril structures.

The GFP reporter screen is such an effective tool that it is now being used to screen compounds from a library of different molecular diversity. Kim et al used the in vivo Aβ42-GFP fusion system to screen a library of triazine derivatives for compounds that block the self association of Aβ42 (Kim, Kim et al. 2006). They demonstrated that one compound in particular effectively blocked Aβ42 self assembly which allowed proper GFP folding and associated fluorescence. This system will likely continue to be used as a powerful method to screen for drugs that may offset or prevent the debilitating neurodegenerative effects of Alzheimer’s disease.

**The Use of Protein Design Algorithms to Focus Protein Libraries:**

Although significant molecular diversity in protein libraries is desirable, especially to insure a higher likelihood of generating the optimal sequence, it is not hard to imagine that many of the sequences in a large randomized library would be non-functional and potentially deleterious. The concept of using protein design algorithms (Dahiyat and Mayo 1997) to virtually screen protein libraries and reduce sequence space down to a region amenable to in vivo or in vitro screening has been applied in the recent past (Voigt, Mayo et al. 2001; Voigt, Martinez et al. 2002; Meyer, Silberg et al. 2003; Patrick and Firth 2005; Denault and Pelletier 2007). For example, Hayes et al used
design algorithms to target specific positions in the engineering of a β-lactamase variant with improved resistance to cefotaxime (Hayes, Bentzien et al. 2002). The algorithms were initially used to redesign (i.e., perform computational mutagenesis on) a subset of 19 residues in proximity to the active site of the enzyme and resulted in one set of mutant residues, and associated rotameric positions, that represented the lowest calculated energy in the context of fixed backbone coordinates (the low energy sequence is termed the global minimum energy conformation - GMEC). To guide library design they determined additional sets of amino acid mutations that had reasonable calculated energies by running Monte Carlo simulated annealing starting with the GMEC sequence. This approach effectively reduced the sequence space for the 19 positions from a possible molecular diversity of ~5.2 x 10^24 down to 172,800 and, upon generation and screening in the laboratory, resulted in a β-lactamase variant that exhibited a 1,280 fold increase in resistance to the targeted antibiotic (Hayes, Bentzien et al. 2002).

Engineered protein libraries, focused with design algorithms at specific positions, were also used to enhance the fluorescent properties of a blue variant of GFP, referred to as BFP (Mena, Treynor et al. 2006; Treynor, Vizcarra et al. 2007). The mutation Y66H of the chromophore of GFP yields a blue fluorescent protein that has the undesirable properties of low quantum yield and relatively rapid photobleaching (Shaner, Steinbach et al. 2005). To improve these features Mena et al used protein design algorithms, and fluorescence activated cell sorting, to target 12 core positions within 7 Å of the imidazole ring of the fluorescent chromophore (Mena, Treynor et al. 2006). Diversity was restricted upon the use of an algorithm that
determines library composition by minimizing a weighted average of conformational energies calculated for each of seven candidate libraries (Treynor, Vizcarra et al. 2007). Combinations of variably focused codons were incorporated into the oligonucleotides used in library engineering. The unintended incorporation of a more inclusive codon at position 224 (i.e., RBG, where R = AG and B = CTG as opposed to the intended RTG codon) resulted in a library that encoded $3.3 \times 10^5$ unique sequences. This level of diversity is greatly reduced relative to the $4.1 \times 10^{15}$ sequences of a library completely randomized at 12 positions. Screening of the focused libraries yielded a variant with enhanced quantum yield (0.55 versus 0.34), reduced pH sensitivity and a 40-fold increase in photobleaching half-life thus illustrating the power of using protein design algorithms to focus library diversity to regions of sequence space most likely to include a variant with the desired properties.

**Engineered Protein Libraries of the Gβ1 Domain:** A proven work horse in the field of protein design is the β1 domain of Streptococcal protein-G (Gβ1) as it has been the subject of design using computational approaches (Dahiyat and Mayo 1997; Malakauskas and Mayo 1998; Ross, Sarisky et al. 2001; Tsai, Tsai et al. 2004; Choi and Mayo 2006) as well as in vivo and in vitro screens (Distefano, Zhong et al. 2002; Kotz, Bond et al. 2004; Alexander, Rozak et al. 2005). In particular, phage display has proved to be a powerful tool for selecting stable Gβ1 variants (as well as other designed proteins (Minard, Scalley-Kim et al. 2001; Chu, Takei et al. 2002; Scalley-Kim, Minard et al. 2003; Bai and Feng 2004; Feng and Bai 2004; chapter 2, section 3)) from fairly diverse engineered protein libraries. For example, in work by Schmid and coworkers phage display was used in combination with in vitro proteolysis to
select for Gβ1 variants with increased intrinsic stabilities (Wunderlich, Martin et al. 2005; Wunderlich and Schmid 2006). The genes for the Gβ1 variants were cloned into the phage gene-3-protein (G3P) and the resulting libraries were subjected to multiple rounds of in vitro proteolysis and amplification in bacteria. In one study, in which this method was compared to a previous computational design of Gβ1 (Malakauskas and Mayo 1998), saturation mutagenesis was used to randomize the codons at four boundary positions (molecular diversity of ~160,000) that were previously identified with computational algorithms. The genes for approximately 100 Gβ1 variants, which exhibited high protease resistance, were sequenced and revealed that amino acids that conferred greater stability were highly degenerate yet biased towards hydrophobic and aromatic residues at three of the four positions (Wunderlich, Martin et al. 2005). Thermodynamic analysis of ~21 variants revealed that two were more stable than the sequence derived from the computational design. This finding is not entirely surprising as a necessary limitation of the computational approach is that the protein backbone must be held rigid for the calculation to remain tractable.

In subsequent work Wunderlich and Schmid used a two step approach in which error-prone PCR was used first on the Gβ1 gene to identify candidate positions with high potential for stabilization. Interestingly, the five positions identified fall within the partially exposed ‘boundary’ category and none were located in the protein’s core (in fact two of the five were identical to positions identified previously with computational methods (Malakauskas and Mayo 1998)). The second step entailed the saturation mutagenesis of the five positions followed by multiple rounds of selections. Manual incorporation of the most stabilizing mutations resulted in a Gβ1 variant that
exhibited an increase in $T_M$ of 35.1º C and an increase in $\Delta\Delta G_D$ of 28.5 kJ mol$^{-1}$ at 70º C (Wunderlich and Schmid 2006).

Significant advances in the field of protein design have come from both computational approaches as well as in vivo and in vitro screens. With the computational approach molecular diversity is generated and screened virtually and normally results in a mutated sequence that reflects the best attempt to mathematically capture the physical chemistry of protein folding (usually represented by a molecular mechanics force field). A caveat to this approach is that many of the physical interactions that give rise to the final folded structure of a protein are difficult to accurately emulate (e.g., variable electrostatics) and/or do not lend themselves readily to the large scale combinatorial analysis of the enormous number of potential side chain interactions (e.g., pair-wise deconvolution of solvation energies). Even with these caveats, computational approaches have consistently proved to be very efficient at getting quite close to the lowest energy mutant sequence (and conformation). On the other hand engineered libraries, and associated screens, have also been used successfully to determine the optimal sequence for specific design goals but have practical limitations regarding library size or identification of optimal amino acid positions for mutagenesis. The emerging trend in which robust computational algorithms are blended with powerful in vivo screens should lead to greater and more rapid success in the field of protein design as each method can be used to effectively address and alleviate the inherent limitations of the other.
Acknowledgement:

This chapter, in part, is a reprint of the material as it appears in Nora H. Barakat, John J. Love, “Molecular Diversity in Engineered Protein Libraries”. Submitted to and accepted by Current Opinion in Chemical Biology.
III. EXPLOITING ELEMENTS OF TRANSCRIPTIONAL MACHINERY TO ENHANCE PROTEIN STABILITY

3.1 Abstract

The correlation between protein structure and function is well established yet the role stability/flexibility plays in protein function is currently being explored. Here we describe an *in vivo* screen in which the thermal stability of a test protein is directly correlated to the transcriptional regulation of a reporter gene. The screen readout is independent of the test protein’s function, proteolytic resistance, solubility or propensity to indiscriminately aggregate and is thus solely dependent on the overall stability of the test protein. The system entails the use of an engineered chimeric construct that consists of three covalently linked domains which includes a constant N-terminal DNA binding domain, a variable central test protein, and a constant C-terminal transcriptional activation domain. The test proteins are mutant variants of the β1 domain of streptococcal protein-G that fairly evenly span a thermal stability range from as low as 38°C to greater than 100°C. When the chimeric construct contains a test variant of low thermal stability the reporter gene is up-regulated to a greater extent relative to that of more stable/less flexible variants. A panel of nine Gβ1 mutant variants was used to benchmark the screen and spectroscopic methods were employed to accurately characterize the thermal and structural properties of each variant.
3.2 Introduction

The field of protein design aspires to engineer proteins of specific function. Function is directly correlated to structure and therefore a specific target fold is normally chosen for a desired function. In addition to structure, protein dynamics also correlate to function (Agarwal 2006) and in recent years increasing experimental evidence (primarily driven by high-field heteronuclear NMR) indicate that overall structural stability or lack thereof, may also play an important role in protein function (Dyson 2004; Dyson 2005; Mittermaier 2006). The overall dynamic regime of protein backbone and side chain atoms correlate to intrinsic stability and thus it reasons that protein flexibility (or plasticity) may also correlate to function. Therefore the ability to tailor a designed protein’s intrinsic stability/flexibility to a desired function may ultimately enhance protein design endeavors. Furthermore, an in vivo screen for protein stability enables the rapid analysis of large portions of a particular fold’s sequence space.

To date considerable efforts have been made to create in vivo screens designed to interrogate large combinatorial libraries of protein variants for increased stability (Bai 2004; Hecht 2004; MacBeath 1998; Magliery 2004) or solubility (Philipps 2003; Auf der Maur 2004; Cabantous, 2005; Waldo, 2003) (for comprehensive reviews see Magliery (Magliery 2004; Magliery 2004)). In addition to these screens, computational design methods have also been used either outright (Korkegian 2005; Ashworth 2006; Dahiyat 1997; Malakauskas 1998; Kuhlman 2003), or in combination with in vivo screens (Wunderlich 2005; Hayes 2002), to design and select for protein variants with increased intrinsic stability and/or improved function. Of particular
relevance to the work reported here are studies in which the compact, highly stable
fold of the β1 domain of streptococcal protein-G (Gβ1) has been the target of design
efforts (Byeon 2003; Byeon 2004; Goehlert 2004; Louis, 2005). Phage display
methods have been developed for the selection of folded Gβ1 variants (Distefano
2002; Kotz 2004; Alexander 2005; O'Neil 1995) (and a close structural homologue -
protein L) (Gu 1995). The general premise of these studies is based on the valid
assumption that destabilized variants do not fold properly and thus do not bind the
constant Fc region of immobilized antibodies as effectively as properly folded
variants. As the relationship between higher stability/enhanced folding and stronger
binding is not strictly absolute (i.e., there are examples of induced fit or folding of
partially disordered proteins upon binding or catalysis (Vamvaca 2004; Love 2004) it
is not unexpected that in some cases phage-based screens were only able to select for
folded variants but not for variants of greater thermal stability (Gu 1995; O'Neil 1995).

Here we report the development of an in vivo screen for protein stability that is
completely independent of the test protein’s function. In addition, screen results
correlate well to the measured stability of protein variants known to self associate
(Shukla 2004) and thus reporter readout is also independent of the test protein’s
solubility or propensity to indiscriminately aggregate. Structure based rational design
was used to create a panel of 9 Gβ1 variants that span a thermal stability range from
38º C to greater than 100º C. These variants were characterized in the context of the
engineered in vivo screen and structural parameters were assessed with circular
dichroism (DC), NMR and in vitro proteolysis.
3.3 Materials and Methods

Materials

All chemicals and reagents were of the highest quality and obtained from either Sigma-Aldrich or Fisher Scientific International Inc. unless otherwise stated. Oligonucleotides were obtained from Integrated DNA Technologies. Pfu Turbo® DNA polymerase was obtained from Stratagene. Restriction enzymes and buffers were obtained from New England Biolabs. The BacterioMatch® Two-Hybrid System was obtained from Stratagene. All cloning was performed with either E.coli XL1-Blue from Stratagene or Top 10 from Invitrogen. Protein expression was performed in BL21(DE3) from Novagen. Point mutations were produced using the QuikChange® method (Stratagene).

Construction of the Chimeric Construct Plasmid

The genes for the chimeric construct (i.e., RNAPα, λcI) were sub-cloned from plasmids (i.e., pTRG, pBT) obtained from a commercially available bacterial two-hybrid system (BacterioMatch® Two-Hybrid System - Stratagene). The gene for RNAP-α was PCR amplified from the pTRG plasmid and cloned into the pBT vector downstream of λcI via EcoRI/BamHI (all oligonucleotide sequences are listed in table 3.1). The genes for the Gβ1 variants were sub-cloned into the chimeric vector with the engineered restriction sites NotI and EcoRI. The λ-operator DNA and the reporter genes are located on the F’ episome which is supplied in the bacterial strain XL1-Blue F’ (Stratagene).
Creation of the G\(\beta\)1 Mutant Variants

Point mutants of G\(\beta\)1-WT and MonB were produced using the QuikChange® method (Stratagene) with oligonucleotide primers (Integrated DNA Technologies) that contained the appropriate DNA base changes. Synthetic DNA oligonucleotides were used for recursive PCR synthesis of the genes for the MonB-ORDES variant and the genes in the randomized library (oligonucleotide sequences are listed Table 3.1).

Assessing Bacterial Growth on Plates Containing Reporter Antibiotic

Chimeric construct plasmids containing the genes for five G\(\beta\)1 variants (i.e., MonA, G\(\beta\)1-WT, MonB-WT, MonB A45V, and MonB A45Y) were transformed into 50 µl of BacterioMatch two-hybrid system reporter strain competent cells (Stratagene). Cells (80 µl) from each transformation reaction were plated on LB control plates (CK) that contained chloramphenicol (12.5 µg/ml) and kanamycin (50 µg/ml). The function of the control plates was to verify colony numbers in the absence of reporter antibiotic. The same volume of cells (80 µl) was plated on reporter antibiotic plates (CCK) that, in addition to chloramphenicol (12.5 µg/ml) and kanamycin (50 µg/ml), also contained carbencillin (both 750 and 1000 µg/ml). All plates were incubated at 37º C for approximately 20 hours.

Assessing Bacterial Growth in Liquid Media Containing Reporter Antibiotic

Plasmids that contained the chimeric construct and the genes for all G\(\beta\)1 variants tested were transformed into BacterioMatch two-hybrid system reporter strain competent cells (Stratagene) and plated on control CK plates that contained
kanamycin (50 µg/ml), and chloramphenicol (12.5 µg/ml). Colonies were picked from these CK plates and grown in 20 ml LB liquid media with kanamycin (50 µg/ml), and chloramphenicol (12.5 µg/ml) overnight at 37º C. From the overnight cultures cells were diluted to OD$_{600}$ of 0.1 in 10 ml of fresh LB containing kanamycin (50 µg/ml), and chloramphenicol (12.5 µg/ml). The cultures were then incubated for 15 min. at 37º C and the OD$_{600}$ was measured again to verify that all cultures were the same density. Carbencillin (3000 µg/ml) was then added and the cultures were incubated for a total of 4 hours at which time the OD$_{600}$ was measured to ascertain the extent of growth.

**Protein Expression and Purification**

For protein expression the genes for the Gβ1 variants were sub-cloned into pET-21a (Novagen) and transformed into BL21(DE3). After growth to an OD$_{600}$ of approximately 1 the cells were induced with 1 mM IPTG. The expressed proteins were isolated using a freeze/thaw method (Johnson, 1994) and purification was accomplished with reverse-phase HPLC using a linear 1% min$^{-1}$ acetonitrile/water gradient containing 0.1% TFA. Concentrations of all variants were determined in 6 M guanidine hydrochloride using standard extinction coefficients for the tryptophan and tyrosine residues. Proteins that were $^{15}$N-labeled for NMR studies were prepared with standard M9 minimal media using $^{15}$N-ammonium sulfate (2 g/L). Protein purity was verified with SDS-PAGE and reverse phase HPLC and the correct molecular weight was confirmed by mass spectrometry.
Circular Dichroism

The CD data were collected on Jasco-810 spectrometer equipped with a thermoelectric unit and using a 0.1 mm path length cell. Protein samples were 50 µM in 50 mM sodium phosphate at pH 6.5. Thermal melts were monitored at 218 nm. Data were collected every 1º C with an equilibration time of 2 min. Far-UV spectra were acquired in the continuous mode at 25º C with 1 nm bandwidth and a 4 s response time. For the thermal denaturation curves the data were normalized by first linearly shifting all points such that \([\theta]_{218}\) value at 5º C equaled zero. Then a scaling factor was obtained for each set by dividing the maximum \([\theta]_{218}\) value for all sets at 95º C (i.e., 53.0) by the \([\theta]_{218}\) value at 95º C for each set. All data points for each set were then scaled by the unique scaling factor calculated for each set.

NMR Spectroscopy

NMR spectra were collected at 293 K on a Varian UnityPlus 600 MHz spectrometer equipped with an HCN-triple-resonance probe with triple-axis pulse field gradients. Protein concentrations were ~1.25 mM in 50 mM sodium phosphate, pH ~6.5. Standard 2D \([^1H, ^15N]\) HSQC spectra were collected for all variants except those selected from the screen. The programs NMRDraw/NMRPipe (Delaglio, Grzesiek et al. 1995) were used to process the NMR data and the program NMRView (One Moon Scientific, Inc.) was used to generate and analyze spectra.
In Vitro Proteolysis Analysis

The enzymes used in this assay were bovine $\alpha$-chymotrypsin and trypsin obtained from Calbiochem® (catalog# 230832) and the trypsin from ICN Biomedicals, Inc™ (catalog# 101171). Each enzyme was reconstituted to a stock concentration of 20 mg/mL in 50 mM sodium phosphate buffer, pH of 6.8. The reactions were carried out at 30º C, constant enzyme concentration, and a constant concentration of each variant (0.625 mM). After fixed amounts of time each reaction was immediately injected onto an HPLC C-18 analytical column.

X-gal indicator plates

MonA and MonB variants in chimeric construct were transformed into BacterioMatch two-hybrid system reporter strain competent cells (Stratagene). The transformation reactions were plated onto fresh prepared X-gal indicator agar plates. The indicator plates are supplemented with X-gal (80 $\mu$g/ml), 0.2 mM $\beta$-galactosidase inhibitor (phenylethyl $\beta$-D-thio galactoside, prepared in dimethyl formamide), kanamycin (50 $\mu$g/ml) (F’ episome resistance), and chloramphenicol (12.5 $\mu$g/ml) (chimeric construct plasmid resistance). Different concentrations of X-gal and $\beta$-galactosidase inhibitor were tested to find the right conditions to see noticeable difference between these two variants (such as X-gal (160 $\mu$g/ml) and 0.4 mM $\beta$-galactosidase inhibitor, X-gal (80 $\mu$g/ml) and 0.4 mM $\beta$-galactosidase inhibitor, or X-gal (40 $\mu$g/ml) and 0.2 mM $\beta$-galactosidase inhibitor).
MacConkey plates

MonA and MonB variants in chimeric construct were transformed into BacterioMatch two-hybrid system reporter strain competent cells (Stratagene). The transformation reactions were plated onto fresh prepared MacConkey agar plates. MacConkey plates are made of 40 g Bacto-MacConkey agar (Difco) added into 1 L of water; autoclaved and cool slightly, then added lactose to 1% final concentration, and chloramphenicol (12.5 µg/ml).

ONPG (O-Nitropheny1 B-D- Galactopyranoside)

MonA, Gβ1-WT, MonB-WT, MonB A45V, and MonB A45Y in the chimeric construct were transformed into BacterioMatch two-hybrid system reporter strain competent cells (Stratagene) and plated onto kanamycin (50 µg/ml), and chloramphenicol (12.5 µg/ml) (CK) LB plates. Colonies were picked from CK plates and grown in 20 ml LB liquid media with kanamycin (50 µg/ml), and chloramphenicol (12.5 µg/ml) for an O.D. 600 of 2.2. 1ml from each culture was spun down for 5 minutes at 400 rpm. One ml of Z buffer and 10ul of toluene were added to each pellet and vortexed for 15 seconds to open up the cells. Pellets were incubated at room temperature for 10 minutes followed by 10 minutes at 28°C. 200 µl of ONPG solution (4 mg/ml in Z buffer) were added to start the reaction. The reactions were stopped after 10 minutes by adding 300 ul of 1 M Na2CO3. Cells were spun down at the maximum speed for 4 minutes. O.D. at 420 reading was measured for the supernatants.
**CENTA™ β-Lactamase Substrate**

MonA and MonB in the chimeric construct were transformed into BacterioMatch two-hybrid system reporter strain competent cells (Stratagene) and plated onto kanamycin (50 µg/ml), and chloramphenicol (12.5 µg/ml) (CK) LB plates. Colonies were picked from CK plates and grown in 20 ml LB liquid media with kanamycin (50 µg/ml), and chloramphenicol (12.5 µg/ml) to an O.D. 600 of 4.0. 1 ml from each culture was spun down for 5 minutes at 400 rpm. One ml of 50 mM phosphate buffer of pH 7 was added to the pellet and vortexed for 3 minutes. 10 µl of CENTA β-Lactamase Substrate (50 mg/ml in methanol) was added to the pellet. The whole reaction was aliquoted into 10 tubes and stored in the dark. The wavelength scan from 250nm to 450 nm for each variant was measured at different time points (i.e. 0 minutes, 30 minutes, 1 hour, 1.5 hours, 2 hours, 2.5 hours, 3 hours, 3.5 hours, 4 hours, and overnight) using the ten different aliquots.
Table 3.1 DNA oligomer sequences. Sequences are listed in 5’ – 3’ direction. Coding sequence is in upper case. Noncoding sequence is in lower case. All oligomers were obtained from Integrated DNA Technologies (IDT)

<table>
<thead>
<tr>
<th>Oligomer</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eco_1HYB_F</td>
<td>gcgtctGAATTCATGCAGGGTTCTGTGACAGAGTTTCT</td>
</tr>
<tr>
<td>bam_1HYB_R</td>
<td>gaattataGATCCGGCCGCCTCTGGTTTCTTTT</td>
</tr>
<tr>
<td>PBT Forward</td>
<td>TCC GTT GTG GGG AAA GTTATC</td>
</tr>
<tr>
<td>PBT Reverse</td>
<td>GGGTAGCCAGCAGCATCC</td>
</tr>
<tr>
<td>Not I W/T</td>
<td>aagagaGCGGCCGCAT CTACTACTT ACAAAATTAATCCTTAA</td>
</tr>
<tr>
<td>ECORI W/T</td>
<td>ggtggtGAATTC -CCTTCAGTAACTGTAAAGGTCTTAGT</td>
</tr>
<tr>
<td>NotI_ProG</td>
<td>cagagaGCGGC -CGCATCTACCTATAAGCTGATTTGAA</td>
</tr>
<tr>
<td>EcoRI_ProG</td>
<td>gacattGAATTC – CCTCCGTCACGGTGTAAGGTCTTTG</td>
</tr>
<tr>
<td>ProG W43Y F</td>
<td>GACAACCGGTGTGGACCGTGTAATATATCTACGACGATGCGACTAAG</td>
</tr>
<tr>
<td>ProG W43Y R</td>
<td>CTTAGTCGATCGTCGTAAGTATATATCCACGCCTACCCGTCAAACACGTGTC</td>
</tr>
<tr>
<td>ProG W43V F</td>
<td>GACAACCGGTGTGGACCGTGTAAGTGACCTACGACGATGCGACTAAG</td>
</tr>
<tr>
<td>ProG W43V R</td>
<td>CTTGTCGACATCGTCGTAAGTCACTTACCGTCAAACACGTGTC</td>
</tr>
<tr>
<td>ProG W43A F</td>
<td>GACAACCGGTGTGGACCGTGTAAGTGACCTACGACGATGCGACTAAG</td>
</tr>
<tr>
<td>ProG W43A R</td>
<td>CTTAGTCGATCGTCGTAAGTCACTTACCGTCAAACACGTGTC</td>
</tr>
<tr>
<td>MonB A45Y F</td>
<td>TTAAGGGTGGAATGGACAGTAGATGAAGCGACCAAGAC</td>
</tr>
<tr>
<td>MonB A45Y R</td>
<td>GTCTTTGTCGCTTCCATCTACTGTCCATTACCCCTTAA</td>
</tr>
<tr>
<td>MonB A45V F</td>
<td>TTAAGGGTGGAATGGACATACGATGAAGCGACCAAGAC</td>
</tr>
<tr>
<td>MonB A45V R</td>
<td>GTCTTTGTCGCTTCCATCTACTGTCCATTACCCCTTAA</td>
</tr>
</tbody>
</table>
3.4 Results

Screen Design

The system developed in my dissertation project is referred to as a ‘Chimeric in vivo screen’. In this system a target protein is expressed as a fusion protein with a DNA binding domain fused at the N-terminus and a transactivation domain at the C-terminus. In Greek mythology a ‘Chimera’ is a fire-breathing monster represented by a composite of three animals: a lion, a goat, and a serpent (Figure 3.1). The three-component ‘chimeric’ fusion protein in the system we have developed acts as a functional transcription factor for a reporter gene (e.g., β-lactamase).

Creation of the chimeric in vivo screen entailed splicing the genes of different mutant variants of the Gβ1 domain between the genes for the proteins used as ‘bait’ and ‘prey’ in a bacterial two-hybrid screen (Dove 1997; Dove 1998; Dove 2004) (BacterioMatch® Two-Hybrid System - Stratagene). The Gβ1 domains all start at amino acid position two (Thr) and end at position 56 (Glu) as per the numbering scheme of the structure deposited in the PDB (accession code 1PGA) (Gallagher 1994). The resulting protein is a three domain chimera where the N-terminal DNA binding domain (full-length bacteriophage λcI, 237 amino acids) is linked to the N-terminus of the Gβ1 variant and the C-terminal transactivation domain [the N-terminal domain of the α-subunit of bacterial RNA polymerase, 248 amino acids)] is linked to the C-terminus of the Gβ1 variant (Figure 3.2). The chimeric construct was created by subcloning the gene for RNAPα into the bait plasmid downstream of λcI. The Gβ1 genes were then subcloned between λcI and RNAPα in the newly created plasmid.
There is a short linker between the C-terminus of λcI and the N-terminus of the Gβ1 variants (Ala3 Ser) and a three amino acid linker (Gly Asn Ser) between the C-terminus of the Gβ1 variant and the N-terminus of RNAP-α (Figure 3.2). The 3-domain chimera acts as a functional transcription factor upon binding the λ operator (located upstream of the reporter genes in the F’ episome) and results in the transcription of reporter genes (e.g., β-lactamase).
Figure 3.1. Chimeric Construct. In Greek mythology a ‘Chimera’ is a fire-breathing monster usually represented as a composite of three animals such as lion, goat, and serpent (A). This image illustrates the 3-domain protein chimera. The genes for the Gβ1 variants were fused between the genes for λcI and RNAPα which resulted in the three domains linked together as follows - λcI-Gβ1-RNAPα (B).
Figure 3.2 Chimeric Vector and Construct. The chimeric construct is a three domain chimera where the N-terminal DNA binding domain (full-length bacteriophage λcI) is linked to the N-terminus of the Gβ1 variant. There is a short linker between the C-terminus of λcI and the N-terminus of the Gβ1 variants (Ala3 Ser). The C-terminal transactivation domain [the N-terminal domain of the α-subunit of bacterial RNA polymerase (RNAPα)] is linked to the C-terminus of the Gβ1 variant. There is a three amino acid linker (Gly Asn Ser) between the C-terminus of the Gβ1 variant and the N-terminus of RNAP-α (A). The chimeric vector and 3-domain protein chimera. The genes for the Gβ1 variants were subcloned into the chimeric vector between the genes for λcI and RNAPα which resulted in the three domains linked together as follows - λcI-Gβ1-RNAPα (B).
Design of Gβ1 ‘parent’ mutants with variable stability

Gβ1-WT is a protein that has only 56 amino acids and exhibits relatively minimal structural perturbation upon fairly extensive mutagenesis. Furthermore, it is monomeric and behaves well in solution. Gβ1-WT has been extensively redesigned and biophysically analyzed. Additionally, its structure has been determined to a high resolution (Gronenborn 1991; Gallagher 1994). Figure (3.3) shows the structure of Gβ1-WT which has a helix that lies on top of a four-stranded β-sheet. The central two strands (beta 1 and beta 4, comprising the NH2- and COOH-termini) are parallel, and the outer two strands (beta 2 and beta 3) are connected by the helix. This novel topology is probably responsible for the extreme thermal stability of this small domain (melting temperature of 85º C).

A number of factors have been attributed to the relatively high thermal stability of Gβ1-WT. For example, aromatic interactions in the core of the protein likely contribute to the observed thermal stability. The role that aromatic clusters have on protein stability have been studied extensively with different thermophilic proteins (Kannan and Vishveshwara 2000) and on Gβ1-WT (Pande and Rokhsar 1999; Ma and Nussinov 2000). Three aromatic amino acids in the core of Gβ1-WT (i.e., Trp 43, Tyr 45 and Phe 52) form an aromatic cluster which likely contributes to its high thermal stability (Figure 3.3). Mutation of any of these amino acids causes a significant decrease in thermal stability (Kobayashi, Honda et al. 2000). Kobayahi, et al. suggested that the rigid structure at the loop region (residues 46 to 51) is another factor for the thermal stability of Gβ1-WT. Cooperativity between these two factors causes a bend in the protein backbone and promotes hydrogen bonds between the two anti-
parallel strands. This is likely the reason why such a short linear peptide forms a stable structure under physiological conditions (Kobayashi, Honda et al. 2000).

**Origin of the MonA and MonB Variants:** The MonA and MonB variants originated from an earlier design project in which a *de novo* protein interface was engineered by computationally docking the normally monomeric Gβ1 domain to itself. The ORBIT suite of protein design algorithms (Dahiyat, 1997) was used to mutate specific interfacial side-chains with the goal of driving specific complex formation (Shukla, 2004; Huang, 2005). This design resulted in a pair of monomers that, upon generation in the laboratory, formed a heterodimer of modest binding affinity. Table (3.2) illustrates the sequence of Gβ1-WT, MonA, and MonB and also highlights the mutated amino acids.

Introduction of the mutations that resulted in the sequences for each monomer altered their biophysical and thermodynamic properties relative to Gβ1-WT. The 12 mutations that resulted in the MonA sequence stabilized it to a hyperthermophile (i.e., Tm > 100° C) while the 8 for MonB were destabilizing resulting in a Tm of 38° C (Gβ1-WT ~85° C). These mutations affected the expression levels of each respective variant (in the context of the original expression vector - pET11M). The expression level for MonA is 40 mg/L while that for MonB is 4 mg/L. The biophysical properties of the Gβ1 variants are summarized in table (3.3).
Figure 3.3. Gβ1-WT Structure and Positions of Specific Residues. The backbone of β1 domain of streptococcal protein-G (Gβ1-WT) is depicted as a blue ribbon for the α-helix and yellow ribbons for the β-sheet. Side chains are shown for the residues that make up the C-terminal aromatic cluster (43, 45, 52). The transparent surface was generated by ‘rolling’ a probe sphere (radius of 1.4Å) over all atoms including side chains (and hydrogen atoms). The vertices and normals for the transparent surface were generated with the program MSMS (Sanner, Olson et al. 1996) (using a probe sphere radius of 1.4Å), the backbone ribbons with the program Molmol (Koradi, Billeter et al. 1996) and the final combined image was rendered with the program POVRay.
Table 3.2 Amino Acid Sequence for Gβ1-WT and the ORBIT Selected Sequences for MonA and MonB.

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WT</td>
<td>- MTYKLILNGKTLGETTTEAVDAATAEKVKQYANDNGVDGEWYDDATKTFTVTE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mon-A</td>
<td>- MTYKLILNGKTLGETTFAEADAALAEYIFRALAKGVDGEWYDDATKTFTVTE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mon-B</td>
<td>- MTYKLILNGKTLGETTTEAVDAADVFAQYADNGVKGETADEATKTFTVTE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(c = \) core, \(s = \) surface, bold = a mutated position, underline = calculated position that is wild-type
Table 3.3 Biophysical properties of the Gβ1 variants.

<table>
<thead>
<tr>
<th></th>
<th>Gβ1-WT</th>
<th>MonA</th>
<th>MonB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Mutations</td>
<td>W/T</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Expression level</td>
<td>~40 mg / liter</td>
<td>~40 mg / liter</td>
<td>~4 mg / liter</td>
</tr>
<tr>
<td>Melting Temperature</td>
<td>~ 85 °C</td>
<td>&gt; 95 °C</td>
<td>~ 38 °C</td>
</tr>
</tbody>
</table>
Differences in N-terminal Processing for the Three Parent Variants: When Gβ1 variants are expressed in bacteria and then purified using reverse phase HPLC, they exhibit differences in the manner in which they elute from the HPLC column (e.g., the MonA variant elutes as two peaks). The first peak corresponds to intact MonA but with the N-terminal methionine still formylated (f-Met). The second peak corresponds to intact MonA where the N-terminal methionine has been processed normally. We theorize that the extreme stability of MonA prevents the enzymatic removal of the formal group from the N-terminal Met. The Gβ1-WT variant elutes from the HPLC column as one peak that corresponds to the intact protein with the N-terminal methionine processed properly. On the other hand, the MonB variant elutes as two peaks. The first peak is very small and corresponds to intact MonB with the N-terminal methionine processed properly. The second peak corresponds to MonB without the N-terminal Met. The fact that the dominant peak for MonB corresponds to the intact protein minus the N-terminal methionine likely reflects the lower overall stability of this variant relative to MonA and Gβ1-WT.

In Vitro Proteolysis Analysis of Gβ1 Variants: In vitro proteolysis experiments were performed on the three parent Gβ1 variants to study their structural integrity. For the in vitro proteolysis assays, we subjected the three variants to degradation with the proteolytic enzymes trypsin and chymotrypsin for time periods of thirty and sixty minutes. The extent of degradation was analyzed with reversed phase HPLC. According to the results depicted in Figure 3.4, the most stable protein variant, MonA, is more resistant to proteolysis relative to Gβ1-WT and MonB. We were thus
able to establish a correlation between protein stability and proteolytic resistance in this *in vitro* assay.

Taking all the above biophysical properties into consideration (melting temperature, expression levels and proteolysis resistance) we confirmed that MonB is the least stable variant. The poor expression yield and low stability of MonB resulted from several factors and we believe that a major factor is due to a specific mutation in the hydrophobic core of MonB. The mutation of Tyr 45 to alanine in the aromatic cluster of MonB acts to destabilize its hydrophobic core and thus the entire protein. Table (3.2) highlights the three key positions in Gβ1 variants: 43, 45, and 52. All the Gβ1 variants have aromatic residues in these three positions except MonB which has the Ala mutation at position 45 and aromatic residues in positions 43 and 52.
Figure 3.4 Correlation between protein stability and proteolysis resistance in this in vitro assay. In the in vitro proteolysis assay, the three Gβ1 variants (Protein G W/T, MonB, and MonA) were subjected to proteolytic enzymes such as, trypsin and chymotrypsin with time increments of thirty and sixty minutes. The extent of proteolysis was analyzed on reversed phase HPLC.
Design of Gβ1 variants of variable stability that originated from the ‘parent’ variants: Nine Gβ1 (β1 domain of streptococcal protein-G) mutant variants with different T_m’s were engineered for testing in the chimeric screen. Six of the nine variants originate from three ‘parent’ variants: the wild-type domain (Gβ1-WT), and two variants termed monomer A (MonA) and monomer B (MonB).

Of the six additional variants four were single point mutants of Gβ1-WT: 1) Gβ1-W43A, 2) Gβ1-W43V, 3) Gβ1-W43Y, and 4) Gβ1-Y45A and two variants were derived from MonB-WT: 1) MonB-A45V and 2) MonB-A45Y. The sequences for all nine variants are listed in Table 3.4. The rationale for the mutations that gave rise to the six mutants stems primarily from work performed by Kobayashi et al. (Kobayashi 2000) and earlier work by Blanco et al (Blanco 1994). Both groups analyzed a 16 residue peptide that encompassed the C-terminal β-hairpin of Gβ1 (residues 41 to 56). It was initially demonstrated that the isolated peptide significantly populates a native-like β-hairpin structure in aqueous solution and that three aromatic residues (W43, Y45 and F52) cluster to form a small hydrophobic core (Blanco 1994) (Figure 3.3). Subsequent work demonstrated that upon mixing the C-terminal peptide with a peptide that comprised the first forty Gβ1 residues a complex was indeed formed and exhibited spectroscopic and thermodynamic properties not unlike that of the intact domain (yet of expected lower stability). Alanine scanning of the C-terminal peptide further illustrated the critical structural role the cluster of aromatic residues plays as mutation of any of these residues to alanine completely abrogated complex formation (Kobayashi 2000).
The rational design of the six additional variants was based on mutation of two of the aromatic residues that make up the C-terminal cluster (\textit{i.e.}, W43 and Y45). Mutation of W43 to alanine reduced the \( T_m \) from the wild-type value of 85º C to 57º C thus confirming the critical role this residue plays in stabilizing the wild-type domain.
Table 3.4: Amino Acid Sequences for Gβ1-WT and mutant variants. Mutation(s) from the sequence just above each line are indicated in bold. The 8 mutations for MonB-WT and the 12 for MonA are bolded. Mutation(s) of MonB-WT sequence are indicated in bold and the MonB-WT positions are underlined.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>Ta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gβ1-WT</td>
<td>-</td>
<td>MYKLLNLGLKGETTTEAVDAATAKVEFQYANDNGVDGEWTYDDATKFTVTE</td>
<td>85°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gβ1</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAATAKVEFQYANDNGVDGEAYDDATKFTVTE</td>
<td>57°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W43A</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAATAKVEFQYANDNGVDGEWTYDDATKFTVTE</td>
<td>66°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gβ1</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAATAKVEFQYANDNGVDGEYDDATKFTVTE</td>
<td>64°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W43V</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAATAKVEFQYANDNGVDGEWTYDDATKFTVTE</td>
<td>55°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gβ1</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAATAKVEFQYANDNGVDGEWTYDDATKFTVTE</td>
<td>55°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y45A</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAATAKVEFQYANDNGVDGEWTYDDATKFTVTE</td>
<td>55°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MonB-WT</td>
<td>-</td>
<td>MYKLLNLGLKGETTTEAVDAADVFAQYAADNGVKGENTDAEXATKFTVTE</td>
<td>38°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MonB</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAADVFAQYAADNGVKGENTYDEATKFTVTE</td>
<td>54°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A45V</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAADVFAQYAADNGVKGENTYDEATKFTVTE</td>
<td>72°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MonB</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAADVFAQYAADNGVKGENTYDEATKFTVTE</td>
<td>64°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A45Y</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAADVFAQYAADNGVKGENTYDEATKFTVTE</td>
<td>64°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MonB</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAADVFAQYAADNGVKGENTYDEATKFTVTE</td>
<td>64°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C0RD5</td>
<td></td>
<td>MYKLLNLGLKGETTTEAVDAADVFAQYAADNGVKGENTYDEATKFTVTE</td>
<td>64°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MonA</td>
<td>-</td>
<td>MYKLLNLGLKGETTTEAVDAADVFAQYAADNGVKGENTYDEATKFTVTE</td>
<td>&gt;95°</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
We predicted that mutation of W43 to valine would have less of an effect on stability as valine is larger and more hydrophobic than alanine and thus would more efficiently fill the space previously occupied by the tryptophan side chain. This prediction proved true as the reduction of the measured $T_m$ was not as drastic and resulted in a $T_m$ of 66º C for the W43V mutation. Interestingly, and somewhat unexpectedly, the mutation of W43 to tyrosine was slightly more destabilizing in comparison to the valine mutant as it resulted in a $T_m$ of 64º C. Finally, the critical role Y45 plays in stabilizing the overall fold was also confirmed upon mutation to alanine as it resulted in the reduction of the $T_m$ to 55º C for the Gβ1-Y45A mutant (Figure 3.5).

The interfacial design that produced the MonB sequence resulted in the substitution of Y45 to alanine which we surmised was the primary cause of the reduction in the thermal stability from 85º C to 38º C. Substitution back to tyrosine confirmed this assumption as the MonB-A45Y revision mutant resulted in an increase in the $T_m$ from 38º C to 72º C. Substitution of the intermediately sized, non-polar residue valine at position 45 (MonB-A45V) resulted in a predictably more modest increase in thermal stability from 38º C to 54º C.

An additional MonB variant, MonB-ORDES, was generated by performing computational mutagenesis on three select positions (23, 27, and 45) using the ORBIT design algorithms. These positions were chosen based on the proven importance of position 45 for the structural integrity of the Gβ1 fold and because positions 23 and 27 are in close spatial proximity to 45 (Figure 3.3). The Gβ1-WT sequence consists of A23, E27 and Y45 whereas the MonB-WT sequence has Ile, Ala and Ala at the
respective positions. The ORBIT calculation performed on the free MonB-WT structure returned Lys, Lys and Phe at these positions and resulted in an increase in the Tm from 38°C to 64°C.
Figure 3.5. Gβ1-WT structure and its variants. Side chains are shown for the residues that make up the C-terminal aromatic cluster (43, 45, 52)
Circular Dichroism

Far ultraviolet (UV) circular dichroism (CD) spectra were collected for ten variants and the spectra collected on the nine mutant variants are highly similar to that of Gβ1-WT (Figure 3.6). These results indicate that the mutant variants most likely maintain the βα fold topology of the wild-type domain. The melting temperatures for all ten variants were measured using standard thermal denaturation monitored by CD at 218 nm (Table 3.4 and Figure 3.7). The slopes of the melting curves are also fairly similar and indicate cooperative unfolding and thus provide further evidence that the mutant variants maintain the general wild-type fold.
Figure 3.6. Far UV Circular Dichroism Scans for Gβ1-WT and Mutant Variants. The CD spectra of the nine variants are nearly identical to that of the wild-type protein: MonB-WT, MonB-A45V, Gβ1-Y45A, Gβ1-W43A, Gβ1-W43Y, MonB-ORDES, Gβ1-W43V, MonB-A45Y, Gβ1-WT, MonA (data not normalized).
Figure 3.7. Circular Dichroism Thermal Unfolding Curves for Gβ1-WT and Mutant Variants. The CD thermal denaturation of the ten variants monitored by CD at 218 nm (normalized as described in Methods). ■ MonB-WT, ▲ MonB-A45V, □ Gβ1-Y45A, □ Gβ1-W43A, □ Gβ1-W43Y, □ MonB-ORDES, □ Gβ1-W43V, □ MonB-A45Y, □ Gβ1-WT, □ MonA.
**Principle of the screen**

The screen entails the use of an engineered chimeric construct that consists of three covalently linked domains. This construct includes a constant N-terminal DNA binding domain, a variable central test protein, and a constant C-terminal transcriptional activation domain ($\lambda$-cI-test protein-RNAP$\alpha$). Upon expression, the resulting 3-domain chimera ($\lambda$-cI-test protein-RNAP$\alpha$) acts as a functional transcription factor and binds the $\lambda$ operator (located upstream of reporter genes located in the F’ episome) through interactions with the $\lambda$cI domain. Endogenous RNA polymerase is recruited through interactions with the covalently attached RNAP-$\alpha$ domain. Binding of the 3-domain chimera turns on the transcription of the reporter genes (i.e., $\beta$-lactamase and $\beta$-galactosidase) and the degree by which the reporter genes are upregulated is proportional to the thermal stability and conformational specificity of the intervening (central) protein variants.

If the central target protein is relatively unstable it is more flexible and thereby allows the DNA binding domain and the transactivation domain to achieve optimal orientation for transcriptional activation of reporter gene (i.e., $\beta$-lactamase). Therefore, unstable variants will survive on increased amounts of carbenicillin plates (Figure 3.8). Conversely, if the central target protein is highly stable it is too rigid to allow the DNA binding domain and the transactivation domain to achieve the proper orientation for transcriptional activation and thus is a less efficient transcription factor. Stable variants will not survive on increased amounts of carbenicillin (Figure 3.9).
Figure 3.8. Chimeric Vector and Construct. The genes for the Gβ1 variants were subcloned into the chimeric vector between the genes for DNA binding domain (λcI) and activation domain (RNAPα) which resulted in the three domains linked together as follows - λcI-Gβ1-RNAPα. Binding of the protein chimera to the λ-operator up-regulates the reporter genes β-lactamase (AMP). If the central target protein is highly stable it is too rigid to allow the DNA binding domain and the transactivation domain to achieve the proper orientation for transcriptional activation and thus is a less efficient transcription factor. Stable variants will not survive on increased amounts of carbenicillin.
Figure 3.9. Chimeric Vector and Construct. The genes for the Gβ1 variants were subcloned into the chimeric vector between the genes for DNA binding domain (λcI) and activation domain (RNAPα) which resulted in the three domains linked together as follows - λcI-Gβ1-RNAPα. Binding of the protein chimera to the λ-operator up-regulates the reporter genes β-lactamase (AMP). If the central target protein is relatively unstable it is more flexible and thereby allows the DNA binding domain and the transactivation domain to achieve optimal orientation for transcriptional activation of reporter gene (i.e., β-lactamase). Therefore, unstable variants will survive on increased amounts of carbenicillin plates.
Screening Mutants on Increasing Amounts of Reporter Antibiotic

The genes for Gβ1-WT, MonA, and MonB-WT were cloned into the chimeric construct and initially tested for growth on agar plates that contained increasing amounts of the reporter antibiotic carbenicillin (Figure 3.10). It is evident from the relatively large disparity in growth rates and colony numbers for these variants that the chimera that contains the least stable variant, MonB-WT, clearly functions as a most effective transcription factor. These results indicate an inverse relationship between growth efficiency and thermal stability as the most stable variant, MonA, gave rise to the fewest colonies, Gβ1-WT slightly more, and the least stable variant, MonB-WT, the greatest number of colonies.

Tests of bacterial growth rates for the additional six mutant variants (Table 3.4) were conducted in liquid media where cells were grown for a set amount of time (e.g., 4 hr.) in media that contained a fixed amount of reporter antibiotic (e.g., 3000 µg/mL carbenicillin). In general the results (Figure 3.11) agree with the trend observed for the plate screening i.e., the chimeras that contain variants of lower thermal stability (and thus potentially higher intrinsic flexibility) up-regulate the reporter genes to a greater extent as compared to those of higher thermal stability. Although a general trend is evident there are a number of variants that up-regulate the reporter gene to an amount that is comparable to the least stable variant, MonB-WT, yet these variants (i.e., MonB-A45V, Gβ1-W43A, Gβ1-Y45A) have Tm’s that are higher by at least 15º C. There are a number of potential reasons for this finding, as for example the fact that screen readout may be more highly sensitive to differences in Tm’s for variants at the higher end of the thermal stability range (e.g., greater than 60º C). Furthermore, in
addition to the screen being sensitive to thermal stability it may also be sensitive to other physical characteristics of the variants, such as intrinsic flexibility. To explore the potential role flexibility plays in this screen, 2D heteronuclear NMR spectra were collected on all variants.
Figure 3.10 Reporter Gene Expression for Gβ1 Variants. The upper most row corresponds to control plates that contain no reporter antibiotic. The middle row contains plates with 750 µg/mL of carbenicillin and the lowest row 1000 µg/mL. The 1st column corresponds to the vector without a Gβ1 variant insert (i.e., λcI fused to RNAPα with a short intervening insert; referred to as PrP) and the other variants are listed below each respective column.
Figure 3.11 Correlation between Bacterial Growth and Protein Thermal Stability. Growth extent for the panel of mutant variants of the Gβ1 domain. The height of each bar corresponds to the bacterial cell density for each mutant after growth for four hours in liquid LB media that contained 3000 µg/mL carbenicillin. The O.D.\_600 value corresponds to the ability to turn on the reporter gene and thus the flexibility of each variant. The parent name, the mutation(s) that gave rise to each variant, and the Tm’s are listed below each bar.
NMR Spectroscopy

[$^1$H, $^{15}$N] HSQC spectra were collected for ten variants (four representative 2D HSQC spectra are illustrated in Figure 3.12). The significant signal dispersion and high resolution exhibited in the spectra recorded for Gβ1-WT is nearly unparalleled and reflects both its compact overall structure and, as noted in the original report of the Gβ1 structure, the fact that approximately 95% of the residues are involved in regular, well-ordered secondary structure elements (Gronenborn 1991). Even though there are differences in apparent conformational flexibility based on the in vivo screen results, and different Tm’s, all variants exhibit comparable signal dispersion relative to Gβ1-WT. This finding is in agreement with the CD results and further indicates that the overall Gβ1 fold topology is likely maintained for all variants.

Although signal dispersion is maintained for all variants there is a trend in signal resolution that generally parallels the screen results, i.e., variants of lower stability exhibit lower resolution (broader line widths) in comparison to those of higher stability. This trend is most evident in the spectra for the least stable variant, MonB-WT, and for the next to least stable variant, MonB-A45V (Figure 3.13 e and 3.13 d respectively). In comparison to the Gβ1-WT spectrum (Figure 3.13 j) there are sections in these spectra in which overlapping peaks obscure the baseline (illustrated by under-bars in Figure 3.13). In addition, MonB-WT exhibits evidence of multiple conformations that exchange slowly on the NMR time scale as there are multiple sets of peaks in the 2D HSQC and at least 4 peaks in the downfield region for the tryptophan indole proton of W43 and the amide proton of F52 (i.e., the peaks between ~10 and ~10.5 ppm in Figure 3.13 c). This finding was not unexpected as MonB-WT
is known to self associate and form amyloid-like fibers when incubated and agitated near its melting temperature (Shukla, Marino et al. 2004).

There is a series of three single point mutants of Gβ1-WT for which the in vivo screen results nicely reflect the peak shape of the associated NMR spectra (i.e., Gβ1-W43A, Gβ1-W43Y, and Gβ1-W43V). The peak corresponding to the backbone amide of F52 is considerably broadened for these mutants relative to Gβ1-WT (Figures 3.13 b and 3.13 g-i). The relatively extreme broadening of this peak indicates greater internal motion of backbone and side chain atoms in this region and further illustrates the important role the centrally located W43 side chain plays in stabilizing the structure of the Gβ1 domain. In light of these mutations the increased dynamics are not unexpected, i.e., mutation of tryptophan to alanine results in an internal cavity that corresponds to the difference between the volume of the tryptophan side chain versus that for alanine (237.6 Å and 91.5 Å respectively) (Chothia 1975). The broadened peak for F52 likely reflects internal motion of backbone and side chain atoms as they rearrange, and/or collapse, to compensate for the resulting cavity. The increased dynamics are due to fluctuations between structural states that are similar in energy and thus not resolved with a single structural solution. The degree of broadening for the F52 amide nicely reflects differences in the melting temperatures as well as the in vivo screen results, i.e., the W43A mutant has the lowest Tₘ (57° C), the highest in vivo screen readout and the broadest F52 NMR peak; the intermediate variant, W43Y, has a higher Tₘ (64° C), exhibits an intermediate in vivo screen readout and also has a broad F52 NMR peak; the most stable variant in this series, W43V, has a slightly higher Tₘ (66° C), reduced in vivo screen readout and an F52 NMR peak that,
although still broad, is sharper than that for the other two variants. The result that the F52 peak shape is considerably broader for these three mutants in comparison to the less stable variants MonB-WT and MonB-A45V is an excellent example of how local flexibility can often be decoupled from overall thermal stability and demonstrates that higher thermal stability does not always correlate to greater conformational specificity (i.e., decreased number of possible conformations). The higher T\textsubscript{m}'s for the W43 mutants likely reflects increased hydrophobic contacts relative to the less stable variants yet the NMR results, in conjunction with the \textit{in vivo} screen readout, reflect the fact that the structural integrities of these variants are fairly similar.

The concurrent behavior in the trends observed for the NMR peak shape and the \textit{in vivo} screen does not hold for one particular variant, G\textbeta1-Y45A. This variant exhibits fairly high \textit{in vivo} screen readout (and has a relatively low T\textsubscript{m} - 55\degree C) yet has relatively sharp NMR lines (Figure 3.13 f). For this variant the F52 amide peak is shifted furthest upfield in both the proton and nitrogen dimensions (Figure 3.13 d). This likely reflects a repositioning of the side chain of F52 (as it is in close proximity to Y45 in the wild-type structure) yet may also be due to additional backbone and side chain rearrangement to compensate for the loss of the tyrosine side chain at position 45.
Figure 3.12. Representative $[^1\text{H}, ^{15}\text{N}]$ HSQC spectra. a) Gβ1-WT, b) Gβ1-W43A, c) MonB-WT, d) Gβ1-Y45A. The insets illustrate either just the peak for the backbone amide of F52 or both peaks that correspond to W43 indole and the backbone amide of F52.
Figure 3.13. 1D Projections of the 2D HSQC Spectra. This figure displays the 1D projections of the 2D HSQC spectra for 10 variants - a) MonA, b) MonB-ORDES, c) MonB-A45Y, d) MonB-A45V, e) MonB-WT, f) Gβ1-Y45A, g) Gβ1-W43Y, h) Gβ1-W43V, i) Gβ1-W43A, j) Gβ1-WT. Regions of relatively poor resolution are indicated by a black under-bar and the resonances for backbone amide proton for F52 are indicated by an arrow.
**Screening Mutants using chromatographic assays**

β-galactosidase and β-lactamase activities were tested with four different substrates. The assays were performed with either MonA or MonB (plate assays) or by using Mon A, MonB and three other variants of different thermal stabilities.

**X-gal indicator plates**

The two variants at the extremes of the Tm range (MonA > 100° C and MonB 38° C) were transformed into BacterioMatch two-hybrid system reporter strain competent cells (Stratagene) and plated on X-gal indicator agar plates (Figure 3.14 a). Regarding the color of the resulting colonies we did not observe a difference in colony color that surpassed the differences in growth rates for these two variants when grown on plates that contained carbenicillin.

**MacConkey plates**

MacConkey medium contains pH indicators that are used to differentiate colonies that ferment sugar compared to colonies unable to ferment sugar. On MacConkey-Lactose plates, colonies that ferment lactose will turn the pH indicator red while colonies that cannot ferment lactose remain white. β-galactosidase hydrolyzes lactose. The MonA and MonB variants were transformed into BacterioMatch two-hybrid system reporter strain competent cells and plated onto MacConkey agar plates (Figure 3.14 b). In a manner similar to that observed for the X-gal indicator plates we did not observe a difference in colony color that surpassed
the differences in growth rates for these two variants when grown on plates that contained carbenicillin.
Figure 3.14. Screening using chromatographic assays: X-gal indicator plates (A). MacConkey plates (B)
ONPG (O-Nitrophenyl B-D-Galactopyranoside)

ONPG (O-Nitrophenyl B-D- Galactopyranoside) is a chromogenic substrate which is colorless but when hydrolyzed by β-galactosidase it is converted to galactose and o-nitrophenol, which is yellow. The amount of o-nitrophenol that has been formed is a function of β-galactosidase activity and is measured by its absorbance at 420 nm. This analysis was performed *ex vivo* and entailed growing the cells to a specified optical density, followed by cell lysis and then incubation of the treated lysate in an appropriate buffer that contained the ONPG substrate. We did observe differences in activity when comparing five Gβ1 variants of different thermal stabilities and the differences did correlate to the thermal stability (Figure 3.15). But in a manner similar to that observed for the plate assays above, the differences did not exceed the differences observed in growth rates for the different variants when grown on plates (or in liquid media) that contained carbenicillin.
Figure 3.15 Correlation between β-galactosidase activity and protein thermal stability. The height of each bar corresponds to optimal density for each mutant at 420 after ONPG testing for 10 minutes. The height of the bar corresponds to the ability to turn on the reporter gene which is β-galactosidase. The parent variants name, the single mutation that gave rise to each variant, and the Tm’s are listed below each bar.
CENTA™ β -Lactamase Substrate

CENTA is a chromogenic β-lactamase substrate that upon hydrolysis of its lactam ring by β-lactamase a color change occurs from a light yellow with a peak maximum of 340 nm to a chrome yellow with a peak maximum of 405 nm. This analysis was performed on two Gβ1 variants of different thermal stability (MonA, and MonB) and differences were observed that also correlated to the thermal stability of each variant. But again, like that observed for the β-galactosidase assays, the measured difference did not exceed the differences observed in growth rates for the different variants when grown on plates (or in liquid media) that contained carbenicillin.
3.5 Discussion

The motivation for the screen design was driven primarily by the significant disparity in the thermal stability between the least and most stable variants (i.e., MonA and MonB-WT). We assumed that the large difference in the measured $T_m$’s corresponded to differences in structural integrity and thus inherent malleability. This assumption was confirmed by preliminary in vitro proteolysis experiments performed on the three ‘parent’ $G_{\beta 1}$ variants (i.e., MonA, MonB-WT and $G_{\beta 1}$-WT). It was initially assumed that, in the context of the chimeric construct, the measured differences in stability might either manifest in variable proteolytic resistance in vivo (i.e., less stable variants would be cleaved at a higher rate and thus up-regulate the reporter genes to a lesser extent) or to differences in the overall structural integrity (stability) of the entire chimeric construct. The results reported herein clearly indicate the latter to be the case. In addition, there were no detectable differences in proteolytic resistance observed in vivo based on western blot analysis of the chimeric constructs that contained the MonA, $G_{\beta 1}$-WT and MonB-WT sequences (data not shown).

A drawback of the current rendition of the screen is that the less stable variants are more efficient at up-regulating the reporter gene and thus have a selective advantage over variants of greater stability. This point raises an important question. Why is it that the more stable variants function as less efficient transcription factors in the context of this chimeric screen? Insights into what might be happening on a molecular level are derived from two sources.
The first stems from work in which Gβ1 was displayed on the surface of filamentous bactériophage with the goal of exploring the structural integrity of a library of mutant variants (O'Neil, Hoess et al. 1995). The screen proved to be successful at selecting folded variants yet did not result in the selection of variants with significantly increased thermal stability. The inability to select for more stable variants may be due to inherent structural properties of the tight Gβ1 fold as an unexpected discovery demonstrated that phage displaying stable Gβ1 variants (e.g., wild type) gave rise to unnaturally small plaques yet unstable mutants gave rise to the normal larger plaque phenotype (O'Neil, Hoess et al. 1995). The final step in phage assembly entails transport of phage coat proteins to the bacterial membrane where they are assembled on viral DNA and extruded through the bacterial membrane. Since the Gβ1 variants were fused to the gpIII coat protein for display purposes it was hypothesized that the presence of the Gβ1 domain may inhibit transport of gpIII across the bacterial membrane. Mutations that disrupt the structural integrity of the tightly folded Gβ1 domain may allow assembly to occur more efficiently imparting a growth advantage for phage harboring less stable variants. In light of the inverse relationship we observe between intrinsic stability and transcriptional activation in the in vivo chimeric screen we consider this explanation to be highly plausible.

In addition to the Gβ1 domain other highly stable proteins have proved refractory to display on filamentous phage (Wilson and Finlay 1998). For example, designed ankyrin repeat proteins that are expressed in soluble form with high yields in E. coli, show high thermodynamic stability and fast cooperative folding and are resistant to proteolysis are not efficiently displayed on filamentous phage (Steiner,
Forrer *et al.* 2006). This problem was successfully alleviated upon altering the secretion pathway from one in which translocation across the bacterial membrane occurs post-translationally to one in which it occurs cotranslationally thus providing direct empirical evidence that highly stable protein folds can inhibit, or partially block, certain cellular processes.

A second source of insight as to why the more stable Gβ1 variants ultimately function as less efficient transcription factors in the chimeric screen is provided upon analyzing the crystal structures of all the components that comprise the chimeric construct (Figure 3.16 c-f). Transcriptional activation begins at the level of DNA where interactions occur between the N-terminal DNA binding domain of λcI and operator DNA (Figure 3.16 f). The N-terminal DNA binding domain of λcI is known to weakly dimerize and the crystal structure (Pabo and Lewis 1982) reveals that there are regions where the domains contact one another and thus may form favorable intermolecular contacts. The N-terminal domain of λcI is separated from the C-terminal ‘dimerization’ domain (Figure 3.16 e) by a relatively unstructured ~40 amino acid linker (not shown). During the phage life cycle the C-terminal domain of λcI mediates dimerization as well as the interactions responsible for the cooperative binding of two repressor dimers to pairs of operator sites (Bell, Frescura *et al.* 2000). The C-terminus of λcI is connected to the N-terminus of the Gβ1 variants through a four amino acid linker (Ala3Ser) and is located at the end of the small symmetrically related helical segments found in close proximity to the dimer interface (illustrated as blue and red ribbons in Figure 3.16 e). Finally the C-terminus of the Gβ1 variants is connected to the N-terminus of RNAPα (Figure 3.16 c) by a three amino acid linker
(Gly Asn Ser). Here again it is quite apparent from the crystal structure (Zhang and Darst 1998) that the N-terminal domain of RNAPα must dimerize properly to form the correct orientation that provides the specific geometry necessary for recruitment of the other highly intertwined components of the RNA polymerase (Figure 3.16 a) (Vassylyev, Sekine et al. 2002).

We believe that the intervening Gβ1 test variants act in a sense as ‘molecular rheostats’ providing variable resistance in allowing the attached domains to achieve the correct dimer orientations. Stable Gβ1 variants likely maintain the tight βαβ fold that possibly inhibits, or partially blocks optimal dimerization of the other components of the chimeric construct. On the other hand the less stable variants are more likely to exist in a dynamic equilibrium between partially unfolded states that more readily allow the optimal intermolecular interactions necessary to achieve the proper shape and chemical complementarily for enhanced recruitment of the endogenous RNA polymerase. In addition, lack of structural stability may enhance reporter gene expression by enabling the chimeric construct, bound at the λcI operator site, to more effectively position RNAPα in proximity to promoter elements, thus stabilizing the binding of RNA polymerase and activating transcription from the test promoter.

The exploitation of readily available, well characterized elements of transcriptional machinery from different organisms has provided a novel method to explore the structural determinants of a particular protein fold (Gβ1) - a fold that continues to provide insights into factors important for protein structure as well as folding. The unique application of these transcriptional elements represents new possibilities for the creation of novel combinatorial screens which should provide yet
more opportunities to rapidly and accurately explore large regions of protein sequence space.

Acknowledgement:

Figure 3.16 Model of the Chimeric Construct. (a) Surface rendering of the crystal structure of the bacterial RNA polymerase holoenzyme from T. thermophilus (1IW7). The coloring scheme for the subunits is as follows: two α subunits - olive green and dark green; β - light gray; β’ - dark gray; ω - orange; σ - red. (b) A static schematic illustrating the connectivity of the domains that make up the chimeric construct. (c-f) Backbone ribbon depictions of the crystal structures of the actual domains of the chimeric construct - (c) N-terminal domain of the α-subunit of the E. coli RNA polymerase (2DF); (d) the β1-domain of protein G (1PGA); (e) the C-terminal dimerization domain of the λcI repressor (1F39); (f) The N-terminal domain of the λcI repressor in complex with λ operator DNA (1LMB).
IV. CHARACTERIZATION OF PROTEIN STABILITY USING EXPERIMENTAL AND COMPUTATIONAL METHODS

4.1 Abstract

One of the primary goals of protein design is to engineer proteins with improved stability. Protein stability is a key issue for chemical, biotech, and pharmaceutical industries. The development of robust proteins / enzymes with the ability to withstand the harsh conditions of industrial operations is of high importance. Two strategies are currently being employed to achieve this goal. The first is directed evolution, in which libraries of variants are searched experimentally for clones possessing the desired properties. The second is computational methods that utilize protein design algorithms to perform in silico screening for stable protein sequences. Here, we used gene libraries of an unstable variant of streptococcal protein G (Gβ1) and an in vivo screening method to identify stabilized variants. Many variants with strongly increased thermal stabilities were isolated and characterized. Concomitantly, computational techniques and protein design algorithms were used to perform in silico screening of the same destabilized variant of Gβ1.
4.2 Introduction

The increased demand for proteins (e.g., enzymes) in the pharmaceutical, chemical and biotech industries has spurred the search for proteins with improved intrinsic stability. The stability of proteins is important when used under a variable conditions, such as high temperature, pH, oxidative and reductive stress. These conditions are typical in harsh industrial processes and biomedical applications. The stability of proteins is therefore an important economical factor (Kirk, Borchert et al. 2002). Many studies have focused on the design of stable, tightly packed proteins with novel sequences using rational design or directed evolution (Magliery, 2004). However, increasing the stability of a protein is a complex process which can be difficult and time consuming when engineered rationally (Eijsink, Gaseidnes et al. 2005). Faster results can sometimes be obtained using directed evolution. Directed evolution is based on a variety of different mutagenesis methods combined with high-throughput screening/selection techniques. To date, considerable efforts have been made to create efficient high-throughput in vivo screening/selection techniques to interrogate large combinatorial libraries of protein variants for increased stability or solubility (MacBeath, Kast et al. 1998; Bai and Feng 2004; Hecht, Das et al. 2004; Magliery and Regan 2004; Barakat, Carmody et al. 2006; Philipps, Hennecke et al. 2003; Waldo 2003; Auf der Maur, Tissot et al. 2004; Cabantous, Pedelacq et al. 2005). For comprehensive reviews see (Magliery and Regan 2004; Magliery and Regan 2004). In addition to these screens, computational design methods have also been used either outright (Dahiyat and Mayo 1997; Malakauskas and Mayo 1998; Kuhlman, Dantas et al. 2003; Korkegian, Black et al. 2005; Ashworth, Havranek et al. 2006), or
in combination with *in vivo* screens (Hayes, Bentzien et al. 2002; Wunderlich, Martin et al. 2005), to design and select for protein variants with increased intrinsic stability and/or improved function.

Here, we compared and contrasted the performance of evolutionary protein stabilization (*i.e.*, using an *in vivo* screening method) to computational stability design (*i.e.*, an *in silico* screening method). The *in vivo* chimeric screen for protein stability was used to identify stabilizing residues at three positions in the relatively unstable MonB variant (Barakat, Carmody et al. 2006, Chapter three). Several stabilized variants were identified, indicating that a number of different combinations of residues at the targeted residue positions (*i.e.*, 23, 27, and 45) lead to higher stability relative to MonB. In parallel, protein design algorithms [*i.e.*, the ORBIT (Optimization of Rotamers Based on Iterative Techniques) suite of protein design algorithms] were used to virtually generate a library by randomizing the same three positions that were physically mutated. The resulting library was computationally screened using a molecular mechanics force field followed by application of the Dead-End Elimination theorem (DEE) which is capable of identifying the optimal sequence (and conformation) of lowest calculated energy. This optimal conformation is referred to as the global minimum energy conformation (GMEC). The *in vivo* selected and computationally derived variants have mostly aromatic and hydrophobic residues at the three mutated positions. These amino acids function to stabilize the hydrophobic core and to fill the cavity that has been created due to the mutation of Tyr 45 to alanine in the core of MonB. Thermodynamic analysis was performed on the variants obtained from the *in vivo* screen, in addition to the computationally derived variant,
and of the 10 variants examined two were more stable than the sequence derived from the computational design.

Starting with the GMEC sequence \((i.e., \text{Lys 23, Lys 27, Phe 45})\), we also applied Monte Carlo simulation annealing to produce a rank-order list of the 5000 low energy sequences. Out of the ten experimentally selected variants seven sequences appeared in the Monte Carlo rank-order list. The three variants that were not listed had either Pro or Gly at position 23. Pro and Gly were omitted from the computational library in order to avoid steric restriction (Pro) or excessive structural freedom (Gly). The energies for the experimentally selected variants were also calculated individually using ORBIT by forcing their amino acids at position 23, 27 and 45. The resulting energies were extremely close to those obtained from the Monte Carlo rank-order list.
4.3 Materials and Methods

Construction of the Monomer B library

The MonB library genes were constructed essentially using Recursive PCR with overlapping oligonucleotides. Four oligonucleotides corresponding to the gene were synthesized as 40-50 mers with 20-nt overlaps and at least 50% GC content at each overlap region (Table 4.1). Two of the oligonucleotides, made with equimolar mixes of phosphoramidites at position 23, 27 and 45. These four oligonucleotides were incubated with 1mM dNTPs (Invitrogen) and 1 Unit of Pfu turbo DNA polymerase (Stratagene). Two amplification runs were done in the PCR. In the first PCR: The oligonucleotides annealed and filled in the gaps in 25 cycles of a cycling scheme of: 30 seconds at 95°C, 30 seconds at 59°C, and 2 minutes at 72°C. 10 µl of this first reaction was added to a second PCR where 50 pmol of the upstream forward and downstream reverse primers were used to amplify the entire gene. The second PCR run included 25 cycles of 1 minute at 95°C, 1 minute at 50°C, and 2 minutes at 72°C.

MonB library genes were created using PCR with overlapping oligonucleotides containing all 8000 mutant sequences. The mutated genes were cloned into pλcI-RNAPα using the engineered restriction sites NotI and EcoRI. However, we found that standard ligation protocols are inefficient, so we used TOPO-TA system to clone the PCR products into TOPO cloning vector (pCR 2.1-TOPO) first. Then the whole TOPO-TA ligation reactions were transformed into high efficiency E-coli (XL1 blue F’ episome competent cells). The transformation reactions were plated on X-gal
plates. White colonies were picked and grown overnight in liquid media at 37º C. The overnight culture of the white colonies was maxi-prepped then subjected to double digestion with NotI and EcoRI. Additionally, 2-domain chimera (λcI-RNAPα) plasmid was digested using NotI and EcoRI enzymes to create sticky ends. Double digested PCR inserts and (λcI-RNAPα) plasmid were subjected to ligation using standard ligation protocols.

**Screening the library using CCK 1000 LB plates (1º screen)**

50 µl of BacterioMatch two-hybrid system reporter strain competent cells (Stratagene) was incubated with 0.85 µl of β-mercaptoethanol on ice for 15 minutes to increase transformation efficiency. 0.85 µl of MonB Library (MonB with random mutations at positions 23, 27, and 45) in chimeric construct (ligation reaction) was added to the competent cells mixed with β-mercaptoethanol. After heat shock, 500 µl of room temp NZY media was added to the cells. Transformation reactions were then incubated in the shaker at 30º C for 1 hour and 45 minutes, at 250 rpm. 90 µl of the transformation reaction was plated onto CCK 1000 LB plates [(kanamycin (50 µg/ml), chloramphenicol (12.5 µg/ml), and Carbencillin (1000 µg/ml))] (~ 5 plates) and one CK LB plate (kanamycin (50 µg/ml), and chloramphenicol (12.5 µg/ml)). The CK LB plate was used to count the number of colonies and to calculate the library size. Plates were incubated overnight at 37º C. Next day, small colonies (slow growing colonies) were picked from CCK1000 LB plates and inoculated into 5ml CK LB media [(LB liquid media with kanamycin (50 µg/ml), and chloramphenicol (12.5 µg/ml)]. Small
colonies cultures were incubated overnight at 37º C, 275 rpm and DNA was isolated using standard DNA isolation protocols.

**CCK 1000 small plates secondary screening (2º screen)**

MonB library plasmids (i.e. isolated plasmids from positive, slow growing colonies), and control plasmids (MonA, MonB-WT, MonB-ORDES in the chimeric construct) were used to transform 10 µl BacterioMatch two-hybrid system reporter strain competent cells (Stratagene) using heat shock transformation methods. 400 µl of NZY media were added after heat shock. Transformation reactions were incubated in the shaker for 45 minutes at 30º C, 275 rpm. 10 µl from each transformation reaction was plated onto small (60 X 15mm plates) CK LB plates (kanamycin (50 µg/ml), and chloramphenicol (12.5 µg/ml)), and small CCK 1000 LB plates (kanamycin (50 µg/ml), chloramphenicol (12.5 µg/ml), and Carbencillin (1000 µg/ml)) LB plates. Plates were incubated overnight at 37º C. Each variant’s CCK 1000 Plate was visualized and the number of colonies was compared to its CK plate and to the control plasmids plates. The comparison helps identify variants that are more stable than MonB-WT. The successful variants (i.e. grow less than MonB-WT on CCK 1000 LB plates) were used in the next screening method (i.e. 96-well plates) to get more accurate measurement of the growth and to eliminate any false positive clones. In addition, selected stable variants were sent for sequencing using PBT-F primer.
96-well plate using CCK 2000 LB media (3\textsuperscript{rd} screen)

MonB library variants (i.e. positive clones that grow less than MonB-WT on small CCK 1000 LB plates) were subjected for the last step of screening using the 96-well plates in CCK 2000 LB media. For each positive MonB library variants, six colonies were picked from CK plates and inoculated in 1ml LB liquid media with kanamycin (50 µg/ml), and chloramphenicol (12.5 µg/ml). Cultures were grown overnight at 37\textdegree C with shaking. Using the overnight cultures (i.e. their OD 600 = 1.0), cells were diluted to O.D. 600 = 0.1 in 180 µl volume (i.e. in 96-wells plate) of LB containing kanamycin (50 µg/ml), chloramphenicol (12.5 µg/ml) and Carbencillin (2000 µg/ml). The 96-well plate was incubated in micro-plate reader for 24 hours, at 37\textdegree C, with continuous shaking at maximum speed. OD 600 was measured for the plate every 15 minutes for 24 hours. Growth curves data for each variant and its six trials were averaged and compared to the average growth curves of MonA, MonB-WT, MonB-ORDES. MonB variants with growth rates the same or less than MonB-ORDES were sub-cloned to express the protein and confirm it stability using CD.

10 ml CCK3000 LB culture screening

All G\textbeta1 variants in the chimeric construct were transformed into BacterioMatch two-hybrid system reporter strain competent cells (Stratagene) and plated onto kanamycin (50 µg/ml) and chloramphenicol (12.5 µg/ml) (CK) LB plates. Colonies were picked from CK plates and grown in 20 ml LB liquid media with kanamycin (50 µg/ml) and chloramphenicol (12.5 µg/ml) overnight at 37\textdegree C. Using the overnight cultures, cells were diluted to O.D. 600 = 0.1 in 10ml volume of LB
containing kanamycin (50 µg/ml) and chloramphenicol (12.5 µg/ml). 10 ml Cultures were incubated at 37º C, 275 rpm for 15 minutes. O.D. 600 was obtained for all cultures to make sure that all cultures had exactly the same O.D. 600 before adding carbencillin to the cultures. Carbencillin (3000 µg/ml) was added and all cultures were incubated in the shaker at 37º C, 275 rpm for 4 hours. O.D. 600 was measured for all cultures after 4 hours.

**Protein Expression and Purification**

For protein expression the genes for the Gβ1 variants were sub-cloned into pET-21a (Novagen) using the restriction enzymes NdeI and EcoRI (oligonucleotides: 5’-GTC CGC GGT CAT ATG ACT ACT TAC AAA-3’ and 5’-GGC GCA G AT GAA TTC TTA TTC AGT AAC TGT AAA-3’) and transformed into BL 21(DE3). After growth and induction the proteins were isolated using a freeze/thaw method (Johnson and Hecht 1994) and purification was accomplished with reverse-phase HPLC using a linear 1% min⁻¹ acetonitrile/water gradient containing 0.1% TFA. Concentrations of all the variants were determined in 6 M guanidine hydrochloride using standard extinction coefficients for the tryptophan and tyrosine residues. Protein purity was verified with standard SDS-PAGE and reverse phase HPLC and the correct molecular weight was confirmed by mass spectrometry.

**Circular Dichroism**

Far ultraviolet (UV) circular dichroism (CD) spectra were collected for all variants and the spectra collected on the nine mutant variants were highly similar to
that of Gβ1-WT. These results indicate that the mutant variants most likely maintain the βαβ fold topology of the wild-type domain. The melting temperatures for all ten variants were measured using standard thermal denaturation monitored by CD at 218 nm. The slopes of the melting curves were also similar and indicate cooperative unfolding and thus provide further evidence that the mutant variants maintain the general wild-type fold.

The CD data were collected on Jasco-810 spectrometer equipped with a thermoelectric unit and using a 0.1 mm path length cell. Protein samples were 50 µM in 50 mM sodium phosphate at pH 6.5. Thermal melts were monitored at 218 nm. Data were collected every 1º C with an equilibration time of 2 min. Far-UV spectra were acquired in the continuous mode at 25º C with 1 nm bandwidth and a 4 s response time. For the thermal denaturation curves the data were normalized by first linearly shifting all points such that [θ]218 value at 5º C equaled zero. Then a scaling factor was obtained for each set by dividing the maximum [θ]218 value for all sets at 95º C (i.e., 53.0) by the [θ]218 value at 95º C for each set. All data points for each set were then scaled by the unique scaling factor calculated for each set.
Table 4.1 Four oligonucleotides corresponding to the MonB Library gene were synthesized as 40-50 mers with 20-nt overlaps and at least 50% GC content at each overlap region

<table>
<thead>
<tr>
<th>Oligo</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oligo: 1 (B1+lib)</td>
<td>5’gcggccgcatctacctataagctgattctgaatggcaagaccctgaaaggtgaaaccagacccgaa</td>
</tr>
<tr>
<td>Oligo: 3 (B3+lib)</td>
<td>5’tatttgcacagtatgcagcggataacggtgtaaggtgatatgg</td>
</tr>
<tr>
<td>Oligo: 2’(B_Lib2)</td>
<td>5’ccgctgcatactgtgcaaatatacNNTgttgttgtggtggtggttacccctt</td>
</tr>
<tr>
<td>Oligo: 4’(B_Lib4)</td>
<td>5’ggaattcctctgtacggtgaacgtcttggtcgcttcgtcNNNtgtccattcacccttaaccacgtt</td>
</tr>
<tr>
<td>Oligo: 5 (B_Primer_F)</td>
<td>5’gcgcactatcgcggccgcacatctac</td>
</tr>
<tr>
<td>Oligo: 6 (B_Primer_R)</td>
<td>5’gcgcacgctcgtggaattcctcttcg</td>
</tr>
</tbody>
</table>
4.4 Results

*In vivo screening strategy*

We used the chimeric *in vivo* screen to identify combinations of stabilizing residues at the three positions described above for the unstable MonB variant (Barakat et al. 2006, Chapter 3). The successful performance of the chimeric *in vivo* screen was verified previously using ten Gβ1 mutant variants that fairly evenly span a melting temperature (T_m) range from 38º C to greater than 100º C (Barakat et al. 2006).

**Creation of the Library of MonB Variants**

A MonB library of mutant variants was created by randomizing the codons of three positions (*i.e.*, 23, 27 and 45) using saturation mutagenesis where all 20 naturally occurring amino acids were allowed at each position. Position 45 was chosen based on the proven importance of position 45 for the structural integrity of the Gβ1 fold. A previous study demonstrated that position 45 is crucial for stability as mutation from Tyr to Ala proved to be a dominant reason for the low stability observed for MonB-WT (mutation back to Tyr increased the T_m from 38º C to 72º C) (Barakat et al. 2006). Therefore, this region of the MonB-WT sequence was targeted for improvements in stability using computational redesign and physical library screening. Positions 23 and 27 were also targeted due to their close proximity to position 45 (Figure 3.3).

The genes for the MonB mutant library were created using PCR with four overlapping oligonucleotides. Two of the oligonucleotides contained degenerate NNN codons at position 23, 27, and 45. The mutated genes were cloned into the chimeric
construct using the engineered restriction sites NotI and EcoRI. During this process we found that normal ligation protocols were inefficient, so a TOPO-TA system was used to clone the PCR products into the TOPO cloning vector (pCR 2.1-TOPO). The whole TOPO-TA ligation reaction was transformed into highly efficient *E. coli* cells (XL1 blue F’ episome competent cells). The transformation reaction was plated on X-gal plates. All white colonies were picked and grown together overnight in liquid media at 37°C. The overnight culture corresponding to the white colonies was maxi-prepped and subjected to double digestion with NotI and EcoRI. Additionally, the chimeric plasmid (λcI-RNAPα) was digested using NotI and EcoRI enzymes. The MonB library was ligated into the 2-domain chimeric plasmid and transformed into *E. coli* competent cells (XL1blue F’ episome reporter gene cells).

To evaluate the diversity of the library, the genes for approximately 60 MonB variants were sequenced and revealed that bases at the random positions were highly degenerate and not biased towards any particular base. At the nucleotide level, we calculated the percentage of each nucleotide (*i.e.*, C, G, T, and A) at the three random positions. Each base was represented by 25% of the total base composition and demonstrated relatively equal diversity in the library at the DNA level. At the amino acids level, none of the sequences analyzed were identical. These results indicate that the designed library likely contained all of the possible mutated sequences and had a molecular diversity of ~8000 different sequences.
In vivo screening for stabilized MonB variants

Screening of the randomized library entailed primary and secondary steps due to the fact that, in the context of the current rendition of the screen, the difference in the growth rate for the most stable variant (MonA) and the least stable variant (MonB) is greater than but still similar to natural variations in bacterial growth rates. The primary screen entailed picking slow growing colonies identified on plates that contained reporter antibiotic. These colonies were then grown in liquid media for the purpose of amplification and isolation of plasmids. In the second screening step the purified plasmids were transformed into the reporter strain and separately plated on small plates to confirm slow growth rates. For colonies that passed the first and second steps an additional step entailed growing the bacteria in a 96 well plate format while monitoring growth rates over 24 hours. Figure 4.2 graphically illustrates the scheme for screening the MonB library.

For the primary screening step approximately 33,000 individual E. coli (XL1blue F’ episome reporter gene cells) colonies, expressing the 3-domain chimera, were plated on media that contained a fixed amount of reporter antibiotic (e.g., 1000 μg/ml carbenicillin). The size of the colonies on reporter plates is generally proportional to the thermal stability of the intervening MonB mutant variant. Small colonies likely represent more stable variants while large colonies represent unstable, more flexible variants. Small colonies (~200), in addition to two large colonies, were picked and grown in liquid media for the purpose of amplifying and isolating the associated plasmids. The isolated DNA that originated from the small colonies was transformed into E. coli reporter cells and re-tested for growth on small reporter plates.
to confirm the slow-growth phenotype observed in the primary screening step. As a control for colony number each of the variants were also grown on plates that contained no reporter antibiotic. Chimeric vectors that contained MonB, MonB-ORDES (the resulting MonB variant derived from the in silico screening with ORBIT, see chapter 3), and MonA were used as controls (Barakat et al. 2006). Colonies that grew more slowly than MonB control colonies were further analyzed by monitoring growth in 96-well plates for 24 hours in fixed amount of reporter antibiotic (e.g., 2000 µg/ml carbencillin). Out of the ~200 variants identified in the primary screening step, 140 passed the secondary step (~60 clones turned out to be false positives upon analysis in the secondary screen) and thus the 140 were grown and observed in the 96-well format. The growth rate measurements in the 96 well format were done in 200 µl volume at 37 ºC, shaking at maximum speed and OD600 readings were measured every 15 minutes for 24 hours. The growth rates of both positive and negative controls were also included in the 96 well plates and consisted of PRP (i.e., the chimeric vector with no insert), MonB, MonB-ORDES, and MonA. Eight variants that consistently exhibited slower growth in comparison to MonB or MonB-OBDES were sequenced and subcloned into the pET21a vector for the purpose of protein expression. Figure 4.1 illustrates the growth extent for all of the variants obtained from the screen in addition to those engineered previously by using rational design methods (Barakat, Carmody et al. 2006).

To test the ability of the screen to identify variants that are less stable than MonB-WT, two fast growing colonies were also isolated, analyzed for growth rates, subcloned into the pET-21a expression vector, expressed, HPLC purified and the
melting temperatures measured with standard CD thermal denaturation. These two variants grew on the reporter plates at a rate that is faster than MonB-WT. In addition, these two variants were analyzed by monitoring their growth in 96-well plate for 24 hours. The two variants (referred to as MonB GND, MonB GKS; the three letter extension refers to the amino acids at positions 23, 27 and 45 respectively) grew faster than MonB-WT. Figure 4.1 illustrates the growth extent for MonB GND, MonB GKS obtained from the screen in addition to those engineered previously by using rational design methods (Barakat, Carmody et al. 2006).
Figure 4.1 Correlation between Bacterial Growth and Protein Thermal Stability. Growth extent for the panel of mutant variants of the Gβ1 domain. The height of each bar corresponds to the bacterial cell density for each mutant after growth for four hours in liquid LB media that contained 3000 µg/mL carbenicillin.
E coli transformation with library

Plasmid isolation from β-lactamase low expression

E coli transformation

Monitoring growth rate for 24 hours in 2000 µg/ml carbencillin

Screen for low growth rate

Cloning into expression vector

Protein expression and purification

Measure Tm using Circular Dichroism

Figure 4.2 *In vivo* screening scheme to identify stable MonB variants
**In silico screening for stabilized MonB variants**

In parallel to generating more stable variants of MonB-WT with the *in vivo* evolution methods described above, computational mutagenesis and screening were also performed on MonB-WT at the same three positions (*i.e.*, 23, 27, and 45). The ORBIT suite of protein design algorithms was used for this purpose (Dahiyat 1997). A MonB mutant library was created and screened by calculating pairwise interaction energies between amino acids side chains and backbone atoms for the three mutated positions (Dahiyat and Mayo 1997). The resulting energies were used as input for a powerful search algorithms called dead-end elimination (DEE) (De Maeyer, Desmet et al. 2000). DEE is capable of identifying one optimal sequence and conformation with lowest energy as determined by the molecular mechanics scoring function. This optimal conformation is referred to as the global minimum energy conformation (GMEC). The resulting sequence consists of Lys, Lys and Phe at positions 23, 27 and 45 respectively and is referred to as MonB-ORDES. Starting from the GMEC sequences, we applied Monte Carlo simulation annealing to produce a rank-ordered list of the 5000 lowest energy sequences.

For the computational calculation, all amino acids (except Gly, Pro, and Cys) were considered at these three randomized positions (*i.e.*, 23, 27, and 45). Prior to running the calculation we determined the identity of 29 residue positions that were in proximity to the three targeted positions. To not bias side chain selection at the three targeted positions, based on ‘wild type’ side chain positioning, the side chains at the additional 29 positions were ‘floated’ (*i.e.*, during the calculation rotameric representations of the ‘wild type’ side chains at the 29 positions were included in the
calculation). The RESCLASS algorithm was used to classify the 29 additional positions as either boundary, core, or surface based on their position in the overall structure. Residues classified as boundary lie at the interface of the buried core and solvent-exposed surface. Positions 1, 18, 25, 29, 31, 33, 43, 50 were classified as boundary residues, positions 3, 5, 20, 26, 30, 34, and 52 were classified as core residues and positions 4, 21, 22, 24, 28, 32, 44, 46, 47, 48, 49, 51, and 53 were classified as surface residues (Dahiyat and Mayo 1997). All these residues were fixed in amino acid identity, but their conformations were allowed to change. Amino acids identity and/or side chain conformations were determined with ORBIT for all 32 residues. All other residues, as well as the backbone, were held fixed. The amino acids sequence of GMEC selected by the ORBIT algorithms contains Lys, Lys and Phe at these positions and, upon generation and analysis in the laboratory, resulted in an increase in the $T_m$ from 38º C to 64º C (Barakat et al. 2006; chapter 3).

In addition to the above side chain selection process we also applied Monte Carlo (MC) simulation annealing, starting with the GMEC sequence and structure, for the purpose of producing a rank-order list of the 5000 lowest energy sequences (Voigt, Gordon et al. 2000). The MC calculations were run so we could explore which alternative amino acid side chains had reasonable energies at the targeted positions and, more importantly, had we obtained such sequences in the in vivo screen. We also desired to determine the calculated energies of the in vivo sequences and ascertain where those energies (and sequences) fell in the MC rank ordered list. The calculated energies of the 5000 rank-ordered sequences ranged from as low as -203.4 (for the original GMEC sequence i.e., KKF) to as high as -152.0 Kcal/mol. The ten variants
selected for from the \textit{in vivo} screen were compared to the 5000 computed variants. Seven out of 10 \textit{in vivo} selected variants were found in the 5000 computed variants. Variants PLL, GKS, GND were not found in the 5000 computed variants due to the fact that they have Pro or Gly residues in the randomized positions and these were not considered during the GMEC or MC calculations.

\textbf{Thermal Stabilities Measured for the Variants Obtained from the \textit{In Vivo} Screen}

Far ultraviolet (UV) circular dichroism (CD) spectra were collected for the 10 variants and the resulting spectra for all but one variant is highly similar to that of G\(\beta\)1-WT (Figure 4.4). These results indicate that the more stable mutant variants most likely maintain the \(\alpha\beta\) fold topology of the wild-type domain. The MonB GND variant is so unstable that the G\(\beta\)1 fold topology is lost and that the structure exists in a random coil state.

The melting temperatures for the nine relatively stable variants were measured using standard thermal denaturation monitored by CD at 218 nm (Table 4.2 and Figure 4.3). The slopes of the melting curves are similar and indicate cooperative unfolding and thus provide further evidence that the mutant variants maintain the overall wild-type fold.

The eight variants (predicted to be more stable than MonB) isolated from the \textit{in vivo} library all have higher melting temperatures than MonB and two are more thermally stable than the variant obtained from the protein design algorithms (\textit{i.e.}, MonB-ORDES). The two \textit{in vivo} variants that originated from fast growing colonies (and thus predicted to be less stable than MonB) have low (\textit{i.e.}, 28\(^\circ\) C for the GKS
variant) to un-measurable melting temperatures (i.e., random coil CD signal for the
GND variant).
Figure 4.3. Far UV Circular Dichroism Scans for Gβ1-WT and Mutant Variants. The CD spectra of the ten variants are nearly identical to that of the wild-type protein except MonB-GND, MonB-GKS, MonBWT, MonB-ISA, MonB-YSV, MonB-YYT, MonB-A45V, MonB-KIY, MonB-PLL, MonB-TQY, MonB-ORDES, MonB-QWF, MonB LYW, MonB-A45Y, Gβ1-WT, MonA.
Figure 4.4. Circular Dichroism Thermal Unfolding Curves for Gβ1-WT and Mutant Variants. The CD thermal denaturation of the ten variants monitored by CD at 218 nm (normalized as described in Methods). MonB-GKS, MonBWT, MonB-ISA, MonB-YSV, MonB-YYT, MonB-A45V, MonB-KIY, MonB-PLL, MonB-TQY, MonB-ORDES, MonB-QWF, MonB LYW, MonB-A45Y, Gβ1-WT, MonA.
Table 4.2 Amino acids at positions 23, 27, and 45 for Gβ1-WT and its variants

| Protein Variants | Variants | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein | Protein |
|------------------|----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
|                  |          | 23      | 27      | 45      | Tm      | Growth in CCK |
| Gβ1-WT           | A E Y    | 85 °    | 0.23 ± 0.01 |
| MonB-WT          | I A A    | 38 °    | 0.95 ± 0.09 |
| MonB OBDES       | K K F    | 64 °    | 0.55 ± 0.03 |
| 1                | L Y W    | 69 °    | 0.41 ± 0.05 |
| 2                | Q W F    | 66 °    | 0.45 ± 0.02 |
| 3                | T Q Y    | 63 °    | 0.62 ± 0.02 |
| 4                | P L L    | 59 °    | 0.63 ± 0.07 |
| 5                | K I Y    | 61 °    | 0.69 ± 0.01 |
| 6                | Y Y T    | 47 °    | 0.81 ± 0.05 |
| 7                | Y S V    | 47 °    | 0.92 ± 0.08 |
| 8                | I S A    | 44 °    | 0.75 ± 0.06 |
| 9                | G K S    | 28 °    | 1.03 ± 0.06 |
| 10               | G N D    | RC      | 1.03 ± 0.15 |
**Measuring protein stability *in silico***

Total energies for each *in vivo* selected variant were calculated using ORBIT. The same computational calculation scheme was used for both the *in silico* mutagenesis and the subsequent DEE selection. However, positions 23, 27, 45 were fixed with the amino acid identity of the sequences obtained from the *in vivo* selected variants and only their conformations (*i.e.*, rotameric descriptions) were allowed to change. The resulting energies were extremely close to those obtained from the MC rank-order list. We found that the average difference between the total energy found using Monte Carlo rank-order list and the total energy calculated using DEE is \( \Delta E = -0.63 \text{ kcal/mol} \).

The ORBIT calculated total energy for each of the 10 variants was compared to the melting temperatures for each *in vivo* variant. Figure 4.5 illustrates the correlation between the ORBIT total energy calculated and the protein thermal stability for each variant. In general, the variants with low total energy (more negative values and thus computational predicted to be more stable) have higher melting temperature while variants with higher total energy (less negative) have lower melting temperatures. Therefore, the computational design algorithm is well suited to estimate the stability of different MonB variants.

In addition, the subtotal energies (*i.e.*, subcategories) for the 10 variants were further analyzed. Energy parameters that were explored and compared include, for example, Van der Waals energies, H-bonding, electrostatics, solvation energy and each relative contribution to the total energy of the structure. We also compared and contrasted the exposed and buried surface areas of each side chain. Table 4.3 contains
the amino acids sequence, the global minimum energy value, and the different energy components of the total energy for each MonB mutant variant. Overall improvements in all these subtotal energies were identified compared to MonB-WT. There is a very favorable correlation between the thermal stabilities of each variant and the total (and subcategories of) energies.
Figure 4.5 Correlation Between protein thermal stability and ORBIT energy calculation
Table 4.3 Amino acids sequence at position 23, 27, and 45, the global minimum energy solution, and the different energy component of the total energy (i.e. van der waals interactions (VDM), H-bonding (Hbond), electrostatics interaction (Elec), polar hydrogen burial (Polar_H), nonpolar exposure (NP_exp) and nonpolar burial (NP-bur) ) for each MonB variant.

<table>
<thead>
<tr>
<th>Protein Variants</th>
<th>23</th>
<th>27</th>
<th>45</th>
<th>Tm</th>
<th>VDW</th>
<th>Elec</th>
<th>Hbond</th>
<th>Polar_H</th>
<th>R_NP_exp</th>
<th>R_NP_bur</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L</td>
<td>Y</td>
<td>W</td>
<td>69 °</td>
<td>-112.9825</td>
<td>-2.8924</td>
<td>-24.1056</td>
<td>32.0000</td>
<td>245.5629</td>
<td>-328.0946</td>
<td>-190.5022</td>
</tr>
<tr>
<td>2</td>
<td>Q</td>
<td>W</td>
<td>F</td>
<td>66 °</td>
<td>-111.9163</td>
<td>-3.0104</td>
<td>-27.6506</td>
<td>34.0000</td>
<td>241.1753</td>
<td>-330.1141</td>
<td>-197.5161</td>
</tr>
<tr>
<td>MonB OEDES</td>
<td>K</td>
<td>K</td>
<td>F</td>
<td>64 °</td>
<td>-111.7790</td>
<td>-3.4017</td>
<td>-24.1056</td>
<td>36.0000</td>
<td>238.5777</td>
<td>-333.6438</td>
<td>-203.3524</td>
</tr>
<tr>
<td>6</td>
<td>Y</td>
<td>Y</td>
<td>T</td>
<td>47 °</td>
<td>-112.9750</td>
<td>-1.9128</td>
<td>-20.2019</td>
<td>32.0000</td>
<td>228.4648</td>
<td>-304.6842</td>
<td>-179.3092</td>
</tr>
<tr>
<td>7</td>
<td>Y</td>
<td>S</td>
<td>V</td>
<td>47 °</td>
<td>-115.1104</td>
<td>-1.9135</td>
<td>-20.2020</td>
<td>32.0000</td>
<td>233.3890</td>
<td>-317.0569</td>
<td>-188.8938</td>
</tr>
<tr>
<td>8</td>
<td>I</td>
<td>S</td>
<td>A</td>
<td>44 °</td>
<td>-105.7937</td>
<td>-2.9049</td>
<td>-24.1056</td>
<td>30.0000</td>
<td>225.1156</td>
<td>-297.9763</td>
<td>-175.8849</td>
</tr>
<tr>
<td>MonB-WT</td>
<td>I</td>
<td>A</td>
<td>A</td>
<td>38 °</td>
<td>-105.0388</td>
<td>-2.9082</td>
<td>-24.1056</td>
<td>30.0000</td>
<td>228.4786</td>
<td>-299.5272</td>
<td>-173.1020</td>
</tr>
</tbody>
</table>
4.5 Discussion

The purpose of this study was to compare two major strategies currently being used to stabilize proteins: *in vivo* and *in silico* screening. We developed gene libraries based on an unstable variant of the Gβ1 domain (MonB-WT) and an *in vivo* screening method, to identify stabilized variants of this mutant Gβ1 variant. For the engineered MonB library, the codons for the three positions 23, 27, and 45 were randomized. Mutant variants of MonB with increased thermal stabilities were found using the *in vivo* chimeric screen. The genes for 10 variants, which exhibited favorable *in vivo* behavior (*i.e.*, relatively high resistance to carbencillin), were sequenced and revealed that amino acids that conferred greater stability were generally degenerate yet biased towards hydrophobic and aromatic at the three randomized positions.

In parallel, protein design algorithms were used in an attempt to generate a stable MonB variant by using sequence optimization at the same positions. The best computed variant ranked third within the panel of the *in vivo* screened variants. This finding is not surprising as a necessary limitation of the computational approach is that the protein backbone must be held rigid for the calculation to remain tractable. Most of the *in vivo* selected variants (seven out of ten) were found in a Monte Carlo rank-ordered list of 5000 additionally computed sequences. Three of the *in vivo* selected variants were not found in the computed list due to the fact that proline, and glycine residues were not considered at the randomized positions.

A comprehensive analysis of the correspondence between the theoretical and experimental stabilities of the selected variants was enabled by the collection of experimental data (*i.e.*, CD) for the computed sequence as well as all the sequences.
obtained from the \textit{in vivo} screen. The thermal stability of each variant was used as a metric for the screen performance and thus the melting temperatures of the selected variants are plotted against growth in carbencillin (Figure 4.6a) as well as the ORBIT calculated energies (Figure 4.6b). The modest correlation between melting temperatures and the ORBIT calculated energies (\textit{i.e.}, $R^2 = 0.65$) indicates that the computational scoring function is capable of screening for and obtaining stable sequences, and predicting actual stabilities (Figure 4.6a). The correlation between the thermal stability of the selected variants and their growth in carbencillin indicates that the \textit{in vivo} chimeric screen is also an excellent predictor of thermal stabilities (Figure 4.6b).

Designing proteins with improved stabilities needs to be a reliable and routine tool. At present, directed evolution methods are fairly robust at generating mutants with improved protein thermostability with confidence and on a reasonable timescale. Concomitantly, the capabilities of computational techniques and \textit{de novo} design are expanding. Emerging design methods that marry the best of the computational and the \textit{in vivo} approaches promise to enable the design of proteins with improved thermostability which is a key tool for protein engineering in the century ahead.
Figure 4.6 Melting temperatures of the selected variants were plotted against their GMEC found in DEE search (a), and their growth in carbencillin (3000 µg/ml) (b)
Acknowledgement:

This chapter, in part, is a reprint of the material as it appears in Nora H. Barakat, Nesreen H. Barakat, John J. Love, “Characterization of protein stability using experimental and computational methods”. Manuscript in final stage of revision.
REFERENCES

Chapter 1


Chapter 2


Chapter 3


Chapter 4


