

Statistical Analysis to Stabilize Proteins

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation
with research distinction in the undergraduate colleges
of The Ohio State University

By

Tran Thi Nguyen

The Ohio State University

May 2010

Project Advisor:

Professor Thomas J. Magliery
Departments of Chemistry and Biochemistry

Committee:

Dr. Thomas J. Magliery

Dr. Amanda S. Simcox

Dr. Jane E. Jackman

Copyright by
Tran Thi Nguyen
2010

Abstract

Consensus design uses statistical information to introduce stabilizing mutations by replacing a wild-type residue with the most common amino acid in a multiple sequence alignment. Consensus residues are found to be stabilizing about 60% of the time, which is much more than the ~1% stabilization from a random mutation. Despite great progress, a comprehensive model to predict protein stabilization is still undetermined. To confront this obstacle, we devised a statistical model using relative entropy (degree of conservation) as a filter to improve consensus design. This knowledge can be used to create an algorithm to improve protein structure, stability, and function. Relative entropy was used to identify mutations in triosephosphate isomerase (TIM) to increase the stability of proteins. TIM protein is an essential component of the ubiquitous glycolytic pathway and its prevalence makes it an excellent model protein for statistical design. TIM mutants at sites with varying relative entropies and secondary structures were engineered to test the hypothesis that more conserved mutations will be more stabilizing. Results from CD thermal melts indicate that highly conserved positions on protein surfaces with low correlations to other positions have the greatest stabilizing effect. To further investigate the mechanism of stabilization, urea melts are being performed to examine unfolding rates. In addition, compensatory mutations are being engineered in high relative entropy mutants with high correlations. Mutations may disrupt the normal interactions between correlated residues, affecting the folding and structure of the protein. This could provide insight into why a few of the high relative entropy mutations did not stabilize the protein as expected. The ability to stabilize proteins has innumerable industrial and therapeutic applications. The use of relative entropy as a statistical analysis can improve techniques for accurately increasing the stability of proteins while obtaining greater depth on the sequence-structure relationship.

Acknowledgements

Thomas J. Magliery, project advisor, Ohio State University departments of Chemistry and Biochemistry

Amanda S. Simcox, oral exam committee, Ohio State University department of Molecular Genetics

Jane E. Jackman, oral exam committee, Ohio State University department of Biochemistry

Brandon Sullivan, OSBP

Magliery Lab

Funding

Ohio State College of Arts and Sciences research scholarships

College of Biological Sciences Dean's Undergraduate research scholarships

Vita

2006.....Pickerington North High School, Pickerington, Ohio
2008-2010.....Undergraduate Researcher, Department of Chemistry,
The Ohio State University

Fields of Study

Major Field: Molecular Genetics

Area of Distinction: Biochemistry

Table of Contents

Abstract.....	ii
Acknowledgements.....	iii
Vita.....	iv
Chapter 1: Introduction.....	1
Protein Stability	1
Current Methods to Stabilize Proteins	3
Consensus Design	6
Statistical Protein Design Using Relative Entropy and Mutual Information..	8
Chapter 2: Materials and Methods.....	10
Selection of Stabilizing Mutants	10
Genetic Engineering.....	12
Protein Purification	14
Circular Dichroism Spectroscopy	15
Additional Relative Entropy Variants and Compensatory Mutations	16
Chapter 3: Results and Discussion.....	17
Engineering of Variants	17
Protein Purification	18
Biophysical Characterization.....	19
Compensatory Mutations	23
Urea Unfolding	24
Remarks	26
References.....	27

Chapter 1- Introduction

Protein Stability

Despite immense research, the fundamental principles of the protein folding problem remain obscure. The protein folding problem was recognized over half a century ago in the works of Anson and Mirsky, who observed that the coagulation of hemoglobin is a reversible process.¹ In the 1950's, Anfinsen's critical research with ribonuclease established that the amino acid sequence of a protein is the sole determinant of its final structure.² He showed that reduced, denatured ribonuclease was capable of spontaneous renaturation *in vitro* upon removal of the denaturant. While this shows that protein structure, and consequently function, is determined by the amino acid sequence, the important question of how the sequence governs protein folding remains unanswered. In addition, proteins can theoretically sample immeasurable numbers of conformations, yet how are they able to fold in fractions of a second? This complexity, known as Levinthal's paradox, states that given the large number of degrees of freedom in an unfolded polypeptide, it is impossible for proteins to assume the innumerable possible conformations to determine their energetic minima.^{3,4} While the general mechanistic understanding of the protein folding problem still eludes scientists,⁵ research has made great progress in protein engineering.

Drexler speculated that it should be possible to design novel proteins before the protein folding problem is solved.⁶ Despite the lack of complete understanding of sequence information and protein folding, great progress has been made to improve protein stability. However, we are far from a model that can reliably predict the effects of mutations on stabilities. Mutations are often accompanied by conformational changes, making the predictions of their effects on stability difficult. Compactness as well as size and shape of the cavity forming residues in the

protein core play critical roles in defining the mutational effects, which generally lead to a lowered stability.⁷

Protein stability characterization is significant in determining changes in stability compared to the wildtype protein. Protein stability is mainly divided into two categories, thermodynamic stability and kinetic stability. While there is a range of methods to measure stability, only a few approaches will be discussed. Circular Dichroism spectroscopy is a valuable tool in measuring thermostability. Alpha-helix and beta-sheet give rise to characteristic CD spectra of varying magnitudes spectrum, allowing scientists to determine whether a protein is folded by its secondary structure. By applying heat at a constant rate, secondary structure and foldedness can be measured. Analysis of the data provides the transition midpoint T_M where 50% of the protein is in its native state and 50% is denatured. In general, the higher the T_M , the more stable the protein. Another method to assess protein stability or the effects of mutations on stability is by equilibrium unfolding using chemical denaturants such as urea⁸ or guanidine hydrochloride. The equilibrium constant, and consequently ΔG , can be determined. In addition, fluorescence of tryptophan can be used to monitor the unfolding of proteins.⁹ Differential scanning calorimetry (DSC) is another powerful analytical tool that can determine the T_M and measures the stability and unfolding of proteins.¹⁰

Proteins are only marginally stable at room temperature.¹¹ In a well-folded protein, entropy and enthalpy yield a very small free energy value.¹² The free energy difference between the native protein and the non-native states is on the order of -10 to $-15 \text{ kcal mol}^{-1}$, depending on the size of the protein.⁷ The marginal stability of native proteins seems to be a biological necessity for rapid disposal when required. Nonetheless, increased stability can provide innumerable industrial, agricultural, and therapeutic applications. For example, tumor suppressor

p53 has been implicated in 50% of human cancers.⁵³ However, the wild-type protein is difficult to study because it is only marginally stable. The engineering of a stable variant of p53 that retains wild-type structure and function has significantly enhanced the research and understanding of the protein.

The ability to stabilize proteins while retaining function has immeasurable advantages. The evolvability of a protein is related to its stability.¹³ Stable proteins are more tolerant of functionally beneficial but destabilizing mutations, aiding in the engineering of new functions. Increased understanding of the protein folding problem and the effects of mutations will advance the design and engineering of novel variants for diverse applications. Industrial and medical applications of proteins often require improved stability.¹⁴ Various experiments have been successful at dramatically stabilizing proteins without sacrificing their biological functions.^{15,16} While protein design and engineering have lead to successful stabilizations, there is no accurate, comprehensive method to predict the thermodynamics of mutations, even if the structure is known.^{7, 17}

Current Methods to Stabilize Proteins

Efforts to investigate protein stability began with a trial-and-error approach implementing random mutagenesis. This unsystematic method in which mutations are selected at random succeeds in only one out of 10^3 to 10^4 point mutations.^{18,19} Directed evolution, which employs the combination of random mutagenesis, followed by a high throughput screening or selection to obtain the desired proteins, has been used to improve thermal stability. While this process has the capacity to uncover stabilizing mutations, it strongly depends on highly sensitive screening assays and selection protocol, both of which may not be available or feasible.⁴³

Technological advances have improved protein engineering by allowing rational amino

acid modifications based on secondary structure predictions and model building. This rational approach can be classified into two main groups. The first approach compares the amino acid sequence of a more thermostable, homologous counterpart to that of the protein of interest. Hyperthermophilic and thermophilic proteins have higher T_M values, and the comparison of amino acid residues and structure may provide insight on factors that allow these proteins to thrive at high temperatures. For example, by analyzing the amino acid sequence of thermophilic homologous proteins, more stable variants of proteases and RNases have been constructed.^{20,21} Another successful stabilization using this rational design was accomplished by Van der Burg *et al.*, who increased the thermostability of thermolysin-like protease from *Bacillus stearothermophilus* by 21 °C.²²

The second rational design approach relies on the three dimensional structure to make structure-based predictions to stabilize the protein. Continuous research has increased our understanding of protein folding and of factors that affect protein stability, providing critical information to aid protein engineering using three-dimensional structures. Rational design using protein structure includes the introduction and modifications of disulfide bonds, salt bridges, secondary structure propensities, and helix capping. The introduction of disulfide bonds has been shown to stabilize proteins,²³ but accurate predictions of these changes are currently not comprehensible. Stabilization with secondary structure propensities has also encountered modest success. Numerous studies on model peptides have analyzed the consequences of placing different amino acids in varying secondary structures, including alpha helices and beta strands.^{24,25,26} While some experiments show successful stabilization, the modifications of secondary structure propensity at single amino acid substitutions are unlikely to have significant effects. For example, beta strand-forming propensities of amino acids at edge positions and

center positions can vary substantially, making protein engineering more difficult.²⁷ Another current modification to increase protein stability utilizes helix capping. Mutations at certain initiation or terminal signal sequence to the preferred amino acid sequence at specific locations may increase stability.²⁸ The introduction of salt bridges can also have stabilizing effect. Many salt bridge stabilizations have been noted in hyperthermophiles. For example, a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase, has a relatively large number of buried atoms and ion-pairs between acidic and basic amino acids that are linked to stability.²⁹ While rational engineering using protein structures have had moderate success in the past, this method has significant limitations.

The use of rational design to predict stabilizing mutations by analyzing tertiary structure has had varied results. Relative ranking of the importance of each method is debatable, and the chance of success is not easily predictable. Another crucial limitation is that these approaches require the protein structure. While a large number of diverse protein structures have been characterized, the vast majority of natural proteins structures are not available. With no tertiary structure, rational mutations cannot be predicted. Not only are larger databases needed, but better stabilization approaches are required. To confront the obstacles of rational design, computational methods have been devised. The technological advancement of computers and programming has paved the way for computational approaches in recent years.

The invention of the computer emerged as a powerful tool for protein stabilization. Dahiyat and Mayo developed a computational algorithm to predict side chain packing and the effect of steric constraints in the core residues. This algorithm was applied on streptococcal protein G β 1 domain to predict seven stabilizing mutations.³⁰ The unfolding temperature of G β 1 domain increased from 81°C to greater than 100°C. More recently, Kuhlman *et al.* used the

program Rosetta to predict the tertiary structure of proteins with modest success. However, computational modeling also has its limitations. It is difficult to calculate with high accuracy due to large margin of error. In marginally stable proteins, the forces that determine folding are large but balancing, giving rise to a small free energy difference between the folded and unfolded states.^{12, 32} The error for computational prediction often exceeds the free energy value of stability of the protein, making it unreliable. Recent research has provided a novel approach to understand how changes in amino acid sequence can lead to stable, functional proteins.

Consensus Design

The genomic era presents a positive prospect using consensus design that may provide valuable insight on protein stabilization. Protein structural and functional information can be deduced from protein sequences and provide great insight on the protein folding problem. Analysis and comparison of the amino acid sequences can show how variation in residues can lead to similar structures and activities. Currently, the most comprehensive use of statistical information for protein engineering is consensus design. Conserved positions were hypothesized to be informationally valuable because Nature samples other residues and selected against them.³⁴ These stabilizing mutations are predicted using a multiple sequence alignment of homologous proteins to obtain the most common amino acid at each position. Therefore, every position is optimized independently. The effects of the individual mutations are small, but large global stabilization can be achieved by additive contributions of many independent mutations. Using the consensus design approach, Steipe *et al.* predicted ten stabilizing mutations in immunoglobulin V κ domain of the anti-phosphorylcholine antibody McPC603 that would improve stability.³³ The nonconsensus amino acids were mutated to the most common amino acid at the respective positions. Six out of these 10 mutations were stabilizing. Replacement with

the consensus residue results in thermostabilization approximately half of the time. This value is an immense improvement compared to the less than 1% chance of improved fitness from random mutations.^{34, 35}

Consensus design has been applied on various proteins with success. Recently, full consensus design on ankyrin repeat domains has been characterized for structure and folding.^{36,37} Surface charge interactions of the consensus protein may explain its exceptional stability.^{38,39} Consensus mutation has also been predicted and tested in murine intrabodies⁴⁰ and humanized antibodies.⁴¹ One of the most notable successes using this approach was experiments performed using fungal phytases.⁴²⁻⁴⁴ The consensus protein was constructed based on thirteen fungal phytase sequence. While the sequence family is rather small and highly homogenous, the consensus enzyme was found to be stabilizing. Consensus phytase exhibited normal catalytic properties with an increase in unfolding temperature 15-22 °C higher than that of each of the parents. To further improve the protein stability using consensus design, additional factors can be included to create a more accurate and comprehensive algorithm.

The additive consequences of a collection of mutations contribute to the overall stabilizing effect in consensus design. One deficiency in the method is why does it only work half of the time? One explanation is that some sites in the proteins are coupled, so not all mutations are additive. In addition, most positions in protein families are not highly conserved. In consensus design, unconserved residues are of little informational value. It is unnecessary to create a full consensus with every single positions mutated to the conserved residue as some amino acids have little consequence to stability or may even have a destabilizing effect. Theoretically, there may be individual mutations that are sufficient to stabilize the entire protein. To investigate this issue, individual mutations of the consensus fungal phytases engineered were

performed separately to analyze the effect of stability for each amino acid changes. Results showed that the individual substitutions always affected thermostability by less than 3 °C.⁴³ Therefore, the observed increase in folding temperature of consensus fungal phytase is unlikely due to a small selection of dominant amino acid substitutions introduced by chance in consensus design, but rather due to the combined effect of multiple amino acid exchanges. However, precedence for single amino acid changes that have major stabilizing effects does exist. A single amino acid exchange was found to increase the melting temperature of *Leishmania* triosephosphate isomerase by 26 °C.⁴⁵ Previous studies that employed partial consensus design chose positions structurally or at random. A method that refines consensus protein engineering by deducing which positions should be mutated to the consensus could significantly advance protein stability.

Statistical Protein Design Using Relative Entropy and Mutual Information

A significant protein design element seems to be the proper placement of hydrophobic residues along the polypeptide chain to form a well-packed core. Information theory for statistical analysis can be used to refine consensus design. A statistical model was devised using relative entropy as a filter to improve consensus design to stabilize protein while retaining proper function. Relative entropy (RE) is a measure of conservation that can easily be calculated and is independent of the number of sequences. We performed this statistical approach using relative entropy on yeast triosephosphate isomerase (TIM) to increase the thermostability of proteins. The variation in each position in TIMs was analyzed by measuring the relative entropy between the amino acid distribution at each site and the distribution of amino acid usage in all proteins in yeast. It was hypothesized that large deviations from the reference state were significant in defining the protein and had a larger contribution to stability.

Previously, a variant of yeast TIM, Sc2TIM, was engineered to test this hypothesis. All positions in Sc2TIM with a relative entropy value of at least two were mutated to the consensus residue if it was not already present in yeast TIM. Of the 63 out of 241 positions with a relative entropy greater than two, only six residues in yeast TIM were not already the consensus amino acid. Sc2TIM was engineered to contain six consensus residues at positions with a relative entropy greater than two. Analysis of Sc2TIM showed that it was destabilized (4.2°C) compared to wild-type yeast TIM. Engineering additional variants containing individual mutations to the consensus residue at highly conserved positions will provide insight on why Sc2TIM was not stabilized as expected.

An essential component of the ubiquitous glycolytic pathway, TIM is a widespread and well studied protein found in numerous organisms. As part of the glycolysis pathway, TIM functions to interconvert dihydroxyacetone phosphate with glyceraldehydes-3-phosphate. Approximately 10% of all natural enzymes have the TIM-barrel architecture⁴⁸ essentially composed of eight alpha-helices and eight beta-strands, resulting in two concentric hydrophobic cores. Natural TIMs are homodimers, with an active site typically in the connecting alpha-beta loops.⁴⁹ In order for catalysis to occur, loop 6 must be in a closed conformation and must be open for product release.^{50,51} There are nearly 1,000 structures of TIM barrels in the PDB, and over 1,200 sequences of TIM in Pfam. It is exceptionally well studied^{46,47} and its prevalence makes TIM an excellent model protein for statistical design.

Chapter 2- Materials and Methods

Selection of Stabilizing Mutants

As previously stated, we hypothesized that the degree of amino acid conservation at a position is related to its contribution to stability. We hypothesize that more conserved positions have a larger affect on protein stability. The degree of conservation for each amino acid position was quantified using relative entropy (RE):

$$RE = \sum_x p_x \ln \frac{p_x}{f_x}$$

The relative entropy equation is the summation of the frequency of each amino acid at a particular position (p_x) multiplied by the natural log of that value over the frequency of the respective amino acid of the reference state (f_x). Essentially, relative entropy is the natural log of the probability of observing a particular distribution if you expect the reference distribution. Amino acid distribution in yeast is used as the reference state to test our hypothesis. Figure 1A shows the amino acid distribution of *Saccharomyces cerevisiae* and positions with varying relative entropies. A relative entropy value of greater than two was considered highly conserved. Figure 1B shows the relative entropy distribution in TIM.

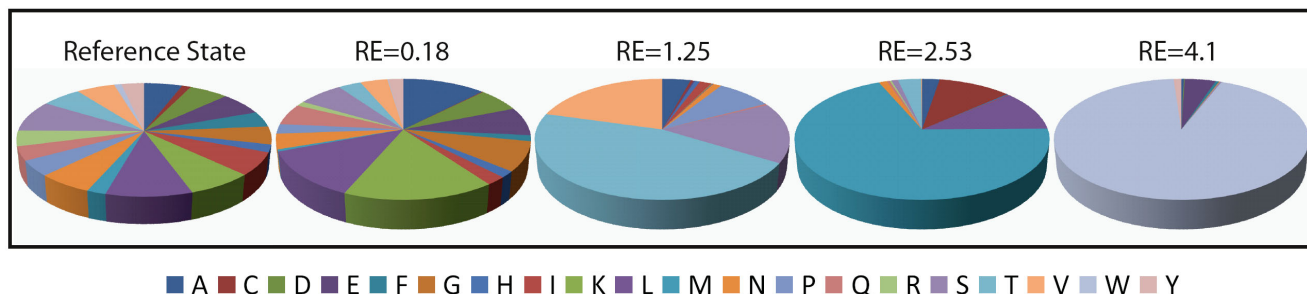


Figure 1A. Amino acid distribution of *S. cerevisiae* (reference state) and positions of varying relative entropies.

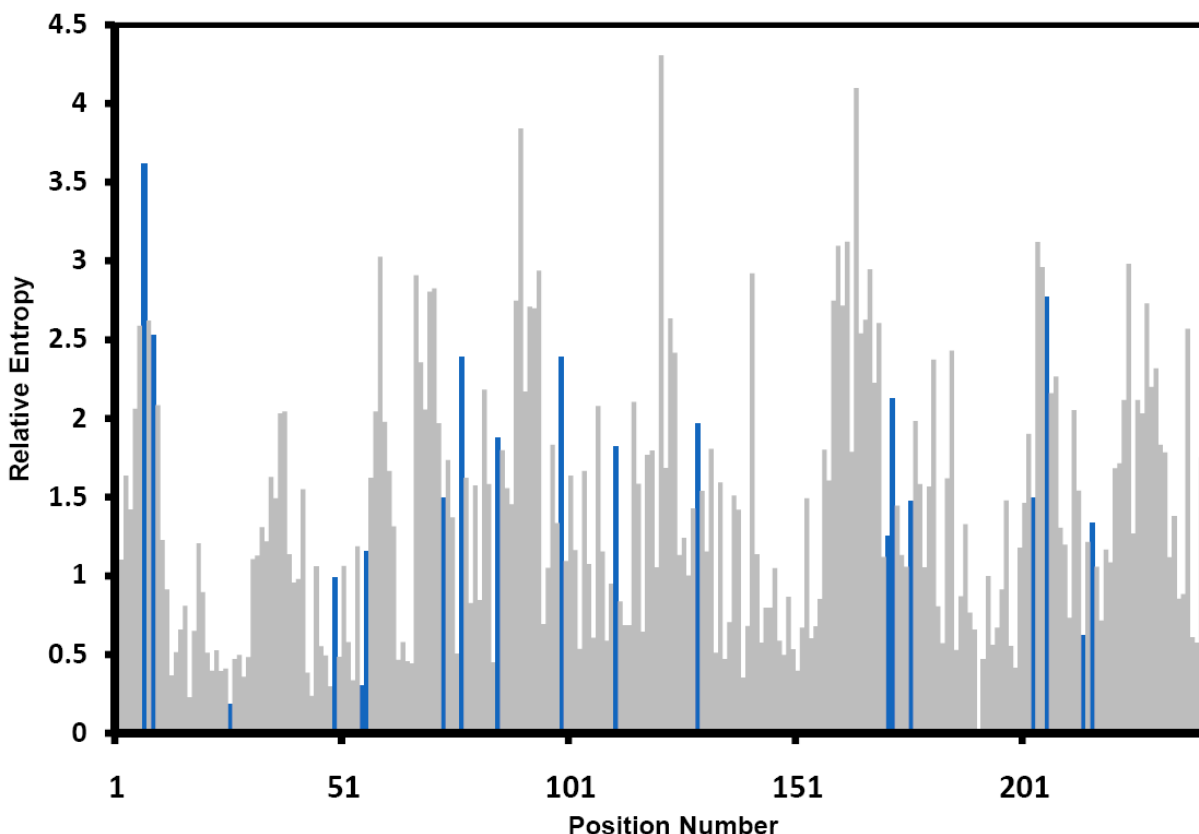


Figure 1B. Relative entropy distribution in TIM. Blue bars indicate positions of high, medium, and low relative entropies selected for analysis.

All six positions in yeast TIM with a relative entropy of two were mutated to the conserved residue simultaneously in TIM (Sc2TIM), as previously noted. However, biophysical characterization of Sc2TIM showed that it was destabilized compared to wild-type TIM. To investigate further as to why Sc2TIM was not stabilized as hypothesized, six variants with individual mutations at high RE values of two were engineered. In addition, three medium RE and three low RE variants were also engineered for comparison and analysis.

In addition to relative entropy, mutual information was also used to select stabilizing mutants. Mutual information is a measure of dependence between amino acids at varying positions. It is a method of measuring relatedness of protein sequences by analyzing the observed

frequency of amino acids at particular positions to that of an expected frequency at the respective positions. The use of mutual information facilitates the selection of stabilizing mutations by ensuring that positions with low correlations are selected to avoid disrupting significant amino acid interactions.

Genetic Engineering

Oligonucleotides were designed to contain the point mutation for site-directed mutagenesis. The oligonucleotides were purchased from Sigma-Genosys and resuspended in purified water to 100 μ M solutions. A 5' primer and 3' primer containing the point mutation was designed for each variant. Common 5' extension and 3' extension primers at the EcoNI and EcoRI sites, respectively, were used for all variants to form individual DNA fragments. Common 5' extension and 3' extension primers at the NcoI and BamHI sites, respectively, were designed for overlap polymerase chain reaction (PCR). The oligonucleotide sequences for each variant are shown in Table 1.

Primer	Sequence: 5' \rightarrow 3'
Medium RE: F229A	CAAGGCTGATGTCGATGGT GCG TTGGTCGGTGGTGCTTCTTTG CAAAGAAGCACCACCGACCAAC CGC ACCATCGACATCAGCCTTG
Medium RE: A175T	GCCATTGGTACCGGTTTG ACAG CTACTCCAGAAGATGCTC GAGCATCTTCTGGAGTAGCT GT CAAACCGGTACCAATGGC
Medium RE: T219E	CTAATGGTAGCAACGCCGTT GAG TTCAAGGACAAGGCTGATG CATCAGCCTTGTCCTTGA ACT CAACGGCGTTGCTACCATTAG
Low RE: T60A	GTTAAGAAGCCACAAGTC GCT GTTCGGTGCTCAAAACGCC GGCGTTTTGAGCACCGAC AGC GA CT TGTGGCTTCTTAAC
Low RE: Y49Q	CCAGCTACCTACTTAGAT CAG TCTGTCTCTTTGGTTAAG CTTAACCAAAGAGACAG ACTG ATCTAAGTAGGTAGCTGG
Low RE: S31K	GAAAGATTGAACACTGCTAA AA TCCCAGAAAATGTCGAAG CTTCGACATTTTCTGGGATTT TA GCAGTGTTCAATCTTTC

Table 1. Primers used for gene construction. Cloning for the individual high relative entropy mutants were previously cloned by Deepti Mathur and Brandon Sullivan.

The PCR reactions were performed using Phusion high-fidelity polymerase and 5X Phusion HF buffer. Typically, PCR conditions consisted of an initial denaturation at 95 °C for two minutes. A subsequent denaturation at 95 °C for 30 seconds was performed, followed by annealing at 60 °C for 45 seconds and elongation at 72 °C for 30 seconds. The denaturation, annealing, and elongation steps were performed for 25 cycles, followed by a final extension at 72 °C for 4 minutes. All PCR products were analyzed using agarose gel electrophoresis for size confirmation. The overview of the reassemble reaction performed by PCR is shown in Figure 2.

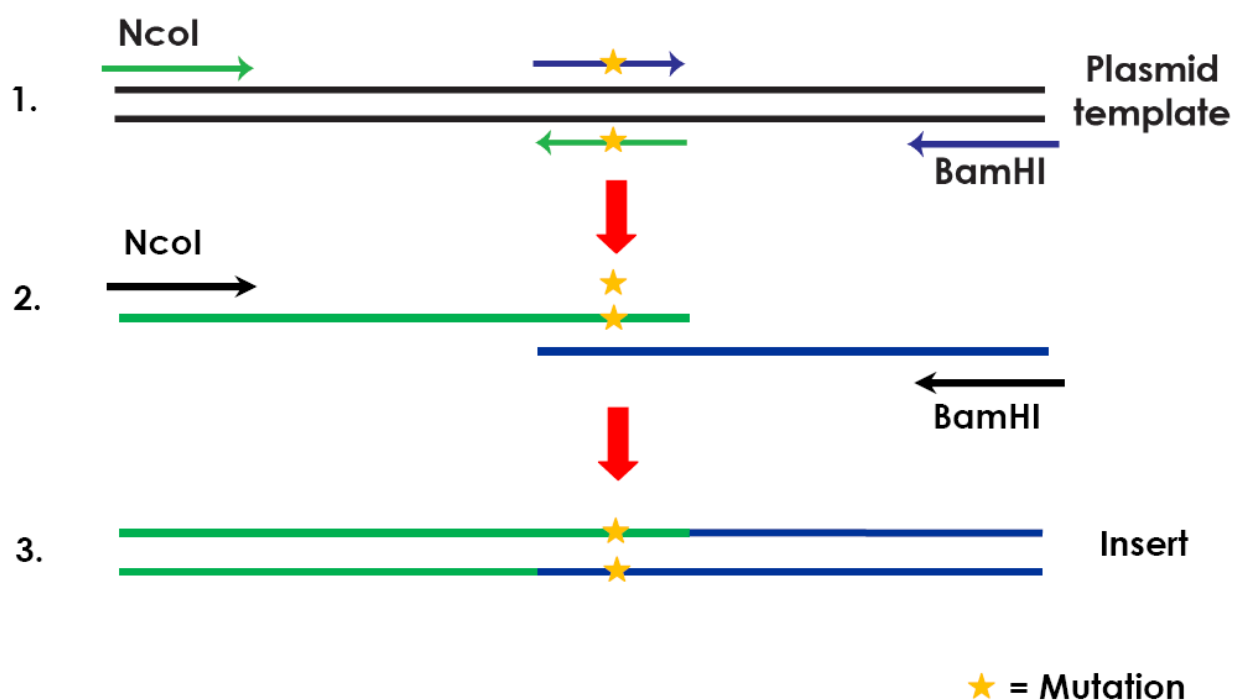


Figure 2. Gene construction using PCR using site-directed mutagenesis and overlap PCR

The fragments were digested with restriction enzyme DpnI to get rid of methylated template DNA, and followed by DpnI denaturation at 80°C and gel purification with a QG buffer cleanup as described by Qiagen to remove DNA fragments produced after restriction digest. The fragments were joined and amplified using overlap PCR to form the desired region of gene insert ranging from the NcoI site to the BamHI site. Restriction digest using NcoI and BamHI was

performed to cleave the ends of the DNA, and the inserts were purified using phenol-chloroform-isoamyl alcohol/chloroform-isoamyl alcohol followed by ethanol precipitation. The gene was ligated into the pHLIC vector, which was previously also digested with NcoI and BamHI, using T4 ligase at 16 °C overnight. The ligation with the pHLIC vector containing the TIM gene was then transformed into electrocompetent DH10B and plated on semisolid LB media containing 1X ampicillin and incubated at 37 °C overnight. Isolated colonies were grown in 5 ml of 2YT media and 5 µl of 1X ampicillin overnight in a 37 °C shaker. The DNA was extracted using a Qiagen Miniprep kit and analyzed using double restriction enzyme digests to ensure correct banding pattern via gel electrophoresis. The gene was then sequence confirmed using a T7 sequencing primer.

Protein Purification

Once sequence confirmed, the gene was transformed in BL21(DE3) and plated on semisolid LB media containing ampicillin. Isolated colonies were picked and grown in 25 ml of 2YT media and 25 µl of 1X ampicillin overnight in the 37 °C shaker. The culture was then inoculated into 1 L of 2YT media and grown to an OD₆₀₀ of approximately 0.75. After induction at a final concentration of 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), the cells were placed in the shaker for an additional 3-4 hours at 37 °C or 6-8 hours at 30°C. The cells were then centrifuged in GS3 tubes at 6,000 rpm for 10 minutes and decanted. Pellets were stored overnight at -80 °C. Pellets were resuspended in 25 ml of lysis buffer (50 mM Tris·HCl pH 8, 300 mM NaCl, 10 mM imidazole, 8 mM β-mercaptoethanol, 1 mM TCEP). To extract the protein, 5 mM MgCl₂, 0.5 mM CaCl₂, 5 mg/mL DNase I, 5 mg/mL RNase A, 0.1% Triton X-100, and 0.3-1 mg/mL HEW lysozyme were added to the solution and allowed to incubate on ice for 30 min. The solution was then sonicated three times at half-power for 30 seconds on ice with 2

minutes between each pulse. The solution was transferred into SS34 tubes and centrifuged at 20,000 rpm for 1 hour at 4 °C. To the supernatant was added 1.5 mL Ni-NTA agarose which binds the N-terminal His₆-site on the proteins and allowed to mix at 4°C for 1 hour. The slurry was poured into a column, washed with 12 mL of wash buffer (50 mM Tris·HCl pH 8, 300 mM NaCl, 20 mM imidazole, 8 mM β-mercaptoethanol, 1 mM TCEP), and the protein was eluted with 3 mL of elution buffer (50 mM Tris·HCl pH 8, 300 mM NaCl, 250 mM imidazole, 8 mM β-mercaptoethanol, 1 mM TCEP). An additional 1 mL of lysis buffer was added to the 3 mL pooled elution and dithiothreitol (DTT) was added to 5 mM. TEV protease⁵² was added to the elution to cleave the N-terminal TEV site and allowed to incubate at room temperature overnight. A second aliquot of TEV protease and DTT was added the following day and incubated at 30 °C for 1-3 hours. The solution was desalted using PD10 columns and exchanged into storage buffer (50 mM sodium phosphate, 300 mM NaCl, pH 8.0, 5 mM BME and 1 mM TCEP) then mixed with 1.5 mL Ni-NTA and allowed to mix at 4°C for 1 hour. The solution was then loaded into a small, pre-fritted column, and the flow-through was collected to obtain the purified TIM protein.

Protein size confirmation was obtained by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) using a 12.5% acrylamide gel. Protein concentration was determined using absorbance at 280 nm ($\epsilon=A/lc$) for aromatic amino acids in protein solutions. Centriplus YM-10 tubes were used for necessary protein concentrations.

Circular Dichroism Spectroscopy

Secondary alpha helical structures were quantified at 222 nm via AVIV Circular Dichroism spectrophotometer. TIM variants were analyzed in storage buffer at 14 μM protein concentrations. The wavelength scan parameters were set between 200.00 to 250.00 nm and sampled every 1 nm with an averaging time of 5.0 seconds for three scans. Thermal

denaturations were performed at temperatures ranging from 25 °C to 95 °C. Ellipticity was monitored at 222 nm and data was collected every 1 °C with 30 seconds equilibrations and an averaging time of 10 seconds.

Chemical denaturations were performed using various concentration of urea. Urea denaturations were also performed in storage buffer at 14 μ M protein. Refractometry was used to determine urea concentrations. Ellipticity was monitored every 5 minutes for 48 hours at 222 nm after the reaction has been equilibrating for 24-48 hours at room temperature.

Additional Relative Entropy Variants and Compensatory Mutations

To further investigate the hypothesis that mutations to the consensus residue at high relative entropy positions have the greatest stabilizing effect, additional mutants were and compensatory mutants were engineered and characterized using the same methods above. The oligonucleotide sequences for each variant are shown in Table 2.

Primer	Sequence: 5'→3'
A61C	CACTGTCGGTGCTCAAAACT G CTACcTGAAGGCTTCTGGTG CACCAGAAGCCTTCa G TAG C AGTTTTGAGCACCGACAGTG
I104V	CACGAAGATGACAAGTTt G TTGCTGACAAGACCAAGTTC GAACTTGGTCTTGT C AGCA A CaAACTTGT C ATCTTCGTG
I78L	GAAAACTCCGTTGACCAACT C AAGGATGTcGGTGCTAAG CTTAGCAC C gACATCCTT G AGTTGGTCAACGGAGTTTT C
V116L	GTTCGCTTTAGGTCAAGGT C TCGGTGT C ATCTTGTGTATC GATACACAAGATGACAC C GAGACCTTGACCTAAAGCGA A C
D176Q	GTTTGGCTGCTACTCCAGA A CAAGCTCAAGATATTCACGCTTC GAAGCGTGAATATCTTGAGCTT G TTCTGGAGTAGCAGCCAA A C
I180V	CCAGAAGATGCTCAAGAT G TT C ACGCTTCCATCAGAAAG CTTTCTGATGGAAGCGTGA A CATCTTGAGCATCTTCTGG
N208K	CTTATACGGTGGTTCaGCT A AAGGTAGCAACGCCGTTACC GGTAACGGCGTTGCTACCTT T AGCtGAACCACCGTATAAG
V221I	CTTCAAGGACAAGGCTGATAT C GATGGTTTCTTGGTCCGGTG CACCGACCAAGAAACCAT C GATATCAGCCTTGTCTTGAAG
I20A compensatory	CTTTAAATTAAACGGTTCCAAACAATCa G CGAAGGAAATcGTTGAAAGATTG CAATCTTTCAACgATTTCCT T CGCtGATTGTTTGGAAACCGTTTAATTTAAAG
G122T compensatory	GCTTTAGGTCAAGGTGT C ACAGTCATCTTaTGTATCGGTG CACCGATACaAAGATGACT G TGACACCTTGACCTAAAGC

Table 2. Primers used for additional RE and compensatory variants.

Chapter 3- Results and Discussion

Engineering of Variants

The central aim is to test the hypothesis that consensus mutations at more conserved positions will have a greater stabilizing effect. Relative entropy as a statistical analysis can improve techniques for accurately increasing the stability of proteins while obtaining greater depth on the sequence-structure relationship. The goal is create more stabilized proteins that maintains wild-type function and properties.

Figure 3 shows the pHLIC vector map with the *Saccharomyces cerevisiae* TIM (YPI) gene and location of the hexahistidine tag, TEV cleavage site. In addition, pHLIC also contains an ampicillin resistance gene for the selection of cells that have taken in the plasmid.

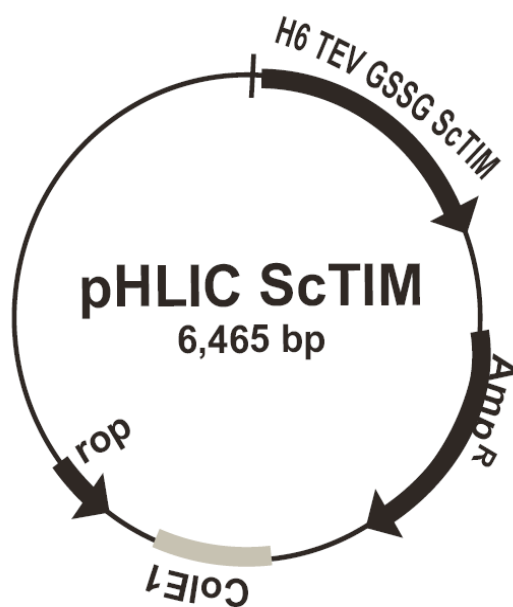


Figure 3. pHLIC vector with TIM gene

All 3 medium relative entropy and 3 low relative entropy variants were successfully cloned using site-directed mutagenesis and overlap PCR. Cloning and biophysical characterization for the six high relative entropy mutants were previously performed by Deepti Mathur. Negative controls

were performed using PCR on the individual fragments to ensure low background. Faint bands or no bands of the negative control PCR indicate miniscule template concentration, which ensures that most of the overlap PCR products are not wild-type. Analytical digests on isolated colonies show a large vector band and an insert band around 800 base pairs, as expected (figure 4). All thirteen variants were sequenced confirmed.

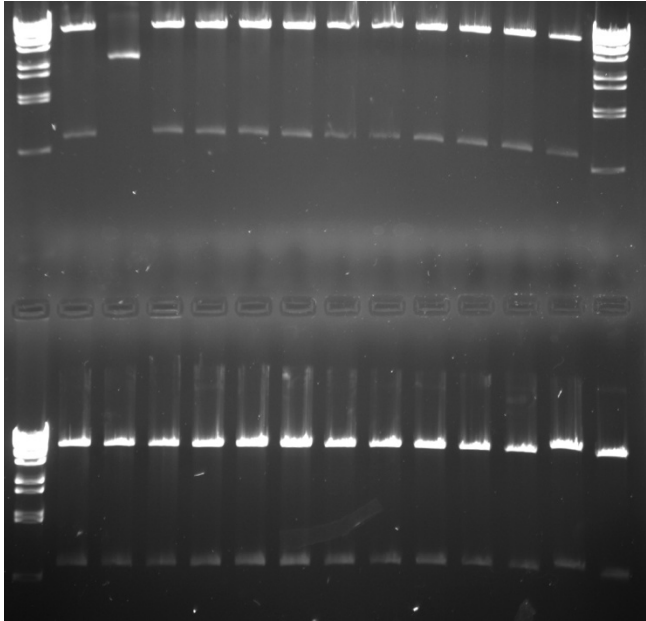


Figure 4. Analytical digests of RE variants using NcoI and BamHI. A large vector band around 5500bp and small insert band around 800bp were as expected. Ladder used is λ DNA BstEII.

Protein Purification

Twelve of the thirteen variants had expression and were successfully purified using nickel affinity chromatography. I was unable to purify medium RE mutant F229A due to lack of expression in BL21(DE3). Various steps of protein purification are shown in figure 5. Wild-type yeast TIM (YPI) was electrophoresed in parallel with all variants as controls.



Figure 5. Steps of YPI purification of a RE variant. P=pellet, S=supernatant, F=flow through, W=wash, E=elution. Purified monomeric YPI protein is in the elution. Ladder is See Blue.

Biophysical Characterization

CD wavelength scans at 222 nm were performed to observe helical content and standardize protein concentration for thermal denaturation. CD thermal melts show that all three low relative entropy mutants were destabilized compared to wild-type yeast TIM. Of the three medium relative entropy mutants, one was destabilizing, and one was stabilizing. Three of the high relative entropy mutants were stabilizing, two were destabilized, and one was the same as wild-type. Figure 6 shows the CD thermal melts of the high, medium, and low relative entropy mutants compared to wild-type.

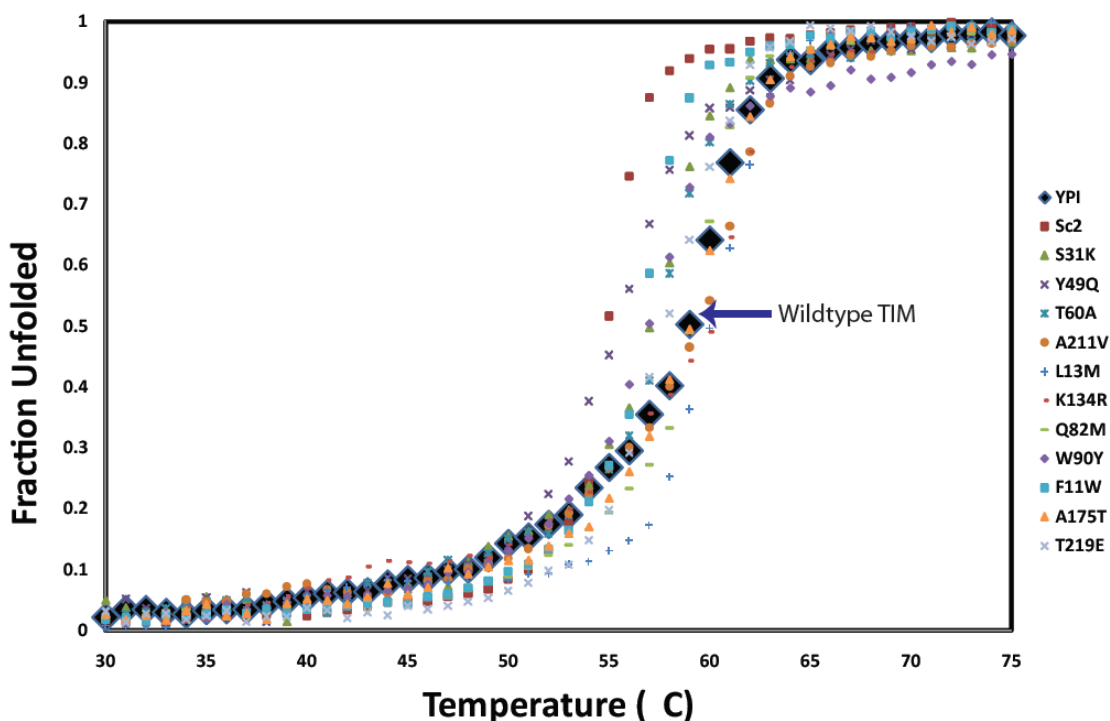


Figure 6. CD thermal melt of high, medium, and low RE variants. Stabilized variants were at high relative entropy positions.

High relative entropy mutants provided the most stabilizing effect. Analysis of the other properties of stabilizing mutants including solvent exposure, secondary structures, BLOSUM score, and statistical correlation were also evaluated. The stabilizing mutations tend to be near the surface of the proteins. Many of the variants located within the protein core were destabilized, possible due to interruption of proper folding. Results suggest that high relative entropy, high solvent exposure, and positive BLOSUM mutants provide the most stability.

From these results, we hypothesize that mutations in high relative entropy positions on protein surfaces will increase protein thermostability. In addition, we also predict that positive BLOSUM values and low mutual information will increase stability. We hypothesize that mutations a positions with a relative entropy greater than 1.25, a positive BLOSUM value, and a

mutual information (MI) value of less than 0.5 will have the greatest stabilizing effect. A low MI value suggests that the amino acid residue has low correlation to other residues in the protein. This avoids mutating positions that may have significant interactions with other residues, and can have destabilizing effects. Eight additional mutants that included these variables were cloned and purified. However, it is critical to note that all high relative entropy positions in YPI either is already the conserved amino acid, or has already been mutated in the previous experiment. The new set of variants follow that above criteria, except they are within the medium relative entropy range. Figure 7 and 8 shows detailed variables of these second set of relative entropy variants, and variables and results for all other mutants. CD thermal melts of these variants show that most of the variants either were neutral mutations, or slightly destabilized compared to wild-type. These results are still consistent with the hypothesis that high relative entropy mutations to the consensus residue are stabilizing.

Mutants	RE	T _m	Solvent %	2° structure	BLOSUM	MI
YPI	N/A	59.3	N/A	N/A	N/A	N/A
Sc2	N/A	55.1	N/A	N/A	N/A	N/A
S31K	0.18	57.3	43	Loop	0	0.47
Y49Q	0.24	55.7	18	Helix	-1	0.73
T60A	0.3	57.9	14	Strand	0	0.54
T219E	1.22	58.0	25	Helix	-1	0.34
A175T	1.25	59.4	39	Loop	0	0.39
W90Y	2.13	56.8	8	Strand	2	0.72
K134R	2.24	60.3	22	Helix	2	0.38
A211V	2.27	59.6	3	Loop	0	0.25
Q82M	2.41	59.3	37	Helix	0	0.45
L13M	2.44	60.5	60	Loops	2	0.48
F11W	3.50	56.9	1	Strand	1	0.29
A61C	1.63	58.6	0.9	Loop	0	0.42
I78L	1.62	55.0	1	Helix	2	0.25
I104V	1.67	59.0	0.5	Helix	3	0.29
V116L	1.59	55.1	1.6	Loop	1	0.18
D176Q	1.47	57.4	17.3	Helix	0	0.51
I180V	1.57	59.4	7.2	Helix	3	0.37
N208K	1.31	58.2	18.3	Helix	0	0.43
V221I	1.71	57.8	1.3	Loop	3	0.32

Figure 7. Parameters of selection for second set of relative entropy mutants in last eight row. RE: >1.25, BLOSUM ≥ 0 , MI ≤ 0.5 . Results and variables for wild-type yeast TIM, Sc2TIM, and previous high, medium, and low RE mutants also listed for comparison.

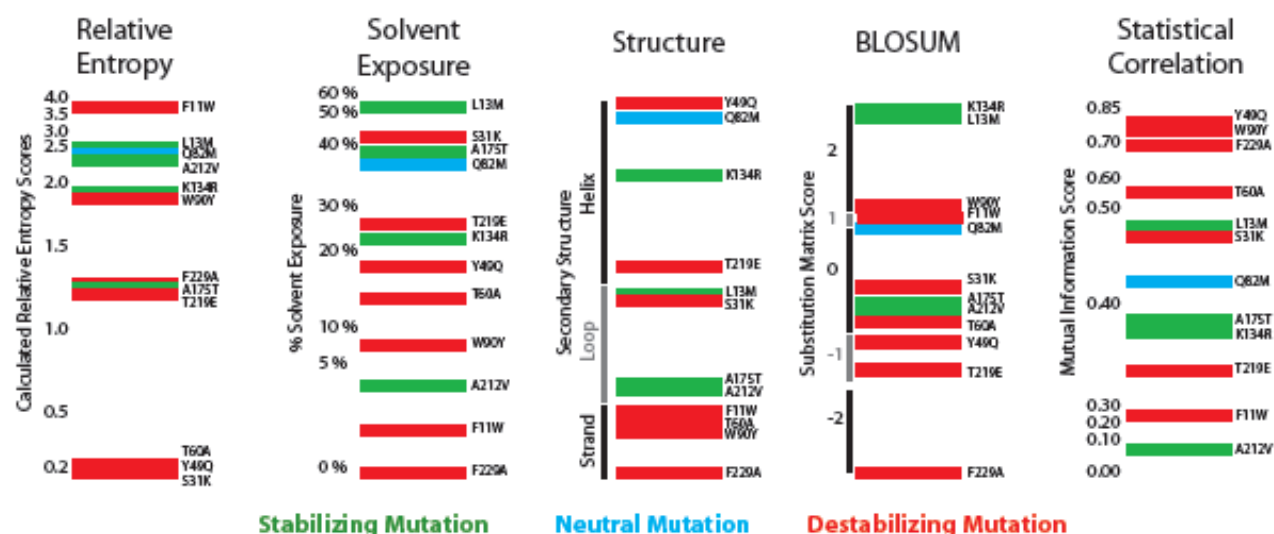


Figure 8. Visual representation of similar data as figure 6. Bars indicate different variants and colors indicate stability compared to wild-type. RE, solvent exposure, structure, BLOSUM values, and MI values are plotted on vertical axis of each block.

Compensatory Mutations

Two previously cloned high relative entropy mutants were shown to be destabilized compared to YPI by CD thermal melt. F11W is a high relative entropy mutant with low solvent exposure and positive BLOSUM value. W90Y is also a high relative entropy mutant with low solvent exposure, and a positive BLOSUM value. More importantly, it has high statistical correlations ($MI=0.72$) with other residues in the protein. Mutating this position may disrupt the normal interactions between correlated residues, affecting the folding and structure of the protein. The mutation from a phenylalanine to a tryptophan may interrupt normal residue interactions, especially at correlated positions. To investigate why these high relative entropy positions were not stabilized as expected, compensatory mutations will be introduced in each variant. Mutation I20A has been selected to compensate for the F11W mutations, and G122T has been selected to compensate for the W90Y mutation. The compensatory I20A and G122T were cloned, and purified. However, W90Y failed to produce sufficient amounts of protein for characterization. The CD thermal melt show that I20A was able to stabilize the F11W mutation significantly. Figure 9 shows the CD thermal melts results for the destabilized high relative entropy F11W variant, and the significantly stabilized F11W-I20A compensatory mutant compared to wild type. This provides logical explanation as to why the high relative entropy mutant was not stabilized as expected. It shows that analyzing correlation is critical in consensus design to stabilize proteins. To ensure that the I20A is not a stabilizing mutation in itself, a variant with only I20A will be analyzed.

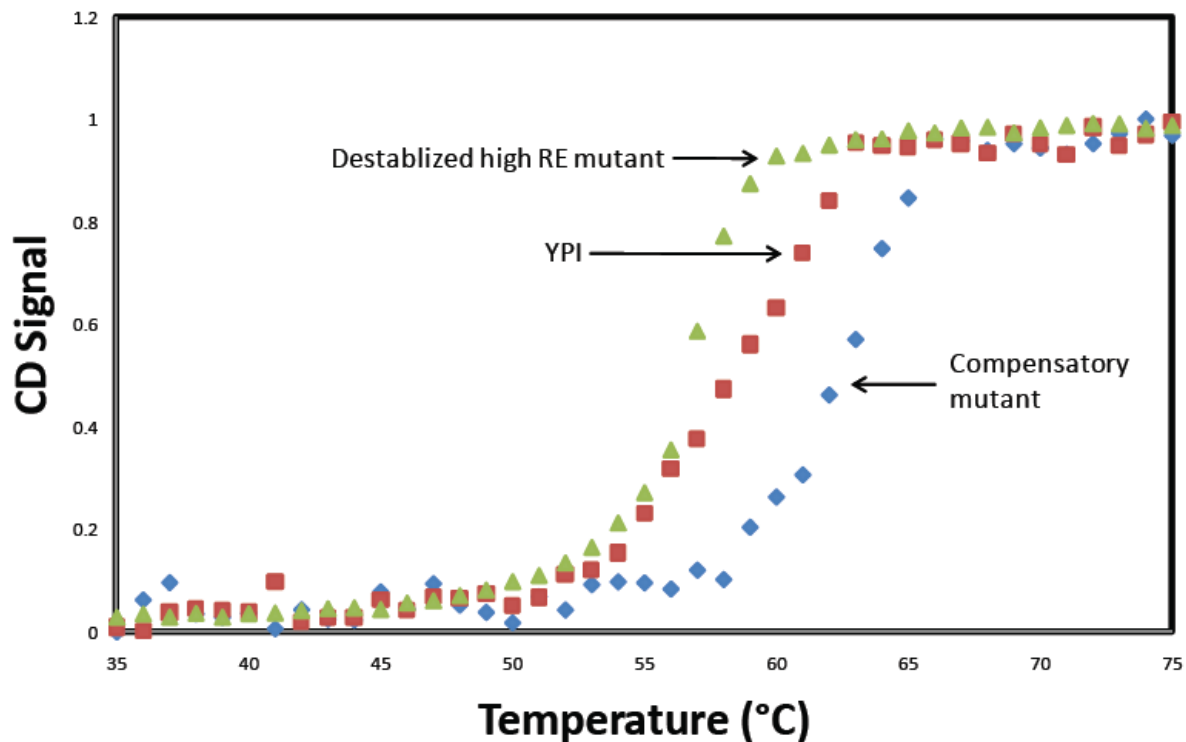


Figure 9. CD thermal melts data of destabilized F11W and F11W-I20A double mutation variant compared to wild-type. I20A mutation was able to compensate for initial destabilized F11W mutation.

Urea Unfolding

In addition to thermostability, kinetic stability of the variants will be analyzed. Although we have found that high relative entropy positions are stabilizing, we do not fully understand the mechanism for this stability. One possible explanation is that the proteins are unfolding at slower rates. Chemical denaturation using urea is performed to obtain the unfolding rates of the variants compared to wild-type TIM. Figure 10 shows urea folding for wild-type TIM using varying concentrations of denaturant over a period of 48 hours. Using this data, we can extrapolate back to 0 M urea, and obtain the protein's rate of unfolding under no denaturant, as shown in Figure 11.

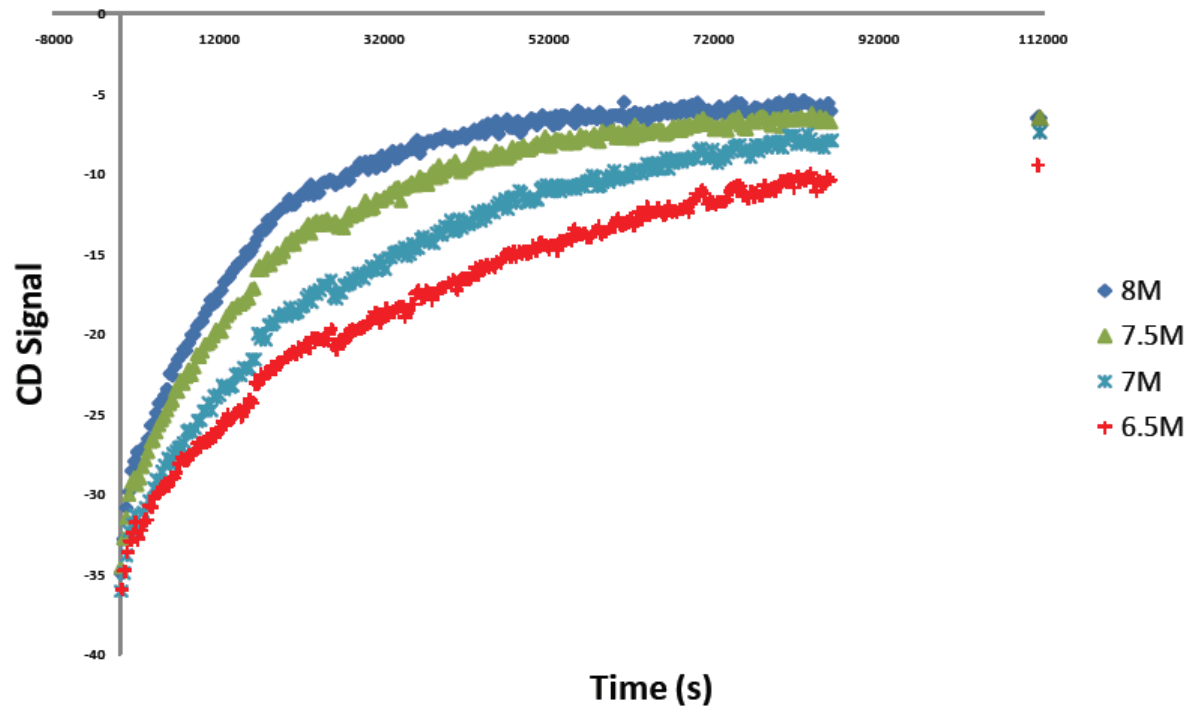


Figure 10. Urea unfolding of YPI over 48 hours using varying concentrations of denaturants.

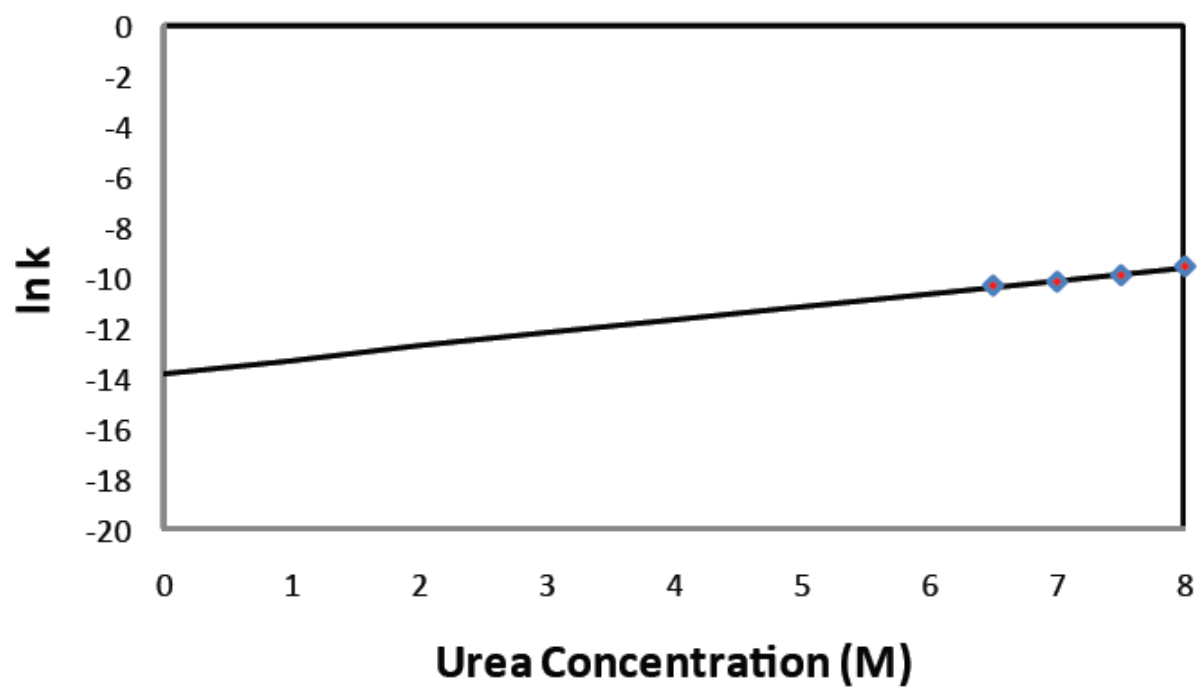


Figure 11. Using data from urea unfolding, the natural rate of protein unfolding can be obtained by extrapolated.

Remarks

Biocharacterization of the high, medium, and low relative entropy mutants supports our hypothesis that changing amino acid residues to the consensus at high relative entropy positions provide stabilizing effects. However, mutations at high relative entropy residues may not be stabilizing at correlated positions due to disruptions in residue interactions, as shown by the compensatory mutations. The use of mutual information in addition to relative entropy can aid in the selection of stabilizing mutations. This method provides a universal approach that can be applied to numerous proteins. One limitation is the lack of available sequences for a specific protein, but in the genomic era, this can easily be overcome.

To confirm that correlation was the destabilizing factors in high relative entropy variants that were not stabilized as expected, the second mutation introduced in the compensatory mutants will be characterized to ensure that the mutation was not stabilizing in itself. Urea unfolding may provide insight on the stabilizing mechanism, and will be performed on all variants. Future directions include analyzing the kinetics of each variants to wild-type and applying this method to another model protein.

References

1. Anson, M.L. & Mirsky, A.E. (1925) On some general properties of proteins. *J. Gen. Physiol*, **9**, 169-179.
2. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223-230.
3. Levinthal, C. (1968).Are there pathways for protein folding? *Extrait du Journal de Chimie Physique*, **65**, 44-45.
4. Dill, K.A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nat Struct Biol*, **4**(1), 10-9.
5. Dill, K.A, et al. (2007) The protein folding problem: when will it be solved?, *Structural Biology*, **17**, 342-346.
6. Drexler, K.E. (1981) Molecular engineering: an approach to the development of general capabilities for molecular manipulation." *Proc. Natl. Acad. Sci. USA*, **78**, 5275-5278.
7. Richards, F.M. (1997) Protein stability: still and unsolved problem. *Cell Mol Life Sci*, **53**, 790-802.
8. Greene, R.F. Jr. & Pace, C.N. (1974) Urea and guanidine hydrochloride denaturation of ribonuclease, lysozyme, alpha-chymotrypsin, and beta-lactoglobulin. *J Biol Chem*, **249**(17), 5388-5393.
9. Monsellier, E. & Bedouelle, H. (2005). Quantitative measurement of protein stability from unfolding equilibria monitored with the fluorescence maximum wavelength. *Protein Engineering, Design & Selection*, **18**(19), 445-456.
10. Lepock, J.R. (2005) Measurement of protein stability and protein denaturation in cells using differential scanning calorimetry. *Methods*, **35**(2), 117-125.
11. Pace, C.N. (1975) The stability of globular proteins. *CRC Crit Rev Biochem*, **3**(1), 1-43.
12. Dill, K.A., Dominant forces in protein folding. *Biochemistry* 1990, **29**(31), 7133-55.
13. Bloom, J.D., Labthavikul, S.T., Otey, C.R. & Arnold, F.H. (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci U S A*, **103**, 5869-5874.
14. Graddis, T.J., Remmele Jr., R.L. & McGrew, J.T. (2002) Designing proteins that work using recombinant technologies. *Current Pharamceutical Biotechnology*, **3**, 285-297.

15. Giver, L., Gershenson, A., Freskgard, P. O., & Arnold, F. H. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 12809–12813.
16. Godoy-Ruiz, R., Perez-Jimenez, R., Ibarra-Molero, B. & Sanchez-Ruiz, J. M. (2004) Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations. *J. Mol. Biol.* **336**, 313–318.
17. Cordes, M.H., Davidson, A.R. & Sauer, R.T. (1996) Sequence space, folding and protein design. *Curr Opin Struct Biol*, **6**, 3-10.
18. Risse B, Stempfer, G., Rudolph, R., Schumacher, G. & Jaenicke, R. (1992) Characterization of the stabilizing effect of point mutations of pyruvate oxidase from *Lactobacillus plantarum*: protection of the native state by modulating coenzyme binding and subunit interaction. *Protein Sci.*, **1**(12), 1710-1718.
19. Arase, A., Yomo, T., Urabe, I., Hata, Y., Katsube, Y. & Okada, H. (1993). Stabilization of xylanase by random mutagenesis. *FEBS Letters*, **316**, 123-127.
20. Imanaka, T., Shibasaki, M. & Takagi, M. (1986) A new way of enhancing the thermostability of proteases. *Nature*, **324**, 695-697.
21. Serrano, L., Day, A.G. & Fersht, A. R. (1993). Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J. Mol. Bio*, **233**, 305-312.
22. Van den Burg, B., Vriend, G., Veltman, O.R., Venema, G. & Eijsink, V.G.H. (1998). Engineering an enzyme to resist boiling. *Proc. Natl Acad. Sci. USA*, **95**, 2056-2060.
23. Mansfeld, J., Vriend, G., Dijkstra, B.W., Veltman, O.R., Van den Burg, B., Venema, G., Ulbrich-Hofmann, R. & Eijsink, V. G. (1997) Extreme stabilization of a thermolysin-like protease by an engineered disulfide bond. *J. Biol. Chem.*, **272**, 11152-6.
24. O'Neil, K. & DeGrado, W.F.(1990). A thermodynamic scale for the helixforming tendencies of the commonly occurring amino acids. *Science*, **250**, 246-250.
25. Horovitz, A., Matthews, J.M. & Fersht, A.R. (1992) Alpha-helix stability in proteins. II. Factors that influence stability at an internal position. *J Mo/BioI*, **227**, 560-568.
26. Kim, C.A. & Berg, J.M. (1993) Thermodynamic beta-sheet propensities measured using a zinc finger host peptide. *Nature*, **362**, 267-270.
27. Minor, D.L. & Kim, P.S. (1994) Measurement of the beta-sheet-forming propensities of amino acids. *Nature*, **367**, 660-663.
28. Richardson, J.S. & Richardson, D.C. (1988) Amino acid preferences for specific locations at the ends of co-helices. *Science*, **240**, 1648-1652.

29. Chan, M.K., Mukund, S., Kletzin, A., Adams, M.W. & Rees, D.C. (1995) Structure of a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase. *Science*, **267**(5203), 1463-9.
30. Dahiyat, B.I. & Mayo, S L. (1997) Probing the role of packing specificity in protein design
Proc. Natl. Acad. Sci. USA, **94**, 10172-7.
31. Liu, Y. & Kuhlman, B. (2006) RosettaDesign server for protein design. *Nucleic Acids Res* **34**, 235-238.
32. Rose, G.D. & Wolfenden, R. (1993) Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annu. Rev. Biophys. Biomol. Struct.*, **22**, 381-415.
33. Steipe, B., Schiller, B., Pluckthun, A. & Steinbacher, S. (1994) Sequence statistics reliably predict stabilizing mutations in a protein domain. *J Mol Biol*, **240**, 188-192.
34. Christians, F.C., Scapozza, L., Cramer, A., Folkers, G. & Stemmer W.P.C. (1999). Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling. *Nature Biotechnol.* **17**, 259.
35. Hermes, J.D., Blacklow, S.C. & Knowles, J.R. (1990) Searching sequence space by definably random mutagenesis: improving the catalytic potency of an enzyme. *Proc Natl Acad Sci U S A*, **87**, 696-700.
36. Kohl, A. et al. (2003) Designed to be stable: crystal structure of a consensus ankyrin repeat protein. *Proc Natl Acad Sci U S A*, **100**, 1700-1705.
37. Mosavi, L.K., Minor, D.L., Jr. & Peng, Z.Y. (2002) Consensus-derived structural determinants of the ankyrin repeat motif. *Proc Natl Acad Sci U S A*, **99**, 16029-16034.
38. Merz, T. et al. (2008) Stabilizing ionic interactions in a full-consensus ankyrin repeat protein. *J Mol Biol*, **376**, 232-240.
39. Wetzel, S.K., Settanni, G., Kenig, M., Binz, H.K. & Pluckthun, A. (2008) Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *J Mol Biol*, **376**, 241-257.
40. Ohage, E. & Steipe, (1999) B. Intrabody construction and expression. I. The critical role of VL domain stability. *J Mol Biol*, **291**, 1119-1128.
41. Knappik, A. et al. (2000) Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. *J Mol Biol*, **296**, 57-86.

42. Lehmann, M. et al. (2000) From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. *Protein Eng*, **13**, 49-57.
43. Lehmann, M., Pasamontes, L., Lassen, S.F. & Wyss, M. (2000) The consensus concept for thermostability engineering of proteins. *Biochim Biophys Acta*, **1543**, 408-415.
44. Lehmann, M. et al. (2002) The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng*, **15**, 403-411.
45. Williams, J.C., Zeelen, J.P., Neubauer, G., Vriend, G., Backmann, J., Michels, P.A.M., Lambeir, A.M. and Wieringa, R.K. (1999) Structural and mutagenesis studies of leishmania triosephosphate isomerase: a point mutation can convert a mesophilic enzyme into a superstable enzyme without losing catalytic power. *Protein Eng*, **12**, 243-250.
46. Alber, T. et al. (1980) On the three-dimensional structure and catalytic mechanism of triose phosphate isomerase. *Philos Trans R Soc Lond B Biol Sci*, **293**, 159-171.
47. Nickbarg, E.B. & Knowles, J.R. (1988) Triosephosphate isomerase: energetics of the reaction catalyzed by the yeast enzyme expressed in Escherichia coli. *Biochemistry*, **27**, 5939-5947.
48. Nagano, N., Orengo, C.A. & Thornton, J.M. (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol*, **321**, 741-765.
49. Lolis, E. et al. (1990) Structure of yeast triosephosphate isomerase at 1.9-A resolution. *Biochemistry*, **29**, 6609-6618.
50. Desamero, R., Rozovsky, S., Zhadin, N., McDermott, A. & Callender, R. (2003) Active site loop motion in triosephosphate isomerase: T-jump relaxation spectroscopy of thermal activation. *Biochemistry*, **42**, 2941-2951.
51. Kempf, J.G., Jung, J.Y., Ragain, C., Sampson, N.S. & Loria, J.P. (2007) Dynamic requirements for a functional protein hinge. *J Mol Biol*, **368**, 131-149.
52. Kapust, R. B. & Waugh, D. S. (2000) Controlled intracellular processing of fusion proteins by TEV protease. *Protein Expr Purif*, **19**(2), 312-318.
53. Joerger, A.C. & Fersht, A.R. (2007) Structural biology of the tumor suppressor p53 and cancer-associated mutants. *Adv Cancer Res*, **97**, 1-23.