THE COMPLEX ROLE OF SEQUENCE AND STRUCTURE IN THE STABILITY AND FUNCTION OF TIM BARREL PROTEINS

A Dissertation Presented

By

YVONNE HOI YAN CHAN

Submitted to the Faculty of the
University of Massachusetts Graduate School of Biomedical Sciences, Worcester
In partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

November 3, 2017

Biochemistry and Molecular Pharmacology

THE COMPLEX ROLE OF SEQUENCE AND STRUCTURE IN THE STABILITY AND FUNCTION OF TIM BARREL PROTEINS

A Dissertation Presented

By

YVONNE HOI YAN CHAN

This work was undertaken in the Graduate School of Biomedical Sciences
Biochemistry and Molecular Pharmacology

Under the mentorship of

C. Robert Matthews, Ph.D., Co-Thesis Advisor

Konstantin B. Zeldovich, Ph.D., Co-Thesis Advisor

Brian A. Kelch, Ph.D., Member of Committee

David G. Lambright, Ph.D., Member of Committee

William E. Royer, Ph.D., Member of Committee

Eugene I. Shakhnovich, Ph.D., External Member of Committee

Daniel N. Bolon, Ph.D., Chair of Committee

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical Sciences

November 3, 2017

Dedication

To my parents,
for your unconditional support.

To my sisters,
for letting me shop at your houses.

To my brother-in-laws,
for reading all my applications.

To my boyfriend,
for being there when I need you the most.

To my friends,
for letting me be myself.

Acknowledgments

This thesis is the product of the guidance from my two co-advisors: Dr. C. Robert Matthews and Dr. Konstantin B. Zeldovich. They trained me with their expertise; they inspired me with their passion for science. Dr. Daniel N. Bolon served as my unofficial third advisor. He welcomed me into his lab and helped me bring my experiments to life. My committee members, Dr. Brian A. Kelch, Dr. David G. Lambright, and Dr. William E. Royer, always asked thought-provoking questions. They pushed me to think critically and refined my research in countless ways. My external committee member, Dr. Eugene I. Shakhnovich, set the bar for seamlessly integrating multiple fields of science to which I aspire.

My friend and collaborator, Dr. Sergey V. Venev, added new dimensions to our paper and greatly improved my bioinformatics skills along the way. Dr. Troy Whitfield provide numerous useful discussions. The past and current members of the Matthews Lab have contributed immensely to my research. Dr. Sagar Kathuria always made time for me and others in the lab. Dr. R. Paul Nobrega led the graduate students with his dedication and intelligence. Kevin Halloran provided invaluable discussion, suggestions, and reminders. Dr. Jill Zitzewitz always knew what to say. Dr. Osman Bilsel had the calmest demeanor and thoughtful suggestions. My interaction with each lab member has trained me to be a better scientist: Noah Cohen, Meme Tran, Dr. Brian Mackness, Dr. Sujit Basak, Dr. Rohit Jain, and Dr. Nidhi Joshi.

Members of the Bolon lab, Dr. Ryan Hietpas, Dr. Ben Roscoe, Dr. Parul Mishra, Dr. Li Jiang, and Dr. Julia Flynn, provided incredibly helpful insights to my project. Adam Choi, a Ph. D. candidate working under the guidance of Dr. Stephen Miller, synthesized a key compound for my enzyme kinetic assay and provided valuable discussions throughout my training career.

Abstract

Sequence divergence of orthologous proteins enables adaptation to a plethora of environmental stresses and promotes evolution of novel functions. As one of the most common motifs in biology capable of diverse enzymatic functions, the TIM barrel represents an ideal model system for mapping the phenotypic manifestations of protein sequence. Limits on evolution imposed by constraints on sequence and structure were investigated using a model TIM barrel protein, indole-3-glycerol phosphate synthase (IGPS). Exploration of fitness landscapes of phylogenetically distant orthologs provides a strategy for elucidating the complex interrelationship in the context of a protein fold.

Fitness effects of point mutations in three phylogenetically divergent IGPS proteins during adaptation to temperature stress were probed by auxotrophic complementation of yeast with prokaryotic, thermophilic IGPS. Significant correlations between the fitness landscapes of distant orthologues implicate both sequence and structure as primary forces in defining the TIM barrel fitness landscape. These results suggest that fitness landscapes of point mutants can be successfully translocated in sequence space, where knowledge of one landscape may be predictive for the landscape of another ortholog.

Analysis of a surprising class of beneficial mutations in all three IGPS orthologs pointed to a long-range allosteric pathway towards the active site of the protein. Biophysical and biochemical analyses provided insights into the molecular mechanism of these beneficial fitness effects. Epistatic interactions suggest that the helical shell may be involved in the observed allostery. Taken together, knowledge of the fundamental properties of the TIM protein architecture will provide new strategies for *de novo* protein design of a highly targeted protein fold.

Table of Contents

Contents

List of Figures

List of Tables

List of abbreviations

ASA: accessible surface area

BASiC: branched aliphatic side chain

CD: circular dichroism

CdRP: 1-(o-carboxyphenylamino)-1-deoxyribulose 5-phosphate

EMPIRIC: Extremely Methodical and Parallel Investigation of Randomized Individual Codon

I: intermediate state

IGP: indole-glycerol phosphate

IGPS: indole-3-glycerol phosphate synthase

ILV: isoleucines, leucines, and valines

k_{cat}: turnover number

k_{eff}: catalytic efficiency

K_m: Michaelis constant

MRE: mean residual ellipticity

N: native state

PCA: principal component analysis

s: selection coefficient

SCA: statistical coupling analysis

ScIGPS: *Saccharomyces cerevisiae* IGPS

SsIGPS: *Sulfolobus solfataricus* IGPS

TIM: triosephosphate isomerase

T_m: melting temperature

TmIGPS: *Thermatoga maritima* IGPS

TtIGPS: *Thermus thermophilus* IGPS

U: unfolded state ensemble

Vmax: maximal velocity

Preface

The work presented in Chapter II has been previously published as *Chan, Y. H., Venev, S. V, Zeldovich, K. B. & Matthews, C. R. Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. Nat. Commun. 8, 14614 (2017)*. This paper was the result of a collaborative effort. Dr. C. Robert Matthews and I designed the experiment. I performed all the experimental work. Dr. Matthews and Dr. Konstantin B. Zeldovich guided the data analysis. Dr. Zeldovich, Dr. Sergey V. Venev, and I analyzed the dataset. I wrote the manuscript with contributions from all authors.

The work presented in Chapter III is a preliminary body of work that explores the phenotypic manifestation of protein sequence at the molecular and cellular level. Some experiments have only been carried out once and will need to be repeated. Dr. C. Robert Matthews and I designed the experiments. The experimental work was carried out by me as well as several summer interns who worked under my guidance: Lorein M. Rodriguez, Grace S. Ahn, Philip Economou, and Katherine Edwards. The substrate used for the functional assay was synthesized by Adam Choi, a Ph. D. candidate working under the guidance of Dr. Stephen Miller. The data interpretation is the work of myself and Dr. Matthews with contributions from Dr. Konstantin B. Zeldovich.

Chapter I – Introduction

General overview: Intersecting protein biophysics and molecular evolution

Proteins are cellular workhorses that perform diverse functions required for survival and proliferation. From chemical reactions to transportation to structural integrity, each protein begins as an unfolded chain of amino acids and, often, must fold to a specific three-dimensional structure for function. Instructions for reaching the native structure and the inherent stability for maintaining that conformation are encoded within the polypeptide. Improperly folded, partially folded, or unfolded proteins can lead to loss of function with catastrophic effects. As such, protein sequences at the molecular level can have organismal impact on fitness. At the extracellular level, environmental variability will determine the costs and benefits of a protein's molecular phenotype such as protein stability, dynamics, and activity. Thus, genetic diversity through genetic drift and selection is the keystone for organism fitness, permitting adaptation to the local environmental stress and leading to evolution over time.

Genetic information is stored in DNA. In 1958, Francis Crick coined the term "central dogma of molecular biology" to describe the flow of genetic information from DNA to RNA to functional protein[1]. The nascent, unstructured polypeptide resulting from transcription of DNA and translation of RNA requires additional processing for proper function including protein folding. Equally important is the maintenance of the protein structure once the native state is reached. Detailing the molecular mechanisms behind these fundamental

processes has sparked numerous studies and lively debates, serving to elucidate the complex relationship between sequence, structure, stability, and function.

Historical highlights and general principles of protein folding

A landmark experiment by Christian Anfinsen in the 1960s demonstrated that proteins fold spontaneously and reversibly[2]. Starting with folded RNase protein, he fully unfolded and reduced the enzyme in urea and 2-mercaptoethanol. Upon removal of both denaturant and reducing agent, Anfinsen observed full recovery of the enzyme activity, demonstrating that the information required for folding is inherent in the sequence. Implicit in this study is that proteins fold in a thermodynamically controlled, pathway independent manner; the folded conformation is the native state representing the global minima.

For biological relevance, a protein must also fold rapidly. After, Cyrus Levinthal put forth a thought experiment questioning the paradoxical discrepancy between long time required for a protein to obtain its native structure by random sampling of conformations and the short time observed during natural protein folding[3]. His solution was that proteins must fold by specific pathways. With kinetically controlled folding, the stable folded conformation is dependent on initial conditions and may be a local minimum rather than the global minimum.

The apparent juxtaposition between the two folding models was largely resolved through the use of funneled-shaped, energy landscape models, where the width of the funnel represents conformational entropy and the height

represents free energy. Landscape theory postulates that the native structure is the thermodynamic consequence of the lowest energy state allowed by the sequence, while, mechanistically, the process folding is kinetically driven by favorable transitions from higher to lower energy states[4]. Proteins exist in an ensemble of states, where each conformation is a single microstate on the energy landscape. Numerous conformations of unfolded proteins populate the maximal energy state at the height and maximal width of the funnel. Rather than a single pre-defined pathway, proteins are free to sample any number of downhill conformations, concomitantly increasing structure content and decreasing conformational entropy, *en route* to the most energetically favored native state.

Protein sequence and stability

For proteins to maintain their structure, favorable stabilizing protein interactions must outweigh the conformation entropic cost of protein folding. Favorable intramolecular interactions dictated by the sequence guide proteins down the energy gradient and determine the stability of the native state, resulting in proteins of the same fold having different stabilities. Most proteins are minimally stable, with a free energy difference between a folded and an unfolded conformation typically ranging from 5 to 10 kcal·mol^{-1} under physiological conditions[5]. Twenty unique natural amino acids are incorporated into proteins and can be broadly classified by their sidechains into three categories: non-polar, polar, and charged. The unique physical and chemical properties of each amino acid type drive protein folding and bias the structure and stability through various

interactions. The stability conferred by specific interactions may differ from the *in situ* bond free energy depending on the local environment of the bonds and how often the bond is formed in the native state[6].

Types of protein interactions

Some protein interactions are specific to amino acid type. Disulfide bonds are covalent bonds formed between the thiol groups of two cysteines. Ionic bonds are formed between amino acids with oppositely charged acidic ($-COO^-$) and basic ($-NH3^+$) R groups (1-3 $kcal \cdot mol^{-1}$)[7]. Two major forces in protein stability are hydrogen bonding and hydrophobic interactions. Hydrogen bonds (H-bonds) are electrostatic attractions between a hydrogen atom (donor) covalently bound to a high electronegative atom such as those found in polar residues and another electronegative atom (acceptor) (2-10 $kcal \cdot mol^{-1}$)[8]. On average, there are 1.1 hydrogen bonds per residue, with a majority of these bonds between peptide groups[9]. H-bonds are critical for forming and maintaining secondary structure. Burial of polar groups by H-bonding increases packing density, which increases Van der Waals interactions[10]. Van der Waals forces are attractions caused by temporary electric dipoles induced by two nearby atoms. While weak, their abundance can provide substantial stability. Similar to H-bonds, hydrophobic forces are critical for protein stability. In addition, they are the major driver of protein folding. Maximal burial of hydrophobic side chains leads to a compact conformation with an aliphatic core and a polar surface. Increasing compactness gives rise to steric constraints, favoring secondary structure formation and

restricting accessible conformations to the unique native state[11]. Hydrophobic interactions stabilize proteins by removing nonpolar side chains from water; burial of a single methylene (-CH_2-) group confers 0.6 to 1.6 kcal·mol^{-1} of stability[9]. Clearly, the relative distribution and abundance of each amino acid dictate their contribution in protein folding and stability for any given environment.

BASiC hypothesis

While a global view of how proteins fold has been largely established, the specific mechanisms by which protein sequence directs protein folding and maintains the protein structure has yet to be elucidated. The Matthews lab has proposed the Branched Aliphatic Side Chain (BASiC) hypothesis that suggests side chain burial of isoleucines, leucines, and valines (ILV) drive protein folding in high energy states and create tightly packed, hydrophobic clusters that form stable cores in the native state through the exclusion of water[12,13]. Amino acids with branched aliphatic side chains are the most hydrophobic residues[14]. Energetically favorable burial of these side chains drive early folding by providing a stable platform for other folding events to build upon[15,16]. Reorganization among these highly aliphatic side chains incurs minimal energy penalty since they can slide by each other to repack[17]. Once folded to the native conformation, large tightly packed clusters of ILV residues effectively exclude water penetration[18]. NMR experiments performed on two structurally conserved TIM barrels, the alpha subunit of tryptophan synthase (αTS) from *Escherichia coli* and indole-3-glycerol synthase (IGPS) from *Sulfolobus solfataricus*, show different

protection patterns that do not correlate with structural elements [19–21]. Rather, the most protected residues correlated with the deeply networked H-bonded branched aliphatic residues. Thus, the cores of the stability seem to vary in location with the ILV clusters[13]. The BASiC hypothesis provides a mechanism by which the primary sequence guides secondary and tertiary structure formation during protein folding as well as a metric for identifying regions of the protein that are likely to be most stable.

Protein evolution and physical constraints

How does protein stability guide molecular evolution? Underlying the plethora of sequences observed is the culmination of millions of years of molecular evolution, constrained by checks and balances of protein biophysics to ensure stability, structure, and function. Protein biophysics links genotype to phenotype, and is a crucial factor between a mutation accumulating randomly due to genetic drift or more selectively due to fitness. Evolution occurs mainly and gradually by divergence through accumulation of point mutations[22]. Most genetic variation is acquired by genetic drift and confers neutral fitness for the organism, where fitness is survival and reproductive success[23]. The marginal stability that is explained thermodynamically as the balance between conformation entropy and favorable interactions can be rationalized longitudinally as a balance between mutation and selection[24]. Introduction of new enzymatic function are often associated with destabilizing mutations, commonly thought of as the activity-stability tradeoff. Neutral mutations with no observable functional

role may offset the destabilizing effects of new functional mutations[25]. Proteins that are more stable may be more evolvable because of their greater tolerance for new functional, destabilizing mutations[26].

While protein stability may promote sequence diversity, selection for protein stability may slow the evolution rate for new structures[27]. The number of unique folds is greatly outnumbered by the vast pool of sequences. For highly stable folds, poor fitness associated with high energy structural intermediates impedes sampling of novel stable structures[27]. Within the universe of protein folds, some folds are heavily represented, while most are rare; this asymptotic power laws distribution for protein folds suggest that there are basic principles for structure designability and evolvability[28]. Design-wise, there are multiple possible strategies for evolvability along the thermodynamic continuum[29]. A “polarized” structure where the active site is loosely connected to a highly stable scaffold encourages functional innovation within a “robust” fold that is tolerant of sequence changes[30]. Loosely packed, less ordered proteins may also be tolerant to mutations because loss of a weak interaction may not have a large thermodynamic effect. Destabilization of the unfolded state by negative design may offer similar overall stability for proteins with high energy native states[31]. At the opposite end of the thermodynamic spectrum, disordered proteins that couple folding with ligand binding have multiple loosely packed native state conformations with similar energies that are highly tolerant to mutation accumulation[32]. In all three cases, mutations may shift the equilibrium from one

conformation to another within the *ensemble* of the native well, lending to promiscuous function or conformation and favoring functional and structural evolution[29].

Inverse protein folding problem

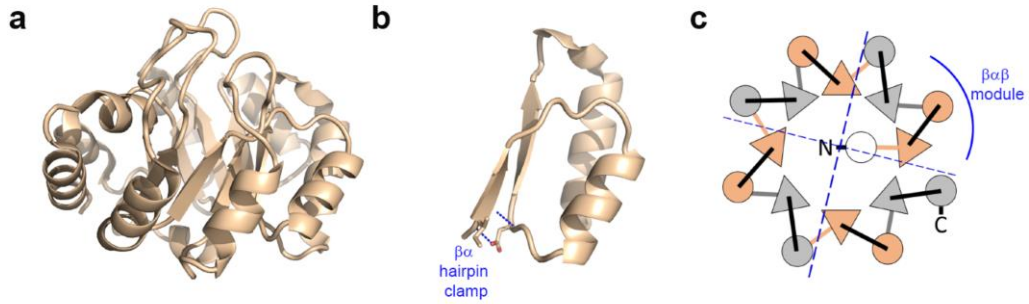
A single amino acid change can tip the equilibrium from one conformation to another, demonstrating the sensitivity of the protein fold to stability perturbation[33]. Valuable insights can be obtained by dissecting the interplay between sequence, structure, and stability from both the sequence and the structure perspective. The challenge in a classic protein folding problem is to predict a single native structure based on a single sequence. The inverse protein folding problem poses a different challenge where multiple optimal sequence solutions are possible for a single structure[34]. Given this degeneracy, determination of relative fitness for similar sequences may help define signature sequence elements and identify inaccessible sequence space for a specific protein fold.

TIM barrels as a model system for intersecting protein biophysics and protein evolution

Phylogenetic reconstruction of a universal protein architecture tree identified the triosephosphate isomerase (TIM) barrel along with the P-loop hydrolase and Rossman folds as the three most ancestral folds[35]. The deep lineage makes TIM barrels an ideal candidate for studying the evolutionary pathways of a protein fold through examination of the sequence and structural constraints leading to its successful incorporation in all branches of life.

In addition to being one of the most ancient folds, the TIM barrel is one of the most common enzyme topologies. The canonical structure consists of eight repeating βα units, where the β-strands form a parallel β-barrel tilted at 36° surrounded by an α-helical shell[36] (Fig. 1a). Loops connect each secondary structure element, with short (3-5 amino acids) αβ-loops found at the N-termini of the β-strands and longer βα-loops at the C-termini of the β-strands. A canonical GXD motif is found in αβ-loops leading to the even numbered strands[37]. Within this motif, the aspartic acid form a clamp with the preceding odd numbered β-strand (Fig. 1b)[38]. This βα-hairpin clamp registers the two consecutive strands and confers several kcal·mol^{-1} of stability, creating a fourfold symmetric βαβα unit. The minimal independently folding unit, the βαβ module, exists within this fourfold symmetry (Fig. 1c).

Figure 1.1 Canonical features of a TIM barrel structure



(**a**) Ribbon diagram of a TIM barrel protein (PDB: 1I4N). (**b**) A canonical βα hairpin clamp formed between a main chain amide H-bond donor in +1 position of the odd β-strand and the side chain acceptor at the −1 position in the following even αβ-loop is highlighted in blue. (**c**) Top view schematic of the TIM barrel. The four-fold symmetric βαβα repeat unit holds the minimal independently folding unit, the βαβ module. Triangles represent β-strands. Circles represent α-helices. Odd strands and helices are colored orange. Even strands and helices are colored gray.

Despite a high structural similarity, the sequences and functions of TIM barrels are highly divergent. Underscoring this sequence diversity, the normalized structural similarity SIMAX score for each of the 33 distinct superfamily ranges from 1.5 to 5 Å, but their sequence similarities are only 5 to 40% identical[37,39–41]. Through the incorporation of cofactors, TIM barrels evolved to catalyze a broad set of biochemical reactions[42]. The stable platform created by the β-barrel and short αβ-loops has high thermodynamic stability, cradling the active site found invariantly near the C-termini of the β-strands or within the adjoining βα-loops. This segregation of the active site on a stable platform is a prime example of a polarized structure that is highly evolvable[30]. Single-domain

TIM barrel proteins impart 13 unique oxidoreductase functions, 2 unique transferase functions, 10 unique hydrolase functions, 5 unique lyase functions, and 4 unique isomerase functions[42]. The result of this rich diversity of sequence and function is the high representation of TIM barrels in over 10% of all known enzymes[37], making this fold an ideal system for identifying sequence constraints on a protein structure, stability, and function.

Structure and stability of IGPS

IGPS is a well-characterized TIM barrel protein found in microbes such as archaea and bacteria as well as fungi and plants, but not in mammals. The stability and folding mechanisms of one archaeal ortholog *Sulfolobus solfataricus* IGPS (SsIGPS) have previously been studied by the Matthews Lab. Like other TIM barrel proteins, SsIGPS have relatively high thermodynamic stability, $\Delta G°_{NI}$ (H_2O) = 8.07±0.24 kcal·mol^{-1} and $\Delta G°_{IU}$(H_2O) = 6.29±0.55 kcal·mol^{-1}. Only one intermediate is detected by urea titration as measured by circular dichroism (CD), but other spectroscopic methods identified the hallmark TIM barrel folding mechanism, $I_{BP} \rightleftharpoons U \rightleftharpoons I1 \rightleftharpoons I2 \rightleftharpoons N$, for SsIGPS[21].

IGPS enzyme chemistry

IGPS catalyzes the third step in the tryptophan biosynthetic pathway[43]. A detailed mechanism of the chemical steps required for conversion of its substrate 1-(o-carboxyphenylamino)-1-deoxyribulose 5-phosphate (CdRP) to its product indole-glycerol phosphate (IGP) steps has been established[44]. The catalytic site of IGPS that performs the enzyme chemistry can be split into two functionally

distinct areas. Ring closure of CdRP is mediated by K110, which acts as a general acid in the condensation step prior to decarboxylation. The dehydration steps involve K53, serving as a general acid, and E51, serving as a general base. Another part of the active site is the phosphate binding pocket, which is mainly localized to the C-terminal end of the eighth β-strand.

Tryptophan biosynthetic pathway and regulation

As the most biochemically expensive amino acid pathway, the multi-step process of tryptophan biosynthesis is the highly regulated[45–48]. In bacteria, the canonical *trp* operon consists of seven genes, five enzymatic genes *trpEDCBA* with seven distinct catalytic domains including IGPS (*trpC*) to synthesize L-tryptophan from chorismate, and two regulatory genes to control expression of the structural genes[47]. In archaeal genomes, the order of *trp* genes are less conserved and may be clustered in several groups instead of a single operon. In *S. solfataricus*, the operon is intact[49]. Regulation of the *trp* operon occurs on the transcription, translation, and protein level through feedback inhibition. Orthologous genes to the prokaryotic tryptophan biosynthetic ones are found in *S. cerevisiae*, but they are located on different chromosomes and are unlinked[50,51]. Dual regulation of the pathway in yeast involves feedback inhibition and enzyme derepression[52].

Depending on the host organism, IGPS can be expressed as either a monofunctional or a bifunctional protein tethered to another tryptophan synthetic enzyme in a single gene. IGPS from archaeon *S. solfataricus* and bacteria *T.*

maritima, and *T. thermophilus* are expressed as single-domain, monofunctional proteins, whereas IGPS from *S. cerevisiae* is translated with anthranilate synthase (AS) component II from a single transcript and IGPS from *E. coli* is translated as a bifunctional enzyme with phosphoribosyl anthranilate isomerase (PRAI) [53–56]. In all cases, the IGPS domain catalyzes a single reaction. Bifunctional proteins can be stably expressed as active monomers[57].

Fitness landscapes

Fitness landscapes map genotypes to fitness and are useful visualization tools to identify local and global fitness maxima and minima of closely related sequences. Exploration of sequence space may provide clues to the accessibility of mutational paths on the fitness landscape during molecular evolution. Technology development has greatly advanced the breadth and depth of sequence space that can be probed in a short amount of time. Methods to probe the sequence-structure-stability-function relationship include sequence exchange experiments, circular permutation, directed evolution, and library screening[58–61]. High throughput mutagenesis schemes to create mutant libraries are gaining traction for parallelizing fitness studies across numerous sequences[62–70].

EMPIRIC fitness screen

Extremely Methodical and Parallel Investigation of Randomized Individual Codon (EMPIRIC) is a systematic approach to experimentally determine fitness landscapes[62]. Using this technique, a plasmid library consisting of all possible individual point mutations for a 10 amino acid region is generated through a

series of molecular manipulations[71]. Functional complementation of a conditional host with mutant gene is selected for during a bulk growth assay where all the mutants compete against each other. Fitness for each mutant is determined by deep sequencing of samples collected throughout the growth competition. Two major advantages to this deep scanning approach are comprehensive library generation for the region of interest and large dynamic range for detection[62].

Scope of project

This project examines the role of protein sequence on stability, structure, and function to ask two separate but related questions: (1) Are fitness landscapes of orthologous proteins correlated and (2) how do sequence and structure influence the molecular mechanisms that shape the fitness landscape of a protein fold? The ideal system for this study is the TIM barrel fold, an ancient and highly represented fold in biology known for its sequence diversity. To minimize complexities in analyses at the offset, a monofunctional enzyme whose biological role has been well characterized was chosen. Available biophysical and biological characterization of IGPS provided valuable insights to the experimental design of the fitness screen. Three thermophilic IGPS orthologs, one archaeal and two bacterial, with high quality structural information available were compared. Importantly, despite sharing only 30-40% sequence identity, high structural similarity between orthologs was observed. To increase the sequence space studied, the EMPIRIC deep mutational scanning approach was applied to each of the orthologs. Focus on understanding protein stability in

relation to sequence led to the choice of known stability elements, the β-barrel region and αβ loops, as the mutagenesis sites. A growth competition with IGPS knockout yeast transformed with the mutant library screened for fitness advantages and defects.

In Chapter II, the fitness landscapes for the three orthologs are compared. Through a combination of sequence, structural, and bioinformatics analyses, the fitness landscapes of evolutionary distant orthologs were shown to be correlated through sequence and structure space. Similar fitness landscapes between orthologs imply that the structure of the TIM barrel fold imposes a specific fitness landscape. An unanticipated discovery of several beneficial point mutations, distal from the active site and mirrored throughout the four-fold protein symmetry, revealed an intricate sequence-structure-fitness relationship.

In Chapter III, ongoing and future studies on the mechanism for the beneficial, allosteric mutations are examined through a series of *in vitro* experiments using single and double point mutants. Allosteric mutations, described here as mutations distal from the active site that affect the catalytic efficiency of enzyme, have been previously reported[72–75]. Long-range effects on enzyme activity imply that distal residues are energetically linked to each other. A single point mutation may induce population changes in the equilibria between different conformations within the ensemble; altered thermal fluctuations may increase sampling of conformations with favorable donor-acceptor distances in

the active site[76]. Higher order energy coupling studied through double mutant cycles may distinguish if interactions are additive, synergistic, antagonistic, or absent[77]. Epistatic interactions can occur between distant sites greater than 20Å[78].

In an effort to provide a holistic view of the allosteric mechanism for the IGPS TIM barrel system, several parameters are considered for each mutant: selection coefficient, secondary structure content, protein stability, and enzyme kinetics measurements. Temperature stress for these thermophilic orthologs working under mesophilic conditions within the yeast host may select for mutants adopting higher energy conformations for increased activity. Mutation of sequence optimized for one environment may reveal long range underlying interactions affecting function. Unraveling the mechanism of the allostery provides potential avenues for protein and drug design.

Chapter IV will summarize the results of this work in the broader context of sequence driven fitness. Taken together, these insights have implication in our understanding of the potential evolutionary pathways of TIM barrels dictated by the sequence and structure and the molecular mechanisms shaping fitness.

Chapter II – Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints

This chapter has been published previously:

Chan, Y. H., Venev, S. V, Zeldovich, K. B. & Matthews, C. R. Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. *Nat. Commun*. **8**, 14614 (2017).

The published work presented in this chapter was a collaborative effort. Dr. C. Robert Matthews and I designed the experiment. I performed all the experimental work. Dr. Matthews and Dr. Konstantin B. Zeldovich guided the data analysis. Dr. Zeldovich, Dr. Sergey V. Venev, and I analyzed the dataset. I wrote the manuscript with contributions from all authors.

Introduction

Proteins carry out diverse and essential functions in all living organisms. Notable among these functions is the catalysis of a host of complex chemical reactions under a broad range of environmental conditions. Based on enzyme classification, proteins catalyse over 5,700 unique biochemical reactions by employing ~1,400 unique folds[79]. This structural redundancy shows that enzymes capitalize on robust structural platforms to introduce novel chemistries through sequence diversification. The TIM barrel fold is one of the oldest and most common motifs in biology[37]. The polarized structure has a well-ordered scaffold that is fused to an autonomous domain for substrate binding, catalysis, and product releases; the spatial segregation of the active site from the stabilizing protein core is an elegant solution for evolving new functions of TIM barrels[30]. The TIM barrel superfamily contains 57 distinct families and encompasses five of the six enzyme commission functional categories catalysing at least 34 unique functions[42].

As sequences and functions of TIM barrel fold proteins emerged through a divergent evolutionary process, a quantitative description of the TIM barrel fitness landscape would be a crucial step in our understanding of protein evolutionary dynamics. Recent deep mutational scanning experiments provided significant insight into the fitness landscapes of individual proteins[62–70] or very closely related homologues[80]. It was found that thermodynamic effects of mutations are not very sensitive to sequence background, prompting biophysical models of

sequence evolution[67,81–85]. Furthermore, mutational scans showed that site-specific amino-acid preferences and, presumably, fitness landscapes are nearly identical in homologues of influenza virus nucleoprotein with 94% identity[80]. To date, it remains unclear whether the fitness landscapes remain similar in homologous proteins with a strongly divergent evolutionary history.

Here we perform a mutational scan of three orthologous TIM barrel fold proteins, indole-3-glycerol phosphate synthase (IGPS), to experimentally determine the fitness landscapes of proteins sharing the same fold and function, but with ancient divergences and low sequence identity. IGPS proteins were mutagenized in eight 10-residue segments, following the fold symmetry and covering the β-barrel core and adjacent elements of the αβ- and βα-loops. A tryptophan auxotrophic *S. cerevisiae* yeast strain, created by deletion of the endogenous IGPS gene, was transformed to prototrophy with each of the three orthologous genes. Relative fitness of the mutants, which depends on IGPS activity, was determined by the abundance of the mutant DNA sequence relative to wildtype (WT) over time[86].

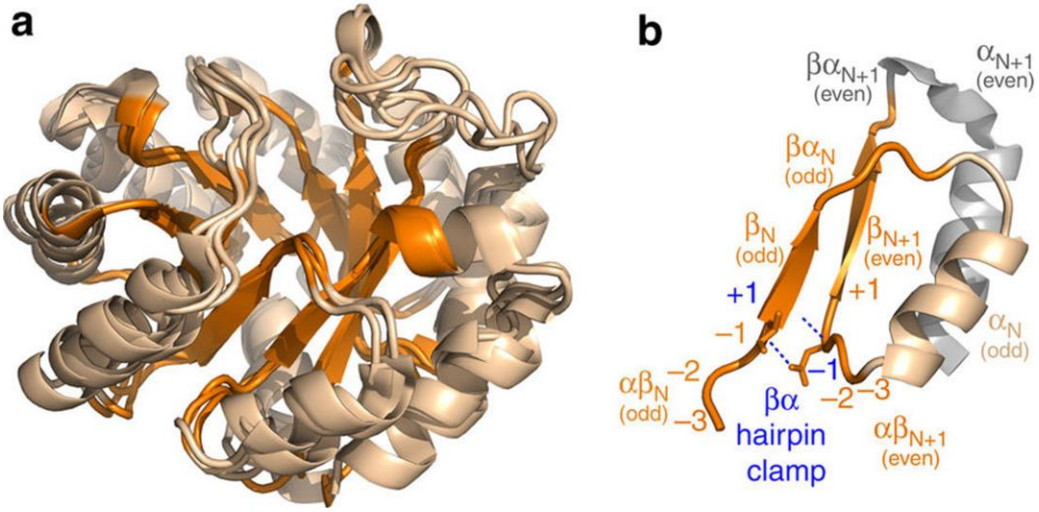
A comprehensive analysis of 5,040 mutations in the three orthologues demonstrated that the fitness landscapes of IGPS orthologues are statistically significantly correlated to each other, despite the sequence identity of approximately 30–40%. Surprisingly, we found that fitness can be dramatically enhanced by mutations distal from the active site in all three orthologues. Fold

geometry and structural elements of the TIM barrel fold, as well as sequence conservation and amino-acid biochemistry, all impose measurable constraints on the fitness landscape. Principal component analyses (PCA) detected commonality between sources of fitness variance in the three IGPS TIM barrel orthologues, while statistical coupling analysis (SCA) revealed evolutionary correlations between the active site and positions of the distal beneficial mutations. These results have significant implications for the design of TIM barrel enzymes with novel functions not found in nature and insights into the unanticipated allostery for the TIM barrel motif.

Results

We applied the EMPIRIC deep mutational scan approach, developed by the Bolon group[71], to explore the fitness landscape of a monofunctional enzyme of the TIM barrel fold. In the canonical TIM barrel structure, eight β-strands and α-helices alternate in sequence, and the β-strands assemble sequentially into a cylindrical core around which the α-helices form a helical shell (Fig. 2a). The TIM barrel scaffold is highly symmetrical, displaying a four-fold βαβα symmetry at the level of the smallest independent folding unit, the βαβ module (Fig. 2b)[38].

Figure 2.1 Orthologous IGPS proteins fold into highly similar canonical TIM barrel structures



(**a**) Ribbon diagrams of structurally aligned SsIGPS (PDB: 2C3Z), TmIGPS (PDB: 1I4N), and TtIGPS (PDB: 1VC4). EMPIRIC mutagenesis library positions are highlighted in orange. The parallel β-strands vary in length from 4–6 residues and the β-barrel structure forms four layers of side chains from alternating β-strands that protrude into the protein core (Supplementary Fig. 2.1). Preceding the N terminus of the β-strands, short αβ-loops, generally 3–4 residues in length, are proposed to play a role in stability. At the C-terminus of the β-strands, long-βα loops, between 5 and 13 residues, link the C termini of the β-strands to the N termini of the subsequent α-helices and invariably form the active site of the enzyme. (**b**) The four-fold symmetric βαβα repeat unit holds the minimal independently folding unit, the βαβ module, highlighted in orange, wheat, and orange, respectively. The even helix shaded in grey does not contribute significantly to stability of the four-fold unit. Within the βαβ module, a canonical βα hairpin clamp formed between a main chain amide H-bond donor in +1 position of the odd β-strand and the side chain acceptor at the −1 position in the following even αβ-loop is highlighted in blue.

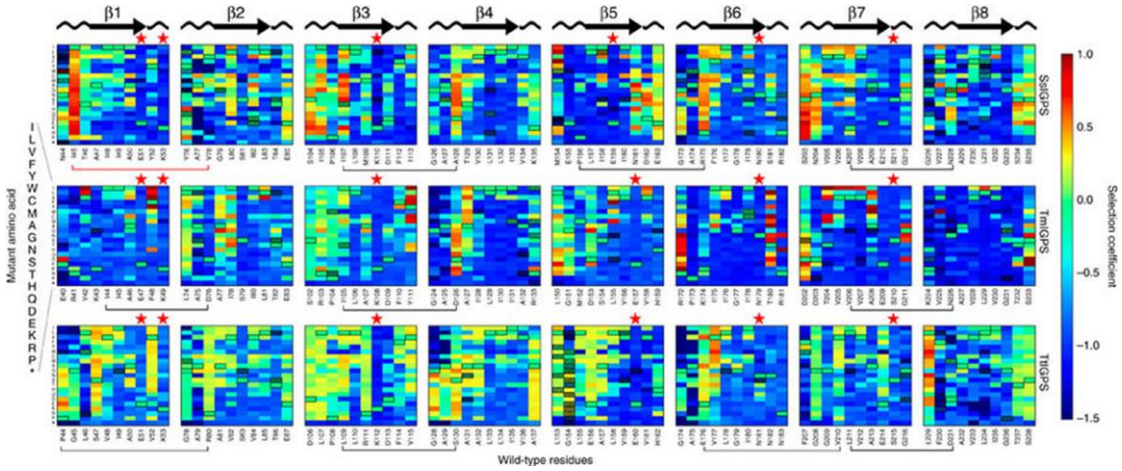
We chose three phylogenetically divergent IGPS orthologues from a thermophilic archeaon, *S. solfataricus* (SsIGPS), and two thermophilic bacteria, *T. maritima* (TmIGPS) and *T. thermophilus* (TtIGPS), Table 2.1.

Table 2.1: Pairwise sequence and structure similarity of the three IGPS orthologues and mutagenized libraries

	SsIGPS	TmIGPS	TtIGPS
SsIGPS	—	**Library sequence:** Identical: 51% Similar: 71%	**Library sequence:** Identical: 49% Similar: 68%
TmIGPS	**Full-length sequence:** Alignment length: 267 Identical: 30% Similar: 49% Structure: Alignment length: 216 r.m.s.d.: 1.58 Å	—	**Library sequence:** Identical: 54% Similar: 73%
TtIGPS	**Full-length sequence:** Alignment length: 271 Identical: 35% Similar: 49% Structure: Alignment length: 214 r.m.s.d.: 1.24 Å	**Full-length sequence:** Alignment length: 277 Identical: 27% Similar: 43% Structure: Alignment length: 241 r.m.s.d.: 1.72 Å	—

For the three proteins, we carried out selection experiments in yeast to determine the fitness of all possible mutations in libraries of 10 positions spanning the 8 αβ-loops, β-strands, and initial portion of the βα-loops (Fig. 2.2).

Figure 2.2 Fitness landscapes of the three IGPS orthologues



Values of the selection coefficient are color-coded on a continuous scale from 1 to −1.5 indicated by the colorbar. WT residues and positions are labelled at the bottom of each panel. Mutant amino acids are indicated on the vertical axis. Within the heatmap, WT residues are indicated by the black outline. Low-quality data filtered from analysis are indicated by the red checkered boxes. Canonical secondary structures are drawn at the top of the panels. Active sites are indicated by the red stars at the top of the position columns. βα hairpin clamps are indicated by the black brackets at the bottom of the position columns. Fitness gains were observed with several mutations of the βα hairpin clamps. For example, more than half the mutations in the three β3α3 hairpin clamps SsIGPS I107 and D128, TmIGPS I105 and D126, TtIGPS I109 and S130 are beneficial (s>0). Two non-canonical interactions analogous to the βα hairpin clamp in SsIGPS are indicated by the red brackets. An ionic interaction is observed between E155 and R175. A hydrophobic stacking interaction is observed between I45, V78, and F40 (not included in library). All mutations of SsIGPS I45, except to stop codons, resulted in fitness advantage over WT.

Mutations were introduced by restriction of the gene followed by its ligation with a cassette containing all 64 codons for a given position, as opposed to transcribing the gene using an error-prone polymerase[66,71]. This approach generates a well-defined, rather than random, sequence diversity in the mutagenized region, reducing experimental noise.

Fitness was quantified as the selection coefficient *s*, that is, the slope of the relative abundance of mutant to WT IGPS over time (see Methods and Supplementary Fig. 2.2). Mutants with selection coefficient of 0 have doubling time equal to WT, while those with a selection coefficient of −1 are lethal. Over 55% of the mutations were deleterious with $s<-0.75$. The β-strands showed a particularly low tolerance to mutation. Presumably, mutations in the core destabilize the native state, resulting in a lower population of functional IGPS enzymes, or distort the active site, reducing catalytic power. Six catalytically important residues form the active site and are located at or near the C termini of β1, β3, β5, β6 and β7. Mutations at these highly conserved residues were strictly not tolerated. Bimodal distributions are often observed with fitness landscapes, describing a thermodynamic ‘cliff’ at which proteins unfold and the mutations are lethal[83,85,86]. While we observed the expected bimodal distribution, the high fitness mode showed a bias towards

beneficial rather than neutral fitness (Supplementary Fig. 2.3). Positive selection coefficients were associated with several mutations in the active site βα-loops at positions that do not directly support enzyme chemistry. Remarkably, fitness can be increased by mutations far from the active site in the αβ-loops, most notably at the positions of βα-hairpin clamps (Fig. 2.2). βα-hairpin clamps are long-range hydrogen bonds between the even numbered αβ-loops and the preceding odd numbered β-strands, enforcing β-strand alignment and providing stability (Fig 2.1b)[87]. Based on the repeating fitness patterns in the minimally independent folding βαβ modules contained within the four-fold βαβα symmetrical units for each of the three orthologues, we analysed the selection coefficients in groups of two consecutive libraries, that is, odd (*n*) and even (*n*+1) strands (Fig 2.1b).

αβ-loops

Within the canonical TIM barrel structure, there is a bias towards the glycine-x-aspartic acid motif in the αβ-loops leading up to the even number β-strands[37,87]. In the even numbered αβ-loops, mutations away from small hydrophobic residues at −2 position (X in the motif) with respect to the β-strand, prevent the tight turn required to dock the preceding helix. We found that this distortion of the βαβα module results in low organismal fitness, likely reflecting destabilization of the native state from poor

packing. In contrast, the odd numbered αβ-loops link adjacent modules and are less constrained in their residue choice. Many mutations at the −3 and −2 positions of the odd strands were beneficial. At the −1 position, the higher fitness in the even versus the odd strand reflects their different accessible surface area (ASA). The average selection coefficient of these positions is correlated with the ASA, where mutations at the more buried even loops resulted in lower fitness than mutations in the more solvent exposed odd loops (Supplementary Fig 2.4). Similar results were found using the relative surface area (RSA)[88] instead of ASA.

βα-hairpin clamp

Loss of the βα-hairpin clamp between the canonical aspartic acid at the −1 position of the even strand and the main chain amide hydrogen beneath a large hydrophobic side chain at the N terminus of the preceding odd numbered β-strand elicited a fitness gain relative to the WT protein (Figs. 2.1b and 2.2). This enhancement is the most dramatic at the β3α3 hairpin clamp in all three orthologues. Significantly, mutation of I107 in the SsIGPS β3α3 hairpin clamp results in high fitness, while just one position away in the β-barrel, L108, mutations are mostly deleterious. The distinctly different responses highlight the different roles of the αβ-loop and the β-strand in protein stability and enzymatic function, each driving fitness of the host organism (Fig. 2.2). The most beneficial mutation in SsIGPS was

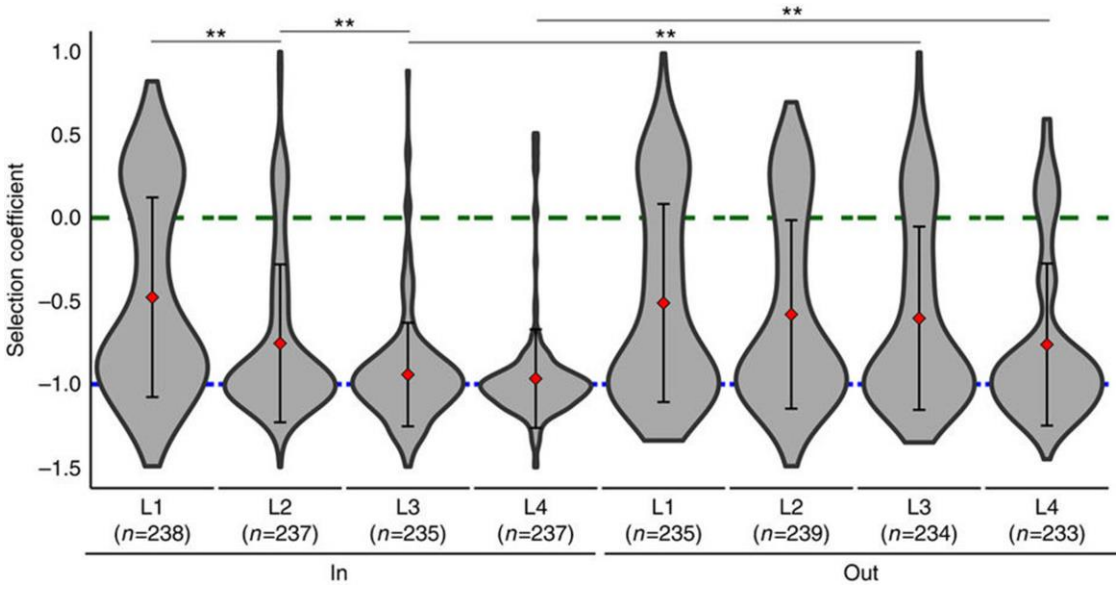
I45Q, with selection coefficient of +0.89. Position I45 is involved in a triple stack of hydrophobic side chains where F40 (α0) is sandwiched between I45 (αβ0-loop) and V78 (αβ1-loop). This bridge between strands β1 and β2 is an alternative solution to the βα-hairpin clamp typically found to stabilize odd and even β-strands. Mutations of the I45 to any other amino acid, except stop codons, are uniformly beneficial (Supplementary Fig. 2.5).

β-strands

The largely deleterious mutations in the β-strands indicate that minimal changes are tolerated within the protein core. Similar to the αβ-loops, the residues in each β-strand can be further stratified by side chain orientation into or out of the β-barrel and by participation in the four layer levels within the β-barrel (Supplementary Fig. 2.1). The evolved structure maximizes buried surface area of side chains, while minimizing steric clashes, by alternating side chain orientation between odd and even strands within a layer and between layers within β-strands[36,89]. Side chains that point inward form a highly stable hydrophobic core, while side chains pointing outward provide docking surfaces for the concentric α-helices by hydrophobic interactions. For side chains pointing outward in layers 2–4, a significant fraction of mutations is beneficial, in contrast to the almost complete lethality of their inward facing counterparts (Fig. 3). Mutations of β-strand residues pointing out of the barrel would be expected to perturb

the α-helix and β-strand interface as well as the intervening βα-loop. Interestingly, inward facing side chains in layer 1 can also increase the fitness for a significant fraction of the mutations. For mutations in the fourth layer, the poor fitness may reflect the known sensitivity of the βα-loops to the interface between the stability core and the active site loops[90].

Figure 2.3 Distribution of fitness values displayed as violin plots stratified by side chain orientation ‘in’ (left) and ‘out’ (right) and by layers (L1 to L4)



Guide lines are drawn at s=0 in green and s=−1 in blue. The mean and s.d. are indicated by the red diamond and error bars. Sample numbers are indicated beneath each layer. Permutation tests (n=10,000) were used to assess statistical significant differences between distributions based on ratio of beneficial (s>0) to total mutations (**P<0.02). See text and Supplementary Fig. 2.1 for details.

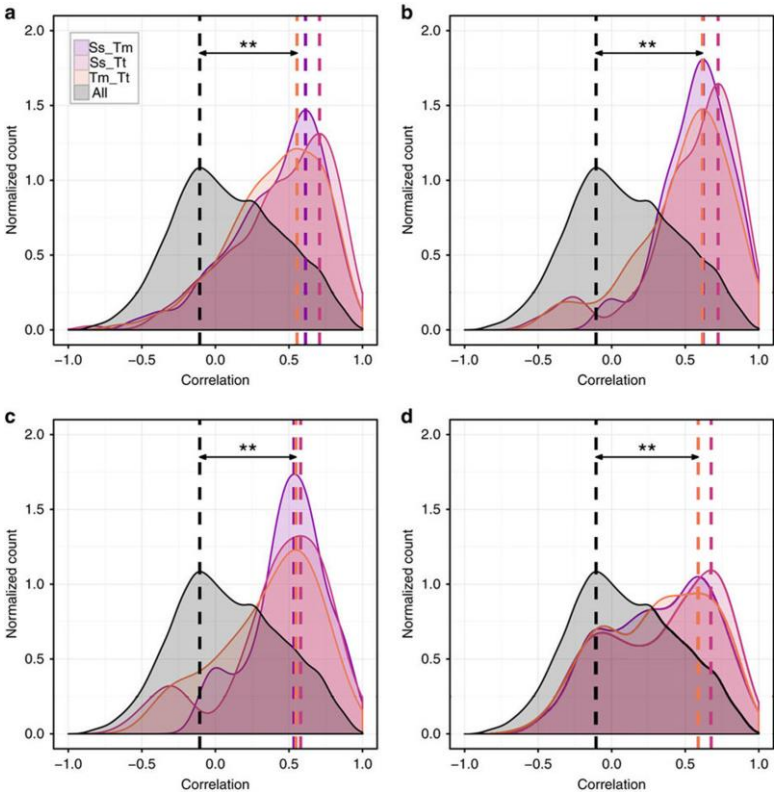
βα-loops and the active site

The active site spans multiple β-strands and subsequent βα-loops (Fig. 2.2). Although mutations at catalytically important residues were uniformly deleterious, some nearby mutations showed improved fitness over WT. Across the three orthologues, beneficial mutations were observed at position 52 (SsIGPS numbering) between the glutamic acid (E51) and lysine (K53) active site residues bridging the β1-strand and βα1-loop interface. Coordination of the substrate to its reactive conformation is mediated by electrostatic interactions between K53 in the βα1-loop and K110 in the βα3-loop[91]. However, molecular dynamics simulations of thermophilic SsIGPS under mesophilic temperatures suggest that strong electrostatic interactions between K53 and the substrate result in a nonproductive substrate geometry[92]. Our fitness results suggest that disrupting the ionic interactions between E51 and the two lysines, K53 and K110, may favor the reactive conformation of the substrate. Other beneficial mutations include the phosphate binding residue, S234, and its adjacent residue, S235. These residues, located in the eighth ‘βα-loop’ form a 3_{10}-helix, and mutation of either may distort the phosphate binding pocket so as to improve substrate binding and/or product release.

Correlation of fitness landscapes between IGPS orthologues

A major question emerging from this screen is whether protein fitness landscapes are conserved across sequence and/or structure space. We compared the fitness landscapes of the orthologues by calculating the Pearson correlation coefficient *R* between fitness values of the 20 mutant amino acids at a pair of positions in two orthologues. The probability distribution of *R* for a specific set of positions was then used to assess the similarity of fitness landscapes. We considered the following four sets of positions: (a) identical WT amino acids in a pair of orthologues irrespective of structural alignment; (b) all structurally aligned positions irrespective of WT amino acid; (c) all structurally aligned position with non-identical WT amino acids; and (d) positions aligned by their four-fold symmetry in the TIM barrel, irrespective of WT amino acid (Fig. 2.4).

Figure 2.4 Fitness landscapes of orthologous proteins are correlated despite their low sequence identity

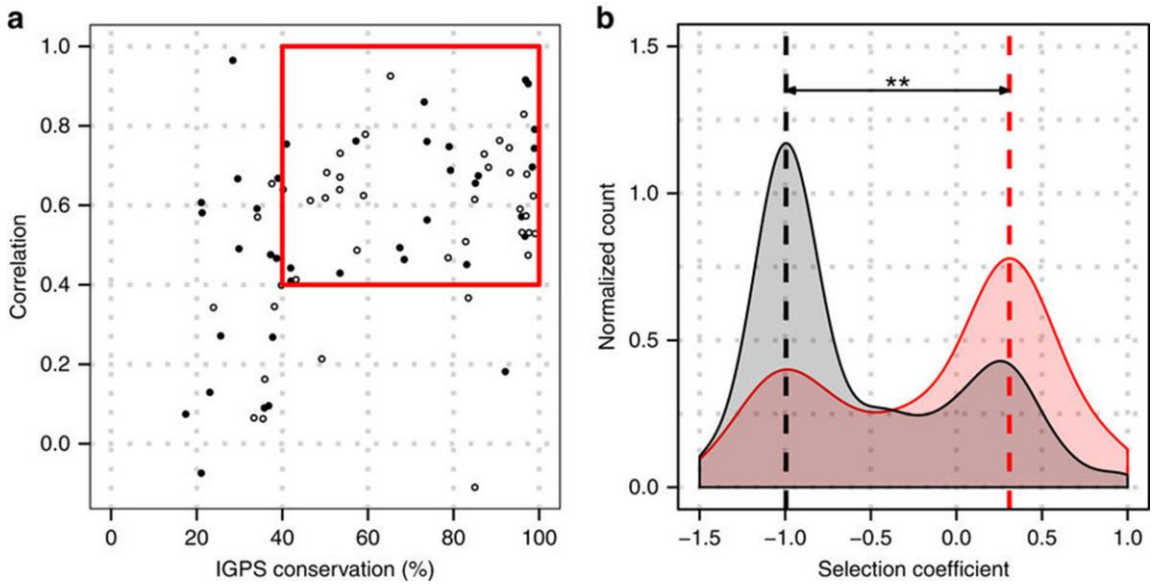


Distributions of Pearson correlation coefficients between fitness landscapes of (**a**) identical WT amino acids in a pair of orthologues irrespective of structural alignment, (**b**) all structurally aligned positions irrespective of WT amino acid; (**c**) all structurally aligned position with non-identical WT amino acids, and (**d**) positions aligned by their four-fold symmetry in the TIM barrel, irrespective of WT amino acid. Fitness landscapes are statistically significantly correlated despite the low sequence identity and ancient divergence of IGPS orthologues (**P<0.02). Statistical significance is detailed in Supplementary Table 2.1.

In all cases, the fitness landscapes of these groups were significantly different from the null model (pairwise correlations between all positions) according to Kolmogorov–Smirnov test ($P<10^{-4}$, Supplementary Table 2.1). The correlations in all three pairs of orthologues were similar in magnitude, and in most cases their distributions were statistically indistinguishable (KS test, $P>0.05$, Supplementary Table 2.2). The modes of these specific distributions corresponded to the Pearson correlation between ~0.5 and 0.7. On the other hand, distributions of correlations for different sets of positions (for example, (a) versus (b) for all orthologues) were different in most cases (KS test, $P<10^{-4}$, Supplementary Table 2.3), suggesting that structure and sequence features define the fitness landscape. These results are statistically significant compared to variation between experimental replicates (see Methods and Supplementary Figs 2.9 and 2.10).

Despite ancient divergence and bacterial versus archaeal origins, we found that fitness landscapes of three IGPS proteins are significantly correlated. Remarkably, for positions exhibiting as low as 40% sequence conservation in the multiple sequence alignment of IGPS proteins, experimentally determined fitness landscapes of the three orthologues remained correlated to each other, $R\geq0.4$ (Fig. 2.5a), complementing earlier observations using sequence alignments and mutational scans[80,93].

Figure 2.5 Sequence conservation and epistasis affect fitness landscapes



(**a**) Fitness landscapes of IGPS orthologues are significantly correlated (*R*>0.4) for IGPS positions ('open circle' odd numbered libraries, 'filled circle' even numbered libraries) displaying 40 to 100% conservation in the multiple sequence alignment (red box): the fitness landscapes of highly conserved positions are strongly correlated. At the same time, the wide range of correlations before the ‘cliff’ of 40% conservation suggests that sequence conservation is not the only determinant of fitness. (**b**) Mutations that ‘transform’ one IGPS orthologue into another primarily have neutral or beneficial effects (pink histogram) compared to all mutations (grey). Epistatic interactions are responsible for the existence of transformative detrimental mutations (minor peak on the pink histogram around *s*=−1). The distribution of selection coefficients for the transformative mutations is significantly different that than of the null distribution (***P*<0.02).

Correlation of fitness landscapes and epistasis

The correlations between fitness landscapes of orthologous proteins are intricately connected to the epistatic interactions in these proteins. If epistasis was absent, we would expect a near-perfect correlation between landscapes, as corresponding sites would have the same amino-acid preferences. Conversely, if epistasis was very strong, the fitness effects of every mutation would be entirely determined by the protein as a whole, and one would expect the fitness landscapes to become uncorrelated. Loosely, correlation of fitness landscapes is inversely proportional to the degree of epistasis. To unravel potential epistatic interactions in IGPS, for each orthologue, we selected ‘transformative’ mutations that correspond to WT residues in another orthologue, ‘transforming’ one protein into another. The distribution of selection coefficients of transformative mutations compared to the distributions of all selection coefficients in the three orthologues is significantly different (Fig. 2.5b). Transformative mutations are significantly enriched in neutral and beneficial mutations; deleterious mutations are depleted. As widely expected, an amino acid that has naturally evolved in one of the orthologues is often tolerated in another orthologue. However, the existence of detrimental transformative mutations suggests epistatic interactions, as their fitness is strongly affected by interactions with the rest of the protein. We found that about

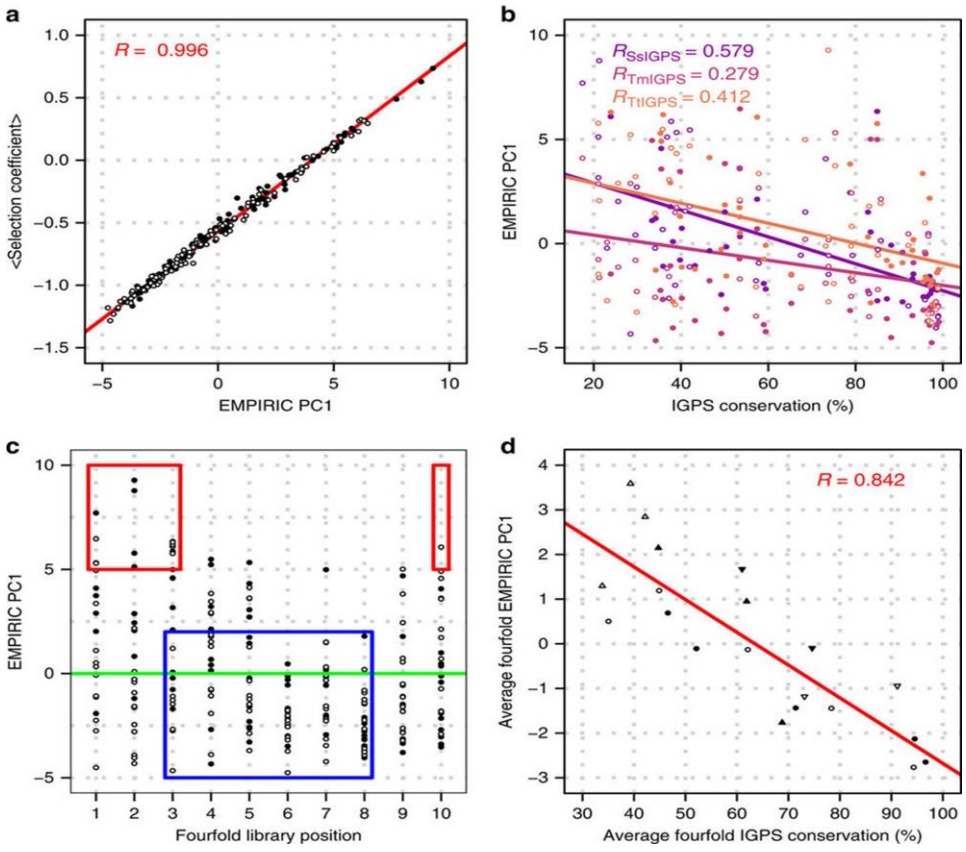
31% of the transformative mutations were detrimental ($s<-0.5$) in IGPS. For comparison, a previous study with isopropymalate dehydrogenase from *E. coli* and *P. aeruginosa* found that 38% of 168 transformative mutations were detrimental *in vitro*[78].

Sources of variance in experimental fitness landscapes of IGPS

To determine if the major sources of fitness variance are equivalent across the three orthologues, we applied PCA to the 80 × 20 matrices of selection coefficients of all mutations at each site surveyed, Supplementary Fig. 2.6. For each orthologue, the first two principal components accounted for almost 70% of observed fitness effects. The first principal component (PC1) explains ∼51% of the fitness variance and, by construction, is proportional to the average fitness of the position (Fig. 2.6a). We did not detect a significant linear relationship of PCA components to features such as ASA, RSA, b-factor, and WT amino-acid sequence conservation, if all three orthologues are considered together (Fig. 2.6b). Still, a periodicity observed in the PC1 scores along the library positions prompted a closer look at the relationship between PC1 and the four-fold symmetric structure (Fig. 2.6c). This scatterplot of PC1 showed differences in the deviation of the principal component scores along the coarse-grained ten library positions. The highest scores, representing the highest average fitness, are found mainly at the first three

positions of the odd strand libraries and the last position of the even strand libraries, associated with the αβ-loops and βα-loops, respectively. The lowest scores, representing the lowest average fitness, were found at positions six and eight, associated with the β-strands whose residues point into the β-barrel. The average four-fold PC1 varied linearly with average four-fold IGPS conservation (R=0.84) (Fig. 2.6d). Thus, fitness must relate to structural conservation of the four-fold symmetry. The second principal component (PC2) explains ~17% of the variance in our fitness data set and is correlated best with the hydrophobicity of the WT residues ($R_{\text{EMPIRIC PC2 to Hydrophobicity}}$~0.67). Hydrophobicity of the WT residues is a major factor affecting the fitness outcome, independent of the residue position in the structure.

Figure 2.6 The first principal component of fitness landscape is related to average four-fold conservation



(**a**) By construction, the first principal component (PC1) of the EMPIRIC PCA is linearly related to average fitness (*R*=0.996, 'open circle' odd numbered libraries, 'filled circle' even numbered libraries). (**b**) A linear relationship between EMPIRIC PC1 and IGPS conservation of varying strength was observed for each orthologue. A linear correlation of *R*=0.408 was observed if all three orthologues were considered together. (**c**) The values of EMPIRIC PC1 vary with four-fold aligned positions, implicating structure in fitness determination. The canonical secondary structures are indicated above the plot. The green line indicates the average score. Highest scores were observed at both ends of the library positions, associated with the αβ and βα-loops (red boxes). Lowest average scores were observed in the intermediate positions, associated with the β-strands (blue boxes). (**d**) Average EMPIRIC PC1 scores based on the four-fold alignment correlate linearly with average four-fold IGPS conservation (*R*=0.842).

Experimental data versus evolved IGPS and TIM sequences

To separate the factors that are important for function from those important for stability in our fitness assay, we performed separate PCA analyses of representative sequences of IGPS and TIM barrel proteins, using amino-acid frequencies at each site as a proxy for their evolutionary fitness (see Methods). Since all IGPS enzymes are TIM barrel proteins, we expected that the major drivers of sequence variance will reflect stability and structure for the TIM PCA and stability, structure, and function for the IGPS PCA. By selecting the 80 aligned positions examined in our fitness assay, we directly compare and contrast the experimental EMPIRIC IGPS fitness landscape to the naturally evolved amino-acid preferences in IGPS and TIM barrels.

The first principal component for both the IGPS PCA and TIM PCAs accounted for ~30% of the variance found in their amino-acid preferences and was correlated with WT amino-acid hydrophobicity ($R_{\text{IGPS PC1 to hydrophobicity}}$~0.70). The two components correlated with each other ($R_{\text{IGPS PC1 to TIM PC1}}$~0.71) and with the second component of the EMPIRIC PCA ($R_{\text{EMPIRIC PC2 to IGPS PC1}}$~0.69, $R_{\text{EMPIRIC PC2 to TIM PC1}}$~0.59), which correlated with hydrophobicity. Thus, stability appears to be the main driver for the residue selection within IGPS and TIM barrels, in general, for the segments examined.

The second principal component accounts for ~20% of the variance observed in IGPS and TIM barrels and is collinear with information content of the sequence alignment, indicating conservation. Similar to PC1, the two PC2 of the

IGPS PCA and TIM PCA were correlated ($R_{\text{IGPS PC2 to TIM PC2}}$∼0.68). At the same time, no correlation was observed between TIM conservation and EMPIRIC fitness in the IGPS orthologues. Presumably, conservation in a sample of 71 representative TIM sequences only reflects very general fold patterns, and has little predictive value for the fitness of mutants in the specific family such as IGPS.

Statistical coupling analysis

While active site residues are highly conserved to preserve enzymatic function, other sites indirectly supporting enzyme activity or those supporting protein stability can also show position-specific sequence constraints and correlated substitutions between positions within a protein family[94–96]. Statistical coupling analysis (SCA) was used to characterize functionally important co-evolving residues in SsIGPS that may inform our fitness results[94]. Two significant sectors were identified. Sector one (46 positions) involved mainly amino acids at the interface between the β-barrel and α-helical shell (Supplementary Fig. 2.7a). The distribution of fitness effects for the 17 positions with known fitness was unimodally deleterious. Over 37% of the residues were branched aliphatic residues and 28% were charged residues, highlighting the importance of hydrophobic and electrostatic interactions in stabilizing the β-barrel/helical shell interface throughout evolution. In contrast, sector two (45 positions) described both stability and functional properties. The distribution of fitness effects for the 25 positions with known selection coefficients was bimodal, s_{modes}∼−0.9 and −0.3,

reflecting highly conserved sites that are intolerant of mutations and sites of high adaptation potential. Active site residues, including the substrate-binding site, made up ~20% of the sector (Supplementary Fig. 2.7b). Of particular interest, two non-active site positions found to improve SsIGPS fitness with mutation were identified in sector two, residues I45 and D128. Corresponding mutations of the β3α3 hairpin clamp in the other two orthologues yielded similar fitness gains. Serving as a putative conduit between the active site and these two residues are select α-helical and β-strand residues in the SCA that span the two ends of the protein (Supplementary Fig. 2.8). The distributed nature of covariation, or SCA sectors, over the IGPS structure is consistent with statistical observations of long-range effects of mutations[97].

Discussion

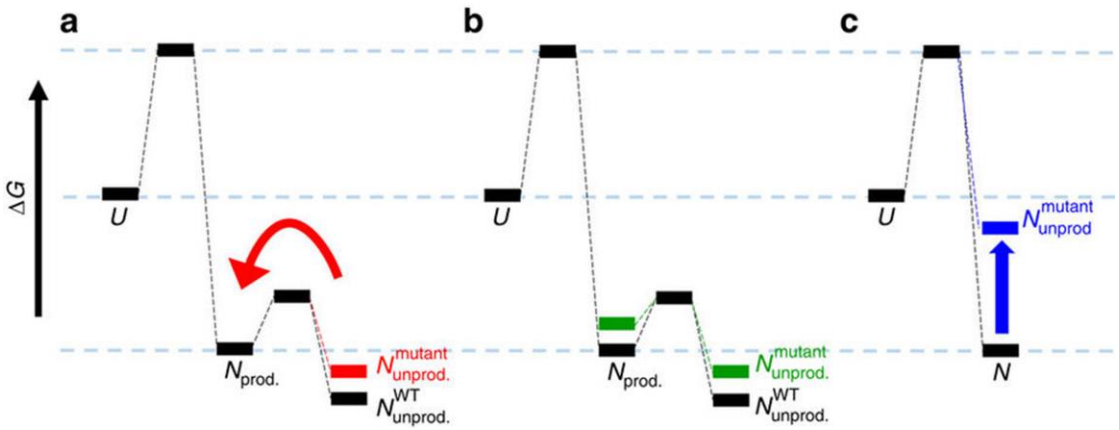
Systematic exploration of fitness landscapes has become possible using EMPIRIC and other high throughout mutagenesis approaches[66–71,80,98]. We have explored the fitness landscapes of three orthologous TIM barrel proteins from archaea and bacteria to answer two important questions at the intersection of biophysics and evolutionary biology: (1) what are the salient sequence and structural correlates of a TIM barrel fitness landscape, and (2) are the fitness landscapes of phylogenetically divergent orthologous proteins correlated with each other and, if so, what is the basis of the correlation?

Protein fitness landscapes depend on stability and enzymatic function, which are both affected by sequence and structure[30,70,84]. Thermophilic proteins

employ multiple mechanisms to maintain their structure at high temperatures, including a higher content of salt bridges and H-bonds relative to their mesophilic counterparts[99,100]. Increased stability is certainly required at high temperature, but it can be a liability at lower temperatures[101]. The native basin of proteins and enzymes contains a dynamic ensemble of rapidly interconverting states, some of which may be more favorable for particular functions, for example, substrate binding or the catalysis of chemical reactions[102]. For IGPS, ring closure of the substrate, 1-(o-carboxyphenylamino)-1-deoxyribulose 5-phosphate (CdRP), requires a minimal distance and favorable orientation between C1 and C2' for the reaction to proceed[91,92]. This conformation is mediated by electrostatic interactions between the substrate and the active site lysine residues[92]. At lower temperatures, the conserved lysine residues show decreased flexibility and restricted structural orientation leading to a greater population of CdRP in an extended, unproductive conformation. At higher temperatures, electrostatic interactions between IGPS and CdRP favor a reactive substrate conformation[92]. Access to these productive higher energy states in the native basin would increase if a mutation leads to a destabilized unproductive native state conformation without a concomitant destabilization of the productive higher energy state (Fig. 2.7a,b). We hypothesize that fitness gains are driven by increased local flexibility and/or dynamics accompanied by a population shift toward the productive catalytic conformation(s). Conversely, deleterious effects may be caused by a decrease in stability and the concomitant decrease in the

concentration of the enzyme or by distortion of the active site and ensuing reduction of catalytic properties (Fig. 2.7c). In fact, largely aliphatic residues in the barrel interior (Supplementary Table 2.4) are essential for structural integrity and incur a high proportion of deleterious mutations (Fig. 2.3). Native state hydrogen exchange protection patterns in the HisF TIM barrel from *T. maritima* revealed that layers 2 and 3 were strongly protected from solvent exchange, implying a major role in stabilizing the native state[12]. The side chains in these two layers participate in large clusters of densely packed ILV residues, known to form cores of stability in globular proteins[13]. We argue that the core of the barrel is critical for the stability of both the unproductive and productive conformations of TIM barrels.

Figure 2.7 Putative effects of mutations on the energy landscape of IGP



Free energy diagrams showing three possible scenarios for the effect of a single mutation on the folding free energy surface ($\Delta G°$, Gibbs free energy of folding; *U*, unfolded state; *N*, native state ensemble). An ensemble of rapidly interconverting productive ($N_{prod.}$) and nonproductive ($N_{unprod.}$) conformations resides in the native basin. At mesophilic temperatures, thermophilic IGPS access the productive conformation to a lesser extent than at their native thermophilic temperatures. (**a**) Beneficial mutations may destabilize the unproductive native state without a concomitant destabilization of the higher energy, productive conformation, resulting in a shift in the population from the unproductive to the productive conformation. Increased activity improves fitness. (**b**) WT-like mutations may destabilize both the unproductive and productive states, resulting in no net change in the population. No change in fitness is observed. (**c**) Deleterious mutations may greatly destabilize the native state, resulting in a population shift to inactive, partially or fully unfolded, states that are susceptible to proteolysis. Poor fitness is associated with loss of enzymatic capabilities.

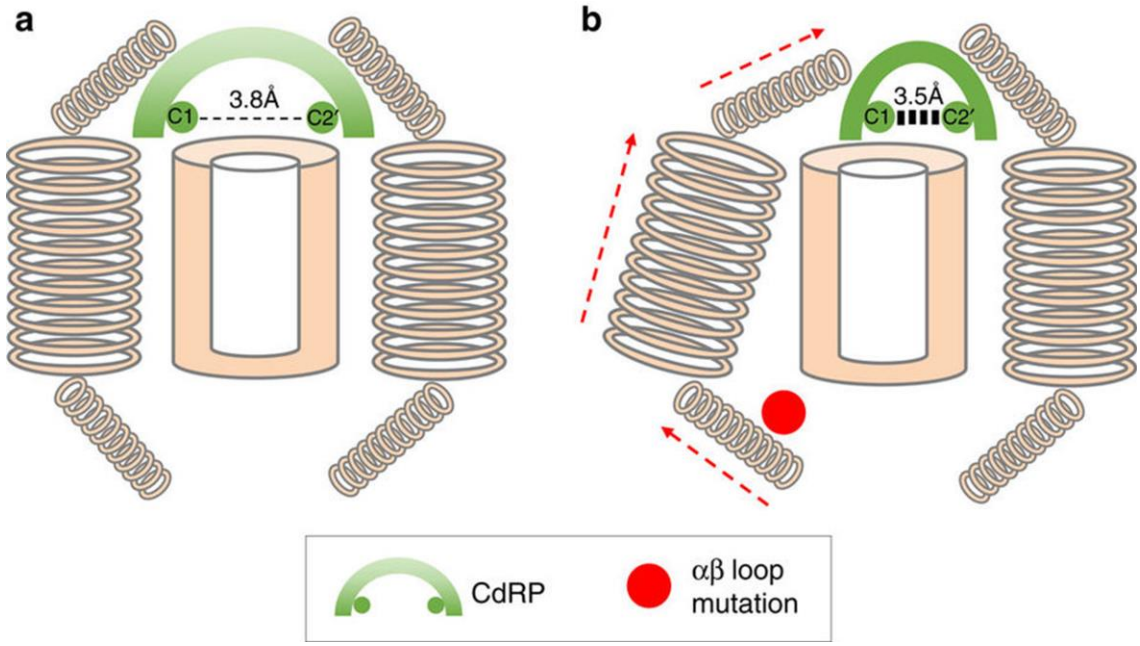
Increased activity of mutant thermophilic proteins, including IGPS, at a lower temperature has been previously reported, and typically involves modifications to the active site or substrate-binding site[103]. In contrast, many positions where mutations increased the fitness in our assay were distal to the active site in the βα-loops. For example, deletion of the canonical β3α3-hairpin clamp, such as SsIGPS I107 and D128, and several other hairpin clamps within the four-fold symmetrical βαβ modules consistently shortened doubling time relative to the WT. Similarly, for the hydrophobic stack between SsIGPS F40, I45 and V78, all mutations at I45, except to stop codons, displayed increased fitness. These long-range effects on fitness revealed an unexpected allostery between the αβ-loops at one end of the TIM barrel and the active site at the opposite end. The biological relevance of this allostery is supported by known cross-activating communication between HisF synthase TIM barrel and HisH glutaminase in the histidine biosynthetic pathway[104]. Activities of these two proteins are coupled through a physical interaction between the two enzymes, where the αβ-loops of the HisF TIM barrel dock onto the oxyanion hole of the HisH glutaminase active site[104]. Signal transduction between the two active sites spans the entire length of the TIM barrel and is mediated by correlated motions of several networks of residues within the β-barrel and α-helices[105].

Although the IGPS fitness landscape observed in our experiments in yeast may differ from that experienced during evolution of these thermophilic proteins, we believe that the basic biophysical and structural constraints we observed hold

true in the natural environment. Moreover, thermophilic orthologues functioning at mesophilic temperatures provide targets of opportunity to detect and map pathways of allostery leading to beneficial fitness under stress conditions. Bernhardt speculated long ago that consecutive enzymes in metabolic pathways might preferentially associate with each other to enhance the throughput of substrates to products[106]. While the assembly status of the three IGPS in our study is monomeric, various combinations of bifunctional and multifunctional enzymes from the tryptophan biosynthetic pathway occur in nature[107]. Thus, the inherent ability of a TIM barrel to stimulate or enhance the activity of another protein in an enzymatic pathway similar to the HisF/HisH pair through allosteric interactions would be advantageous.

Our fitness data for IGPS suggest that mutations in the αβ-loops induce changes in the distal active site βα-loops to favor the catalytically active conformation (Fig. 2.8). They may do so by mimicking the physiological stimulation by partner proteins or by inducing the conformation naturally preferred at their respective optimal growth temperatures. The greater fitness observed for mutations in the outward facing side chain positions in the β-barrel (Fig. 2.3) compared to mutations with inward facing side chains, implicates perturbations in the α-helical shell as the conduit for the allostery. This conjecture is supported by the SCA results that point to a possible pathway of signal transduction between the active site and the αβ-loops in SsIGPS, mediated by intervening helical elements (Supplementary Fig. 2.8).

Figure 2.8 The active site of IGPS coordinates the ring closure event converting substrate CdRP to product IGP



The active site orients the C1 and C2' atoms of the substrate to a specific distance and geometry, favoring conversion of CdRP to IGP. Correlated motions of grouped amino acids dispersed throughout the protein influence conformation of the active site. This access to the productive enzyme conformation dictates enzyme activity. Under mesophilic conditions, thermophilic IGPS enzymes are less dynamic, less flexible and less active than at their native thermophilic temperatures. (**a**) During yeast growth at 30 °C, the active site of thermophilic IGPS favors an unproductive conformation, where CdRP is oriented in an extended, non-reactive conformation. Reduced activity results in reduced fitness of the WT thermophilic IGPS compared to the WT mesophilic IGPS. (**b**) An allosteric beneficial effect on fitness was observed with certain mutations of the αβ-loop, resulting in greater fitness of the mutants compared to the thermophilic WT IGPS proteins at mesophilic temperatures. These mutations transmit their allostery via correlated protein breathing motions to favor the productive conformation of the active site, where CdRP is oriented optimally in a reactive conformation. Improved fitness observed for β-strand residues whose side chain point to the β-strand/α-helices interface implicates the α-helices (Fig. 2.3) as the conduit for the allostery between the active site βα-loops and the αβ-loop.

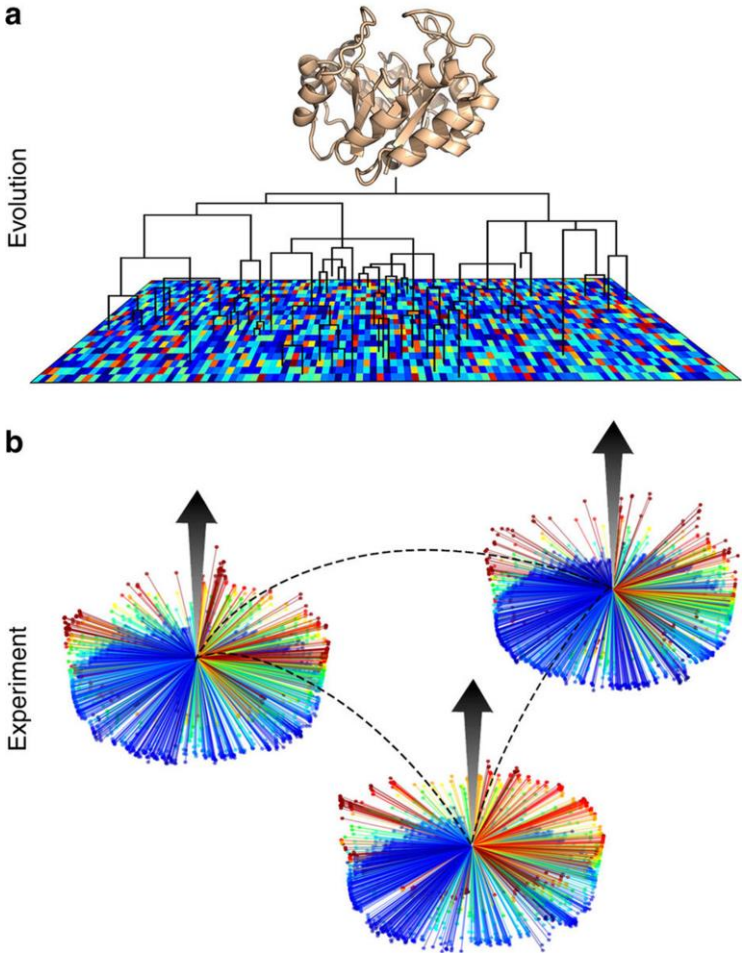
Experimental mapping of fitness landscapes has progressed from sampling of some or all combinations of diallelic loci[108,109] towards large-scale mutational scans[66–71,80,98]. Although there is evidence about the roughness of fitness landscapes for small deviations from the WT sequence[63,110,111], our finding of correlations between the fitness landscapes of divergent IGPS orthologues suggests that the common fold and function act together to smooth the landscape. In fact, low throughput experiments have shown that the effects of mutations on biophysical properties such as Tm and $\Delta\Delta G°$ are largely conserved across homologues of influenza nucleoprotein[112], and modern versus ancestral thioredoxins and β-lactamases[113]. These studies suggest that amino-acid preferences at a given position in the structure are mostly conserved during evolution, contrary to a modelling prediction of Pollock *et al*[114]. However, these works focused on specific biophysical properties of the proteins, rather their full functional consequences. This limitation was lifted in a deep mutational scan of two influenza virus nucleoproteins (94% sequence identity) in their full physiological context[80]. Again, the amino-acid preferences were found to be conserved at most sites, suggesting strongly correlated fitness landscapes of aligned positions. In agreement, we found significantly higher correlation of fitness landscapes of aligned positions even when the WT amino acids differed between orthologues pairs (Fig. 2.4c,d).

Beyond constraints imposed by the structure, ancestral sequence reconstruction of ribonuclease H1 suggests that various molecular mechanisms

are employed to stabilize proteins during evolution in response to environmental stress[115]. Diverse distributions of ionic, H-bond, and hydrophobic interactions are observed in our three orthologues[54,55,116], indicating that each protein sampled different sequence space and evolutionary paths as part of its 'thermodynamic system drift'[115]. At the same time, high correlation of fitness landscapes between positions with identical WT amino acids irrespective of structural alignment (Fig. 2.4a) highlight the biochemical and physical constraints required for properly maintaining the protein fold.

In summary, we found a very strong correlation of organism-level fitness landscapes of three extant proteins at 30–40% sequence identity, and of bacterial versus archaeal origin. Two related interpretations of this result appear possible. First, conservation of amino-acid preferences does persist across the two phylogenic domains and low sequence identity. Second, in a complementary way, we interpret this conservation as translocation of fitness landscapes in sequence space: fitness landscapes of single point-mutants can be successfully translocated to a different starting point in sequence space. We propose to visualize translocation by presenting the fitness landscapes of single point-mutants as pinwheels in sequence space (Fig. 2.9). All mutants are one change away from the WT, with the color and height of each mutant representing its fitness. In this analogy, translocation means that pinwheels for orthologues centered far apart in sequence space still maintain a similar shape and color profile.

Figure 2.9 Translocation of fitness landscapes in the sequence space of orthologous TIM barrels



a) Over evolutionary time, orthologous proteins adapt to changing environment for optimal fitness. Steps in sequence space select for protein stability and activity under their native conditions, all while retaining the TIM barrel fold and IGPS function. (**b**) Experimentally derived fitness landscapes mapped from point mutations represent single steps from WT sequence. Despite significant divergence of WT in sequence space, the fitness landscapes of IGPS orthologues remain correlated (dashed lines). Rather than traditional two-dimensional heatmaps, fitness values are displayed on a three-dimensional pinwheel, highlighting the wide range of possible fitness effects of a single sequence step. The profiles of the pinwheels are similar, indicating the correlation of fitness landscapes, even if WT sequences (centers of the wheels) are only ∼40% identical and widely separated. PCA demonstrates a correlation between experimental fitness landscapes and amino-acid preferences in evolved sequences.

Understandably, deviations from the perfect correlation between the landscapes of orthologues result from epistatic effects within the protein. As mentioned previously, the magnitude of epistasis is inversely related with the correlation between the fitness landscapes. For proteins of relatively small divergence, biophysical models achieve a significant power to predict the effects of mutations[117,118], especially if informed by high-throughput mutational scan data[67,81]. Our findings suggest that such approaches may be extended to proteins with a greater degree of sequence divergence. Further development of the models and metrics for comparing the fitness landscapes[119,120] will produce appropriate tools for quantifying the landscapes at various degrees of divergence between sequences.

Translocation of fitness landscape is facilitated by plasticity of the TIM barrel fold, both in natural sequences and *in vitro*, as stable TIM barrels can be created by fusions of natural half barrels, $(\beta\alpha)_4$ (refs [121,122]). Indeed, it is well established that the *inverse* protein folding problem, designing sequences for a given template structure, has multiple solutions[34]. Recently, the first successful *de novo* TIM barrel design employed a completely symmetric four-fold repeat[89,123]. Observations of the four-fold symmetry in both our experimental fitness landscape and *in silico* design serve to validate both approaches in understanding the fundamental properties of TIM protein architecture. From an experimental perspective, fitness studies on other common proteins platforms

have the potential to reveal unanticipated sequence-structure-fitness relationships and provide new strategies for *de novo* protein design[89,123].

Methods

Strains and culture conditions

S. cerevisiae strain BY4742 Δ*IGPS::KanMX* was produced using the same PCR-generated deletion strategy described by the Saccharomyces Genome Deletion Project[124]. The last 810 bp of the TRP3 gene encoding IGPS were replaced with the KanMX gene. Deletion of IGPS with the KanMX gene was confirmed by Sanger sequencing.

The pRS416 vector carrying the auxotrophic URA3 marker was a gift from Daniel Bolon's lab. A lower expressing Tma19 promoter, also provided by the Bolon laboratory, was used to increase the sensitivity of the fitness assay. Three silent mutations were introduced into the plasmid at the URA3 marker, the ampicillin resistance marker, and the Tma19 promoter to disrupt BSAI recognition sites. The BSAI enzyme was used to create the saturating mutagenesis libraries.

IGPS genes were purchased from Genscript. An N-terminal 6 × His tag and Tev protease recognition site were added to each construct in anticipation of future studies requiring protein abundance measurements and protein purification. In anticipation for *in vitro* folding studies, the non-canonical N-terminal α00 of each gene was deleted to reduce aggregation during refolding of purified proteins (SsIGPS Δ1–26, TmIGPS Δ1–31, TtIGPS Δ1–34). To prevent

non-specific cleavage when using Tev protease, position 18 was mutated from arginine to serine in SsIGPS. To prevent disulfide bonds and oxidation by molecular oxygen, position 102 was mutated from a cysteine to a serine in TmIGPS. IGPS genes were cloned into the pRS416 vector using restriction sites, SpeI and BamHI.

For each of the orthologues, we created 8 plasmid libraries where a 10 amino-acid region was mutagenized using the EMPIRIC method, see Hietpas *et al*[86] for a detailed protocol. Briefly, inverted BsaI restriction sites are introduced by PCR within the region of interest. BsaI digestion is followed by directional sticky-end ligation of oligonucleotide cassettes containing a single randomized site containing all 64 codons. Thus, each library contained 640 sequence variants corresponding to all possible amino-acid (and codon) mutations at each position. The sequence variance was deterministic, as no error-prone polymerases were used.

Yeast transformation was performed using the LiAc/SS carrier DNA/PEG method as described by Gietz and Schiestl[125]. Yeast cells were grown in rich media with G418 to select for IGPS knockout yeast. Transformed cells were selected on synthetic minimal media lacking uracil. Selection for IGPS activity was achieved through growth of transformed yeast (one plasmid library per culture) in synthetic drop-out medium lacking tryptophan. All growth experiments were performed at 30 °C. Liquid cultures were maintained in log phase

throughout the fitness assay by periodic dilution. G418 selection was maintained throughout the growth. The oligonucleotide sequences for mutant library generation and primer sequences used for creating the plasmid libraries and processing the deep-sequencing samples are available at https://github.com/yvehchan/TIM_EMPIRIC.

EMPIRIC data processing

Illumina deep sequencing (36 base single reads) was performed by Elim Biopharmaceuticals. In-house analysis of the deep-sequencing results was performed using custom software. Reads were stringently filtered based on several criteria: Phred score>20 across all 36 bases (error probability of 0.01), valid barcode match, single reference sequence match for anticipated single codon mutations, and absence of MmeI recognition site. MmeI enzyme was used to create overhangs for ligation of time-stamping barcodes. Sequence counts were tracked on the nucleotide and amino-acid level. A total of 5,040 (10 AA/library × 8 libraries/orthologue × 3 orthologues × 21 mutation types) amino-acid mutations were created. Raw fitness was calculated as the slope of the $\log_2$ relative abundance of the mutant to WT versus time over 6 to 10 time points spread over 4 doubling periods of the yeast,

$$w_i = \frac{\mathrm{d}}{\mathrm{d}t}\log_2\left(\frac{N_i(t)}{N_i^{\mathrm{WT}}(t)}\right), \qquad (1)$$

where $N_i(t)$ is the abundance (count) of mutant i at time t, and $N_i^{\mathrm{WT}}(t)$ is the count of WT amino acid at the corresponding position. Slope was determined

using the linear regression, Supplementary Fig. 2.2. From the raw fitness *w*, selection coefficients were determined according to

$$s = -\frac{w}{w_{\mathrm{STOP}}}, \qquad (2)$$

where w_{STOP} is the average raw fitness of all mutations to a stop codon within a 10 amino-acid region, corresponding to an individual selection experiment, $w_{\mathrm{STOP}} = \frac{1}{10}\sum_{i=1}^{10} w_{i,\mathrm{STOP}}$. This normalization ensures that on average, stop codons have a selection coefficient of −1. Fitness calculations were performed on all values except 79 mutations leading to MmeI recognition sites. Seven mutations had poor coverage in our mutagenesis libraries and were also removed from analysis: SsIGPS A174W, TmIGPS I131M, TmIGPS I178C, TmIGPS I178W, TtIGPS G78M, TtIGPS L153C and TtIGPS L153Y. After filtering, a total of 4,954 mutations were analysed.

To assess the reproducibility of our fitness results, we compared selection coefficients of two full biological replicates of regions β3, β4 of SsIGPS comprising 399 mutations (R=0.947, Supplementary Fig. 2.9a).

Sequence and structural analyses

Pairwise sequence alignments of the three orthologues were performed using Clustal Omega provided by EMBL-EBI[126]. Sequence identities and similarities were calculated using Ident and Sim Program in the Sequence Manipulation Suite provided by bioinformatics.org. PDB accession codes used

for structural analyses were 2C3Z for SsIGPS[116], 1I4N for TmIGPS[54] and 1VC4 for TtIGPS[55]. RMSDs of structural alignments were performed using SPalign[127]. Structural figures were generated using PyMOL Molecular Graphics System v1.8.2.0. Accessible surface area (ASA) was calculated using EBI PISA tool. RSA was calculated from ASA using an empirical scale[88].

Comparison of fitness distributions

For structural analyses, specific groups of residues were compared to identify differences in their distributions of fitness effects. The fraction of beneficial mutations was used to describe the shape of fitness distribution. A permutation test was used to determine if two fitness distributions differed significantly. For each subset, the original fitness measurements for 19 amino-acid substitutions at a given position were reassigned to a random position and the corresponding fraction of beneficial mutations was determined. Statistical significance (*P* value) of the observed fraction of beneficial mutations was calculated from a distribution (*N*=10,000) of the randomized values. Plots were generated in R version 3.2.0.

Correlations of fitness landscapes

The Pearson correlation coefficient was used to assess the similarity in response to the set of 19 amino-acid substitutions for a given pair of residues. To assess the statistical significance, we compared the fitness landscapes of two full biological replicates of regions β3, β4 of SsIGPS comprising 20 positions (Supplementary Fig. 2.9b). The average correlation between fitness landscapes

of the replicates was R=0.89, much higher than all the correlations between the orthologues (Supplementary Fig. 2.10). The difference between the distributions of correlation coefficients of orthologues and that of the replicates was highly statistically significant (KS test, $P<10^{-4}$).

Principal component analysis

PCA was used to identify major sources of variance in our data set. PCA analyses were performed using custom scripts written in Python version 2.7.6. For the EMPIRIC PCA, fitness values were normalized for every position to have the same mean and s.d. PCA was performed separately for each of the orthologues. For the IGPS PCA, the multiple sequence alignment (MSA), 1,744 representative sequences was obtained from PFam (accession code PF00218). Two positions in the fifth library of SsIGPS had no structural match with positions of the TmIGPS and TtIGPS libraries. Therefore, three separate alignments were created using positions corresponding to all 80 library positions in each of the orthologues. The minimal pairwise overlap between the three alignments is >60 positions. Amino-acid frequencies corresponding to our library positions were extracted and the resulting three matrices of size 80 × 20 were log-transformed and normalized for PCA application. No significant differences in the IGPS-PCA results were observed between the three sequence alignments. For the TIM PCA, structural alignment of 71 non-redundant TIM-barrel proteins[128] was constructed using SPalign[127]. Pairwise structural alignment was performed for each of template structures SsIGPS, TmIGPS, and TtIGPS to the 71

representative TIM barrel structures. As with the IGPS-PCA, we extracted the corresponding 80 library positions from the pairwise alignment to obtain three multiple structure sequence alignments (MSSA) for each of the templates. No significant differences in the TIM-PCA results were observed between the three structural alignments.

Statistical coupling analysis

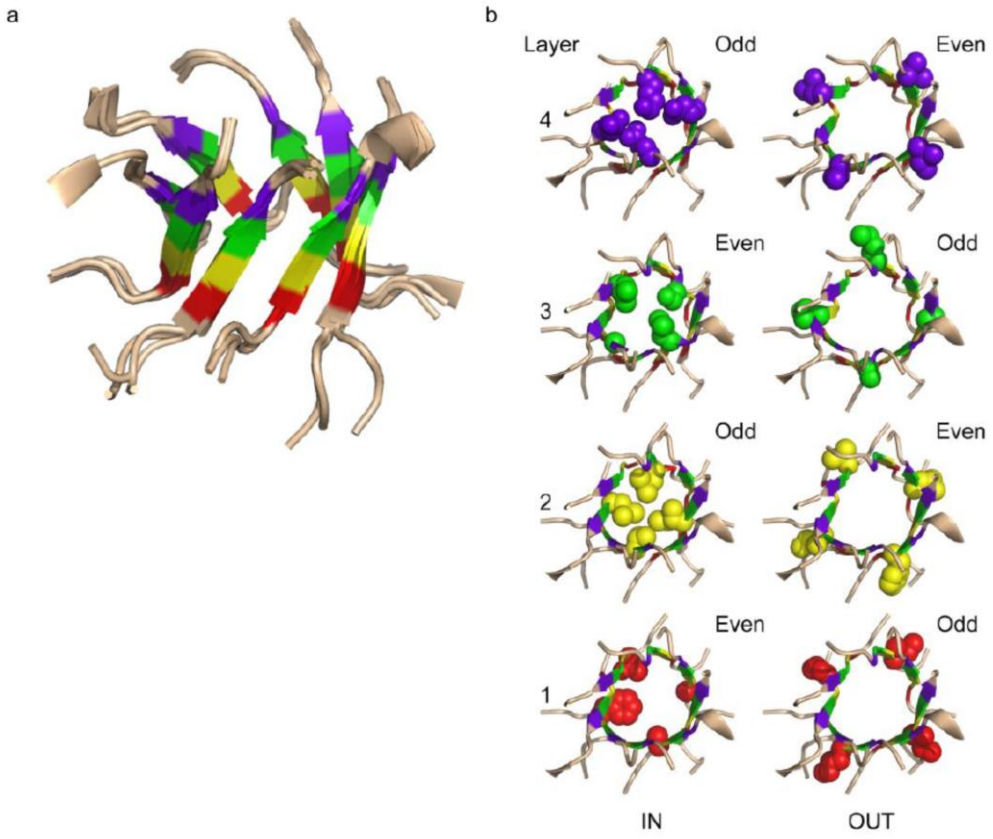
SCA was used to identify groups of co-evolving residues that have functional roles in protein activity and stability. A modified IGPS-MSA was generated for the SCA in order to minimize large gaps in the alignment. The full-length sequence of SsIGPS was used as a BLAST search seed in the NCBI non-redundant database. COBALT[129] was used to align 1,000 protein sequences with length between 200 and 300, and redundancy filter reduced number of sequence down to 537 (35 to 95% sequence identity). Aligned positions with >80% gaps were removed from the alignment. This modified alignment was passed into SCA v5.0 MATLAB toolbox[94] and analysed using the default parameters. Two eigenvalues from the SCA positional correlation matrix exceeded the significance threshold. Therefore, the expected number of sectors was set to k_{max}=2. Spectral analysis of SCA sequence correlation matrix revealed no phylogenetic and/or sequence sampling biases, permitting functional interpretation of the two identified protein sectors.

Data availability

All data, primer sequences, alignments, and scripts for data analyses are available at https://github.com/yvehchan/TIM_EMPIRIC or from the corresponding author upon request.

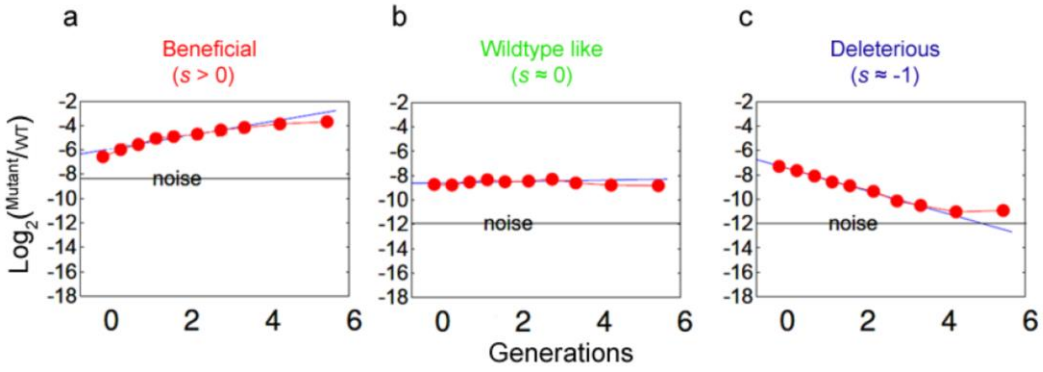
Supplementary information

Supplementary Figure 2.1 Canonical layers of the β-barrel stabilize the protein core and provide surface area for docking the α-helices



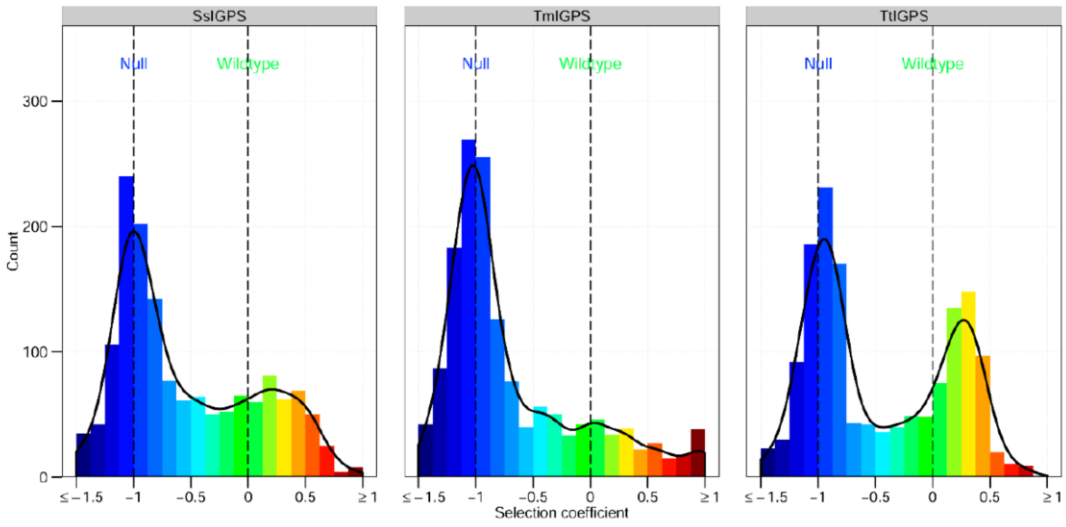
(a) Four layers of the β-strands forming the TIM barrel core are colored red, yellow, green, and purple, from the N-terminus to C-terminus. **(b)** Orientation of side chains alternate in and out of the β-barrel per strand and per layer to maximize packing and to reduce steric clashes. Alternating side chain orientation evenly distributes stabilizing hydrophobic interactions within the β-barrel core and between the β-strands and surrounding α-helices.

Supplementary Figure 2.2 Selection coefficient is determined by the slope of the relative abundance of mutant to WT IGPS over time



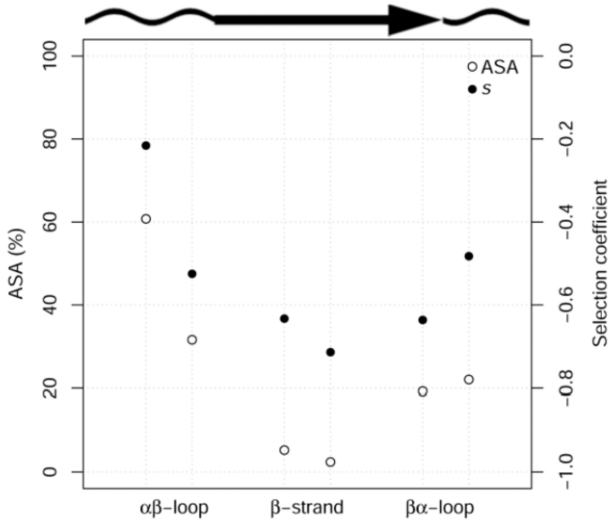
Representative examples of the three observed fitness phenotypes are displayed on a $\log_2$ scale to show doubling time. **(a)** Beneficial mutations have selection coefficient greater than 0, indicated by the positive slope of the relative abundance of mutant to WT over time. **(b)** WT-like mutations have selection coefficient approximately 0, indicated by the flat character of the relative abundance of mutant to WT over time. **(c)** Deleterious mutations have selection coefficient of approximately -1, indicated by the negative slope of the relative abundance of mutant to WT over time. The noise level line indicates the abundance of the mutant sequence obtained from deep sequencing of a WT sample.

Supplementary Figure 2.3 Distribution of fitness values for three orthologous IGPS proteins



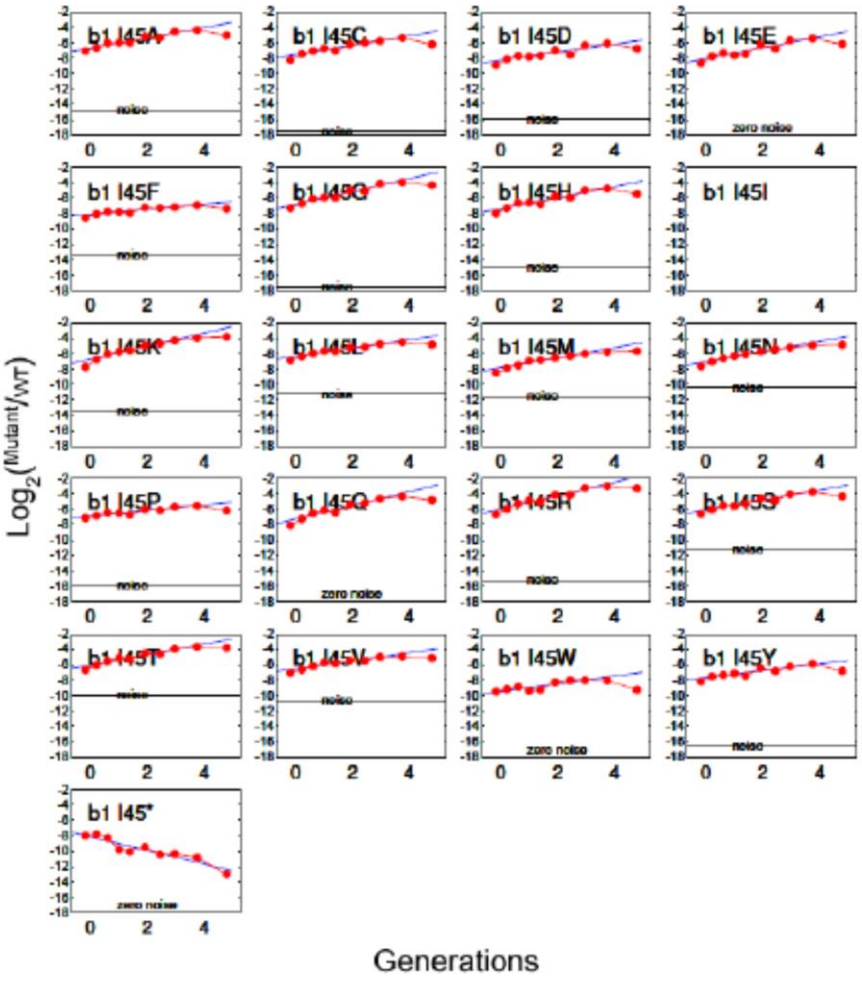
Distribution of fitness values for SsIGPS (left), TmIGPS (center), and TtIGPS (left) plotted on a histogram. Bimodal distributions were observed for all three orthologs, centered at s = -1 and centered above s = 0. Fitness values above 0 indicate a fitness gain due to mutation.

Supplementary Figure 2.4 Accessible surface area and fitness vary by secondary structure and strand parity



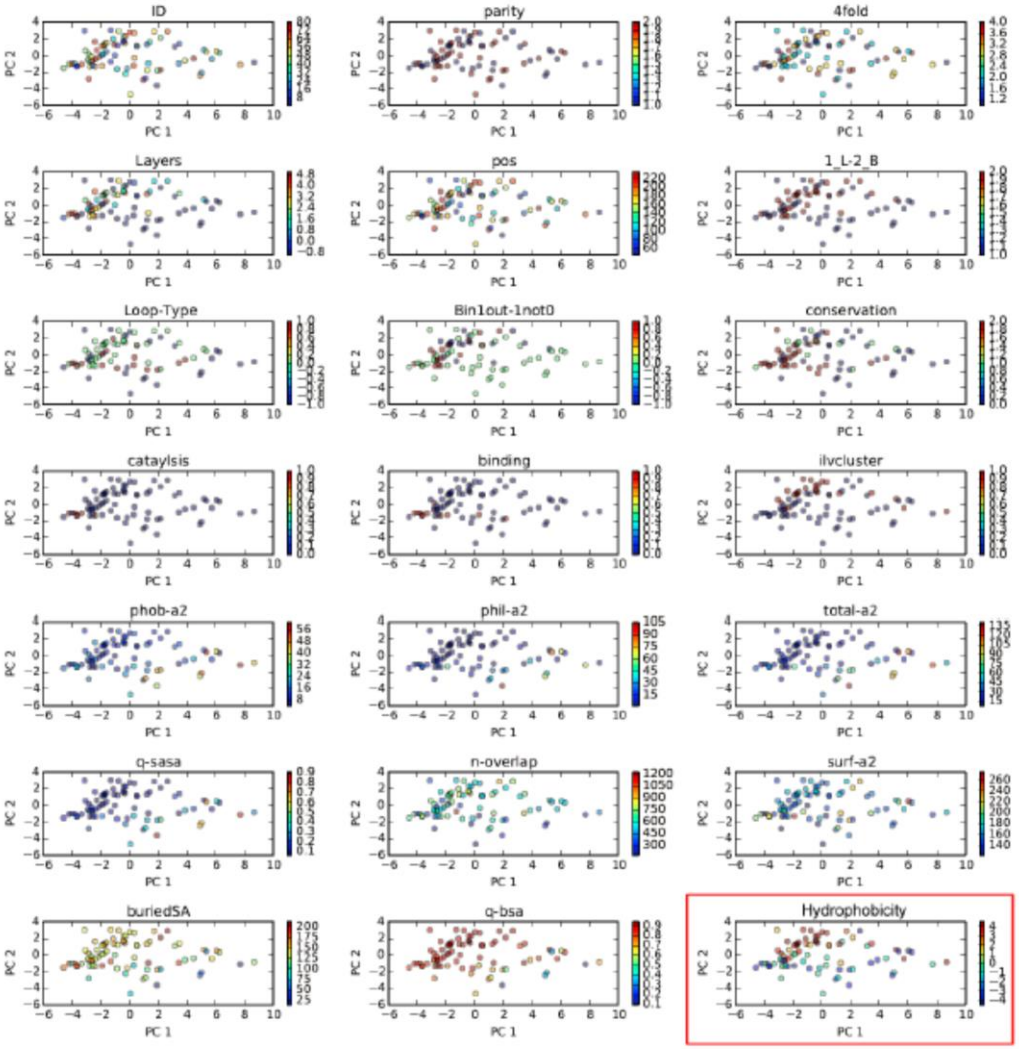
Average fitness varied depending on both secondary structure and strand parity (● Selection coefficient, right axis). Average fitness correlates to ASA of these stratified groups (○ ASA, left axis).

Supplementary Figure 2.5 The effect of mutations at SsIGPS I45 on fitness



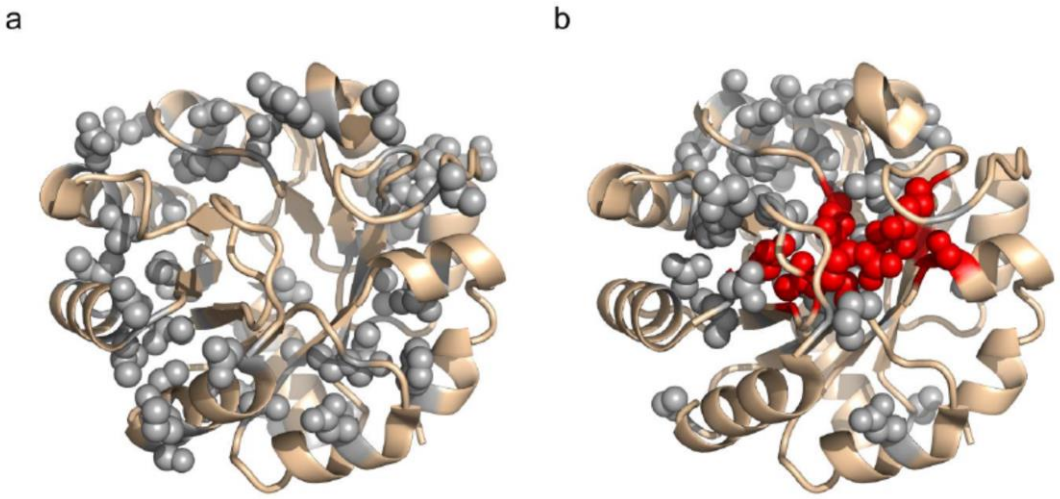
Mutations to SsIGPS I45 were uniformly beneficial, except for mutation to the stop codon. Tryptophan biosynthesis is a highly regulated process, where tryptophan accumulation leads to feedback inhibition at the first step of synthesis[51]. By the fourth generation, mutant and WT abundances were comparable, suggesting that a steady state of tryptophan concentration has been reached.

Supplementary Figure 2.6 Representative biplot of the secondary principal component (PC2) vs. the first principal component (PC1) of the PCA for SsIGPS



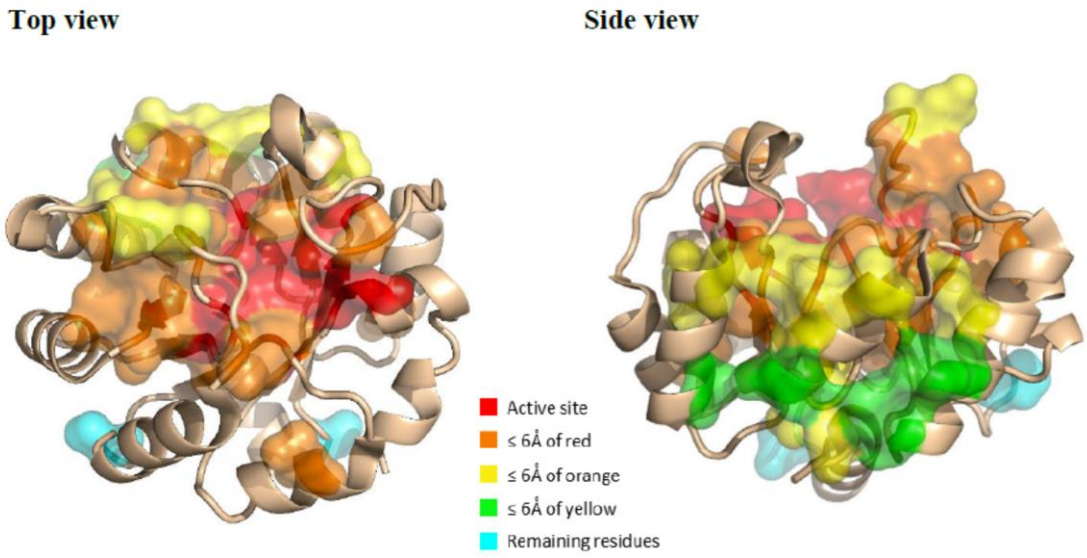
Several biochemical and structural features were examined to identify major sources of fitness variance in our EMPIRIC dataset. Biplots did not reveal a direct relationship between PC1 and the features examined. Hydrophobicity based on the Kyte-Doolittle scale was identified as the second largest factor influencing fitness, indicated by the monotonic color change along the PC2 axis (last plot highlighted by the red box).

Supplementary Figure 2.7 SCA sectors in IGPS TIM barrel proteins represented on SsIGPS



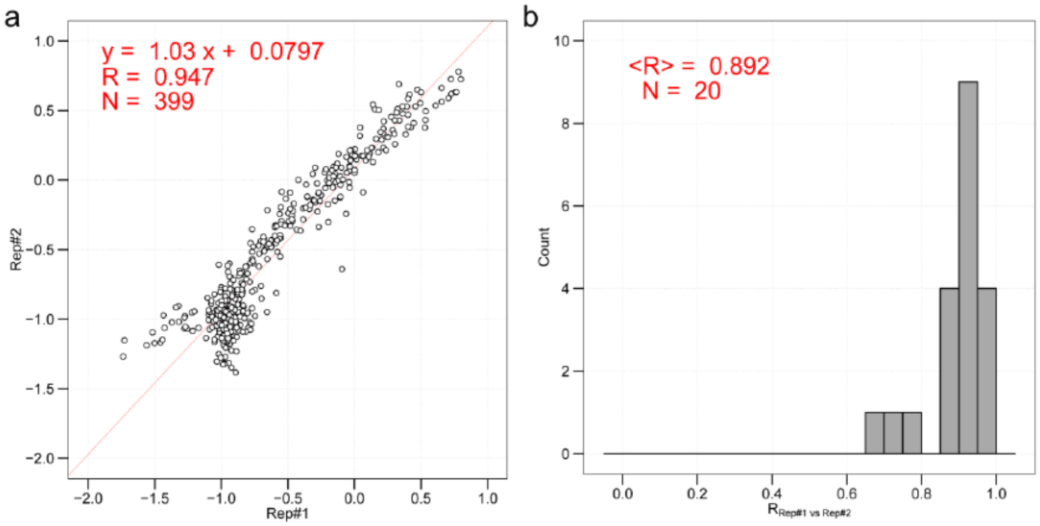
(a) Residues in sector one are highlighted with gray spheres. These residues are involved mainly in the β-strand/α-helical interface and α-helical/α-helical interface, required for stabilizing the tertiary structure. **(b)** Residues in sector two are highlighted with gray and red spheres. These residues are involved in protein function and stability. Residues highlighted in red are active site residues. All other residues in sector two are colored gray.

Supplementary Figure 2.8 Proposed conduit for allostery identified by SCA and fitness data



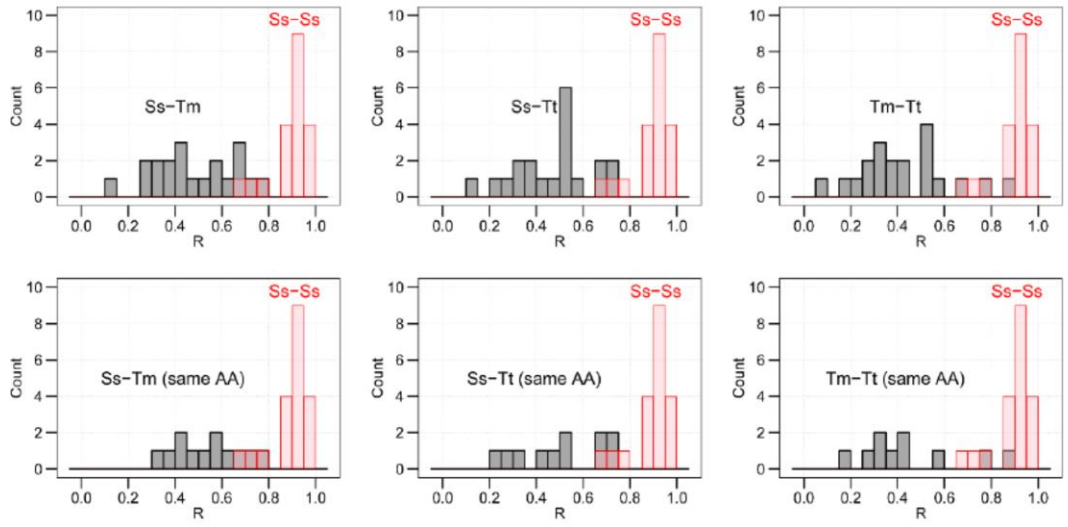
Sector two residues displayed by surface representation, (left) top view (right) side view, to show potential path for communication from active site to αβ-loops. Surface representation is color coded based on distance from active site residues in sector two. Active site residues are colored red. Residues within 6Å of the active site residues in sector two are colored orange. Sector 2 residues within 6Å of the orange residues are colored yellow. Sector two residues within 6Å of the yellow residues are colored green. All remaining residues in sector two are colored cyan.

Supplementary Figure 2.9 Reproducibility of EMPIRIC fitness results and correlation of fitness landscapes of biological replicates of β3 and β4 libraries



(a) Fitness results of full biological replicates of SsIGPS β3 and β4 are highly reproducible (*R* = 0.947). **(b)** Correlation of fitness landscape between biological replicates (20 residue positions) has a high mean *R* value.

Supplementary Figure 2.10 Distribution of Pearson correlation R between fitness landscapes of SsIGPS biological replicates and of orthologs



Pearson correlation *R* between the fitness landscapes of 20 residues in SsIGPS biological replicates (red) is much stronger than the correlations between the landscapes of the orthologs (gray, all positions or positions with matching WT amino acids, $P < 10^{-4}$).

Supplementary Table 2.1 Comparison of correlation distributions to null distribution

Pairwise comparison to Null			
Identical AA	**Count (n)**	**r_{mode}**	**p-value**
SslGPS vs TmlGPS	502	0.613	< 1E-4
SslGPS vs TtlGPS	504	0.708	< 1E-4
TmlGPS vs TtlGPS	520	0.555	< 1E-4
Null - all correlations	19200	-0.106	
Structurally aligned positions			
SslGPS vs TmlGPS	78	0.621	< 1E-4
SslGPS vs TtlGPS	78	0.724	< 1E-4
TmlGPS vs TtlGPS	80	0.617	< 1E-4
Null - all correlations	19200	-0.106	
Structurally aligned positions, different AA			
SslGPS vs TmlGPS	308	0.588	< 1E-4
SslGPS vs TtlGPS	308	0.676	< 1E-4
TmlGPS vs TtlGPS	314	0.589	< 1E-4
Null - all correlations	19200	-0.106	
Four-fold aligned positions			
SslGPS vs TmlGPS	35	0.537	< 1E-4
SslGPS vs TtlGPS	38	0.580	< 1E-4
TmlGPS vs TtlGPS	37	0.548	< 1E-4
Null - all correlations	19200	-0.106	

P-values obtained from performing a two-sided Kolmogorov-Smirnov test comparing the distribution of correlations within specific subsets of grouped positions to the null distribution of all possible pairwise correlations.

Supplementary Table 2.2 Comparison of correlation distributions between orthologs

Pairwise comparison between sets of orthologs	
Identical AA	**p-value**
Ss-Tm vs Ss-Tt	0.003
Ss-Tm vs Tm-Tt	0.235
Ss-Tt vs Tm-Tt	0.002
Structurally aligned positions	
Ss-Tm vs Ss-Tt	0.422
Ss-Tm vs Tm-Tt	0.311
Ss-Tt vs Tm-Tt	0.105
Structurally aligned positions, different AA	
Ss-Tm vs Ss-Tt	0.827
Ss-Tm vs Tm-Tt	0.238
Ss-Tt vs Tm-Tt	0.772
Four-fold aligned positions	
Ss-Tm vs Ss-Tt	0.108
Ss-Tm vs Tm-Tt	0.970
Ss-Tt vs Tm-Tt	0.126

P-values obtained from performing a two-sided Kolmogorov-Smirnov test comparing the distribution of correlations within specific subsets of grouped positions for one ortholog pair to another ortholog pair. Similar distributions are indicated by the high p-values.

Supplementary Table 2.3 Comparison of correlation distributions between subsets libraries of SsIGPS

Pairwise comparison to structure and sequence groups	
Identical AA vs structurally aligned positions	<1E-4
Identical AA vs structurally aligned positions, different AA	0.668
Identical AA vs four-fold aligned positions	<1E-4
Structurally aligned positions vs structurally aligned positions, different AA	<1E-4
Structurally aligned positions vs four-fold aligned positions	<1E-4
Structurally aligned positions, different AA vs four-fold aligned positions	<1E-4

P-values obtained from performing a two-sided Kolmogorov-Smirnov test comparing the distribution of correlations between specific subsets of grouped position.

Supplementary Table 2.4 Amino acid composition for the four canonical β-barrel layers represented in the three orthologous IGPS proteins

	Branched aliphatics	Small aliphatics	Aromatic	Acidic	Basic	Hydoxylic	Sulfur containing
In	20	11	1	9	3	4	0
Layer 1	2	8	1	0	0	1	0
Layer 2	12	0	0	0	0	0	0
Layer 3	6	3	0	0	0	3	0
Layer 4	0	0	0	9	3	0	0
Out	33	10	1	0	2	1	1
Layer 1	8	2	0	0	1	1	0
Layer 2	10	1	1	0	0	0	0
Layer 3	3	7	0	0	1	0	1
Layer 4	12	0	0	0	0	0	0

The β-barrel is composed largely of branched aliphatic residues followed by small aliphatic residues. The second layer pointing into the barrel and the fourth layer pointing out of the barrel are completely composed of branched aliphatic residues, whose hydrophobic interactions stabilize the protein core and active site, respectively. The fourth layer pointing into the β-barrel is composed entirely of charged residues, where long range electrostatic interactions orient and maintain the active site, supporting catalysis.

Chapter III – Molecular mechanism of beneficial allosteric mutations in SsIGPS

This chapter is a preliminary body of work that explores the phenotypic manifestation of protein sequence at the molecular and cellular level. Some experiments have only been carried out once and will need to be repeated. Dr. C. Robert Matthews and I designed the experiments. The experimental work was carried out by me as well as several summer interns who worked under my guidance: Lorein M. Rodriguez, Grace S. Ahn, Philip Economou, and Katherine Edwards. The substrate used for the functional assay was synthesized by Adam Choi, a Ph. D. candidate working under the guidance of Dr. Stephen Miller. The data interpretation is the work of myself and Dr. Matthews with contributions from Dr. Konstantin B. Zeldovich.

Introduction

Enzymes are cellular workhorses that catalyze chemical reactions required for metabolism, respiration, and other life processes. External stresses on the cell such as temperature or resource availability may require these proteins to function at an altered rate, to recruit new substrates, or to catalyze new biochemical reactions for survival. Some cells within a population will have enzymes with altered sequence capable of responding to the selective pressure based on biophysical properties encoded by their protein sequence. Neutral mutations that have accumulated through drift may become beneficial or deleterious. Responsive cells will be more fit compared to those that cannot, where fitness is generally described by reproductive success. Over time, the population of cells with proteins capable of conferring the desired function will increase under selective pressure for that particular phenotype and the allele becomes fixed.

Fitness is observed on the population level based on a selective phenotype, but the driver of fitness is determined on the molecular level by the genotype. Proteins are encoded by a sequence of amino acids from which intramolecular interactions between these residues determine the biophysical properties such as the native fold, stability, dynamics, and for enzymes, their kinetic properties. One model of molecular evolution suggests that proteins evolve though a mutation-selection balance, where mutations introduced by

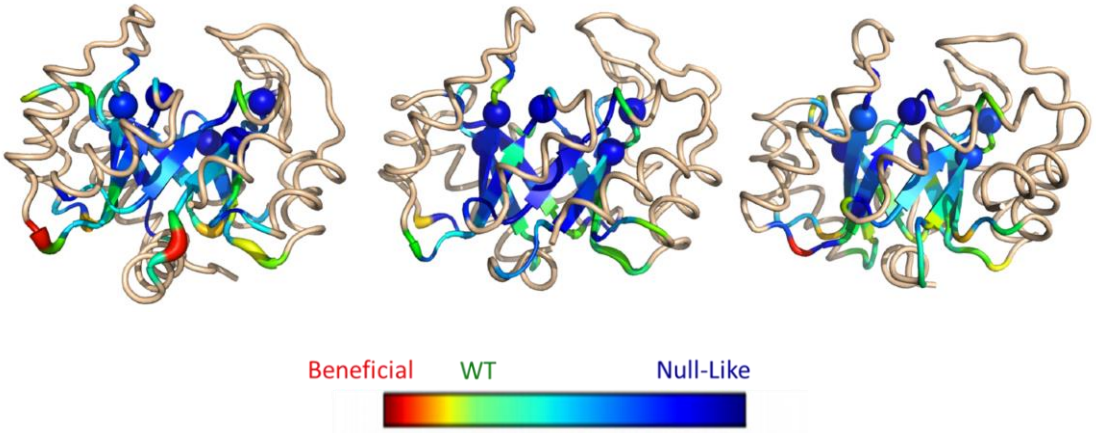
genetic drift are mainly under selection for protein stability and proper folding in order to maintain function[83,85,130].

We previously carried out an EMPIRIC fitness screen to probe the effect of sequence on protein stability, activity, and fitness[131]. In this experiment, indole-3-glycerol synthase (IGPS), the enzyme responsible for catalyzing the third step in the tryptophan biosynthetic pathway, was knocked out, yielding a tryptophan auxotrophic *S. cerevisiae* yeast strain. Functional complementation was achieved by transformation of orthologous IGPS from each of the three phylogenetically divergent sources: hyperthermophilic archaeon *S. solfataricus*, hyperthermophilic bacteria *T. maritima*, and thermophilic bacteria *T. thermophilus*. IGPS from all three sources are highly similar in structure, folding to the canonical TIM barrel structure, but their sequences vary widely with only 30-40% sequence identity. The comparison of fitness between orthologs permitted variation in sequence, while holding structure and function constant. Sequence space was greatly expanded through the EMPIRIC deep mutational scanning method. Saturating point mutations were introduced to the stability elements of the TIM barrel, the βα-loops and the β-barrel, in order to focus the study on relationship between sequence, structure, and stability. To a first approximation, the fitness readout is reporting on protein activity, however, given the location of our mutations, the major factor influencing catalytic efficiency will be the ability for the enzyme to attain its structure and catalyze its reaction. Protein stability is the major Other factors also come into play since cellular

fitness is determined on the organism level, where metabolic regulation and protein homeostasis networks are actively engaged.

An interesting observation from the fitness screen revealed regular patterns of beneficial mutations reflected throughout the four-fold symmetry of the TIM barrel architecture for all three orthologs. Low-temperature stress for the thermophilic enzymes may have provided the opportunity for selection of mutants whose activity exceeds the wildtype at the mesophilic host temperature. While the presence of beneficial mutations can be rationalized, the location of these mutations was surprising. Mutations in residues within the canonical βα-hairpin clamps at the C-terminal end of the β-barrel show the greatest fitness at the opposite terminus of the β-barrel (Fig. 3.1). These residues are far from the active site which lies on the N-terminal face of the barrel.

Figure 3.1 Orthologous IGPS proteins showed similar patterns of fitness response at structurally aligned positions



Ribbon diagrams of SsIGPS (Left, PDB: 2C3Z), TmIGPS (Center, PDB: 1I4N), and TtIGPS (Right, PDB: 1VC4) color-coded by their average selection coefficient. Spheres denote active site residues. Mutations of βα-hairpin regions show the highest beneficial fitness response. Alternating light and dark blue colors on the β-barrels demonstrate the relative fitness response of mutations for residues pointing into and out of the barrel, respectively.

Previously unassociated with any catalytic role, the βα-hairpin clamps are long range side-chain main-chain hydrogen bonds between even numbered αβ-loops and the preceding odd numbered β-strands that are important for β-strand alignment and providing stability[87]. The minimally independently folding unit, the βαβ-module, may reflect the stabilizing role of the βα-hairpin within the fourfold βαβα symmetric structure[38]. Under mesophilic temperature, a non-productive conformation is favored for these thermophilic enzymes[92,132]. Loss of the canonical βα-hairpin may result in beneficial fitness effects by destabilizing the native state, such that a higher energy, productive

conformation is adopted resulting in a population shift to a more catalytic competent native state ensemble or increased flexibility required for catalysis. Another set of mutations that showed beneficial effects were found for residues in the β-strands whose side chains point out of the barrel towards the α-helical shell (Fig 3.1). These side chains interact with the α-helices, effectively docking the outer helical shell. Considering these two classes of beneficial mutations together, we hypothesized that a pathway of communication through high order energetic coupling leads from the αβ-loops towards the active site at the βα-loops via the α-helical shell.

In this study, our goal was to determine the molecular mechanism behind the allosteric effect of these beneficial mutations. Fitness for single and double point mutants within βα-hairpin clamps and α-helices were determined to investigate possible long-range connectivity between the αβ-loops and βα-loops via the α-helical shell. Double mutants of βα-hairpins were studied to explore the interaction between the four symmetrical folds. To complement the fitness assays and probe the molecular basis of observed changes in fitness, a series of structural, thermodynamic, and enzymatic assays were performed. Elucidation of the pathway by which the disruption of the βα-hairpin increased fitness within a population would be a major step forward in our understanding of complex interrelationship between protein sequence, structure, stability, and function with a highly represented biological system.

Results

Fitness of individual point and double mutants

To probe for potential networks of long-range interactions, fitness was measured for single and double point mutations introduced throughout the β/α interface of the TIM barrel of SsIGPS. High order energetically linked interactions along the length of the protein and between the fourfold symmetry are investigated using double mutants. Alanine was chosen as the standard for mutations because its small size avoids steric clashes. Wildtype alanines were mutated to isoleucines. While isoleucines are usually considered helix breakers, isoleucines have an unusually large propensity in helices of TIM barrels, potentially to increase the energy penalty of an α-helix undocking from the β-barrel[133].

Fitness is referred to here as the growth rate (inverse of doubling time) of tryptophan auxotrophic IGPS knockout yeast transformed with the mutant SsIGPS construct. Functional complementation rescues the auxotrophy during selective growth in media lacking tryptophan. Quantification of the selection coefficient, *s*, compares the relative growth rate of the mutant to our SsIGPS wildtype (SsWT). A mutant with selection coefficient of 0 has a comparable growth as SsWT, whereas a selection coefficient of -1 has a comparable lethal phenotype as null or uncomplemented yeast. Growth rates exceeding SsWT have selection coefficients greater than 0, and are considered beneficial mutations. A table of mutants, the locations of the mutated sites on the structure,

the interaction we are probing, and selection coefficients is provided (Table 3.1 and Table 3.2).

Table 3.1 Point mutations introduced throughout SsIGPS resulted in a wide range of fitness response

Point mutation	Location	Potential interaction of interest	*s*
F40A	α0	Hydrophobic clamp	0.01
I45A	α0β1-loop	Hydrophobic clamp	0.04
I45K	α0β1-loop	Hydrophobic clamp	0.08
S70A	α1		0.02
M73A	α1		-0.14
M73I	α1		-0.23
V78A	α1β2-loop	Hydrophobic clamp	-0.27
L96A	α2		-0.20
I99A	α2		0.01
I107A	β3	βα-hairpin clamp-Donor	0.07
I107K	β3	βα-hairpin clamp-Donor	0.08
L108A	β3		-0.53
L108K	β3		-0.47
A122I	α3		-0.92
D128A	α3β4-loop	βα-hairpin clamp-Acceptor	0.05
D128K	α3β4-loop	βα-hairpin clamp-Acceptor	-0.11
R150A	α4	Ionic bond	-0.47
E155A	α4β5-loop	Ionic bond	-0.88
D165N	α5		-0.63
L166A	α5		-0.03
R175A	α5β6-loop	Ionic bond	-0.04
L197A	α6		-0.02
K207A	β7	βα-hairpin clamp-Donor	-0.11
E218A	α7		-0.18
N228A	α7β8-loop	βα-hairpin clamp-Acceptor	0.00

Table 3.2 Fitness response from double mutants suggest some distal sites are energetically linked

Double mutations	Location	Potential interaction of interest	*s*
I45A/S70A	α0β1-loop/α1	Mod1-clamp/Mod1-helix	0.09
I45A/M73A	α0β1-loop/α1	Mod1-clamp/Mod1-helix	-0.11
I45A/I107A	α0β1-loop/β3	Mod1-clamp/Mod2-clamp	0.10
S70A/A122I	α1/α3	Mod1-helix/Mod2-helix	-0.99
M73A/A122I	α1/α3	Mod1-helix/Mod2-helix	-0.99
I107A/D128A	β3/α3β4-loop	βα-hairpin clamp-Acceptor/Donor pair	-0.01
I107A/E155A	β3/α4β5-loop	Mod2-clamp/Mod3-clamp	-0.99
I107A/K207A	β3/β7	Mod2-clamp-Donor/Mod4-clamp-Donor	-0.09
D128A/R175A	α3β4-loop/α5β6-loop	Mod2-clamp-acceptor/Mod4-clamp-Acceptor	-0.23
D165N/R175A	α5/α5β6-loop	Mod3-helix/Mod3-clamp	-0.97

Several point mutations were deleterious: E155A (*s* = -0.88) and A122I (*s*=-0.92). Likewise, double mutants containing either one of those mutations resulted in a lethal phenotype. E155 in αβ-loop 5 forms a non-canonical hairpin clamp through a salt bridge triad with R175 in αβ-loop 6 and R150 in α-helix 4.

Located on α-helix 3, the methyl group of A122 faces the βα interface. Due to the intrinsic tilt of β-barrels with a shear number of 8, A122 is within 4Å of M109 at the C-terminus of β3 and V130 at the N-terminus of β4.

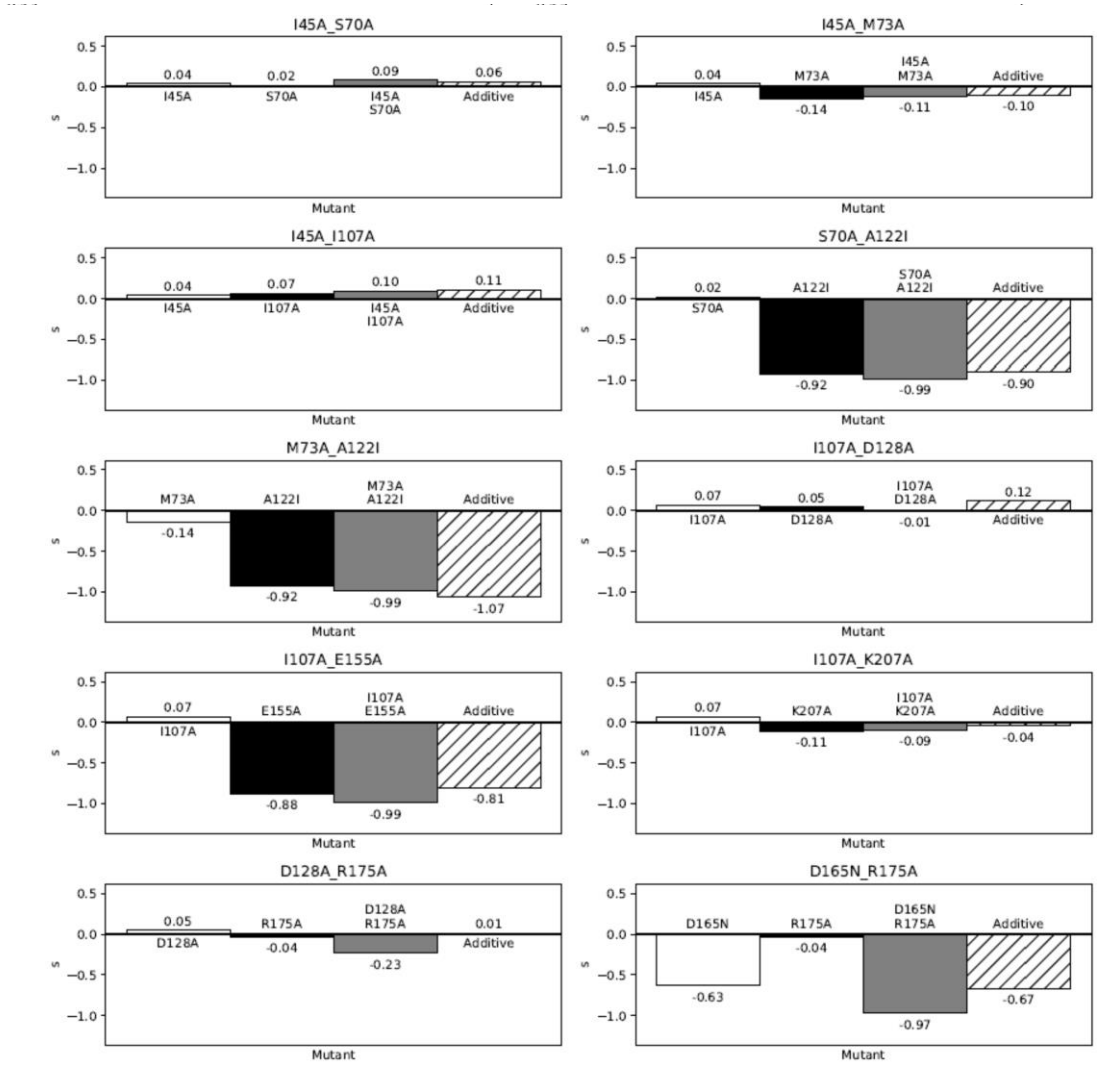
Other mutations had moderate fitness defects (-0.2 > *s* > - 0.6), while others improved fitness. From the previous fitness screen, the most beneficial point mutations were associated with the βα-hairpins within the first and second modules of all three TIM barrels. For SsIGPS, these mutations and their selection coefficients from individual growth assay are I45K (*s* = 0.08), I107K (*s* = 0.07), I107A (*s* = 0.07), D128A (*s* = 0.05), I45A (*s* = 0.04). The results of these individual growth assays agree with the fitness results obtained from the bulk competition EMPIRIC screen.

Only two mutations in the α-helices were found to be beneficial, both in the first βαβ-module: S70A in α-helix 1 (*s* = 0.02) and I99A in α-helix 2 (*s* = 0.01). Interestingly, S70 is buried at the interface of α1 and α2 and its hydroxyl is hydrogen bonded to S102 to avoid an energetic penalty. Two essential catalytic residues, E51 and K53, are found at the C-terminus of the first β-strand and the first βα-loop stemming from the β-strand, respectively. A few residues away from S70, the mutation of M73 to either alanine or isoleucine resulted in decreased fitness relative to SsWT (s_{M73A} = -0.14, s_{M73I} = -0.23). The sidechain of M73 points in the opposite direction as S70 and is within 4Å of two residues in α-helix

8, L236 and I243. These results illustrate to the sensitivity of fitness to side chain replacements.

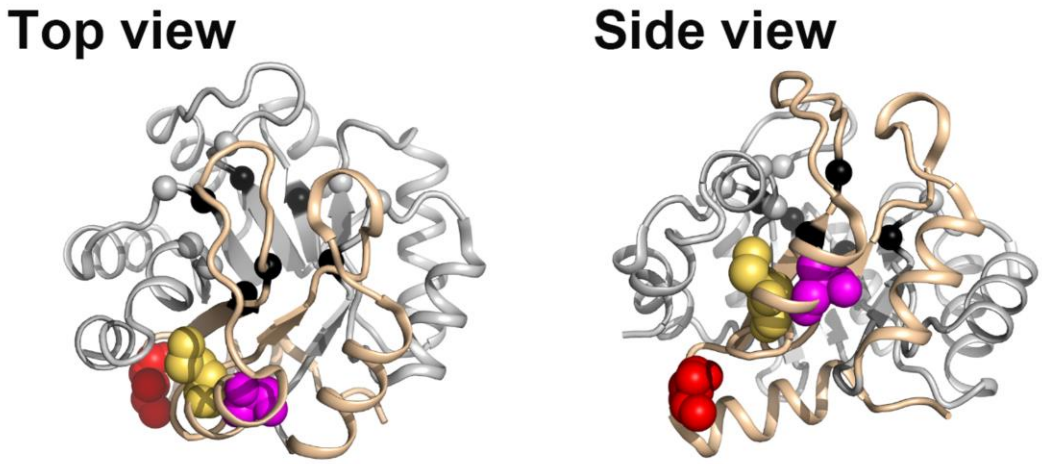
Double mutants are often used to probe possible interaction between two sites (Fig 3.2). Additive effects are attributed to two independent sites, while non-additive effects are attributed to interacting sites. As expected, the two residues that form the canonical βα-hairpin clamp in module 2, I107 and D128, showed nonadditivity in the double mutant I107A/D128A. Of particular interest, two double mutants, I45A/S70A (s_{I45A} = 0.04, s_{S70A} = 0.02, $s_{I45A/S70A}$ = 0.09) and D165N/R175A (s_{D165N} = -0.63, s_{R175A} = -0.04, $s_{D165N/R175A}$ = -0.97), have fitness exceeding the expectations from the additive effects of individual mutations. These results suggest that the clamp and the helix are energetically linked. Within the same turn of the helix as S70A, mutation I45A/M73A did not show the same synergistic effect as I45A/S70A (Fig 3.3).

Figure 3.2 Most double mutations were non-additive



Selection coefficient values for the single and double mutations are plotted on to a bargraph. The white and black bars represent the two point mutants that make up the double mutant, represented by the gray bar. The hatched bar represents the hypothetical additive value of the two individual point mutations.

Figure 3.3 Residues S70A and M73A are situated on the same turn of α-helix 1

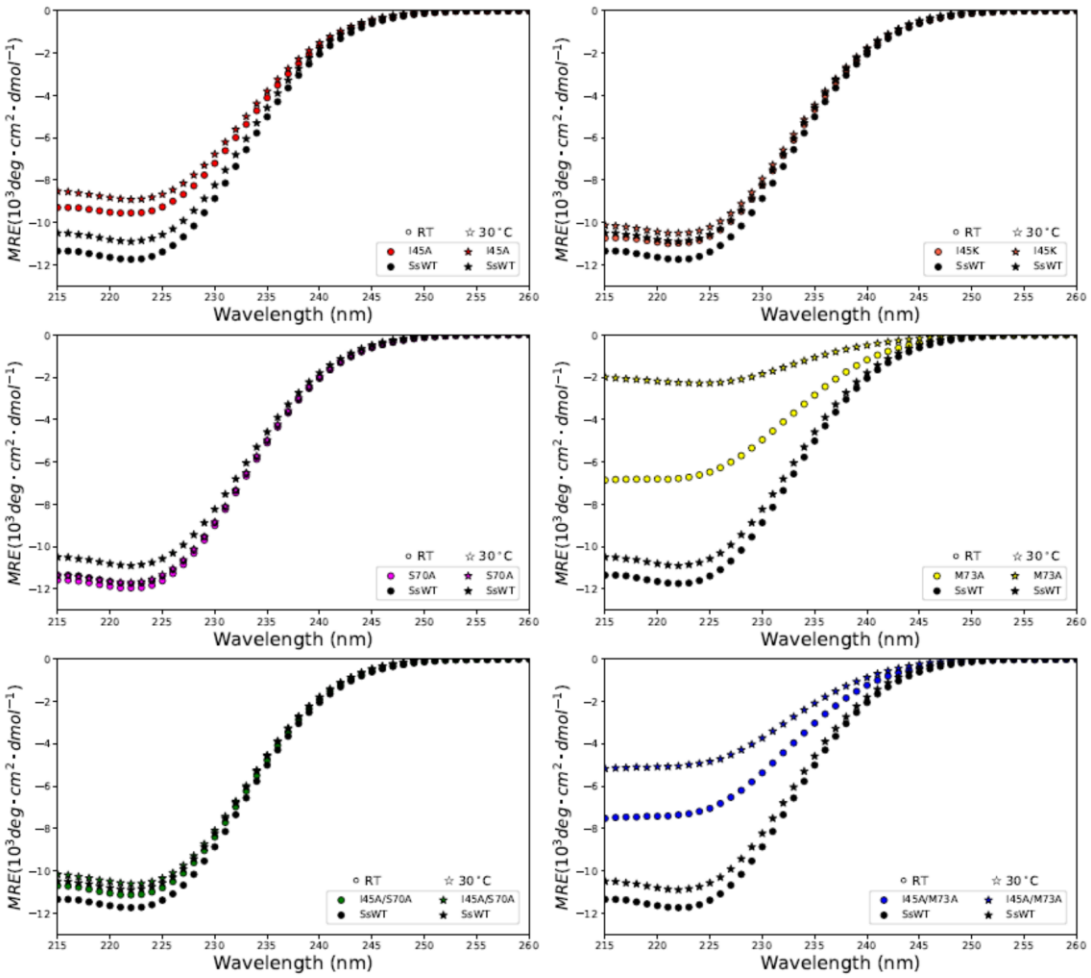


Ribbon diagram highlighting residues I45 (red), S70 (magenta), M73 (yellow). The wheat color denotes the boundaries of the one of the fourfold symmetric unit. Spheres denote active site residues (black – enzyme chemistry, gray – substrate/product binding). Two active site residues are located N-terminal to α-helix 1.

Perturbation of secondary structure

The far-UV CD spectra of SsIGPS mutants show the characteristic minimum at ~222 nm for structures with large α-helical secondary structure content. Compared to SsWT, the MRE signal is reduced for several mutations, in particular, I45A and M73A, and the double mutant pair. Of note, while all spectra here were collected several days following purification, M73A and I45A/M73A showed highly reduced MRE signal compared to spectra collected the day following purification (data not shown). Furthermore, a dramatic decrease in signal is observed in spectra collected at room temperature, 22°C, compared to the yeast host temperature, 30°C, for mutants M73A (~67% loss in θ_{222} signal) and I45A/M73A (~31% loss in θ_{222} signal), reflecting lower thermal stability. SsWT and other mutants displayed between a 2 to 7% loss in signal when equilibrated to 30°C (Fig 3.4).

Figure 3.4 Ellipticity of SsWT and SsIGPS variants demonstrate temperature sensitivity of some mutants

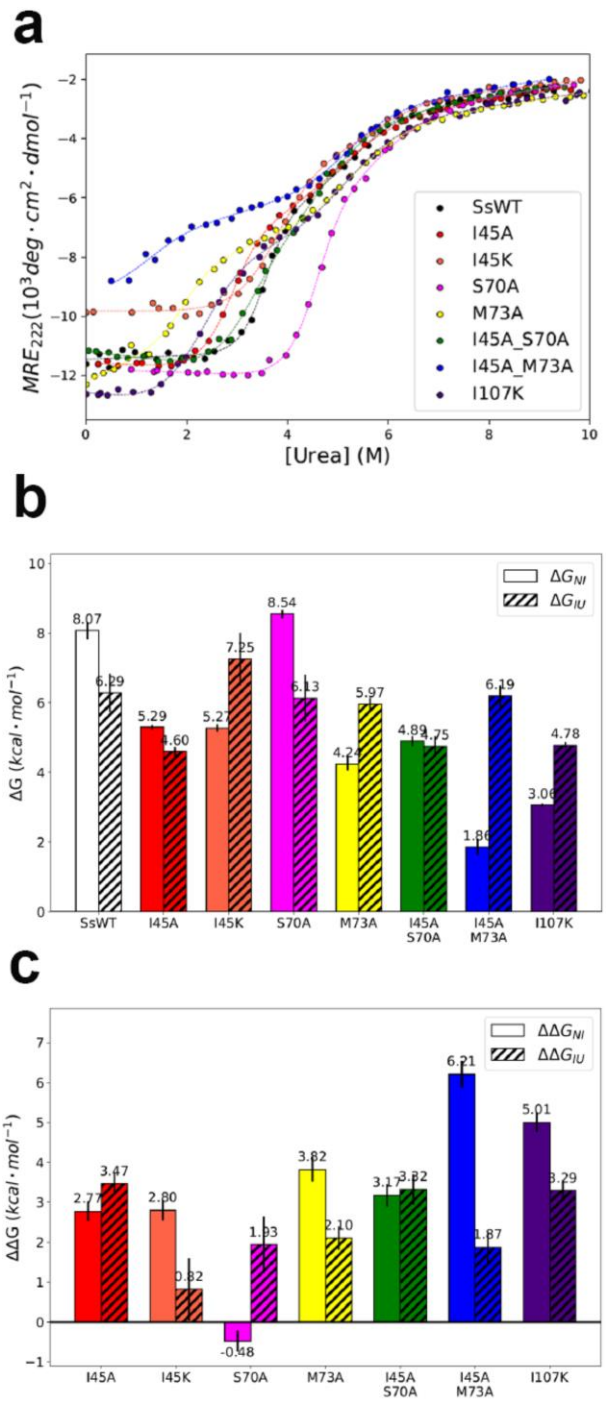


Far-UV CD spectra of SsWT and mutants collected at room temperature, ~22°C, (○) and ~30°C (☆) in 10mM KPi, pH 7.2. Mutant spectra are shown in color. SsWT spectra are shown in black for reference. The prominent negative band at 222 nm is indicative of the presence of α-helix secondary structure. The reduced CD signals of M73A (yellow) and I45A/M73A (blue) at ~22°C were further reduced at 30°C.

Protein unfolding by urea denaturation

Equilibrium urea denaturation curves were carried out for select mutants. Protein unfolding displayed a 3-state process, Native (N) ⇌ Intermediate (I) ⇌ Unfolded (U), for all mutants examined (Fig. 3.5a). The SsWT titration collected at 30°C could not be accurately fit, thus, for the purpose of this exploratory analysis, a titration of SsWT collected at room temperature, 22°C, was used. Once a suitable titration has been collected, this dataset will be reanalyzed with the properly matched wildtype reference. Most mutations resulted in decreased stability for both the N to I transition and the I to U transition (Fig. 3.5b). Reduced MRE signal for mutants, I45A/M73A (blue) and I45K (orange), suggest some loss in secondary structure in the native state. Remarkably, I45K maintains a flat native baseline beyond 2 M urea, indicating that the altered native state structure is stable. In contrast, a steep native baseline for M73A (yellow) indicates a highly unstable native state. Excluding the previously mentioned mutants displaying reduced MRE, the most striking loss in native state stability were observed for I107K (purple) (Fig. 3.5b). Interestingly, mutation S70A showed increased native state stability. In all cases, mutations appear to impact the N to I transition more than the I to U transition (Fig. 3.5c, Table 3.3). A linear relationship was observed between ΔG_{NI} and ΔG_{Total} ($r = 0.94$). A weaker correlation was observed between ΔG_{IU} and ΔG_{Total} ($r = 0.55$).

Figure 3.5 SsIGPS variants displayed a destabilized native state



The entire CD spectra as a function of urea for each variant was globally fit to a 3-state model. Collected sample reads are indicated by the filled circles. Fits to the data are indicated by the dash lines. (A) Urea melts observed by CD at 222 nm show stabilities of mutants differed from SsWT for both the native and intermediate states. (B) The free energy differences of unfolding for the N to I transition, $\Delta G°_{NI}$ (solid fill), and for the I to U transition, $\Delta G°_{IU}$ (hatched fill) are graphically represented in a bar graph. (C) Changes in protein stability induced by the mutations, $\Delta\Delta G°_{NI}$ and $\Delta\Delta G°_{IU}$, were determined as the difference in free energy between SsWT and the SsIGPS variant for the respective NI and IU transitions.

Table 3.3 Thermodynamic parameters of SsIGPS variants

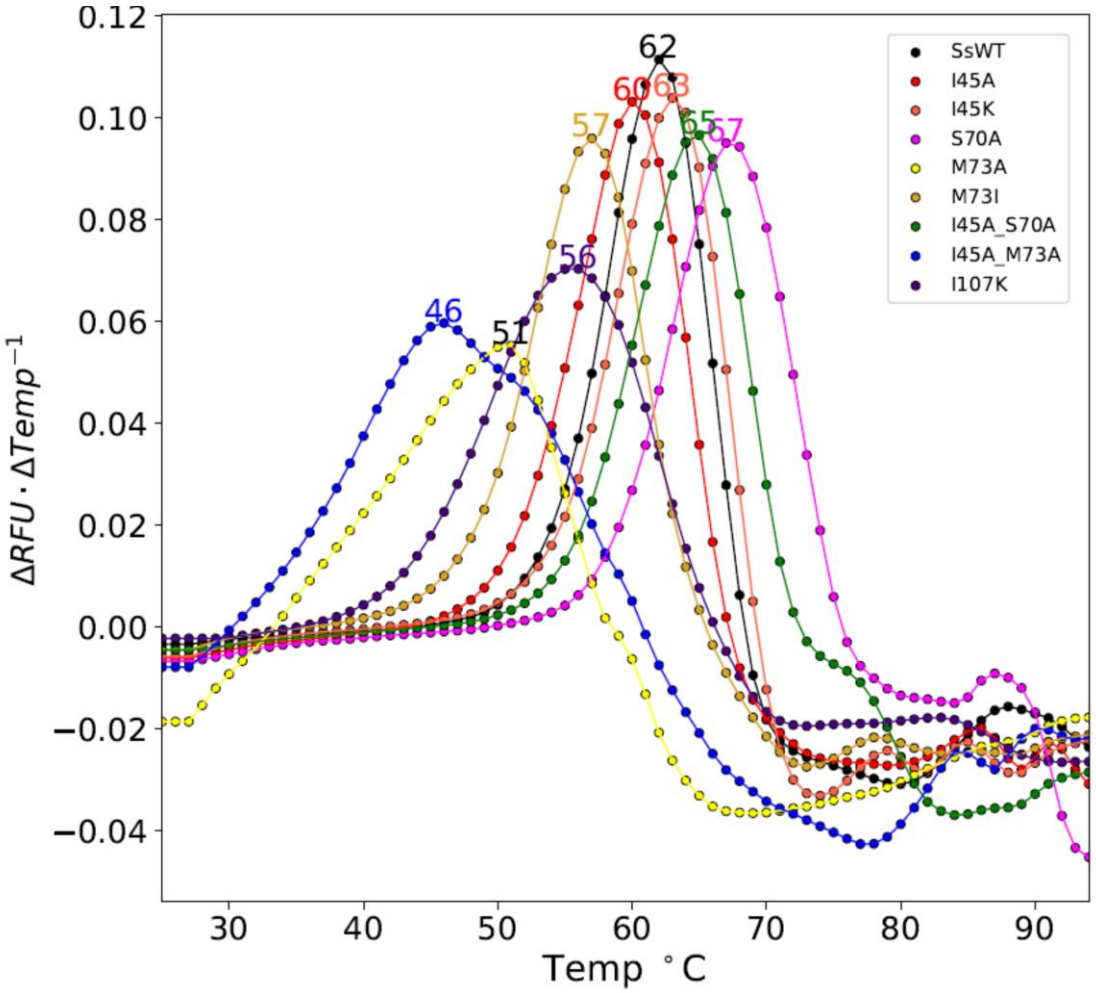
SsIGPS variant	ΔG_{NI}	ΔG_{IU}	ΔG_{Total}	$\Delta\Delta G_{NI}$	$\Delta\Delta G_{IU}$	δ_{NI} (kcal·mol^{-1})	δ_{IU} (kcal·mol^{-1})
SsWT	8.07±0.24	6.29±0.55	14.35±0				
I45A	5.29±0.06	4.6±0.12	9.9±-2.77	2.77±0.24	3.47±0.27		
I45K	5.27±0.1	7.25±0.75	12.52±-2.8	2.8±0.26	0.82±0.79		
S70A	8.54±0.11	6.13±0.66	14.68±0.48	-0.48±0.26	1.93±0.71		
M73A	4.24±0.18	5.97±0.17	10.21±-3.82	3.82±0.3	2.1±0.29		
M73I	-	-	-	-	-		
I107K	3.06±0.04	4.78±0.08	7.84±-5.01	5.01±0.24	3.29±0.25		
I45A/S70A	4.89±0.14	4.75±0.28	9.64±-3.17	3.17±0.27	3.32±0.37	0.88±0.45	-2.08±0.84
I45A/M73A	1.86±0.23	6.19±0.3	8.05±-6.21	6.21±0.33	1.87±0.38	-0.39±0.51	-3.69±0.55

M73I could not be fit due to low ellipticity signal.

Protein unfolding by thermal denaturation

Given the sensitivity of the ellipticity to temperature (Fig. 3.4), we probed the thermal transitions of the same set of variants with differential scanning fluorimetry. Thermal denaturation curves were carried out for the same mutants as with the urea denaturation. In this assay, native protein is incubated with an environmentally sensitive dye, SYPRO red, to monitor thermally induced denaturation. When proteins unfold upon addition of heat, SYPRO red binds to the newly exposed hydrophobic patches, emitting fluorescence and serving as a reporter for the thermal denaturation reaction. The melting temperatures (T_m) are obtained from the maximum value of the first derivative of the melting curve (Fig 3.6).

Figure 3.6 Melting temperatures varied greatly by mutation



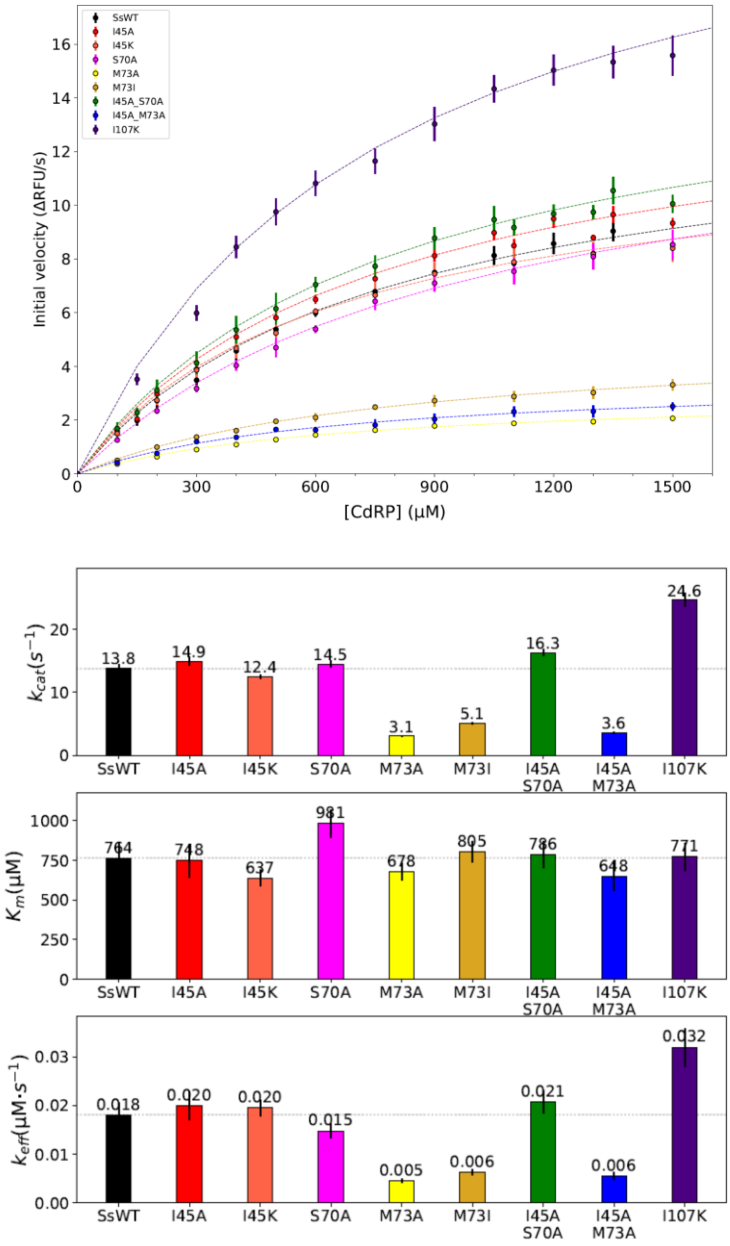
SsIGPS variants were melted as function of temperature, from 25°C to 95°C in 1°C·min^{-1} increments in 10 mM KPi, pH 7.2. Protein unfolding is reported by fluorescence upon SPYRO red binding. Mutants with higher melting temperatures are more thermal stable.

Consistent with the urea denaturation data, S70A was most resistant to denaturation, surpassing SsWT. I45A/M73A with the highly reduced MRE and M73A with the steep native baseline were the first variants to melt, followed by I107K, which had a greatly reduced ΔG_{NI}. I45K, which also had reduced MRE signal, but displayed resistance to urea melting, had a comparable melting temperature to SsWT.

Enzyme kinetics

Another critical metric for assessing the molecular basis for the phenotypic effects of mutations on fitness is enzymatic catalysis. Kinetic parameters were determined for the ability of SsIGPS variants to convert CdRP to IGP. This conversion yields a naturally fluorescent product that emits at the same wavelength as the pathway's end-product, tryptophan. Initial velocities of each of the SsIGPS variants were calculated at multiple concentrations of substrate to derive kinetic parameters, V_{max}, k_{cat}, K_m, and the catalytic efficiency $k_{eff} = k_{cat}/K_m$, using the Michael-Menten equation (Fig 3.7a, b).

Figure 3.7 Initial velocity curves identified several SsIGPS mutants that were more catalytically efficient than SsWT



Initial velocity measurements were collected to determine the catalytic efficiency of each SsIGPS variants at 30°C in 10 mM KPi, pH 7.2. **(a)** Fluorescence readings from the product formation (○) were fit to the Michaelis-Menten equation (---) to determine the kinetic parameters, k_{cat}, K_m, and k_{eff}. **(b)** The kinetic parameters were plotted as bargraphs. The gray dashed line is a visual guide for the value associated with SsWT. Several mutants showed greater catalytic efficiency compared to SsWT. Mutations at residue M73 resulted in poor catalytic efficiency.

Some mutations such as I107K had increased k_{cat} parameters, whereas other mutations, such as those involving residue M73, showed distinctly decreased k_{cat}. In contrast, most mutations had comparable K_m values to SsWT with a few exceptions. S70A showed the greatest deviation, with a 30% increase in Km, while I45K showed a 16% decrease in K_m. Evidently, these side chain replacements distal from the active site can impact both the substrate turnover as well as substrate binding affinity. A table of all the features is provided to summarize the results (Table 3.4).

Table 3.4 Kinetic parameters of SsIGPS variants

SsIGPS variant	kcat (s^{-1})	Km (μM)	kcat/km ($\mu M^{-1} \cdot s^{-1}$)	$\Delta\Delta G^{\ddagger}_{kcat}$ (kcal·mol^{-1})	$\Delta\Delta G^{\ddagger}_{keff}$ (kcal·mol^{-1})	δ_{kcat} (kcal·mol^{-1})	δ_{keff} (kcal·mol^{-1})
SsWT	13.79±0.58	764±98	0.018±0.0024				
I45A	14.92±0.67	748±103	0.0199±0.0029	0.047	0.060		
I45K	12.44±0.33	637±51	0.0195±0.0017	-0.062	0.048		
S70A	14.46±0.53	981±87	0.0147±0.0014	0.028	-0.122		
M73A	3.07±0.09	678±55	0.0045±0.0004	-0.905	-0.833		
M73I	5.07±0.16	805±66	0.0063±0.0006	-0.602	-0.633		
I107K	24.62±1.01	771±88	0.0319±0.0039	0.349	0.343		
I45A/S70A	16.26±0.58	786±84	0.0207±0.0023	0.099	0.082	0.023	0.144
I45A/M73A	3.59±0.16	648±88	0.0055±0.0008	-0.810	-0.711	0.047	0.061

Discussion

Organism fitness is subject to the local environment. A beneficial phenotype in one setting may be an impediment in another. In this fitness screening system, a tryptophan biosynthetic gene, IGPS, from hyperthermophilic Archaeon *S. solfataricus* (SsIGPS) is transformed into and rescues a tryptophan auxotrophic IGPS knockout *S. cerevisiae* yeast. Expression of SsIGPS protein functionally complements the missing endogenous IGPS gene activity, completing the multi-step tryptophan biosynthetic pathway and permitting growth in synthetic media lacking this amino acid. While SsIGPS can complement the knockout condition, preventing lethality in synthetic *-trp* dropout media, the fitness is lower than when the endogenous ScIGPS gene is reintroduced using the same complement system, with a doubling time of ~3 hrs for ScIGPS vs ~4-5 hrs for SsIGPS (Data not shown).

Despite sharing a common structure and function, the different doubling times suggest that the catalytic efficiencies are not equal between these two orthologous IGPS TIM barrel protein within the yeast cell. The hyperthermophilic SsIGPS enzyme is not optimized for the mesophilic temperature of the yeast, introducing a temperature stress. Consequently, there is potential for a mutation to improve enzyme activity.

Recently, we carried out a large-scale EMPIRIC mutagenesis fitness screen using this complementation system and identified several point mutations

that conferred beneficial fitness effects for SsIGPS as well as for two other thermophilic orthologs, *T. maritima* and *T. thermus*[131]. Beneficial mutations at analogous sites within the TIM barrel structure across the three orthologs pointed to a common mechanism. The most distinctive sites detected occur at a canonical TIM barrel feature, the βα-hairpin clamps, found in each of the fourfold symmetrical units of the protein structure. The mutations at these sites are particularly interesting, not only because they confer some of the fastest doubling time, but they are also distal from the active site with no known connection to catalytic function. To investigate the molecular mechanism behind these beneficial allosteric mutations, we selected several sites for detailed biological, biophysical, and biochemical characterization. This multilevel approach is essential for bridging the effects of mutation on the molecular level and the selection process on the organism level.

Mutation sites were sampled throughout the TIM barrel of SsIGPS. The fitness effects of the mutations are directly compared to SsWT using the selection coefficient. Individual fitness determinations for the point mutations of the canonical β3α3 hairpin clamp and the non-canonical hydrophobic clamp in the βαβ1 module recapitulated the results of the bulk EMPIRIC assay. New mutations within the helical shell and double mutants probed the potential network of interactions conferring the beneficial fitness effects.

We investigated the potential interactions between and within modules based on the selection coefficients of double mutants. Mutation of clamp residues in different modules typically resulted in less than additive fitness effects (I107A/E155A, I107A/K207A, D128A/R175A). These results hint at interaction between the modules. Likewise, mutation of both partners of a clamp (I107A/D128A) resulted in near neutral fitness compared to the beneficial response of each individual mutation. Mutation of helix residues in two different helices (S70A/A122I, M73A/A122I) resulted in lethal phenotype, likely following the lethal phenotype of the individual mutation, A122I. This epistatic effect is reminiscent of a thermodynamic cliff, where protein structure can no longer be maintained after loss of a certain threshold of stability[83,85,86]. Within the same module, D165N/R175A resulted in a greater growth defect than the individual mutations would predict. Likely, there is crosstalk between distant residues within the same module and some interactions affect fitness more than others. Also within a single module, positive epistasis is observed for I45A/S70A, while I45A/M73 appears to have an additive interaction. Considering that both the beneficial S70A and the near-neutral M73A mutation are within the same turn of α-helix 1, the different fitness response in their individual and double mutation poses an interesting juxtaposition. To probe the beneficial nature of mutation S70A and the synergistic effect of I45A/S70A, individual and double mutations of residues I45, S70, and M73 were selected for follow-up *in vitro* studies to explore

possible long range interactions. The highly beneficial mutation I107K in the canonical βα-hairpin clamp was also included for further studies.

The selection coefficients of the point mutants demonstrate that the partnered residues within a clamp do not contribute equally to fitness, and the mutation type matters. The local stereochemical environment and the residue's role in supporting protein stability will contribute to the benefit or detriment of a mutation. Lodged between I45 and V78, loss of the aromatic ring for F40 had minimal impact on fitness, likely because hydrophobic interactions may still persist between I45, F40A, and V78. Loss of V78 may prevent α0 from docking and properly capping the N-terminus of the β-barrel. Mutation I45A resulted in a positive selection coefficient, perhaps by inducing conformational changes upstream to the β1 and β1α2-loop, each holding an essential catalytic residue (highlighted in black) or by reduced interaction with I247 which is downstream of the phosphate binding site in β8α8 loop. Strikingly, replacement of the isoleucine with a bulkier, charged lysine, resulted in better fitness than I45A, highlighting the importance of sequence in determining populations within the native well ensemble and its functional consequences. The charged amino acid must interact with the aqueous environment differently than the branched hydrophobic, likely distorting the structure and positioning the active site in a more reactive form for catalysis or product release. A similar effect is observed with residue I107, where mutation to I107K results in better fitness than I107A, both of which are cases of beneficial mutations.

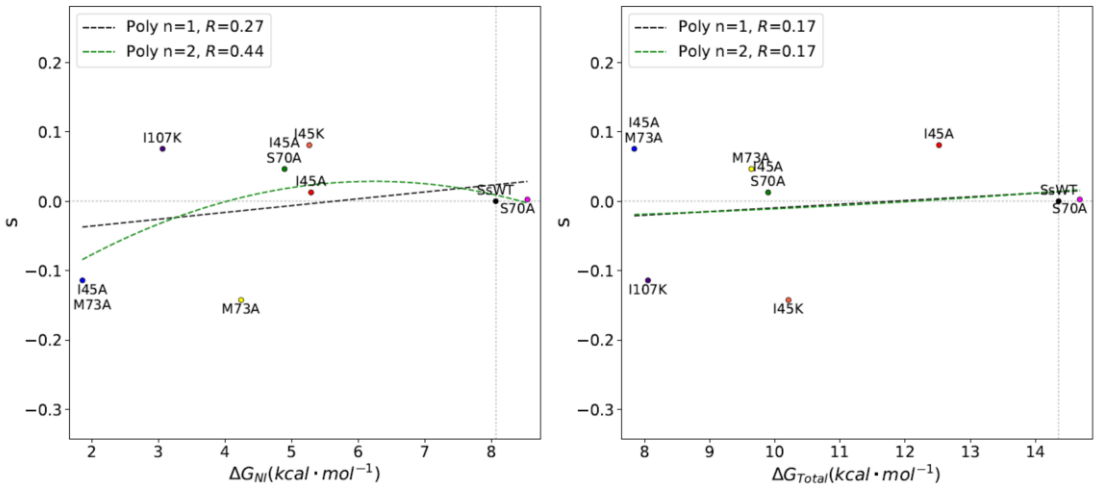
Previously, we observed beneficial mutations for residues within the β-strands pointing towards the α-helical shell and hypothesized that communication between the clamps in the αβ-loops and the active site in the βα-loops occurs via the α-helices. Double mutants were screened to determine if the two sites are acting within the same network of communication. Double mutant I45A/S70A (*s* = 0.09) appear to have a synergistic beneficial effect relative to the individual I45A (*s* = 0.04) and S70A (*s*=0.02). In contrast, M73A shows near neutral fitness and the fitness effect of I45A/M73A appears additive (Fig. 3.2).

Biophysical and biochemical measurements were analyzed to understand the underlying effects of protein stability and activity on fitness. Similar secondary structure content is found for I45A/S70A as compared to SsWT at both room temperature (RT) ~ 22°C and ~30°C (Fig. 3.4). However, I45A/M73A displayed a loss in secondary structure content. A key structural difference between S70A and M73A is that S70A interacts with α-helix 2, whereas M73A interacts with α-helix 8. Loss of the methionine to an alanine may weaken the interaction with nearby L236 and I243, increasing the likelihood of the α-helical shell opening. A similar loss in MRE signal for the M73A single mutation supports this notion.

In agreement with the reduced secondary structure content, stability measurements by urea titration showed a greatly reduced ΔG_{NI} (Table 3.2). A survey of all the mutants revealed that most of these mutants have a reduced ΔG_{NI}, and to a smaller extent a reduced ΔG_{IU}. S70A stood out as a special case

that had a greater ΔG_{NI}, but a comparable ΔG_{IU}. Since these beneficial and near neutral mutations seem to impact ΔG_{NI} more than ΔG_{IU}, we examined the relationship between *s* and ΔG_{NI} and between *s* and ΔG_{Total} (Fig. 3.8). Rather than a linear relationship, there appears to be a broad range of ΔG_{NI} around ~6 kcal·mol^{-1} resulting in optimum fitness. A weak linear correlation is observed between *s* and ΔG_{Total}.

Figure 3.8 Non-linear relationship observed between *s* and ΔG_{NI}

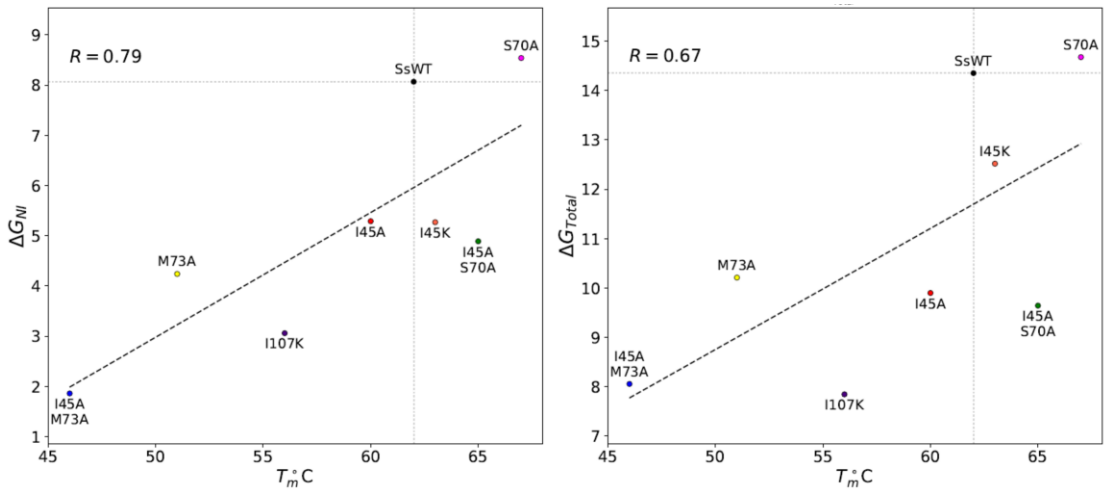


The selection coefficient of the SsIGPS variants were plotted against their stabilities, *s* vs ΔG_{NI} (left) and *s* vs ΔG_{Total} (right). Comparison between a linear (--) and quadratic (--) model showed a better fit with the 2nd order polynomial regression for ΔG_{NI}. In contrast, a first order regression was a better fit for ΔG_{Total}.

By comparing double and single mutants, the interaction free energy (δ) between two sites as readout through a protein unfolding reaction can be calculated by $\delta = \Delta\Delta G_{double} - (\Delta\Delta G_{mut1} + \Delta\Delta G_{mut2})$. A nonzero δ value means that two sites are interacting, either directly or through the protein matrix. The δ_{NI} and δ_{IU} for I45A/S70A and I45A/M73A were both nonadditive (Table 3.3). This result indicates that the non-canonical hairpin clamps between β1 and β2 perturbs the local environment of both S70A and M73 in α1. Thus, α1 provides a potential conduit for allostery between the active site in the βα-loops at the C-terminus of the β-strands and the hairpin clamps at the opposing end of the β-barrel.

Temperature sensitivity of some mutants prompted us to examine the apparent melting temperatures (T_m) of these enzymes. Tm is linearly correlated with both ΔG_{NI} and ΔG_{Total} obtained by urea denaturation ($R_{Tm\ and\ \Delta GNI} = 0.79$, $R_{Tm\ and\ \Delta GTOTAL} = 0.67$), but is a better reporter for native state unfolding than for the intermediate.

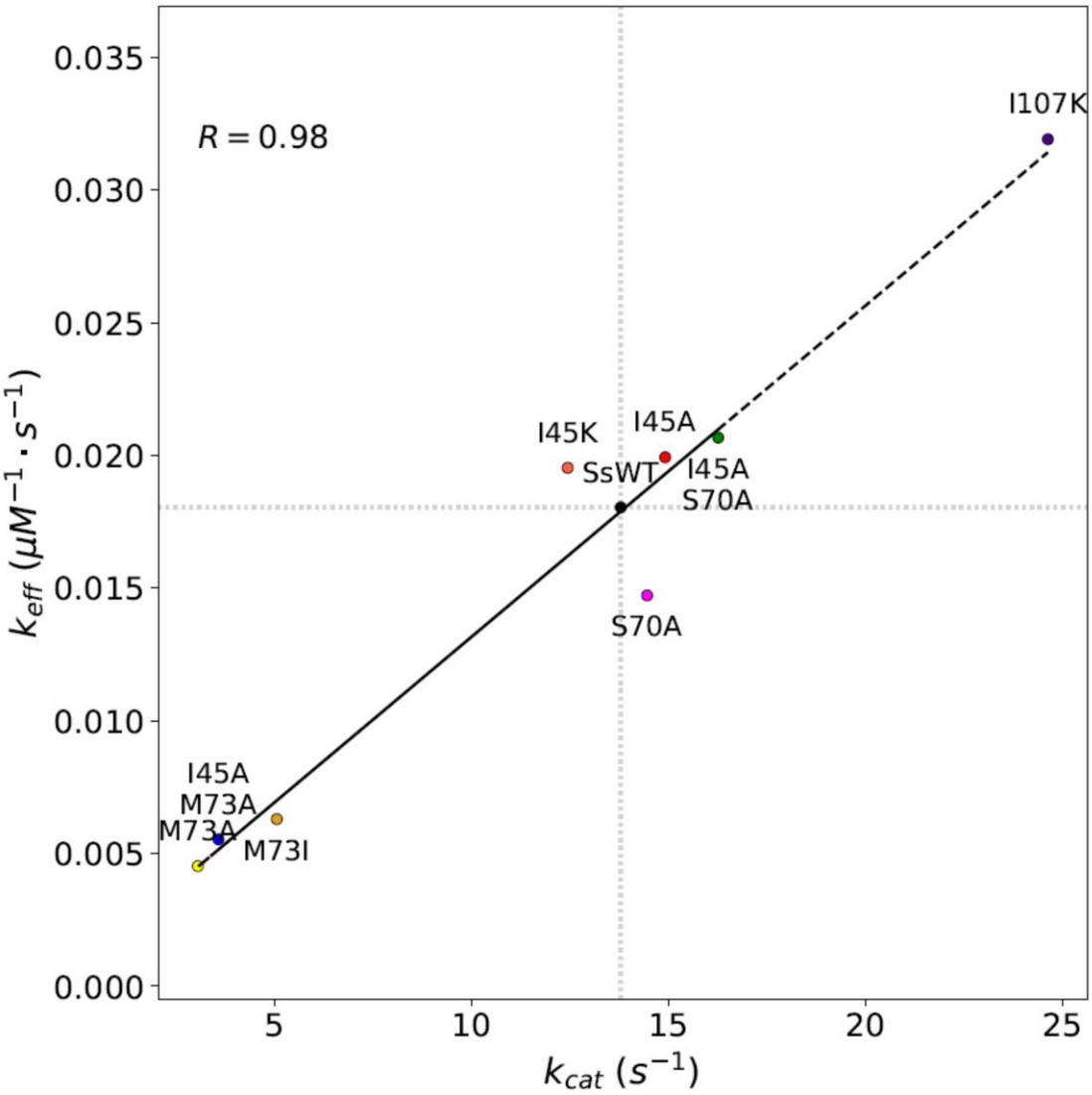
Figure 3.9 Linear relationship observed between ΔG_{NI} and T_m



The stabilities of the SsIGPS variants were plotted against their melting temperatures, ΔG_{NI} vs T_m (left) and ΔG_{Total} vs T_m (right). A linear relationship was observed between stability and melting temperature in both cases. The fit is denoted by the black dashed line (--).The gray dashed lines are visual guides for the values associated with SsWT.

In addition to protein structure and stability, we examined the impact of mutation on enzyme function. Some mutations resulted in greater catalytic efficiency, while others had the opposite effect. A strong correlation (r = 0.98) between k_{eff} and k_{cat} suggest that mutations affect enzyme activity primarily through modification of the substrate turnover rate (Fig 3.10). Two mutations, I45K and S70A, appear to affect substrate binding more than the enzyme chemistry in driving catalytic efficiency.

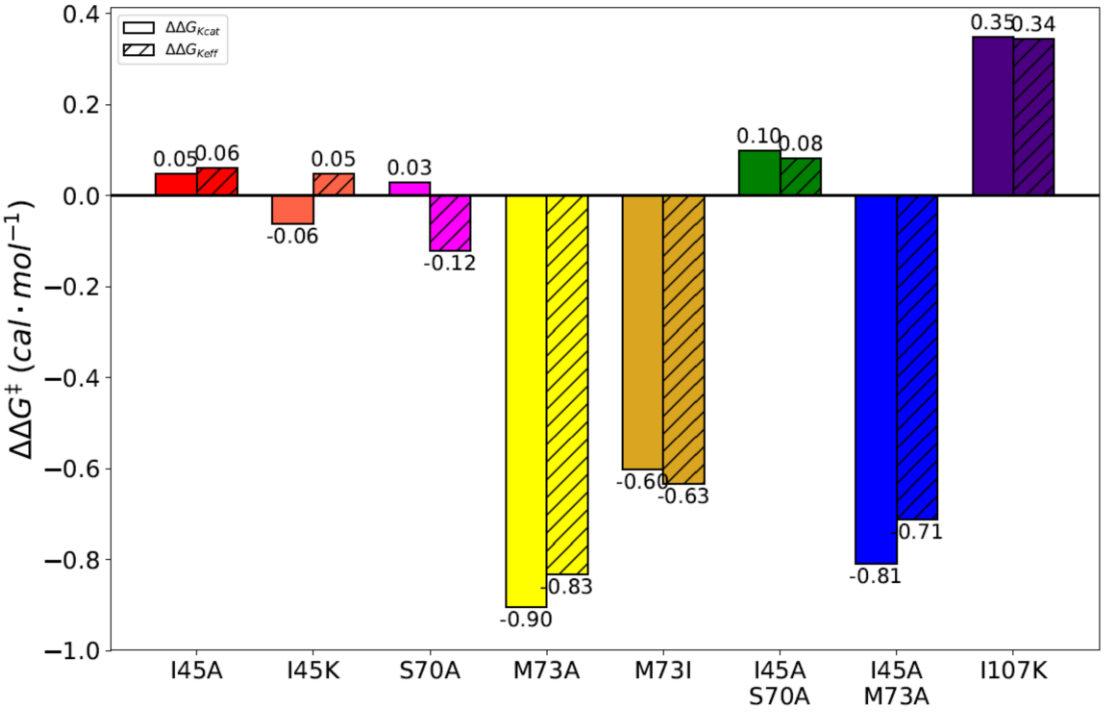
Figure 3.10 Linear relationship observed between k_{eff} and k_{cat}



The k_{eff} was plotted as a function of k_{cat} for each of the SsIGPS variants. A linear relationship was observed between the turnover number and the catalytic efficiency. The fit is denoted by the black dashed line (--).The gray dashed lines are visual guides for the values associated with SsWT.

Similar to the determination of ΔΔG for protein stability change, the difference in catalytic rate between SsWT and the mutant can be used to calculate the change in the activation energy barrier, $\Delta\Delta G^{\ddagger}$. A positive value in $\Delta\Delta G^{\ddagger}$ indicates a lower energy barrier associated with the mutation. Interestingly, I45A and S70A had opposite effects on both the $\Delta\Delta G^{\ddagger}_{kcat}$ and $\Delta\Delta G^{\ddagger}_{keff}$ (Fig 3.11). I45A had a lower activation energy if we consider the whole enzymatic process, but had a higher activation energy for the enzyme chemistry. The reverse was true for S70A. The double mutant I45A/S70A received both benefits, having faster product turnover as well as greater catalytic efficiency. The non-zero interaction energies suggest that there are interactions between both the I45/S70 pair and the I45/M73 pair (Table 3.4). Intuitively, mutations that impact enzyme efficiency likely affect fitness. We found that the selection coefficient increases linearly with enzymatic efficiency (R = .87).

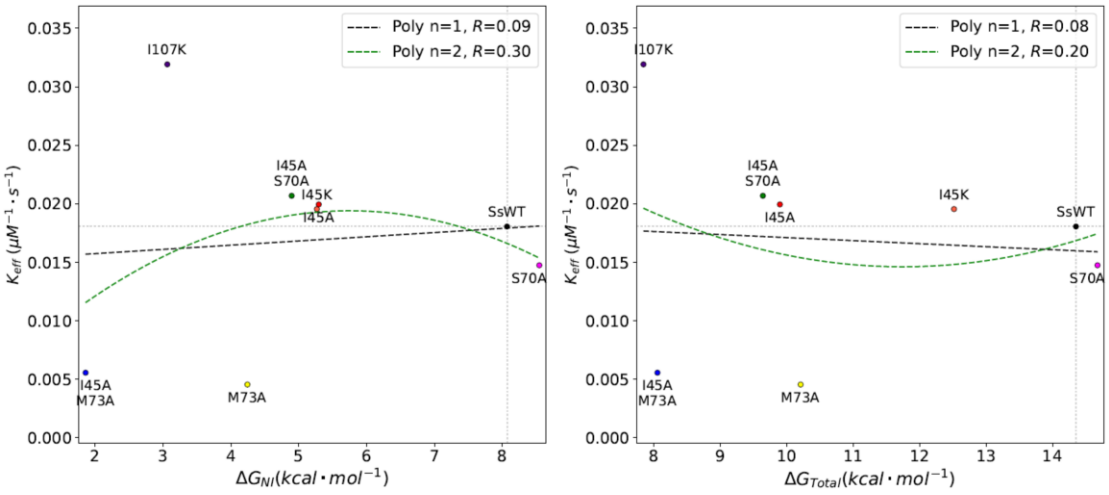
Figure 3.11 Decreased activation energy associated with some mutations



The $\Delta\Delta G^{\ddagger}_{kcat}$ and $\Delta\Delta G^{\ddagger}_{keff}$ associated with mutation are plotted on a bargraph. A positive $\Delta\Delta G$ is indicative of reduced activation energy.

To bridge the biophysical properties with enzyme activity, we examined the catalytic efficiency as a function of native state stability. In this stability performance curve for activity, we observe optimal activity around ΔG_{NI} ~6 kcal·mol^{-1}, which is on par with the stability performance curve for fitness. Enzyme performance is harder to predict when we consider overall stability.

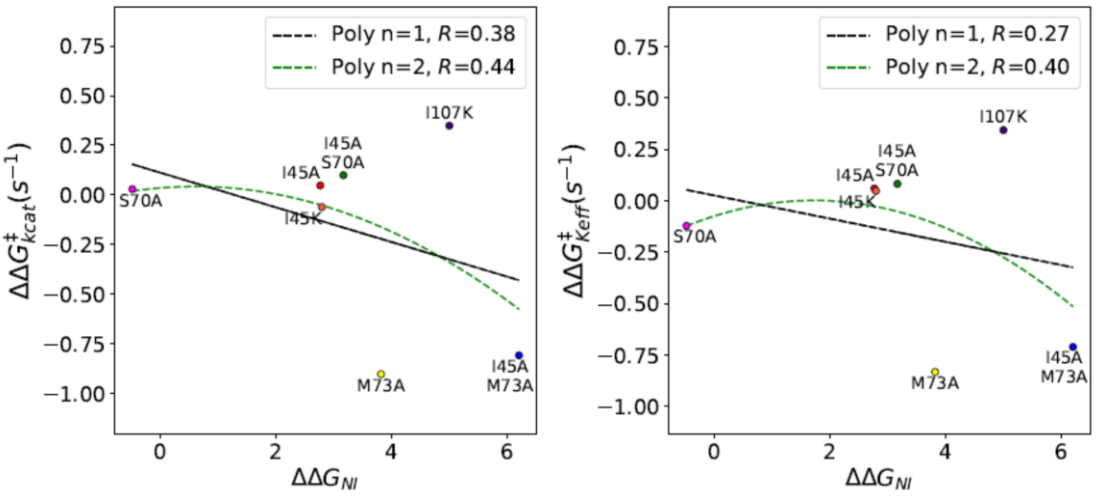
Figure 3.12 A non-linear relationship is observed between ΔG_{NI} and k_{eff}



The k_{eff} of the SsIGPS variants were plotted against their stabilities, k_{eff} vs ΔG_{NI} (left) and k_{eff} vs ΔG_{Total} (right). Comparison between a linear (--) and quadratic (--) model showed a better fit with the 2nd order polynomial regression for ΔG_{NI} and for ΔG_{Total}. The fit is denoted by the black dashed line (--).The gray dashed lines are visual guides for the values associated with SsWT.

Due to the cold temperature stress of SsIGPS within the yeast host, we questioned whether changing the native state stability lowered the activation energy of the enzyme reaction. We found that loss of several kcal·mol^{-1} of native state stability may result in lower activation energy to a point, but a great loss may result in the opposite effect of increasing the activation energy (Fig. 3.13).

Figure 3.13 A non-linear relationship is observed between $\Delta\Delta G_{NI}$ and $\Delta\Delta G^{\ddagger}$



The $\Delta\Delta G^{\ddagger}$ of the SsIGPS mutants were plotted against their $\Delta\Delta G_{NI}$, $\Delta\Delta G^{\ddagger}_{kcat}$ vs $\Delta\Delta G_{NI}$ (left) and $\Delta\Delta G^{\ddagger}_{keff}$ vs $\Delta\Delta G_{Total}$ (right). Comparison between a linear (--) and quadratic (--) model showed a better fit with the 2nd order polynomial regression for both $\Delta\Delta G^{\ddagger}_{kcat}$ and $\Delta\Delta G^{\ddagger}_{keff}$. The fit is denoted by the black dashed line (--).

The sequence of SsWT enzyme has evolved to perform efficiently in extremely hot environments. Within the host cell, complementation of the IGPS knockout condition by SsWT results in reduced fitness compared to yeast ScIGPS, suggesting that SsWT may be working suboptimally. While the chemical process is the same between the orthologs, the rate limiting step in catalysis differs. In fact, for SsIGPS, the rate determining step for catalysis is temperature-dependent. Long loops at the C-terminus of the β-barrel protect the active site from the solvent environment, but also limit substrate access to the active site[134].

At mesophilic temperature, displacement of these loops permitting product release is rate determining[135,136], while at the optimal growth temperature of the thermophilic archeon, the chemical reaction is rate limiting[44,137].

Mutations have a rippling effect that begin at the molecular level and propagates to fitness at the organism level. Using mutational analysis, we dissected the effects of native state stability and catalytic efficiency on fitness. In our system, there is an optimum level of native state destabilization that may reduce the energy barrier for enzyme catalysis, improving fitness. Nonadditivity of fitness, stability, and activity within double mutant cycles suggests that α1 communicates with both the hairpin clamps in the αβ-loops and the active site near the βα-loops. The interaction between β1α1 and β2α2 loops have been shown to have functional consequences on catalysis[138,139]. The hydrogen bond between R54 and N90 in the first and second loop, respectively, coordinates the microenvironment created by active site residues K51 and K53 and the π-π interaction between F89 and the anthranilate moiety of CdRP[138,139]. At mesophilic temperatures, these two long loops are less flexible and strong electrostatic interactions between K53 and the C1 and C2' of CdRP result in an unproductive extended substrate conformation[92,132]. At higher temperatures, a larger population of K53 is found to have a greater distance between K53 and the substrate, resulting in a productive orientation of CdRP[92,132].

Taken together, we developed a mechanistic model for the observed allosteric beneficial fitness effects of the βα-hairpin clamp and non-canonical clamps in the αβ-loops. Mutation of the clamp destabilizes the native state, perturbing the canonical helix placement and the active site upstream of the helix via the long βα-loops forming the microenvironment of the active site. This displacement lowers the activation energy by orienting the active site in a catalytic competent state, improving catalytic efficiency by increasing the maximal turnover rate. Minimal changes in K_m suggest that substrate binding is not impacted with these mutations. The phosphate binding pocket is primarily associated with the β7α7 and β8α8 loops[44,91].

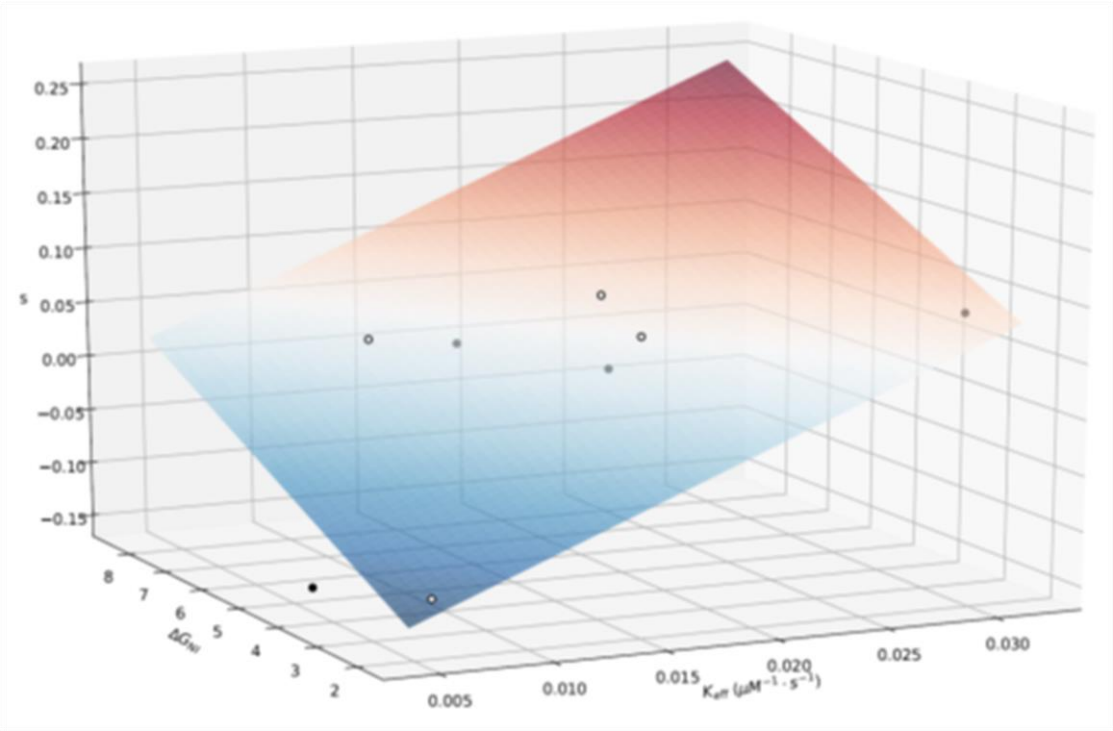
The outcome of this mutant population with the enhanced catalytic capabilities is the observed beneficial fitness over SsWT. A non-linear relationship was found between the native state stability and catalytic efficiency. A summary of all reported values is provided in Table 3.5. Using multiple regression, we created a quantitative model to describe the fitness landscape as a function of the thermodynamic energy and enzyme kinetic landscapes (Fig. 3.14). As we continue studying this system, additional features will continue improving our understanding of this complex relationship.

Table 3.5 Experimentally determined fitness, stability, and kinetic values for SsIGPS variants are listed for point mutations (top) and double mutations (bottom)

Point mutation	Doubling time	MRE_{222} 22°C	MRE_{222} 30°C	ΔG_{NI}	ΔG_{IU}	ΔG_{Total}	kcat (s^{-1})	Km (μM)	kcat/km ($\mu M^{-1} \cdot s^{-1}$)
SsWT	3.80	-11737	-10899	8.07±0.24	6.29±0.55	14.35±0	13.79±0.58	764.4±98.31	0.018±0.0024
I45A	3.62	-9546	-8919	5.29±0.06	4.6±0.12	9.9±-2.77	14.92±0.67	748.3±103.71	0.0199±0.0029
I45K	3.32	-10974	-10504	5.27±0.1	7.25±0.75	12.52±-2.8	12.44±0.33	637.01±51.28	0.0195±0.0017
S70A	3.71	-11947	-11701	8.54±0.11	6.13±0.66	14.68±0.48	14.46±0.53	981.8±87.32	0.0147±0.0014
M73A	4.30	-6779	-2248	4.24±0.18	5.97±0.17	10.21±-3.82	3.07±0.09	678.16±55.24	0.0045±0.0004
M73I	-	-	-	-	-	-	5.07±0.16	805.15±66.57	0.0063±0.0006
I107K	3.36	-	-	3.06±0.04	4.78±0.08	7.84±-5.01	24.62±1.01	771.35±88.87	0.0319±0.0039
I45A_S70A	3.47	-11135	-10612	4.89±0.14	4.75±0.28	9.64±-3.17	16.26±0.58	786.54±84.44	0.0207±0.0023
I45A_M73A	4.18	-7356	-5051	1.86±0.23	6.19±0.3	8.05±-6.21	3.59±0.16	648.59±88.77	0.0055±0.0008

Point mutation	*s*	$\Delta\Delta G_{NI}$	$\Delta\Delta G_{IU}$	δ_{NI} ($kcal \cdot mol^{-1}$)	δ_{IU} ($kcal \cdot mol^{-1}$)	$\Delta\Delta G^{\ddagger}_{kcat}$ ($kcal \cdot mol^{-1}$)	$\Delta\Delta G^{\ddagger}_{keff}$ ($kcal \cdot mol^{-1}$)	δ_{kcat} ($kcal \cdot mol^{-1}$)	δ_{keff} ($kcal \cdot mol^{-1}$)
I45A	0.042	2.77±0.24	3.47±0.27			0.047	0.060		
I45K	0.081	2.8±0.26	0.82±0.79			-0.062	0.048		
S70A	0.022	-0.48±0.26	1.93±0.71			0.028	-0.122		
M73A	-0.142	3.82±0.3	2.1±0.29			-0.905	-0.833		
M73I	-0.229					-0.602	-0.633		
I107K	0.075	5.01±0.24	3.29±0.25			0.349	0.343		
I45A_S70A	0.088	3.17±0.27	3.32±0.37	0.88±0.45	-2.08±0.84	0.099	0.082	0.023	0.144
I45A_M73A	-0.114	6.21±0.33	1.87±0.38	-0.39±0.51	-3.69±0.55	-0.810	-0.711	0.047	0.061

Figure 3.14 Model of fitness as a function of stability and activity



The selection coefficient is plotted as a multiple regression of ΔG_{NI} and k_{eff}. The surface plot is colorcoded from high (red) to low (blue) fitness. Scatter points of the actual values are denoted as above (○) or below (●) the surface.

Methods

Yeast strain and culture conditions

S. cerevisiae strain BY4742 Δ*IGPS::KanMX* was produced using the same PCR-generated deletion strategy described by the Saccharomyces Genome Deletion Project[124]. The last 810 bp of the TRP3 gene encoding IGPS were replaced with the KanMX gene. Deletion of IGPS with the KanMX gene was confirmed by Sanger sequencing.

The same expression vector used in the previously described EMPIRIC fitness assay was used for the individual growth assays. The pRS416 vector carrying the auxotrophic URA3 marker was a gift from Daniel Bolon's lab. A lower expressing constitutive Tma19 promoter, also provided by the Bolon laboratory, was used to increase the sensitivity of the fitness assay. Three silent mutations were introduced into the plasmid at the URA3 marker, the ampicillin resistance marker, and the Tma19 promoter to disrupt BSAI recognition sites. The BSAI enzyme was used to create the saturating mutagenesis libraries.

The wildtype SsIGPS gene was purchased from Genscript. An N-terminal 6 × His tag and Tev protease recognition site were added for protein purification and protein abundance measurements. The non-canonical N-terminal α00 (residues 1 to 26) of each gene was deleted to reduce aggregation during refolding of purified proteins. To prevent non-specific cleavage when using Tev protease, position 18 was mutated from arginine to serine. Point mutations were

introduced into the wildtype background through site-directed mutagenesis using the recommended PCR protocol for Phusion® High Fidelity DNA polymerase.

Yeast transformation was performed using the LiAc/SS carrier DNA/PEG method as described by Gietz and Schiestl[125]. Yeast cells were grown in rich media with G418 to select for IGPS knockout yeast. Transformed cells were selected on synthetic minimal media lacking uracil. Selection for IGPS activity was achieved through growth of transformed yeast (one plasmid library per culture) in synthetic drop-out medium lacking tryptophan. All growth experiments were performed at 30 °C. Liquid cultures were maintained in log phase throughout the fitness assay by periodic dilution. G418 selection was maintained throughout the growth. Fitness, w_i, was calculated as the slope of the log2 relative abundance of the mutant to WT versus time over 6 to 10 time points spread over 6 to 8 WT doubling. Selection coefficients, s, were calculated as $1 - w_i$ and normalized to the vector only control $-(s_i/s_{vector})$.

Protein expression and purification

The SsIGPS construct used for the fitness assay was cloned into the pGS21a expression vector using restriction sites, SacI and BAMHI. Plasmids were transformed into *E. coli* strain BL21 Codonplus®(DE3)RIL from Agilent or NiCo21(DE3) from NEB. Protein expression was induced with 1 mM IPTG for four hours. All proteins were isolated from inclusion bodies by dissolving the insoluble fraction of the cell lysate in 10 M urea and 10 mM imidazole, followed by sonication. Cell debris was removed by centrifugation and the soluble fraction

bound to nickel resin for one hour at room temperature. The protein-bound resin was washed with 10 column volume of the same resuspension buffer. The protein was eluted with a step gradient of 20 mM and 250 mM imidazole. Pure fractions were pooled and dialyzed into 10 mM KPi, pH 7.2. The protein was further purified through a gradient elution off the Q-column. Protein purity was confirmed by Coomassie-staining of SDS-PAGE gels.

Protein sequence

MSGSHHHHHHSSDIENLYFQGQRPIISLNERILEFNKSNITAIIAEYKRKSPSGLD
VERDPIEYSKFMERYAVGLSILTEEKYFNGSYETLRKIASSVSIPILMKDFIVKESQ
IDDAYNLGADTVLLIVKILTERELESLLEYARSYGMEPLIEINDENDLDIALRIGARF
IGINSRDLETLEINKENQRKLISMIPSNVVKVAESGISERNEIEELRKLGVNAFLIG
SSLMRNPEKIKEFIL*

Circular dichroism structure analysis

Far-UV CD spectra were collected on a JASCO model J810 CD spectrophotometer. All samples were buffered with 10 mM KPi, pH 7.2 at 30°C. Measurements were taken in a 0.5 cm path length curvette with a bandwidth of 2.5 nm and a step size of 0.5 nm from 210 to 260 nm. Three spectra were collected, averaged, and buffer subtracted for each sample with a total averaging time of 3s per wavelength. A protein concentration of ~3 μM was used for wavelength scans.

Equilibrium unfolding

Unfolded (~10 M buffered urea) and folded (10 mM KPi, pH 7.2 buffer) stocks of 3 µM protein were mixed to yield samples with titrated concentrations of urea from 0 M to 9 M. Each sample was thoroughly mixed and equilibrated overnight in a 30°C incubator. CD spectra were collected as before. CD data were globally fit to a three state model, N ⇋ I ⇋ U, as a function of urea using custom in-house Savuka software.

Thermal melts

Protein samples were made by combining to a final concentration the following: 20 µM native protein and 10X SYPRO red in 10 mM KPi, pH 7.2 to a total volume size of 20 µL. Samples were run in triplicates and organized on a 96-well white PCR plate. A BioRad CFX96 Touch Real-Time PCR Detection System was used for thermal melting and for fluorescence measurements. A temperature ramp rate of 1°C·min^{-1} was used. The fluorescence in each well was collected at the same rate using the Texas Red channel (excitation: 560-590 nm, emission: 610-650 nm). Fluorescence reads were normalized and the replicates averaged before the first derivative of each trace was calculated. The melting temperature is determined as the temperature at which the first derivative of the melting curve is at the maximum.

Enzyme kinetic assays

Protein samples were made by combining to a final concentration of the following: 1 µM native protein and varying concentrations of CdRP [0-1500 µM] in 10 mM KPi, pH 10 mM to a total volume size of 100 µL. Twelve samples of

with varying substrate concentrations were organized on a 96-well black flat bottom plate. Prior to reaction initiation, proteins were incubated at 30°C. The chemical reaction is initiated upon addition of CdRP and the samples read immediately in a Tecan Infinite M1000 Pro plate reader, preheated to 30°C. A single orbital shake lasting 1 s at 306 rpm is applied before the first read. The fluorescence in each well was read for 80 kinetic cycles at an excitation of 280/5 nm and emission 348/5 nm. The initial velocity is calculated as the RFU over time. The kinetic parameters were determined using the Michaelis-Menten equation (1), where v_0 is the initial velocity, v_m is the maximal rate of the reaction, [s] is the concentration of substrate, and K_m is the Michaelis constant.

$$(1)\ v_0 = v_m * \left(\frac{[s]}{[s]+K_m}\right)$$

Multiple regression model

The multiple regression model describes the selection coefficient as weighted sum of the catalytic efficiency and native state stability, equation (2):

$$(2)\ s = \beta_0 + \beta_1 k_{eff} + \beta_2 \Delta G_{NI} + \beta_3 (\Delta G_{NI})^2$$

Coefficient values were minimized using ordinary least square (β_0= -0.218, β_1 = 8.03, β_2 = 0.0254, β_3 = -0.0017).

Data availability

All data and scripts for data analyses are available at https://github.com/yvehchan/TIM_EMPIRIC.

Chapter IV – Discussion

Summary

The goal of this research is to elucidate the role of sequence on protein stability and activity. We selected the TIM barrel as our model system for several reasons. The canonical $(\beta\alpha)_8$ structure is an ancient, highly represented fold achieved by a considerable diversity of sequences. Members of the TIM barrel fold are found in all three superkingdoms and catalyze diverse reactions. The stable platform with a segregated active site makes an attractive target for protein engineering. The use of orthologs permitted us to constrain protein structure and function, while exploring the sequence space. IGPS is a well-studied TIM barrel with a single function, whose enzyme chemistry has been previous characterized. Crystal structures of several orthologs structures have been determined to high resolution and the folding mechanism has been examined in detail.

Using the EMPIRIC deep mutational scanning approach, the fitness landscapes of orthologous TIM barrel fold proteins were experimentally determined. Fitness effects of point mutations in three phylogenetically divergent IGPS proteins during adaptation to temperature stress were probed by auxotrophic complementation of yeast with prokaryotic, thermophilic IGPS. A combination of sequence, structural, and bioinformatics analyses of the 5,040 mutations of three extant orthologs demonstrated that the fitness landscapes are significantly correlated to each other, despite low sequence identity ~30-40% and

deeply separated bacterial *vs.* archaeal origins. Fold geometry and stability elements, as well as sequence conservation and amino acid biochemistry, all imposed measurable constraints on the fitness landscape. Principal component analyses (PCA) was performed to determine if experimentally derived fitness values can be extrapolated to naturally evolved proteins; commonality between sources of fitness variance were detected in the three orthologs and in the amino acid usage derived from a multiple sequence alignment (MSA) of IGPS.

Two related conclusions were drawn from this study. First, conservation of amino acid preferences does persist across low sequence identity and, at least, two phylogenetic domains. In a complementary way, this conservation can be interpreted as translocation of fitness landscapes in sequence space: fitness landscapes of single point mutants can be successfully translocated to a different starting point in sequence space. Additionally, an unanticipated discovery of several beneficial point mutations, distal from the active site and mirrored throughout the four-fold protein symmetry, revealed the important interplay between sequence and structure. Statistical coupling analysis (SCA) of 537 IGPS sequences supported evolutionary correlations between the active site and positions of the distal beneficial mutations.

We investigated the molecular mechanism of these beneficial mutations through a battery of fitness, biophysical, and biochemical assays on SsWT and select point and double mutants. Point mutations were introduced throughout the

α-helix/β-barrel interface to identify potential crosstalk between residues of the nominal stability and active site loops mediated by members of the α-helices. Previously identified beneficial mutations of SsIGPS from the EMPIRIC screen were confirmed with individual growth assays compared to SsWT. Two helical mutations were used as test cases for residues within the communication network between the opposing ends of the protein.

Most mutations reduce protein stability. For these beneficial and near neutral mutations we examined, the native state appears to be impacted to a greater extent than the intermediate state. Urea and thermal-induced unfolding revealed a linear relationship between ΔG_{NI} and T_m. A non-linear relationship was observed between ΔG_{NI} and *s*, where a broad peak for *s* is found following a slight loss in ΔG_{NI}. Double mutants, I45A/S70A and I45A/M73A, having nonzero δ values suggest the partner pairs are energetic linked. Similar to the selection coefficient, improved catalytic efficiency was improved with reduced ΔG_{NI}. The catalytic efficiency is linearly correlated with k_{cat}, suggesting that decreasing the native state stability may speed up the substrate turnover. Thus, the observed beneficial fitness response may be a product of a single sequence change increasing enzyme function through modification of protein stability.

Conclusion

Protein sequence can greatly impact protein stability. For enzymes, the consequence of changes in stability is reverberated through changes in protein structure and activity on the molecular level. On the cellular level, protein

abundance is maintained and regulated by protein homeostasis systems. Over evolutionary timescales, mutations are introduced by genetic drift and may have no observable phenotype. Once an organism is placed under stress and a particular phenotype placed under selection, mutations may enhance essential cellular functions and greatly influence fitness. Prolonged exposure to selection may cause a mutation to become fixed.

Through a combination of sequence, structural, and bioinformatics analyses, I demonstrated that the fitness landscapes of evolutionary distant orthologs are correlated through sequence and structure space. An unanticipated discovery of several beneficial point mutations, distal from the active site and mirrored throughout the four-fold protein symmetry, revealed a delicate balance between stability, activity and fitness that is modulated by sequence.

Future direction

Numerous paths are possible for this project. The richness of the EMPIRIC dataset combined with the select point and double mutations provide high dimensional data for studying epistasis within a protein. Examination of other double mutations may provide further confirmation of our working model for the molecular mechanism of beneficial allostery. Specifically, mutation I107K resulted in the greatest positive change in *in vitro* activity and showed almost comparable doubling time as the yeast IGPS complementation. Investigating double mutations of I107K with a helix or clamp mutation may provide interesting metrics to the extent of possible beneficial effects on protein activity and

organism fitness. Testing new environmental conditions, such as elevated or depressed temperatures may also add new dimension to this dataset by defining the conditions for which the beneficial effects are observed.

The overarching goal of this project is to understand the role of sequence on stability and other phenotypic manifestations. In parallel with this research, a fellow graduate student Kevin Halloran is studying the effect of sequence on folding and main chain dynamics. A collaboration with a graduate student Sravya Kotaru in Dr. Joshua Wand's lab at UPenn is examining the role of sequence on side chain dynamics. Other potential avenues for exploration include crystallography and MD simulations to obtain high quality structural information and atomistic structural, dynamic, and solvation details, respectively. These experiments along with our current analyses will provide fundamental knowledge for mapping genotype to phenotype to the TIM barrel superfamily.

Perspective

Evolutionary pathways of fitness landscapes

The evolutionary path of a protein is guided by constraints imposed by the sequence and structure, manifested through protein stability, dynamics, function, expression, and numerous other characteristics. Proteins evolve through adoption of mutations that lead to new functions. Point mutation fitness landscapes such as the ones generated here describe the genotype-phenotype relationship of a given system, forecasting potential divergent points and dead ends. While robustness to evolution is dictated in part by stability capable of

tolerating new mutations, epistasis has confounded predictability of protein evolution. Generation of double or multiple mutation fitness maps would provide a mechanistic understanding of the observed epistasis. Within the context of a protein fold, correlation of fitness landscapes may provide general principles that identifies key features restricting trajectories accessibility and, in a complementary way, illuminating access of paths towards new structural fold.

Medicine

Protein misfolding and aggregation are the underlying mechanism for several neurodegenerative diseases such as Alzheimer's and ALS as well as for certain cancers. Point mutations may cause changes in the stability such that the protein may not fold or the structure can no longer by maintained. A basic understanding of how sequence relates to protein stability will elucidate key features leading to successful protein folding and provide useful guidelines for drug targeting.

Protein engineering

The ability to fine-tune protein stability through sequence manipulation is a holy grail of protein engineering. Protein stability is relevant for therapeutic and industrial applications. Biopharmaceuticals seek methods to increase protein stability for practical storage conditions, enhanced shelf-life, and longer half-life. Enzymes are found in common commercial products such as detergents and fermented foods. Enhanced protein stability may increase efficiency of enzymatic activity under harsh conditions such as elevated temperatures or non-neutral pH,

resulting in increased throughput and decreased byproducts. As one of the most common motifs in biology with highly tractable features, the TIM barrel fold is a prime candidate for protein engineering. Moreover, a stable platform such as the TIM barrel permits limitless possibilities for microbial technology development through engineering of enzymes with novel activity.

Basic research aimed at understanding the relationship between protein sequence and stability can have a big impact on our daily lives.

References

1. Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12,** 138–163 (1958).
2. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science (80-.).* **181,** 223–230 (1973).
3. Levinthal, C. Are there pathways for protein folding? *J. Chim. Phys.* **65,** 44–45 (1968).
4. Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nat. Struct. Mol. Biol.* **4,** 10–19 (1997).
5. Pace, C. N. Conformational stability of globular proteins. *Trends Biochem. Sci.* **15,** 14–17 (1990).
6. Sharp, K. A. & Englander, S. W. How much is a stabilizing bond worth? *Trends in Biochemical Sciences* **19,** 526–529 (1994).
7. Fersht, A. R. Conformational equilibria in α- and δ-chymotrypsin. The energetics and importance of the salt bridge. *J. Mol. Biol.* **64,** 497–509 (1972).
8. Pauling, L. *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry. Journal of the American Chemical Society* **82,** (1960).
9. Nick Pace, C., Martin Scholtz, J. & Grimsley, G. R. Forces stabilizing proteins. *FEBS Letters* **588,** 2177–2184 (2014).
10. Schell, D., Tsai, J., Scholtz, J. M. & Pace, C. N. Hydrogen bonding increases packing density in the protein interior. *Proteins Struct. Funct. Genet.* **63,** 278–282 (2006).
11. Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry* **29,** 7133–7155 (1990).
12. Gangadhara, B. N., Laine, J. M., Kathuria, S. V., Massi, F. & Matthews, C. R. Clusters of branched aliphatic side chains serve as cores of stability in the native state of the HisF TIM barrel protein. *J. Mol. Biol.* **425,** 1065–1081 (2013).
13. Kathuria, S. V., Chan, Y. H., Nobrega, R. P., Özen, A. & Matthews, C. R. Clusters of isoleucine, leucine, and valine side chains define cores of stability in high-energy states of globular proteins: Sequence determinants of structure and stability. *Protein Sci.* **25,** 662–675 (2016).
14. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132 (1982).
15. Rose, G. D. & Wolfenden, R. Hydrogen Bonding, Hydrophobicity, Packing, and Protein Folding. *Annu. Rev. Biophys. Biomol. Struct.* **22,** 381–415 (1993).
16. Pace, C. N., Trevino, S., Prabhakaran, E. & Scholtz, J. M. Protein structure, stability and solubility in water and other solvents. *Philos. Trans. R. Soc. London Ser. B-Biological Sci.* **359,** 1225–1234 (2004).

17. Foulkes-Murzycki, J. E., Scott, W. R. P. & Schiffer, C. a. Hydrophobic sliding: a possible mechanism for drug resistance in human immunodeficiency virus type 1 protease. *Structure* **15,** 225–233 (2007).
18. Das, P., Kapoor, D., Halloran, K. T., Zhou, R. & Matthews, C. R. Interplay between drying and stability of a TIM barrel protein: A combined simulation-experimental study. *J. Am. Chem. Soc.* **135,** 1882–1890 (2013).
19. Wu, Y., Vadrevu, R., Kathuria, S., Yang, X. & Matthews, C. R. A Tightly Packed Hydrophobic Cluster Directs the Formation of an Off-pathway Sub-millisecond Folding Intermediate in the ?? Subunit of Tryptophan Synthase, a TIM Barrel Protein. *J. Mol. Biol.* **366,** 1624–1638 (2007).
20. Gu, Z., Rao, M. K., Forsyth, W. R., Finke, J. M. & Matthews, C. R. Structural Analysis of Kinetic Folding Intermediates for a TIM Barrel Protein, Indole-3-glycerol Phosphate Synthase, by Hydrogen Exchange Mass Spectrometry and Gō Model Simulation. *J. Mol. Biol.* **374,** 528–546 (2007).
21. Gu, Z., Zitzewitz, J. A. & Matthews, C. R. Mapping the Structure of Folding Cores in TIM Barrel Proteins by Hydrogen Exchange Mass Spectrometry: The Roles of Motif and Sequence for the Indole-3-glycerol Phosphate Synthase from Sulfolobus solfataricus. *J. Mol. Biol.* **368,** 582–594 (2007).
22. MAYNARD SMITH, J. Natural Selection and the Concept of a Protein Space. *Nature* **225,** 563–564 (1970).
23. Kimura, M. *The Neutral Theory of Molecular Evolution*. *The neutral theory of molecular evolution* (1983). doi:citeulike-article-id:4441469
24. Bershtein, S., Serohijos, A. W. & Shakhnovich, E. I. Bridging the physical scales in evolutionary biology: from protein sequence space to fitness of organisms and populations. *Current Opinion in Structural Biology* **42,** 31–40 (2017).
25. Tokuriki, N., Stricher, F., Serrano, L. & Tawfik, D. S. How protein stability and new functions trade off. *PLoS Comput. Biol.* **4,** (2008).
26. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 5869–5874 (2006).
27. Gilson, A. I., Marshall-Christensen, A., Choi, J. M. & Shakhnovich, E. I. The Role of Evolutionary Selection in the Dynamics of Protein Structure Evolution. *Biophys. J.* **112,** 1350–1365 (2017).
28. Koonin, E. V., Wolf, Y. I. & Karev, G. P. The structure of the protein universe and genome evolution. *Nature* **420,** 218–223 (2002).
29. Tokuriki, N. & Tawfik, D. S. Protein Dynamism and Evolvability. *Science (80-.).* **324,** 203–207 (2009).
30. Tóth-Petróczy, Á. & Tawfik, D. S. The robustness and innovability of protein folds. *Current Opinion in Structural Biology* **26,** 131–138 (2014).
31. Tokuriki, N., Oldfield, C. J., Uversky, V. N., Berezovsky, I. N. & Tawfik, D. S. Do viral proteins possess unique biophysical features? *Trends Biochem. Sci.* **34,** 53–59 (2009).

32. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6,** 197–208 (2005).
33. Yeates, T. O. Protein Structure: Evolutionary Bridges to New Folds. *Current Biology* **17,** (2007).
34. Yue, K. & Dill, K. a. Inverse protein folding problem: designing polymer sequences. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 4163–4167 (1992).
35. Caetano-Anollés, G. & Caetano-Anollés, D. An evolutionarily structural universe of protein architecture. *Genome Research* **13,** 1563–1571 (2003).
36. Murzin, A. G., Lesk, A. M. & Chothia, C. Principles determining the structure of β-sheet barrels in proteins I. A theoretical analysis. *J. Mol. Biol.* **236,** 1369–1381 (1994).
37. Nagano, N., Orengo, C. A. A. & Thornton, J. M. M. *One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions*. *Journal of Molecular Biology* **321,** 741–765 (2002).
38. Zitzewitz, J. A., Gualfetti, P. J., Perkons, I. A., Wasta, S. A. & Matthews, C. R. Identifying the structural boundaries of independent folding domains in the alpha subunit of tryptophan synthase, a βα barrel protein. *Protein Sci.* **8,** 1200–9 (1999).
39. Fox, N. K., Brenner, S. E. & Chandonia, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.* **42,** D304–D309 (2014).
40. Copley, R. R. & Bork, P. Homology among (βα) 8 barrels: implications for the evolution of metabolic pathways 1 1Edited by G. Von Heijne. *J. Mol. Biol.* **303,** 627–641 (2000).
41. Sillitoe, I. *et al.* CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* **43,** D376–D381 (2015).
42. Goldman, A. D., Beatty, J. T. & Landweber, L. F. The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism. *J. Mol. Evol.* **82,** 17–26 (2016).
43. Creighton, T. E. & Yanofsky, C. Indole-3-glycerol phosphate synthetase of Escherichia coli, an enzyme of the tryptophan operon. *J. Biol. Chem.* **241,** 4616–4624 (1966).
44. Zaccardi, M. J., Yezdimer, E. M. & Boehr, D. D. Functional identification of the general acid and base in the dehydration step of indole-3-glycerol phosphate synthase catalysis. *J. Biol. Chem.* **288,** 26350–26356 (2013).
45. Bentley, R. & Haslam, E. The Shikimate Pathway — A Metabolic Tree with Many Branche. *Crit. Rev. Biochem. Mol. Biol.* **25,** 307–384 (1990).
46. Akashi, H. & Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 3695–3700 (2002).
47. Yanofsky, C. Attenuation in the control of expression of bacterial operons.

Nature **289,** 751–758 (1981).
48. Gong, F., Ito, K., Nakamura, Y. & Yanofsky, C. The mechanism of tryptophan induction of tryptophanase operon expression: tryptophan inhibits release factor-mediated cleavage of TnaC-peptidyl-tRNA(Pro). *Proc. Natl. Acad. Sci. U. S. A.* **98,** 8997–9001 (2001).
49. Merkl, R. Modelling the evolution of the archeal tryptophan synthase. *BMC Evol. Biol.* **7,** 59 (2007).
50. Hütter, R., Niederberger, P. & DeMoss, J. A. Tryptophan biosynthetic genes in eukaryotic microorganisms. *Annu. Rev. Microbiol.* **40,** 55–77 (1986).
51. Braus, G. H. Aromatic amino acid biosynthesis in the yeast Saccharomyces cerevisiae: a model system for the regulation of a eukaryotic biosynthetic pathway. *Microbiol. Rev.* **55,** 349–70 (1991).
52. Miozzari, G., Niederberger, P. & Huetter, R. Tryptophan biosynthesis in Saccharomyces cerevisiae: control of the flux through the pathway. *J. Bacteriol.* **134,** 48–59 (1978).
53. Hennig, M., Darimont, B., Sterner, R., Kirschner, K. & Jansonius, J. N. 2.0 å structure of indole-3-glycerol phosphate synthase from the hyperthermophile Sulfolobus solfataricus: possible determinants of protein stability. *Structure* **3,** 1295–1306 (1995).
54. Knochel, T., Pappenberger, A., Jansonius, J. N. J. . & Kirschner, K. The crystal structure of indoleglycerol-phosphate synthase from Thermotoga maritima. Kinetic stabilization by salt bridges. *J. Biol. Chem.* **277,** 8626–8634 (2002).
55. Bagautdinov, B. & Yutani, K. Structure of indole-3-glycerol phosphate synthase from Thermus thermophilus HB8: Implications for thermal stability. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67,** 1054–1064 (2011).
56. Wilmanns, M., Priestle, J. P., Niermann, T. & Jansonius, J. N. Three-dimensional structure of the bifunctional enzyme phosphoribosylanthranilate isomerase: Indoleglycerolphosphate synthase from Escherichia coli refined at 2.0 Å resolution. *J. Mol. Biol.* **223,** 477–507 (1992).
57. Eberhard, M., Tsai-Pflugfelder, M., Bolewska, K. & Hommel, U. Indoleglycerol Phosphate Synthase—Phosphoribosyl Anthranilate Isomerase: Comparison of the Bifunctional Enzyme from Escherichia coli with Engineered Monofunctional Domains. *Biochemistry* **34,** 5419–5428 (1995).
58. Cadwell, R. C. & Joyce, G. F. Randomization of genes by PCR mutagenesis. *Genome Res.* **2,** 28–33 (1992).
59. Yu, Y. & Lutz, S. Circular permutation: A different way to engineer enzyme structure and function. *Trends in Biotechnology* **29,** 18–25 (2011).
60. Höcker, B. Directed evolution of (betaalpha)(8)-barrel enzymes. *Biomol. Eng.* **22,** 31–8 (2005).
61. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein

science. *Nat. Methods* **11,** 801–807 (2014).
62. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. a. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 7896–7901 (2011).
63. Bank, C., Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. A Systematic Survey of an Intragenic Epistatic Landscape. *Molecular Biology and Evolution* **32,** 229–238 (2015).
64. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a Function of Purifying Selection in TEM-1 β-Lactamase. *Cell* **160,** 882–892 (2015).
65. Starita, L. M. *et al.* Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics* **200,** 413–422 (2015).
66. Doud, M. B. & Bloom, J. D. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses* **8,** (2016).
67. Bloom, J. D. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol. Biol. Evol.* **31,** 1956–1978 (2014).
68. Roscoe, B. P. P., Thayer, K. M. M., Zeldovich, K. B. B., Fushman, D. & Bolon, D. N. A. N. A. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.* **425,** 1363–1377 (2013).
69. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19,** 1537–51 (2013).
70. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533,** 397–401 (2016).
71. Ryan Hietpas*, Benjamin Roscoe*, Li Jiang, and D. N. A. B. Fitness Analyses of All Possible Point-Mutants for Regions of Genes in Yeast. *Nat. Protoc.* **7,** 1382–1396 (2013).
72. Valentini, G. *et al.* The allosteric regulation of pyruvate kinase. *J. Biol. Chem.* **275,** 18145–52 (2000).
73. Latallo, M. J., Cortina, G. A., Faham, S., Nakamoto, R. K. & Kasson, P. M. Predicting allosteric mutants that increase activity of a major antibiotic resistance enzyme. *Chem. Sci.* (2017). doi:10.1039/C7SC02676E
74. Natarajan, C. *et al.* Epistasis Among Adaptive Mutations in Deer Mouse Hemoglobin. *Science (80-.).* **340,** 1324–1327 (2013).
75. Khersonsky, O. & Fleishman, S. J. Incorporating an allosteric regulatory site in an antibody through backbone design. *Protein Sci.* **26,** 807–813 (2017).
76. Lee, J. & Goodey, N. M. Catalytic contributions from remote regions of enzyme structure. *Chemical Reviews* **111,** 7595–7624 (2011).
77. Horovitz, A. Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Fold. Des.* **1,** R121–R126 (1996).
78. Lunzer, M., Brian Golding, G. & Dean, A. M. Pervasive cryptic epistasis in molecular evolution. *PLoS Genet.* **6,** 1–10 (2010).
79. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
80. Doud, M. B., Ashenberg, O. & Bloom, J. D. Site-specific amino acid

preferences are mostly conserved in two closely related protein homologs. *Mol. Biol. Evol.* **32,** 2944–2960 (2015).
81. Bloom, J. D. An experimentally informed evolutionary model improves phylogenetic fit to divergent lactamase homologs. *Mol. Biol. Evol.* **31,** 2753–2769 (2014).
82. Bloom, J. D., Wilke, C. O., Arnold, F. H. & Adami, C. Stability and the Evolvability of Function in a Model Protein. *Biophys. J.* **86,** 2758–2764 (2004).
83. Zeldovich, K. B., Chen, P. & Shakhnovich, E. I. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 16152–7 (2007).
84. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19,** 596–604 (2009).
85. Wylie, C. S. & Shakhnovich, E. I. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 9916–21 (2011).
86. Hietpas, R., Roscoe, B., Jiang, L. & Bolon, D. N. A. Fitness analyses of all possible point mutations for regions of genes in yeast. *Nat. Protoc.* **7,** 1382–1396 (2012).
87. Yang, X., Kathuria, S. V., Vadrevu, R. & Matthews, C. R. βα-Hairpin clamps brace βαβ modules and can make substantive contributions to the stability of TIM barrel proteins. *PLoS One* **4,** (2009).
88. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilites of residues in proteins. *PLoS One* **8,** (2013).
89. Huang, P.-S. *et al.* De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat Chem Biol* **12,** 29–34 (2016).
90. Ochoa-Leyva, A. *et al.* Protein Design through Systematic Catalytic Loop Exchange in the (β/α)8 Fold. *J. Mol. Biol.* **387,** 949–964 (2009).
91. Hennig, M., Darimont, B. D., Jansonius, J. N. & Kirschner, K. The catalytic mechanism of indole-3-glycerol phosphate synthase: Crystal structures of complexes of the enzyme from Sulfolobus solfataricus with substrate analogue, substrate, and product. *J. Mol. Biol.* **319,** 757–766 (2002).
92. Mazumder-Shivakumar, D., Kahn, K. & Bruice, T. C. Computational Study of the Ground State of Thermophilic Indole Glycerol Phosphate Synthase: Structural Alterations at the Active Site with Temperature. *J. Am. Chem. Soc.* **126,** 5936–5937 (2004).
93. Meyer, A. G., Dawson, E. T. & Wilke, C. O. Cross-species comparison of site-specific evolutionary-rate variation in influenza haemagglutinin. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **368,** 20120334 (2013).
94. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* **138,** 774–786 (2009).
95. Vijayabaskar, M. S. & Vishveshwara, S. Insights into the fold organization

of tim barrel from interaction energy based structure networks. *PLoS Comput. Biol.* **8,** (2012).
96. Ovchinnikov, S. *et al.* Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* **4,** e09248 (2015).
97. Jack, B. R., Meyer, A. G., Echave, J. & Wilke, C. O. Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. *PLoS Biol.* **14,** (2016).
98. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7,** 741–6 (2010).
99. Kumar, S., Tsai, C.-J. & Nussinov, R. Factors enhancing protein thermostability. *Protein Eng.* **13,** 179–191 (2000).
100. Szilágyi, A. & Závodszky, P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: Results of a comprehensive survey. *Structure* **8,** 493–504 (2000).
101. Závodszky, P., Kardos, J., Svingor, Á. & Petsko, G. A. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc. Natl. Acad. Sci.* **95,** 7406–7411 (1998).
102. Kalbitzer, H. R., Spoerner, M., Ganser, P., Hozsa, C. & Kremer, W. Fundamental link between folding states and functional states of proteins. *J. Am. Chem. Soc.* **131,** 16714–16719 (2009).
103. Suzuki, T., Yasugi, M., Arisaka, F., Yamagishi, A. & Oshima, T. Adaptation of a thermophilic enzyme, 3-isopropylmalate dehydrogenase, to low temperatures. *Protein Eng. Des. Sel.* **14,** 85–91 (2001).
104. Douangamath, A. *et al.* Structural evidence for ammonia tunneling across the (β/α)8 barrel of the imidazole glycerol phosphate synthase bienzyme complex. *Structure* **10,** 185–193 (2002).
105. Rivalta, I. *et al.* Allosteric pathways in imidazole glycerol phosphate synthase. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E1428–E1436 (2012).
106. Srivastava, D. K. & Bernhard, S. A. Biophysical chemistry of metabolic reaction sequences in concentrated enzyme solution and in the cell. *Annu. Rev. Biophys. Biophys. Chem.* **16,** 175–204 (1987).
107. Farber, G. K. & Petsko, G. A. The evolution of α/β barrel enzymes. *Trends Biochem. Sci.* **15,** 228–234 (1990).
108. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science (80-.).* **312,** 111–114 (2006).
109. Olson, C. A. A., Wu, N. C. C. & Sun, R. A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain. *Curr. Biol.* **24,** 2643–2651 (2016).
110. Hinkley, T. *et al.* A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat. Genet.* **43,** 487–9 (2011).
111. Kondrashov, D. A. & Kondrashov, F. A. Topological features of rugged fitness landscapes in sequence space. *Trends in Genetics* **31,** 24–33 (2015).

112. Ashenberg, O., Gong, L. I. & Bloom, J. D. Mutational effects on stability are largely conserved during protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 21071–6 (2013).
113. Risso, V. A. *et al.* Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol. Biol. Evol.* **32,** 440–455 (2015).
114. Pollock, D. D., Thiltgen, G. & Goldstein, R. a. Amino acid coevolution induces an evolutionary Stokes shift. *Proc. Natl. Acad. Sci.* **109,** E1352–E1359 (2012).
115. Hart, K. M. *et al.* Thermodynamic System Drift in Protein Evolution. *PLoS Biol.* **12,** (2014).
116. Schneider, B. *et al.* Role of the N-Terminal Extension of the $(\beta\alpha)_8$-Barrel Enzyme Indole-3-glycerol Phosphate Synthase for Its Fold, Stability, and Catalytic Activity †, ‡. *Biochemistry* **44,** 16405–16412 (2005).
117. Bershtein, S. *et al.* Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria. *PLoS Genet.* **11,** (2015).
118. Rodrigues, J. V *et al.* Biophysical principles predict fitness landscapes of drug resistance. *Proc. Natl. Acad. Sci. USA* **113,** E1470-1478 (2016).
119. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15,** 480–490 (2014).
120. Ferretti, L. *et al.* Measuring epistasis in fitness landscapes: The correlation of fitness effects of mutations. *J. Theor. Biol.* **396,** 132–143 (2016).
121. Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the (βα)8-Barrel Enzyme Fold. *Chem. Rev.* **105,** 4038–4055 (2005).
122. Gerlt, J. A. & Babbitt, P. C. Barrels in pieces? *Nat Struct Mol Biol* **8,** 5–7 (2001).
123. Woolfson, D. N. *et al.* De novo protein design: How do we expand into the universe of possible protein structures? *Current Opinion in Structural Biology* **33,** 16–26 (2015).
124. Winzeler, E. A. Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis. *Science (80-.).* **285,** 901–906 (1999).
125. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2,** 31–34 (2007).
126. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
127. Yang, Y., Zhan, J., Zhao, H. & Zhou, Y. A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. *Proteins Struct. Funct. Bioinforma.* **80,** 2080–2088 (2012).
128. Gromiha, M. M., Pujadas, G., Magyar, C., Selvaraj, S. & Simon, I. Locating the Stabilizing Residues in (β/α)8 Barrel Proteins Based on Hydrophobicity, Long-Range Interactions, and Sequence Conservation. *Proteins Struct. Funct. Genet.* **55,** 316–329 (2004).

129. Papadopoulos, J. S. & Agarwala, R. COBALT: Constraint-based alignment tool for multiple protein sequences. *Bioinformatics* **23,** 1073–1079 (2007).
130. Bloom, J. D., Raval, A. & Wilke, C. O. Thermodynamics of neutral protein evolution. *Genetics* **175,** 255–266 (2007).
131. Chan, Y. H., Venev, S. V, Zeldovich, K. B. & Matthews, C. R. Correlation of fitness landscapes from three orthologous TIM barrels originates from sequence and structure constraints. *Nat. Commun.* **8,** 14614 (2017).
132. Mazumder-Shivakumar, D. & Bruice, T. C. Molecular dynamics studies of ground state and intermediate of the hyperthermophilic indole-3-glycerol phosphate synthase. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 14379–84 (2004).
133. Berezovsky, I. N., Zeldovich, K. B. & Shakhnovich, E. I. Positive and negative design in stability and thermal adaptation of natural proteins. *PLoS Comput. Biol.* **3,** 0498–0507 (2007).
134. Rozovsky, S. & McDermott, a E. The time scale of the catalytic loop motion in triosephosphate isomerase. *J. Mol. Biol.* **310,** 259–70 (2001).
135. Schlee, S. *et al.* Kinetic mechanism of indole-3-glycerol phosphate synthase. *Biochemistry* **52,** 132–142 (2013).
136. Merz, A. *et al.* Improving the catalytic activity of a thermophilic enzyme at low temperatures. *Biochemistry* **39,** 880–889 (2000).
137. Zaccardi, M. J., Mannweiler, O. & Boehr, D. D. Differences in the catalytic mechanisms of mesophilic and thermophilic indole-3-glycerol phosphate synthase enzymes at their adaptive temperatures. *Biochem. Biophys. Res. Commun.* **418,** 324–329 (2012).
138. Zaccardi, M. J. *et al.* Loop-loop interactions govern multiple steps in indole-3-glycerol phosphate synthase catalysis. *Protein Sci.* **23,** 302–311 (2014).
139. O'Rourke, K., Jelowicki, A. & Boehr, D. Controlling Active Site Loop Dynamics in the (β/α)8 Barrel Enzyme Indole-3-Glycerol Phosphate Synthase. *Catalysts* **6,** 129 (2016).