Combined use of experimental and computational screens to characterize protein stability

Nora H.Barakat, Nesreen H.Barakat and John J.Love

Department of Chemistry and Biochemistry, San Diego State University, 5500 Campanile Dr, San Diego, CA 92182-1030, USA

Received January 4, 2010; revised July 4, 2010; accepted July 17, 2010

Edited by Stephen Mayo

One of the primary goals of protein design is to engineer proteins with improved stability. Protein stability is a key issue for chemical, biotechnology and pharmaceutical industries. The development of robust proteins/enzymes with the ability to withstand the potentially harsh conditions of industrial operations is of high importance. A number of strategies are currently being employed to achieve this goal. Two particular approaches, (i) directed evolution and (ii) computational protein design, are quite powerful yet have only recently been combined or applied and analyzed in parallel. In directed evolution, libraries of variants are searched experimentally for clones possessing the desired properties. With computational methods, protein design algorithms are utilized to perform in silico screening for stable protein sequences. Here, we used gene libraries of an unstable variant of streptococcal protein G (G\textsubscript{bb}) and an in vivo screening method to identify stabilized variants. Many variants with notably increased thermal stabilities were isolated and characterized. Concomitantly, computational techniques and protein design algorithms were used to perform in silico screening of the same destabilized variant of G\textsubscript{bb}. The combined use, and critical analysis, of these methods promises to advance the field of protein design.

Introduction

A major goal of the field of protein design is to increase the thermal stability of target proteins. Proteins designed with greater thermal stabilities will generally possess enhanced properties such as greater resistance to proteolysis and lower propensity to aggregate; both normally resulting in improved expression profiles. In the case of enzymes, if greater thermal stability does not impair the catalytic properties of the enzyme, then lower concentrations of enzyme can be used in certain applications as the desired properties will increase the enzyme’s half-life. This is especially important for industrial applications where improvements in enzyme stability can lead to economic savings. In a manner similar to the field of protein folding, much progress has been made in understanding the forces that govern protein stability, but the precise relationship among amino acid sequence, structure and stability is still not fully understood. Therefore, considerable effort is currently being made to explore new methods, and enhance existing methods, to increase the thermal stability of target proteins.

Three broad categories of approaches have been applied to increase the stability of proteins: rational-design methods, computational design methods and directed evolution. All these approaches have led to advances in our understanding of the parameters that contribute to protein stability. The methods utilized for each of these categories are not mutually exclusive and, in fact, their combined use is now more common (Voigt et al., 2001; Hayes, 2002; Funke, 2005; Mena et al., 2006; Trenynor et al., 2007; Shivange et al., 2009), and is a major focus of the work reported herein. Rational-design approaches (defined here as design decisions based on the three-dimensional structure of the targeted protein and driven by human interpretation and rationalization) have been practiced the longest and thus the body of literature describing these methods is quite extensive (Hellinga, 1997; Boroschuer and Pohl, 2001; Salvador and Luis, 2004; Bommarius et al., 2006; Mayer et al., 2007; Khoo et al., 2009). Directed evolution (Bloom and Arnold, 2009; Gerlt and Babbitt, 2009; Tracewell and Arnold, 2009) and computational design methods (Lazar et al., 2003; Vizzcarra and Mayo, 2005; Das and Baker, 2008; Alvizo et al., 2007; Alvizo and Mayo, 2008; Lippow and Tidor, 2007; Damborsky and Brezovsky, 2009; Havranek and Baker, 2009; Kim et al. in press) have not been practiced as long but have also made significant contributions to the field of protein design. It is our belief that the continued practice of combining these methods will lead to greater advances in the field of protein design and thus enhance our ability to specifically alter the thermal stability of target proteins.

Research on the stability of proteins, which has particular relevance to this work, includes design efforts performed on small test proteins (i.e. proteins that are less than 100 amino acids). Increased understanding of the forces that drive protein folding and factors that contribute to protein stability has been garnered from the total design of novel folds (Kortemme et al., 1998; Kuhlman, 2003; Bosche, 2009), the total re-design of natural proteins (based on fixed wild-type backbone positions) (Dahiya and Mayo, 1997; Shah, 2007) and partial designs of test proteins that resulted in significantly higher thermal stabilities (Malakauskas and Mayo, 1998; Marshall et al., 2002).

A small test protein that has proved to be a workhorse of the field of protein design is the \( \beta_1 \) domain of streptococcal protein G (G\textsubscript{bb}). The physical properties of this protein, which consists of four \( \beta \)-strands and an overlying \( \alpha \)-helix (\( \beta\beta\alpha \beta \)), are the reason why it has been used extensively in many protein design efforts (Malakauskas and Mayo, 1998; Alexander et al., 2005, 2007; Wunderlich et al., 2005; Wunderlich and Schmid, 2006; Scott and Daggett, 2007; Crowhurst and Mayo, 2008; He et al., 2008; Thoms in press). It is relatively small (56 amino acids) and the melting temperature \( T_m \) of the wild-type G\textsubscript{bb} is fairly high (\(~85^\circ C\)). Its three-dimensional structure has been solved to high resolution...
with both X-ray crystallography and NMR, it expresses well in *Escherichia coli*, and has proven to be quite permissible to mutagenesis while maintaining its overall fold. It is this protein domain that we used to develop the in vivo screen for protein thermal stability described previously by Barakat et al. (2007) and in the work described herein.

The engineered in vivo screen entails the use of a chimeric protein that consists of three domains (or proteins) expressed as one protein. The three domains consist of a constant N-terminal DNA-binding domain (full-length bacteriophage λCI), a variable central test protein (e.g. Gβ1 variants) and a constant C-terminal transcriptional activation domain [the N-terminal domain of the α-subunit of bacterial RNA polymerase (RNAPα)]. The three-domain chimera acts as a transcription factor upon binding the λ operator and results in the up-regulation of transcription for reporter genes (e.g. β-lactamase). The constant N- and C-terminal domains were derived from a bacterial two-hybrid screen (Dove et al., 1997; Dove and Hochschild, 1998; Dove and Hochschild, 2004) (BacterioMatch® Two-Hybrid System, Stratagene).

The system was initially benchmarked with a series of 10 mutant variants of Gβ1 used as the variable central test protein. The 10 variants fairly evenly span a thermal stability ranged from 38°C to >100°C. Variants of lower thermal stability (and presumably greater malleability) up-regulate the reporter genes to a greater extent relative to those of higher thermal stability (and potentially greater rigidity). Therefore, variants of lower thermal stability are able to survive on greater amounts of reporter antibiotic relative to those of higher thermal stability. We believe that the Gβ1 test variants, expressed as the intervening variable domain, act in a sense as ‘molecular rheostats’ in providing variable resistance for the attached domains to possibly achieve the correct dimer orientations necessary for optimal recruitment of the endogenous transcription complex. Stable Gβ1 variants likely maintain the tight ββαββ fold that possibly inhibits and thus blocks optimal dimerization of the other components of the chimeric construct. On the other hand, the less stable variants are more likely to exist in a dynamic equilibrium between partially unfolded states that allow the proper intermolecular interactions necessary to achieve the optimal shape and chemical complementarity for enhanced recruitment of the endogenous RNA polymerase.

In this work we describe 10 additional mutant variants of Gβ1 obtained from the in vivo screen (both stable and unstable variants) and compare their physical parameters to Gβ1 computational designs (and associated energies) generated with the suite of protein design algorithms collectively known as ORBIT (Optimization of Rotamers By Iterative Techniques; Dahiyat and Mayo, 1997).

Materials and methods

Construction of the MonB library

The monomer B (MonB) library of genes that varied at position 23, 27 and 45 was constructed using recursive PCR with overlapping oligonucleotides. Four oligonucleotides corresponding to the gene were synthesized as 40–50 base oligonucleotides with 20-nucleotide overlaps and at least 50% GC content at each overlap region. Two of the oligonucleotides were synthesized with equimolar mixes of the four DNA bases at positions 23, 27 and 45. All other positions of the genes contained codons for the MonB ‘wild-type’ amino acids. Saturation mutagenesis at positions 23, 27 and 45 resulted in a library complexity of 8000 mutant sequences. Standard ligation protocols proved to be inefficient, so the TOPO-TA system was used to first clone the PCR products into TOPO cloning vector (pcR 2.1-TOPO). White colonies (which contain the PCR insert) were grown overnight and then maxi-prepped and double digested with NotI and EcoRI. The resulting library was cloned into the chimeric plasmid (λcI-RNAPα) using standard ligation protocols.

Primary screening of the MonB library on CKC 1000 plates

Fifty microliters of BacterioMatch two-hybrid system reporter strain competent cells (Stratagene) were incubated with the MonB library (random codons at positions 23, 27 and 45). Following standard transformation protocols, 90 μl of the transformation reaction were plated on CKC 1000 plates. The nomenclature ‘CKC 1000’ corresponds to plates that contain chloramphenicol (12.5 μg/ml), kanamycin (50 μg/ml), and carbenicillin (1000 μg/ml). Chloramphenicol is used to maintain the reporter construct on the F' episome, and kanamycin is used for selection of bacteria that contain the plasmid that expresses the three-domain chimeric construct. Upon expression, the three-domain chimeric construct functions to up-regulate the reporter genes located on the F' episome (i.e. β-lactamase and β-galactosidase). The nomenclature ‘CK plates’ corresponds to plates that contain just chloramphenicol and kanamycin but no reporter antibiotic (i.e. carbenicillin). The CK plates were used as controls for colony counts and to estimate the library size (total number of colonies). Small, slow-growing colonies were picked from the CKC 1000 plates and inoculated into 5 ml LB media for the purpose of overnight growth and plasmid recovery.

Secondary screening of the MonB library on CKC 1000 plates

MonB library plasmids that were isolated from slow-growing colonies (the primary step described above) and control plasmids (the genes for MonA, MonB-WT and MonB-ORDES in the chimeric construct) were transformed into the reporter strain competent cells. Transformation reactions were plated onto small CK plates and small CKC 1000 plates. The resulting colonies were compared with both the control CK plates and the control variants (MonA, MonB-WT and MonB-ORDES). This comparison was used to identify variants that grew more slowly than the MonB-WT colonies. Successful variants were used in the secondary screening process (i.e. the 96-well plate format described next) to eliminate false positives.

Tertiary screening of the MonB library in 96-well plates using CKC 2000

The MonB library variants that passed the primary and secondary screens were subjected to a final step of screening using 96-well plates that contained carbenicillin at a concentration of 2000 μg/ml. For each of the selected MonB library variants, six colonies were picked form CK plates and inoculated in 1-ml CK media. Cultures were grown overnight and diluted the next day to an OD 600 of ~0.1 in a volume of 180 μl of LB containing carbenicillin at a concentration of 2000 μg/ml. The 96-well plates were incubated in a
micro-plate reader for 24 h, at 37°C, with continual shaking and OD 600 readings recorded every 15 min. Growth curves data for the six trials for each variant were averaged and compared with the average growth curves of MonA, MonB-WT and MonB-ORDES. The MonB library variants with growth rates equal to or less than MonB-WT were selected for sequencing and for sub-cloning into the expression vector for the purpose of producing pure protein selected for sequencing and for sub-cloning into the vector for the purpose of producing pure protein for circular dichroism (CD) analysis and melting temperature determination.

**Large-scale screening in 10-ml cultures containing CKC 3000**

In addition to the three screening step described above, the variants that met the slow-growth criteria were further analyzed in a large volume format to confirm the slow-growth phenotype. All Gβ1 and MonB library variants in the chimeric construct were transformed into BacterioMatch two-hybrid system reporter strain competent cells and plated on CK plates. Colonies were picked and grown overnight in 20-ml CK media. The next day the overnight cultures were diluted to OD 600 of 0.1 in 10-ml volume of CK media and allowed to grow for 15 min. The OD 600 readings were measured to insure that all cultures had the same OD 600 before adding carbenicillin. Carbenicillin (3000 μg/ml) was then added to all cultures and incubated at 37°C, 275 rpm for 4 h. The OD 600 readings were measured for all cultures after 4 h. The OD readings illustrated in Fig. 2 were obtained from this large volume growth format.

**Computational design**

The ORBIT suite of protein design algorithms by Dahiyat and Mayo (1997) was used to determine optimal amino acid choices, and associated rotameric atomic coordinates, for ‘calculated’ positions in the context of the MonB-WT sequence. The calculated positions for the work reported herein correspond to positions 23, 27 and 45. The process by which the ORBIT algorithms are used to obtain the global minimum energy conformation (GMEC) divides into two overall steps. The first step entails the use of a molecular mechanics force field [i.e. DREIDING, Mayo et al. (1990)] to calculate pair-wise energies for two sets of interaction energies for each rotamer at each calculated position. The first set of interaction energies includes the interaction of the rotamer side chain with the template (i.e. the fixed backbone and fixed side-chain atoms), and the second set the interaction of the rotamer side chain with all other possible rotamers at every other calculated position. Empirical potential energy functions were used to calculate the pair-wise energies with parameters that include, for example, standard scale factors for van der Waals interactions, terms for the solvation and hydrogen bond potentials, a benefit term for burial of nonpolar residues and penalty terms for polar burial and non-polar exposure. After all pair-wise energies are calculated, the second main step entails the application of the dead end elimination algorithm (DEE; Desmet et al., 1992) to eliminate rotamers that are not consistent with the GMEC. A rotamer determined to be incompatible with the GMEC is termed ‘dead ending’ and is eliminated from further consideration. The GMEC is attained through iterative elimination of dead-ending rotamers until only a single rotamer remains at each calculated position.

During the ORBIT calculations on the three calculated positions, all amino acids (except Gly, Pro and Cys) were considered and incorporated backbone-dependent side-chain rotamers derived from the library of Dunbrack and Karplus (Dunbrack and Karplus, 1993). Prior to running the calculation, we determined the identity of 29 residue positions that were in proximity to the three targeted positions. To not bias side-chain selection at the three targeted positions, based on ‘wild-type’ side-chain positioning, the side chains at the additional 29 positions were ‘floated’ (i.e. during the calculation of rotameric representations of the ‘wild-type’ side chains at the 29 positions were included in the calculation). All these residues were fixed in amino acid identity, but their conformations were allowed to change. Amino acid identities and/or side-chain conformations were determined using ORBIT for all 32 residues. All other residues, as well as the backbone, were held fixed. The amino acid sequence, of the GMEC, selected by the ORBIT algorithms contains Lys, Lys and Phe at the respective positions and, upon generation and analysis in the laboratory, resulted in an increase in the $T_m$ from 38 to 64°C (Barakat et al., 2007).

In addition to calculating the lowest energy GMEC sequence, the ORBIT algorithms were also used to run Monte Carlo (MC) calculations using the GMEC sequence (and associated rotameric conformations) as the starting structure. This was necessary because the algorithms used to determine the GMEC sequence do not return the second, third, etc., lowest energy sequences as the amino acid choices (and rotameric descriptions) associated with those sequences may have been eliminated during the step in which the DEE criterion was applied. The MC calculations were run to determine the energies of sequences and structures near in energy to the lowest energy GMEC sequence. The ORBIT algorithm used to run the MC calculations incorporates a standard Metropolis approach by Metropolis et al. (1953) with simulated annealing (Kirkpatrick et al., 1983). The MC algorithm was used to sample sequence space around the GMEC sequence (i.e. 23K, 27K, 45F) using as input the pair-wise energies previously calculated for determining the GMEC sequence. The temperature was cycled 1000 times using 4000 and 150 K as temperature limits and each cycle consisted of 10 000 steps. Random changes to the sequence were kept or discarded according to a probability calculated using a Boltzmann criterion based on the energy difference of the change and the pseudo-temperature at that point of the annealing cycle. Approximately 5000 sequences were obtained and rank ordered according to their calculated energies.

**Protein expression and purification**

For protein expression the genes for the Gβ1 variants were subcloned into the pET-21a plasmid (Novagen) using the restriction enzymes NdeI and EcoRI (oligonucleotides: 5'-GTC CGT CAT ATG ACT ACT TAC AAA-3' and 5'-GGG GCA GAT GAA TTC TTA TTC AGT AAC TGT AAA-3') and transformed into BL21(DE3). After growth and induction the proteins were isolated using a freeze/thaw method and purification was accomplished with reverse-phase HPLC using a linear 1% min$^{-1}$ acetonitrile/water gradient containing 0.1% trifluoroacetic acid. Concentrations of all the variants were determined in 6 M guanidine hydrochloride using standard extinction coefficients for the tryptophan, tyrosine and...
phenylalanine residues. Protein purity was verified with standard SDS-PAGE and reverse phase HPLC and the correct molecular weight was confirmed by mass spectrometry.

**Circular dichroism**

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

**Results**

**In vivo screening strategy**

We used the engineered chimeric in vivo screen to identify combinations of stabilizing residues at three select residue positions for an unstable Gβ1 variant referred to as MonB (Barakat et al., 2007). This work deals primarily with this variant, and mutants thereof, and thus it is referred to herein as MonB-WT. The MonB-WT variant originated from an earlier design project in which a de novo protein interface was engineered by computationally docking the normally monomeric Gβ1 domain to itself followed by the use of the ORBIT suite of protein design algorithms (Dahiyat and Mayo, 1997) to mutate specific interfacial side chains with the goal of driving specific complex formation (Shukla et al., 2004; Huang et al., 2005; Huang et al., 2007). This design resulted in a pair of monomers that, upon generation in the laboratory, formed a heterodimer of modest binding affinity (the monomers are referred to as MonA and MonB). Introduction of the mutations that resulted in the MonA sequence stabilized it to a hyperthermophile (i.e. $T_m >100°C$) while the 8 for MonB were destabilizing resulting in a $T_m$ of 38°C (Gβ1-WT $\approx$85°C). The amino acid sequences of these variants are listed in Table I.

The successful performance of the chimeric in vivo screen was verified previously using 10 rationally designed Gβ1 mutant variants that fairly evenly span a melting temperature ($T_m$) ranged from 38 to $>100°C$ (Barakat et al., 2007). The screen has been used to consistently differentiate Gβ1 variants that possess different thermal stabilities: variants with higher thermal stabilities exhibit a slow-growth phenotype in the presence of the reporter antibiotic while those with lower thermal stabilities exhibit a faster growth phenotype.

**Creation of the library of MonB variants**

A MonB library of mutant variants was created by randomizing the codons of three positions (i.e. 23, 27 and 45) using saturation mutagenesis where all 20 amino acids were allowed at each position. Position 45 was chosen based on the proven importance of this position for the structural integrity of the Gβ1-fold (Barakat et al., 2007). Therefore, this region of the MonB-WT sequence was targeted for improvements in stability using both computational design and physical library screening. In addition to position 45, positions 23 and 27 were targeted due to their close proximity to 45 (Fig. 1).

To evaluate the diversity of the library, the genes for approximately 60 MonB library variants were randomly chosen and sequenced and the results revealed that bases at the random positions highly degenerate and not biased towards any particular base. At the nucleotide level, we calculated the percentage of each nucleotide (i.e. A, C, G and T) at the three positions. Each base was represented by $\approx$25% of the total base composition and demonstrated relatively equal diversity in the library at the DNA level. At the amino acids level, none of the sequences analyzed were identical and no unexpected bias was detected for any particular amino acid at any of the three positions. These results indicate that the designed library likely contained all of the possible mutated sequences and had a molecular diversity of $\approx$8000 different sequences.

**In vivo screening for stabilized MonB variants**

Screening of the randomized library entailed primary, secondary and tertiary steps (see section Materials and methods) due to the fact that the difference in the growth rate for the most stable variant (MonA; a Gβ1 variant that has a $T_m$...
value of $>100^\circ$C) and the least stable variants is greater than but still similar to natural variations in bacterial growth rates. The primary screen entailed picking slow-growing colonies identified on plates that contained reporter antibiotic. These colonies were then grown in the liquid media for the purpose of amplification and isolation of plasmids. In the second screening step, the purified plasmids were transformed into E. coli reporter cells and re-tested for growth on reporter plates to verify the slow-growth phenotype observed in the primary step. For comparison of colony size and growth rates, chimeric vectors that contained the genes for MonB, MonB-ORDES (a MonB variant derived from the in silico screening with ORBIT, described in the next section; Barakat et al. (2007), and MonA were used as controls. Colonies that did grow more slowly than MonB control colonies were further analyzed by monitoring growth in 96-well plates for 24 h in a fixed amount of reporter antibiotic (e.g. 2000 $\mu$g/ml carbenicillin). Out of the $\approx 200$ variants identified in the primary screening step, 140 passed the secondary step ($\approx 60$ clones turned out to be false positives) and thus the 140 were grown and observed in the 96-well format. The growth rates of both positive and negative controls were also included in the 96-well plates and consisted of PRP (i.e. the chimeric vector with no insert), MonB, MonB-ORDES and MonA. We analyzed eight variants that exhibited the most consistent slow growth in comparison
with MonB and MonB-ORDES. These were sequenced and subcloned into the pET21a plasmid for the purpose of large-scale protein expression and purification. Figure 2 illustrates the growth extent for all of the variants obtained from the screen in addition to those engineered previously using rational design methods (Barakat et al., 2007).

To further verify the correlation between growth rate and thermal stability, two relatively large, fast-growing colonies were also isolated and analyzed. These two colonies proved to grow more robustly in comparison with MonB-WT and thus their genes were subcloned into the pET-21a expression vector, expressed, purified and the melting temperatures measured with standard CD thermal denaturation analysis. The two variants are referred to as MonB GND, MonB GKS (the three letter extension refers to the amino acids at positions 23, 27 and 45, respectively). The two leftmost bars in Fig. 2 illustrate the increased growth rate for these variants. The lower thermal stabilities for these two variants were confirmed with CD thermal denaturation analysis. The measured melting temperature for MonB GKS is 28°C and the other fast-growing variant, MonB GND, proved to be a random coil and thus it was not possible to measure its melting temperature.

The total number of variants obtained from the in vivo screen is 10: eight that exhibit higher thermal stability than MonB-WT and two from fast-growing colonies that proved to be less stable. The \( T_m \)'s, growth extent (as reflected in the OD 600 readings from the larger 10-ml cultures) and amino acid sequences for positions 23, 27 and 45 are listed in Fig. 2. The amino acid sequence for positions 23, 27 and 45 for MonB-WT are Ile, Ala and Ala, respectively, and the amino acids for these respective positions for wild-type Gβ1 are Ala, Glu and Tyr.

**In silico screening for stabilized MonB variants**

In parallel with generating more stable variants of MonB-WT using the in vivo screening methods described above, computational mutagenesis and screening were also performed on MonB-WT at the same three positions (i.e. 23, 27, and 45). The ORBIT suite of protein design algorithms was used for this purpose (Dahiyat and Mayo, 1997). A MonB mutant library was created and screened by using a molecular mechanics scoring function to calculate pairwise interaction energies between rotameric descriptions of amino acid side chains and backbone atoms for the three selected positions. The resulting matrix of pairwise energies was used as input for the DEE algorithm that was utilized to identify the one optimal sequence and conformation with the lowest energy (i.e. the GMEC). The resulting sequence consists of Lys, Lys and Phe at positions 23, 27 and 45, respectively, and is referred to as MonB-ORDES.

In addition to the above side-chain selection process, we also applied MC simulation annealing, starting with the GMEC sequence and structure, for the purpose of producing a rank-order list of approximately 5000 lowest energy sequences (Voigt et al., 2000). The MC calculations were run so we could explore which alternative amino acid side chains had reasonable energies at the targeted positions and, more importantly, had we obtained such sequences in the in vivo screen. We also desired to determine the calculated energies of the in vivo sequences and ascertain where those energies (and sequences) fell in the MC rank ordered list. The calculated energies of the ~5000 rank-ordered sequences ranged from as low as -203.4 (for the original GMEC sequence, i.e. KKF) to as high as -152.0 kcal/mol. Of the 5000 sequences approximately 4000 were unique (the other ~1000 were alternative rotameric descriptions of redundant sequences). The 10 variants selected from the in vivo screen were compared with the ~4000 computed variants. Five of 10 in vivo selected variants were found in the ~4000 computed variants (six if the GMEC and MonB-WT sequences are considered). The sequences for positions 23, 27 and 45, the MC rank orders and energies, and the \( T_m \)'s are listed in Table II. It is of interest to note that while none of the variants isolated from the screen returned Gβ1 wild-type amino acids (i.e. 23A, 27E, 45Y) at positions 23 and 27, the variants with the five highest melting temperatures all had aromatic residues at position 45 (Table II). This is yet additional evidence of the structural importance of the aromatic cluster formed by amino acids at position 43, 45 and 52 (i.e. W, Y, F, respectively).

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Rank Order</th>
<th>Energy</th>
<th>( T_m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>LYW</td>
<td>n.f.</td>
<td>n.d.</td>
<td>69°</td>
</tr>
<tr>
<td>QWF</td>
<td>n.f.</td>
<td>n.d.</td>
<td>66°</td>
</tr>
<tr>
<td>KKF</td>
<td>1</td>
<td>-203.35</td>
<td>64°</td>
</tr>
<tr>
<td>TYQ</td>
<td>2262</td>
<td>-178.90</td>
<td>63°</td>
</tr>
<tr>
<td>KIY</td>
<td>n.f.</td>
<td>n.d.</td>
<td>61°</td>
</tr>
<tr>
<td>PLL</td>
<td>n.f.</td>
<td>n.d.</td>
<td>59°</td>
</tr>
<tr>
<td>YYT</td>
<td>755</td>
<td>-188.78</td>
<td>47°</td>
</tr>
<tr>
<td>YSV</td>
<td>2515</td>
<td>-177.39</td>
<td>47°</td>
</tr>
<tr>
<td>ISA</td>
<td>2880</td>
<td>-175.12</td>
<td>44°</td>
</tr>
<tr>
<td>IAA</td>
<td>3313</td>
<td>-171.87</td>
<td>38°</td>
</tr>
<tr>
<td>GKS</td>
<td>n.f.</td>
<td>n.d.</td>
<td>28°</td>
</tr>
<tr>
<td>GND</td>
<td>n.f.</td>
<td>n.d.</td>
<td>RC</td>
</tr>
</tbody>
</table>

The three positions that were targeted in the in vivo and in silico screening are listed in the left most column, followed by the rank order from the MC calculations, the MC energies and finally the measured melting temperature (n.f., not found; n.d., not determined; RC, random coil).

aKKF is the calculated GMEC sequence and IAA is the sequence for MonB-WT.
To be able to compare all of the variants isolated from the *in vivo* screen to the computational design, we used ORBIT to separately calculate the energies of each of the 10 variants that were obtained from the *in vivo* screen (the ORBIT-calculated energies are graphically represented in Fig. 3). The means by which this was accomplished is described below and in the Materials and methods section.

**Measuring protein stability in silico**

Total energies for each *in vivo* selected variant were calculated using ORBIT. The same computational calculation scheme was used for both the *in silico* mutagenesis and the subsequent DEE selection. However, positions 23, 27, 45 were fixed with the amino acid identity of the sequences obtained from the *in vivo* selected variants and only their conformations (i.e. rotameric descriptions) were allowed to change. The resulting energies were extremely close to those obtained from the MC rank-order list.

The ORBIT-calculated total energy for each of the 10 variants was compared with the melting temperatures for each *in vivo* variant. Figure 3 illustrates the correlation between the calculated ORBIT total energy and the protein thermal stability for each variant. In general, the variants with low total energy (more negative values and thus computationally predicted to be more stable) have higher melting temperature whereas variants with higher total energy (less negative) have lower melting temperatures. Therefore, it appears that the computational design algorithm is fairly well suited to estimate the stability of different MonB variants.

**Thermal stabilities measured for the variants obtained from the *in vivo* screen**

Far ultraviolet (UV) CD spectra were collected for the 10 variants and the resulting spectra for all but one variant is highly similar to that of Gβ1-WT (Supplementary data, Fig. S1). These results indicate that the more stable mutant variants most likely maintain the ββαββ fold topology of the wild-type domain. The MonB GND variant is so unstable that the Gβ1-fold topology is lost and that the structure exists in a random coil state. The melting temperatures for the other nine variants were measured using standard thermal denaturation monitored by CD at 218 nm (Supplementary data, Fig. S2). The slopes of the melting curves are similar and indicate cooperative unfolding and thus provide further evidence that the mutant variants maintain the overall wild-type fold.

The eight variants (predicted to be more stable than MonB) isolated from the *in vivo* library all have higher melting temperatures than MonB and two are more thermally stable than the variant obtained from the protein design algorithms. The two variants that originated from fast-growing colonies (and thus predicted to be less stable than MonB) have low (i.e. 28°C for the GKS variant) to un-measurable melting temperatures (i.e. random coil CD signal for the GND variant).

**Discussion**

The primary purpose of this study was to compare two strategies currently being used in the field of protein design to stabilize proteins: namely, *in vivo* and *in silico* screening. We developed gene libraries based on an unstable variant of the Gβ1 domain (MonB-WT) and utilized an *in vivo* screening method to identify stabilized variants of this mutant Gβ1 variant. For the engineered MonB library, the codons for just three positions (i.e. 23, 27 and 45) were randomized. Mutant variants with increased thermal stabilities were indeed identified using the *in vivo* chimeric screen. The genes for eight variants, which exhibited favorable *in vivo* behavior (i.e. slow-growth phenotype in the presence of carbenicillin and thus indicative of a relatively stable variant), were sequenced and revealed that amino acids that conferred greater stability were generally diverse yet, not unexpectedly, biased towards hydrophobic and aromatic amino acids at the three randomized positions.

To prove the corollary of the screen (i.e. since small, slow-growing colonies represent stable MonB variants, then fast-growing colonies should represent variants that are relatively unstable), we analyzed two fast-growing colonies from the *in vivo* screen. Upon sequencing of the genes for these two variants it was determined that the amino acids at positions 23, 27 and 45 were GKS and GND, respectively. Although we initially assumed that these two variants would be too unstable to produce an *E. coli* expression system, we were indeed able to produce enough material (i.e. milligram quantities) for CD thermal denaturation analysis. The measured $T_m$ for the GKS variant was 28°C and, since the other GND variant proved to be so unstable its CD signal indicated random coil, and thus the $T_m$ was too low to measure (CD spectra and thermal denaturation melts are shown in Supplementary data, Fig. S1). We find it quite surprising that it is possible to produce these variants in *E. coli* especially in light of the fact that other small test proteins, which have considerably higher $T_m$s, do not express at an appreciable level in *E. coli*. The secondary structures of the other test proteins (e.g. protein A) is primarily α-helical and we speculate that it may be the presence of residual β-hairpins in the structure of the destabilized Gβ1 variants that potentially impart a greater resistance to *in vivo* proteolysis within the context of the *E. coli* expression strain. We are currently in the process of testing this hypothesis by
swapping the β-hairpins regions of these test proteins and analysis of potential residual structure with NMR.

In parallel with the in vivo and experimental analysis, protein design algorithms were used in an attempt to generate a stable MonB variant by using computational sequence optimization at the three targeted positions. The best computed variant has a measured $T_m$ that ranks third within the panel of the in vivo screened variants and consists of KKF at the respective positions. The fact that this variant was not the most stable is not surprising as a necessary limitation of the computational approach is that the protein backbone must be held rigid for the calculation to remain tractable. During the ORBIT calculations, the necessity for a rigid backbone limits amino acid choices to those that fit well into that defined backbone geometry and return favorable van der Waal contact energies, whereas, in the context of the chimeric screen, backbone atoms can move and accommodate a greater variety of amino acid side chains. Approximately half of the in vivo selected variants (five out of ten) were found in an MC rank-ordered list of ~5000 computed sequences. Three of the in vivo selected variants were not found in the computed list due to the fact that proline and glycine residues were not considered at the randomized positions.

A comprehensive analysis of the correspondence between the theoretical and experimental stabilities of the selected variants was enabled by the collection of experimental data (i.e. CD) for the computed sequence as well as all the sequences obtained from the in vivo screen. The thermal stability of each variant was used as a metric for the screen performance and thus the melting temperatures of the selected variants are plotted against growth in carbencillin (Fig. 4B) as well as the ORBIT calculated energies (Fig. 4A). The modest correlation between melting temperatures and the ORBIT-calculated energies (i.e. $R^2 = 0.65$) indicates that although the computational scoring function is capable of screening for and obtaining stable sequences, and predicting potential stabilities, ongoing improvements need to be made. The correlation between the thermal stability of the selected variants and their growth in carbencillin indicates that the in vivo chimeric screen is also an excellent predictor of thermal stability (Fig. 4B).

Designing proteins with improved stabilities need to be a reliable and routine tool. At present, directed evolution methods are fairly robust at generating mutants with improved protein thermostability and on reasonable time scales yet there are aspects that are relatively laborious and successful outcomes are not guaranteed. Overall, computational design approaches utilize fewer wet-lab resources and therefore it is our belief that the in vivo screen results may be used to iteratively improve the scoring function and thus enhance the ability to generate computationally designed proteins with greater thermal stabilities. For example, the correlation between measured thermal stabilities and in vivo screen readout is fairly robust and thus it should be possible to use the sequences obtained from the screen to fine tune the weights of adjustable parameters in the force field used by the ORBIT algorithms. New sequences can be computed using the adjusted weights and rapidly tested for increased stability in the context of the chimeric screen. In addition, the ORBIT algorithms could be used to narrow amino acid choices for select positions by running a MC calculation (starting with the GMEC sequence) to determine subsets of amino acids with similar characteristics (e.g. predominantly hydrophilic and/or charged amino acids) at the targeted positions. This would effectively reduce the sequence space tested in the physical screen as the resulting library could be generated by utilizing codons biased towards the desired amino acid characteristics. The resulting library could be interrogated using the chimeric screen and isolated stable sequences would again be used to fine tune adjustable weights for parameters in the applied force field. It is our belief that emerging design methods that marry the best of the computational and in vivo approaches promise to enhance the design of proteins and thus should continue to be a key tool for the field of protein engineering.

**Supplementary data**

Supplementary data are available at PEDS online.

**Funding**

This work was supported by the National Science Foundation under Career grant no. 0448670 and the California Metabolic Research Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Nora Barakat is a recipient of an Arne N. Wick Pre-doctoral Research Fellowship from the California Metabolic Research Foundation.