Preface

For many, proteins are something you eat. For others, they serve as the “workhorses” of our bodies, carrying out virtually all of the tasks needed in order to function properly. So, it seems fairly obvious why so much time and energy has been spent studying proteins, due to their fundamental roles in biology. Typically in order for a protein to carry out its vital work, it must be both stable and adopt a highly specific three-dimensional structure that allows it to interact with other molecules for the purposes of cellular chemistry, signaling, etc…

This is essentially the classical view of proteins—a static three dimensional structure. However, there are many domains (regions) of important proteins that carry out the main function of the protein but have no defined tertiary structure. Unfortunately, as much as we might understand about proteins, we still have yet to elucidate how proteins “know” how to adopt the correct three-dimensional structure and further why diverse sequences can adopt the same or nearly similar structures (motifs). As a consequence of these issues with the classical view, emphasized by new experimental results (obtained especially from NMR spectroscopy), a new paradigm was needed to further understand the mechanism of proteins.

One must begin to consider proteins not as static structures but as a distribution of conformations (microstates) ranging from completely folded conformations, to fully folded conformations, or rather, an ensemble. The fully folded protein is often the most thermodynamically stable structure and consequently is observed most frequently.

To test hypotheses about nature’s polymers, protein design has been used to construct experimental tests and to understand the effects of primary sequence changes on folding kinetics and overall protein stability. The question being answered here is, can we use this new information and leverage it to design proteins? And finally, can we then use these new insights to understand loss of function in medically relevant proteins?

Abstract

A new computational method is developed to create novel amino acid sequences that fold into specified protein structures. We utilize an ensemble model that generates a range of accessible protein conformations as well as a simplified “alphabet” of “thermodynamic environments.” This design method was implemented on the well-studied SH3 domain of C. elegans as the design target. Four randomly selected designs exhibit primary, secondary, and tertiary structure consistent with the target in silico. Further, an unexpected correlation between the fitness, as measured by the design algorithm, of the most buried residues in the native state and protein stability was found.
Because the method is facile, general, rapid, and computationally inexpensive, it should be applicable to a wide range of protein folds. Future improvements to the method will include explicit consideration of solubility, preservation of function, and optimization of stability. We propose that this method can be used to help understand medically relevant protein mutations that can result in a loss of function due to changes in stability.

Introduction

The wild-type protein used as the design target is a small protein sequence, Sem-5, found in C. elegans, that is homologous to two SH3 domains found in the human GRB2 adapter protein (1-3). GRB2 is one of the most studied adapter proteins because of its involvement in RAS signaling pathway, which, when overactive, is known to promote cancer (4). GRB2 does not have any catalytic activity; however, it uses its SH3 domains (Sem5-like) to bind to small proline-rich polypeptide sequences, such as SOS, which further stimulates activated RAS-GTP (1). Localization of GRB2 in the neuronal cell body has also been discovered in the brains of patients diagnosed with Alzheimer’s disease which suggests its involvement (5). Understanding the structure of this protein, using our design method, could possibly shed light as to why GRB2 is able to play a role in many different pathways by reliable modeling, and eventually prediction, of protein-protein interactions.

If all folding information is contained within the primary sequence (6) it should be possible to construct identically folded proteins with diverse sequences, provided the thermodynamic landscape is preserved. This is the basic theory behind our design methodology. Understanding protein folding, the physiological process by which a newly synthesized amino acid chain adopts its functional shape, is one of the greatest challenges in molecular biology (7). This challenge has been historically framed as having two distinct unsolved questions (8, 9): why does a protein adopt a unique ground-state conformation out of the approximately 1000 known natural folds (10), and how is that conformation thermodynamically stabilized against unfolding? This “specificity vs. stability” framework suggests that one way to approach the whole problem is to study each part separately. Such thinking has guided the field of protein design and has resulted in the current generation of accurate practical structure prediction algorithms (11-13). These algorithms have generally employed sophisticated knowledge-based specificity information (i.e. sequence and structure similarity) at the expense of physically realistic stability information. Culmination of vigorous theoretical development has resulted in many achievements (14), such as design of proteins with novel enzymatic functions (15) and design of never-before-seen protein folds (16, 17). Yet, these other design strategies are non-trivial to utilize by other laboratories or those without strong backgrounds in computational biophysics. In contrast, one envisions a more readily accessible method of protein design because it can have many direct applications. Consequently our goal is to develop a simpler design method that is readily useful as a tool for other laboratories.

However even when protein structures can be accurately predicted from sequence, the thermodynamic stability of the structure cannot be accurately predicted (18-20). The main reason for this shortcoming is that the myriad interatomic interactions that comprise the enthalpy and the
entropy of a typical protein, which add to hundreds of kilocalories each, and the stability of the folded protein, the difference of enthalpy and entropy, is usually only a few kilocalories (20, 21). Thus, small errors in physical energy functions can quickly obscure any stability calculation. The current generation of algorithms usually incorporate greatly simplified energy functions (if thermodynamic stability is even estimated at all). Active research is continually improving these calculations, but even for single amino acid changes in well-studied proteins stability prediction is still a challenge (14, 19).

This work approaches the problem of understanding protein folding through protein design in a new way, combining specificity and stability information into a single, simple algorithm. The design method leverages the thermodynamic information in the protein conformational ensemble to generate novel amino acid sequences compatible with a given fold. For the example case of the C. elegans SH3 domain, structure predictions, combined with experimental fluorescence and circular dichroism measurements, suggest that four novel amino acid sequences are indeed folded proteins compatible with an SH3 fold, providing one answer to the specificity question. Furthermore, it is demonstrated that a fitness function for buried (hydrophobic) core residues is sufficient to quantitatively explain the experimental thermodynamic stability of the designs, providing one answer to the stability question. The synthesis of stability and specificity information would allow us to understand mutations that lead to a loss in function due to stability, because for the SH3 domain there is a strong correlation between conserved residues and stability (22).

**Results and Discussion**

*Detailed Description of Design Algorithm*

The proposed algorithm uses as input a protein structure (Figure 1). A model of the conformational ensemble of the input structure is generated by COREX-BEST (23-25), and this ensemble is Boltzmann-weighted using a surface area based energy function to provide ensemble-averaged values for the thermodynamic stability, and its enthalpic and entropic contributions, at each residue position in the protein. (23) Since the residue stabilities computed by COREX have been shown to correlate well with experimental hydrogen exchange measurements (23, 26, 27), the computed ensemble can provide detailed thermodynamic information not immediately apparent from the single input structure.

This detailed thermodynamic information is leveraged into the design algorithm as follows. For any globular protein, discrete combinations of stability, enthalpy, and entropy have been observed; these combinations are termed “thermodynamic environments” (28). Thus, any globular protein structure can be simply described as a collection of “native state thermodynamic environments” (29, 30). The denatured (unfolded) state of the same protein is naturally part of the COREX-BEST ensemble, and a separate set of “denatured state thermodynamic environments” has been derived (31).
The native and denatured thermodynamic environments have been previously demonstrated to contain information on the compatibility of amino acid sequence with structure, and statistical counts of the frequencies of the amino acids in each environment permit a “fitness measure” (i.e., a log-odds score) of each of the 20 natural amino acids in each of the eight native state and eight denatured state thermodynamic environments (30, 31). Importantly, the native and denatured state environments appear to contain different types of information about protein structure and stability.

The environments are input into the algorithm (Figure 2), and novel amino acids at each position in a putative designed protein are scored according to a previously developed (29-31) fitness measure in each respective native (Figure 3) and denatured environment. The particular choice of amino acid at each position is constrained by an ad-hoc device termed a “sequence mask” (Table 1). The mask constrains the number of allowed residues by limiting a particular wild-type residue that, for example, has positive score in its environment. Any substitutions are limited to those residues that also score positive in that same environment. A similar process is applied to the denatured state environments, and the final mask is the intersection of the native and denatured state sub-masks (Table 1). The mask dramatically filters the possible sequence space for the SH3 domain by many orders of magnitude, so that on average less than 50% of the original 20 amino acids are permitted at any given position in a designed sequence.

The second filter is based on the summed fitness scores of the putative amino acid sequence in its native (Figure 3) and denatured environments. As mentioned, previous work has established the efficacy of high fitness scores in fold recognition, and on average it is observed that each residue achieves a certain threshold average positive score. Designed sequences must also meet this score threshold.

Given a sequence mask, the algorithm then proceeds by randomly choosing one of the allowed amino acids at every position. When the entire putative sequence design is chosen, the native and denatured state fitness scores are summed and compared to the thresholds. If the sequence meets the score thresholds and contains few repeated amino acids (a characteristic shared by most known globular proteins), it is output as a candidate sequence. For these SH3 domains, tens of millions of generated sequences are tried and discarded for every one that passes the mask and score filters.

**Computational Characterization of SH3 Designed Sequences**

Primary, secondary, and tertiary structural characteristics of designed sequences were assessed to understand similarities and differences as compared to known SH3 domains. Conservation of primary sequence, both in general physico-chemical side chain properties as well as specific preservation of individual residue types for functional reasons, are well-known characteristics diagnostic for protein family membership (21, 32). Figure 4 shows sequence alignments of known SH3 domains (33), novel proteins from the design algorithm, and randomly generated proteins of identical length. There is a clear correspondence between conservation of hydrophobic and polar regions between the natural SH3 domains and the designed sequences, and this correspondence is weakened in the randomly generated proteins, despite the inclusion of a similar amount of
sequence identity. This result is interpreted as evidence that the mask and scoring filters indeed bias allowed sequence space to favor amino acid choices that reflect the larger SH3 domain family.

Stronger evidence for this bias is obtained by attempting to computationally predict (13) the secondary and tertiary structure for five randomly chosen novel sequence designs. These five are indeed foldable with high confidence using the I-TASSER software (Figure 5). Notably, the secondary structure as well as tertiary are well-modeled by the software (Figure 5, right hand side), with very little evidence of alpha-helical structure as expected for the all-beta SH3 barrel protein. As a control for the correct usage and functioning of the software, the wild-type design target indeed folds to the expected structure (rightmost point).

Against a background of randomly generated sequences of increasing amino acid identity to the target protein, there is a clear threshold of sequence similarity, between 40-50% identity, at which I-TASSER exhibits the ability to fold to the correct structure (Figure 5, colored points). The designed sequences, however, have a sequence identity of only 34-45% (Figure 5, labeled points), suggesting that the choices made for new amino acids in the designs are adding additional information relevant to adopting the target structure.

**Experimental Characterization of SH3 Designed Sequences**

Do these designed proteins experimentally fold into SH3 beta barrels? To answer this question, these four novel sequences were engineered to contain an N-terminal His-tag, expressed in *E. coli* and purified (Table 2). As a control, the wild-type protein design target was prepared in parallel. A sixth sequence, a single substitution of the wild-type W52L, was also studied for reasons described below.

Consistent with characteristics of known globular proteins (21), Tryptophan emission spectra showed evidence of buried fluorophores which are exposed to solvent upon heating for all designs except for Design 9 and Design 11 W52L (Figure 6a). Closer inspection of the putative locations of the two Tryptophans in the Design 9 sequence (Table 2) suggested that they would be solvent exposed in the target structure, rationalizing the deviant behavior; similarly the position of the single Tryptophan W51 in the W52L mutant is solvent-exposed in the models (as well as a known functionally important binding residue) (34). Although the biological function of the SH3 domain, binding of a Proline-rich ligand, was not explicitly considered in the design algorithm, addition of ligand resulted in an increase in Tryptophan emission signal, diagnostic of a bound complex (Figure 6b). (35)

Circular dichroism spectroscopy provides perhaps the most informative data concerning the secondary structure of a protein (36). The wild-type protein exhibited a spectrum similar to that previously reported (35), with a weak overall negative ellipticity and a local maximum around 220 nm (Figure 6c). However, spectra for the designs (Figure 6c) were different than the wild-type spectrum, exhibiting weak signals of a single minimum between 205 and 220 nm. All spectra, whether design or wild-type, were inconsistent with alpha-helical structure. Other naturally occurring SH3 domains that are known to adopt the characteristic beta-barrel fold, such as Abl,
Fyn, and alpha-spectrin, exhibit non-superimposable far-UV circular dichroism spectra at neutral pH when the proteins are folded (37).

Deconvolution of the Figure 6c spectra (38) revealed that all designed proteins were predicted to contain similar amounts of beta structure, with very little alpha (Table 3). Moreover, this secondary structure content was consistent with that expected from the target structure, and the inferred fold for all proteins was the same expected fold classification for the SH3 domain: “mainly-beta” (38, 39). Taken together with the observation that the point mutant W52L dramatically changes the circular dichroism spectrum as compared to wild-type (Figure 6c), it is concluded that the most likely interpretation is that all designed proteins exhibit the expected secondary structure of the design target, and that local effects on the circular dichroism spectrum unique to the wild-type protein are caused by the presence of two neighboring Tryptophan residues (Table 2) (36, 40).

Cooperative Thermal Denaturation of Designed Proteins

One hallmark of a folded globular protein is cooperative reversible denaturation (8). To test whether any designed proteins exhibited this behavior, the circular dichroism signal at 222 nm was monitored as a function of temperature. All designed proteins, as well as the wild-type protein, displayed cooperative denaturation that was at least 85% reversible upon cooling (Figure 7a). Melting temperature, $T_m$, and enthalpy at the melting temperature, $\Delta H_m$, were not greatly different than values obtained for other known SH3 domains and variants (Figure 7b), establishing that the denaturation thermodynamics for the designed proteins were consistent with those expected from proteins of similar size. Under the assumption of a size-dependent heat capacity (8), the folded stabilities of the designed proteins were estimated with the Gibbs-Helmholtz equation (Figure 7c) to span a range of 1.1 (Design 2) to 5.8 +/- 1.0 kcal/mol (Design 11 = wild-type) at 25 ºC (Table 4), pH 7.8.

Thermal Stabilities of Designed Proteins Correlate with Thermodynamic Environments Scores

An unexpected correlation was observed between the native state log-odds scores and the experimental stability of the protein: residues of the designs that were modeled to occupy the buried core of the protein exhibited fitness scores that scaled with the measured stability (Figure 8). A maximum value of the correlation between native state fitness score and stability was observed for those buried residues exposing 20 % or less of the possible maximum van der Waals’ surface area for the entire residue (data not shown), suggesting that this result was robust for the SH3 domain structure. This observation is interpreted as both validation of the COREX ensemble methodology (since the experimental stability of a protein as measured from hydrogen-deuterium exchange experiments (23, 27) often correlates with the exchange behavior of the most buried residues in the protein (41)), and as a possible practical means of increasing protein stability. This hypothesis, which is currently being tested on a larger database of varied protein structures with varied stabilities, implies that engineered point mutations at buried positions that increase the
native state fitness score relative to the wild-type protein will increase the engineered protein’s stability.

Future Directions

1. Crystallization trials are in progress to directly obtain structural information and verify if these designed proteins fold to the SH3 beta-barrel domain.
2. A computational stability predictor based on the buried subset of thermodynamic environments’ scores is under development.
3. Comprehensive characterization of the extended SH3 beta barrel domain sequence space, under the constraint of the design principles articulated here, may provide insight into the evolution of this protein family and may permit more effective sequence-based detection of natural SH3 domains.
4. Improvements to the design method may focus on preservation of functionally important residues and optimization of protein solubility.
5. Use this method to help understand medically relevant protein mutations that can result in a loss of function due to changes in stability.

Materials and Methods

Computational Methods

A model of the protein ensemble for the design target C. elegans signal transduction protein SH3 domain (34) was generated from the Protein Data Bank (42) structure 1sem using the COREX-BEST algorithm (23, 25). Parameters for the software were: Window Size = 5 residues, Minimum Window Size = 4 residues, Simulated Temperature of 25.0 °C. Entropy weighting factors of 0.5 and 1.5 were used to estimate residue stability constants for the native and denatured states of the protein, respectively. Native and denatured state thermodynamic environments were assigned from the enthalpic and entropic components of the COREX stability by mapping each residue’s ensemble-averaged $[\Delta G, \Delta H_{ap}, \Delta H_{pol}, T\Delta S_{conf}]$ to the closest average values for the databases of proteins reported previously (30, 31).

The design mask was for the SH3 domain target was computed from the native and denatured thermodynamic environments by tabulating the sign of the log-odds score table (30, 31) for the wild-type 1sem amino acid target (separately for the native and denatured states) and permitting only those residues in the score table that exhibited the same sign for the score value. The native state log-odds score table is shown in Figure 3. The denatured state score table, and perl scripts for creating the design mask for any protein, are available from the authors.

I-TASSER 4.2 (13) predictions were run using a local copy of the software, with all default parameters except the benchmark flag, which was set to “on” in order to reduce the influence of sequence identity on the structure models.
Protein sequences (Table 2) were codon-optimized for *E. coli* expression, given an N-terminal HisTag, and cloned into *pJ414* (T7-promoter, ampicillin-resistance) expression vectors by DNA 2.0 (Menlo Park, CA). The auto-induction method of Studier (43) was used to express the proteins. Routinely, freshly transformed single overnight BL21(DE3) colonies, that exhibited ampicillin resistance on LB plates, were selected and grown for 24 hours in 1 L of auto-induction medium at 37 ºC, 160 rpm. Cell pellets were stored at -20 C for at least 12 hours, then thawed at 4 ºC, resuspended and sonicated in 50 - 100 mL of 7 M urea, 100 mM sodium phosphate monobasic, 10 mM Tris, pH 7.8. The lysate was cleared at 15,000 rpm for 1 hour at 4 ºC. The supernatant (cleared lysate) was then passed over 1 – 2 mL of HisPur Nickel or Cobalt Resin (ThermoFisherScientific, Rockford, IL) that was previously equilibrated with at least 20 column volumes of 7 M urea buffer. The column was washed with at least 50 column volumes of 7 M urea buffer, and the 280 nm absorbance signal was monitored until baseline was reached. Protein was eluted with 3 – 12 mL of 7M urea buffer brought to 0.5 – 1 M imidazole, pH re-adjusted to 7.8. Eluted material was exhaustively dialyzed vs 4 x 4L of 20 mM sodium chloride, 50 mM sodium phosphate, pH 7.8, 10% glycerol. Yields were estimated to be 1 – 10 mg / L as computed by the theoretical extinction coefficients given by the SWISS-PROT ProteinParam widget at expasy.org; the values were 13980, 15470, 13980, 13980, and 22460 M⁻¹ cm⁻¹ for Designs 11 (wild-type), 10, 9, 8, and 2, respectively. The computed extinction coefficient for the Design 11 W52L mutant was 8480 M⁻¹ cm⁻¹. The HisTag was generally not removed for experimental characterization, although a pilot experiment on the wild-type protein containing the HisTag compared to the FactorXa HisTag-removed version of the same protein revealed little difference in the observed circular dichroism spectrum or the melting temperature (data not shown). Proteins were judged to be > 90% pure by 20% SDS-PAGE, and the correct sequences of the wild-type and W52L proteins were confirmed by trypsin digest MALDI and ESI mass spectrometry. D10 was observed to be soluble to no more than 10 µM at neutral pH, and intermittent difficulty of obtaining 1 mg mL⁻¹ solutions of D2 and D8 were noted, suggesting that solubility may be an important overlooked factor in the design algorithm (data not shown). The point mutant W52L was engineered using primers obtained from Integrated DNA Technologies (Coralville, IA) to guide PCR based site-directed mutagenesis (Phusion System, New England Biolabs, Ipswich, MA) of the wild-type DNA; the correct product was confirmed by sequencing.

**Experimental Methods: Biophysical Characterization**

Tryptophan fluorescence emission was monitored with an AVIV Model 410 Fluorimeter, excitation set to 295 nm, bandwidths for excitation and emission were always equal and set to 1.0 – 3.0 nm, depending on protein concentration. SosY (34, 44) peptide concentration was determined using the theoretical extinction coefficient of 1490 M⁻¹ cm⁻¹ given by the SWISS-PROT ProteinParam widget at expasy.org. Circular dichroism spectra were obtained on an AVIV Model 420 using a 0.1 cm pathlength cell. BeStSel software (38) was used for deconvolution analysis, but CDNN deconvolution software (45) gave qualitatively similar results (data not shown). Thermal denaturation was performed at step sizes of 1 ºC, 20 ºC min⁻¹ heating rate, 0.1
min equilibration and 30 secs data collection, monitoring signal at 222 nm. All samples were > 85% reversible from ambient temperature to 80 °C as judged by the return of the 222 nm signal. Protein concentrations were generally from 1 – 20 µM in 20 mM sodium chloride, 50 mM sodium phosphate, pH 7.8, 10% glycerol.

Acknowledgements

We thank Alex Chin, Dr. Hesam Motlagh, Jordan White for assistance with mass spectrometry, and Dr. Katie Tripp for assistance with core facility instruments. Financial support from the Johns Hopkins University Dean’s Undergraduate Research Award (DURA), the NSF MCB-1330211, and the NIH R01-GM063747 are all gratefully acknowledged.

References


**Figure and Table Legends**

**Figure 1.** Design target SH3 domain sequence, structure, and thermodynamic environments.
Cartoon representation (pyMOL, Schroedinger LLC, New York, NY) of 1sem.pdb is shown; beta-strands are colored yellow, 3-10 helix in red, and coil regions in green. Amino acid sequence in one-letter code is shown underneath, with residue positions classified as beta-strand (E), 3-10 helix (H) and coil (no letter) annotated just above the sequence. Also shown are the native and denatured state thermodynamic environment classifications; these classifications are numbered 1 – 8 in for each state, where environment 1 is predicted to be relatively least stable and environment 8 is predicted to be relatively most stable in the COREX-BEST protein ensemble.

**Figure 2.** Workflow of simple protein design algorithm.
Three adjustable parameters are contained within the algorithm; these are calibrated based on observed characteristics of the COREX-BEST ensembles of many diverse globular proteins of known structure, as described in the text.

**Figure 3.** Native state thermodynamic environments log-odds score table.
The values in the table represent log (observed / expected) counts of amino acids observed in the COREX-BEST ensembles of a large number of diverse protein folds (29, 30). Values greater than 1 represent statistical enrichment of amino acids, values less than 1 represent depletion. Rainbow colors indicate relative stability of the environment of the protein in its folded structure: cooler colors (environments 1-4) are lower stability and warmer colors (environments 5-8) are higher stability.

**Table 1.** Allowed sequence “mask” for the target SH3 domain.
At each of the 58 residue positions in the 1sem SH3 domain sequence, only a subset of the 20 amino acids are permitted to occur at a position in a putative sequence design. Whether or not an amino acid is permitted is
based on native and denatured state ensemble characteristics, quantified as statistical log-odds scores, as described in the text.

Figure 4. Comparison of designed novel sequences with naturally occurring SH3 domain sequences. Eighteen naturally occurring sequences are shown according to the structure-based alignment of Larson & Davidson (33). Just beneath are 20 novel sequences that pass all mask and score filters of the design algorithm. Conservation of hydrophobic regions (yellow highlight), polar/charged regions (blue and green font), and conserved individual charged residues (magenta) are evident in the designed sequences; highly conserved positions between the two sets of proteins are marked by asterisks and the letter “B”. In contrast, randomly generated sequences containing approximately the same amount of identity (~25%) to the design target as do the designed sequences display much less common coloring (bottom 20 sequences). Coloring and annotation was performed by CHROMA (46).

Figure 5. Tertiary structure modeling of selected designed sequences as compared to randomly generated sequences. One hundred randomly generated sequences of identical length to the SH3 target domain were created at various amounts of sequence identity to the target: cooler colors represent lower levels of identity and warmer colors represent the highest levels of identity. I-TASSER structure predictions (13) were run for each sequence and the quality of the five resulting models as compared to the target were assessed with TM-Score (47); higher values of TM-Score indicate more structure similarity to the target, points are average +/- stdev of the five TM-Scores. For comparison, novel designed proteins Design 2, Design 8, Design 9, Design 10, Design 11, were run and their model qualities were also plotted (larger points). The designed proteins clearly fold properly due to a higher level of sequence information, despite a lower level of sequence identity, suggesting that the design algorithm is selecting residues important for the foldability of the sequences. Examples of the first I-TASSER models are given on the right side of the figure.

Table 2. Amino acid sequences of novel designed proteins used for biophysical characterization. The first 12 residues contain the HisTag and a FactorXa cut site.

Figure 6. Fluorescence and circular dichroism characterization of designed proteins. A. Tryptophan emission spectra at low and high temperature. Loss of signal and redshift upon heating is evident for most proteins. Design 9 and Design 11 W52L appear to be exceptions, possibly due to exposure of the fluorophore to solvent in the native state of the protein, as described in the text. Buffer conditions were 20 mM sodium chloride, 50 mM sodium phosphate, 10 % glycerol pH 7.8. B. Increase in Tryptophan emission signal upon binding is observed for Design 11, Design 11 W52L, and Design 10. Design 2 did not show an increase in signal. C. Far-UV circular dichroism spectra show differences between the designed proteins that are inconsistent with alpha-helical or random coil structure. Buffer conditions were 20 mM sodium chloride, 50 mM sodium phosphate, 10 % glycerol pH 7.8.

Table 3. Estimated secondary structure content for designed SH3 proteins. Deconvolution of the circular dichroism spectra shown in Figure 6c was performed with the BeStSel software
Figure 7. Cooperative thermal denaturation profiles for designed proteins. A. Designed proteins show a range of cooperative behavior consistent with globular protein denaturation. Buffer conditions were 20 mM sodium chloride, 50 mM sodium phosphate, 10% glycerol pH 7.8. Smooth curves are best fits to the following equation: $y = \frac{[(ax+b)+(cx+d)\exp(-\Delta H_m/R * (1/x - 1/T_m))]}{[1+\exp(-\Delta H_m/R * (1/x – 1/T_m))]}$. Values for $\Delta H_m$ and $T_m$ so obtained are given in Table 4. B. Enthalpy $\Delta H_m$ and melting temperature $T_m$ of designed proteins compared to those of known SH3 domains. Values for the designed proteins are not incompatible with those measured for naturally occurring SH3 domains. C. Gibbs-Helmholtz protein stability as a function of temperature. Curves were computed from the following equation (8): $\Delta G(T) = \Delta H_m(1-T/T_m)+\Delta C_p[(T-T_m)-T\ln(T/T_m)]$. The heat capacity for each protein was computed from the four separate equations in Table 5 of Robertson & Murphy (8) and the results were averaged and reported in Table 4.

Table 4. Thermodynamic parameters for designed proteins. $\Delta H$ and $T_m$ were obtained from fits to the experimental data shown in Figure 7a. $\Delta C_p$ was estimated from the equations found in Table 5 of Robertson & Murphy (8), using the I-TASSER (13) structure models for each design. $\Delta G$ is the value from Figure 7c at 25 °C.

Figure 8. Correlation between experimental stability measurements and native state log-odds scores of buried residue positions. A buried position is defined as less than or equal to 10% solvent exposed surface area of the entire residue, as compared to the blocked tripeptide value (23). Native state log-odds scores (Table 3) for the residues of designed proteins modeled to be buried in the protein core were summed and plotted on the x-axis. The y-axis plots the experimental stability from Table 4. The inset shows a spacefill I-TASSER model of Design 2 (gray), illustrating the 10% or less exposed residues (red) that contribute to the x-axis score.
Figure 1.

SecStruc     EEEE     EEEEE     EEEEE     EEEEEHHHHEEEEE  
             ETKFVQLFDFNPQESGELAFKRGDVITLINKDDPNWWEGQLNNRGGFPSNYVAPYN  
Native      4444477777442122222222244444443322223483333333357447741111  
Denatured   77773511144553353366666611222222113333313777776264777772222  

Figure 2.

INPUT: Native and Denatured Thermodynamic Environments of Target

“Mask” defines allowed amino acids

ADJUSTABLE PARAMETERS:
1. Average (0.10) and Variance (0.20) of Native Scores
2. Average (0.10) and Variance (0.20) of Denatured Scores
3. Maximum Number of Doublets (2)

OUTPUT: Amino Acid Sequences Compatible with the Design Target
Figure 3.
Figure 4.
Figure 6.

A. 

B. 

C.
Figure 7.

A.

B.

C.
Figure 8.
Table 1.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>T</td>
<td>K</td>
<td>F</td>
<td>V</td>
<td>Q</td>
<td>A</td>
<td>L</td>
<td>F</td>
</tr>
<tr>
<td>W</td>
<td>L</td>
<td>C</td>
<td>F</td>
<td>Y</td>
<td>V</td>
<td>C</td>
<td>F</td>
<td>Y</td>
</tr>
<tr>
<td>W</td>
<td>N</td>
<td>A</td>
<td>G</td>
<td>P</td>
<td>C</td>
<td>Q</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>M</td>
<td>T</td>
<td>C</td>
<td>L</td>
<td>A</td>
<td>N</td>
<td>Q</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>S</td>
<td>P</td>
<td>H</td>
<td>R</td>
<td>Q</td>
<td>D</td>
<td>R</td>
<td>A</td>
<td>I</td>
</tr>
<tr>
<td>Q</td>
<td>D</td>
<td>E</td>
<td>H</td>
<td>R</td>
<td>K</td>
<td>A</td>
<td>L</td>
<td>I</td>
</tr>
</tbody>
</table>

Translation:

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>T</td>
<td>K</td>
<td>F</td>
<td>V</td>
<td>Q</td>
<td>A</td>
<td>L</td>
<td>F</td>
</tr>
<tr>
<td>W</td>
<td>L</td>
<td>C</td>
<td>F</td>
<td>Y</td>
<td>V</td>
<td>C</td>
<td>F</td>
<td>Y</td>
</tr>
<tr>
<td>W</td>
<td>N</td>
<td>A</td>
<td>G</td>
<td>P</td>
<td>C</td>
<td>Q</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>M</td>
<td>T</td>
<td>C</td>
<td>L</td>
<td>A</td>
<td>N</td>
<td>Q</td>
<td>D</td>
<td>R</td>
</tr>
<tr>
<td>S</td>
<td>P</td>
<td>H</td>
<td>R</td>
<td>Q</td>
<td>D</td>
<td>E</td>
<td>H</td>
<td>R</td>
</tr>
<tr>
<td>Q</td>
<td>D</td>
<td>E</td>
<td>H</td>
<td>R</td>
<td>K</td>
<td>A</td>
<td>L</td>
<td>I</td>
</tr>
<tr>
<td>Design</td>
<td>Sequence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 (Wild-Type)</td>
<td>MHHHHHHIQRHMETKFQALFDNQPESGELAFKRGTDVTILKDPNWWEGQLNNRGIFPSNYVAPYN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MHHHHHHIQRHMDLKYVAAYFRFNPQEPGEIAWNAHGIVTVSGKTNWHEQVNYKIEAYPENWVAGMS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>MHHHHHHIQRHMRPHYLQFLLLDFNTEQKNELAFGRGTDVTIGKGDTHQEDLKYIDLFPPWNFIAPMQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MHHHHHHIQRHMLKFLAQFLSYGQETDQLFDRGDDIVIVVEDKGTNMHWGSLQNFIFIKPFPEKYIPPPK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>MHHHHHHIQRHMDLNPQPGAFDYNPQEPGEIVWKRGKVVIDGTDNPWHKDNGKAPKLYLETYAPMR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 W52L</td>
<td>MHHHHHHIQRHMETKFQALFDNQPESGELAFKRGTDVTILKDPNWLEGQLNNRGIFPSNYVAPYN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>D8</td>
<td>D9</td>
<td>D10</td>
<td>D11</td>
<td>D11 W52L</td>
<td>1sem.pdb</td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>----------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>α-Helix</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
<td>6%</td>
<td>7%</td>
<td>5%</td>
<td></td>
</tr>
<tr>
<td>AntiParallel-β</td>
<td>33</td>
<td>35</td>
<td>37</td>
<td>36</td>
<td>27</td>
<td>32</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Parallel-β</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Turn</td>
<td>14</td>
<td>15</td>
<td>15</td>
<td>14</td>
<td>20</td>
<td>13</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Random Coil</td>
<td>50</td>
<td>48</td>
<td>46</td>
<td>47</td>
<td>47</td>
<td>48</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>BeStSel / CATH Fold Classification</td>
<td>“Mainly-β”</td>
<td>“Mainly-β”</td>
<td>“Mainly-β”</td>
<td>“Mainly-β”</td>
<td>“Mainly-β”</td>
<td>“Mainly-β”</td>
<td>“Mainly-β”</td>
<td></td>
</tr>
</tbody>
</table>
### Table 4.

<table>
<thead>
<tr>
<th></th>
<th>$\Delta H$ (kcal/mol)</th>
<th>$T_m$ (C)</th>
<th>$\Delta C_p$ (kcal/mol K)</th>
<th>$\Delta G$ (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design 2</td>
<td>31.7</td>
<td>37.4</td>
<td>0.660 ± 0.315</td>
<td>1.1</td>
</tr>
<tr>
<td>Design 8</td>
<td>47.9</td>
<td>69.3</td>
<td>0.669 ± 0.319</td>
<td>4.2</td>
</tr>
<tr>
<td>Design 9</td>
<td>47.2</td>
<td>80.3</td>
<td>0.636 ± 0.322</td>
<td>4.5</td>
</tr>
<tr>
<td>Design 10</td>
<td>35.0</td>
<td>62.0</td>
<td>0.639 ± 0.320</td>
<td>2.5</td>
</tr>
<tr>
<td>Design 11</td>
<td>58.6 ± 5.6</td>
<td>72.3 ± 0.8</td>
<td>0.655 ± 0.317</td>
<td>5.8 ± 0.8</td>
</tr>
<tr>
<td>Design 11 W52L</td>
<td>50.0</td>
<td>52.6</td>
<td>0.635 ± 0.319</td>
<td>3.5</td>
</tr>
</tbody>
</table>