Combinatorial library approach to decipher the stability effects of loop mutagenesis in model protein scaffold via high throughput methods.

Shiladitya Sen

Introduction

In folded globular proteins, loops are regions joining elements of secondary structure that are known to play a variety of roles in the evolution,[1] structure,[2] stability, folding[3] and function[4] of proteins. They are usually the regions where insertions/deletions and mutations in the primary amino acid sequence occur in evolution, giving rise to considerable variations in specificity (as seen for antibody engineering), thermal stability, proteolytic susceptibility[5] and overall protein structure. But understanding the relationship between protein stability and its primary sequence is one of the fundamental problems in chemical biology. Most studies in current literature has focused on the effects on the thermodynamic stability by mutating residues lying in the core of the protein, but there are very few studies to elucidate the effect of loop mutations and insertions/deletions on protein stability. Most computational algorithms have failed to estimate the detailed effects of dynamic loops due to limitations in estimating the force field/potential functions[6] and challenges in incorporating backbone flexibility in such regions. In a Bayesian-model based statistical study[7], it has been shown that the distribution of the backbone dihedral angles in loop residues is a function of the identity and dihedral angle of many neighboring residues in their folded state. Such restrictions make computational modeling a lot challenging and erroneous.

Experimental evidence suggests contrasting evidences about the effects of inserting residues within loops. In one study,[8] insertion of a big loop did not affect the thermal stability of a SH2 domain protein despite the belief that restricting the conformational entropy of a flexible loop would reduce stability. In another study,[2] the length and conformation of the loop for beta-hairpins has to be optimal. A single insertion or a deletion is detrimental for the overall fold and function of the protein. It is known that amino acids glycine and proline are commonly found in loop regions[9] due to their favorable backbone dihedral angles, but another study [10] indicate that these residues only play a crucial role for loop length less than ten residues.

All inferences drawn from the above references involve a limited dataset and the generality of such predictions can only be valid by analyzing a more statistically relevant dataset. In order to analyze large number of mutants we generate combinatorial libraries within a model protein Rop. Rop or Repressor Of Primer is a compact four helix bundle protein that comprises of two monomers (Helix-turn-helix), each of 63 amino acids, arranged in an antiparallel fashion. [11] The biochemical function of Rop is to control the copy number of ColE1 plasmid by interacting and stabilizing the complex formed by two RNA molecules[12]. Extensive biophysical characterization has been carried out on various mutants of Rop to probe their effects on the folding, stability, kinetics and structure. [13-16] In one study,[17] involving mutations of two loop residues D30 and A31 to glycines of varied length it was observed that the mutant with one and two glycine in place of both the loop residues were thermodynamically the most stable. All other mutants with 3-10 glycine residues were substantially destabilized with an inverse correlation between stability and loop length indicating Rop having an optimized tight loop. Other studies,[18] have indicated that a D30G mutation substantially enhances the thermal stability of Rop by 10^{o}C without affecting its overall fold while another mutation[19] A31P changes the overall typology of Rop with one monomer bisecting the other at right angle.

To select optimally folded (functional) variants of Rop from large libraries we utilize a Green fluorescent protein (GFP) dependent *invivo* screen[20] based on the activity of Rop. Functional Rop variants are fluorescent while ill-behaved variants are non-fluorescent. For biophysical characterization we have recently developed a high throughput method[21] to evaluate the stability of 96 variants at a time by utilizing an extrinsic fluorophore. In this study we have utilized a cysteine free version of Rop (AV Rop) as our scaffold where both the cysteines in wild type Rop has been mutated to an alanine and a valine respectively[22] with no change in structure and stability to alleviate problems associated with cysteine residues. In this paper we will refer to AV-Rop as R55.

In this study we have investigated the effect of loop randomization on the thermal stability of a folded protein. We have randomized positions 29, 30, 31 and 32 in context of AV Rop and a variant (R55I) that is itself destabilized. We have also probed into the effects of inserting an extra residue in the loop. Results indicate that all mutants without the additional residue are destabilized in both the AV(R55) and the R55I context while some selected mutants have almost similar or enhanced stability than the starting scaffold by inducing a favorable charge-charge interaction and lowering the ΔCp value or the change in the heat capacity of unfolding.

Results

Construction of library, sequence and stability.

The 1ROP crystal structure (Protein Data Bank) backbone depicts that positions 29, 30, 31 and 32 constitute the turn or loop of the helix-turn-helix scaffold of AV-Rop. AV-Rop is a Cys free mutant of wt Rop[22]. These positions are Leu, Asp, Ala and Asp respectively in Rop. Since these residues were all or mostly surface exposed, these residues were randomized into all 20 amino acids using the degenerate NNK codon at each position. The theoretical size of the library was $1.6x10^5$. The NNK-AV or the NNK-R55 library was cloned into the pMRH6 vector (detailed cloning methodology and selection in the materials and method section) and 48 positives were randomly chosen from the screen[20] that is routinely used to identify active/natively folded variants of Rop. The fraction of positives as visualized from the screening plates was around 15%. The sequencing data suggested a bias towards negatively charged amino acid in position 32. Out of the 48 sequences, 42 had either an Asp or a Glu at that position. There was also a strong preference of hydrophobic amino acids in position 29 with Ala, Val, Leu and Ile being present in 44 of the sequenced variants. The other positions i.e. 30 and 31 had a broad diversity in the distribution of amino acids as was expected. These 48 mutants were purified in a high throughput fashion and their relative stabilities were estimated in a 96 well plate using HTTS[21] that utilizes an extrinsic fluorophore as a probe to measure thermal stabilities of protein. All Tm's were in close proximity to each other with the maximum destabilization (by 7^oC with respect to AV-Rop) observed for the variant TDNE. None of the variants were found to be more stable than AV-Rop. This result provides evidence that during the course of evolution, the loop of Rop has been optimized for the wt sequence fold (helix-turn-helix scaffold).

To attain more flexibility and entropic freedom in the loop, we made another library (NNK-R55 Insertion) by randomizing positions 30, 30a, 31 and 32. 30a is an extra insertion in the otherwise tight loop. Position 29 was not randomized since a closer picture of Leu29 in the 1ROP crystal structure reveals that side chain of Leu to be buried in the hydrophobic core even though the backbone forms a part of the loop. Moreover any positive that had a polar or a non-hydrophobic residue in that position from the previous library was destabilized the most. Again, 48 sequences were randomly picked from the screen, sequenced and characterized to find their Tm's. The variation in Tm's were lower (the most destabilized variant was 4^{o}C lower in Tm than AV-Rop) than the previous library but none of them were more stable than AV-Rop. This indicated that relaxing the strain/entropy of the loop with an additional residue was not enough to incorporate additional stability into the scaffold. The sequencing depicted distribution of polar/charged amino acids in positions 30, 30a and 31 but still there was still a strong preference of negatively charged amino acid in position 32. (39 out of 48 had Asp or Glu). This presented an interesting question about which other position really dictates the amino acid preference at position 32? Close retrospection of the crystal structure reveals a possible electrostatic interaction between Arg 55 and Asp 32. By heptad repeat numbering, position 55 should be a part of the hydrophobic core but it sticks out to the surface and instead Phe 56 positions itself in the core. We constructed a 20 member NNK55 library in the AV-Rop scaffold to identify an amino acid that would get rid of the charge bias without affecting the overall structure and stability. We identified R55I mutant that destabilizes the AV-Rop by 3^{o}C but is overall folded. Since it's a charged to a neutral mutation we expect the charged bias to be eliminated and hence present a better picture of the amino acid distribution within the loop.

We constructed two different libraries, one in which positions 29, 30, 31 and 32 were randomized in the R55I scaffold (NNK-R55I library) and the other in which positions 30, 30a(insertion), 31 and 32 were randomized (NNK-R55I Insertion library). 48 positives from each library were randomly picked and their thermal stabilities were estimated with HTTS. It was surprising to see that every mutant from the NNK-R55I library was destabilized relative to R55I by 0.2-3^{o}C, since nature has optimized the loop of Rop in the context of Arg in position 55 and not Ile. However, in the NNK-R55I insertion library, 23 out of the 48 actives analyzed had T_m's greater than the Tm of R55I. This suggests that the extra inserted residue in the otherwise tight loop is required to compensate for the loss in stability due to the R55I mutation. This also suggests that the existing optimized loop of the Rop is too tight and has evolved irrespective of the wild type scaffold. Sequencing of the 48 variants from both these libraries did not have any sequence bias in position 32 like before. Interestingly for the NNK-R55I Insertion library, there was a strong correlation between positions 30a and 32 i.e. whenever there was a positively charged residue in position 30a there was a negative charge in 32 and vice versa.

Detailed characterization of library.

To understand the detailed thermodynamic parameters of unfolding, we picked four variants from each of the libraries and characterized them individually. The variants were selected based on their physical properties listed in table below. They were characterized biophysically using Circular Dichroism (CD) spectroscopy. The thermal melts of unfolding showed a two-step sigmoidal curve for all indicating that the additional residue in the loop does not affect the kinetic pathway of Rop unfolding. In previous literature, the kinetics of Rop unfolding occurs via a single step from folded dimeric protein to unfolded monomers, while for folding it occurs via a fast bimolecular association step followed by a slow (rate determining) rearrangement of the

Mutant	**Why it was chosen?**	**Mutant**	**Why it was chosen?**
LDVE **(NNK-R55)**	Consensus	LDAQD **(NNK-R55-Insertion)**	Consensus
LKND	Most stable	LDKHE	Most stable
TDNE	No Hydrophobic on position 29	LSVNQ	No Charge on 32
LNAN	No charge on 32	LDAHR	Inversion of charge on 32
LRAE **(NNK-R55I)**	Consensus	LDEQK **(NNK-R55I-Insertion)**	Consensus
SDKN	No Hydrophobic on position 29	LVKHE	Reversal of charge
VNHT	No charged residue	LAVAN	No charge, least stable
IKNR	Two positives	LQEHR	Most stable

individual monomers. Since we did not observe any intermediates or non-sigmoidal unfolding curves using both HTTS and CD, it could be concluded that an additional residue in the strained loop of Rop does not affect the overall kinetics of folding/unfolding. It is the hydrophobic packing of the core residues that dictate the kinetic pathway of folding/unfolding. The T_m's calculated from CD correlated well with the $T_{1/2}$'s. The estimated stability via chemical melts (D_{50} values) also followed the same trend. Since we were potentially measuring a very narrow stability difference among all these mutants (2-6^{o}C thermally), correlation of all the mentioned biophysical methods was a convincing result. The free energy change (ΔG^{H2O}), extrapolated from urea melts at 25^{o}C, also correlated well with thermal stabilities, with higher ΔG^{H2O} values of unfolding corresponding to more stable variants.

To determine the individual contributions of unfolding from Enthalphy (ΔH_u) and Entropy (ΔS_u), temperature dependent urea melts (Gibbs-Helmholtz analysis) was carried out for all these individual mutants. Few interesting trends were observed. Mutants from the same library had a variation in both ΔH and ΔS values for different reasons. For instance in NNK-R55 library LDVE, LKND and LNAN had a positive correlation of ΔH_u with respect to their T_m's but very similar ΔS_u values. But for TDNE where we have a Thr (alcohol) buried in place of Leu (hydrophobe) at position 29 which prefers a hydrophobe, the ΔH_u value is drastically reduced. The change in ΔCp is negligible among all these variants indicating that such mutations did not form or destroy any favorable strong H-bonding or electrostatic interactions. It is well established that a reduction in ΔCp in protein indicates a formation of a strong H bonding network or an electrostatic like interaction that enhances the ΔG^{H2O} of unfolding.

In NNK-R55-Insertion library, where we have an extra residue in the loop, there is a small rise in the ΔH_u when compared to the variants from the NNK-R55 library indicating some kind of an additional favorable non-covalent interaction in the folded state that is absent in the NNK-R55 library variants. Again within the variants LDAQD, LSVNQ and LDAHR there was a positive correlation between ΔH_u and T_m , with both ΔH_u and ΔS_u values being similar. Surprisingly the variant LDKHE had a lower ΔC_p value and a higher $\Delta H_{u,}$ and ΔS_u value with respect to the other variants within the same library, indicating some form of a strong non-covalent interaction formed in the folded state of the protein. Also for LDAQD the ΔCp value was lower than that of LSVNQ and LDAHR since in the last two variants we have broken the favorable 32-55 interaction.

For the variants of R55I-NNK library, variants LRAE, VNHT and IKNR have a lower ΔH_u and ΔS_u values compared to the mutants LDVE, LKND and LNAN from the R55-NNK library. This further indicates the loss of the favorable interaction between position 32 and R55. A similar lowering of ΔH_u is also observed between R55 and R55I. There is also an increase in the ΔC_p

Scaffold	**Mutant**	**T_M (^{o}C)**	**$T_{1/2}$ (^{o}C)**	**D_{50} (M)**	**m**	**ΔG^{H_2O} (25^{o}C) (kcal/mol)**	**ΔH_u (kcal/mol)**	**ΔC_p (kcal /mol*deg)**	**ΔS_u (kcal /mol*deg)**
	R55	70	60.2	4.2	1.6	11.8	58	1.12	0.82
	R55I	64.3	55.9	3.5	1.4	8.4	39	1.31	0.61
R55-NNK	LDVE	68.2	59.2	3.8	1.2	10.2	49	1.16	0.72
	LKND	69.4	59.8	3.9	1.5	11.1	55	1.13	0.79
	TDNE	62.4	53.9	3.4	0.9	7.8	34	1.21	0.54
	LNAN	66.1	56.2	3.6	1.1	9.1	46	1.28	0.70
R55-NNK (Insertion)	LDAQD	68.2	59.1	3.9	1.4	10.6	55	1.13	0.78
	LDKHE	69.9	59.8	4.1	1.6	11.8	69	0.87	1.02
	LSVNQ	66.2	56.3	3.6	1.2	9.4	49	1.29	0.68
	LDAHR	64.1	55.1	3.5	1.1	8.6	40	1.35	0.64
R55I-NNK	LRAE	63.7	55.1	3.4	1.3	8.2	38	1.30	0.60
	SDKN	60.8	52.2	2.9	0.8	6.4	19	1.38	0.31
	VNHT	61.9	53.7	3.1	0.9	7.1	32	1.34	0.52
	IKNR	62.8	54.2	3.2	1.1	7.6	35	1.33	0.56
R55I-NNK (Insertion)	LDEQK	67.8	58.7	3.7	1.3	9.8	66	0.97	0.97
	LVKHE	67.3	58.5	3.7	1.4	9.4	59	1.01	0.89
	LAVAN	63.3	54.6	3.2	0.7	8.1	34	1.34	0.53
	LQEHR	69.1	59.3	4.0	1.6	10.6	64	0.92	0.94

values of the mutants when compared to the mutants from R55-NNK library which further validates the loss of the favorable interaction, resulting in destabilized mutants. Similar trend is observed when we compare SDKN and TDNE, where both mutants have an alcohol containing residue instead of a hydrophobic residue in position 29. The change in the ΔC_p values within the mutants from the library was negligible suggesting none of them had lost or gained any strong non-covalent interactions.

For the variants of R55I-NNK-Insertion library, LDEQK, LVKHE and LQEHR had a lower ΔCp, a higher ΔSu and a higher ΔHu value than LAVAN. As mentioned earlier, similar result was obtained for LDKHE. This strongly suggests a strong interaction between the second and the fourth residue (i.e. position 30a and position 32) in both the insertion libraries. This also explains in the sequence data analysis that why many sequences in the R55I-NNK-Insertion library had a strong preference of charged residues in position 30a and 32. Limited number of variants with such strong correlations was also found within R55NNK insertion library, but all had a negative charge at 32 since arginine was present in position 55.

Materials and methods.

Library construction

Four different loop libraries of Rop variants were used for this study. The NNK-AV library consists of variants in which residues 29, 30, 31, and 32 were randomized to all the 20 amino acids. The NNK-AV Insertion library were randomized with all the 20 amino acids in position 30, 30a, 31 and 32, where 30a implies an additional residue. Two similar libraries, NNK-R55I and NNK-R55I Insertion were constructed, with respect to NNK-AV and NNK-AV Insertion respectively. The first two libraries were synthesized using an engineered cysteine-free Rop sequence, AV-Rop. The last two libraries used the same AV scaffold that had an additional R55I mutation. All these four libraries were created by initial PCR overlap reassembly of four synthetic oligonucleotides (table SX), that each represent almost one-third of the approximately 200 basepair gene. The four randomized positions for each library were thus located on just one of the oligonucleotide. NNK codon was used for each of the randomized position that encodes for all the 20 amino acids but with only one stop codon. Following the PCR reassembly with Phusion polymerase for 5 cycles at an annealing temperature of 55^{o}C, the product was PCR amplified for 25 cycles (again with Phusion polymerase) with the pMRTEV5 and pMRH6TEV3 primers (table SX) that appended the restriction site and TEV protease sequences to the gene library. This purified and digested PCR product was ligated between the Afl III and BamHI sites of pMRH6, fusing an hexahistidine tag amino-terminal of the TEV cleavage site. To ensure high-efficiency subcloning, a digestion with EcoRI and Nde I was used to eliminate stuffer background as evidenced by a negative control ligation lacking insert. The cloned library in pMRH6 was transformed into DH10B (DE3) containing pUCBADGFPuv for confirmation of active phenotype, as well as high-throughput expression and sequencing. *E.coli* strains, DH10B, BL21(DE3) cells were purchased from Stratagene. The DH10B (DE3) strain of *E. coli* was lysogenized previously in the Magliery lab with the DE3 lamboid phage using a Novagen Individual kit.

High-throughput screening and sequencing

Individual libraries transformed into DH10B (DE3) containing pUCBADGFPuv (Amp resistant) was grown on a LB-kan-amp-0.0005% arabinose (KAA) agar plate at 42^{o}C for 16 hours to screen for the *in vivo* activity. According to the cell based screen, an active Rop variant displays a phenotype depicting high fluorescence, while an inactive variant does not fluoresce. We applied this method to sort active and inactive variants from each of the libraries. 48 variants from each library, that showed high fluorescence, were picked for high throughput protein expression.

These active variants were plated again onto 0.0005% arabinose (KAA) agar plate for sequencing. These colonies, exhibiting obvious fluorescence, were then picked with pipette tips and resuspended in 25 μl of sterile H_20. This suspension was then spotted onto LB-kan plates in 2 μl volumes using an 8-channel pipette manifold to create an array of 48 clones on a single agar plate. Sequencing directly from the spotted clones was done by Genewiz Inc. (South Plainfield, New Jersey).

High-throughput protein purification

Individual Rop variant seeds in DH10B (DE3) were grown in 1.5 mL 2YT media in each well of a 2 mL, 96 square deepwell TiterBlock plate (USA Scientific) covered with a porous membrane at 37 °C for 18 h. The seeds were diluted to 2 mL 2YT and OD_{600} = 0.75-1.0, induced with 10 μM IPTG, and overexpressed at 30 °C for 18 h. Cell pellets were resuspended in 200 μL lysis buffer and lysed by adding 100 μg mL^{-1} lysozyme, 0.5 μg DNase I, 40 ng RNase A, 5 mM $MgCl_2$, 0.5 mM $CaCl_2$, and 20 μL PopCulture reagent (Novagen), followed by incubation at RT for 30 min. The plate was then covered with an Axymat (Axygen) and was incubated for atleast one hour with occasional vortexing. Soluble fractions were mixed with 50 μL NiNTA magnetic beads (Qiagen) and incubated at RT for 1 h. The bound resin was washed (lysis buffer with 20 mM imidazole) and resuspended in 25 μL lysis buffer. The proteins were cleaved off the resin by 10 μg TEV protease (in 0.5 μl) and 11 mM βME with incubation at 30 °C for 3 hours.

High-Throughput Thermal Scanning and data fitting.

The SYPRO™ Orange dye (Invitrogen, Carlsbad, CA) was supplied in DMSO at 5000× the working concentration for PAGE staining. Spectra were obtained on a Bio-Rad CFX96 thermal cycler Real-Time Detection System. Samples of 20 µL per well were prepared by mixing 1 □L of 200× SYPRO™ Orange (final concentration 10×) with 19 □L of protein (0.1 mg mL^{-1} or 2.1 µM), in lysis buffer (50 mM Tris•HCl, 300 mM NaCl, 10 mM imidazole, pH 8), loaded into Bio-Rad 96-well 0.2 thin-wall PCR plates, and sealed with optical quality sealing tape (Bio-Rad). Thermal denaturations (ramp rate of 1 °C min^{-1} at 0.2 °C intervals with an equilibration of 5 seconds at each temperature prior to measurement) were acquired by measuring fluorescence intensities using the FRET channel with excitation from 450-490 nm and detection from 560-580 nm. All data were exported and plotted in Microsoft Excel 2007.

For HTTS data processing, melting data were normalized and truncated from their initial room-temperature fluorescence values to their maxima. We then fitted the data to a modified Clarke & Fersht equation (1), which accounts for a non-flat pre-transition baseline. Here, α_F and β_F are the intercept and slope of baseline for the folded state, and m is an exponential factor associated with the slope of the transition at the apparent melting temperature T_m.

$$Signal = \frac{(\alpha_F + \beta_F T) + e^{m(T-T_m)}}{1+ e^{m(T-T_m)}} \quad \mathbf{(1)}$$

All the parameters α_F, β_F , m and T_m were evaluated by least squares fitting using the Solver plug-in of Microsoft Excel by minimizing the sum of the squared differences between each data point and each fitted point.

Derivative plots (change in fluorescence per change in temperature) of both the raw and the fitted data were evaluated using the slope as determined from a sliding seven point window around each temperature value. The derivative plot of the raw data was then fitted to a standard Gaussian equation (2)

$$Fitted\ value\ = a * e^{-\frac{(T-b)^2}{2c^2}}\ \mathbf{(2)}$$

in which the constant *a* defines the height of the curve's peak, *b* defines the position of the center of the peak (this point corresponds to the temperature where the derivative attains its maximal value) and *c* defines the width of the Gaussian. All the constants *a*, *b* and *c* were evaluated by least square fitting using the Solver plug-in of Microsoft Excel.

The Gaussian fitted derivative plot of the raw data matched well to the derivative plot of the data that was fitted to the Clark & Fersht equation. This indicated that the temperature corresponding to *b,* as computed from the Gaussian fit, correlated well with T_m, calculated from the Clarke & Fersht equation.

Protein Purification of selected variants

Rop variants were overexpressed in BL21(DE3) in 1L 2YT media grown to an OD_{600} = 0.7-0.9, induced with 0.1 mM IPTG, and incubated at 30°C for 14-16 h. Cells were harvested by centrifugation and the cell pellet was resuspended in lysis buffer (50 mM Tris•HCl, 300 mM NaCl, 10 mM imidazole, 2 mM βME , pH 8). Cells were lysed by adding 0.1 mg mL^{-1} lysozyme, 2 μg mL^{-1} DNase I, 200 ng mL^{-1} RNase A, 5mM $MgCl_2$, 0.5 mM $CaCl_2$, and 0.1% Triton X-100. The cell suspension was sonicated and centrifuged at 30,000 *g*. The soluble fraction was mixed with 1500 μL NiNTA agarose slurry (Qiagen) and incubated for 1 h at 4 °C. The bound resin was washed (lysis buffer with 20 mM imidazole), and the protein was eluted with 3 mL elution buffer (lysis buffer with 300 mM imidazole). The His_6 fusion tag was cleaved twice with 0.5 mg of rTEV protease (in 100 μl) with addition of 5 mM DTT, followed by overnight incubation at RT or 30 °C for 3 h. After diluting to 5 mL, the protein was exchanged into lysis buffer using two PD10 column (GE Healthcare) and mixed with 1000 μL NiNTA agarose slurry at 4 °C. After incubation for 1 h at 4°C, the filtrate containing the free protein was stored in lysis buffer. The protein was concentrated by centrifugation through a YM3 filter (Millipore) and exchanged into the appropriate buffer using a PD10 column for subsequent characterization by CD spectroscopy.

CD Spectroscopy (Thermal and Urea melts)

Spectra were obtained on an Jasco J-815 spectrometer (Ohio State University Department of Chemistry). Experiments were conducted at 50 μM protein monomer, determined by UV absorption at 280 nm, in CD buffer (50 mM sodium phosphate, 300 mM NaCl, pH 6.3). Thermal denaturations were acquired with a ramp rate of 1 °C min^{-1} at 222 nm, 25 to 90ºC, 6 second equilibriation at every temperature. Urea denaturations were performed in the same conditions but with 0, 1, 2, 3, 4, 5, 6, or 7 M urea, with spectra acquired after equilibrating 28h at RT, at 222 nm. Urea stock solutions were made at approximately 9.5 M by dissolving the appropriate quantity in diluted 10X CD buffer and determining the exact concentration by refractometry. For temperature dependent Urea denaturation at 222 nm, samples with similar conditions were prepared at 1.2, 2.2, 3.3, 4.3, 5.4 and 6.4M Urea. The samples were equilibrated for 28 hours at RT. Each sample were further equilibrated at every temperature (25^{o}C, 30^{o}C, 35^{o}C, 40^{o}C, 45^{o}C and 50^{o}C) for 20 minutes before acquiring the scan, going from low to high values of temperature.

$\Delta G^{H}2^{O}$ estimation and stability curves of proteins

To calculate the $\Delta G^{H}2^{O}$, the thermodynamic stability the protein in buffer (i.e., zero denaturant), we followed the method of Dalal *et al.*, assuming an equilibrium between the folded dimer (N) and two unfolded monomer (U).

$$N2 \Leftrightarrow 2U \qquad \textbf{(3)}$$

The equilibrium constants for each urea concentration D were determined from fraction unfolded (F_U), fraction folded (F_N), and total protein concentration (C_T) using:

$$K_D = C_T \frac{F_U^2}{F_N} \qquad \textbf{(4)}$$

These were converted to ΔG values from –RT $\ln K_D$ and the $\Delta G^{H}2^{0}$ and m-values were determined by fitting a line:

$$\Delta G = \Delta G^{H_2O} - m[D] \qquad \textbf{(5)}$$

Such $\Delta G^{H}2^{0}$ values were computed at six different temperatures (25°C, 30°C, 35°C, 40°C, 45°C and 50°C) to fit into the Gibbs-Helmholtz equation (G-H equation) below to equate the ΔC_p (Molar heat capacity at constant pressure) and change in enthalphy of unfolding (ΔH_u) at T_m.

$$\Delta G^{H_2O} = \Delta H_u \left(1 - {}^{\mathrm{T}}/_{\mathrm{T_m}}\right) - \Delta C_p \left((\mathrm{T_m} - \mathrm{T}) + \mathrm{T} \times \ln\left({}^{\mathrm{T}}/_{\mathrm{T_m}}\right)\right) \qquad \textbf{(6)}$$

Where T_m is the melting temperature at which 50% of the protein is unfolded and T is any temperature. At T_m since $\Delta G^{H}2^{0} = 0$, hence $\Delta S_u = \Delta H_u/T_m$.

Discussion

Investigations based on functional screens done previously depict that random mutagenesis in the loop usually results in a fully folded functional variant of the protein,[23] but the effects on stability due to such mutations were unknown. In this study, using the highly efficient cell based screen for Rop to identify natively folded variants we have also gained insight about the effects on overall thermodynamic stability by changing the primary structure of the loop. Our model protein, Rop is a simple well studied stable protein so correlating the sequence-stability relationship is more readily revealed. We have addressed three essential questions. Can randomization and additional insertion of a naturally evolved loop result in folded/active proteins? How and by what extent can such mutations affect protein stability without affecting the fold? Can such randomizations compensate or enhance the stability of a destabilized variant? To answer these questions we have made two versions of combinatorial libraries (one by varying positions 29, 30, 31 and 32 and the other by varying positions 30, 30a, 31 and 32 where 30a is an insertion) by utilizing R55 and a destabilized R55I variant as scaffolds.

For libraries in which four residues were randomized without an insertion, positions 29 (in the context of both R55 and R55I) and 32 (in the R55 context) had a strong bias for the wild type residue at that position. This illustrated that 29 is more a core residue with its side chain involved in packing within the core (as evident from the crystal structure and the previous study where only positions 30 and 31 were mutated as core residues[17]), while it is the electrostatic interaction between R55 and D32 that forces position 32 to prefer a negatively charged residue. The spread of T_m's for both the libraries were low (6-7^{o}C) and none of the variants were more stable than the starting scaffold, even for R55I which was itself destabilized than R55. This indicates that the loop of Rop is amazingly tight (optimized) and has evolved over time independent of D32-R55 interaction since the wild type sequence of loop in R55I context was the most stable variant and the rest were all destabilized.

To eliminate the charge bias at position 32, we identified variant R55I as a scaffold that was active from the screen and was 7°C destabilized than R55 from the R55NNK library. R55Q, which was an inactive (unfolded or misfolded) from the screen, was also used as a scaffold for both the versions of the library but no variants were found that passes the screen, implying loops alone cannot restore the fold and function in Rop.

For the insertion library, in R55 the spread of T_m's was even lower (3-4°C) and none of the selected variants were more stable than the starting scaffold, indicating the additional residue that introduces flexibility in the tight loop does not help to stabilize. Interestingly in the R55I scaffold, there were more than 28 out of the 48 sequences analyzed that were 6-7°C more stable than R55I but none were more stable than R55. More specifically all these 28 sequences had a stabilizing electrostatic interaction between positions 30a and 32 that occupied oppositely charged amino acids. Even for the R55 scaffold, the most stabilized variants had similar stabilizing salt bridge interactions but none had a positively charged residue in position 32 due to the presence of arginine at position 55. But none were more stable than R55 probably due to a stronger D32-R55 interaction. The plausible cause could be that 32 and 55 are distant in primary structure and hence in the unfolded state there lies no favorable interaction between them, whereas 30a and 32 are close in sequence space and even in the unfolded state might possess some interaction.

To further probe the detailed thermodynamic contribution of such stabilizing electrostatic interaction, we carried out Gibbs-Helmholtz analysis on selected variants from all the four different libraries. It is evident between R55 has a higher ΔH_u, lower ΔC_p value and a higher ΔS_u

than R55I indicating the favorable D32-R55 interaction. For all variants that had the potential favorable salt bridge within the insertion library in both R55 and R55I scaffold, there was a considerable decrease in ΔC_p value along with an increase in ΔS_u value. Interestingly the ΔΔHu value (19kcal/mol) between R55 and R55I was greater than ΔΔHu value (13 kcal/mol) between LDKHE (one with 30a and 32 salt bridge) and LDAQD (one without the salt bridge). This indicated that R55 and D32 being far away in the primary sequence has a greater impact on enthalphy change when the interaction is removed compared to 30a and 32 that lie much close in sequence space. Although it's conceived that the ΔΔHu difference between 30a and 32 would be minimal due to its close proximity in primary sequence but a previous study[7] has illustrated that for small loops (<8 residues) the backbone dihedral angles are function of the neighboring residues in the folded state. Such neighboring residues (that form favorable interactions within the core) can force the side chains within the loop to interact in a specific orientation which is absent in the unfolded state. Such interactions are weaker than R55-D32 as indicated by $\Delta\Delta H_u$ value differences and hence we could not find any mutant having the 30a-32 interaction being stable than R55.

The result of our study on loop randomization has important consequences and ramifications for designing stable proteins with loop engineering. This is one of the most exhaustive studies on sequence-stability relationship within a loop without affecting the overall fold/function of the protein Though we focus upon a single protein Rop, our results should have applicability in systems with α-helical, β-sheet, or mixed secondary structure provided the loop under consideration is small (less than 10 residues) and there lies considerable intraloop and loop-protein contacts. Correct orientation of the side chains and favorable interactions within the loop can enhance the stability without affecting the function.

References

(1) Panchenko, A. R. and Madej, T. *BMC Evolutionary Biology* **2005**, *5*, 1.

(2) Sibanda, B. L. and Thornton, J.M. *Nature* **1985**, *316*, 170.

(3) Ybe, J. H. *Protein Sci.* **1996**, *5*, 814.

(4) Fetrow, J. S., Cardillo, T.S. and Sherman, F. *Proteins* **1989**, *6*, 372.

(5) Pickersgill, R., Varvill, K., Jones, S., Perry, B., Fischer, B., Henderson, I., Garrard, S., Sumner, I. & Goodenough, P., *FEBS Lett.* **1994**, *347*, 199.

(6) Potapov V. *PEDS* **2009**, *22*, 553.

(7) Ting D, Shapovalov M, Mitra R, Jordan MI, Dunbrack RL Jr. *PLoS computational biology* **2010**, *6*, 100763.

(8) Scalley-Kim M. and Baker, D., *Protein sci.* **2003**, *12*, 197.

(9) Wilmot, C. M. and Thompton, J.M., *JMB* **1988**, *203*, 221.

(10) Florian Krieger and Thomas Kiefhaber *J. Am. Chem. Soc.* **2004**, *127*, 3346.

(11) Banner D.W. and Tsernoglou D *JMB* **1987**, *196*, 657.

(12) Castagnoli, L., Scarpa, M., Kokkinidis, M., Banner, D.W., Tsernoglou,; D. & Cesareni, G. *EMBO J* **1989**, *8*, 621.

(13) Munson M, Fleming K.G., Nagi, A.D., Sturtevant J.M., Regan, L. *Protein Sci.* **1996**, *5*, 1584.

(14) Munson M, and Regan L *Folding & Design* **1997**, *2*, 77.

(15) Munson M, O. B. R., Sturtevant JM, Regan L *Protein Sci.* **1994**, *3*, 2015.

(16) Nagi AD, A. K., Regan L *JMB* **1999**, *286*, 257.

(17) Athena, D. N. and Regan, L. *Folding & Design* **1996**, *2*, 67.

(18) Paul, F. P., Agrawal,V., Axel, T., Brunger and Lynne, Regan *Nature Structural Biology* **1996**, *3*, 55.

(19) Glykos, N. M., Cesareni, G., Kokkinidis, M. *Structure* **1999**, *7*, 597.

(20) Magliery, T. J. and Regan, L. *Protein Eng. Des. Select.* **2004**, *17*, 77.
(21) Lavinder, J. J. Hari, S.B., Sullivan, B.J. and Magliery, T.J. *J. Am. Chem. Soc.* **2009**, *131*, 3794.

(22) Hari, S. B. B., C.; Lavinder, J.J. and Magliery, T.J. *Protein Sci.* **2010**, *19*, 670.

(23) Brunet, A. P., Huang, E.S., Huffine, M.E., Loeb, J.E., Weltman, R.J. and Hecht, M. H. *Nature* **1993**, *364*, 355.

Author affiliations

Shiladitya Sen[1] and Thomas J. Magliery[1,2]

[1]Department of Chemistry and Biochemistry, The Ohio State University, 100 West 18th Avenue, Columbus, OH 43210, USA

[2]Ohio State Biochemistry Program, The Ohio State University, 100 West 18th Avenue, Columbus, OH 43210, USA

Corresponding author: Magliery, Thomas J. Email-(magliery@chemistry.ohio-state.edu)