Chapter 4

Development and Validation of Methods for Multi-State Design and Combinatorial Library Design

Adapted from a manuscript coauthored with Alex Nisthal and Stephen L. Mayo.

Abstract

The stability, activity, and solubility of a protein sequence are determined by a delicate balance of molecular interactions in a wide variety of conformational states, including competing states and native conformational states. Even so, most computational protein design methods model sequences in the context of a single conformation representing the native state. Despite the potential for improved simulation accuracy when the native state is represented by an ensemble of related structures, such calculations have not been attempted due to the lack of sufficiently powerful optimization algorithms for multi-state design. Here, we have applied our multi-state design algorithm to study the potential utility of various forms of input structural data for design.

To facilitate this analysis, we developed new methods for the design and high-throughput stability determination of combinatorial mutation libraries based on protein design calculations. The application of these methods to the core design of a small model system produced many variants with improved thermodynamic stability, and showed that multi-state design methods can be applied to large structural ensembles without requiring the use of different rotamer libraries, energy functions, or design strategies. Stabilized variants were found in libraries based on each type of structural data we tested. Our library design method produced degenerate codon libraries that represented the underlying design calculations, and exhaustive screening of these libraries helped to clarify several sources of error in our designs that would have otherwise been difficult to ascertain.

The complete lack of correlation between our experimental and simulated stability values shows clearly that a design procedure need not reproduce experimental data

directly to generate many successful variants. This surprising result suggests a potential new direction for the improvement of protein design technology.

Introduction

During the past two decades, protein-engineering efforts based on directed evolution have met with considerable success.[1–3] In tandem, structure-based computational protein design (CPD) methods have been developed to allow screening for desirable sequences to be performed *in silico*.[4–6] Despite a number of high-profile results that demonstrate the potential of CPD,[7–14] the routine computational design of functional proteins remains elusive. Thus, many current efforts focus on the improvement of CPD methodology or on the synergistic application of CPD with experimental high-throughput screening or selection.[15] These lines of inquiry need not be orthogonal; the computational design and experimental screening of mutant libraries can facilitate a more thorough evaluation of CPD than studies that focus on the comparison of individual designed sequences.

Here, we have applied this type of hybrid approach to investigate the degree to which X-ray crystallographic structures, NMR solution structures, and ensembles derived from molecular dynamics simulations can serve as useful sources of structural information for CPD. This study was made possible by the development of new methods for the computational design and high-throughput experimental stability determination of combinatorial protein libraries. The results we report here provide simultaneous experimental validation for (1) the application of multi-state protein design methods to large conformational ensembles, (2) the transformation of arbitrary CPD results into combinatorial mutation libraries, and (3) the experimental stability determination of these

libraries by high-throughput gene assembly, protein expression, purification, and screening.

Our work here was motivated by a desire to address one of the major approximations of CPD: the reliance on a single, rigid main-chain conformation. Although the stability, solubility, and activity of a protein depend on the relative energetic contributions of many conformational states, including ensembles of native, unfolded, and aggregated structures,[16] most CPD methods evaluate sequences based on their energies in the context of one fixed backbone structure. This simplification has made design results undesirably sensitive to slight changes in main-chain and side-chain conformation, and has made difficult the selection of sequences with amino acid composition similar to naturally occurring protein. These issues have been approached via the use of high-resolution structural templates, expanded rotamer libraries,[17, 18] energy functions with softened repulsive terms,[11, 19, 20] iteration between structural refinement and sequence design,[11, 21] and composition-based reference energies.[11, 22] Although these strategies can help to mitigate the impact of the fixed-backbone approximation, they do not address the fundamental reality that protein fitness depends on a diverse range of conformational states.

In a handful of cases, multi-state design (MSD) procedures have been used to find sequences that simultaneously stabilize or destabilize a combination of a few different conformational states.[23–25] However, MSD techniques have not yet been applied to ensembles with many conformational states that might better reflect the flexibility of real proteins. The degree to which various energy functions, rotamer libraries, and structural templates of single-state design (SSD) might be appropriate for this type of MSD

calculation is heretofore unknown. We recently developed a framework for MSD that allows for efficient sequence optimization given hundreds of conformational states. Here, we have applied this framework to test the applicability of current CPD methods to large structural ensembles, and to investigate whether the use of such ensembles might result in the selection of more desirable sequences by CPD.

With limited exceptions,[26] a unique native state with at least marginal stability is required for protein function as we understand it today. Accordingly, the most basic goal of CPD has been to optimize interactions between amino acids side chains to promote thermodynamic stability of the native state. Unfortunately, the experimental validation of a new design procedure on this basis is often beset with uncertainty. Standard methods for the measurement of protein stability are too laborious to allow the testing of more than a few designed variants, and the top-scoring sequence produced by a new design procedure does not (yet) sufficiently reflect its general utility. To facilitate the experimental evaluation of larger numbers of designed sequences, higher throughput is required in the assembly of genes, the expression and purification of proteins, and the measurement of stabilities. Fortunately, recent progress in these areas has allowed us to construct an efficient pipeline for the basic evaluation of new procedures in CPD. Gene libraries assembled from degenerate oligonucleotides, a frameshift selection scheme that reduces contamination by erroneous genes,[27] and economical sequence verification make tenable the production of numerous specific designed genes. Commercial microtiter plates for the growth of expression cultures and the purification of hexahistidine-tagged proteins allow sufficiently pure protein to be produced easily from these genes. Finally, liquid-handling robotics[28] expedites the preparation of a chemical denaturation series for

each protein in 96-well format, and the fraction of protein unfolded in each well is assayed in a plate reader measuring tryptophan fluorescence.[29] The integration of these technologies has allowed us to assess the stability of hundreds of designed protein variants with minimal experimenter intervention and limited incremental expense.

Given several design procedures to evaluate and a high-throughput experimental assay, we needed a general and rigorous method to choose a limited number of representative sequences to test from each design. Fortunately, structure-based computational protein design methods have been enlisted previously to focus high-throughput screening and selection on desirable subsets of sequence space. For example, CPD can be used to help identify positions amenable to site-saturation mutagenesis[30] and site-directed recombination.[31, 32] When a protein engineering effort is intended to help evaluate CPD procedures, as in this case, designed combinatorial mutation libraries are more appropriate because they reflect more strongly the sequence preferences of CPD. Although several useful computational protein library design methods have been developed, none reported so far takes directly into account CPD energies, allows control over library size and possible sets of amino acids, and eschews heuristics that can introduce bias into the libraries it produces.[33–36] So that our experimental results might better reflect the results of the underlying CPD calculations, we developed a new library design procedure, called Combinatorial Libraries Emphasizing And Reflecting Scored Sequences (CLEARSS), which satisfies all of these criteria.

We used standard single-state design (SSD) and MSD to redesign the core of the small, stable domain Gβ1 based on several sources of structural information, including a crystal structure, an NMR structure, and MD simulations. Our efforts were motivated by

a curiosity about the relative merits of different sources of structural data for design, and the hypothesis that use of a structural ensemble might help to correct for design failures observed in SSD. Because the imperfect nature of CPD limits the conclusions that can be drawn from a comparison of single sequences, we developed the CLEARSS algorithm to make combinatorial libraries based on the lists of scored sequences produced by CPD. We applied this algorithm to the results of our design calculations, and assayed the designed libraries using a new protocol for the expression, purification, and stability assessment of protein libraries with high throughput.

We found that all three sources of structural data resulted in designed libraries with multiple stabilized variants. The designed libraries based on an NMR ensemble were extremely similar, whether a single representative structure or all 60 ensemble members were used for modeling. The most promising results by far were achieved using a constrained 128-member MD-ensemble, which produced a designed library with no significantly destabilized and many stabilized variants. Despite the apparent success of this design, there was no correlation observed between the simulation energies and the experimental stabilities of any of these variants.

Our results suggest that the basic principles of CPD extend beyond the design of single sequences to the design of combinatorial libraries, and that the rigorous screening of such libraries can help to pinpoint sources of error in a design procedure. They show that MSD methods are applicable to large structural ensembles when used with standard rotamer libraries and energy functions, inspiring optimism about more ambitious future applications for MSD. They also hint that the use of structural ensembles could help to alleviate problems that occur when targeting a single, fixed input structure. Furthermore,

they illustrate clearly that the success of CPD does not hinge on its ability to directly correlate simulation energies with experimental measures of fitness. This surprising property of CPD may suggest a new possible direction of inquiry for the improvement of CPD.

Results and discussion

Designed libraries

To simplify the validation of our multi-state design and combinatorial library design methods, we applied them to a previously studied set of core positions (Figure 1) in the small model system Gβ1, and relied on a set of energy functions that previously found stabilized variants based on this design.[19] Given the requirements for purified protein of our stability assay, we chose to design and screen a 24-member library based on each of the following sources of structural information: a crystal structure (xtal-1), an NMR-constrained minimized average (NMR-1), an NMR ensemble (NMR-60), a constrained MD ensemble (cMD-128), and an unconstrained MD ensemble (uMD-128).

The sequence of steps used to design the combinatorial libraries we tested experimentally is depicted in Figure 2. First, the standard design procedure was applied to each structural input, and optimization was performed with SSD-FASTER or MSD-FASTER to give a list of amino acid sequences and their CPD energies for each design. The CLEARSS library design algorithm was then applied to each list of sequences to give a rank-ordered list of combinatorial mutation libraries. All amino acid sequences in each of the top 20 CLEARSS libraries were instantiated and evaluated by rotamer optimization. The CLEARSS library to test experimentally for each structural input was chosen by objective criteria based on the energies of the rescored sequences, as described in the methods section.

All five designed libraries comprise relatively conservative sets of mutations away from the wild-type sequence (Table 1). All libraries other than uMD-128 share many characteristics in common. Each of these libraries chose only the wild-type amino acid at positions A20, A26, F30, and A34. Every member of each of these four libraries contained the single-mutant Y3F, which previous experiments have shown to be well tolerated by the structure. These four libraries all allowed the wild-type amino acid at every other position, and all contain the most stable Gβ1 core variant previously characterized (Y3F+L7I+V39I).

The two NMR libraries were extremely similar to each other: both chose the amino acids FILV at position 52, and directed the remaining diversity to positions 7 and 39. In contrast, xtal-1 and cMD-128 allowed only the wild-type Phe at position 52, and instead allocated diversity towards positions 7, 39, and 54. xtal-1 differs from cMD-128 in that it gave up L7F and V39L to allow L5I. The unconstrained MD ensemble library uMD-128 was the least conservative, specifying a size reversal of two nearby residues via mutations L5A and A34F, and diversity at residue 30, a position untouched in the other libraries.

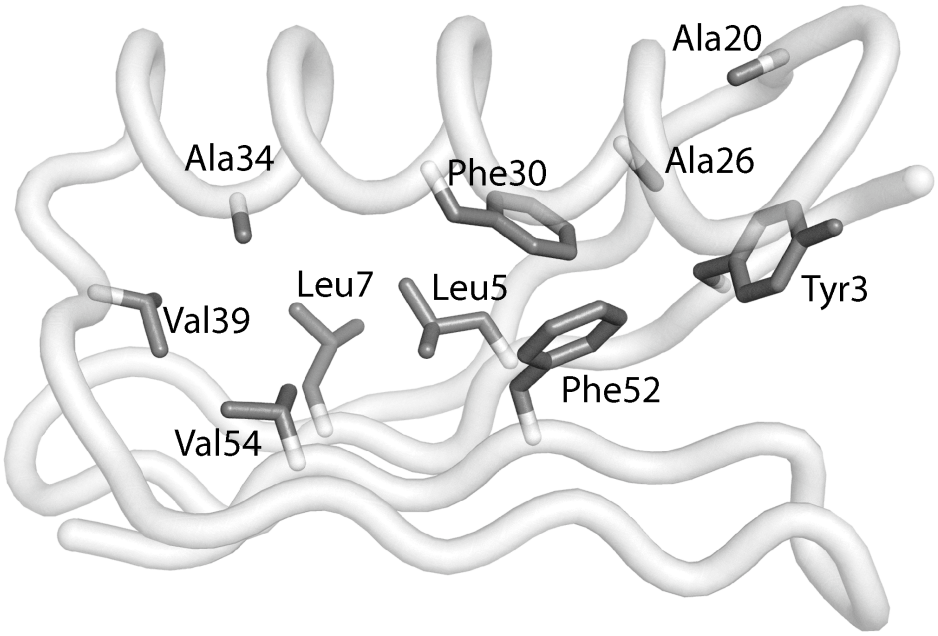

Figure 1: The core residues of Gβ1 designed in this study. Each of these positions was allowed to assume various rotamers of the hydrophobic amino acids Ala, Val, Ile, Leu, Phe, Tyr, and Trp. Position Trp43 (not shown) was additionally allowed to change rotamer but not amino acid type. All other side chains and the main chain were fixed in the input conformation for the state being modeled in each case.

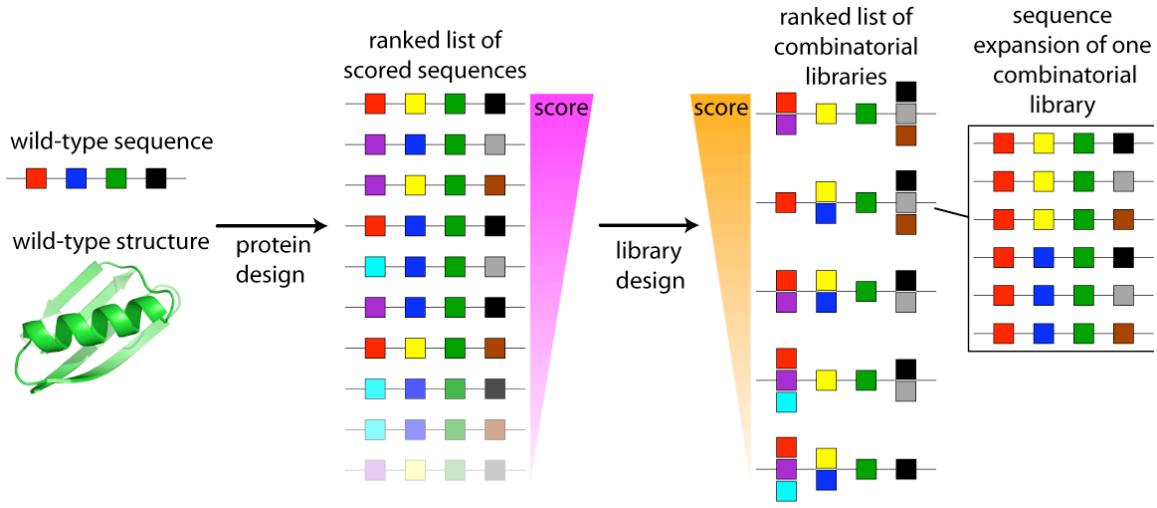

Figure 2: The general scheme used to design combinatorial mutation libraries based on computational protein design calculations. A line of boxes indicates a protein sequence; each box represents a position in the protein chain. Different colored boxes represent different amino acids. The set of sequences on the far right represent the expansion of a particular combinatorial library into the set of sequences it represents. The energies of the sequences in the expansions are used to decide which combinatorial library to test experimentally, as described in the Methods section.

Table 1: Combinatorial libraries designed from different sources of structural information. **xtal-1**: the designed library based on single-state design of the crystal structure. **NMR-1**: the library based on single-state design of the constrained minimized average NMR solution structure. **NMR-60**: the library based on multi-state design of the 60-member NMR structural ensemble. **cMD-128**: the library based on multi-state design of the constrained molecular dynamics ensemble. **uMD-128**: the library based on multi-state design of the unconstrained molecular dynamics simulation.

Residue	WT	xtal-1	NMR-1	NMR-60	cMD-128	uMD-128
3	Y	F	F	F	F	F
5	L	IL	L	L	L	A
7	L	ILV	ILV	IL	FILV	FL
20	A	A	A	A	A	A
26	A	A	A	A	A	A
30	F	F	F	F	F	FIL
34	A	A	A	A	A	F
39	V	IV	IV	ILV	ILV	IL
52	F	F	FILV	FILV	F	F
54	V	IV	V	V	IV	AV

Experimental characterization of designed libraries

Each library was constructed using a modification of the traditional gene assembly protocol[37] that minimizes oligonucleotide overlap. These changes were intended to limit oligonucleotide costs and allow degenerate nucleotides to be placed in non-overlapping regions, limiting library composition biases produced by differential annealing effects. Expensive and time-consuming oligonucleotide purification was omitted; instead, a frameshift selection plasmid pInSALect was applied to correct for errors introduced during oligonucleotide synthesis and PCR assembly.[27] Over-sequencing (4x) of a 24-member library typically gave 85% correctly inserted, non-mutated sequences (see supplemental materials), out of which ~ 80% of each desired library could be recovered. Missing library members were generated by standard quick-change mutagensis.

The libraries were then expressed, purified, and denatured as described in the methods. Control experiments verifying the accuracy and precision of the microtiter plate-based stability assay showed excellent agreement with denaturation experiments monitored by circular dichroism (see supplemental materials). Future improvements in the throughput of stability determination can come from the usage of robotics platforms for variant construction, colony picking, and protein purification. Shifting the focus from sequencing towards stability screening could quickly produce information about the best mutants, as is common in directed evolution protocols. However, since a comprehensive screening of each designed library was desired, a lower level of throughput was tolerated.

Experimental screening of the xtal-1 library (Figure 3) showed two distinct sets of variants. The 12 library members with wild-type Leu at position 5 all exhibited stabilities similar to or better than the wild-type sequence, while the 12 with Ile at position 5 were all significantly destabilized. Screening of the NMR-based libraries (Figures 4 and 5) showed a similar dichotomy. In each case, the 6 library members with the wild-type Phe at position 52 exhibited wild-type-like stability or better. The remaining 18 variants from each NMR-based library were highly destabilized, and many lacked enough of a pretransition to be fit to the two-state unfolding model.

Evaluation of the MD libraries indicated that all 24 variants from the constrained library, cMD-128, had stability similar to the wild type or better (Figure 6). In contrast, all 24 variants from the uMD-128 library failed to produce any significant change in fluorescence signal across the denaturation series, and thus may be unfolded or structurally perturbed, as discussed below. A comparison of all five experimentally characterized libraries (Figure 7) indicates clearly that the cMD-128 design successfully produced a variety of stabilized mutants, whereas every other designed library specified at least one problematic substitution that rendered many of its sequences destabilized or otherwise unlike the wild type.

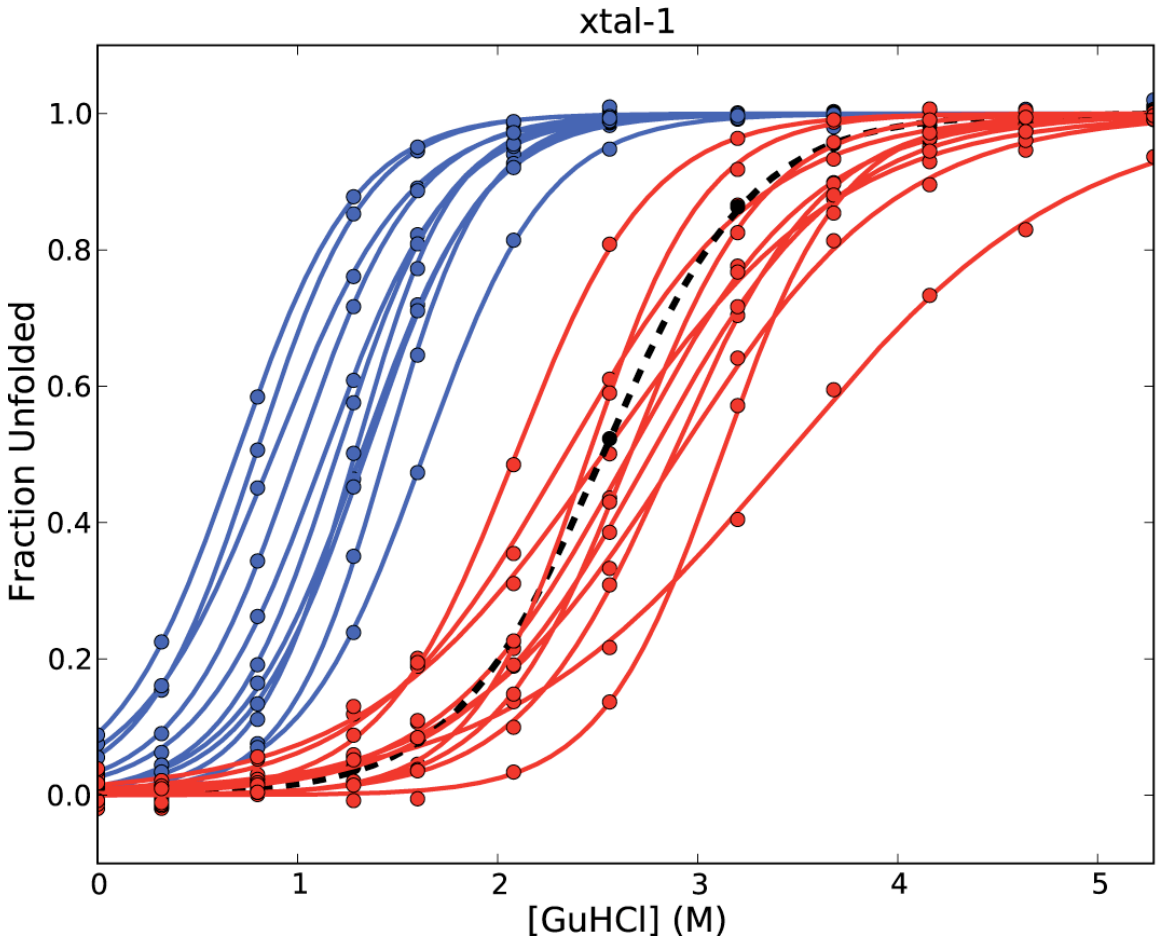

Figure 3: Fraction-unfolded curves derived from the stability determination of library xtal-1. The dashed black curve denotes variant Y3F, which is the closest library member to the wild type in terms of sequence, and which is known to have a stability very similar to the wild type. Red curves denote variants with C_m > 2.0 M, and correspond to all variants with Leu at position 5. Blue curves denote variants with C_m < 2.0 M, and correspond to variants with Ile at position 5. Not pictured: variant Y3F+L5I+L7I, which did not give a signal that could be fit to a two-state unfolding model.

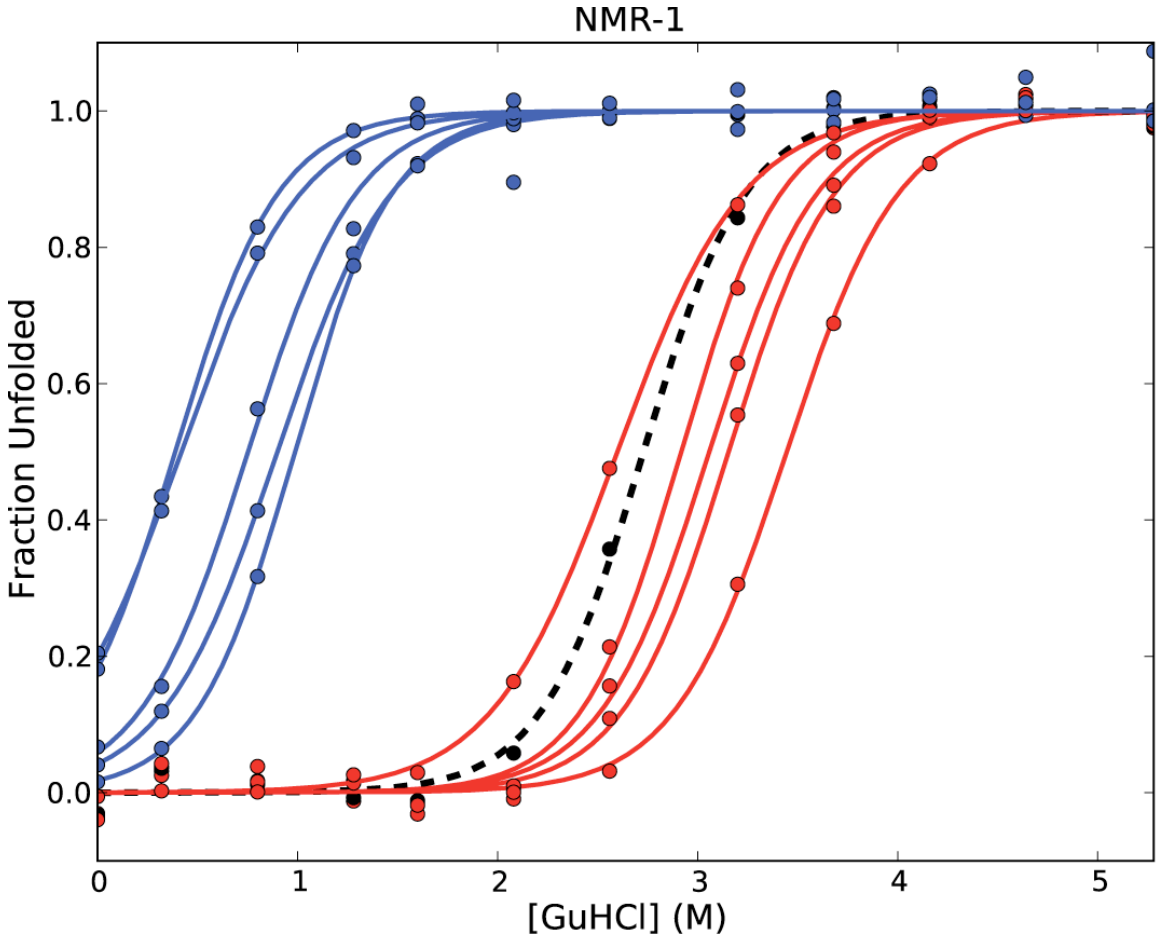

Figure 4: Fraction-unfolded curves derived from the stability determination of library NMR-1. The dashed black curve denotes variant Y3F, which is the closest library member to the wild type in terms of sequence, and which is known to have a stability very similar to the wild type. Red curves denote variants with C_m > 2.0 M, and correspond to all variants with Phe at position 52. Blue curves all represent variants with C_m < 2.0 M, which lack Phe at position 52, and have Val at position 39. Not pictured: 13 variants that lack Phe at position 52.

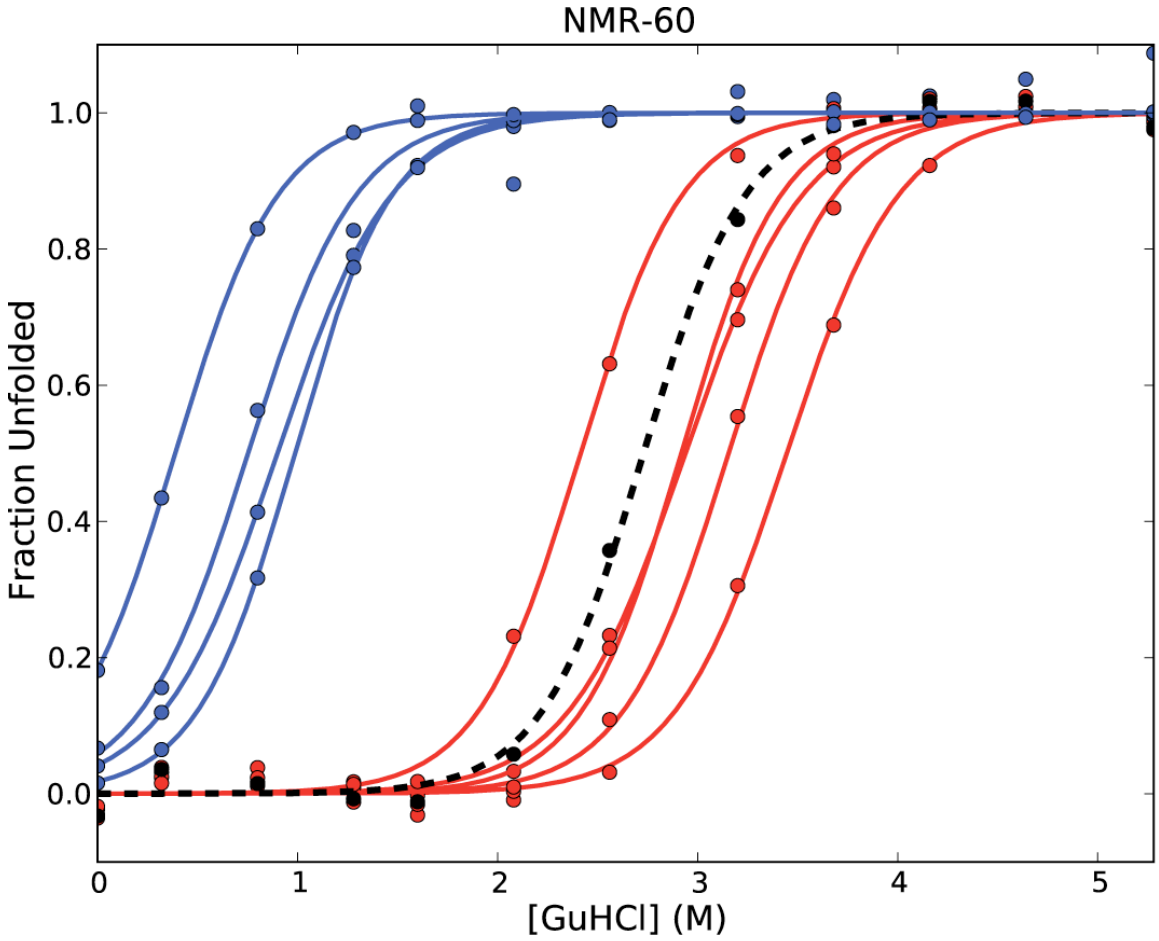

Figure 5: Fraction-unfolded curves derived from the stability determination of library NMR-60. The dashed black curve denotes variant Y3F, which is the closest library member to the wild type in terms of sequence, and which is known to have a stability very similar to the wild type. Red curves denote variants with $C_m > 2.0$ M, and correspond to all variants with Phe at position 52. Blue curves all represent variants with $C_m < 2.0$ M, which lack Phe at position 52, and have Val at position 39. Not pictured: 14 variants that lack Phe at position 52.

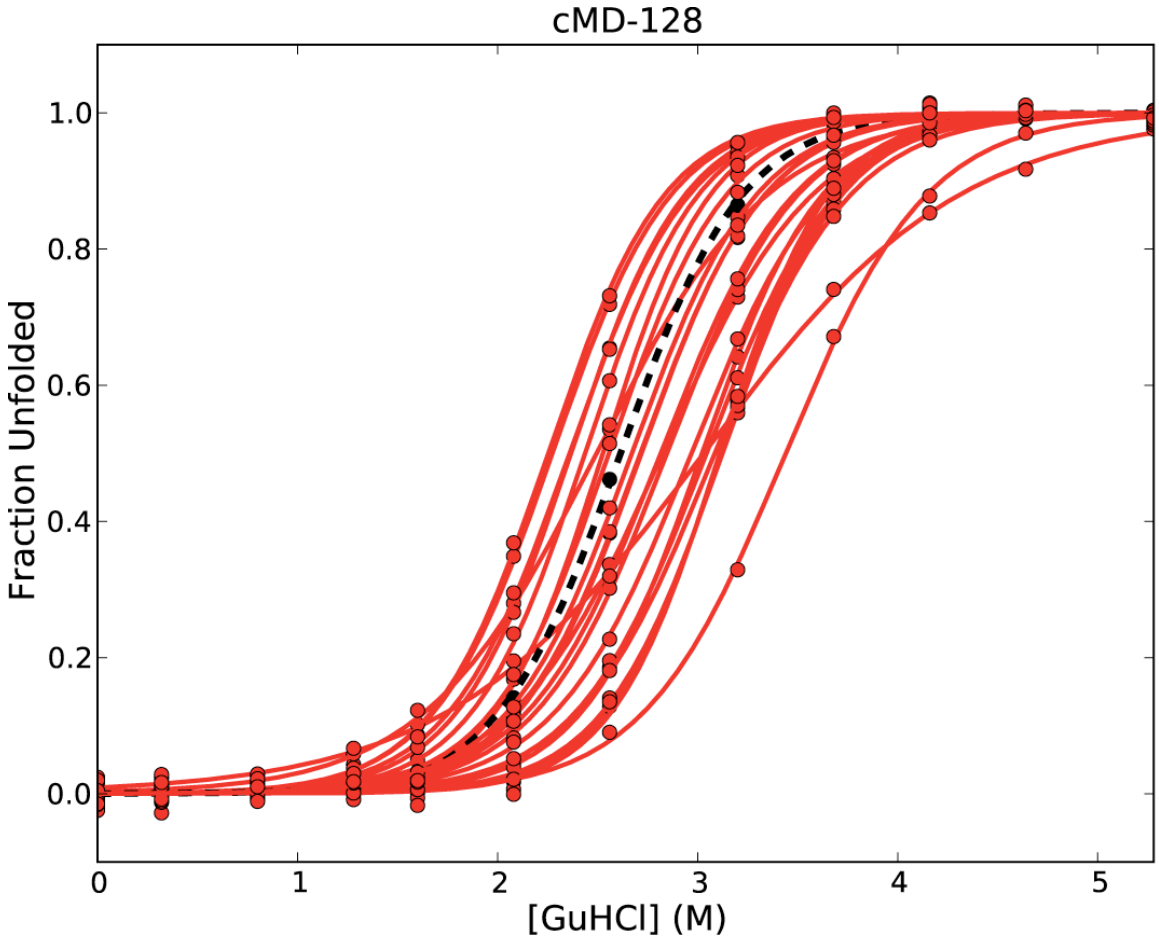

Figure 6: Fraction-unfolded curves derived from the stability determination of library cMD-128. The dashed black curve denotes variant Y3F, which is the closest library member to the wild type in terms of sequence, and which is known to have a stability very similar to the wild type. Red curves denote variants with C_m > 2.0 M, and correspond to all 24 variants in the library.

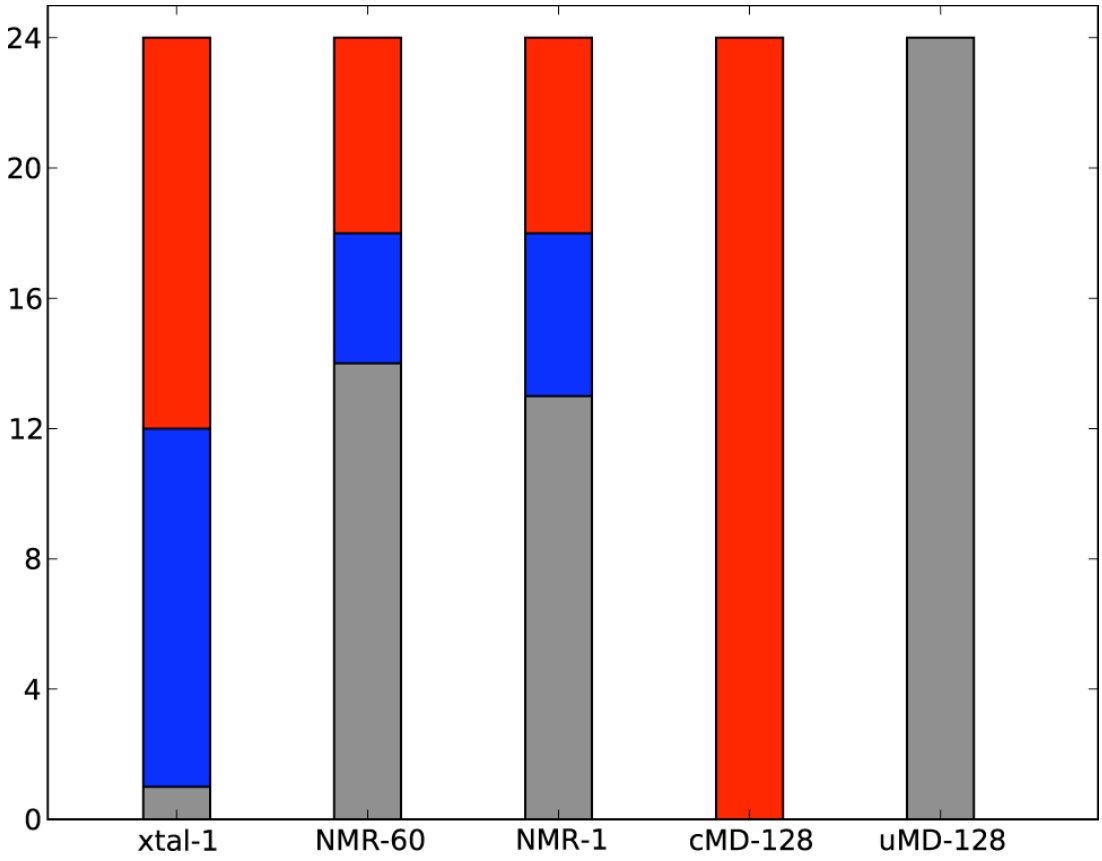

Figure 7: Each library partitioned into three stability groups. The colors match those in Figures 3–6: red (stable, $C_m > 2.0$), blue (destabilized, $C_m < 2.0$ M), grey (did not give a signal that could be fit to a 2-state model; not pictured in Figures 3–6).

Origin of destabilizing mutations

With experimental screening results in hand, we can return to the calculations that inspired them and ask why mutations such as L5I, F52ILV, and A34F were chosen by the design procedure. These mutations were all present in high-scoring sequences from the original design calculations, and thus are not artifacts introduced by the library design process.

The selection of the amino acids FILV at position Phe52 in the two NMR-based libraries resulted in three quarters of each library being significantly destabilized. In the context of the NMR structures, no Phe rotamer in the library was able to fit perfectly at position 52, encouraging the selection of smaller amino acids. If the set of rotamers at this position is supplemented with the observed rotamer in each structure, the design chooses to allocate diversity to positions 7 and 39, resulting in libraries similar to xtal-1. This result highlights how dramatically the rotameric approximation can influence the results of a design. It suggests that, at the very least, rotamers optimized for the wild-type sequence should be included when the goal is to find particular desirable sequences. In this case, we omitted the structurally observed rotamer at each position in order to limit the significant bias towards the wild-type sequence that these rotamers tend to cause. In the context of a real protein engineering project, this choice would have considerably reduced our chances of success.

The L5I mutation, which caused half of the xtal-1 library members to be destabilized relative to the wild-type sequence, may have been selected due to a failure of the softened repulsive contact potential that is used to counteract unrealistic rigidity introduced by the CPD model. The γ methyl group of Ile5 bumps into a Thr residue on

an adjacent β strand and is scored as a serious clash using unscaled van der Waals radii, but appears innocuous with the atomic radius scaling factor of $\alpha = 0.9$ that we used for the designs evaluated here. Repeating the design calculations with radii scaled by intermediate values such as 0.925 and 0.95 prevents Ile from being chosen at position 5, but also increases the frequency with which smaller residues are chosen at position Phe52. Interestingly, the recommendation of $\alpha = 0.9$ is derived from previous experiments based on the same set of Gβ1 core positions that were designed here. The earlier work drew conclusions based only on the best-scoring sequences produced by the design calculations, and found no difference between scaling atomic radii by 0.9 or 0.95.[19] Our results here indicate that the quality of sequences produced by the design procedure varies significantly with values of α between 0.9 and 0.95 when more sequences are taken into account. Given this, a more rigorous investigation of the most appropriate α value for design seems both tenable and warranted.

To analyze the uMD-128 data, it is important to note that our stability assay reports on the environment of the single Trp residue of Gβ1. Changes in packing caused by substitutions at other positions could alter the native-state environment of Trp43 enough to flip its side chain out into solution or change its fluorescence properties, crippling our ability to monitor unfolding by fluorescence. This interpretation seems unlikely for the destabilized members of the crystal structure and NMR libraries, for which a partial unfolding transition is clearly indicated by the raw data. However, the members of the uMD-128 library fail to show even a hint of such a transition, rendering the validity of our assay more suspect in this case.

Interestingly, others have investigated a 5-fold core variant of Gβ1 that bears substitutions similar to those in our uMD-128 library, including the A34F mutation. Structural characterization of this variant by NMR and X-ray crystallography indicated a domain-swapped tetrameric structure; the fluorescence emission maximum of this sequence was blue-shifted by almost 20 nm.[38] Related variants with the A34F substitution, including the A34F single mutant of the wild-type sequence, have also been shown to assume domain-swapped or side-by-side dimeric conformations in solution.[39, 40] Given these reports, the variants in our uMD-128 library, which all bear the A34F mutation, might also plausibly assume one of these oligomeric conformations. In this case, the library sequences could easily exhibit fluorescence emission spectra incompatible with our assay parameters, which were developed based on the characteristics of the wild-type sequence. Ultimately, the structural features of the uMD-128 library are unknown without additional experimental characterization. However, the published investigations of Gβ1 variants with the A34F substitution suggest that our uMD-128 library sequences are likely to assume conformations other than those modeled in our design calculations.

Influence of the designed library selection method

At this point, it is important to address the degree to which serendipity in designed library selection might affect the conclusions we may draw from our experiments. The CLEARSS library design procedure was developed with an understanding that many different combinatorial libraries may similarly represent a given list of scored sequences. Thus, its default mode of operation is to produce a list of the

top-scoring designed combinatorial libraries that satisfy all constraints, and to let the user choose between them. In general, this choice might be influenced by chemical intuition or prior mutational data, and thus partially account for properties of the system that are not modeled during the design procedure. To make our evaluation of input structural data sources as fair as possible, we chose to ignore such influences and apply an objective strategy based on the energies of the sequences in the libraries. Nevertheless, we must ask how other reasonable libraries generated by CLEARSS would have fared in our experimental assay.

Each of the top 20 designed libraries based on the NMR ensemble, as well as each based on the single average NMR structure, assigned smaller residues than the wild-type Phe to position 52. The remaining diversity of each library was occupied by various combinations of the other mutations present in the xtal-1, NMR-1, and NMR-60 libraries we screened in this work. It seems very likely, then, that the screening of any of the top NMR-based libraries from our designs would have resulted in stability data quite similar to that shown in Figures 4 and 5. Similarly, all of the top 20 designed libraries based on the unconstrained MD ensemble contained mutations L5A and A34F, and would be expected to exhibit similar fluorescence characteristics to the library uMD-128 we tested here.

A more interesting case is provided by the designs based on the crystal structure and constrained MD ensemble. Our analysis of the libraries xtal-1 and cMD-128 produced by these designs seems to indicate that cMD-128 was more successful, since a much greater fraction of its members were shown to be highly stable. However, when the top 20 libraries from each design were inspected in aggregate, it became clear that

both designs had produced a variety of libraries with various expected properties. The xtal-1 library and the cMD-128 library were each found in the top 20 libraries produced by both designs. Furthermore, each design produced several libraries with diversity at position 52, like NMR-1 and NMR-60. It seems clear that small changes to the constrained MD ensemble or to our energy functions might have reversed any potential conclusions about the usefulness of structural ensembles compared to single structures for the purposes of CPD.

The nature of approximation in computational protein design

In addition to helping validate the use of multi-state and combinatorial library design methods for computational protein design, our experimental results also allowed some unexpected insight into protein design itself. Plots of experimental stability versus simulation energy for the cMD-128 library (Figure 8) failed to yield any correlation, despite the apparent success of this design calculation. Likewise, the design calculations for xtal-1 and the NMR libraries failed to predict the pronounced destabilizing effects of mutations L5I or F52L, even though these designs also found a variety of stabilized variants. The design problem we chose is not simply too trivial for our purposes: the uMD-128 library and many previous reports attest to the myriad ways in which this system can be broken.[19, 38–42]

With a multiplicity of approximate methods available for computing the relative stabilities of protein sequences, the difficulty of solving this problem generally and accurately is sometimes overlooked. The stability of a sequence depends on the equilibrium between a relatively well-defined ensemble of native state conformations and

a vaguely defined ensemble of competing states. Our ability to find the relevant low-energy states is constrained by the vastness of protein conformational space and the extremely rugged energy landscape produced by our energy functions. Amino acid substitutions alter this energy landscape unpredictably, limiting the utility for design of structural information gathered for individual sequences. Current approaches tend to model native states at high resolution using whatever structures happen to be available, and account for competing states implicitly using statistical and heuristic terms.

Such methods have been surprisingly effective, given the approximations they rely upon. One perspective is that a CPD method is successful only to the extent that it can accurately predict or rank the stabilities of the variants it simulates, and that improvements in designed sequences will follow from improvements in ranking ability.[43] Accordingly, several groups have taken on large-scale forcefield parameterization efforts based on thermodynamic databases.[44, 45] In our research group, a forcefield tuned to offer significantly improved correlation between simulated and experimental stability differences did not exhibit improved performance for combinatorial design methods that allow large jumps in sequence space.[45] We can infer the same about the tuned forcefield of another group, given several reports of successful designs based on iterative one-by-one design and none based on combinatorial design methods.[46–50] The ability to reproduce experimental stability rankings is apparently not sufficient for accurate combinatorial protein design, at least in the range of ranking accuracy that has been achieved so far. The results of our work here furthermore suggest that this property is not even necessary for effective design.

This perspective prompts a modified view of the factors that make structure-based protein design possible in the first place. As discussed above, protein structures relax to accommodate mutations, and the computational difficulty of simulating these relaxations accurately has so far rendered intractable the stability ranking of sequence variants with many mutations. Fortunately, this malleability also means that sequences chosen to fit into a rigid protein model, even using approximate energy functions, will likely be tolerated by whatever relaxed structure results from the mutations they contain. In this way, the soft material properties of proteins impede the development of the quantitative protein design method we seek, but also make possible the more qualitative methods we can apply today.

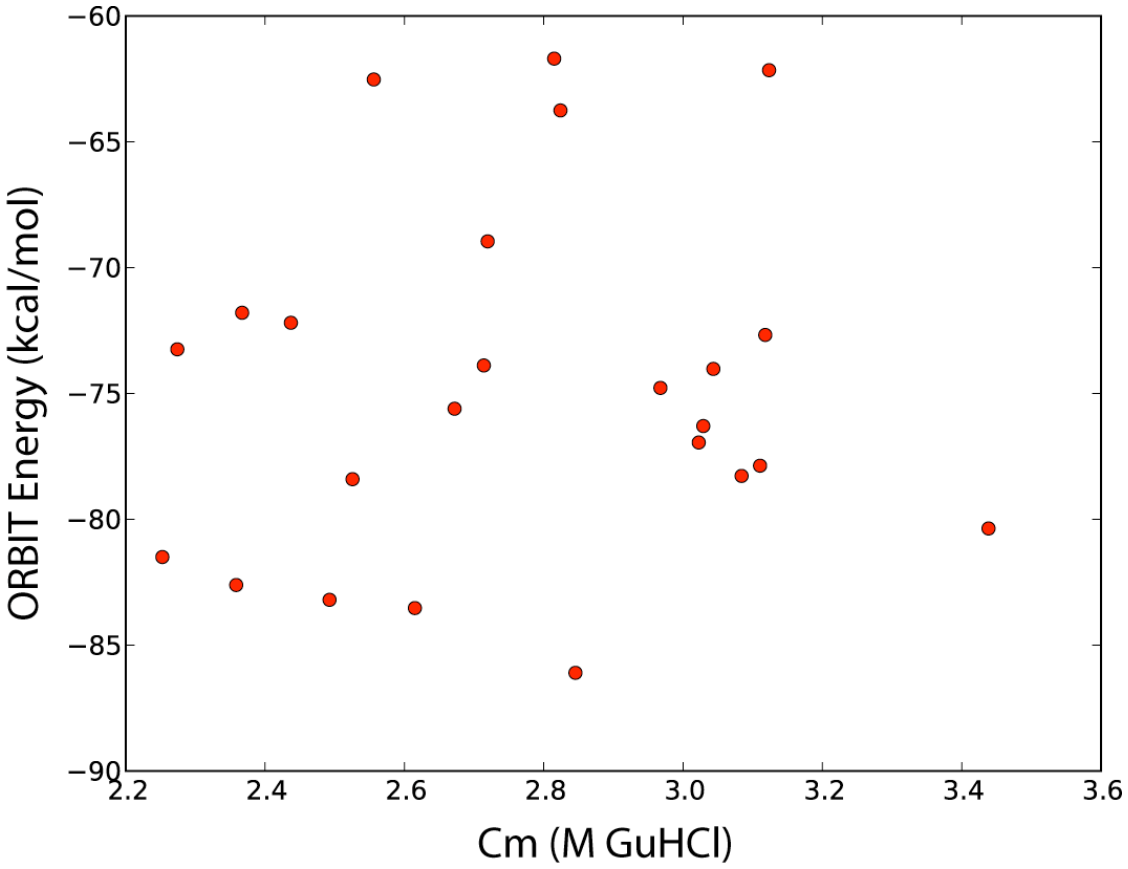

Figure 8: Correlation between simulation energy and experimental stability for the cMD-128 library. No correlation was observed between the experimentally measured fitness of the sequences and simulation energies that were used to select them for experimental screening.

Conclusions

Here, we have reported the development of new methods for the design and stability screening of combinatorial libraries based on lists of scored sequences. These methods were enlisted to test the application of multi-state design procedures to several structural ensembles, and to compare the resulting designs to those based on single structures. Designed libraries gave multiple stabilized variants when based on a crystal structure, an MD trajectory from that crystal structure, an NMR ensemble, and a single structure derived from the NMR ensemble. Our single-state and multi-state designs based on NMR data produced similar sets of libraries; likewise did those based on crystallographic data. Although an MD-based library gave superlative results, we cannot definitively conclude that the use of a structural ensemble provides any particular advantage over a single high-resolution structure for the purposes of design. Nevertheless, this initial success seems intriguing and warrants additional study. It seems clear that the energy functions and rotamer libraries developed for single-state modeling are equally applicable to the multi-state design of large structural ensembles. This result has important ramifications for future methods in CPD: even if structural ensembles fail to prove useful in the modeling of native states, they are expected to be crucial in the accurate modeling of competing states, which are undoubtedly more diverse.

In addition to validating the idea of design based on large structural ensembles, our work has provided further support in favor of rigorously screening an area of sequence space discovered by simulation, and has helped in vetting our new, general method for library design. For some designs that specified undesired destabilizing mutations, library screening suggested underlying causes for design failure that would not

have been apparent via the ad-hoc testing of individual sequences. Because our library design procedure is specifically intended to faithfully represent its input scored sequence list, and is indifferent to the origin of the list, it should be more useful for the evaluation of new design procedures than its predecessors.

Finally, the observed lack of correlation between experimental and simulated stabilities in our relatively successful sets of designed sequences may suggest a modified approach to protein design. Current design procedures seem to find stable sequences by selecting mutations that are likely to be accommodated by a relaxed version of the template structure, and not by accurately ranking the mutations relative to each other. In this view of design, finding sequences that satisfy the native state is relatively easy, while deciding which sequences satisfy it best is considerably more difficult. Given that stability is a function of nonnative states as much as native ones, the implication is that additional effort should be directed more toward eliminating sequences that can favorably assume competing states and less toward attempting to accurately predict which will best satisfy the native state. Since the relevant competing states under nondenaturing conditions likely exhibit significant residual structure, their treatment will probably require more sophisticated techniques than the composition-based heuristic terms used today. An interesting initial approach might be to perform multi-state design with an ensemble of native states as the positive design target and an ensemble of perturbed or expanded native states as the negative design target. The hypothesis is that selecting sequences to satisfy the compact native state and to not satisfy an expanded native state would tend to promote the desired specificity of a well-folded protein. Whether or not this type of strategy proves successful depends on the degree to which nonnative states

influence free energies of folding in a sequence-dependent (rather than composition-dependent) manner, and on the accuracy with which negative design can be performed against a computationally tractable set of competing states. Ultimately, techniques for native-state structural refinement will be crucial in the improvement of variant ranking; such methods may profitably be applied to produce appropriate nonnative ensembles as well. The next steps along the road to more accurate protein design thus include the development of methods for the construction and validation of useful nonnative ensembles, and the integration of structure refinement techniques with multi-state design methods. The validation provided here for our multi-state design, library design, and high-throughput stability screening methods represents a significant step towards the development of future methods that live up to the initial promise of computational protein design.

Materials and methods

Input structural data

Input atomic coordinates for the β1 domain of Streptococcal protein G (Gβ1) were taken from the 2.2 Å crystal structure 1pga,[51] the 60-member NMR structural ensemble 1gb1, and a constrained, minimized average structure generated from the ensemble 2gb1.[52] Hydrogens (if any) were stripped from each structure, and new hydrogen positions were optimized along with side-chain amide and imidazolium group flips using REDUCE.[53] Each structure was then standardized with 50 steps of conjugate gradient minimization using the DREIDING force field.[54] An unconstrained 128-member molecular dynamics (MD) ensemble was generated from the minimized crystal structure by running a 12.8 ps MD trajectory at 300 K using the DREIDING force field and saving the coordinates every 0.1 ps. The constrained MD trajectory was generated by the same procedure, using an additional harmonic point restraint with a force constant of 100 kcal/mol/Å^2 applied to keep C_α atoms near their initial positions. Each MD snapshot was standardized as described above. After standardization, the NMR, constrained MD, and unconstrained MD ensembles exhibited average pairwise main-chain RMSDs of 0.25, 0.12, and 0.84 Å, respectively.

Sequence Design Specifications and Energy Calculations

In the sequence designs, ten core positions of Gβ1 (3, 5, 7, 20, 26, 30, 34, 39, 52, and 54), were allowed to assume any of the hydrophobic amino acids A, V, L, I, F, Y,

and W. Tryptophan 43 was allowed to change conformation but not amino acid type, so that our fluorescence-based stability assay would not be compromised. Allowed side-chain conformations at the variable positions were taken from the Dunbrack backbone-dependent rotamer library with expansions of ±1 standard deviation around $\chi 1$ and $\chi 2$.[17] To avoid bias toward the wild-type sequence, this set was not supplemented with the side-chain coordinates from the input structure, except at position 43. All other side chains and the main chain were fixed in the input conformation. Pairwise energies were computed for each structure or ensemble member using energy functions described previously,[55, 56] with the polar hydrogen burial term omitted.

Sequence optimization

FASTER was used to find optimized sequences in the single-state design of the crystal structure and the NMR constrained minimized average.[57] Multi-state sequence optimization of the NMR, unconstrained MD, and constrained MD ensembles was performed using a method similar to several that have been described.[23, 25] These methods implement a combinatorial search through amino acid sequence space in which sequences are scored by performing rotamer optimization in the context of each state and these energies are combined to yield a single ensemble score. Our implementation uses FASTER for both the search through amino acid sequence space and for the rotamer optimization on each state (Chapter 3). Here, the energies of a sequence in the context of several states were combined into a single score by computing the free energy of the ensemble system at 300 K:

$$A = -kT\log(\sum_j e^{-E_j/kT})$$

where each E_j is the energy of the sequence when threaded on member j of the ensemble.

Combinatorial library design

To choose combinatorial sequence libraries for experimental screening, we used a new algorithm reported here (see supplementary material). Given a list of scored sequences, a list of allowed sets of amino acids, and a range of desired library sizes, the method evaluates all possible combinations of sets of amino acids at different positions that lead to a library with a size in the desired range. Each position in each library is scored by summing the Boltzmann weights of the sequences in the list that contain a library-specified amino acid at that position. The position scores are then summed to give an overall library score. Our algorithm is able to consider all possible libraries because it treats positions independently, and because it ignores amino acid sets that are unnecessarily large in the context of a given position. In this work, a temperature of 300 K was used in the Boltzmann weighting, and the target library size was 24. We allowed only those sets of amino acids that can be specified by degenerate codons that do not include codons observed with low frequency in *E. coli*.

After applying this algorithm to the lists of sequences produced by the computational designs, we instantiated the 20 best-scoring libraries from each design and rescored all of the amino acid sequences in each library by rotamer optimization. Each library we inspected contained the best-scoring sequence from the design it was based on, although this is not required by our method. From each design, we chose for

experimental testing the library in the top 20 with the smallest energy spread between its best-scoring and worst-scoring sequence.

Library construction, expression, and purification

Oligonucleotides (desalted, Integrated DNA Technologies) ranging from 45 to 60 bp containing ~ 18 bp overlapping segments were assembled via a modified Stemmer method[37] using KOD Hot Start Polymerase (Novagen) to generate full-length streptococcal Gβ1 with an N terminal His_6 tag. Secondary structure content and annealing temperatures were verified by NUPACK.[58, 59] The following procedure was repeated for each library constructed. Oligonucleotides containing the desired single mutation or degenerate codon were swapped into the assembly mixture to generate the diversity of each library. If a degenerate codon could not account for the desired residue diversity, equimolar ratios of applicable single mutation oligonucleotides were added to the assembly mixture. Standard subcloning techniques were performed to insert the library into a frameshift selection plasmid (pInSALect),[27] and after miniprepping the selected harvested colonies, the library was inserted into an expression plasmid (pET11a). The library was transformed into BL21 Gold DE3 cells (Stratagene) by heat shock and colonies were picked into 96-well plates for plasmid miniprepping and sequencing (Agencourt Biosciences). Missing library members were generated by standard quick-change protocols. Sequence-verified library members were pulled from replicated glycerol stocks and inoculated into 5 mL of Instant TB media (Novagen) in 24-well plates. After overnight incubation at 37ºC, cells were pelleted by centrifugation at 5,000 x *g* for 20 min. Pellets were freeze/thawed once and resuspended in lysis buffer

(50 mM $NaPO_4$, 300 mM NaCl, 1x CelLytic B (Sigma-Aldrich), 2.5 mM imidazole, pH 8) before another identical centrifugation step. Cell lysates were loaded onto an equilibrated HIS-Select filter plate (Sigma-Aldrich), washed twice and eluted with buffer containing 250 mM imidazole, pH 8.

Microtiter plate-based stability determination

Appropriate amounts of 8 M GdmCl (Sigma-Aldrich), Milli-Q water, eluted protein, and 50 mM $NaPO_4$ buffer, pH 6.5, were added to maintain a fixed volume in each well of 96-well Costar UV transparent flat bottom plates by a Freedom EVO liquid handling robot (Tecan). Mutant proteins were subjected to a 12-point GdmCl gradient across the columns of the plate where each row contained a separate denaturation experiment. Only twenty-seven 96-well plates were needed for all libraries, including duplicates. The plates were equilibrated for at least one hour and shaken at 900 rpm on a microtiter plate shaker (Heidolph).

Tryptophan fluorescence measurements were taken on a fluorescence plate reader (Tecan) with a plate stacker attachment. Ideal parameters were empirically determined for wild-type Gβ1 and later used for every library assayed. Excitation was performed at 295 nm and emission measured at 341 nm with 10 nm bandwidths. Data were fit as a two-state unfolding transition using the linear extrapolation method[60] in Pylab. The GdmCl concentration at the midpoint of denaturation, C_m, was estimated numerically based on the fraction-unfolded curve fit.

Supplementary information

Combinatorial library design

Structure-based computational protein design (CPD) methods can be harnessed to expedite the engineering of proteins by directed evolution. Several methods have been developed to allow the design of combinatorial mutation libraries to be informed by the results of CPD calculations (Figure 2). These approaches allow many specific variants chosen by CPD to be tested experimentally, and can facilitate assessment and improvement of the design procedure. Hayes et al. described a method in which a list of low-energy sequences found by CPD is used to generate a table of frequencies for each amino acid type at each position, and then a frequency cutoff is applied to limit the library to only those amino acids found more frequently than the cutoff value at each position.[33] Mena and Daugherty developed a similar procedure that produces libraries that include as many of the sequences in the CPD list as possible, while using only those sets of amino acids that can be encoded using degenerate codons.[35] This feature helps to ensure that the resulting combinatorial gene libraries can be synthesized quickly and inexpensively. Treynor et al. developed a computational library design method analogous to CPD in which interactions between sets of amino acids at various positions are scored, and this system of interactions is sampled using standard CPD optimization algorithms to find the most favorable degenerate codon sequence.[36]

In our view, a procedure that couples CPD to the design of combinatorial protein libraries should provide at least the following:

1. **Explicit consideration of CPD energies**. Methods that ignore CPD energies lead to a weaker correspondence between the final libraries and the original design calculations, limiting the predictive capability of the library design procedure and making improvement of CPD through library screening and analysis more difficult.
2. **Direct specification of the range of library sizes that should be produced**. In general, the desired library size will be a direct function of experimental screening capacity. A method that does not allow the user to specify the library size will either require repeated manual rerunning in an attempt to generate the desired library size, or will waste potentially prohibitive amounts of compute time analyzing libraries with irrelevant sizes.
3. **Control over which sets of amino acids are allowed**. Users with limited resources will usually prefer sets of amino acids that can be encoded using degenerate codons, because the resulting gene libraries can be synthesized in a single reaction with a relatively small number of inexpensive oligonucleotides. Those who can afford larger numbers of oligonucleotides and liquid-handling robots will be able to test libraries made with arbitrary sets of amino acids, which in general should more accurately reflect the sequence preferences of CPD calculations. A robust library design method must therefore handle whatever sets of amino acids the user deems appropriate.
4. **Consideration of all user-allowed sets of amino acids at each position**. Some methods use heuristics to remove from consideration particular sets of amino acids at each position. Although this process can reduce the computational cost of

the library design procedure, it can also result in the elimination of desirable libraries.

Because no previously reported algorithm that we know of satisfies all these criteria, we developed one that does. The new algorithm takes several inputs: (1) a list of scored sequences; (2) a list of allowed sets of amino acids (e.g., those that can be encoded using degenerate codons); (3) a range of preferred library sizes; (4) a simulation temperature that controls the degree of preference for sequences with better scores; and, optionally, (5) sets of amino acids that are to be required or prohibited at particular positions. Based on these inputs, the algorithm produces a list of combinatorial libraries that are ranked according to the degree to which they satisfy the input list of scored sequences.

The process used by the algorithm to produce a list of combinatorial libraries from a list of scored sequences can be conceptually separated into three steps (Figure 9).

Step **A**. Scan through the input list of scored sequences, and generate a "total diversity" library that includes, at each position, every amino acid seen in the list at that position. This library represents the list optimally but ignores the user's preferred library size and allowed sets of amino acids. If later steps indicate that the size of the problem with this total diversity is insurmountably large, the user can request that the total diversity library be constructed from a subset of the input sequence list. For example, given a list of length 10,000, the user might decide to consider only the best 1,000 sequences in the list during this step.

Step **B**. Enumerate all possible amino acid size configurations that lead to combinatorial libraries within the range of sizes specified by the user. A size configuration is simply a specific number of amino acids at each position in the protein (e.g., 3 amino acids at position 1, 4 amino acids at position 2, etc.). An amino acid set size need not be considered at a particular position if it is larger than the smallest set that includes all amino acids found at that position in the total diversity library. This greatly reduces the total number of size configurations that need to be generated in this step and scored in the next step.

Step **C**. For each size configuration, determine the best set of amino acids of the required size at each position. This is done for each position independently by computing a partition function for each amino acid set with the given size. Amino acid sets that lack user-required amino acids or contain user-prohibited amino acids can be skipped here. Given a position and an allowed set of amino acids, iterate through the list of scored sequences, and for each sequence add to a cumulative partition function the Boltzmann-weight, $\exp(-E/kT)$, where E is the score of the sequence, k is the Boltzmann constant, and T is the simulation temperature. If the amino acid at that position in the current sequence is not found in the amino acid set of interest, nothing is added to the partition function. If the simulation temperature is low, the best-scored sequences will contribute most strongly to the partition function; if the temperature is high, all sequences in the list will contribute similarly. At each position, the set of amino acids with the most favorable partition function (position library score) is chosen. This procedure produces an optimal combinatorial library for each size configuration. The optimal libraries of each possible

size configuration can then be ranked based on the sums of their position library scores across all positions.

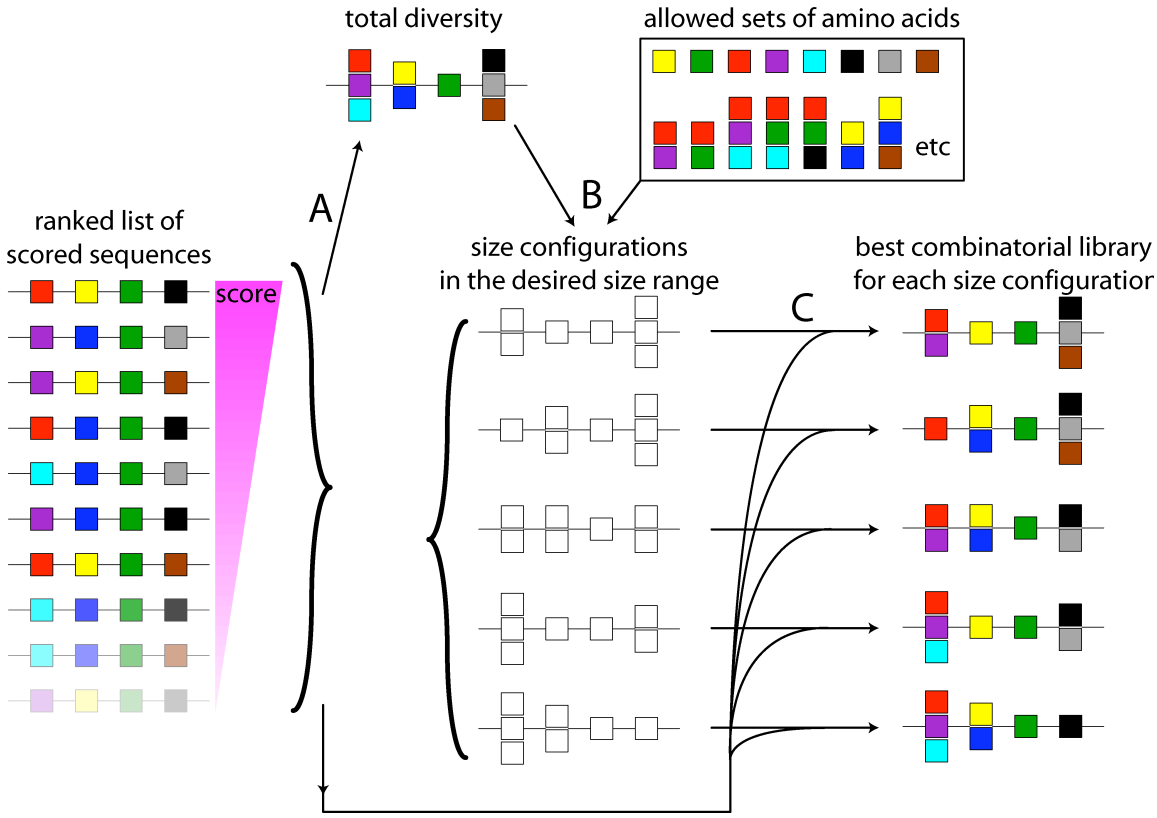

Figure 9: Detail of the library design method. (A) The list of scored sequences defines an initial "total diversity" library that is typically much larger (10^3 – 10^{15}, or even more) than the desired library size (10^2–10^6). (B) This total diversity library and the allowed sets of amino acids are used to construct a set of size configurations that lead to libraries in the desired range of sizes. The boxes in the list of size configurations are unfilled, indicating that the particular amino acids at each position have not yet been determined at this step. (C) For each size configuration generated in the previous step, the original list of scored sequences is used to find the optimal set of amino acids of the required size at each position.

Microtiter plate-based stability assay controls

The fluorescence profiles of the GdmCl gradient and the elution buffer show no effect on the shape of the unfolding transition of wild-type Gβ1 (Figure 10). Sample signal below the elution buffer was interpreted as expression failure; any data that could not be fit yet whose signal was above the elution buffer was deemed expressed but unstable/unfolded (but see discussion above). In order to test the accuracy of the microtiter plate-based denaturation assay, Gβ1 unfolding was monitored by circular dichroism (Aviv Biomedical) and tryptophan fluorescence in a fluorimeter (Photon Technology International). The denaturation profiles from these low-throughput experiments were compared to results from the fluorescence plate reader (Figure 11). The overlapping data points support the use of a two-state unfolding fit during our stability calculations and verify the accuracy of the assay. Next, the unfolding curves from several protein preparations from different concentrations confirmed the assay's precision (Figure 12). These results support some assumptions that the stability determination method described here makes in order to maintain a high level of throughput. First, we never assay for protein concentration before setting up the GdmCl gradient, relying on the fraction-unfolded plot to remove any concentration bias/effects. Second, the high concentration (250 mM) of imidazole in elution buffer is never dialyzed out of the eluted protein solution. Figures 11 and 12 show that these discrepancies in protein preparation have no significant effect on fraction unfolded plots for the wild-type protein.

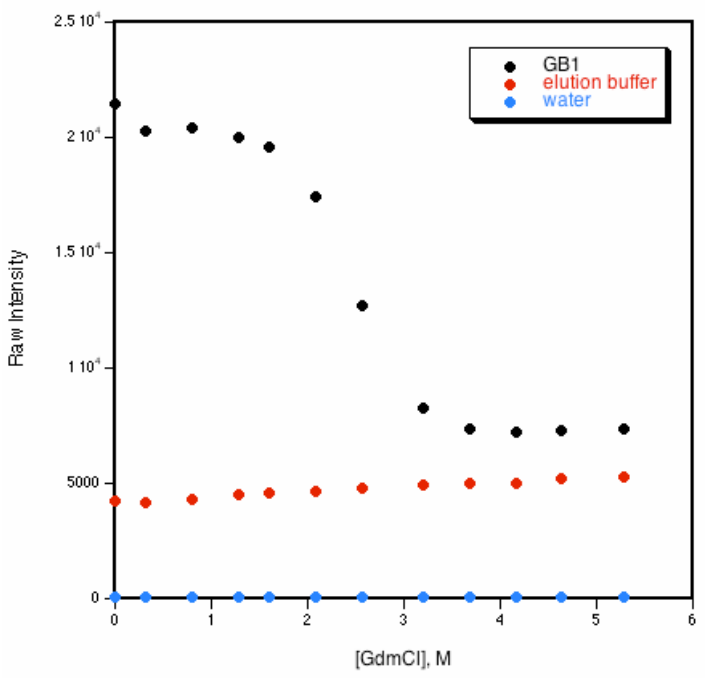

Figure 10: Denaturation gradient and elution buffer fluorescence profiles. Gβ1 (black) was expressed in a 5 mL culture, purified, and eluted with 500 μL of elution buffer (50 μM $NaPO_4$, 300 mM NaCl, 250 mM imidazole, pH 8). Since each point of the Gβ1 denaturation profile contains 35 μL of eluted protein, the elution buffer profile (red) substitutes protein with 35 μL of elution buffer. Similarly, the water profile (blue) adds 35 μL of water to make up the final volume. Each denaturation profile contains an increasing gradient of GdmCl, 50 μM $NaPO_4$ buffer at pH 6.5, and water.

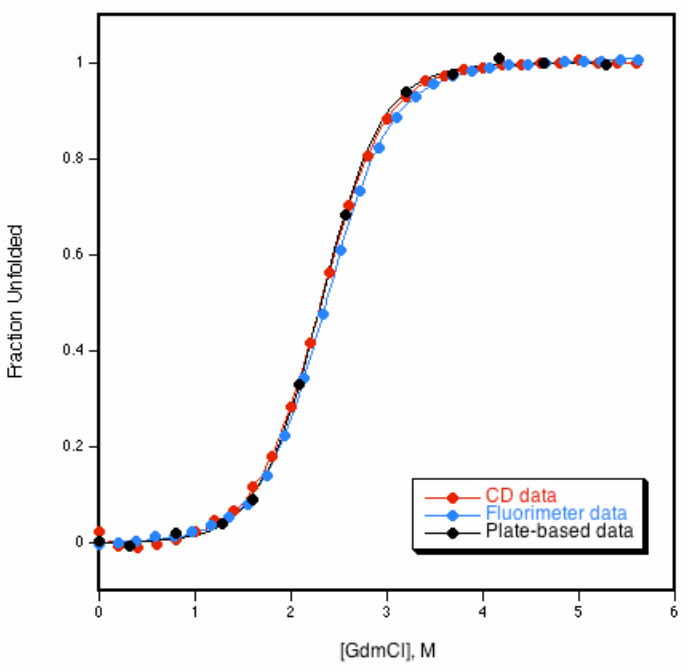

Figure 11: Fraction-unfolded profiles between different modes of detection. CD data (red) measured 5 μM Gβ1 titrated with a 5 μM Gβ1/8 M GdmCl solution in 0.2 M steps at 218 nm. Fluorimeter data (blue) measured 5μM Gβ1 titrated as in the CD experiment with excitation performed at 295 nm and emission recorded at 341 nm with 4 nm bandwidths. Plate-based data (black) measured 12 separate solutions of 10 μM Gβ1 in response to increasing amounts of 8 M GdmCl with fluorescence parameters identical to the fluorimeter data except for 10 nm bandwidths. All samples were measured at 25°C in 50 μM $NaPO_4$ buffer at pH 6.5.

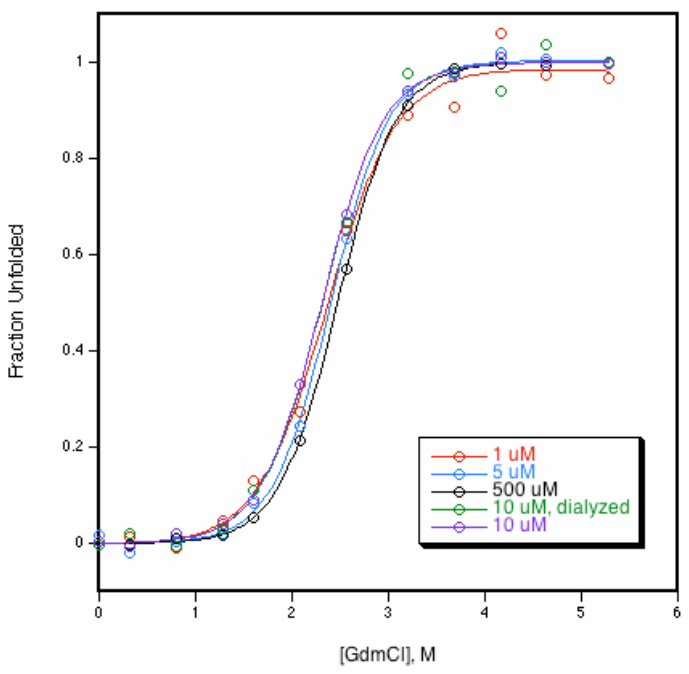

Figure 12: Fraction-unfolded profiles between different protein preparations. Gβ1 was expressed in 100 mL cultures, purified and diluted to 1, 5, 10, and 500 μM in 50 μM $NaPO_4$ buffer at pH 6.5. Another expression culture was dialyzed overnight (Pierce Biotechnology) after purification and diluted to 10 μM in the same buffer. All measurements were taken on a fluorescence plate reader as described in the text.

References

1. Arnold, F. H., Combinatorial and computational challenges for biocatalyst design. *Nature* **2001,** *409* (6817), 253–257.

2. Jackel, C.; Kast, P.; Hilvert, D., Protein design by directed evolution. *Annual Reviews of Biophysics* **2008,** *37*, 153–173.

3. Bershtein, S.; Tawfik, D. S., Advances in laboratory evolution of enzymes. *Current Opinion in Chemical Biology* **2008,** *12* (2), 151–158.

4. Alvizo, O.; Allen, B. D.; Mayo, S. L., Computational protein design promises to revolutionize protein engineering. *Biotechniques* **2007,** *42* (1), 31–35.

5. Lippow, S. M.; Tidor, B., Progress in computational protein design. *Current Opinion in Biotechnology* **2007,** *18* (4), 305–311.

6. Schueler-Furman, O.; Wang, C.; Bradley, P.; Misura, K.; Baker, D., Progress in modeling of protein structures and interactions. *Science* **2005,** *310* (5748), 638–642.

7. Bolon, D. N.; Mayo, S. L., Enzyme-like proteins by computational design. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** *98* (25), 14274–14279.

8. Dahiyat, B. I.; Mayo, S. L., De novo protein design: Fully automated sequence selection. *Science* **1997,** *278* (5335), 82–87.

9. Dwyer, M. A.; Looger, L. L.; Hellinga, H. W., Computational design of a Zn2+ receptor that controls bacterial gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **2003,** *100* (20), 11255–11260.

10. Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Rothlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D., De novo computational design of retro-aldol enzymes. *Science* **2008,** *319* (5868), 1387–1391.

11. Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D., Design of a novel globular protein fold with atomic-level accuracy. *Science* **2003,** *302* (5649), 1364–1368.

12. Looger, L. L.; Dwyer, M. A.; Smith, J. J.; Hellinga, H. W., Computational design of receptor and sensor proteins with novel functions. *Nature* **2003,** *423* (6936), 185–190.

13. Malakauskas, S. M.; Mayo, S. L., Design, structure and stability of a hyperthermophilic protein variant. *Nature Structural Biology* **1998,** *5* (6), 470–475.

14. Rothlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K.

N.; Tawfik, D. S.; Baker, D., Kemp elimination catalysts by computational enzyme design. *Nature* **2008,** *453* (7192), 190–U4.

15. Chica, R. A.; Doucet, N.; Pelletier, J. N., Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Current Opinion in Biotechnology* **2005,** *16* (4), 378–384.

16. Shortle, D., The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB Journal* **1996,** *10* (1), 27–34.

17. Dunbrack, R. L.; Cohen, F. E., Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science* **1997,** *6* (8), 1661–1681.

18. Lassila, J. K.; Privett, H. K.; Allen, B. D.; Mayo, S. L., Combinatorial methods for small-molecule placement in computational enzyme design. *Proceedings of the National Academy of Sciences of the United States of America* **2006,** *103* (45), 16710–16715.

19. Dahiyat, B. I.; Mayo, S. L., Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences of the United States of America* **1997,** *94* (19), 10172–10177.

20. Grigoryan, G.; Ochoa, A.; Keating, A. E., Computing van der Waals energies in the context of the rotamer approximation. *Proteins* **2007,** *68* (4), 863–878.

21. Hu, X.; Wang, H.; Ke, H.; Kuhlman, B., High-resolution design of a protein loop. *Proceedings of the National Academy of Sciences of the United States of America* **2007,** *104* (45), 17668–17673.

22. Pokala, N.; Handel, T. M., Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *Journal of Molecular Biology* **2005,** *347* (1), 203–227.

23. Ambroggio, X. I.; Kuhlman, B., Computational design of a single amino acid sequence that can switch between two distinct protein folds. *Journal of the American Chemical Society* **2006,** *128* (4), 1154–1161.

24. Boas, F. E.; Harbury, P. B., Design of protein-ligand binding based on the molecular-mechanics energy model. *Journal of Molecular Biology* **2008,** *380* (2), 415–424.

25. Havranek, J. J.; Harbury, P. B., Automated design of specificity in molecular recognition. *Nature Structural Biology* **2003,** *10* (1), 45–52.

26. Dyson, H. J.; Wright, P. E., Intrinsically unstructured proteins and their functions. *Nature Reviews of Molecular and Cell Biology* **2005,** *6* (3), 197–208.

27. Gerth, M. L.; Patrick, W. M.; Lutz, S., A second-generation system for unbiased reading frame selection. *Protein Engineering Design & Selection* **2004,** *17* (7), 595–602.

28. Cox, J. C.; Lape, J.; Sayed, M. A.; Hellinga, H. W., Protein fabrication automation. *Protein Science* **2007,** *16* (3), 379–390.

29. Aucamp, J. P.; Cosme, A. M.; Lye, G. J.; Dalby, P. A., High-throughput measurement of protein stability in microtiter plates. *Biotechnology and Bioengineering* **2005,** *89* (5), 599–607.

30. Voigt, C. A.; Mayo, S. L.; Arnold, F. H.; Wang, Z. G., Computational method to reduce the search space for directed protein evolution. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** *98* (7), 3778–3783.

31. Endelman, J. B.; Silberg, J. J.; Wang, Z. G.; Arnold, F. H., Site-directed protein recombination as a shortest-path problem. *Protein Engineering Design & Selection* **2004,** *17* (7), 589–594.

32. Voigt, C. A.; Martinez, C.; Wang, Z. G.; Mayo, S. L.; Arnold, F. H., Protein building blocks preserved by recombination. *Nature Structural Biology* **2002,** *9* (7), 553–558.

33. Hayes, R. J.; Bentzien, J.; Ary, M. L.; Hwang, M. Y.; Jacinto, J. M.; Vielmetter, J.; Kundu, A.; Dahiyat, B. I., Combining computational and experimental screening for rapid optimization of protein properties. *Proceedings of the National Academy of Sciences of the United States of America* **2002,** *99* (25), 15926–15931.

34. Kono, H.; Saven, J. G., Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *Journal of Molecular Biology* **2001,** *306* (3), 607–628.

35. Mena, M. A.; Daugherty, P. S., Automated design of degenerate codon libraries. *Protein Engineering Design & Selection* **2005,** *18* (12), 559–561.

36. Treynor, T. P.; Vizcarra, C. L.; Nedelcu, D.; Mayo, S. L., Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proceedings of the National Academy of Sciences of the United States of America* **2007,** *104* (1), 48–53.

37. Stemmer, W. P.; Crameri, A.; Ha, K. D.; Brennan, T. M.; Heyneker, H. L., Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **1995,** *164* (1), 49–53.

38. Kirsten Frank, M.; Dyda, F.; Dobrodumov, A.; Gronenborn, A. M., Core mutations switch monomeric protein GB1 into an intertwined tetramer. *Nature Structural Biology* **2002,** *9* (11), 877–885.

39. Byeon, I. J.; Louis, J. M.; Gronenborn, A. M., A protein contortionist: core mutations of GB1 that induce dimerization and domain swapping. *Journal of Molecular Biology* **2003,** *333* (1), 141–152.

40. Jee, J.; Byeon, I. J.; Louis, J. M.; Gronenborn, A. M., The point mutation A34F causes dimerization of GB1. *Proteins* **2008,** *71* (3), 1420–1431.

41. Gronenborn, A. M.; Frank, M. K.; Clore, G. M., Core mutants of the immunoglobulin binding domain of streptococcal protein G: stability and structural integrity. *FEBS Letters* **1996,** *398* (2–3), 312–316.

42. Su, A.; Mayo, S. L., Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Science* **1997,** *6* (8), 1701–1707.

43. Mendes, J.; Guerois, R.; Serrano, L., Energy estimation in protein design. *Current Opinion in Structural Biology* **2002,** *12* (4), 441–446.

44. Guerois, R.; Nielsen, J. E.; Serrano, L., Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of Molecular Biology* **2002,** *320* (2), 369–387.

45. Zollars, E. S. Force Field Development In Protein Design. Caltech, Pasadena, CA, 2006.

46. Fajardo-Sanchez, E.; Stricher, F.; Paques, F.; Isalan, M.; Serrano, L., Computer design of obligate heterodimer meganucleases allows efficient cutting of custom DNA sequences. *Nucleic Acids Research* **2008,** *36* (7), 2163–2173.

47. Tur, V.; van der Sloot, A. M.; Reis, C. R.; Szegezdi, E.; Cool, R. H.; Samali, A.; Serrano, L.; Quax, W. J., DR4-selective tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) variants obtained by structure-based design. *Journal of Biological Chemistry* **2008,** *283* (29), 20560–20568.

48. van der Sloot, A. M.; Mullally, M. M.; Fernandez–Ballester, G.; Serrano, L.; Quax, W. J., Stabilization of TRAIL, an all-beta-sheet multimeric protein, using computational redesign. *Protein Engineering Design & Selection* **2004,** *17* (9), 673–680.

49. van der Sloot, A. M.; Tur, V.; Szegezdi, E.; Mullally, M. M.; Cool, R. H.; Samali, A.; Serrano, L.; Quax, W. J., Designed tumor necrosis factor-related apoptosis-inducing ligand variants initiating apoptosis exclusively via the DR5 receptor. *Proceedings of the National Academy of Sciences of the United States of America* **2006,** *103* (23), 8634–8639.

50. Szczepek, M.; Brondani, V.; Buchel, J.; Serrano, L.; Segal, D. J.; Cathomen, T., Structure-based redesign of the dimerization interface reduces the toxicity of zinc-finger nucleases. *Nature Biotechnology* **2007,** *25* (7), 786–793.

51. Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L., 2 Crystal-Structures of the B1 Immunoglobulin-Binding Domain of Streptococcal Protein-G and Comparison with Nmr. *Biochemistry* **1994,** *33* (15), 4721–4729.

52. Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M., A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* **1991,** *253* (5020), 657–661.

53. Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C., Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology* **1999,** *285* (4), 1735–1747.

54. Mayo, S. L.; Olafson, B. D.; Goddard, W. A., Dreiding — a Generic Force-Field for Molecular Simulations. *Journal of Physical Chemistry* **1990,** *94* (26), 8897–8909.

55. Gordon, D. B.; Hom, G. K.; Mayo, S. L.; Pierce, N. A., Exact rotamer optimization for protein design. *Journal of Computational Chemistry* **2003,** *24* (2), 232–243.

56. Gordon, D. B.; Marshall, S. A.; Mayo, S. L., Energy functions for protein design. *Current Opinion in Structural Biology* **1999,** *9* (4), 509–513.

57. Allen, B. D.; Mayo, S. L., Dramatic performance enhancements for the FASTER optimization algorithm. *Journal of Computational Chemistry* **2006,** *27* (10), 1071–1075.

58. Dirks, R. M.; Pierce, N. A., A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry* **2003,** *24* (13), 1664–1677.

59. Dirks, R. M.; Pierce, N. A., An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *Journal of Computational Chemistry* **2004,** *25* (10), 1295–1304.

60. Santoro, M. M.; Bolen, D. W., Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. *Biochemistry* **1988,** *27* (21), 8063–8068.