DOI: 10.1002/cbic.201300326

Highly Diverse Protein Library Based on the Ubiquitous $(\beta/\alpha)_8$ Enzyme Fold Yields Well-Structured Proteins through in Vitro Folding Selection

Misha V. Golynskiy, John C. Haugner, III, and Burckhard Seelig*[a]

Proper protein folding is a prerequisite for protein stability and enzymatic activity. Although directed evolution can be a powerful tool to investigate enzymatic function and to isolate novel activities, well-designed libraries of folded proteins are essential. In vitro selection methods are particularly capable of searching for enzymatic activities in libraries of trillions of protein variants, yet high-quality libraries of well-folded enzymes with such high diversity are lacking. We describe the construction and detailed characterization of a folding-enriched protein library based on the ubiquitous $(\beta/\alpha)_8$ barrel fold, which is found in five of the six enzyme classes. We introduced seven randomized loops on the catalytic face of the monomeric, thermostable $(\beta/\alpha)_8$ barrel of glycerophosphodiester phosphodiesterase (GDPD) from *Thermotoga maritima*. We employed in vitro folding selection based on protease digestion to enrich intermediate libraries containing three to four randomized loops for folded variants, and then combined them to assemble the final library (10^{14} DNA sequences). The resulting library was analyzed by using the in vitro protease assay and an in vivo GFP-folding assay; it contains $\sim 10^{12}$ soluble monomeric protein variants. We isolated six library members and demonstrated that these proteins are soluble, monomeric and show $(\beta/\alpha)_8$-barrel fold-like secondary and tertiary structure. The quality of the folding-enriched library improved up to 50-fold compared to a control library that was assembled without the folding selection. To the best of our knowledge, this work is the first example of combining the ultra-high throughput mRNA display method with selection for folding. The resulting $(\beta/\alpha)_8$ barrel libraries provide a valuable starting point to study the unique catalytic capabilities of the $(\beta/\alpha)_8$ fold, and to isolate novel enzymes.

Introduction

Directed evolution experiments have generated numerous commercially valuable enzymes and have helped gain insight into the origin and evolution of enzymatic function. The success of any directed-evolution experiment fundamentally depends on the diversity and quality of the starting library of protein variants. A protein library is considered of high quality if a substantial fraction of the library consists of well-folded, soluble, and stable proteins that contain a diverse set of mutations and potential active sites for a variety of desired activities. In vitro selection strategies generally outperform in vivo or screening approaches by several orders of magnitude in terms of library diversity, and are preferred for the isolation of very rare mutants, for example, novel enzymes.[1] However, high quality enzyme libraries that can harness the ultra-high throughput of in vitro methods are currently lacking.

The ubiquitous $(\beta/\alpha)_8$ or "TIM barrel" fold is a promising scaffold for a general-purpose protein library that could be used for the isolation of new enzymatic activities and the understanding of the origins of enzymatic function. This versatile fold is used in five of the six enzymatic classes and is highly favored by natural enzymes to catalyze a wide array of reactions, in some cases at the diffusion rate limit.[2] In the $(\beta/\alpha)_8$ barrel fold, the main structural and catalytic elements are spatially separated. The barrel itself is formed by eight alternating alpha helices and beta strands and provides the structural foundation; the eight loops connecting helices and strands on one side of the barrel are responsible for substrate binding and catalysis and are known as the catalytic face of the barrel. These features are favorable for enzyme engineering, as modification of functional elements is less likely to affect the structural stability of the overall scaffold.[3] In a few cases, the catalytic activities of $(\beta/\alpha)_8$ barrel enzymes have been successfully swapped through protein engineering, to understand how the $(\beta/\alpha)_8$ barrel fold could be recruited to perform new activities.[4] Although, in some other cases, desired activities have been obtained by altering the substrate specificity of existing enzymes through targeted mutagenesis,[5] the introduction of novel activities often necessitated more extensive protein remodeling.[1b,c,6] In an effort to enable more divergent sequence exploration (well beyond that obtainable from point mutations), the tolerance of $(\beta/\alpha)_8$ scaffolds to the insertion of different natural $(\beta/\alpha)_8$ loop fragments was investigated.[7] Furthermore, the enzymatic activity of existing $(\beta/\alpha)_8$ barrel proteins has been improved or modified by a combination of rational

[a] Dr. M. V. Golynskiy,+ J. C. Haugner, III,+ Prof. B. Seelig
BioTechnology Institute & Department of Biochemistry
Molecular Biology and Biophysics, University of Minnesota, Twin-Cities
1479 Gortner Ave, St. Paul, MN 55108 (USA)
E-mail: seelig@umn.edu

[+] These authors contributed equally to this work.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/cbic.201300326.

design and directed evolution, similarly to proteins of other folds.[2c,8] In addition, rational design approaches for de novo enzymes have repeatedly favored the $(\beta/\alpha)_8$ barrel fold over others, likely because of the ease of positioning appropriate catalytic and substrate-binding residues.[9] This is particularly significant as, despite recent success in the rational redesign of enzymes, de novo design of enzymes is still considered a formidable task.[9b,10] In summary, the combination of valuable $(\beta/\alpha)_8$ barrel protein features, such as catalytic versatility, efficiency, stability, structural modularity, and plasticity, make this fold an ideal scaffold for enzyme engineering.

Herein we report the construction of a highly diverse library of $(\beta/\alpha)_8$ barrels (~10^{14} unique DNA sequences) that contains seven randomized loops, and the enrichment for well-folded, soluble proteins. Unfortunately, the deleterious effect of mutations on stability is a major constraint in protein evolvability[11] and is implicated in limiting the speed of evolution in nature.[12] Previous studies have predicted that the probability of a protein retaining its structure declines exponentially with the number of mutations.[13] An additional concern during the creation of a highly diverse protein library is the unavoidable occurrence of frameshifts and unintended stop codons, caused by imperfect chemical synthesis of the respective DNA library; this can greatly reduce the number of full-length library members.[14] To generate a high quality library, we employed two complementary strategies. In the first strategy we removed stop codons and frameshifts from shorter library cassettes through in vitro selection by mRNA display.[14] In the second strategy we selected for folded protein variants by using protease digestion, which removes poorly folded proteins as they are more susceptible to proteolysis.[15] We combined these two strategies by assembling our final library in vitro and step-wise from intermediate libraries preselected for folded variants and the absence of frameshifts or premature stop-codons (Figure 1). Although the selection procedures reduce the number of protein variants in the intermediate libraries, diversity is regenerated in the final library by recombining these preselected intermediate libraries. Unlike previous $(\beta/\alpha)_8$ library construction attempts in which 49 amino acids were simultaneously inserted into all eight loops in the catalytic face of the $(\beta/\alpha)_8$ fold and likely caused unfolding of the substantial fraction of the final library,[14,16] our conservative step-wise assembly approach aimed to significantly improve the overall library quality. In order to assess the impact of our folding selection, we additionally prepared a control library (without the folding selection). The quality of the two libraries was assessed independently by orthogonal in vitro and in vivo folding assays. These libraries will be used for isolating de novo activities as well as for studying the origins of enzymatic function, the role of folding on the emergence of activity, and the adaptability of the omnipresent TIM barrel fold for different catalytic functions.

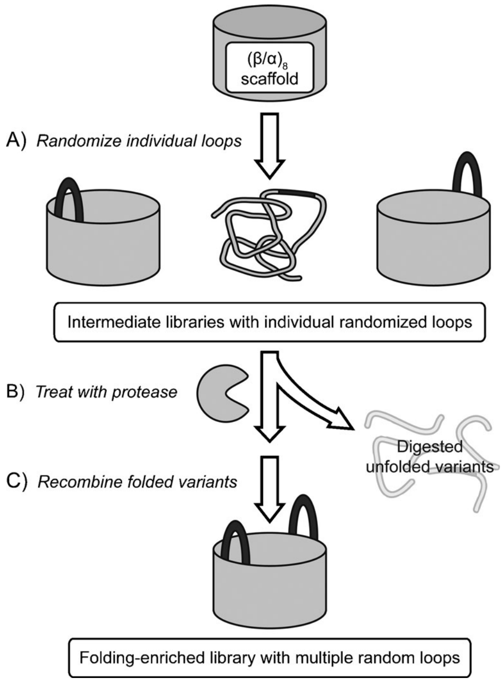
(β/α)8 scaffold
A) Randomize individual loops
Intermediate libraries with individual randomized loops
B) Treat with protease
Digested unfolded variants
C) Recombine folded variants
Folding-enriched library with multiple random loops

Figure 1. General strategy for the stepwise construction of the folding-enriched library based on the $(\beta/\alpha)_8$ scaffold. Selection for folded proteins (protease digestion of unfolded variants) is followed by recombination of folded variants to generate the final $(\beta/\alpha)_8$ library with seven randomized loops.

Results

Identification and characterization of a $(\beta/\alpha)_8$ scaffold protein and an unfolded control

We first sought to identify a suitable $(\beta/\alpha)_8$ scaffold candidate as a starting point for the library design. We desired a highly stable, cysteine-free, monomeric protein with a crystal structure, and chose glycerophosphodiester phosphodiesterase (GDPD) from the hyperthermophile *Thermotoga maritima* as the starting scaffold as it meets all these criteria (Figure 2).[17] We hypothesized that the overall structure of the GDPD protein would be sufficiently stable to tolerate the replacement of loops on the catalytic face of the barrel with random sequences, and even the insertion of additional amino acids. The GDPD catalytic face consists mainly of short loops and could potentially accommodate larger active sites with minimal steric clashes, similarly to recent experiments that changed TIM barrel activities.[7c] To optimize our protocols for folding assessment and selection (essential to our library assembly strategy), we prepared GDPDmut, a destabilized GDPD construct that lacked the parental tertiary structure. In particular, two adjacent large hydrophilic substitutions (G31R/V32E) were introduced into the $(\beta/\alpha)_8$ barrel to disrupt the parent GDPD structure (GDPDwt) through steric clashing and the insertion of unfavorable charge into the tightly packed core of the barrel.

To confirm that the mutant construct did not have the parental $(\beta/\alpha)_8$ structure, GDPDwt and GDPDmut were expressed, purified by His_6 tag chromatography, and characterized in solution. Unlike GDPDwt, GDPDmut did not express solubly, but it could subsequently be solubilized by a refolding step after

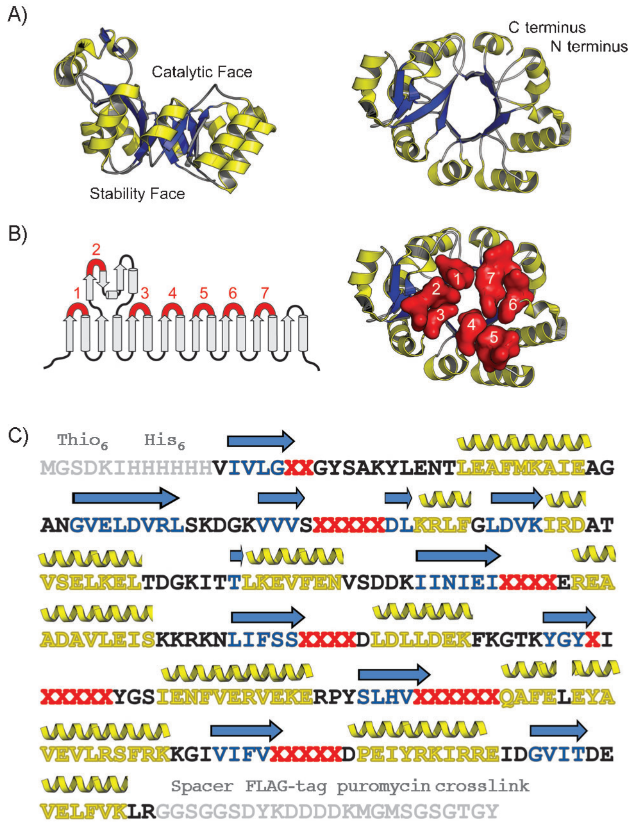
A)
Catalytic Face
Stability Face
C terminus
N terminus
B)
1
2
3
4
5
6
7
C)
Thio6
His6
MGSDKIHHHHHHVIVLGXXGYSAKYLENTLEAFMKAIEAG
ANGVELDVRLSKDGKVVVSXXXXXDLKRLFGLDVKIRDAT
VSELKELTDGKITTLKEVFENVSDDKIINIEIXXXXEREA
ADAVLEISKKRKNLIFSSXXXXDLDLLDEKFKGTKYGYXI
XXXXXYGSIENFVERVEKERPYSLHVXXXXXXQAFELEYA
VEVLRSFRKKGIVIFVXXXXXDPEIYRKIRREIDGVITDE
Spacer
FLAG-tag
puromycin crosslink
VELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY

Figure 2. Design of the $(\beta/\alpha)_8$ library based on the GDPD protein scaffold. A) Side and top views of crystal structure of the GDPD $(\beta/\alpha)_8$ scaffold that was used as a starting point for the library construction (PDB ID: 1O1Z); α-helices and β-strands are shown in yellow and blue, respectively. B) Secondary-structure representation and GDPD scaffold. loops 1–7, which were randomized during library construction, are numbered and shown in red. C) Sequence of the GDPD library, showing positions randomized with the NNG/C codon (red, "X"), β-strands (blue), α-helices (yellow), and non-native residues added to the termini of the GDPD scaffold (purification tags, spacers, and puromycin-crosslinking region needed for mRNA display; gray).

purification under denaturing conditions. In contrast to the monomeric GDPDwt, GDPDmut exists almost exclusively as an oligomeric species in solution, as shown by size exclusion chromatography (Figure S1 A). Analysis of the secondary structure by far-UV circular dichroism (CD) demonstrated that both constructs possess defined, yet differing, elements of secondary structure, based on the similarities at 208 nm and differences at 222 nm (wavelengths associated with α-helical structure in the far-UV CD; Figure S1 B). In order to gain greater insight into the overall folding of the two GDPD constructs, we probed the tertiary structure through 1-anilinonaphthalene-8-sulfonic acid (ANS) fluorescence and near-UV CD (Figure S1 C and S1 D). Both methods showed that GDPDmut has substantially less tertiary structure and a more exposed hydrophobic surface area, relative to GDPDwt. After establishing that GDPDmut lacks the tertiary and quaternary structure of the parent GDPD scaffold, the two constructs were used to establish and optimize the dynamic range of the protease digestion folding selection.

Optimization of the folding selection by in vitro protease digestion

In order to employ protease digestion selection to reduce the fraction of poorly folded protein variants in our $(\beta/\alpha)_8$ library, we first optimized the selection conditions to successfully discriminate between GDPDwt and GDPDmut. Selection based on protease digestion with phage and ribosome display has successfully enriched protein libraries for folded members.[15a,c] Although generally used to improve the stability of a single protein, in one case this approach was applied to improve qualities of de novo libraries based on specific secondary modules.[15b] Throughout our assembly protocol we used mRNA display, an in vitro selection and evolution method that employs the small molecule puromycin to covalently attach proteins to their own mRNA.[18] This method had been previously used to isolate an enzyme de novo from a noncatalytic scaffold with two randomized loops,[1b,c,19] and is excellently suited for the long term goal of isolating enzymatic activities from large protein libraries such as described here.

In pilot experiments, mRNA-displayed proteins were first treated with several proteases known to have preferences for hydrophobic residues (data not shown), and then His_6 tag purified by immobilized metal affinity chromatography (Figure S2). We hypothesized that hydrophobic residues would serve as a good criterion for removing unfolded proteins from the library, as such residues are preferentially buried in the protein core (less likely to be surface-exposed in well-folded proteins).[20] Chymotrypsin, which cleaves adjacent to large hydrophobic residues, showed the largest discrimination between the two control constructs in the pilot experiments; the method was further optimized to yield ~140-fold enrichment of GDPDwt over GDPDmut ((92 ± 1.2) vs. (0.67 ± 0.12)% survival). "Percent survival" is defined as the ratio in His_6 tag purification yield between protease-treated and untreated samples. Furthermore, mRNA-displayed fusions of GDPDmut and a GDPDmut control lacking the His_6 tag were analyzed by the same protocol to determine the level of nonspecific background binding. The optimized chymotrypsin protocol was utilized for the selection and analysis of the $(\beta/\alpha)_8$-based libraries.

Construction of intermediate libraries with randomized loops

Intermediate libraries with several randomized loops were used as building blocks during the step-wise assembly of the final folding-enriched library (Figure S3). To further increase the diversity of the libraries, we also inserted one to four additional amino acids into these loops, with the exception of loop 1 (Table 1). We generated seven libraries, each with a single randomized loop (corresponding to loops 1 to 7 on the catalytic face of the scaffold). In the next step, these libraries were used to assemble intermediate libraries with multiple randomized loops (Figure S3 A). Specifically, fragments of the GDPD gene were PCR amplified to introduce two to six NNS (S = G/C) randomized codons at the desired loop positions; the resulting fragments were digested with restriction enzyme and ligated

Table 1. Comparison of loop length in the GDPDwt scaffold to the randomized loops used in assembling the $(\beta/\alpha)_8$ libraries.

Loop	1	2	3	4	5	6	7	8
GDPDwt loop size	2aa	3aa	1aa	1aa	5aa	2aa	4aa	1aa
library loop size	2aa	5aa	4aa	4aa	6aa	6aa	5aa	wild type

together to generate libraries encoding full-length proteins that contain one or two random loops. Next, "half-libraries" (three or four random loops) were generated, by combining PCR-amplified fragments of the libraries with one or two random loops. Loop 8 was omitted from the library assembly as its location is distant from the core of the $(\beta/\alpha)_8$ barrel and, therefore, unlikely to contribute to the formation of a potential active site with the rest of the randomized regions.

The introduction of multiple loops into the GDPD protein was expected to substantially destabilize the starting scaffold and reduce the fraction of folded proteins in a given library. To guide the library assembly process and decide at which step to perform either the whole folding selection or the mRNA display alone, we first analyzed the protease digestion rates of several intermediate libraries (described below). The mRNA display procedure removed proteins with unintended stop codons (introduced by the use NNS codons for randomization, and by errors during DNA primer synthesis).[14] The mRNA display therefore increased the quality of a library, which was beneficial for subsequent folding selection.

Folding selections of the intermediate libraries by in vitro protease digestion

To evaluate the tolerance of the GDPD scaffold to amino acid insertion and randomization, several libraries containing one or two randomized loops were treated with chymotrypsin to assess the fraction of surviving library members (Table 2). As expected (likely due to steric clashes between random loops from different libraries), the survival rate for libraries with two randomized loops was lower than the product of the survival rates of the two parent libraries with a single randomized loop each. The survival rates for libraries containing one or two randomized loops were significantly above the GDPDmut background (Table 2). Therefore, to preserve some spatial context of the randomized loops, we subjected those libraries only to mRNA display to remove stop codons, and then recombined them into the two half libraries, termed "L1–4" and "L5–7" (randomized loops 1–4 and 5–7, respectively). Our goal was to enrich these two libraries for folded proteins until the survival rate was well above that of GDPDmut in as few rounds of selection as possible, to preserve library diversity. These libraries (possessing four and three randomized loops, respectively) were therefore subjected to folding selection (Figure S3 A). Library L5–7 exhibited 52 % survival rate, and L1–4 showed a significantly lower rate (1.4 %); the surviving variants were subjected to a second round of folding selection, thereby yielding a final survival rate of 9.2 %. The increase in survival rate (well above background) implies that both half libraries were indeed enriched for folded sequences. Additional rounds of folding selection would decrease the diversity of enriched sequences without necessarily improving folding much further (Table 2).

Table 2. Results of the folding selection by in vitro protease digestion.

	Digested species	Survival [%][a]
control constructs	GDPDwt	92 ± 1.2
	GDPDmut	0.67 ± 0.12
	GDPDmut ($-$His$_6$)[b]	0.4
	L3 ($-$His$_6$)[b]	0.3
analytical selections[c]	L3	28
	L4	78
	L5	80
	L3–4	10
preparative selections[d]	L1–4 (1st round)	1.4
	L1–4 (2nd round)	9.2
	L5–7 (1st round)	52
final libraries	folding-enriched	6.6 ± 1.1
	control	2.2 ± 0.3[e]

[a] % Survival is defined as fraction of mRNA-displayed species that are not digested during the chymotrypsin treatment and is calculated as the ratio (Ni-NTA purification yield of chymotrypsin treated species)/(Ni-NTA purification yield of undigested species). [b] Constructs lacking the His$_6$ tag needed for Ni-NTA purification. [c] Small scale selections to assess tolerance of GDPDwt to the insertion of one or two loops to guide the library assembly process. [d] Preparative selections performed to generate intermediate libraries used for the assembly of the final folding-enriched library. [e] The % survival for the control library (loops 1–7 randomized) is higher than for library L1–4. This result is counter-intuitive and likely due to an artifact in the protease assay, possibly caused by unfolded proteins that escaped the protease digestion by aggregating (false-positives).

Assembly of the final folding-enriched library

The stop-codon-free, folding-enriched variants from libraries L1–4 and L5–7 that survived the protease digestion selection (~10^9 and 10^{10} sequences, respectively) were used to assemble the final folding-enriched library, with a total of 32 randomized amino acid positions. Although combining these intermediate libraries could theoretically produce ~10^{19} unique sequences, the actual amount is limited to the sub-milligram quantities of DNA that can be synthesized in the lab. Our final library contained 1.6×10^{14} unique DNA sequences and is at the upper limit of library sizes compatible with in vitro selection methods such as mRNA and ribosome display.

Analysis of stability of folding-enriched library and comparison to control library by using the protease assay (in vitro)

In order to assess the benefits of the folding selection, a control library was prepared from the same seven single-loop libraries used during the construction of the folding-enriched library (Figure S3 B). The resulting library shared the same randomized elements as the folding-enriched library (and a comparable 2.9×10^{14} complexity), but had not been preselected to maintain the parent $(\beta/\alpha)_8$ fold. A single round of mRNA display was employed to remove the stop codons and frameshifts immediately prior to the final recombination step. Rather than using the full-length GDPD gene, only half-gene fragments of the L1–4 and L5–7 libraries were subjected to a round of mRNA display. By using only these fragments instead of the whole parent scaffold, we aimed to avoid a bias of the randomized loops towards the folded parent structure and thus allow maximum diversification. To assess the impact of the folding selection by protease digestion, we directly compared a small fraction (~10^{10} sequences) of the control and folding-enriched libraries by our protease protocol. The folding-enriched library had a 6.6% survival rate: threefold higher than the control library assembled from the L1–4 and L5–7 fragments that had not been selected for folding (Table 2).

Assessment of folding of the final libraries by a GFP-fused reporter assay (in vivo)

In order to confirm the efficacy of the protease digestion folding selection with an independent method, we analyzed a fraction of our libraries with a GFP-fused folding reporter system. In this system, the proteins were expressed as N-terminal fusions of GFP. The GFP fluorescence of the protein–GFP constructs is dependent on the soluble expression of the folded cargo protein and correlates with the stability to intracellular degradation.[21] This approach had been employed to enrich smaller protein libraries (up to 10^8) for folded variants in vivo, and thus is an alternative to our in vitro folding selection.[21a] We first analyzed the several intermediate libraries that were used to construct the control library and compared them to the GDPDwt and GDPDmut controls (Figure S4). These intermediate libraries contained one to four randomized loops and had not been selected for folding. As GDPDwt was shown to be solubly expressed and well behaved in solution, we were interested in the fraction of our libraries that exhibited fluorescence similar to that of the GDPDwt–GFP fusion. In addition, we determined the mode of the GFP fluorescence as a qualitative metric for general library trends as GFP fluorescence correlates with intracellular stability. Flow cytometric analysis of *Escherichia coli* BL21(DE3) cells expressing GDPD–GFP constructs showed a near base-line separation between GDPDwt and GDPDmut, both of which exhibited significantly higher fluorescence than cells transformed with an empty vector (control plasmid). Analysis of the non-preselected libraries showed that libraries with randomized loops in the N-terminal half of the $(\beta/\alpha)_8$ barrel (libraries L1–2, L3–4, and L1–4) exhibited lower GFP fluorescence and lower GDPDwt-like fraction, compared to the libraries with randomized loops in the C-terminal half of the $(\beta/\alpha)_8$ barrel (libraries L5, L6–7, L5–7; Figure S4, Table S1). The folding-enriched and control libraries were analyzed in the *E. coli* BL21(DE3) Rosetta strain, which provides additional tRNA for enhanced eukaryotic protein expression, as the assembly process potentially enriched for eukaryotic codons that are suboptimal in conventional bacterial expression. We observed improvements in the folding-enriched library relative to the control library in both the mode of GFP fluorescence (the most frequently found fluorescence value) and the fraction of GDPDwt-like variants (Figure 3, Table 3).

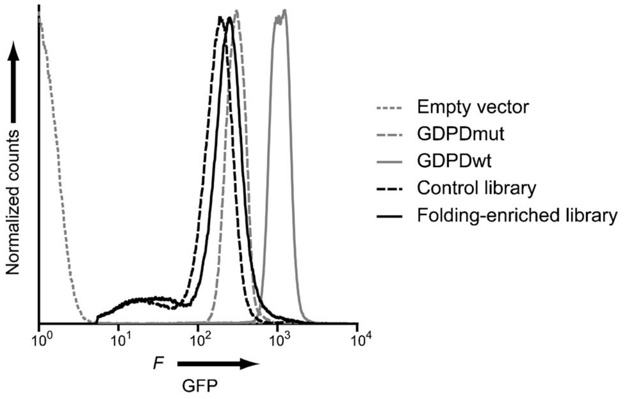

Figure 3. Assessment of folding by GFP-fusion assay. Fluorescence histograms of *E. coli* BL21(DE3) Rosetta cells expressed GFP fused to library members or control proteins as indicated. The empty vector population was gated out on the histograms of cells transformed with the GDPD constructs.

Table 3. GFP-fused in vivo folding assessment of the final $(\beta/\alpha)_8$ fold-based libraries.[a]

Species	Mode of GFP fluorescence[b]	Cells with GDPDwt-GFP fluorescence [%][c]
GDPDmut	24.6	0.01
control library	15.4	1.4
folding-enriched library	19.8	5.4
GDPDwt	100	98.5

[a] Constructs were transformed into *E. coli* BL21(DE3) Rosetta cells. Prior to analysis, data were gated to exclude cell populations that matched fluorescence and scatter profiles of cells transformed with empty vector control plasmid. [b] Values normalized to the mode of GFP fluorescence of GDPDwt-GFP. [c] Wild-type cells were gated on the forward scatter versus GFP contour plot to include ~98% of all wild-type cells.

Isolation of well-folded members of the final libraries by cell sorting

To confirm that soluble expression of GFP fusions is indeed closely correlated with the GDPDwt-like GFP fluorescence, control and folding-enriched libraries were sorted by fluorescence-activated cell sorting (FACS; Figure S5). We subdivided the GDPDwt-like GFP fluorescence window into a low and a high GFP signal during sorting, as the control library exhibited a discrete peak at high signal within this region (gate H in Figure S5 B). Cells with such high GFP profile could be false positives due to either insoluble aggregates or truncated proteins as noted in previous reports that have used the GFP reporter system.[21a]

Analysis of soluble library-GFP fusions by western blotting and SDS-PAGE

The four sorted populations (low and high GFP signal, of each of the control and folding-enriched libraries; Figure S5) were regrown in liquid culture under sorting conditions, and the respective amounts of soluble full-length library-GFP fusion proteins were compared by anti-GFP western blotting (Figure S6).

A fraction of these cultures was also plated, individual GFP-positive clones were isolated, expressed, and the soluble protein fraction of each clone was analyzed by SDS-PAGE (data not shown). The SDS-PAGE and western blot results showed similar trends, and were in good agreement with each other (Table S2). The folding-enriched library populations contained a higher fraction of soluble GFP fusions in both the low and high populations compared to the control library populations. Western blot analysis also showed that the high-GFP populations for both libraries contained at least 50% false positive clones (expressed GFP alone). Based on the fraction of full length, soluble library–GFP fusions in the FACS-sorted populations, we calculated that the soluble library members comprised between 1.0 and 1.2% of the folding enriched library and between 0.02 and 0.033% of the control library. This corresponds to an overall 35- to 50-fold improvement in library quality, based on the fraction of soluble sequences. Therefore, the final folding-enriched $(\beta/\alpha)_8$ fold library contained about 10^{12} soluble protein variants (Table S2).

Biophysical characterization of soluble library clones

We sought to further investigate and compare the solubility of protein variants from the control and folding-enriched libraries selected at random, as well as the folding-enriched variants isolated by FACS (above). All constructs were cloned into a protein expression plasmid to express the FACS-sorted library-GFP constructs without the GFP. Only sequences from the FACS-sorted folding-enriched library produced soluble proteins (data not shown), six of which were purified for further characterization. Similarly to the initial GDPDwt and GDPDmut characterization, we performed size-exclusion chromatography and measured the near-UV CD and ANS fluorescence to investigate the quaternary, secondary, and tertiary structure of these library variants (Figure 4). All of these proteins were monomeric in solution, maintained CD signatures similar to that of GDPDwt, and showed ANS profiles intermediate between those for GDPDmut and GDPDwt.

Sequence analysis of library clones

To better understand the underlying changes that occurred upon our selection for folding, we sequenced randomly chosen individual clones from the control and folding-enriched libraries, as well as the soluble folding-enriched library clones acquired by FACS sorting of the GFP-fused library. We analyzed the amino acid distribution of the 1393 sequenced NNS codons and did not observe any stop codons, thus confirming that these were removed during the mRNA-display step (Table 4). We further grouped the sequenced codons into classes of amino acids based on their properties, and then compared the distributions of these classes for the control library (randomly chosen clones) and the folding-enriched library (randomly chosen clones, and soluble clones; Figure 5). To evaluate whether the detected distribution changes were statistically significant ($p < 0.05$), we performed pairwise t-test comparisons of the grouped codons from the folding-enriched library sequences (random and soluble clones) against the control library sequences (random clones). We observed a significant decrease in aromatic residues in the folding-enriched library relative to the control library. The soluble library clones from the folding-enriched library, isolated during FACS sorting experiment, exhibited the same decrease in aromatic residues, and, in addition, showed an increase in polar residues at the expense of aliphatic residues.

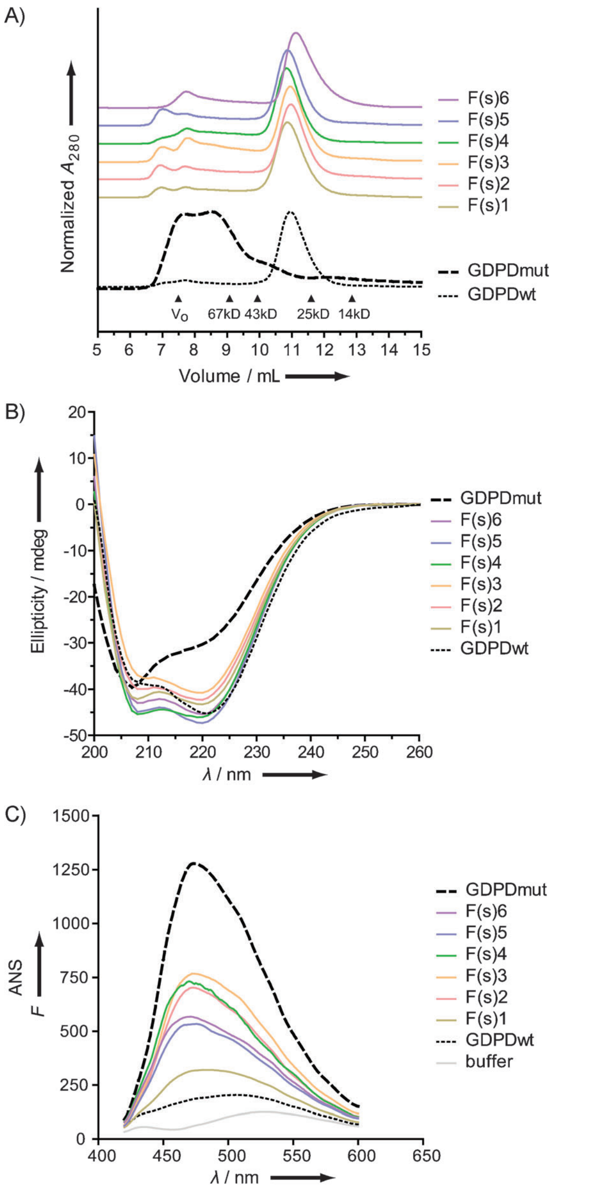

Figure 4. Biophysical characterization of six soluble folding-enriched library clones from the FACS-sorted high GFP population. GDPDwt and GDPDmut data are included for reference. A) Size exclusion chromatography (quaternary structure). B) Far-UV circular dichroism spectroscopy (secondary structure). C) ANS fluorescence (tertiary structure).

Discussion

The objective of this study was to generate and characterize a high quality protein library based on the $(\beta/\alpha)_8$ fold, by com-

Table 4. Amino acid (aa) distribution for NNS codons, shown in %.

Amino acid		NNS (−stop)[a]	Control library[b] random clones	Folding-enriched library[b] random clones	Folding-enriched library[b] soluble clones
polar	Asn	3.2	3.8	5.3	3.4
	Gln	3.2	3.2	3.8	2.8
	Ser	9.7	7.0	8.4	15.5
	Thr	6.5	7.0	5.1	6.6
basic	Arg	9.7	10.2	10.1	11.0
	His	3.2	3.2	4.2	3.1
	Lys	3.2	5.1	5.1	3.8
acidic	Asp	3.2	4.0	4.4	4.5
	Glu	3.2	2.2	2.7	3.8
aliphatic	Ala	6.5	6.1	6.9	6.6
	Ile	3.2	2.9	2.7	3.1
	Leu	9.7	9.7	11.2	8.3
	Met	3.2	4.3	2.9	2.4
	Val	6.5	6.7	5.5	3.1
aromatic	Phe	3.2	4.1	1.9	3.1
	Trp	3.2	3.0	2.9	1.4
	Tyr	3.2	2.9	1.9	2.1
structural	Cys	3.2	1.3	1.3	1.4
	Gly	6.5	5.4	6.3	8.3
	Pro	6.5	8.0	8.0	5.9
	stop	0	n.o.	n.o.	n.o.
codons sequenced			628	475[c]	290[c]

[a] Theoretical aa distribution for NNS(−stop) was calculated from the expected NNS distribution lacking a stop codon. [b] Experimentally observed values from sequencing analysis of individual library clones. [c] Loops containing wild-type sequences were omitted from analysis. n.o.: not observed.

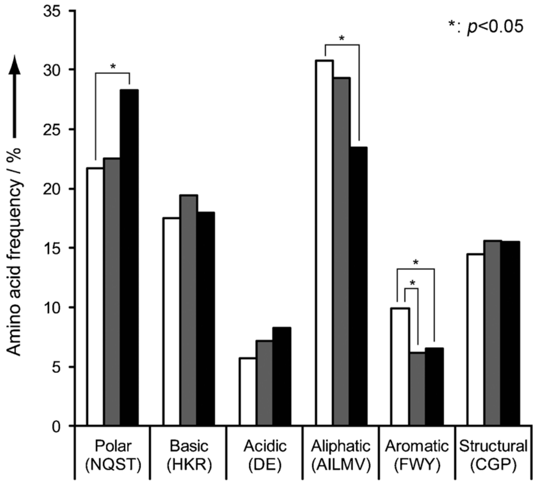

Figure 5. Amino acid composition of randomized loop regions. Amino acids are grouped according to chemical properties; compositions were calculated from sequencing data. Control library, randomly picked clones (white); folding-enriched library, randomly picked clones (gray); folding-enriched library, soluble clones (black). Statistically significant differences, as determined by pairwise t-test, are indicated by a star.

bining step-wise assembly with in vitro folding selection. We further sought to evaluate the efficacy of such an approach by comparing a representative fraction of members of the libraries with two orthogonal methods for the assessment of folding.

Our in vitro and in vivo folding assessment methods provided different metrics to measure folding stability; these were survival rates during protease digestion, the mode of fluorescence of GFP-fused library members, and the fraction of library members that behave like GDPDwt in the GFP assay. All three metrics displayed similar trends for the intermediate libraries, and showed a substantial improvement in the quality of the folding-enriched library compared to the control library, thus demonstrating the success of our folding selection strategy. Although these metrics were useful to characterize the libraries in bulk and assess the library construction process, they were only indirect measures for determining how much the library was enriched for soluble, well-folded protein variants that behaved like the starting $(\beta/\alpha)_8$ scaffold. To quantify directly the fraction of desired library variants, we cloned and expressed 20 randomly chosen proteins from both libraries in *E. coli*. We did not obtain any soluble proteins from this small sample size, thus indicating that the fraction of soluble variants in each library was below 5%. We therefore sorted a fraction of the GFP-fused libraries by FACS and were able to isolate library members that readily expressed in bacteria, were monomeric, and exhibited behavior similar to that of GDPDwt in solution. Sequencing results suggested that, as expected, improved solubility correlates with increased presence of polar amino acids at the expense of aliphatic residues. Furthermore, the occurrence of aromatic amino acids was reduced in the folding-enriched library compared to the control library; this might in part be a result of the selection process (disfavoring those residues because of chymotrypsin's preference to cleave next to aromatic amino acids). Based on the number of soluble, GDPDwt-like clones we obtained from the sorting experiment and the biochemical characterization of individual clones, we calculated that soluble, monomeric and folded sequences comprise about 1% of the folding-enriched library (~10^{12} variants)—an increase over the control library of up to 50-fold.

The in vitro and in vivo folding methods employed in our work required the fusion of the $(\beta/\alpha)_8$ library proteins to either their own mRNA or a GFP reporter protein; this could, in principle, alter stability or solubility of the proteins. To minimize this during the in vitro protease digestion, the mRNA was reverse-transcribed to generate a linear mRNA–cDNA hybrid, thereby preventing the mRNA from folding and affecting the digestion by obscuring protease sites. Furthermore, we assessed whether the fusion to GFP affected solubility of the library proteins by expressing soluble library-GFP constructs without the GFP fusion and analyzing them by SDS-PAGE: all proteins remained soluble. Notably, during the FACS sorting experiment we encountered a substantial number of false-positive highly fluorescent cells; these had resulted from clones that had lost their GDPD library cargo, thereby leading to the expression of GFP alone. It has been proposed in earlier work that such false positives result from either truncated or highly aggregated and insoluble species.[21a] We were able to exclude these false-positives by analyzing the soluble fraction of the expressed proteins by gel electrophoresis. The folding selection by protease digestion likely also allowed some false-positive protein variants to be selected. For example, we envisioned that certain unfolded proteins could escape protease digestion through aggregation, as such proteins would be inaccessible to the protease. We counteracted this possibility by including detergents and denaturants (Triton X-100 and SDS) in our buffers. Yet, as we cannot rule out residual selection bias of this kind, we deliberately chose not to further enrich the intermediate libraries L1–4 and L5–7 beyond the initial one or two rounds of selection. Finally, the biophysical characterization of individual soluble library members confirmed that our protease selection protocol successfully enriched for folded variants with a structure similar to the parental $(\beta/\alpha)_8$ scaffold.

The final folding-enriched library contains up to 32 randomized amino acid positions distributed over seven loops. The soluble library variants isolated by FACS exhibited some variability in the location and number of loops that were randomized. Interestingly, randomization in loops 2 and 3 was disfavored in the folding-enriched library, as we frequently recovered the parent GDPDwt sequence in these loops (~80 and ~40% parent sequence, respectively, in randomly picked clones). All soluble clones isolated in the FACS experiments showed the parent sequence in loops 2 and 3, while containing other randomized loops. In addition, libraries that contained randomized loop 2 and/or 3 also exhibited lower protease survival rates and lower GFP fluorescence; this is further evidence that randomization here is detrimental to the stability of the $(\beta/\alpha)_8$ barrel. We suspect that the wild-type loops observed in the final library arose during the step-wise library assembly. Although initially present only at very low levels in the intermediate libraries, these variants were enriched by the folding selections. However, the $\sim 10^{12}$ soluble members of the folding-enriched library had at least three randomized loops and at least 13 randomized amino acids. For comparison, a recent study described the switch of one $(\beta/\alpha)_8$ scaffold enzyme to an unrelated $(\beta/\alpha)_8$ activity by a single loop insertion.[7c] If a new enzymatic activity can be found with the exchange of a single loop (as these results suggest), our library of soluble proteins with three and more randomized loops has an even greater potential to contain different enzymatic activities. In addition, some of the less soluble library members might also be exploitable by in vitro selection methods as, for example, the mRNA display has been shown to help keep poorly soluble proteins in solution through the attachment of a large highly-soluble RNA molecule. However, the solubility of such proteins would subsequently need to be improved through directed evolution, unlike the $\sim 10^{12}$ already soluble library members. In summary, we demonstrated that the soluble clones had retained most of the overall structural features of the parent $(\beta/\alpha)_8$ fold, despite the introduction of multiple randomized stretches of amino acids. To the best of our knowledge, this is the first report of a high-quality library based on the $(\beta/\alpha)_8$ enzyme fold with such high complexity.

Our work also allowed us to make several observations regarding the behavior of the GDPDwt $(\beta/\alpha)_8$ fold, the role of randomized loop positions, and the impact of combining individual loop libraries. We observed that single, entirely randomized loop insertions into GDPDwt resulted in libraries with 30–80% survival in protease-digestion folding selection. Interestingly, prior in vivo work demonstrated that single known loops inserted into an unrelated $(\beta/\alpha)_8$ barrel resulted in similar tolerance with regards to folding.[7b] The authors suggested that it was the site of insertion and not the inserted sequence that had the greatest influence on the stability of the resulting protein chimera. The results we present here strongly support this notion and suggest that other $(\beta/\alpha)_8$ barrels might exhibit similar tolerance to single loop insertion, regardless of whether the loop sequence had been favored previously in nature or is entirely random. In fact, previous work suggests that random regions are beneficial in adapting known loops to the context of a new $(\beta/\alpha)_8$ barrel structure.[7a] When we combined two libraries with different folding stabilities, the resulting library displayed lower folding stability than the less-stable input library, as evidenced by both protease digestion and the GFP-fusion assay for multiple libraries. We observed a general trend that the N-terminal half of the barrel appears more vital for folding stability than the C-terminal half. This finding was inferred from the low GFP fluorescence, the high protease digestion rates, and the sequencing results for libraries containing randomized N-terminal loops. Similar positional preferences were observed in previous experiments on another $(\beta/\alpha)_8$ scaffold.[7b] Although we were initially concerned that the introduction of several randomized loops into the GDPDwt scaffold would drastically unfold the $(\beta/\alpha)_8$ structure, by all our metrics the data indicate that this scaffold is tolerant to multiple loop insertions, particularly in the C-terminal half of the barrel. In summary, our results support the hypothesis that the core of a hyperthermophile $(\beta/\alpha)_8$ barrel fold provides sufficient stability to offset the effects of destabilizing loops of the catalytic face, and thus render the $(\beta/\alpha)_8$ fold an attractive scaffold in enzyme engineering by loop insertion.

Conclusions

The high quality and complexity of the libraries reported here are expected to provide an invaluable starting point for the engineering of novel enzymes and the understanding of the origins of enzymatic function in the $(\beta/\alpha)_8$ fold. By introducing randomized elements onto a stable scaffold in step-wise fashion and enriching for folded variants, we increased the probability of finding novel enzymes with diverse activities. These initial, potentially low-enzymatic activities could subsequently be evolved further under appropriate selection conditions to give rise to more efficient specialist enzymes.[16,22] Many $(\beta/\alpha)_8$ enzymes act on substrates with a phosphate group, and some soluble variants of the folding-enriched library have retained the residues that comprise the native phosphate binding site. This site can be used as a handle to improve substrate binding or to study the role of such handles in the evolution of enzymes. Furthermore, isolating novel activities from these libraries that are unrelated to the original GDPD function will help to elucidate whether the $(\beta/\alpha)_8$ barrel fold is predestined for certain activities, how it can be adapted to perform new functions, and what impact a library preselected for folding might have on isolation of enzymatic activity. Finally, an estimated 1% of our folding-enriched library contains molecules that are solubly expressed in *E. coli* and show substantial diversity in the number and positioning of randomized loops. Our libraries are thus compatible with in vitro and in vivo evolution methods. Work is underway to interrogate the libraries for de novo enzymes by using mRNA display, and to study the $(\beta/\alpha)_8$ fold adaptability through bacterial selections.

Experimental Section

All chemicals were purchased from Sigma–Aldrich unless otherwise stated. All restriction enzymes, Alkaline Phosphatase, Calf Intestinal, T7 RNA ligase, T4 DNA ligase and Phusion High Fidelity DNA polymerase were purchased from New England Biolabs (Ipswich, MA). All PCR reactions were performed with Phusion High Fidelity DNA polymerase. If available, high-fidelity versions of the restrictions enzymes were employed. Gel extraction, PCR purification and DNA mini-prep kits were purchased from Qiagen (Valencia, CA). Sequencing reactions were performed either by ACGT, Inc. (Wheeling, IL) or the University of Minnesota BioMedical Genomics Center (St. Paul, MN).

Cloning and expression of GDPDwt and GDPDmut constructs: The synthetic gene encoding GDPD flanked by purification tags, optimized for dual expression in rabbit reticulocyte and *E. coli*, was purchased from GenScript (Piscataway, NJ). Specifically, the gene coded for $Thio_6His_6$ tag–GDPDwt–$(GGS)_2$ spacer–FLAG epitope–pyromycin crosslinking region. This construct was PCR amplified and cloned into pET28a vector (Novagen/Millipore). GDPDmut was generated by using standard mutagenesis protocols with pET28/GDPDwt as the template. For protein expression, plasmids were transformed into BL21(DE3) Rosetta *E. coli* (Novagen) and clones were grown on LB medium in the presence of kanamycin ($34\,mg\,L^{-1}$) and chloramphenicol ($34\,mg\,L^{-1}$). Overnight cultures were diluted 1:1000 into fresh LB media and grown to $OD_{600}=1$ prior to induction with IPTG (1 mM). Cells were grown for an additional 4 h at 37 °C prior to harvesting and storage at −20 °C. Frozen cell pellets were resuspended in lysis buffer (Tris·HCl (50 mM pH 8.0), NaCl (50 mM)) and lysed by using an S-450D Digital Sonifier (Branson, Danbury, CT). Cell debris was removed by centrifugation, and the His_6-tagged proteins were purified by affinity chromatography on Ni-NTA Superflow resin (Qiagen) under native conditions for GDPDwt and buffers containing denaturant (guanidinium chloride (6 M)) for GDPDmut, according to the manufacturer recommendation. Elution fractions containing GDPDmut were dialyzed to remove denaturants by first diluting 1:4 in dialysis buffer (Tris·HCl (50 mM, pH 7.5), NaCl (100 mM)), then dialyzing overnight in 7 kDa MWCO Snake Skin Dialysis Tubing (Pierce/Thermo Scientific) in dialysis buffer. Protein purification was evaluated by SDS-PAGE in precast 4–12% gradient gels (Invitrogen/Life Technologies).

Circular dichroism (CD) spectroscopy: All CD experiments were performed on a J-815 spectropolarimeter (Jasco). For far-UV experiments, ellipticity of protein samples (20 μM in Tris·HCl (10 mM, pH 7.5), NaCl (20 mM)) was measured from 190 to 260 nm ($50\,nm\,min^{-1}$) in a quartz cuvette (1 mm path length). Each spectrum represents the average of ten accumulations. For near-UV experiments, ellipticity was measured as for far-UV, except a quartz cuvette with a 10 mm path length was used, and the wavelength was 260–350 nm.

1-Anilinonaphthalene-8-sulfonic acid (ANS) fluorescence measurements: ANS is an environmentally sensitive dye which exhibits increased fluorescence upon interaction with hydrophobic protein surfaces and has been previously used to indirectly report on protein tertiary structure.[23] Measurements were performed on SpectraMax M2 or M5 plate readers (Molecular Devices, Sunnyvale, CA) in black flat-bottom 96-well NUNC Maxisorp plates. Samples containing protein (5 μM) and ANS (1 mM) in dialysis buffer (Tris·HCl (50 mM, pH 7.5), NaCl (100 mM)), were measured at 1 nm intervals ($\lambda_{ex}=403$ nm, $\lambda_{em}=430$–600 nm). Data were smoothed with Kaleidograph software (Synergy Software, Reading, PA).

Size-exclusion chromatography: Ni-NTA purified protein samples were loaded onto a Tricorn column (10 mm×300 mm; GE Healthcare) packed with Superdex 75 resin (GE Healthcare) and analyzed in an AKTA FPLC system (GE Healthcare) in dialysis buffer (Tris·HCl (50 mM, pH 7.5), NaCl (100 mM)). The column was calibrated by using an Amersham low molecular weight calibration kit (GE Healthcare).

Library assembly: All loop libraries were assembled by a three step process: PCR amplification, restriction digest, and ligation (Figure S3). All PCR reactions employed fixed primers at the 5′ and 3′ termini and internal primers containing a restriction site (Tables S3 and S4). Loop randomization and insertion was carried out at the single or double loop library level by amplifying two fragments of GDPDwt from the pET28/GDPDwt template and introducing randomized NNS codons in one of the primers. Assembly of half libraries and final libraries was performed by using internal primers that did not introduce any randomized nucleotides. Following PCR amplification, DNA was phenol/chloroform extracted and ethanol precipitated, by following standard molecular biology protocols.[24] DNA was digested with appropriate restriction enzyme (Table S4) and purified on 2% agarose gel. Purified digested fragments were ligated with T4 DNA ligase (16 °C, overnight). The ligation product was purified on 2% agarose gel and PCR amplified with external primers to generate approximately ten copies of the full length template for use in the next round of library construction. During the construction of folding-enriched library, the L1–2, L3–4, L5, and L6–7 libraries were subjected to a single round of mRNA display (below) to remove stop codons and frameshifts and then combined to generate L1–4 and L5–7 libraries. These libraries were subjected to protease-based selection (below) and then combined to assemble the final folding-enriched library. During the control library assembly, half-gene fragments (L1–4 and L5–7 libraries) were mRNA-displayed to minimize artifacts related to folding. In the final assembly step for both the control and folding-enriched libraries, 10^9–10^{10} L1–4 and L5–7 DNA sequences were amplified on 20 mL scale to generate $\sim 5\times10^{14}$ starting sequences. Because of the increased scale of the BsaI-HF digest and final ligation reactions, DNA purification was performed through a 4.5% native PAGE gel, extracted under UV, and electroeluted with an S&S Elutrap (Schleicher & Schuell, Dassel, Germany).

mRNA display: Creation of mRNA displayed fusions was performed as previously described,[18b] but with the following alteration: RNA was produced from the DNA library with T7 RNA polymerase (template (5 nM), HEPES (200 mM, pH 7.5), $MgCl_2$ (35 mM), spermidine (2 mM), NTP (5 mM each), BSA ($0.1\,mg\,mL^{-1}$), DTT (40 mM), inorganic pyrophosphatase ($1\,U\,mL^{-1}$), RNaseOUT ($150\,U\,mL^{-1}$; Invitrogen), T7 RNA Polymerase ($50\,U\,mL^{-1}$)) and incubated at 37 °C for 3 h. RNA was precipitated by LiCl (1/3 equivalent of LiCl (8 M)) at −20 °C for at least 30 min. The RNA pellet was washed with ice cold ethanol (70%) and dissolved in water. RNA was photo crosslinked by combining transcribed RNA (3 μM) in buffer (HEPES (20 mM, pH 7.5), KCl (100 mM), Spermidine (1 mM), EDTA (1 mM)) with Psoralen–puromycin oligo 5′-X(tagccggtg)AAAAAAAAAAAAAAAZZACCP-3′[18b] (7.5 μM; X = psoralen C6, lowercase letters = 2′-OMe, Z = triethylene glycol, P = puromycin, stretch of A's and ACC are DNA) under 365 nm light on ice for 20 min (efficiency ~ 50 %). Crosslinked RNA was ethanol precipitated and dissolved in water. For translation, the mixture (200 μL or 1 mL) contained crosslinked RNA (200 nм), nuclease-treated rabbit reticulocyte lysate (40 %; Promega), amino acid mix (25 μм), ^{35}S-methionine (25 nм, > 1000 Ci mmol^{-1}; PerkinElmer) KCl (120 mм), and $Mg(OAc)_2$ (0.6 mм), and was incubated at 30 °C for 10 min followed by high-salt incubation (addition of KCl (to 550 mм), $MgCl_2$ (50 mм)) for 5 min at 23 °C. The translation mixture was diluted tenfold into oligo(dT) binding buffer (Tris·HCl (20 mм, pH 8), EDTA (10 mм), NaCl (1 м), Triton X-100 (0.2 %)) and incubated with oligo(dT) cellulose (0.2 mg cellulose per μL translation; GE Healthcare) with rotation for 15 min at 4 °C. The oligo(dT) cellulose was washed on a chromatography column (Bio-Rad Hercules, CA) with oligo(dT) binding buffer, then with oligo(dT) wash buffer (Tris·HCl (20 mм, pH 8) NaCl (0.3 м)), and eluted with elution buffer (Tris·HCl (2 mL, 2 mм, pH 8)). The eluent was spin filtered through a 0.45 μm filter (Millipore) to remove any residual oligo(dT) cellulose and added 10x PBSTr (Na_2HPO_4 (80.6 mм), KH_2PO_4 (19.4 mм), KCl (27 mм), NaCl (1.37 м), Triton X-100 (0.1 %), pH 7.4) to a final concentration of 1×. The mixture was added to Anti-Flag M2-Agarose Affinity Gel (25 μL, equilibrated according to the manufacturer's instructions) and incubated with rotation for at least 1 h at 4 °C. Flag resin was washed on a chromatography column (Bio-Rad) with 1× PBSTr followed by Flag wash buffer (HEPES (50 mм, pH 7.4), NaCl (150 mм), Triton X-100 (0.01 %)) where the final wash was performed in batch in a microcentrifuge tube. Elution was performed by incubating Flag resin with Flag peptide (56 μм in Flag wash buffer) for 10 min at 4 °C with rotation, and filtering through a 0.45 μm filter (Millipore) to remove any Flag resin. The eluent was diluted with Flag elution buffer until mRNA-displayed fusions reached 3×10^8 fusions/μL as measured by scintillation counting (LS6500 multipurpose scintillation counter; Beckman). This was followed by reverse transcription with Superscript II (1.5×10^8 fusions/μL, 50 nм RT-primer (5′-TTTTTT TTTTTT TTTNCC AGATCC AGACAT TCCCAT-3′), Tris·HCl (50 mм, pH 8.3), $MgCl_2$ (3 mм), 2-mercaptoethanol (10 mм), dCTP (0.5 mм), dGTP(0.5 mм), dTTP (0.5 mм), dATP (5 μм), RNaseOUT (100 U mL^{-1}; Invitrogen/Life Technolgies), Superscript II (500 U mL^{-1}; Invitrogen/Life Technolgies)). A sample (10 μL) was removed to serve as a non-radiolabeled control prior to the addition of [α-^{32}P]dATP (16 μм final concentration, 3000 Ci mmol^{-1}; PerkinElmer) to the reverse transcription. Both tubes were incubated at 42 °C for 30 min and the control was stored at −20 °C. The reverse transcription product was treated with alkaline phosphatase, calf intestinal (30 U mL^{-1}) at 37 °C for 10 min. Reverse-transcribed fusions were then dialyzed in a 20 K MWCO Slide-A-Lyzer (Pierce/Thermo Scientific) three to four times against dialysis buffer (Tris·HCl (50 mм, pH 7.5), NaCl (100 mм)) until all unincorporated [^{32}P]dATP had been removed.

In vitro folding selection by protease digestion: The dialyzed fusions were subjected to our folding selection. Triton X-100 (0.1 % (*w/v*)) and sodium dodecyl sulfate (0.05 % (*w/v*)) were added, and fusions were incubated with chymotrypsin (6 μg mL^{-1}; Princeton Separations, Adelphia, NJ) at 30 °C for 5 min. The digest was stopped by the sequential addition of phenylmethylsulfonyl fluoride (2 mм) and KCl (5 mм) and incubated on ice for 10 min. The potassium dodecyl sulfate precipitate was removed by an Ultrafree-MC 0.45 μm Spinfilter (Millipore) at 4 °C followed by addition of three volumes of Ni-NTA binding buffer (sodium phosphate (100 mм, pH 8), Tris·HCl (10 mм, pH 8), NaCl (250 mм), guanidinium chloride (6 м; Amresco, Solon, OH), Triton X-100 (0.1 %)). The mixture was added to one volume of Ni-NTA agarose (Qiagen) pre-equilibrated with Ni-NTA binding buffer, and incubated with rotation for at least 1 h at 4 °C. The Ni-NTA agarose was washed on a chromatography column (Bio-Rad) with more Ni-NTA binding buffer, followed by 5 washes with a mixture of Ni-NTA binding buffer and Ni-NTA native wash buffer (Tris·HCl (10 mм, pH 8), NaCl (250 mм), Triton X-100 (0.01 %)) with final guanidinium concentrations of 4.5, 3, 1.5, 0.5, and 0.25 м. The Ni-NTA agarose was washed with additional Ni-NTA native was buffer followed by elution in Ni-NTA elution buffer (Tris·HCl (50 mм, pH 8), NaCl (50 mм), imidazole (500 mм), Triton X-100 (0.01 %)). The eluent was concentrated to a third of its original volume in a SpeedVac concentrator, then ethanol precipitated, and dissolved in Tris·HCl (10 mм, pH 8) by heating to 80 °C. cDNA was amplified by PCR with Phusion polymerase and primers to add a 5′-UTR (untranslated region) for the next round of mRNA display. Yields from each purification step were determined through Cerenkov counting on an LS6500 multipurpose scintillation counter (Beckman).

GFP-based folding assay: The GFP-based folding assay was based on the pER13a reporter plasmid previously employed to isolate protein variants with improved folding.[21a] The plasmid was kindly provided by R. Sterner lab and contains an out-of-frame GFP. A fraction of the library of interest (~ 10^8–10^9 sequences) was PCR amplified with Phusion polymerase and cloned into pER13a plasmids at NdeI and NotI restriction sites to generate N-terminal fusions to GFP. Libraries were ligated into the digested pER13a plasmid by using T4 DNA ligase. Ligation reaction mixtures were purified through spin columns (PCR Purification Kit, Qiagen) prior to electroporation into electrocompetent NEB 5-alpha cells (New England Biolabs). Following 1 h incubation at 37 °C, cells were plated and grown overnight on kanamycin-containing agar plates. Approximately 10^4–10^5 independent colonies were washed off the plates and their plasmids were isolated (QIAprep Spin Miniprep kit, Qiagen). BL21(DE3) and BL21(DE3) Rosetta cells (Novagen) were used for GFP-fused expression of intermediate (Table S1) and final libraries (Table 3), respectively. Electrocompetent cells were prepared by standard molecular biology protocols,[24] then transformed with ~ 10^8 DNA sequences and grown overnight at 37 °C in LB medium (50 mL) supplemented with kanamycin (75 mg L^{-1}) and chloramphenicol (34 mg L^{-1}). An overnight culture was used to inoculate the same medium (10 mL), and cells were grown to OD_{600} ≈ 0.6, then cooled to 30 °C for 30 min prior to addition of IPTG (0.5 mм). Growth was continued for 6 h at 30 °C. An aliquot (1.5 mL) was pelleted in an Eppendorf 5424 centrifuge (1900*g*, 3 min, RT), washed with PBS (1 mL) and resuspended in PBS (500 μL). Flow cytometry experiments were performed at the University Flow Cytometry Resource (University of Minnesota). Samples were analyzed on a FACSCalibur (BD Biosciences) with 488 nm excitation and emission through a 530/30 nm bandpass filter. The FlowJo software package (TreeStar Inc, Ashland, OR) was used for data analysis. The population of cells transformed with the empty vector was gated out from all experiments before determining the GFP mode (most frequently found fluorescence value) for the remaining cells. Cell sorting experiments were performed on FACSAria (BD Biosciences). Sorting gates were defined by side-scatter versus GFP fluorescence dot plots. The GDPDwt-like population gate was set based on the cells transformed with GDPDwt-GFP construct, while gates for low GFP and high GFP populations were set based on the cells transformed with the GFP-fused control library. Sorted cells were used to inoculate LB medium containing kanamycin and chloramphenicol and re-grown again under sorting conditions as above, for western blot analysis. An aliquot of the regrown cells was removed prior to IPTG induction and plated on LB agar plates containing kanamycin and chloramphenicol. Individual clones picked at random from these plates were grown in liquid culture and analyzed by SDS-PAGE for soluble expression of library-GFP variants. Six clones from the high GFP population of the GFP-fused folding-enriched library (out of 20 soluble clones identified by SDS-PAGE) were subcloned into pET28a and expressed for further characterization, as above for the GDPD control constructs.

Western blot analysis: Cell pellets were lysed by using a BugBuster protein extraction reagent (Novagen/Merck Millipore) according to the manufacturer's recommendations. The insoluble fraction was pelleted and resuspended in the original volume of the BugBuster reagent. Samples were mixed with equal volume of twofold Laemmli sample buffer (BioRad), heated for 5 min at 95 °C and spun down. A fraction of all samples was removed, diluted tenfold and run on 4–12% gradient gel (Invitrogen). Western blotting was performed according to standard protocols[24] by using polyclonal rabbit anti-GFP primary antibody (1:5000 dilution; #290, Abcam, Cambridge, UK). Anti-rabbit secondary antibody labeled with DyLight 800 (1:20 000 dilution; #5151,Cell Signaling Technology, Danvers, MA) was used, and visualized with an Odyssey infrared imaging system (LI-COR Biosciences, Lincoln, NE). Images were analyzed with the Image J software package (NIH) to quantify intensities of anti-GFP stained bands.

Acknowledgements

We thank the R. Sterner lab for providing the pER13a plasmid, P. Champoux for assistance with the FACS experiments, members of J. M. Ervasti lab for access to the Li-Cor imaging system, and B. J. Hackel, R. J. Kazlauskas, M. D. Lane, D. Morrone, K. N. Golynskiy and M. B. Quin for comments on the manuscript. This work was supported in part by the US National Institutes of Health (NIH, T32 GM08347 to J.C.H.), the Biocatalysis Initiative of the BioTechnology Institute at the University of Minnesota (to M.V.G. and B.S.) and the Minnesota Medical Foundation (to B.S.).

Keywords: (beta/alpha)$_8$ barrel · GFP-folding reporter · mRNA display · protein engineering · protein folding

[1] a) H. Leemhuis, V. Stein, A. D. Griffiths, F. Hollfelder, *Curr. Opin. Struct. Biol.* **2005**, *15*, 472–8; b) M. V. Golynskiy, J. C. Haugner III, A. Morelli, D. Morrone, B. Seelig, *Methods Mol. Biol.* **2013**, *978*, 73–92; c) B. Seelig, J. W. Szostak, *Nature* **2007**, *448*, 828–831; d) F.-A. Chao, A. Morelli, J. C. Haugner III, L. Churchfield, L. N. Hagmann, L. Shi, L. R. Masterson, R. Sarangi, G. Veglia, B. Seelig, *Nat. Chem. Biol.* **2013**, *9*, 81–83.

[2] a) N. Nagano, C. A. Orengo, J. M. Thornton, *J. Mol. Biol.* **2002**, *321*, 741–765; b) R. K. Wierenga, *FEBS Lett.* **2001**, *492*, 193–198; c) R. Sterner, B. Höcker, *Chem. Rev.* **2005**, *105*, 4038–4055; d) S. C. Blacklow, R. T. Raines, W. A. Lim, P. D. Zamore, J. R. Knowles, *Biochemistry* **1988**, *27*, 1158–1167.

[3] J. A. Gerlt, F. M. Raushel, *Curr. Opin. Chem. Biol.* **2003**, *7*, 252–264.

[4] a) B. Höcker, J. Claren, R. Sterner, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 16448–16453; b) S. Leopoldseder, J. Claren, C. Jürgens, R. Sterner, *J. Mol. Biol.* **2004**, *337*, 871–879; c) J. Claren, C. Malisi, B. Höcker, R. Sterner, *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 3704–3709; d) S. Evran, A. Telefoncu, R. Sterner, *Protein Eng. Des. Sel.* **2012**, *25*, 285–293.

[5] U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore, K. Robins, *Nature* **2012**, *485*, 185–194.

[6] a) H.-S. Park, S.-H. Nam, J. K. Lee, C. N. Yoon, B. Mannervik, S. J. Benkovic, H.-S. Kim, *Science* **2006**, *311*, 535–538; b) D. S. Tawfik, *Science* **2006**, *311*, 475–476; c) C. Heinis, K. Johnsson, *Methods Mol. Biol.* **2010**, *634*, 217–232.

[7] a) A. Ochoa-Leyva, F. Barona-Gómez, G. Saab-Rincón, K. Verdel-Aranda, F. Sánchez, X. Soberón, *J. Mol. Biol.* **2011**, *411*, 143–157; b) A. Ochoa-Leyva, X. Soberón, F. Sánchez, M. Argüello, G. Montero-Morán, G. Saab-Rincón, *J. Mol. Biol.* **2009**, *387*, 949–964; c) G. Saab-Rincón, L. Olvera, M. Olvera, E. Rudiño-Piñera, E. Benites, X. Soberón, E. Morett, *J. Mol. Biol.* **2012**, *416*, 255–270; d) H. Ma, T. M. Penning, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 11161–11166; e) E. Campbell, S. Chuang, S. Banta, *Protein Eng. Des. Sel.* **2013**, *26*, 181–186.

[8] a) A. D. Griffiths, D. S. Tawfik, *EMBO J.* **2003**, *22*, 24–35; b) J. E. Vick, D. M. Z. Schmidt, J. A. Gerlt, *Biochemistry* **2005**, *44*, 11722–11729; c) P.-C. Tsai, N. Fox, A. N. Bigley, S. P. Harvey, D. P. Barondeau, F. M. Raushel, *Biochemistry* **2012**, *51*, 6463–6475.

[9] a) L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas III, D. Hilvert, K. N. Houk, B. L. Stoddard, D. Baker, *Science* **2008**, *319*, 1387–1391; b) D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, D. Baker, *Nature* **2008**, *453*, 190–195; c) H. K. Privett, G. Kiss, T. M. Lee, R. Blomberg, R. A. Chica, L. M. Thomas, D. Hilvert, K. N. Houk, S. L. Mayo, *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 3790–3795.

[10] a) J. B. Siegel, A. Zanghellini, H. M. Lovick, G. Kiss, A. R. Lambert, J. L. St. Clair, J. L. Gallaher, D. Hilvert, M. H. Gelb, B. L. Stoddard, K. N. Houk, F. E. Michael, D. Baker, *Science* **2010**, *329*, 309–313; b) D. Baker, *Protein Sci.* **2010**, *19*, 1817–1819; c) M. V. Golynskiy, B. Seelig, *Trends Biotechnol.* **2010**, *28*, 340–345.

[11] a) N. Tokuriki, D. S. Tawfik, *Curr. Opin. Struct. Biol.* **2009**, *19*, 596–604; b) J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 5869–5874.

[12] K. B. Zeldovich, P. Chen, E. I. Shakhnovich, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 16152–16157.

[13] J. D. Bloom, J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami, F. H. Arnold, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 606–611.

[14] G. Cho, A. D. Keefe, R. Liu, D. S. Wilson, J. W. Szostak, *J. Mol. Biol.* **2000**, *297*, 309–319.

[15] a) V. Sieber, A. Plückthun, F. X. Schmid, *Nat. Biotechnol.* **1998**, *16*, 955–960; b) T. Matsuura, A. Plückthun, *Origins Life Evol. Biospheres* **2004**, *34*, 151–157; c) T. Matsuura, A. Plückthun, *FEBS Lett.* **2003**, *539*, 24–28; d) F.-X. Schmid, *ChemBioChem* **2011**, *12*, 1501–1507.

[16] O. Khersonsky, D. S. Tawfik, *Annu. Rev. Biochem.* **2010**, *79*, 471–505.

[17] E. Santelli, R. Schwarzenbacher, D. McMullan, T. Biorac, L. S. Brinen, J. M. Canaves, J. Cambell, X. Dai, A. M. Deacon, M.-A. Elsliger, S. Eshagi, R. Floyd, A. Godzik, C. Grittini, S. K. Grzechnik, L. Jaroszewski, C. Karlak, H. E. Klock, E. Koesema, J. S. Kovarik, et al., *Proteins* **2004**, *56*, 167–170.

[18] a) R. W. Roberts, J. W. Szostak, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 12297–12302; b) B. Seelig, *Nat. Protoc.* **2011**, *6*, 540–552.

[19] G. S. Cho, J. W. Szostak, *Chem. Biol.* **2006**, *13*, 139–147.

[20] S. Moelbert, E. Emberly, C. Tang, *Protein Sci.* **2004**, *13*, 752–762.

[21] a) T. Seitz, R. Thoma, G. A. Schoch, M. Stihle, J. Benz, B. D'Arcy, A. Wiget, A. Ruf, M. Hennig, R. Sterner, *J. Mol. Biol.* **2010**, *403*, 562–577; b) J. J. Graziano, W. Liu, R. Perera, B. H. Geierstanger, S. A. Lesley, P. G. Schultz, *J. Am. Chem. Soc.* **2008**, *130*, 176–185; c) J.-D. Pédelacq, E. Piltch, E. C. Liong, J. Berendzen, C.-Y. Kim, B.-S. Rho, M. S. Park, T. C. Terwilliger, G. S. Waldo, *Nat. Biotechnol.* **2002**, *20*, 927–932.

[22] a) D. Amar, I. Berger, N. Amara, G. Tafa, M. M. Meijler, A. Aharoni, *J. Mol. Biol.* **2012**, *416*, 21–32; b) S. D. Copley, *J. Biol. Chem.* **2012**, *287*, 3–10.

[23] L. Stryer, *J. Mol. Biol.* **1965**, *13*, 482–495.

[24] *Molecular Cloning: A Laboratory Manual*, 3rd ed. (Eds.: J. Sambrook, D. W. Russell), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, **2001**.

Received: May 22, 2013
Published online on August 16, 2013

CHEMBIOCHEM

Supporting Information

Highly Diverse Protein Library Based on the Ubiquitous$(\beta/\alpha)_8$ Enzyme Fold Yields Well-Structured Proteins through in Vitro Folding Selection

Misha V. Golynskiy, John C. Haugner, III, and Burckhard Seelig*[a]

cbic_201300326_sm_miscellaneous_information.pdf

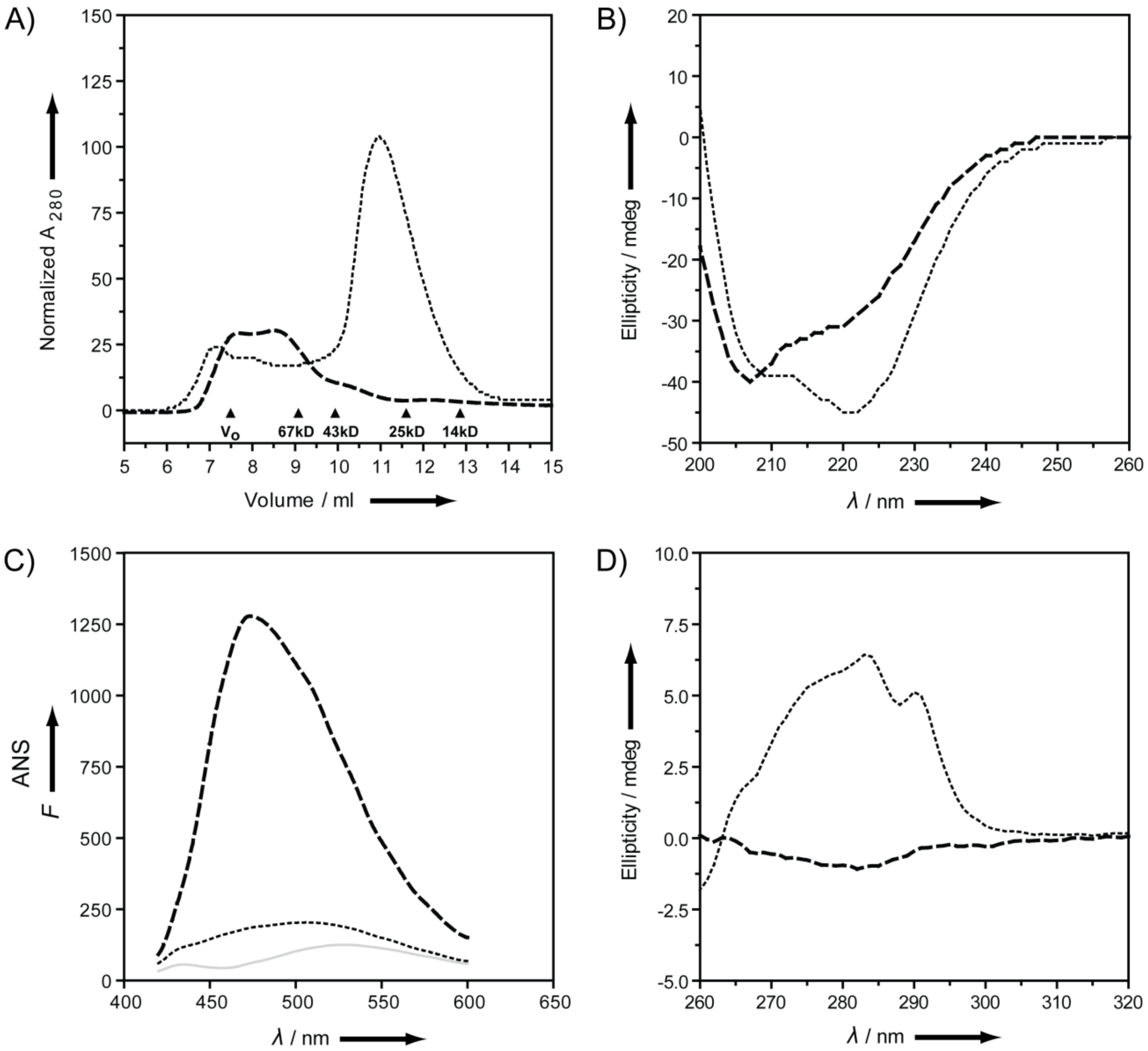

Figure S1. Biophysical characterization of GDPDwt (black dotted line) and GDPDmut (black dashed line) proteins. A) Investigation of quaternary structure by size exclusion chromatography. Monomeric GDPDwt elutes at 11 ml while oligomeric GDPDmut elutes at 8 ml. B) Far-UV circular dichroism spectroscopy to study the secondary structure. GDPDwt exhibits characteristic alpha-helical peaks at 208 nm and 222 nm while GDPDmut has lost some secondary structure, as indicated by signal decrease in 222 nm. C) 1-Anilinonaphthalene-8-sulfonic acid (ANS) fluorescence measurements (tertiary structure). ANS fluorescence is known to increase upon interaction with exposed hydrophobic protein patches, suggesting increased presence of exposed hydrophobic patches in GDPDmut, associated with the loss of the native GDPDwt tertiary structure. Buffer is shown as solid light gray line. D) Near-UV circular dichroism spectroscopy reveals that GDPDmut has significantly less tertiary structure relative to GDPDwt.

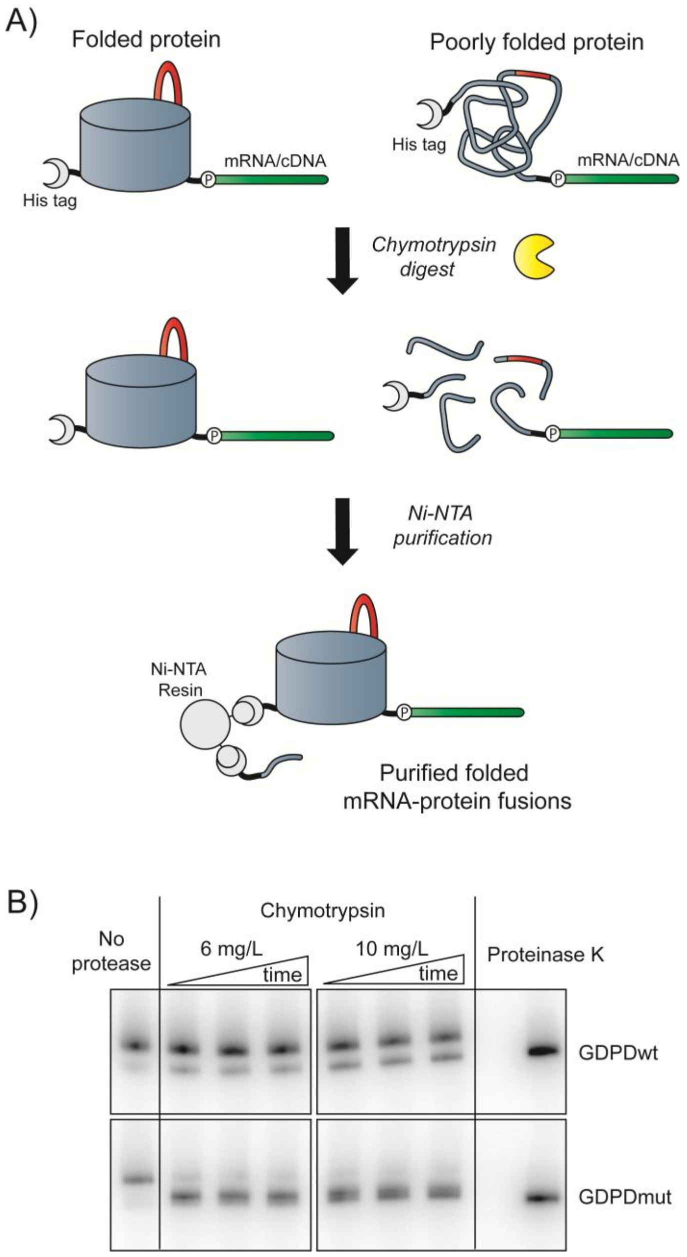

Figure S2. Folding selection by *in vitro* protease digestion. A) Schematic of the protease-digestion based selection by mRNA display. A mixture of folded and unfolded proteins that are covalently linked to their encoding mRNA/cDNA hybrid via puromycin (P) is subjected to chymotrypsin digest and then purified under denaturing conditions via Ni-NTA affinity chromatography. Only the cDNA of the well folded proteins is immobilized on the Ni-NTA resin, and amplified by PCR for downstream applications. In contrast, cleavage of unfolded proteins severs the link between His_6-tag and cDNA, thereby preventing immobilization of the cDNA. B) Analysis of chymotrypsin digestion during initial optimization steps. mRNA-displayed GDPDwt and GDPDmut proteins were incubated with chymotrypsin for 10, 15 and 20 min prior to analysis of crude digest reactions by SDS-PAGE. Undigested mRNA-protein fusions and fusions treated with proteinase K (to degrade proteins non-specifically) were used as controls for 0% and 100% digestion, respectively. Under these conditions GDPDmut is preferentially digested by chymotrypsin. Final optimization steps included the additional use of detergents and an analysis of digestion based on comparison of His_6-tag purified digested and undigested samples via scintillation counting.

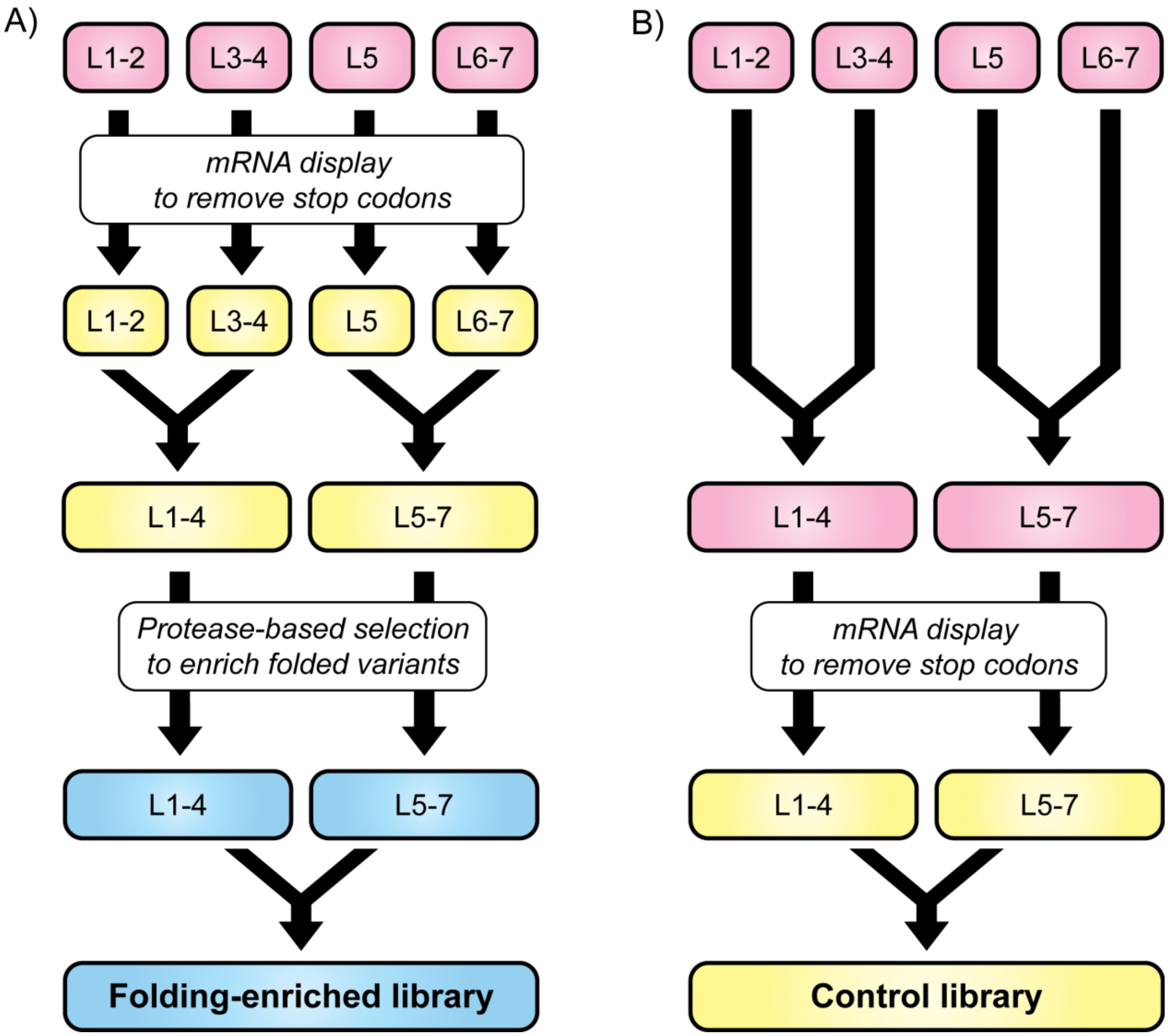

Figure S3. Step-wise strategy for the construction of libraries based on the (β/α)$_8$ fold. Each box represents an individual library with randomized loop(s) (e.g. library L1-2). Libraries containing only a single randomized loop are not shown for clarity. The libraries were mRNA-displayed to remove stop codons or, in addition, were subjected to a folding selection by *in vitro* protease digestion (shown in yellow and blue, respectively). A) Construction of the folding-enriched library. Full length (β/α)$_8$ barrel libraries were used in mRNA display and folding selection. B) Construction of the control library. (β/α)$_4$ half-gene fragments of the barrel libraries were used in mRNA display to prevent the native (β/α)$_8$ fold from biasing the randomized regions in the control library.

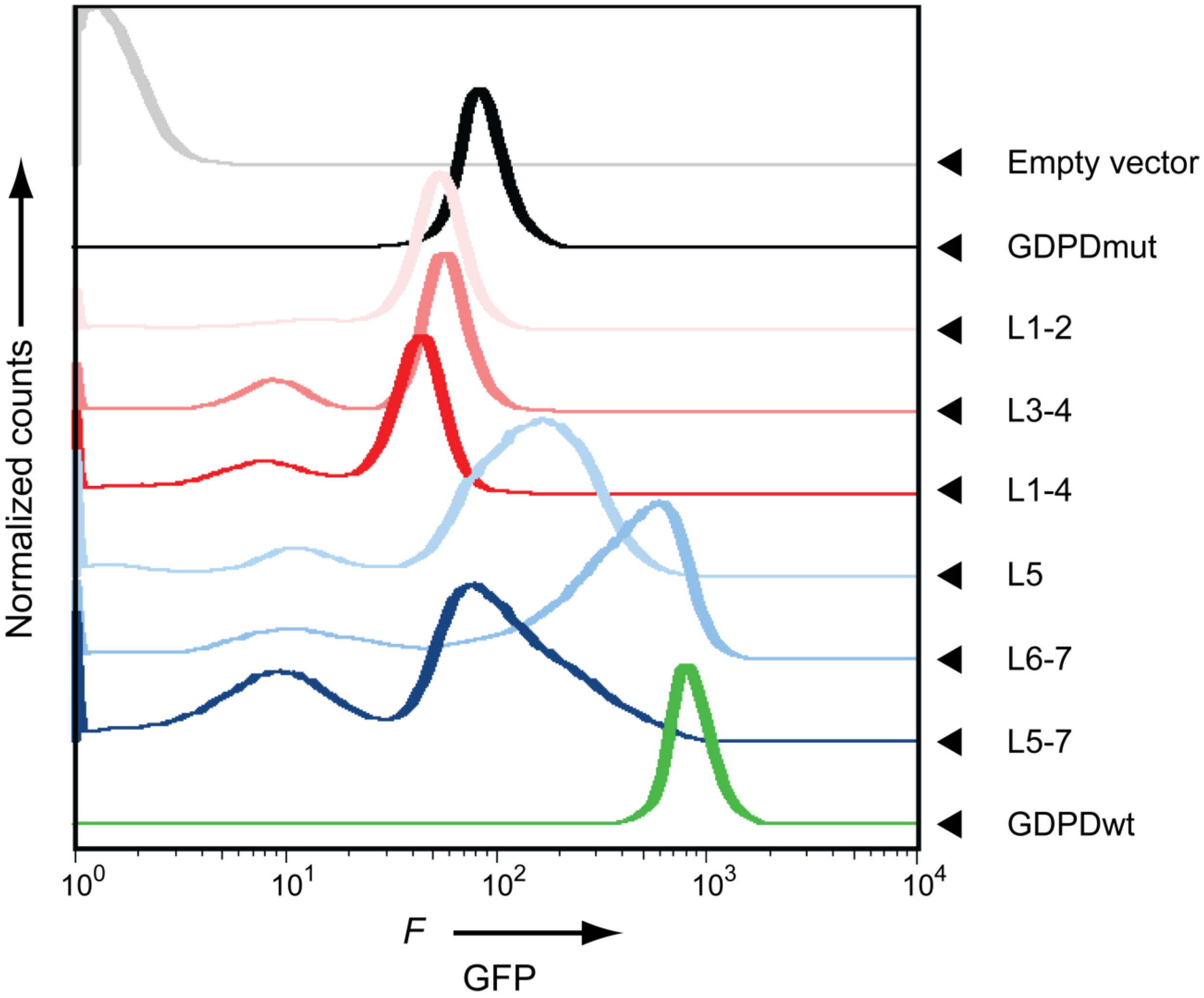

Figure S4. Assessment of folding by GFP-fusion assay. Fluorescence histograms of *E. coli* BL21(DE3) cells expressing the GFP-fused control constructs and intermediate libraries. The empty vector population was gated out on the histograms of cells transformed with the GDPD constructs. None of the libraries shown here had been selected yet for folding.

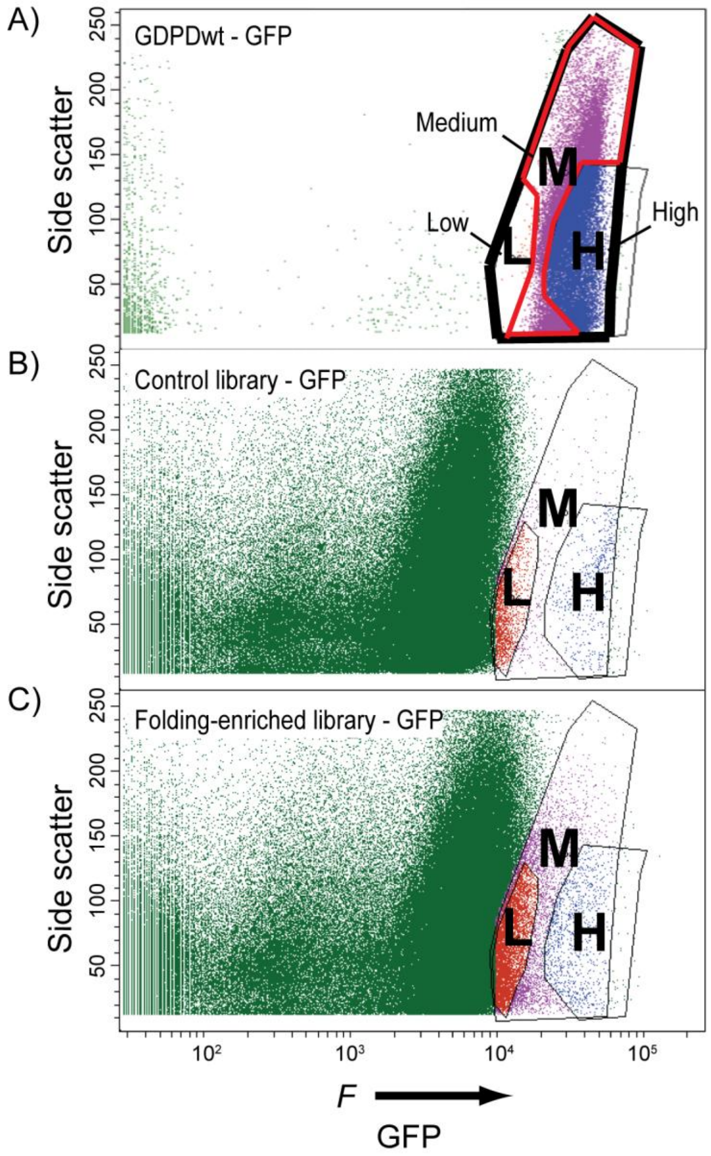

Figure S5. Analysis of library populations by fluorescence-activated cell sorting experiments (FACS). Side scatter versus GFP fluorescence dot plots are shown for A) GDPDwt-GFP, B) control library and C) folding-enriched library. In the top panel, the GDPDwt-like population of cells is framed with a thick black line. The low (L) and high (H) populations of library-GFP cells shown in the lower two panels were collected during the sorting experiment for further analysis. The number of cells in the medium (M) population was estimated (see Table S2).

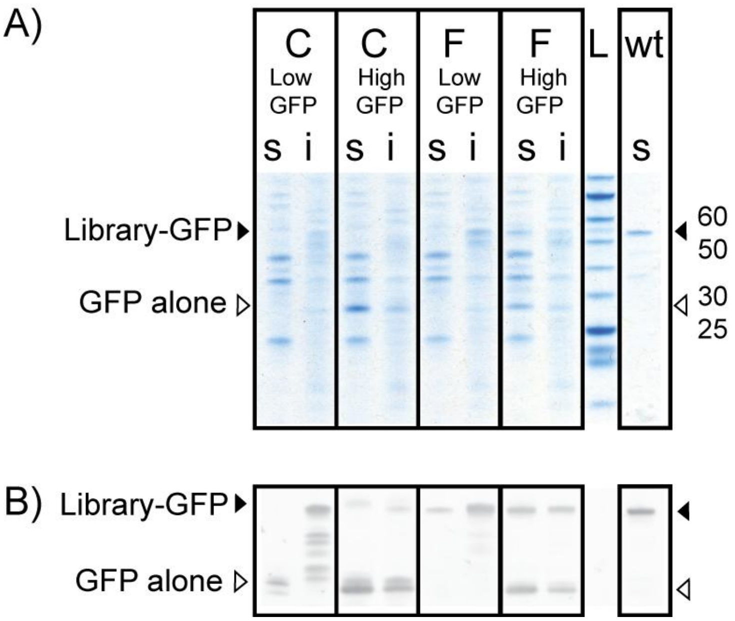

Figure S6. Comparison of the soluble and insoluble fractions of the FACS sorted library populations. A) SDS-PAGE gel Coomassie-stained. B) Western blot of SDS-PAGE gel using polyclonal anti-GFP antibody. Approximate molecular weights for library-GFP fusions (57 kD, black triangle) and for GFP alone (27 kD, white triangle). Equal amounts of soluble (s) and insoluble (i) fractions were loaded for each library. The soluble fraction of over-expressed GDPDwt-GFP fusion is shown as positive control.

Table S1. GFP-fused *in vivo* folding assessment of intermediate libraries. [a]

Species	Mode of GFP fluorescence [b]	% of cells with GDPDwt-GFP fluorescence [c]
GDPDmut	10.4	0.02
L1-2	6.7	0.90
L3-4	7.2	0.05
L1-4	5.4	0.09
L5	21.3	9.3
L6-7	72.3	51.6
L5-7	10	6.9
GDPDwt	100	98.5

[a] Constructs were transformed into *E.coli* BL21(DE3) strain. Stop-codon containing libraries that have not been subjected to mRNA display. Prior to analysis all data were gated to exclude cell populations that matched fluorescence and scatter profiles of cells transformed with empty plasmid.
[b] Values normalized to the mode GFP fluorescence of GDPDwt-GFP.
[c] Wild type cells were gated on the forward scatter versus GFP contour plot to include ~98% of all wild type cells.

Table S2. Fraction of soluble, GDPDwt-like, library-GFP fusions in the FACS-sorted populations.

		% soluble GFP fusions		
		of sorted cells		of all cells [e]
Library population [a], [b]	% Non-empty cells [c]	By Western blot [d]	By SDS-PAGE	
Control, low GFP	0.86%	0.24%	<5% (0/18 clones)	0.002%
Control, medium GFP	0.23% [f]	0.24% - 5.76% [g]	not available	0.001% - 0.013%
Control, high GFP	0.31%	5.76%	<5% (0/19 clones)	0.018%
Estimated % soluble proteins in control library				0.020% - 0.033% [h]
Folding-enriched, low GFP	3.18%	18.20%	29% (4/14 clones)	0.579%
Folding enriched, mid GFP	1.68% [f]	18.20% - 26.70%[g]	not available	0.306% - 0.449%
Folding-enriched, high GFP	0.54%	26.70%	24.6% (20/81 clones)	0.144%
Estimated % soluble proteins in folding-enriched library				1.029% - 1.172% [h]
Enrichment of soluble proteins after folding selection [i]				35 to 50-fold

[a] Library populations fell into the GDPDwt-like profile window defined by side scatter vs. GFP fluorescence dot plot of GDPDwt-GFP construct (Figure S5 A).
[b] Only the low and high GFP populations were sorted during the FACS experiment.
[c] % Non-empty cells defined as the ratio (# of cells analyzed - # of cells with empty-vector) / (# of cells analyzed). A total of 1.4% of the control library and 5.4% of the folding-enriched library fell into the GDPDwt-like window used for sorting and analysis.
[d] % Soluble GFP fusions of sorted cells calculated from Western blot analysis (Figure S6) using Image J to quantitate the intensities of all anti-GFP stained bands. Defined as the ratio (intensity of soluble GFP-fusions) / Σ (intensity of all anti-GFP stained bands).
[e] % Soluble GFP fusions of all cells calculated as (% non-empty cells x % soluble GFP fusions of sorted cells estimated from Wester blot) for the population of interest.
[f] % Non-empty cells for the unsorted medium GFP population was calculated as the difference in % non-empty cells populations (GDPDwt-like – low GFP – high GFP).
[g] Values are lower and upper estimates based on % soluble GFP fusions of sorted cells in the low and high GFP populations.
[h] Estimated % soluble proteins in a library calculated as Σ (% soluble GFP fusions of all cells) for low, medium and high GFP populations.
[i] Fold improvement defined as the ratio (estimated % soluble proteins in the folding-enriched library)/ (estimated % soluble proteins in the control library).

Table S3. List of primers used during library construction.

Primer	Sequence
041B [a]	GCCTTCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGGGCAGCGATAAGATCCACC
042 [a]	TTAATAGCCGGTGCCAGATCCAGACATTCC
018	TATGACCAAGCTTCCAGGGTGTTTTCCAGATACTTGGCGGAGTAACCSNNSNNGCCCAGCACAATCAC
019	CTGGAAAACACCCTGGAAG
047	AACAACAAGGTCTCAGGCGCTTTAAATCSNNSNNSNNSNNSNNGCTCACGACCACCTTGCC
048	AACAACAAGGTCTCGCGCCTGTTCGGTCTGGACG
028	AACAACAAGGTCTCCCTCACGTTCSNNSNNSNNSNNGATTTCGATGTTGATGATCTTG
029	AACAACAAGGTCTCGTGAGGCCGCGGACGCAGTGCTGGAGATCAGCAAAAAGCGTAAG
022	TGGAGGTCCTTGGTACCCTTGAATTTTTCATCCAGCAGGTCCAGATCSNNSNNSNNSNNGGAGCTGAAAATCAGGTTCTTAC
023	GATCTGGACCTGCTGGATG
012	AGTAGAGGTACCAAATACGGTTATNNSATCNNSNNSNNSNNSNNSTACGGTTCCATTGAAAATTTCG
015	CTTCGTCGATCAGATAACCG
003	AGTAAAGAGCTCAAAGGCCTGSNNSNNSNNSNNSNNSNNCACGTGCAGAGAGTACGGAC
017	AGGCCCTATCAGGCCTTTGAG
013	AGTAGATACGTATTTTGCGGTAGATTTCCGGATCSNNSNNSNNSNNSNNCACAAAAATCACGATGCC
016	CTGAAAGAGCTGACCGATGG
039	TTAATAGCCGGTGCCAGATCCAGACATTCCCATTTTGTCATCGTCATCCTTATAGTCGGAGCCACCGGTCTCACCCTTGAATTTTTCATCCAGC
040B	TTCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGCACCATCACCATCACCATGGTCTCAAGGGTACCAAATACGGTTAT
035B	GACTGAACTGATGCCTATATGCTTGTCCGTCCAGATTGGTCTCGTGTTTGAGAACGTGTCCGATG
036B	AACTCAAGTATCGCTATGCCGGTCAATAACCGAGGTCTCAAACACTTCCTTCAGGGTGG
049	CTGCGTGGTAGCTCGTAGCACAGACTCAGCGGATACACACAGAGGTCTCAAAAATTCAAGGGTACCAAATACGG
050	GCACTCCGCTTAGATAGATAGCCAGAAGACAGACAAGGTCTCATTTTTCATCCAGCAGGTCCAG
037B	AAGTAGCATAGAGTGTCGCTCTGGATGTCAAGGTCTCAAAAGGAGCGTCCGTACTCTCTG
038B	ATGATAGCAGATGGACTTAGATTTCCGGTCAGGTGCCGAGGTCTCCCTTTTCCACGCGCTCCACG
GDPDx_001 [b]	TCTGTAAACCATGGATGGGCAGCGATAAGATCCAC
GDPDx_002 [b]	CTGTGCGCTCGAGTTAATAGCCGGTGCCAGATCC
GDPDx_003 [c]	GAAGGAGATATACATATGGGCAGCGATAAGATC
GDPDx_004 [c]	GGAGCCAGCGCGGCCGCCATAGCCGGTGCCAGATCCAG
GDPDmut_Fw [d]	GAAGCCGGCGCGAATCGTGAGGAGCTGGATGTG
GDPDmut_Rev [d]	CACATCCAGCTCCTCACGATTCGCGCCGGCTTC
030 [e]	TTCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGGGCAGCGATAAGATCGTGATTGTGCTGGGCCATCGCGG

[a] Primers used as standard primers to amplify any full length GDPD-based template during the library construction.
[b] Primers used to amplify template DNA for insertion into the pET28a plasmid for protein expression.
[c] Primers used to amplify template DNA for insertion into the pER13 plasmid for the GFP-fusion assay.
[d] Primers used to generate pET28/GDPDmut from the pET28/GDPDwt template.
[e] Primers 030 and 042 were used to generate GDPDmut(-His$_6$) from the pET28/GDPDmut template.

Table S4. Fragments used in library assembly. [a], [b]

	5'-Fragment				3'-Fragment				
Library [d]	Name	Template	FW	REV	Name	Template	FW	REV	Restriction enzyme used [c]
L1	L1_A	pET28/GDPDwt	041B	018	L1_B	pET28/GDPDwt	019	042	*HindIII*
L2	L2_A	pET28/GDPDwt	041B	047	L2_B	pET28/GDPDwt	048	042	*BsaI*
L3	L3_A	pET28/GDPDwt	041B	028	L3_B	pET28/GDPDwt	029	042	*BsaI*
L3(-His$_6$)	L3_A(-His$_6$)	pET28/GDPDwt	030	028	L3_B	pET28/GDPDwt	029	042	*BsaI*
L4	L4_A	pET28/GDPDwt	041B	022	L4_B	pET28/GDPDwt	023	042	*KpnI*
L5	L5_A	pET28/GDPDwt	041B	015	L5_B	pET28/GDPDwt	012	042	*KpnI*
L6	L6_A	pET28/GDPDwt	041B	003	L6_B	pET28/GDPDwt	017	042	*SacI*
L7	L7_A	pET28/GDPDwt	041B	013	L7_B	pET28/GDPDwt	016	042	*SnaBI*
L1-2 [e]	L1-2_A	L1	041B	047	L2_B	pET28/GDPDwt	048	042	*BsaI*
L3-4 [e]	L3-4_A	pET28/GDPDwt	041B	028	L3-4_B	L4	029	042	*BsaI*
L6-7 [e]	L6-7_A	L6	041B	034	L6-7_B	L7	033	042	*BsaI*
L1-4 [f]	C-L1-4_A	L1-2	041B	036B	C-L1-4_B	L3-4	035B	042	*BsaI*
L5-7 [g]	C-L5-7_A	L5	041B	038B	C-L5-7_B	L6-7	037B	042	*BsaI*
L1-4 (m)	F-L1-4_A	L1-2 (m)	041B	036B	F-L1-4_B	L3-4 (m)	035B	042	*BsaI*
L5-7 (m)	F-L5-7_A	L5 (m)	041B	038B	F-L5-7_B	L6-7 (m)	037B	042	*BsaI*
Control library	C-frag_A	L1-4 frag (m)	041B	050	C-frag_B	L5-7 frag (m)	049	042	*BsaI*
Folding-enriched library	F-frag_A	L1-4 (m&p)	041B	050	F-frag_B	L5-7 (m&p)	049	042	*BsaI*

[a] Individual libraries were generated by amplifying the 5'- and 3'-fragments of the library using the specified template, and FW /REV primer pair.
[b] Libraries denoted with (m) have been subjected to mRNA display; libraries denoted with (p) have been subjected to *in vitro* folding selection by protease digestion.
[c] 5'- and 3'-fragments were digested with the indicated restriction enzyme (RE), gel-purified and ligated to produce the individual libraries. The *BsaI* restriction site, which is absent in the parent GDPDwt scaffold, was introduced into the 5-' and 3' fragments by PCR amplification and was removed again during the gel purification of digested fragments.
[d] Individual libraries were gel purified after the ligation reaction. A fraction of the purified library was used as a template for the successive steps in library assembly.
[e] Libraries were subjected to mRNA display during folding-enriched library construction and then PCR amplified with 041B/042 primer pair to restore T7 promoter sequence lost during the transcription process.
[f] Library was used as template with 041B/039 primer pair to generate the control library fragment (L1-4 frag) subjected to mRNA display.
[g] Library was used as template with 040B/042 primer pair to generate the control library fragment (L5-7 frag) subjected to mRNA display.

.

Supplementary Data

DNA sequence of the GDPDwt scaffold used as template for library assembly

```
GCCTTCTAATACGACTCACTATAGGGACAATTACTATTTACAATTACAATGGGCAGCGAT
AAGATCCACCATCACCATCACCATGTGATTGTGCTGGGCCATCGCGGTTACTCCGCCAAG
TATCTGGAAAACACCCTGGAAGCTTTCATGAAAGCGATCGAAGCCGGCGCGAATGGTGTG
GAGCTGGATGTGCGCCTGTCTAAAGACGGCAAGGTGGTCGTGAGCCATGATGAAGATTTA
AAGCGCCTGTTCGGTCTGGACGTCAAAATCCGTGACGCCACCGTGTCTGAACTGAAAGAG
CTGACCGATGGCAAAATTACCACCCTGAAGGAAGTGTTTGAGAACGTGTCCGATGACAAG
ATCATCAACATCGAAATCAAGGAACGTGAGGCCGCGGACGCAGTGCTGGAGATCAGCAAA
AAGCGTAAGAACCTGATTTTCAGCTCCTTTGATCTGGACCTGCTGGATGAAAAATTCAAG
GGTACCAAATACGGTTATCTGATCGACGAAGAGAACTACGGTTCCATTGAAAATTTCGTG
GAGCGCGTGGAAAAGGAGCGTCCGTACTCTCTGCACGTGCCCTATCAGGCCTTTGAGCTC
GAATATGCGGTGGAGGTGCTGCGCTCCTTCCGTAAAAAGGGCATCGTGATTTTTGTGTGG
ACCCTGAATGATCCGGAAATCTACCGCAAAATACGTAGAGAGATCGATGGTGTGATTACC
GACGAAGTGGAGCTGTTTGTGAAACTGCGTGGCGGCAGCGGTGGCTCCGACTATAAGGAT
GACGATGACAAAATGGGAATGTCTGGATCTGGCACCGGCTATTAA
```

Color code:
T7 transcription promoter / TMV translation enhancer
Thio$_6$His$_6$
GDPDwt
(GGS)$_2$ spacer
FLAG tag
Puromycin-crosslinking region

Primers 041B and 042 were used as standard primers to amplify any full length GDPD-based template during the library construction. Primers GDPDx_001 and GDPDx_002 were used to amplify template DNA for insertion into the pET28a plasmid for protein expression. Primers GDPDx_003 and GDPDx_004 were used to amplify template DNA for insertion into the pER13 plasmid for the GFP-fusion assay. Primers GDPDmut_Fw and GDPDmut_Rev were used to generate pET28/GDPDmut from the pET28/GDPDwt template. Primers 030 and 042 were used to generate GDPDmut(-His$_6$) from the pET28/GDPDmut template.

Sequence alignment of the six soluble F(s) clones characterized in this manuscript (Figure 4).

Alignment was performed using Clustal Omega server, Clustal O(1.1.0)

Loop residues randomized during library construction are highlighted in green.

* = column contains identical amino acid

: = column contains different but highly conserved amino acids

```
GDPDwt          MGSDKIHHHHHHVIVLGHRGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)1           MGSDKIHHHHHHVIVLGSRGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)2           MGSDKIHHHHHHVIVLGHRGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)3           MGSDKIHHHHHHVIVLGRLGYSAKYLENTLEAFMRAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)4           MGSDKIHHHHHHVIVLGNGGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)5           MGSDKIHHHHHHVIVLGRVGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
F(s)6           MGSDKIHHHHHHVIVLGRLGYSAKYLENTLEAFMKAIEAGANGVELDVRLSKDGKVVVSHDEDLKRLFG
                *****************  ***************:**********************************

GDPDwt          LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)1           LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKTINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)2           LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)3           LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)4           LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)5           LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
F(s)6           LDVKIRDATVSELKELTDGKITTLKEVFENVSDDKIINIEIKEREAADAVLEISKKRKNLIFSSFDLDL
                *********************************** *********************************

GDPDwt          LDEKFKGTKYGYLIDE-ENYGSIENFVERVEKERPYSLHVP----YQAFELEYAVEVLRSFRKKGIVIF
F(s)1           LDEKFKGTKYGYLIDE-ENYGSIENFVERVEKERPYSLHVTPTLLSQAFELEYAVEVLRSFRKKGIVIF
F(s)2           LDEKFKGTKYGYKISLWASYGSIENFVERVEKERPYSLHVSSTKDAQAFELEYAVEVLRSFRKKGIVIF
F(s)3           LDEKFKGTKYGYKIGRGGGYGSIENFVERVEKERPYSLHVYSGSPLQAFELEYAVEVLRSFRKKGIVIF
F(s)4           LDEKFKGTKYGYIISLKDTYGSIENFVERVEKERPYSLHVQRASFKQAFELEYAVEVLRSLRKKGIVIF
F(s)5           LDEKFKGTKYGYGIAEGLVYGSIENFVERVEKERPYSLHVELEFMIQAFELEYAVEVLRSFRKKGIVIF
F(s)6           LDEKFKGTKYGYLIDE-ENYGSIENFVERVEKERPYSLHVAVGRVLQAFELEYAVEVLRSFRKKGIVIF
                ************ *     *********************      **************:********

GDPDwt          VWT-LNDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)1           VKNNVCDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)2           VPCLRCDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)3           VASSTHDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)4           VAPDLPDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)5           VRADMSDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
F(s)6           VTSVTRDPEIYRKIRREIDGVITDEVELFVKLRGGSGGSDYKDDDDKMGMSGSGTGY
                *     ***************************************************
```