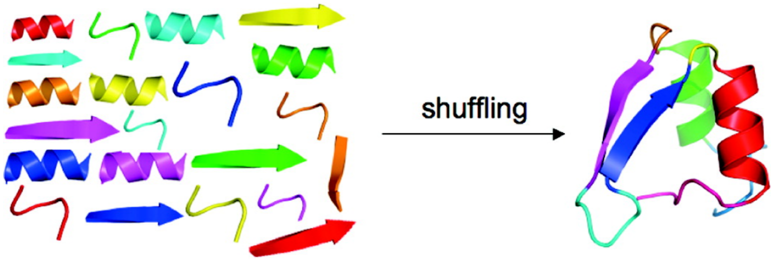
Article

Selecting Folded Proteins from a Library of Secondary Structural Elements

James J. Graziano, Wenshe Liu, Roshan Perera, Bernhard H. Geierstanger, Scott A. Lesley, and Peter G. Schultz

J. Am. Chem. Soc., **2008**, 130 (1), 176-185 • DOI: 10.1021/ja074405w



Journal of the American Chemical Society is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

J|A|C|S
ARTICLES
Published on Web 12/08/2007

Selecting Folded Proteins from a Library of Secondary Structural Elements

James J. Graziano,† Wenshe Liu,† Roshan Perera,† Bernhard H. Geierstanger,*,‡ Scott A. Lesley,*,†,‡ and Peter G. Schultz*,†,‡

Department of Chemistry and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037 and Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Drive, San Diego, California 92121

Received June 17, 2007; E-mail: schultz@scripps.edu; slesley@gnf.org; bgeierstanger@gnf.org

Abstract: A protein evolution strategy is described by which double-stranded DNA fragments encoding defined *Escherichia coli* protein secondary structural elements (α-helices, β-strands, and loops) are assembled semirandomly into sequences comprised of as many as 800 amino acid residues. A library of novel polypeptides generated from this system was inserted into an enhanced green fluorescent protein (EGFP) fusion vector. Library members were screened by fluorescence activated cell sorting (FACS) to identify those polypeptides that fold into soluble, stable structures *in vivo* that comprised a subset of shorter sequences (∼60 to 100 residues) from the semirandom sequence library. Approximately 10^8 clones were screened by FACS, a set of 1149 high fluorescence colonies were characterized by dPCR, and four soluble clones with varying amounts of secondary structure were identified. One of these is highly homologous to a domain of aspartate racemase from a marine bacterium (*Polaromonas sp.*) but is not homologous to any *E. coli* protein sequence. Several other selected polypeptides have no global sequence homology to any known protein but show significant α-helical content, limited dispersion in 1D nuclear magnetic resonance spectra, pH sensitive ANS binding and reversible folding into soluble structures. These results demonstrate that this strategy can generate novel polypeptide sequences containing secondary structure.

Introduction

Despite the large sequence diversity present in the proteomes of known organisms, most functional proteins characterized to date assume one of relatively few distinct folds,[1−5] suggesting that nature uses a limited number of stable, soluble folds relative to what is theoretically possible in protein sequence space. These folds likely represent the divergent products of a limited set of ancient protein folds. However, protein sequences that fold into stable, unique topologies but are not encoded by any sequenced genome may also exist. It may be possible to identify other stable folds that simply have not yet been sampled in the course of evolution by means of an artificial selection and molecular evolution process. Methods such as DNA shuffling[6] that mimic the combinatorial diversity mechanism of the immune system are among the most efficient methods to modify or enhance protein activity. Unfortunately, the structural diversity generated in a library of shuffled homologues is relatively small,[7−11] although newer methods for random fragment assembly that overcome sequence homology requirements may lead to libraries with increased structural diversity.[12,13] Alternatively, natural or *in vitro* combinatorial assembly of distinct protein subunits (e.g., subdomains, exons, etc.) can create significant structural and functional diversity.[14−19] For example, the mammalian blood clotting proteins plasminogen, protein C, urokinase, and prothrombin are all derived from different combinations of 5 exons.[14] Sequence and structural studies also suggest that many existing TIM barrel proteins evolved from the combinatorial assembly of subunits from more primitive 8-fold βα barrels.[15,16] Other approaches for creating libraries of novel sequences that

† The Scripps Research Institute.
‡ Genomics Institute of the Novartis Research Foundation.

(1) Greene, L. H.; Lewis, T. E.; Addou, S.; Cuff, A.; Dallman, T.; Dibley, M.; Redfern, O.; Pearl, F.; Nambudiry, R.; Reid, A.; Sillitoe, I.; Yeats, C.; Thornton, J. M.; Orengo, C. A. *Nucleic Acids Res.* **2007**, *35*, D291−D297.
(2) Orengo, C. A.; Sillitoe, I.; Reeves, G.; Pearl, F. M. *J. Struct. Biol.* **2001**, *134*, 145−165.
(3) Thornton, J. M.; Orengo, C. A.; Todd, A. E.; Pearl, F. M. *J. Mol. Biol.* **1999**, *293*, 333−342.
(4) Wolf, Y. I.; Brenner, S. E.; Bash, P. A.; Koonin, E. V. *Genome Res.* **1999**, *9*, 17−26.
(5) Zeldovich, K. B.; Berezovsky, I. N.; Shakhnovich, E. I. *J. Mol. Biol.* **2006**, *357*, 1335−1343.
(6) Stemmer, W. P. *Nature* **1994**, *370*, 389−391.
(7) Crameri, A.; Raillard, S. A.; Bermudez, E.; Stemmer, W. P. *Nature* **1998**, *391*, 288−291.
(8) Crameri, A.; Whitehorn, E. A.; Tate, E.; Stemmer, W. P. *Nat. Biotechnol.* **1996**, *14*, 315−319.
(9) Kikuchi, M.; Harayama, S. *Methods Mol. Biol.* **2002**, *182*, 243−257.
(10) Kurtzman, A. L.; Govindarajan, S.; Vahle, K.; Jones, J. T.; Heinrichs, V.; Patten, P. A. *Curr. Opin. Biotechnol.* **2001**, *12*, 361−370.
(11) Zhao, H.; Arnold, F. H. *Nucleic Acids Res.* **1997**, *25*, 1307−1308.
(12) Bittker, J. A.; Le, B. V.; Liu, D. R. *Nat. Biotechnol.* **2002**, *20*, 1024−1029.
(13) Bittker, J. A.; Le, B. V.; Liu, J. M.; Liu, D. R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 7011−7016.
(14) Harayama, S. *Trends Biotechnol.* **1998**, *16*, 76−82.
(15) Lang, D.; Thoma, R.; Henn-Sax, M.; Sterner, R.; Wilmanns, M. *Science* **2000**, *289*, 1546−1550.
(16) Nagano, N.; Orengo, C. A.; Thornton, J. M. *J. Mol. Biol.* **2002**, *321*, 741−765.
(17) Davidson, A. R.; Sauer, R. T. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 2146−2150.
(18) Reidhaar-Olson, J. F.; Sauer, R. T. *Science* **1988**, *241*, 53−57.
(19) Matsuura, T.; Ernst, A.; Pluckthun, A. *Protein Sci.* **2002**, *11*, 2631−2643.

10.1021/ja074405w CCC: $40.75

Table 1. 5′ and 3′ Oligonucleotide Linker Sequences for Fragment Library

	linker name	linker sequence
5′	Helix	GAGCTCGAAGACAGCGGCCGG
5′	Loop H_	GAGCTCGAAGACCGGCCGCT
5′	Loop S_	GAGCTCGAAGACTGGTGATC
5′	Strand	GAGCTCGAAGACTGGCTAAC
5′	Terminator Helix	GAGCTCGTCGACCCATGGATCTGATAAAATTCTTCCTTGATGTCTTCGAGCTC
5′	Terminator Loop H_	GAGCTCGTCGACCCATGGATCTGATAAAATTCTTCGCCGATGTCTTCGAGCTC
5′	Terminator Loop S_	GAGCTCGTCGACCCATGGATCTGATAAAATTCTTCGTGAATGTCTTCGAGCTC
5′	Terminator Strand	GAGCTCGTCGACCCATGGATCTGATAAAATTCTTCGCTAATGTCTTCGAGCTC
3′	Helix	GAGCTCGAAGACAGCGGCCGG
3′	Loop _H	GAGCTCGAAGACCTCAAGGAG
3′	Loop _S	GAGCTCGAAGACGTTAGCCAG
3′	Strand	GAGCTCGAAGACGATCACCAG
3′	Terminator Helix	GAGCTCGAAGACCGGCCGCTGCCGGCCGATCGGGATCCGAGCTC
3′	Terminator Loop _H	GAGCTCGAAGACCGCTTGCTGCCGGCCGATCGGGATCCGAGCTC
3′	Terminator Loop _S	GAGCTCGAAGACCGGCTACTGCCGGCCGATCGGGATCCGAGCTC
3′	Terminator Strand	GAGCTCGAAGACCGGTGACTGCCGGCCGATCGGGATCCGAGCTC

fold into native-like protein structures include the use of binary patterned residues designed around an existing known fold and the random assembly of cassettes composed of as few as three amino acid residues.[17,18] A more recent report details the use of cassettes of binary patterned residues around known protein motifs that are ultimately randomly assembled without fitting any designed folding constraint.[19] However, these latter methods have not yet yielded stable, novel folds under physiological conditions.[20]

Despite the considerable effort to identify or evolve new folds, only a limited number of distinct protein folds have been identified to date.[18] Here we describe a system for the synthesis of polypeptides with potentially novel folds in which existing sequence and structural data from bacterial proteins are used to assemble a pool of secondary structural elements (α-helices, β-strands, and loops). These secondary structural elements are in turn recombined into a library of *de novo* sequences that may be screened or selected for folded proteins. Proteins that fold into a native state acquire measurable biophysical characteristics such as solubility, stable secondary structure, well-dispersed 1D nuclear magnetic resonance (NMR) spectra indicative of packed side chains, and reversible two-state unfolding behavior. In this work, we chose to initially screen the library for polypeptides that are soluble in aqueous solution using a GFP fusion protein reporter that had been previously reported.[21] Indeed, several sequences with stable secondary structure are identified.

Materials and Methods

Library Construction. All secondary structural elements in the shuffling pool were selected from unique *Escherichia coli* proteins with structures on file in the Protein Data Bank (PDB).[22] PDB files of these proteins were submitted to the program PROMOTIF,[23] and the output files annotating the elements of secondary structure and the primary sequence of each element, as well as loop primary sequence data from the Sloop database,[24] were submitted to DeCypher FrameSearch BLAST (Active Motif, Inc.) to obtain the *E. coli* nucleotide sequences encoding each element. Elements comprising fewer than 5 residues were not considered.

Oligonucleotide primers for each secondary structural element (α-helix, β-strand, loop) or chain terminator are comprised of an element or terminator specific sequence and a sequence that can hybridize to an intervening linker. The linkers for the different secondary structure elements are shown in Table 1.

Double-strand (ds) library elements were produced by either polymerase chain reaction (PCR) amplification from *E. coli* genomic DNA or by single-strand (ss) oligonucleotide hybridization and extension reactions with Taq polymerase according to the following cycling parameters: 95 °C for 3 min then 30 cycles of 95 °C for 30 s, 56 °C for 45 s, 72 °C for 1 min, followed by a final extension at 72 °C for 5 min. The products of these reactions were inserted into pBAD-Thio (Invitrogen). DNA fragments encoding the secondary structural elements and the 5′ and 3′ terminators were isolated from plasmid DNA by restriction digest with Bbs I alone (secondary structural elements) or Bbs I and AsiS I (terminators) followed by purification on a 2.5% agarose gel. The purified fragments were combined into three pools and normalized to 250 μg/mL as follows: Pool 1: Helix/Loop/Strand fragments in 1:1:1 ratio; Pool 2: 5′ terminators; and Pool 3: 3′ terminators. Ligation of the fragments into a library of oligomers was accomplished with a 100000:100000:1 ratio of Pools 1, 2, and 3, respectively, in 20 μL reactions with T4 ligase in buffer containing 15% w/v PEG 6000 and 25 μM ATP, at 20 °C for 4 h. Ligated products were amplified by PCR with primers specific for the invariant sequences of the 5′ and 3′ terminators (fwd: GAGCTCGTCGACCCATGGG, rev: GAGCTCGGATCCCGATCGG). The resulting fragments (between 400 bp and 3000 bp in size) were isolated and purified from a 1% agarose gel and were inserted into the screening vector JG1 (see below). Competent TOP10 cells (Invitrogen) were transformed by electroporation with 5 μg aliquots of DNA and 500 μL of competent cells and were subsequently pooled. An aliquot was drawn to determine the average transformation efficiency for the library after 1 h of recovery. The remaining pooled cells were pelleted and then resuspended in 10 mL of Luria−Bertani (LB) media supplemented with carbenicillin at a final concentration of 50 μg/mL. This stock was used to inoculate 1 L of LB containing 50 μg/mL carbenicillin and then cells were incubated overnight at 37 °C. Glycerol stocks were prepared and stored at −70 °C.

The screening vectors JG1, JG1TEV, JG2, and JG2TEV were derived from a previously reported GFP fusion vector (GFP Folding Reporter[21]) and constructed as follows: the gene for enhanced green fluorescent protein (EGFP) was amplified from pEGFP-C1 (Clonetech) with the addition of upstream and downstream linkers added by overlap PCR.[25] The upstream linker includes an Nco I site, a 6-Thio/6× His expression tag, and Sal I and BamH I sites for inserting the shuffled fragment

(20) Hecht, M. H. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 8729−8730.
(21) Waldo, G. S.; Standish, B. M.; Berendzen, J.; Terwilliger, T. C. *Nat. Biotechnol.* **1999**, *17*, 691−695.
(22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(23) Hutchinson, E. G.; Thornton, J. M. *Protein Sci.* **1996**, *5*, 212−220.
(24) Burke, D. F.; Deane, C. M.; Blundell, T. L. *Bioinformatics* **2000**, *16*, 513−519.
(25) Hayashi, N.; Welschof, M.; Zewe, M.; Braunagel, M.; Dubel, S.; Breitling, F.; Little, M. *Biotechniques* **1994**, *17*, 310, 312, 314−315.

library in frame with EGFP downstream; the downstream linker incorporates a Pac I restriction site. This EGFP fragment was inserted into the Nco I/Pac I site of plasmid pMH4.[26] The pre-existing BamH I site in pMH4 was silenced by Quickchange mutagenesis (Stratagene). The downstream linker for JG1 and JG1TEV incorporate an additional C-terminal 6× His tag. A Tobacco Etch virus (TEV) protease cleavage site (ENLYFL—G) was inserted between the BamH I site and the start codon of EGFP in JG1 and JG2 by overlap PCR mutagenesis to afford JG1TEV and JG2TEV, respectively. Vectors JG5 and JG5TEV were constructed from pMH4 as described above with the exception that EGFP is deleted, and the TEV site in JG5TEV is inserted between the 6-Thio/6× His expression tag and the Sal I restriction site.

FACS Sorting. *E. coli* cells for fluorescence activated cell sorting (FACS) were prepared based on the protocol of Santoro et al.[27] Induced cells containing the library were washed twice with 1× PBS buffer, then diluted into phosphate buffered saline (PBS) buffer to an OD_{600nm} of 0.1, and stored on ice until sorting on a FACSVantage DiVa (Becton Dickinson). Selection criteria were set to sort positive cells into one of two tubes: high (GFPuv fluorescence >80 RFU) or low fluorescence (10 RFU < GFPuv fluorescence <80 RFU). Cells were recovered and amplified by overnight growth at 37 °C in LB with 50 μg/mL carbenicillin and stored as glycerol stocks at −70 °C. The sorted library was plated onto large format (30 cm × 30 cm square) plates (Genetix) containing LB agar supplemented with 50 μg/mL carbenicillin and 0.02% arabinose at a dilution sufficient to produce well-separated colonies. Fluorescent colonies were picked into LB media supplemented with 50 μg/mL carbenicillin, aliquoted into 96-well microtiter plates, and grown overnight at 37 °C.

Library Screening. Colonies (1 μL of saturated cell culture from each well) were screened based on the presence and size of inserted library clones by multiplex colony dPCR (1 unit Taq polymerase, 0.2 mM dNTPs, PCR buffer, 10 μM Insert Verification primers (fwd: CATCATCATCACGTGGTCGACCCATGGG and rev: GCCAGCG-GATCCCGATCGGCC)). Up to three 96-well plates were combined in multiplex format per PCR screen. The results of the combined dPCR reaction were deconvoluted by identifying those wells with positive inserts of desirable length and repeating the reactions under the same conditions with cells from a single well per dPCR reaction. Target clones identified from insert screening were examined for detectable expression of soluble fluorescent protein. Clones in 96-well microplates were grown to OD_{600nm} 0.75−1 in LB supplemented with 50 μg/mL carbenicillin at 37 °C and induced by addition of 0.02% arabinose (final concentration) for 3 h. At the 1.5 h point, a cocktail of ceftriaxone and cefotaxime (10 μg/mL final concentration of each) was added to each well to facilitate cell lysis for the soluble protein screen described below. After 3 h at 37 °C, the plates were centrifuged (4000× *g*, 15 min) and the supernatant was divided and transferred to clear bottom, black 96-well microplates (Corning), as well as clear, flat-bottomed 96-well microplates (Falcon) or Maxisorp ELISA plates (Nunc). The black plates were read on a fluorescent plate reader (Molecular Devices) (ex: 488 nm, em: 508 nm). Screening of target clones for soluble protein was performed in parallel with small-scale expression testing by established protocols.[28] Briefly, β-galactosidase activity was determined by adding 50 μL of 4× Z-buffer (180 mM Na_2HPO_4, 120 mM NaH_2PO_4, 30 mM KCl, 3 mM $MgSO_4$, 150 mM β-mercaptoethanol) and 50 μL of 4× ONPG (2.7 mg/mL ONPG, 60 mM K_2HPO_4, 33 mM KH_2PO_4, 8 mM $(NH_4)_2SO_4$, 2 mM sodium citrate) to the clear microplates containing 25 μL of soluble lysate. Upon development of a yellow color in positive control wells (EGFP), the reaction was quenched with 75 μL of 1 M Na_2CO_3, pH 8. The A_{420nm} and A_{550nm} of each plate were recorded with the time of quenching. The final 75 μL of lysate was bound overnight to the Maxisorp plates at 4 °C after dilution to a final volume of 100 μL with TBS (100 mM Tris, pH 7.5, 150 mM NaCl). The next day, buffer was removed and the plates were blocked with 1% (w/v) BSA in TBS for 4 h at 25 °C. The wells were subsequently washed with TBST (TBS with 0.1% v/v Tween-20), 100 μL of Ni-HRP conjugate (KPL Labs) was then added at a dilution of 1:2500 for 1 h at 25 °C, and the wells were again washed with TBST. One-hundred microliters of the HRP substrate (KPL Labs) was added, and color was allowed to develop until the positive control (EGFP) well was deep blue. The reaction was quenched with 100 μL of 1 M HCl, and the A_{420nm} for each plate was recorded. Solubility scores were calculated and assigned by first weighting the Ni-HRP A_{420nm} readings such that the mean was 1 order of magnitude greater than the mean of the β-galactosidase activity scores. The solubility score for each well was then calculated as the corrected Ni-HRP absorbance divided by the β-galactosidase activity score for the well.

Polypeptide Expression and Purification. One liter flasks of each clone were grown to OD_{600nm} 1 at 37 °C and induced by the addition of arabinose (0.02%). The temperature was then lowered to 18 °C for 12 h, and the cells were pelleted for storage or purification. Cells were ruptured by sonication in lysis buffer (50 mM Tris, pH 8, 500 mM NaCl, 1 mM $MgCl_2$, 5% v/v glycerol, 1 mM TCEP, 1 mM PMSF and 1 mM pepstatin). The soluble fraction of the cell lysate was isolated by ultracentrifugation for 1 h at 100000× *g*. Purification of all proteins was carried out at 4 °C using perfusion chromatography with Poros 20 MC resin charged with Ni^{2+} ions on a BioCAD Sprint (Applied Biosystems) in 50 mM Tris, pH 8.0, 500 mM NaCl, 1 mM TCEP. Proteins were eluted with a buffer containing 500 mM imidazole, 50 mM Tris, 500 mM NaCl, 1 mM TCEP, pH 6.0. Eluted fractions were analyzed by SDS-PAGE, and those containing the desired protein were pooled and exchanged into buffer A (10 mM Tris, pH 8, 100 mM NaCl, 1 mM TCEP). Desalted protein was subjected to size exclusion chromatography (SEC) on either Superdex 200 HR (>25 kD) or Superdex 75 HR (<25 kD) in buffer A. Proteins subsequently digested with TEV protease were subjected to an additional round of purification as described below. Pooled fractions were again concentrated as before, and final protein concentrations were determined both by Bradford assay (Pierce) and by A_{280nm} measurement. For circular dichroism (CD) or NMR analysis, proteins were buffer exchanged into PBS.

To cleave the fusion proteins, TEV protease was added at a ratio of 0.2 μg TEV protease per microgram of protein, incubated for 4 h at 25 °C in 10 mM Tris, pH 8.0, 100 mM NaCl, and then cooled to 4 °C for 8 h. Some fusion proteins required a 24-hour incubation at 4 °C. Proteolysis was monitored by SDS-PAGE. The cleaved target protein was purified by metal chelating chromatography and buffer exchange as described above.

bisANS Binding. Measurements of 5,5′-bis(8-anilino-1-naphthalene-sulfonate (bisANS) binding to purified protein were performed with a fluorescence plate reader (Molecular Devices) as previously described.[29] Stock solutions of 1 mM bisANS in DMSO, 100 mM potassium acetate buffer (pH 3.6, 4.6, and 5.6), 100 mM potassium phosphate buffer (pH 6.6 and 7.6), and 100 mM CHES buffer (pH 8.6 and 9.6) were prepared. Assay conditions were: 2 μM protein, 20 mM buffer, 100 mM NaCl, 2 mM DTT, and 15 μM bisANS. Bovine serum albumin (BSA) (2 μM) served as the positive control: negative control samples contained no protein. Each target protein and control was assayed at pH 3.6, 4.6, 5.6, 6.6, 7.6, 8.6, and 9.6 in triplicate in a clear bottom, black 96-well microplate (Corning). Upon mixing, the reactions were incubated for 10 min at 25 °C, and then emission was measured from 440 to 550 nm (excitation at 395 nm).

Circular Dichroism. CD measurements were carried out with an Aviv stopped flow CD spectrophotometer (model 202SF) using 0.2 cm path length quartz cuvettes with 6 μM samples in PBS buffer. Scans (200−250 nm) were carried out at 0.5 nm increments and averaged

(26) Lesley, S. A.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 11664−11669.
(27) Santoro, S. W.; Schultz, P. G. *Methods Mol. Biol.* **2003**, *230*, 291−312.
(28) Lesley, S. A.; Graziano, J.; Cho, C. Y.; Knuth, M. W.; Klock, H. E. *Protein Eng.* **2002**, *15*, 153−160.
(29) Offredi, F.; Dubail, F.; Kischel, P.; Sarinski, K.; Stern, A. S.; Van de Weerdt, C.; Hoch, J. C.; Prosperi, C.; Francois, J. M.; Mayo, S. L.; Martial, J. A. *J. Mol. Biol.* **2003**, *325*, 163−174.

for 5 scans per sample. Thermal denaturation of proteins was monitored by recording ellipticity at 222 nm with 3 scans per sample and temperature point. Thermal melts were performed from 20 to 90 °C, and back to 20 °C, with a 2 min pause between measurements taken at 2 °C intervals (dead band 0.1 °C, equilibration time 6 s, averaging time 2 s). Mean residue molar ellipticity ($[\Theta]_{MRW}$) was calculated from the spectrophotometer output (mdeg) using the formula: $[\Theta]_{MRW} = \Theta/(10 \times c_r \times l)$ where Θ is the CD signal (mdeg), c_r is the mean residue molar concentration, and l is the path length of the cell in centimeters. CD data was used to compute relative proportions of secondary structure components as previously published.[30] Thermal melt data was fit to a five parameter sigmoidal model using SigmaPlot 2000 (SPSS, Inc.).

1D NMR. NMR spectra were collected at 300 K on a Bruker Advance 600 MHz instrument equipped with a $^1H/^{13}C/^{15}N$−TXI CryoProbe (Bruker Biospin, Billerica, MA). Samples were typically in 500 μL PBS, pH 7 with 50 μL D_2O added as lock solvent. Samples of FK506 binding protein (FKBP12) in 20 mM sodium phosphate buffer at pH 6.75 were measured in the presence of various amounts of urea. FKBP12 consists of a five-stranded β-sheet with a short α-helix and connecting loops.[31] The protein has 124 amino acids including a His_6 N-terminal expression and purification tag, and a molecular weight of 13 757.6 Da. FKBP12 was expressed in *E. coli* and purified by Ni-affinity chromatograpy. Chemical shifts are relative to TSP (at 0.00 ppm) added as an internal standard or in a reference sample. 1D ^1H spectra were acquired using excitation sculpting with gradients for water suppression.[32] For samples in urea, an additional low power presaturation pulse was applied to suppress signals from the urea amino groups. The recycle delay was 2 s, and 16384 complex points were collected over a sweep width of 13.97 ppm. NMR spectra were processed in TOPSPIN 1.3 (Bruker Biospin, Billerica, MA) applying an exponential line-broadening function of 1 Hz to all spectra.

Results and Discussion

Library Design. Five primary elements, α-helices, β-strands, loops, chain initators, and chain terminators, were combined in a library to semirandomly assemble polypeptides of up to 800 amino acids in length, which were then fused to EGFP protein for rapid screening of folded proteins in *E. coli*[21,26]. A set of α-helices, β-strands, and loops of five or more residues each was chosen from 190 nonredundant *E. coli* protein structures in the PDB to assemble a database of secondary structural elements. These proteins represent each of the major classifications of protein fold topologies (all α, all β, α/β, and α+β).[22,33] The program PROMOTIF, which extracts and reformats PDB header information from each file into corresponding primary and secondary structural elements, was used to simplify construction of the database. The sLOOP database was used to identify the sequences of all loops in the pool and further classify each loop as to the secondary structures it joins (i.e., helix to helix, strand to helix, etc.). The initial database of secondary structural elements contains 4389 helices (5−55 residues in length), 2054 strands (5−21 residues in length), and 246 loops (5−21 residues in length).

A library consisting of 605 helices, 328 strands, and 246 loops was assembled as a representative pool of secondary structural elements from the above database. To construct the library, the list of primary sequence elements in the database was converted to the associated *E. coli* DNA sequence using DeCypher FrameSearch BLAST (Active Motif, Inc.). Separate PCR reactions were used to generate dsDNA encoding each element of the library. Random recombination of these structural elements could lead to a large number of nonproductive ligation products. For example, the possibility of fragment inversion during ligation is 50% per fragment if blunt-end ligation is employed. As such, only 1 in 2^n (where n is the number of fragments) ligation products would consist of structural elements entirely in the sense orientation. Therefore, short linkers, which incorporate a Bbs I Type II−S restriction enzyme recognition site that leaves 4-base sticky-end overhangs, were added to both ends of each library element. As a consequence, the sense strand orientation is maintained throughout fragment ligation. The sticky-end sequences were also designed such that a loop fragment was inserted between every helix and strand in the encoded polypeptide, and such that the loop fragments maintain the linkage orientation between helices and strands that existed in the source proteins (e.g., a loop that connects a helix at the amino-terminus and a strand at the carboxy-terminus would connect the same type of fragments in the library). As a result of this cloning strategy, the linker sequences insert 3 nonnative residues (GRA links helix to loop, VDH links loop to helix, LRP links strand to loop, and PAR links loop to strand) between each secondary structural element upon ligation. The linker residues were chosen based on a statistical propensity for joining a helix or a strand to a loop. Finally, chain terminator fragments were designed to incorporate restriction sites (a Sal I site at the 5′ terminus and a BamH I site at the 3′ terminus) for directional cloning of the ligated library and to provide a means of modulating the sizes of the final ligation products.

Library Construction and Characterization. Oligonucleotides incorporating the complementary Bbs I restriction sites were synthesized, and 1248 separate PCR reactions were carried out to generate dsDNA encoding each element of the library. A schematic illustration of the library assembly and cloning is shown in Figure 1. The products of the PCR reactions were individually inserted into the pBAD-Thio TOPO vector system as stocks. Although 13% (164/1248) of the ligation reactions failed, the resulting library still had considerable diversity—DNA sequencing verified 96% (1040/1084) of the remaining clones. Ligation reactions were performed on DNA fragments isolated from plasmid stocks using blunt ligation conditions (although the fragments all had sticky-end overhangs) to favor intermolecular over intramolecular ligation of compatible sticky-ends.[34] Identical stoichiometries of the helix, strand, and loop encoding fragments were used in the ligation reactions. In addition, 5′ terminators were also used at the same stoichiometry to inhibit intramolecular ligation (by "capping" the extending polymer at one end). The other terminator fragment was included in the reaction at a 100000-fold lower concentration. This ratio was empirically determined to give fragments of the desired length and composition of secondary elements.

Ligated fragments of the desired length (≥300 bp) were isolated by agarose gel electrophoresis and inserted into an EGFP fusion vector for screening and subsequent preliminary characterization. Library integrity was verified by analysis of 480 clones (from a library of 5.3×10^9 elements) for insert size and the presence of the invariant terminator sequences.

(30) Greenfield, N.; Fasman, G. D. *Biochemistry* **1969**, *8*, 4108−4116.
(31) Michnick, S. W.; Rosen, M. K.; Wandless, T. J.; Karplus, M.; Schreiber, S. L. *Science* **1991**, *252*, 836−839.
(32) Hwang, T. L.; Shaka, A. J. *J. Magn. Reson., Ser. A* **1995**, *112*, 275−279.
(33) Orengo, C. A.; Flores, T. P.; Taylor, W. R.; Thornton, J. M. *Protein Eng.* **1993**, *6*, 485−500.
(34) Upcroft, P.; Healey, A. *Gene* **1987**, *51*, 69−75.

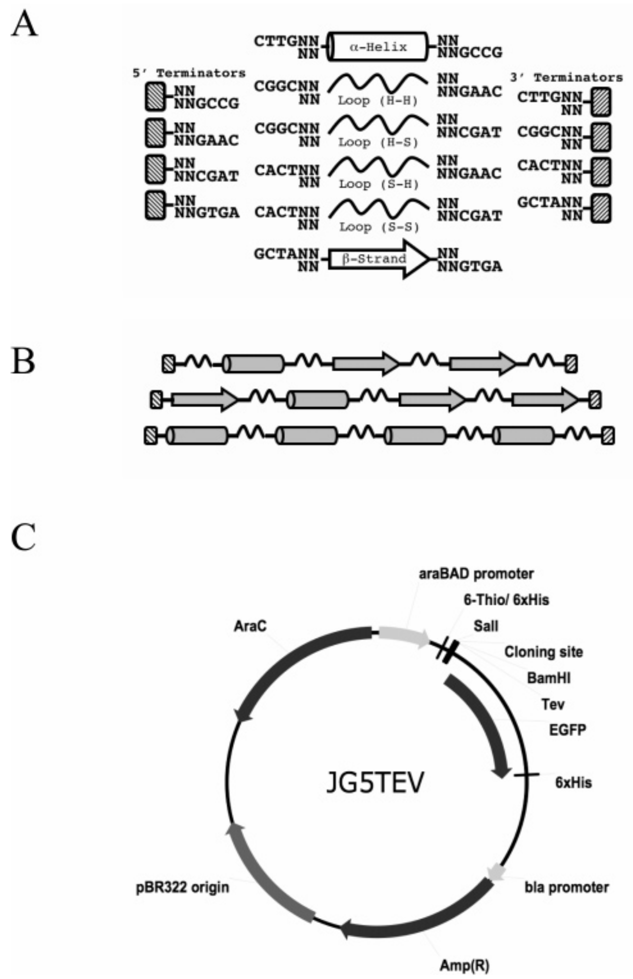


Figure 1. Schematic representation of the library synthesis scheme. (A) Secondary structural elements are identified and oligoDNA primers are designed to add Bbs I restriction sites unique to the 5′ and 3′ ends of each element family. Additional synthetic 5′ and 3′ terminators include complementary Bbs I restriction sites on one end. (B) Secondary structure elements are amplified and then digested, purified, and ligated in a single reaction. The ligated, polymerized library is PCR amplified and gel purified for cloning into the EGFP fusion vector for FACS. (C) Vector map of the vector JG5 for protein expression and characterization (in the absence of the EGFP fusion protein).

Additionally, 96 clones were fully sequenced. Insert sizes ranged from 200 to 2700 bp with an average insert size of 300−400 bp; 74% (71 of 96) of the clones were the product of the desired ligation reaction with 5′ and 3′ terminators at the ends of library elements. The number of fragments per clone ranged from 5 to 17 with an average of 7. In addition, 73% (52 of 71) of these clones had point mutations focused predominantly at the ligation sites; 17% (9 of 52) were point mutations that caused frame shifts. The high frequency of mutations at the ligation junctions may result from incomplete ligation leaving nicked DNA that is further propagated as mutations at these sites by multiplexed PCR amplification of the ligated library. The introduction of point mutations at the connections of loops with helices or sheets was not expected to adversely affect the quality of the library since this phenomenon has been shown to occur naturally as examination of exon joints in spliced eukaryotic mRNA has revealed that these splice points are highly susceptible to mutation.[35]

Library Screening. Previously, a strategy was reported[21] to screen for the expression of soluble proteins that involves fusion of the target protein to the N-terminus of GFP. Fusion proteins that are soluble and stable to proteolysis afford a high fluorescence signal, whereas insoluble proteins aggregate, resulting in reduced signal from the sequestered GFP. This is a particularly attractive approach since a large library of GFP fusion polypeptides can be rapidly screened by FACS ($\sim 10^8$ cells per hour throughput).[8,27,36,37] To this end, a modified GFP fusion vector was created in which a 6-Thio/6× His expression leader sequence was added upstream of the library cloning site to facilitate subsequent characterization of fluorescent clones.[26] Several variations of this vector were constructed to allow flexibility in the processing of the fusion proteins including the presence or absence of an additional C-terminal 6× His tag (JG1 or JG2, respectively) or the EGFP fusion protein (JG1/2 or JG5, respectively). Additionally, vectors were generated that contain a TEV protease cleavage site either between the cloning site and EGFP (JG1TEV and JG2TEV) or between the 6-Thio/6× His expression leader and the cloning site (JG5TEV).

The library of randomly ligated fragments was inserted into the EGFP fusion vector JG1, transformed into *E. coli* TOP 10 cells ($\sim 10^9$ efficiency) and screened by FACS. Induced cells containing the library could be binned into two distinct fluorescent fractions labeled high fluorescence (FACS GFPuv fluorescence gated >80 RFU) and low fluorescence (FACS GFPuv fluorescence gated >10 RFU and <80 RFU). On the basis of 1×10^8 sorted cells, approximately 5.7×10^7 cells were identified as low fluorescence (57% of the total cell population in the induced library) and 1.6×10^6 cells were identified as high fluorescence (1.6% of the total). An initial characterization of the FACS sorted library was carried out by colony dPCR to verify the presence and relative size of the library insert. A set of 1149 high fluorescence colonies was picked from sorted clones plated on solid media. A total of 299 of 1149 clones with library inserts (26%) screened by colony dPCR had inserts in the size range of 200−400 bp with an overall average insert size of 500 bp for all 1149 colonies screened. DNA sequencing of 24 of these clones showed 17 of 24 clones (71%) were desired products of the ligation protocol. Twelve of the 17 inserts (70%) contained point mutations focused at the ligation points.

Because of the large number of high fluorescence clones initially identified by FACS, additional assays amenable to high throughput analysis were used to identify those clones most likely to possess the characteristics of soluble, folded proteins. A semiquantitative assay of fluorescence emission for each induced clone should correlate with the amount of soluble, folded protein (because GFP is not an environmentally sensitive fluorophore). Analysis of the 1149 clones with library inserts identified 44 clones (4%) with measurable fluorescence (greater than the baseline of 1000 RFU) (Figure 2A). The level of measured fluorescence from this expression assay differs by 2 orders of magnitude over that measured in the FACS instrument due to the difference between the FACS assay (examining the fluorescence emission intensity from single cells) versus the microplate protein expression screen (examining the fluorescence emission of millions of cells per well simultaneously). A

(35) Gilbert, W.; de Souza, S. J.; Long, M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 7698−7703.
(36) Becker, S.; Schmoldt, H. U.; Adams, T. M.; Wilhelm, S.; Kolmar, H. *Curr. Opin. Biotechnol.* **2004**, *15*, 323−329.
(37) Winson, M. K.; Davey, H. M. *Methods* **2000**, *21*, 231−240.

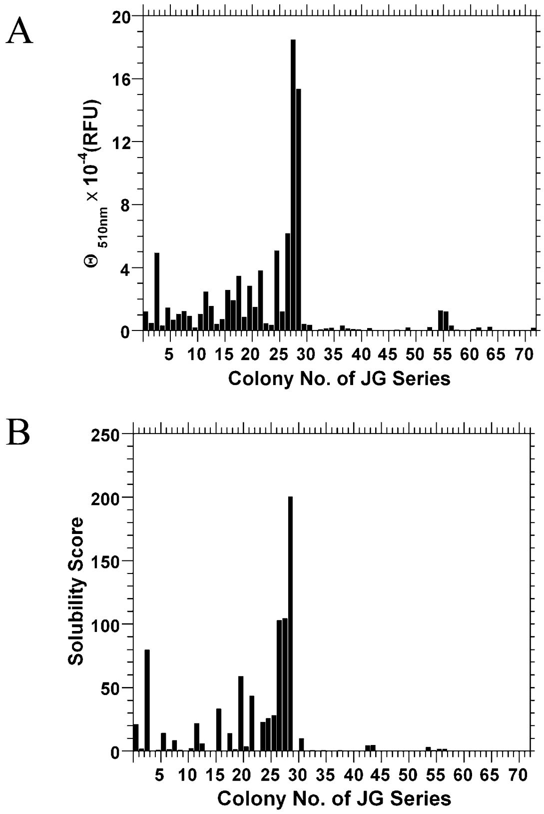


Figure 2. Secondary library screening. Library screening of insert positive clones identified and isolated by FACS. Clones are screened by measurement of (A) fluorescence emission at 510 nm and (B) solubility scores for amplified cultures of 70 clones from both the Low and High Fluorescence libraries. Fluorescence emission of the induced library detects induced EGFP fusion proteins. The two-part solubility screen measures the expression of soluble His-tagged protein.

second assay, which combines a β-galactosidase (β-gal) reporter assay for expression of misfolded protein with a nickel conjugated horse radish peroxidase (HRP) colorimetric assay that indicates the presence of soluble protein with 6× His tags,[28] was used to independently verify protein expression and solubility (Figure 2B). Twenty-two clones were found to have significant solubility scores, 14 of which were already identified in the fluorescence expression assay (Figure 2B). In addition, 18 clones that neither showed measurable fluorescence nor had significant solubility scores, but did have high Nickel-HRP assay scores, were considered for further characterization. Lesley et al.[28] had noted in their description of the solubility screen that it was more likely that false negative results due to high β-gal activity coupled with moderate or high Nickel-HRP response could fail to identify truly soluble proteins than the likelihood of a false negative resulting from high Nickel-HRP alone. These 70 clones were inserted into the appropriate expression and screening vectors for further characterization.

Characterization of Selected Polypeptides. Seven of these 70 clones that were expressed as nonfusion proteins with a 6-Thio/6× His tag (JG5 vector) produced sufficient protein to be visualized as Comassie-stained bands by SDS-PAGE after a single nickel chelating FPLC purification step. Four of these seven clones were found in the fluorescence screen and all seven were in the solubility screen, indicating that the solubility screen may be a better judge of soluble, folded protein. DNA sequencing of the seven clones showed that clones 5.6 and 5.12 were identical and that clone 5.24 contained a nonsense mutation. Therefore, clones 5.1, 5.6, 5.26, 5.29, and 5.31 were further characterized. Each of the five clones (5.1, 5.6, 5.26, 5.29, and 5.31) was between 198 and 273 bp (65 and 90 amino acids, respectively) in length (Figure 4 and Table 2). The translated protein sequence of each clone was submitted to a BLAST search (blastp and blastn) and InterProScan to detect any DNA contamination from the experimental system.[38,39] The elements of the secondary structure library were easily identified by performing a blastn search optimized for short (25 bases or fewer), nearly identical sequences.[40] However, without this constraint, four of the five clones possess no global sequence homology to any known *E. coli* protein or any other protein sampled in the BLAST and InterProScan searches. Noteworthy is the fact that clone 5.6 has global sequence homology with a family of environmental bacterial aspartate racemases with *E* values from the blastp search of 2×10^{-17}. The a priori likelihood that the clone 5.6 sequence of assembled fragments exists in the library is 2.4×10^{-17} (246 loops × 328 strands × 246 loops × 605 helicies × 246 loops × 328 strands × 246 loops) while there are only approximately 6×10^{15} DNA molecules in the ligation reaction to start (250 ug/mL total DNA corresponds to approximately 11 pmol of each element which is 6×10^{12} molecules × 1179 fragments). It would seem unlikely that this sequence exists more than once in the final ligated library pool. Figure 3 displays an alignment of 5.6 with its 5 closest homologues. The BLAST search of clone 5.6 identified fragments of the gene that exist in the *E. coli* fragment library but not a single *E. coli* gene with detectable homology. Therefore, the possibility of environmental contamination is unlikely. Conversely, the search identified highly homologous sequences from several environmental bacteria, though none are identical. The closest related sequence (by phylogeny) is from a psychrophillic bacteria (*Polaromonas sp.*), which is highly unlikely to be a contaminant in our more temperate laboratory. Thus, clone 5.6 demonstrates that libraries created from randomly recombined secondary structural elements can at a minimum recapitulate known protein structural domains.

The protein sequences of clones 5.1, 5.6, 5.26, 5.29, and 5.31 including the expression and purification tag are listed in Figure 4. With the exception of clone 5.31, all proteins have predicted p*I* values > 9.6; clone 5.31 is the only neutrally charged protein in the group (Table 2). Compared to the amino acid composition expected from the codon usage in *E. coli* (http://cmr.tigr.org/tigr-scripts/CMR/shared/Menu.cgi?menu = genome), clones 5.1 and 5.29 contain disproportional numbers of Arg, Pro, and Ser residues. Similarly, clones 5.6 and 5.26 are rich in Ala and Arg whereas 5.31 is rich in Pro and Gly.

The biophysical properties of all five clones (5.1, 5.6, 5.26, 5.29, and 5.31) were characterized by a number of methods including bisANS binding, CD spectrometry, thermal denaturation, and 1D NMR.[19,41−44] The fluorescent dye 5, 5′-bis(8-anilino-1-naphthalenesulfonate) (bisANS), which binds to ex-

(38) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(39) Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. *Nucleic Acids Res.* **2005**, *33*, W116−W120.
(40) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403−410.
(41) Betz, S. F.; Raleigh, D. P.; DeGrado, W. F. *Curr. Opin. Struct. Biol.* **1993**, *3*, 601−610.
(42) Davidson, A. R.; Lumb, K. J.; Sauer, R. T. *Nat. Struct. Biol.* **1995**, *2*, 856−864.
(43) Hecht, M. H.; Das, A.; Go, A.; Bradley, L. H.; Wei, Y. *Protein Sci.* **2004**, *13*, 1711−1723.

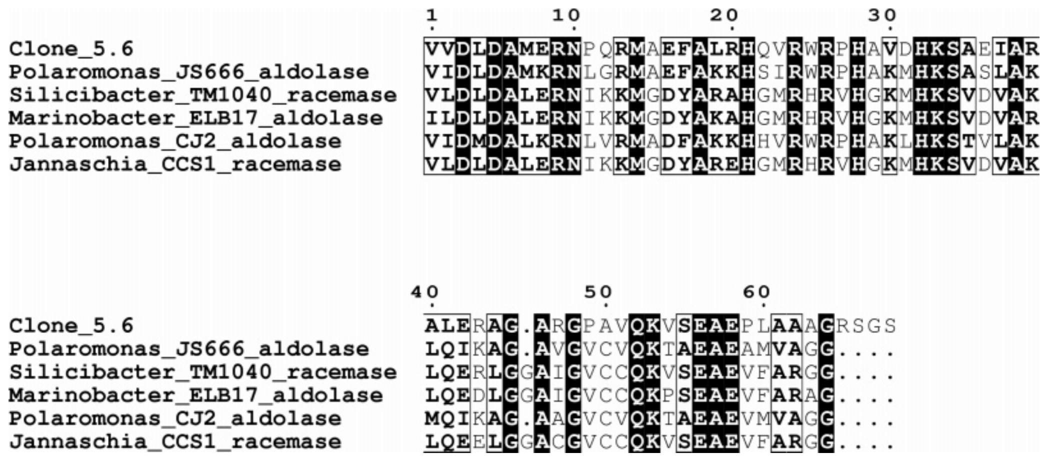


Figure 3. Primary sequence alignment of clone 5.6 with the 5 closest homologous proteins identified by blastp search[40] using ClustalW[58] with default settings on the World Wide Web service of the Euorpean Bioinformatics Institute (http://www.ebi.ac.uk/clustalw). Graphical image prepared using ESPript 2.2.[59] Identical residues are white, and conserved residues are boxed and bold.

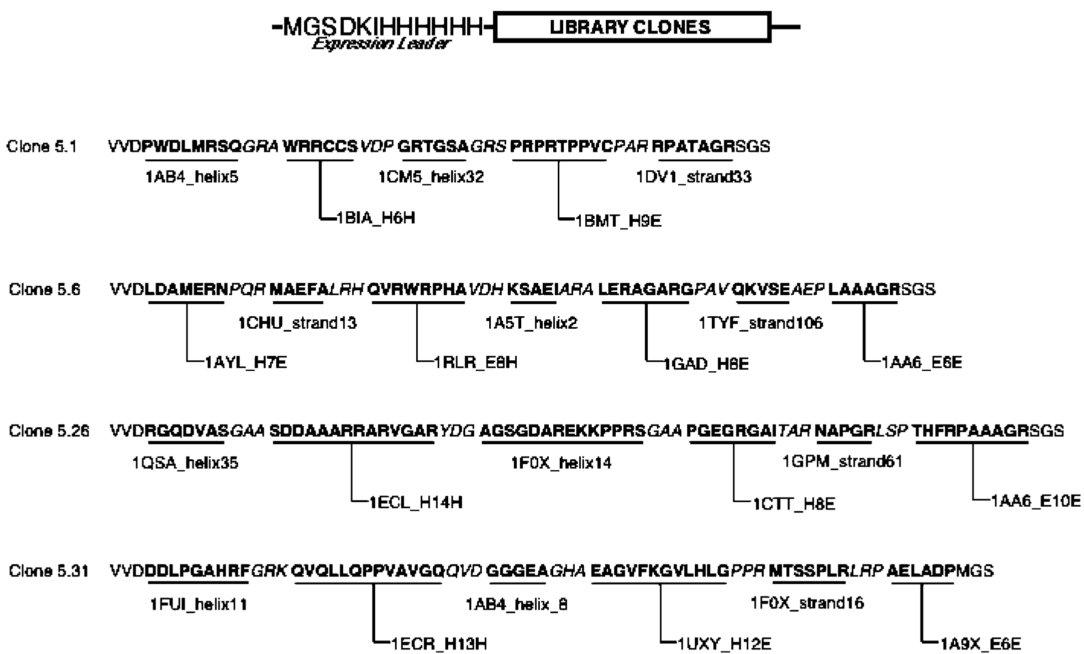


Figure 4. Sequences and fragment origins of clones 5.1, 5.6, 5.26, and 5.31. The invariant sequence from the expression vector is represented schematically above each of the clones. The secondary structure elements are bold, underlined, and labeled as to the original protein structures (Protein Data Bank code and fragment identifier). H/E-number-H/E defines the loop fragments by the type of secondary structure element joined (H-helix, E-strand), and the number reflects the number of residues in the loop. The inserted synthetic linkers from the Bbs I restriction sites are in italics.

Table 2. Molecular Weight, p*I*, and Thermal Stability of Target Clones

clone	residues	est. mol. wt. (Da)	est. p*I*	ΔG_u (kcal/mol)	T_u (°C)
5.1	65	7156.0	11.58	3.9	68.5 ± 2
5.6	79	8749.7	9.68	3.6	66.2 ± 1
5.26	90	9280.1	11.09	not measured	58.7 ± 3
5.29	69	8141.3	11.00	not measured	not measured
5.31	84	8916.0	6.74	4.1	71.2 ± 1

[a] Molecular weight and p*I* estimates are based on the sequences including the expression and purification tag but without the N-terminal methionine.

posed or accessible hydrophobic surfaces of proteins,[45] has been used as a probe of protein folding. BisANS fluorescence increases upon binding to unfolded as well as molten globule proteins.[46] Environmental sensitivity of bisANS binding implies structural fluctuations or unfolding transitions, whereas insensitivity to changes in pH, temperature, or denaturant infers either a lack of significant secondary or tertiary structure (random coil) or a highly stable compact tertiary structure. Therefore, the binding of bisANS to clones 5.1, 5.6, 5.26, 5.29, and 5.31 was measured as a function of pH. At pH 7.0, the maximum fluorescence was at 495 nm for all of the proteins (Figure 5); this wavelength was therefore used to monitor emission intensity when the samples were assayed from pH 3.6 to 9.6. The bisANS binding data showed marked differences in pH sensitivity for the different clones. All clones showed a significant increase in fluorescence at acidic pH with clone 5.26 having the highest degree of change in ANS binding. Therefore, it is likely that these clones are undergoing pH-induced structural changes.

(44) Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. *Science* **2003**, *302*, 1364−1368.
(45) Smoot, A. L.; Panda, M.; Brazil, B. T.; Buckle, A. M.; Fersht, A. R.; Horowitz, P. M. *Biochemistry* **2001**, *40*, 4484−4492.
(46) Goto, Y.; Fink, A. L. *Biochemistry* **1989**, *28*, 945−952.

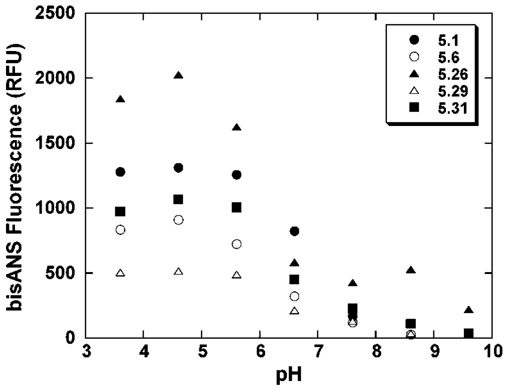


Figure 5. bisANS binding as function of pH as measured by fluorescence emission at 495 nm for clones 5.1, 5.6, 5.26, 5.29, and 5.31; 2 μM of each protein was combined with 15 μM bisANS and incubated for 10 min at 25 °C in acetate (pH 3.6, 4.6, and 5.6), phosphate (pH 6.6 and 7.6), and CHES (pH 8.6 and 9.6) buffers in triplicate prior to measuring fluorescence emission.

The far-UV CD spectra (250−200 nm) of clones 5.1, 5.6, 5.26, 5.29, and 5.31 were then determined to assess the presence of secondary structure in the clones. Clone 5.29 lacks regular secondary structure (0% helical content) consistent with a random coil topology. This data correlates well with the low degree of bisANS binding upon pH induced unfolding. For the other four clones, CD spectra revealed distinct minima at 208 and 222 nm, corresponding to varying degrees of largely α-helical secondary structure. Calculation of the secondary structure for clones 5.1, 5.6, 5.26, and 5.31 (Figure 6A) from the CD data indicated they have 40% (5.1), 29% (5.6), 15% (5.26), and 47% (5.31) α-helical structure.[30] The experimentally determined helical content appears to contrast with the expected fraction of helical content based on the secondary structures of the initial library elements (assuming the fragments adopt similar secondary structures in the selected polypeptides). The latter values predict that the polypeptides would be 22% (5.1), 6% (5.6), 23% (5.26), and 17% (5.31) α-helical. Secondary structure predictions with AGADIR[47] suggest there is little to no α-helical structure in any of the polypeptide sequences. AGADIR predicts 1.2% (5.1), 7.8% (5.6), 4.8% (5.26), and 0.3% (5.31) α-helical content which does not correlate to either the CD data or the estimated percentage of α-helical content based on the original fragment origins (see below). The estimates based on the fragment origins would seem to correlate more closely with the CD data when all three estimates are examined together suggesting that some degree of residual secondary structure may be retained by the fragments outside the context of their native sequence. The small dataset presented here is insufficient to draw any conclusions with regard to the correlation of structural propensities to the polypeptides in solution. The lack of correlation between the fragment origins and the structure they appear to assume in the selected polypeptides is not too surprising because it is well known that primary sequence is not the sole determinant of secondary structure formation.[52−56] Notably, although the bisANS binding data suggests that clone 5.26 is the most conformationally sensitive to pH changes, it displays the least degree of helical character of the four remaining clones.

The α-helical character of clones 5.1, 5.6, 5.26, and 5.31 provided a means to study unfolding by monitoring the change in ellipticity at 222 nm as a function of temperature (Figure 6B). For clones 5.1 and 5.6, the ellipticity decreased to 37 and 35% of the maximal value at 20 °C, respectively. The changes were much smaller for clones 5.26 and 5.31 with 84 and 80% of the CD signal remaining at 90 °C, respectively. Clones 5.1, 5.6, and 5.31 were found to unfold reversibly (Figure 6C and D). T_u (midpoint temperature for unfolding) and ΔG_u (free energy of folding) values (Figure 6C and Table 2) were determined by fitting the reversible denaturation data to a five-parameter model via nonlinear regression. Values of ΔG_u ranged from 3.6 to 4.1 kcal/mol and are similar to other de novo "synthetic proteins".[17,48−50] Clones 5.1, 5.6, and 5.31 have near superimposable unfolding and refolding profiles that can be modeled accurately with a two-state transition suggesting cooperative unfolding. Clones 5.1, 5.6, and 5.31 may therefore have some tertiary interactions. In contrast, clone 5.26 unfolded irreversibly with little change in ellipticity as the temperature was lowered after initial unfolding (Figure 6B and D). Both 5.26 and 5.31 retained a large proportion of the CD signal even at the highest temperature. Common to all four clones is the high helicity that cannot be explained by simple addition of the expected CD signal from individual elements of secondary structure and is also substantially higher than the α-helcial content predicted by AGADIR.

The chemical shift dispersion of the 1D NMR spectra of proteins is a qualitative measure of protein folding and tertiary packing of protein side chains.[57−59] The 1D NMR spectra of clones 5.6, 5.26, and 5.31 measured at neutral pH are shown in Figure 7A, B, and C, respectively. Data collection for clones 5.1 and 5.29 proved difficult due to insufficient quantities of concentrated, soluble protein. The NMR spectra of the three other clones are different from each other but display little dispersion in both the amide (6−10.5 ppm) and methyl proton regions (−0.5−2.5 ppm) compared to the spectra of FKBP12, a 13.9 kDa, compact predominantly β-sheet protein (Figure 7D).[31] Consistent with its high α-helical content as measured by CD, the amide resonances of clone 5.31 (Figure 7C) show less dispersion than the other two clones. Significant line broadening is observed for the amides of clone 5.26 (Figure 7B) and to some extent also for clone 5.6 (Figure 7A), suggesting conformational exchange rather than a highly stable secondary and tertiary structure. This conclusion is supported by the observation of only limited protection of amide resonances from solvent exchange in 1D experiments with presaturation of the solvent (Supplemental Figure S1). Addition of solid urea to a sample of 5.26 induces only small chemical shift

(47) Munoz, V.; Serrano, L. *Biopolymers* **1997**, *41*, 495−509.
(48) Makhatadze, G. I.; Privalov, P. L. *Adv. Protein Chem.* **1995**, *47*, 307−425.
(49) Privalov, P. L. *Adv. Protein Chem.* **1979**, *33*, 167−241.
(50) Wei, Y.; Liu, T.; Sazinsky, S. L.; Moffet, D. A.; Pelczer, I.; Hecht, M. H. *Protein Sci.* **2003**, *12*, 92−102.
(51) Lau, K. F.; Dill, K. A. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 638−642.
(52) Dinner, A. R.; Sali, A.; Smith, L. J.; Dobson, C. M.; Karplus, M. *Trends Biochem. Sci.* **2000**, *25*, 331−339.
(53) Dobson, C. M. *Nat. Rev. Drug Discovery* **2003**, *2*, 154−160.
(54) Looger, L. L.; Dwyer, M. A.; Smith, J. J.; Hellinga, H. W. *Nature* **2003**, *423*, 185−190.
(55) Rojas, N. R.; Kamtekar, S.; Simons, C. T.; McLean, J. E.; Vogel, K. M.; Spiro, T. G.; Farid, R. S.; Hecht, M. H. *Protein Sci.* **1997**, *6*, 2512−2524.
(56) Mossing, M. C.; Bowie, J. U.; Sauer, R. T. *Methods Enzymol.* **1991**, *208*, 604−619.
(57) Wei, Y.; Hecht, M. H. *Protein Eng.* **2004**, *17*, 67−75.
(58) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. *Nucleic Acids Res.* **1994**, *22*, 4673−4680.
(59) Gouet, P.; Robert, X.; Courcelle, E. *Nucleic Acids Res.* **2003**, *31*, 3320−3323.

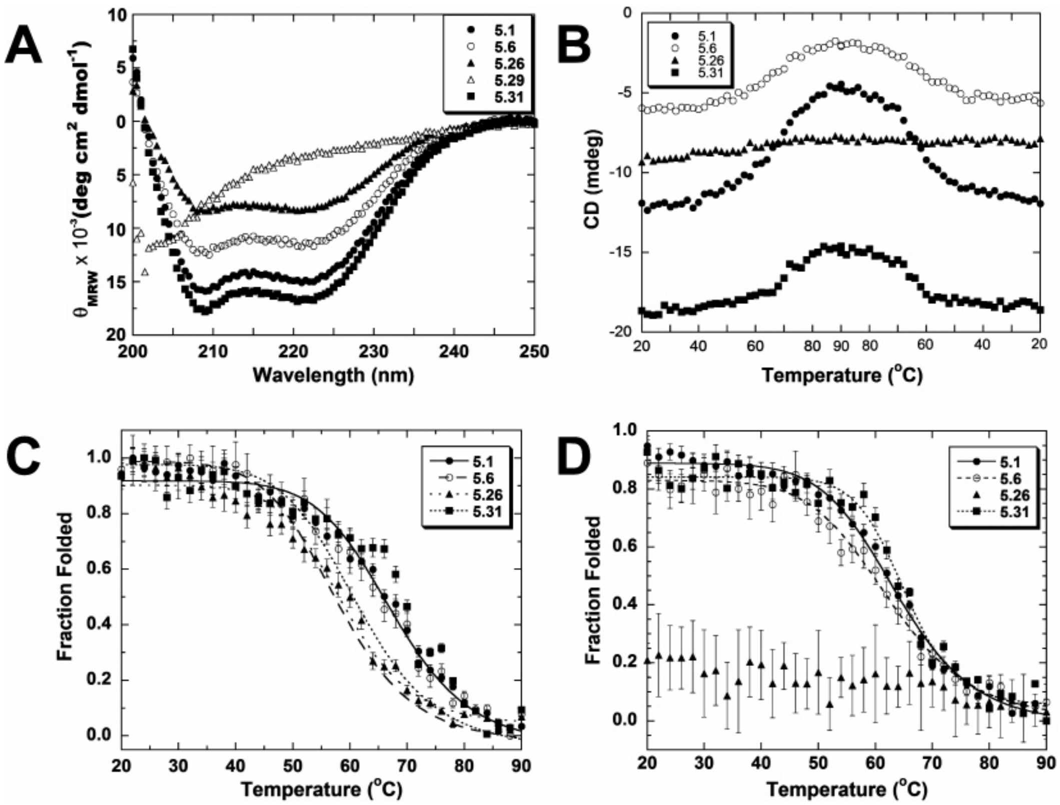


Figure 6. Circular dichroism measurements. (A) Far-UV CD spectra for clones 5.1, 5.6, 5.26, 5.29, and 5.31. Spectra of each protein at 6 μM collected at 25 °C in PBS buffer averaged over five scans. (B) Thermal unfolding and refolding of clones 5.1, 5.6, 5.26, and 5.31 measured by loss of ellipticity at 222 nm. (C) Thermal unfolding and (D) thermal refolding expressed as the fraction of protein folded at a given temperature. Spectra were collected with 6 μM protein in PBS buffer from 20 to 90 °C, and then from 90 back to 20 °C. Clone 5.26 did not reversibly unfold. The data were fitted to a sigmoidal equation.

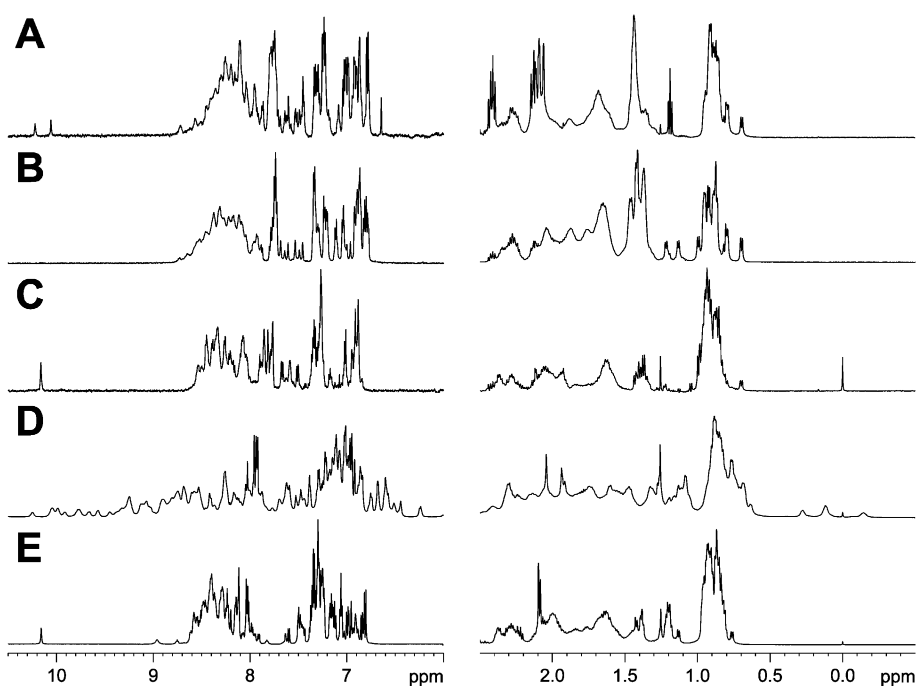


Figure 7. 1D NMR spectra of the amide regions (6−10.5 ppm) and methyl regions (−0.5−2.5 ppm) of clones (A) 5.6, (B) 5.26, and (C) 5.31. For comparison, the spectra of folded FKBP12 and of unfolded FKBP12 in 6.7 M urea are shown in (D) and (E), respectively. Spectra were collected at 300 K as described in Materials and Methods.

changes (Figure S2) and different sets of resonances for folded and unfolded protein (as is the case for FKBP12; Figure S3) cannot be identified. The NMR spectra of the selected clones are similar to that of FKBP12 unfolded in 6.7 M urea (Figure 7E) but CD strongly supports stable, helical secondary structures beyond what would be expected based on the amino acid sequence alone. The selected clones should therefore not be considered "native-like" proteins but rather "molten globule-like",[46] perhaps as an intermediary stage in the evolution from a random assembly of secondary structural elements toward a compact stable protein. The selection process utilized in this study identified successful combinations of secondary structural elements that could be combined to form a soluble assembly of secondary structure. Further mutagenesis with an alternate selection specific to promoting the formation and stabilization of a hydrophobic would be necessary to achieve the next stage of forming a compact "native-like" protein.

By selecting elements from known folded proteins that combine into novel polypeptide sequences, we have identified a number of soluble polypeptides after screening only a small fraction of the first generation library (0.0001%; 1149 clones from a library of $\sim 10^9$). Of the 1149 clones screened, 4 sequences (0.3%; clones 5.1, 5.6, 5.26, and 5.31) were identified as proteins with significant amounts of secondary structure. Assuming that cell sorting and FACS analysis were not statistically biased, our sampling would suggest that the library should contain on the order of 2.25×10^6 novel proteins with significant secondary structure (3 of the 4 clones represented sequences without any homologs). This is certainly an overestimate as amplification of the library prior to cloning and transformation would have slightly biased the diversity by introduction of duplicate sequences. However, the total number of unique sequences in this library with the desired characteristics appears to be significant nonetheless. Previous theoretical estimations by Lau and Dill[51] predict that the fraction of protein sequence space represented by those sequences (average chain length 70 residues) that fold into stable, native structures to be between 10^{-4} and 10^{-7}, suggesting that the system described here is more efficient for creating novel folded protein sequences. A previous report describing libraries of similar size constructed from semirandom assembly of secondary structure elements identified only sequences that required high ionic strength or chaotropic agents to induce measurable secondary structure and native-like protein character.[19] The secondary structure elements used in the creation of those libraries were based on binary patterning instead of native sequences, suggesting that the use of native sequences imparts a greater degree of secondary structure upon the product polypeptide sequences. Most surprising was our identification of a sequence (5.6) that largely recapitulates a conserved domain from species of environmental bacteria without any homologous sequences within the *E. coli* genome. This gives rise to hope that libraries constructed from elements of secondary structure may improve the likelihood of identifying novel proteins possessing detectable activity over the use of random sequences.

Conclusions

We have described an experimental system by which secondary structure elements are combinatorially assembled into a library that contains soluble proteins with significant secondary structure. Most noteworthy, the initial screen of this library identified one protein (clone 5.6) that is highly homologous to the N-terminal domain of a family of known proteins. Three other proteins were identified that had no homologous sequence in any of the available genomes. The high rate (0.3%) of identifying folded proteins by screening only a small proportion of our library suggest that the semirandom assembly of secondary structural elements results in a significant likelihood of identifying novel protein sequences. Additional optimization of the library construction and screening protocols, as well as inclusion of additional structural elements (such as independently folding domains and sequences that can function as folding nuclei[52,53]), may further improve the number and quality of proteins selected from the screens. Alternatively, the generation of additional point mutations in the library by iterative rounds of error-prone PCR and/or DNA shuffling might be expected to lead to improved packing or hydrogen bonding interactions between the secondary structural elements. It should also be possible to screen the next generation libraries for functional activities such as heme and DNA binding[54−56] and esterase activity.[57]

Acknowledgment. We thank Heath Klock and Drs. Jeff Kelly, Ray Stevens, Dave Goodin, Jack Johnson, Dave Wemmer, Glen Spraggon, Andreas Kreusch, Chris Lee, and Mike DiDonato for helpful discussion and use of resources. Additionally, we thank Alan Saluk and Cheryl Kim for FACS assistance and Ted Foss and Luke Wiseman for assistance with CD measurements. We thank Susan Crown for samples of FKBP12 protein. This work was supported by NIH grant 5R01 GM56528.

Supporting Information Available: Additional NMR data and complete ref 26. This material is available free of charge via the Internet at http://pubs.acs.org.

JA074405W