CECsoil) × 100] with an equation approximating soil cation exchange capacity (soil CEC), which gives TEB = (BS%/100) × {3.5 × OC% + [(Clay% × CECclay)/100)]}; where OC% is percent organic carbon in the soil, Clay% is the percent clay content, and CECclay is the approximate cation exchange capacity for the dominant clay mineral (for example, kaolinite, allophane) or an average value of 40 meq per 100 g for mixed clay mineralogy. Soil pH values are based on the average weighted value of measured pH over the top 100 cm of 400 soil profiles from the FAO soil units. [AGLS Soil Resources Group, *Soil Properties and Qualities Estimation Based on Soil Groups by the Soil Map of the World* (FAO, Rome, Italy, 1992)].

29. I thank S. Buol, D. DeAngelis, G. Marland, J. McBrayer, W. Post, and D. Richter for critical review of the manuscript. F. Nachtergaele of the FAO Soil Resources Working Group in Rome provided the information on soil properties used to generate Fig. 3. Supported by the Program for Ecosystem Research of the U.S. Department of Energy's Office of Health and Environmental Research under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc. This is Oak Ridge National Laboratory Environmental Sciences Division publication 4172.

RESEARCH ARTICLE

Protein Design by Binary Patterning of Polar and Nonpolar Amino Acids

Satwik Kamtekar, Jarad M. Schiffer,* Huayu Xiong, Jennifer M. Babik, Michael H. Hecht†

A general strategy is described for the de novo design of proteins. In this strategy the sequence locations of hydrophobic and hydrophilic residues were specified explicitly, but the precise identities of the side chains were not constrained and varied extensively. This strategy was tested by constructing a large collection of synthetic genes whose protein products were designed to fold into four-helix bundle proteins. Each gene encoded a different amino acid sequence, but all sequences shared the same pattern of polar and nonpolar residues. Characterization of the expressed proteins indicated that most of the designed sequences folded into compact α-helical structures. Thus, a simple binary code of polar and nonpolar residues arranged in the appropriate order can drive polypeptide chains to collapse into globular α-helical folds.

The essential features for the design of proteins are suggested by two familiar properties of natural proteins. First, globular, water-soluble proteins invariably fold into structures that bury extensive hydrophobic surface area while simultaneously exposing polar side chains to solvent (*1*). Second, these structures typically contain an abundance of hydrogen bonded secondary structure (α helices and β strands).

The dual constraints of forming regular secondary structure while burying hydrophobic side chains (and exposing hydrophilic ones) are met most directly by designing novel amino acid sequences capable of forming amphiphilic α helices, β strands, or both. For a segment of polypeptide chain to form an α helix or a β strand with one hydrophilic face and one hydrophobic face, the sequence must be designed with a periodicity of polar and nonpolar residues that matches the repeat for that type of secondary structure. For example, for the design of a stable β sheet protein, the sequence must be composed predominantly of alternating polar and nonpolar residues. Conversely, for an α-helical protein, the periodicity of polar and nonpolar residues must approximate the 3.6-residue repeat that is characteristic of α helices.

We describe a general strategy for protein design that is based on the assumption that the ability of a sequence to form amphiphilic secondary structures may actually suffice to drive a designed polypeptide chain to fold into a compact native-like structure. Our strategy is based on the premise that formation of stably folded structures does not require the explicit design of specific interresidue contacts; the precise packing of the three-dimensional jigsaw puzzle need not be specified a priori. Only the sequence location, but not the identity, of the polar and nonpolar residues must be specified explicitly.

We produced a collection of protein sequences that satisfy these criteria. This collection was generated by constructing a degenerate family of synthetic genes. Each gene encoded a different amino acid sequence, but all sequences shared the same periodicity of polar and nonpolar residues. The sequence degeneracy in this family of genes was made possible by the organization of the genetic code: Wherever a nonpolar amino acid was required, the degenerate codon NTN was used (where N represents a mixture of A, G, T, and C). Wherever a polar amino acid was required, the degenerate codon NAN was used. With these degenerate codons, positions requiring a nonpolar amino acid were filled by Phe, Leu, Ile, Met, or Val, whereas positions requiring a polar amino acid were filled by Glu, Asp, Lys, Asn, Gln, or His.

Design criteria. To test our design strategy, we focused on a structural motif that was small and simple, yet large enough to form a globular structure with a well-defined hydrophobic core and an abundance of secondary structure. To simplify the design and characterization of the novel proteins, we chose a fold that was composed of only one type of secondary structure, α helices. Previous work on the design of peptides and proteins has indicated that helices are easier to design than β structures (*2–5*). This probably reflects both the greater modularity of helical structures and the greater tendency of β structures to form insoluble aggregates. For these reasons we designed four-helix bundles. The four-helix bundle is a common fold among natural proteins (*6*) and has also been the target structure in previous efforts directed toward the design of novel proteins (*2–4*). Two representations of an idealized four-helix bundle are shown in Fig. 1.

To sample a large section of sequence space, we used degenerate nonpolar codons at each of the buried positions and degenerate polar codons at each of the surface positions in the α helices of our designed proteins. In the four helices shown in Fig. 1B, there are 24 buried (hydrophobic) positions and 32 surface (hydrophilic) positions. Because each buried position can be occupied by one of five different nonpolar side chains (Phe, Leu, Ile, Met, or Val) and each surface position can be occupied by any one of six different polar side chains (Glu, Asp, Lys, Asn, Gln, or His), a total of $5^{24} \times 6^{32} = 4.7 \times 10^{41}$ different amino acid sequences are theoretically possible. Our design strategy is based on the premise that a substantial fraction of the sequences fitting this pattern will actually fold into proteins that are compact, stable, and α-helical.

In our collection of sequences (Fig. 2) the interhelical turn sequences were not degenerate. It was necessary to define explicitly some regions of our sequence in order to facilitate gene construction, and

S. Kamtekar, H. Xiong, and M. H. Hecht are in the Department of Chemistry and J. M. Schiffer and J. M. Babik are in the Department of Molecular Biology at Princeton University, Princeton, NJ 08544.

*Present address: Department of Biology, University of California, San Diego, La Jolla, CA 92093.
†To whom correspondence should be addressed.

we chose to confine the nonvariable regions to the interhelical turns. The precise sequence of these interhelical regions probably is not crucial in dictating the structure of a four-helix bundle, as suggested in an earlier study demonstrating that virtually any sequence can be tolerated in an interhelical turn of the natural four-helix bundle protein cytochrome b562 (*7*). Nonetheless, because our synthetic strategy required the interhelical regions to be fixed (not variable) sequences, we chose sequences that would be reasonable at these positions. For the positions at the amino and carboxyl ends of the α helices (the "N-cap" and "C-cap"), we used glycine because its flexibility might be advantageous, given that we could not predict the exact orientation of the helices relative to one another. Also, glycine is found frequently at these positions in natural proteins (*8*). At the position after the C-cap (that is, at the beginning of the three-residue interhelical turn) we chose proline, because it is a strong helix breaker (*9*, *10*) and is the most commonly found residue at that position in natural proteins (*8*). The other two residues of the interhelical turn differ for the three turns. The first and third turns were Pro^{18}-Asp^{19}-Ser^{20} and Pro^{56}-Arg^{57}-Ser^{58}, respectively. These sequences were chosen to make the turns fairly polar and to allow for the possibility of a salt bridge between Asp^{19} of the first turn and Arg^{57} of the third turn. The sequence of the middle turn (Pro^{37}-Ser^{38}-Gly^{39}) was dictated by the placement of a restriction endonuclease site for Bgl I in the center of the gene.

The length of the designed interhelical turns was also considered. By inserting three residues between the C-cap glycine of one helix and the N-cap glycine of the next helix, we designed sequences to favor the formation of interhelical turns and disfavor the formation of a single long amphiphilic helix. This feature of negative design causes the hydrophobic stripe of one helix to be offset relative to the hydrophobic stripe of the preceding helix (*2*, *11*). Consequently, if the interhelical sequences failed to form turns, then the resulting helix would be very long, but would no longer be capable of forming an amphiphilic structure with continuous polar and nonpolar faces.

Construction of a degenerate library of designed sequences. A degenerate library of genes was constructed so that all members of the library encoded protein sequences that fit the generic pattern (Figs. 1 and 2), yet each particular member of the library encoded a different amino acid sequence. The proposed interhelical turns were encoded by constant sequences. However, the polar and nonpolar residues of the α helices were encoded by the degenerate codons NAN and NTN, respectively.

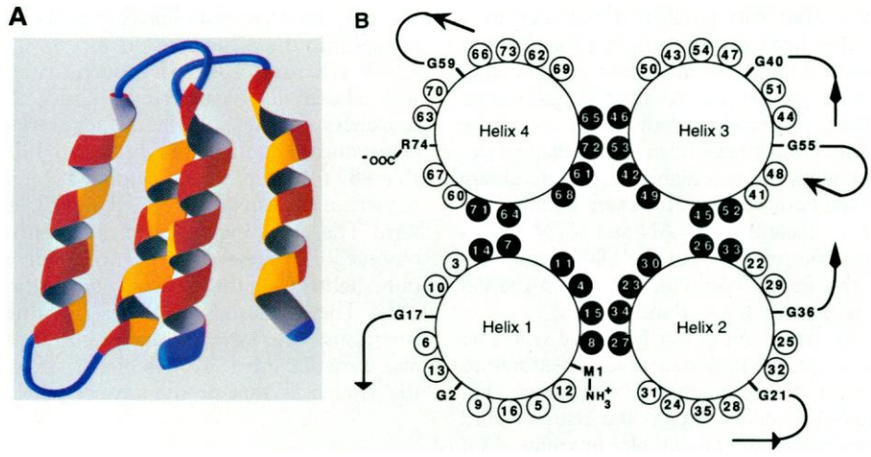


Fig. 1. (**A**) Ribbon diagram (*43*) showing the periodicity of polar (red) and nonpolar (yellow) residues in a four-helix bundle. (**B**) Head-on representation of a four-helix bundle. Individual helices are shown in the helix wheel representation with a repeat of 3.6 residues per turn (*44*). The binary pattern of the design is illustrated by showing generic polar residues as white circles and generic nonpolar residues as black circles (*39*). The identities of the N-cap and C-cap residues at the ends of each helix are shown explicitly. The interhelical turns are represented by arrows.

The degenerate library of genes was constructed by assembling four synthetic oligonucleotides (Fig. 3). Each of these oligonucleotides was designed to encode a segment of protein comprising a single helix and adjacent turn regions. The oligonucleotides were synthesized with a mixture of bases at each of the N positions. However, the interhelical regions were not degenerate and thus could be annealed together to serve as priming sites for the enzymatic synthesis of complementary strands (Fig. 3). These procedures were done twice: once to construct a restriction fragment encoding the first two helices (and adjoining turns) and once to construct a fragment encoding the last two helices (and adjoining turns). The double-stranded DNA was cut with Bgl I. This restriction enzyme was chosen because it generates nonpalindromic "sticky

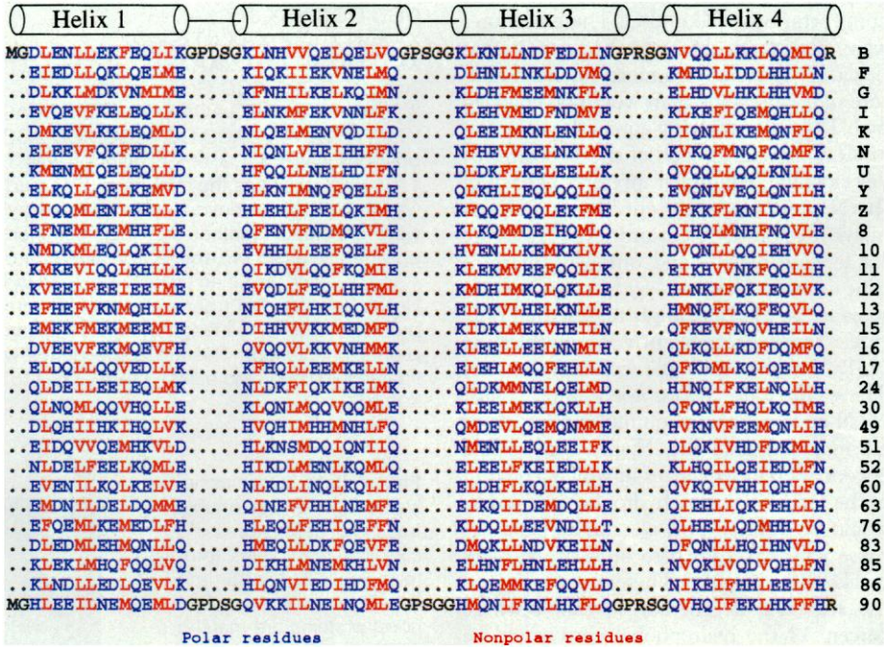


Fig. 2. De novo sequences that have been expressed as soluble and protease-resistant proteins. Within the α helices, the polar and nonpolar residues are shown in blue and red, respectively (*13*). Nondegenerate residues are shown in black. Each sequence is identified by a letter or number shown on the right. Single-letter abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

ends" that can be ligated together in a specific head to tail fashion to generate a gene encoding the entire four-helix bundle. Because the sequence of the full-length gene reflects the semi-random sequences of the original oligonucleotides, this overall procedure gives rise to a highly degenerate library of synthetic genes. However, because the precise locations of NAN and NTN codons were controlled, all of the DNA sequences in this library were consistent with the generic pattern (Figs. 1 and 2).

In the design of our library of synthetic genes we incorporated several features to control the degeneracy of the library. For the NAN (polar) triplets the first and third bases consisted of equimolar mixtures of G, C, and A (*12*). Thymine was not included in the mixture because its presence in the first position would give rise to the stop codons TAA or TAG. By also omitting T in the third position, our collection of sequences was biased to favor some amino acids over others. In particular, glutamine would occur more frequently than histidine, lysine more than asparagine, and glutamic acid more than aspartic acid (*13*). These biases are found in the sequences of natural proteins (*14, 15*) and particularly in the sequences of α-helical proteins (*9*); therefore we also constrained the degeneracy in our library of genes.

For the NTN (nonpolar) triplets, the first base comprised a mixture containing an A:T:C:G molar ratio of 3:3:3:1 (*12*). The third base mixture contained equimolar concentrations of G and C. These mixtures were biased in this way because it was necessary to generate a more balanced mixture of amino acids than would be encoded by NTN codons having equimolar mixtures of all four bases at each of the N positions. For example, equimolar mixtures at each of the N positions would encode six times as many leucines as methionines. (There are six leucine codons and only one methionine codon.) However, by biasing the mixtures as we did, leucine is represented only three times as frequently as methionine (*13*). Another example concerns valine. There are four codons for valine. Thus, NTN codons having equimolar mixtures of all four bases at the N positions would encode protein sequences in which a quarter of the residues in their hydrophobic interiors would be valine. This was undesirable because valine is known to be a strong helix breaker (*9, 10, 16*). Therefore we limited the sequence degeneracy so that only 10 percent of the hydrophobic core would be composed of valine side chains. By including only C and G (rather than all four bases) in the third position, we have biased the library to favor codons used by *Escherichia coli* to express its most abundant proteins (*17*).

After assembly of the full-length DNA (Fig. 3), the degenerate library of genes was cloned into the T7 expression vector pRB-12.65 (*18*) and 108 independent clones were chosen for further study. The DNA sequences of these synthetic genes were determined, and the sequence data revealed that 48 of the 108 clones contained correct sequences consistent with the designed pattern. The overall amino acid composition observed for these 48 sequences differed only slightly from the expected composition (*13*). The remaining 60 clones contained insertions, deletions, or aberrant ligations and were discarded. The 48 clones harboring sequences that fit the correct pattern were then assayed for their ability to express novel proteins.

Expression of stably folded proteins. Because each of our synthetic genes has a different DNA sequence (*19*), each clone has the potential to express a different protein sequence. However, cloning a designed DNA sequence does not ensure expression of a protein that is compact, soluble, and resistant to intracellular proteases. Overexpression of many natural proteins has been hindered by difficulties with folding, stability, or solubility (*20*). In attempting to express a de novo designed sequence, three possible outcomes must be considered:

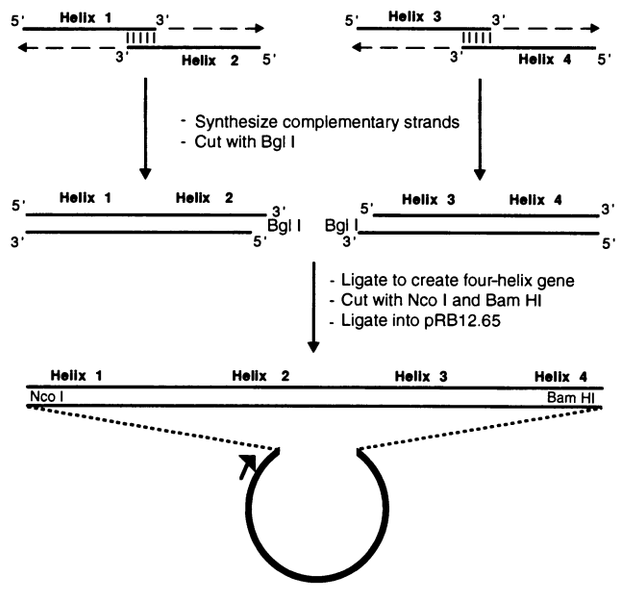
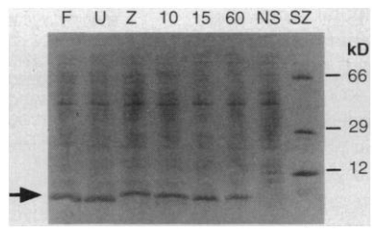


Fig. 3. Construction of a degenerate library of genes designed to encode four-helix bundle proteins. Four deoxyoligonucleotides were synthesized by solid-phase phosphoramidite chemistry (*45*). The sequences of the semirandom oligonucleotides are: Hel-1, 5′ ACG.CTC.CCC.ATG.GGC.XAX.JTZ.XAX.XAX.JTZ.JTZ.XAX.JTZ.XAX.XAX.JTZ.JTZ.XAX.GGT.CCA.GAC.TCT.GGT. 3′; Hel-2, 5′ CAG.TTG.CGC.GCC.TCT.CGG.GCC.YTY.ZAK.ZAK.YTY.YTY.ZAK.YTY.YTY.ZAK.ZAK.YTY.YTY.ZAK.ZAK.YTY.YTY.ZAK.YTY.ACC.AGA.GTC.TGG.ACC. 3′; Hel-3, 5′ C.TGT.CAG.GGC.CCG.AGC.GGC.GGC.XAX.JTZ.XAX.XAX.JTZ.JTZ.XAX.JTZ.XAX.XAX.JTZ.JTZ.XAX.GGT.CCT.CGT.AGC.GGT. 3′; and Hel-4, 5′ GC.AGA.CGG.ATC.CTA.ACG.YTY.ZAK.ZAK.YTY.YTY.ZAK.YTY.YTY.ZAK.ZAK.YTY.YTY.ZAK.YTY.ACC.GCT.ACG.AGG.ACC. 3′. The mixture of T, C, G, and A used at J positions was 3:3:1:3; at K positions it was 3:1::3:3; at X positions the mixture of C, G, and A was 1:1:1; at Y positions the mixture of T, C, and G was 1:1:1; at Z positions the mixture of C and G was 1:1. Oligonucleotides were purified by denaturing PAGE. Subsequently, Hel-1 and Hel-2 were annealed together and their complementary strands were synthesized by Klenow DNA polymerase as shown above. This procedure generated double-stranded DNA coding for helices 1 and 2, which was then cut with Bgl I. An identical set of operations was performed on Hel-3 and Hel-4. Because Bgl I generates nonpalindromic overhangs, ligation produced only head to tail products (that is, full-length genes containing Hel-1 and Hel-2 ligated to Hel-3 and Hel-4 in the proper orientation). The full-length genes were subsequently cut with Nco I and Bam HI and ligated into linearized plasmid pRB12.65 (*18*). The genes were then transformed into *E. coli* strain X90/DE3.



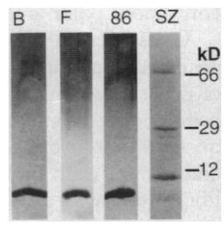


Fig. 4 (left). Coomassie-stained gel of the soluble fraction of cells expressing six different proteins (proteins F, U, Z, 10, 15, and 60). The highly expressed novel proteins are marked with an arrow on the left side of the figure. Lane 7, a control in which a mutation generates a nonsense (NS) at the third codon; lane 8, size markers (SZ): cytochrome c (12 kD), carbonic anhydrase (29 kD), and bovine serum albumin (66 kD). Soluble fractions were prepared as described (*25*). **Fig. 5 (right).** Silver-stained gel of the purified proteins B, F and 86. Size markers are as in Fig. 4.

1) No expression is observed. If the designed sequence does not fold into a stable compact structure then it will be proteolyzed in vivo and will fail to accumulate in the cell (*21*). It is also possible that some proteins fail to be expressed because transcription or translation rates are diminished by particular RNA structures or codon usage patterns.

2) Expression is observed, but the protein forms insoluble inclusion bodies. Insoluble aggregates of misfolded chains are frequently observed when natural proteins are expressed in large amounts (*20, 22*). Although the polypeptide chains sequestered in inclusion bodies are resistant to intracellular proteolysis, no conclusions can be drawn about their folded structures.

3) Expression of soluble protein is observed. The accumulation of a soluble protein requires that it escape degradation by cellular proteases. The ability of a soluble protein to resist proteolysis is far greater if it folds into a compact and stable structure than if it exists as an unfolded polypeptide chain (*21*). Thus the ability of a novel protein to withstand proteolysis in vivo is evidence for the formation of a stable compact structure.

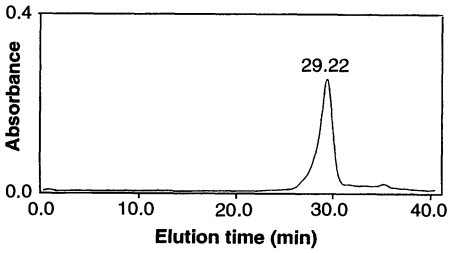


Fig. 6 (left). Size-exclusion chromatography demonstrating that protein F is compact and monomeric. Absorbance at 230 nm is shown as a function of elution time. The column was calibrated using globular proteins whose molecular masses are known from their amino acid sequences. These proteins, their masses, and the observed elution times are: bovine serum albumin (66 kD, 24.39 min), carbonic anhydrase (29 kD, 27.62 min), and cytochrome b562 (12.3 kD, 28.61 min). The novel 74–amino acid protein elutes at 29.22 min. This is later than cytochrome b562, which is a natural four-helix bundle of 106 amino acids. Thus protein F has an apparent molecular mass that is smaller than cytochrome b562. Proteins were chromatographed on a Superose 12 gel filtration column (Pharmacia).

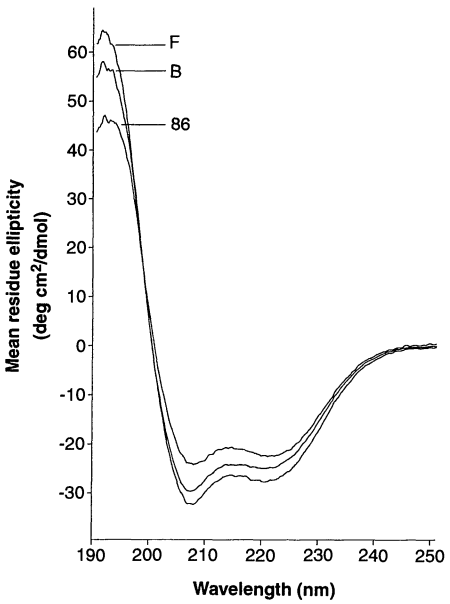


Fig. 7 (right). The CD spectra of proteins B, F, and 86. Spectra were recorded at 20°C in 10 mM $NaPO_4$ (pH 7.5), 50 mM NaF. Protein concentrations were 23, 15, and 16 μM, respectively, as determined by quantitative amino acid analysis (*28*). Spectra were recorded on an Aviv Circular Dichroism Spectrometer model 62 DS with a 1-mm path length cuvette. Readings were taken every 0.2 nm and averaged over a period of 2 seconds.

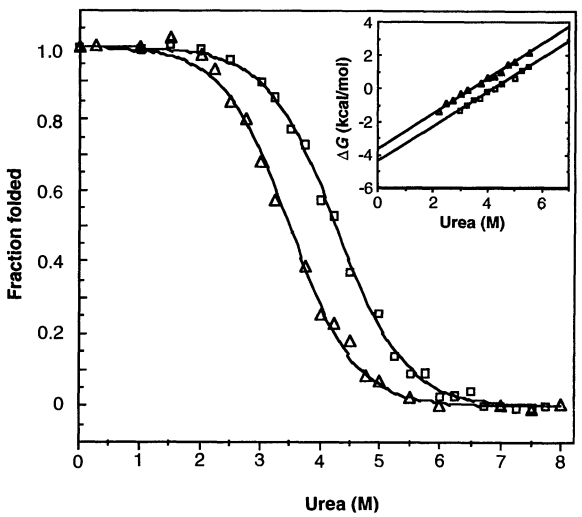


Fig. 8. Urea denaturation curves of proteins B (triangles) and F (squares). Ellipticity at 222 nm was monitored as a function of urea concentration with a 1-cm path length cuvette in an Aviv Circular Dichroism Spectrometer model 62 DS. Spectra were recorded at 20°C in 50 mM $NaPO_4$ (pH 7.5), 100 mM NaCl. To calculate the fraction folded, upper and lower baselines were determined and the formula $f = (ub - s)/(ub - lb)$ was used where f is the fraction folded, ub is the value of the upper baseline, lb is the value of the lower baseline, and s is the measured ellipticity. The ΔG and midpoint values were determined, as shown in the inset, by plotting $-RT\ln K$ as a function of urea concentration and fitting a least squares line after discarding outlying points. The value of K is defined as $(ub - s)/(s - lb)$. The solid curves shown in the main figure were obtained with these ΔG and midpoint values.

We tested all 48 of the correct sequences for expression of novel proteins. Expression was monitored by means of SDS–polyacrylamide gel electrophoresis (PAGE) to assay for a new band of ~8 kD in either the soluble or the insoluble fraction of the cells. Overall, we found that 29 of the 48 sequences (60 percent) expressed protein that was both soluble and resistant to intracellular degradation. Soluble proteins from six of these clones are shown in Fig. 4.

The ability of these designed amino acid sequences to resist proteolytic degradation suggests that they fold into stable globular structures. Although protease resistance is a crude assay for structure, it is nevertheless demanding. For many natural proteins the ability of a structure to resist proteolytic degradation can be completely undermined by single amino acid substitutions that destabilize the native structure (*21*). The thermal stability of the native structure is a key determinant of the proteolytic susceptibility of a protein both in vivo and in vitro (*21, 23*). Thus the accumulation of soluble protein that resists proteolytic degradation provides initial evidence that most of our semirandom sequences (29 out of 48) adopt three-dimensional structures that are both compact and stable.

Purification and characterization of the designed proteins. Three of our designed proteins (denoted B, F, and 86) were purified and assayed for their abilities to form stable α-helical structures (*24*). Protein was overexpressed in the T7 expression system (*25*). The protein of interest was separated from bulk cellular contaminants by a freeze-thaw protocol (*7, 26*). Proteins were then purified by standard procedures, including acid precipitation and cation-exchange chromatography (*27*). Because the proteins isolated from the freeze-thaw procedure were both abundant and relatively free of contaminants, the final purifications were straightforward and the designed proteins were purified to the point where they could be seen as single bands on a silver-stained SDS-PAGE gel (Fig. 5). The purified proteins were subjected to amino acid analysis and mass spectrometry and determined to be covalently intact and not proteolyzed (*28*).

The size and oligomeric state of the purified proteins were determined by size exclusion chromatography. Several natural proteins, known to form compact globular structures were used as calibration standards. Protein F eluted as a single peak from a Superose 12 gel filtration column (Fig. 6). The location of this peak is consistent with the expected molecular mass of this protein (8335 daltons on the basis of its amino acid sequence). All three of the novel proteins eluted later than cytochrome b562, which is itself a monomeric four-helix bundle having a molecular mass of 12,300 (*29*). If the

novel proteins had formed either extended conformations or dimers, then they would have eluted with larger apparent molecular sizes. Therefore, the designed sequences fold into structures that are both compact and monomeric.

The secondary structures of the designed proteins were probed by circular dichroism (CD) spectroscopy. This technique is particularly diagnostic for α-helical proteins, whose spectra include a maximum at 190 nm, a crossover at 200 nm, and negative minima at 208 and 222 nm (*30*). The CD spectra of all three of our purified proteins displayed these α-helical features (Fig. 7). These CD spectra were comparable both in shape and magnitude to the spectra of natural four-helix bundles such as cytochrome b562 (*7, 31*), or growth hormone (*32*). Furthermore, the spectra of proteins B, F, and 86 all indicated helical contents similar to or greater than those measured for the designed proteins α_4 (*3*) or Felix (*2*).

The stabilities of the proteins to denaturation by urea were measured by monitoring the mean residue ellipticity at 222 nm as a function of denaturant concentration (Fig. 8). As would be expected for proteins with different amino acid sequences, proteins B and F unfold with different transition midpoints, 3.5 M and 4.3 M urea, respectively. Protein 86 was only marginally stable, having a midpoint of ~1.5 M urea (*33*). The free energy associated with the transition between the folded and unfolded forms of a protein in the absence of denaturant can be obtained by extrapolating to a urea concentration of 0 molar (*34*). By this method, these two designed proteins are stabilized by 3.7 kcal/mol (B), and 4.4 kcal/mol (F). For comparison, several natural proteins that have similar, or only slightly greater stabilities are α-lactalbumin (4.4 kcal/mol) (*35*), ribonuclease T1 (5.6 kcal/mol), dihydrofolate reductase (5.9 kcal/mol), and staphylococcal nuclease (6.1 kcal/mol) (*36*).

A binary code for protein design. Experimental studies of natural proteins have shown that their structures are remarkably tolerant to amino acid substitution (*37*). This tolerance, however, is limited by a need to maintain the hydrophobicity of interior side chains. Thus, although the information necessary to encode a particular protein fold is highly degenerate, this degeneracy is constrained by a requirement to control the locations of polar and nonpolar residues (*38*). Theoretical work by Dill and colleagues has suggested that the location of hydrophilic and hydrophobic residues in a linear chain provides the major force that drives collapse into a unique compact structure (*39*).

We have built upon these earlier findings to develop a general strategy for protein design. This strategy is based on a simple binary code of polar and nonpolar amino acids: Our sequences were designed to position polar side chains on the surface of a four-helix bundle and nonpolar side chains in the interior. However, the precise identity of each polar or nonpolar residue was allowed to vary extensively. We have begun characterization of a collection of 48 different amino acid sequences whose design was based on this binary code and have shown that most of them fold into structures that are sufficiently compact and stable to escape cellular proteases. Purification and detailed characterization of three of these proteins show that they indeed fold into structures that are compact, monomeric, and predominantly α-helical. Denaturation studies show, not surprisingly, that different sequences give rise to proteins with different stabilities. Of the three proteins purified thus far, two have stabilities comparable to natural proteins.

The sequences of our designed proteins resemble those of natural proteins both because they are complex rather than repetitive and because individual members of the collection differ substantially from one another (for example, protein U contains 15 leucines whereas protein 15 contains only 3). Overall, our sequences contain 15 of the 20 naturally occurring amino acids. Eleven different amino acids are possible in the helices, and an additional four are present in the fixed interhelical regions. The sequences of our novel proteins are also native-like in that they use a hydrophobic-hydrophilic pattern inherent in the genetic code. Furthermore, the binary code we have postulated for de novo design suggests a mechanism for the early evolution of natural proteins.

The results described above indicate that the purified proteins fold into compact structures that contain an abundance of α-helical secondary structure. However, our results do not prove these proteins actually have the appropriate number and orientation of α helices that define a four-helix bundle. Furthermore, we have not yet demonstrated how many of our proteins actually possess native-like interiors. Studies of the folding of natural proteins (*40*) and recent attempts to design proteins (*2–5, 41*) have shown that polypeptide chains can collapse into globular structures that are significantly more flexible than truly native protein structures. These "molten globule" states resemble native proteins in that they are compact and contain significant amounts of secondary structure, yet they differ from native proteins by having interiors that are flexible and less ordered than those of truly native structures (*40*). The flexibilities of interior side chains in a protein can be compared experimentally by multidimensional nuclear magnetic resonance spectroscopy (*42*) and our collection of 29 proteins represents an attractive system for making such comparisons. Although we cannot yet say which of our designed proteins possess flexible interiors, and which are native-like, we anticipate that among our collection of compact stable structures there will be a range of behaviors. Some will probably have relatively flexible interiors, others will be somewhat less flexible, and still others will exhibit the complementary packing of side chains that is typical of the jigsaw puzzle–like interiors of native proteins.

REFERENCES AND NOTES

1. W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959); K. A. Dill, *Biochemistry* **29**, 7133 (1990).
2. M. H. Hecht, J. S. Richardson, D. C. Richardson, R. C. Ogden, *Science* **249**, 884 (1990).
3. L. Regan and W. F. DeGrado, *ibid.* **241**, 976 (1988).
4. T. Sasaki and E. T. Kaiser, *J. Am. Chem. Soc.* **111**, 380 (1988); K. W. Hahn, W. A. Klis, J. M. Stewart, *Science* **248**, 1544 (1990); M. R. Ghadiri, C. Soares, C. Choi, *J. Am. Chem. Soc.* **114**, 4000 (1992).
5. C. G. Unson, B. W. Erickson, D. C. Richardson, J. S. Richardson, *Fed. Proc.* **93**, 1837 (1984); D. G. Osterman and E. T. Kaiser, *J. Cell. Biochem.* **29**, 57 (1985); J. S. Richardson and D. C. Richardson, in *Protein Engineering*, D. L. Oxander and C. F. Fox, Eds. (Liss, New York, 1987), pp. 149–163.
6. P. C. Weber and F. R. Salemme, *Nature* **287**, 82 (1980); J. S. Richardson, *Adv. Protein Chem.* **34**, 167 (1981); C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland, New York, 1991), pp. 33–42.
7. A. P. Brunet *et al.*, *Nature* **364**, 355 (1993).
8. J. S. Richardson and D. C. Richardson, *Science* **240**, 1648 (1988).
9. P. Y. Chou and G. D. Fasman, *Annu. Rev. Biochem.* **47**, 251 (1978); G. D. Fasman, *Prediction of Protein Structure and the Principles of Protein Conformation* (Plenum, New York, 1989).
10. K. T. O'Neil and W. F. DeGrado, *Science* **250**, 646 (1990); P. C. Lyu, M. I. Liff, L. A. Marky, N. R. Kallenbach, *ibid.*, p. 669; M. Blaber, X.-j. Zhang, B. W. Matthews, *ibid.* **260**, 1637 (1993).
11. J. S. Richardson and D. C. Richardson, *Trends Biochem. Sci.* **14**, 304 (1989).
12. These compositions refer to the coding strands for helices 1 and 3. When non-coding strands were synthesized, as was done for helices 2 and 4, complementary mixtures of bases were used (Fig. 3).
13. Overall, at those positions where the NTN codon was used, the expected (and observed) amino acid compositions were (percent): Phe, 15 (15); Leu, 45 (41); Ile, 15 (14); Met, 15 (17); and Val 10 (13). At positions where the NAN codon was used the expected (and observed) amino acid compositions were (percent): His, 11 (12); Gln, 22 (24); Asn, 11 (13); Lys, 22 (18); Asp, 11 (10); and Glu, 22 (23). In our collection of 48 correct sequences representing 3552 amino acids, there were only six "bonus" mutations. None of these bonus mutations inserted a charged residue at a nonpolar position, and no individual sequence contained more than one bonus mutation.
14. P. McCaldon and P. Argos, *Proteins: Struct. Funct. Genet.* **4**, 99 (1988).
15. T. E. Creighton, *Proteins: Structures and Molecular Properties* (Freeman, New York, ed. 2, 1993), p. 4.
16. J. E. Alter, R. H. Andreatta, G. T. Taylor, H. A. Scheraga, *Macromolecules* **6**, 564 (1973); S. Padmanabhan, S. Marqusee, T. Ridgeway, T. M. Laue, R. L. Baldwin, *Nature* **344**, 268 (1990).
17. H. A. De Boer and R. A. Kastelein, in *Maximizing*

Gene Expression, W. Reznikoff and L. Gold, Eds. (Butterworth, Stoneham, MA, 1986), pp. 225–285.
18. R. M. Breyer, A. D. Strosberg, J. G. Guillet, *EMBO J.* **9**, 2679 (1990).
19. For 24 buried positions that can be occupied by any one of five nonpolar residues, and 32 surface positions that can be occupied by any one of six polar residues, a total of $5^{24} \times 6^{32} = 5 \times 10^{41}$ different amino acid sequences are theoretically possible. However, because only 0.1 μmol of DNA was actually synthesized, the total number of sequences in our collection cannot exceed 6×10^{16}, a number 25 orders of magnitude smaller than the number of different amino acid sequences that are theoretically possible. Therefore it is unlikely that any given sequence occurs more than once in our collection.
20. F. A. O. Marston, *Biochem. J.* **240**, 1 (1986).
21. A. A. Pakula, V. B. Young, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8829 (1986); D. A. Parsell and R. T. Sauer, *J. Biol. Chem.* **264**, 7590 (1989).
22. A. Mitraki and J. King, *Bio/Technology* **7**, 690 (1989).
23. M. H. Hecht, J. M. Sturtevant, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5685 (1984).
24. The three proteins were chosen for practical reasons. Cell extracts from 10 candidates were placed in 50 mM sodium acetate buffer (pH 4.0). The proteins in four of these cases remained in solution. These four were thus promising candidates for purification at low pH over cation exchange columns. Mass spectrometry indicated that one of them was not pure after being run on such a column and therefore further characterization was done only on the remaining three. Thus the proteins were not chosen on the basis of stability per se.
25. F. W. Studier and B. A. Moffatt, *J. Mol. Biol.* **189**, 113 (1986); F. W. Studier, A. H. Rosenberg, J. J. Dunn, J. W. Dubendorff, *Methods Enzymol.* **185**, 60 (1990).
26. J. A. Ybe, B. H. Johnson, M. H. Hecht, unpublished material.
27. *Escherichia coli* cells (strain X90 DE3) were grown in 2xYT broth in the presence of ampicillin (100 μg/ml). When the culture achieved an optical density at 600 nm of between 0.7 and 1.0, the cells were induced by the addition of isopropyl-β-D-thiogalactopyranoside to 100 μg/ml, and growth was continued for 3 hours more. Cells were harvested by centrifugation, washed in a buffer containing 50 mM tris-HCl (pH 8.0), 200 mM NaCl, and recentrifuged. Cell pellets were then subjected to three cycles of freeze-thawing (by alternating them between a dry ice, ethanol bath for 5 min and a 10°C water bath for 10 min). The pellets were then gently resuspended in ice-cold 0.5 mM $MgCl_2$ and incubated for 1 hour. The cell debris was removed by centrifugation and discarded. The pH of the supernatant was lowered to 4.0 by the addition of sodium acetate buffer at pH 4.0 to a final concentration of 50 mM. Precipitates appeared and were removed by centrifugation and filtration (Acrodisc; pore size, 0.2 μm; Gelman Sciences). The cleared solution was then separated on an S-Sepharose Fast Flow ion exchange column (Pharmacia) and eluted with a gradient where buffer A was 50 mM sodium acetate, pH 4.0, 0.5 mM EDTA, and buffer B was 50 mM sodium acetate, pH 4.0, 0.5 mM EDTA, 1.0 M NaCl. Peak fractions were pooled, dialyzed into appropriate buffers, and assayed for purity by silver-stained SDS-PAGE, amino acid analysis, and mass spectrometry.
28. Laser desorption mass spectrometry and quantitative amino acid analysis were performed by T. Thannhauser and R. Sherwood at the Cornell Biotechnology Center.
29. E. Itagaki and L. P. Hager, *J. Biol. Chem.* **241**, 3687 (1966).
30. N. Greenfield and G. D. Fasman, *Biochemistry* **8**, 4108 (1969); W. C. Johnson Jr., *Proteins Struct. Funct. Genet.* **7**, 205 (1990).
31. P. A. Bullock and Y. P. Myer, *Biochemistry* **17**, 3084 (1978).
32. L. A. Holladay, R. G. Hammonds Jr., D. Puett, *ibid.* **13**, 1653 (1974).
33. S. Kamtekar and M. H. Hecht, unpublished data.
34. C. N. Pace, *Crit. Rev. Biochem.* **3**, 1 (1975).
35. F. Ahmed and C. C. Bigelow, *Biopolymers* **25**, 1623 (1986).
36. C. N. Pace, *Trends Biochem. Sci.* **15**, 14 (1990).
37. W. A. Lim and R. T. Sauer, *Nature* **339**, 31 (1989); J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, R. T. Sauer, *Science* **247**, 1306 (1990).
38. J. U. Bowie and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152 (1989); J. U. Bowie, N. D. Clarke, C. O. Pabo, R. T. Sauer, *Proteins Struct. Funct. Genet.* **7**, 257 (1990).
39. K. F. Lau and K. A. Dill, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 638 (1990); H. S. Chan and K. A. Dill, *J. Chem. Phys.* **95**, 3775 (1991); K. Yue and K. A. Dill, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 4163 (1992).
40. D. A. Dolgikh, A. P. Kolomiets, I. A. Bolotina, O. B. Ptitsyn, *FEBS Lett.* **165**, 88 (1984); K. Kuwajima, *Proteins: Struct. Funct. Genet.* **6**, 87 (1989); D. N. Brems and H. A. Havel, *ibid.* **5**, 93 (1989); J. Baum, C. M. Dobson, P. A. Evans, C. Hanley, *Biochemistry* **28**, 7 (1989).
41. S. F. Betz, D. P. Raleigh, W. F. DeGrado, *Curr. Opin. Struct. Biol.* **3**, 601 (1993); T. M. Handel, S. A. Williams, W. F. DeGrado, *Science* **261**, 879 (1993); D. P. Raleigh and W. F. DeGrado, *J. Am. Chem. Soc.* **114**, 10079 (1992).
42. L. K. Nicholson *et al.*, *Biochemistry* **31**, 5253 (1992); G. Wagner, *Curr. Opin. Struct. Biol.* **3**, 748 (1993).
43. M. Carson, *J. Appl. Cryst.* **24**, 958 (1991).
44. M. Schiffer and A. B. Edmundson, *Biophys J.* **7**, 121 (1967).
45. M. Caruthers, *Accounts Chem. Res.* **24**, 278 (1991).
46. We thank L. Gillis, M. E. Huffine, and S. Marla for help with the initial experiments; B. Johnson and J. Swan for advice; R. Bryer for plasmid pRB12.65; H. Li and G. Montelione for modeling Fig. 1 (A); and M. Linnell and M. Flocco for oligonucleotide synthesis. Supported by the Whitaker Foundation and by a Beckman Foundation Young Investigator Award (M.H.H.).

11 August 1993; accepted 9 November 1993