J. Mol. Biol. (1997) **273**, 789–796

JMB

COMMUNICATION

De Novo Protein Design: Towards Fully Automated Sequence Selection

Bassil I. Dahiyat[1], Catherine A. Sarisky[1] and Stephen L. Mayo[2]*

[1]*Division of Chemistry and Chemical Engineering and* [2]*Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, Pasadena CA 91125, USA*

**Corresponding author*

Several groups have applied and experimentally tested systematic, quantitative methods to protein design with the goal of developing general design algorithms. We have sought to expand the range of computational protein design by developing quantitative design methods for residues of all parts of a protein: the buried core, the solvent exposed surface, and the boundary between core and surface. Our goal is an objective, quantitative design algorithm that is based on the physical properties that determine protein structure and stability and which is not limited to specific folds or motifs. We chose the ββα motif typified by the zinc finger DNA binding module to test our design methodology. Using previously published sequence scoring functions developed with a combined experimental and computational approach and the Dead-End Elimination theorem to search for the optimal sequence, we designed 20 out of 28 positions in the test motif. The resulting sequence has less than 40% homology to any known sequence and does not contain any metal binding sites or cysteine residues. The resulting peptide, pda8d, is highly soluble and monomeric and circular dichroism measurements showed it to be folded with a weakly cooperative thermal unfolding transition. The NMR solution structure of pda8d was solved and shows that it is well-defined with a backbone ensemble rms deviation of 0.55 Å. Pda8d folds into the desired ββα motif with well-defined elements of secondary structure and tertiary organization. Superposition of the pda8d backbone to the design target is excellent, with an atomic rms deviation of 1.04 Å.

Keywords: protein design; NMR; force field; sequence optimization; Dead-End Elimination

De novo protein design has received considerable attention recently, and significant advances have been made toward the goal of producing stable, well-folded proteins with novel sequences. Several groups have applied and experimentally tested systematic, quantitative methods to protein design with the goal of developing general design algorithms (Hellinga *et al*., 1991; Hurley *et al*., 1992; Desjarlais & Handel, 1995; Harbury *et al*., 1995; Klemba *et al*., 1995; Betz & Degrado, 1996; Dahiyat & Mayo, 1996). To date, these techniques, which screen possible sequences for compatibility with the desired protein fold, have focused mostly on the redesign of protein cores. We have sought to expand the range of computational protein design by developing quantitative design methods for residues of all parts of a protein: the buried core, the solvent exposed surface, and the boundary between core and surface. A critical component of the development of these methods has been their experimental testing and validation. Our goal is an objective, quantitative design algorithm that is based on the physical properties that determine protein structure and stability and which is not limited to specific folds or motifs. This work reports the initial computational and experimental results of combining our core, surface, and boundary methodologies for the design of a small protein motif.

In selecting a motif to test the integration of our design methodologies, we sought a protein fold

Abbreviations used: DEE, Dead-End Elimination; CD, circular dichroism; PDB, Protein Data Bank; TOCSY, total correlation spectroscopy; NOESY, nuclear Overhauser enhancement spectroscopy; NOE, nuclear Overhauser enhancement; DQF-COSY, double quantum-filtered correlation spectroscopy; COSY, correlation spectroscopy; FMOC, 9-fluorenylmethoxycarbonyl; TFA, trifluoroacetic acid.

0022–2836/97/440789–08 $25.00/0/mb971341

that would be small enough to be both computationally and experimentally tractable, yet large enough to form an independently folded structure in the absence of disulfide bonds or metal binding sites. We chose the ββα motif typified by the zinc finger DNA binding module (Pavletich & Pabo, 1991). Though it consists of less than 30 residues, this motif contains sheet, helix, and turn structures. Further, recent work by Imperiali and co-workers who designed a 23 residue peptide, containing an unusual amino acid (D-proline) and a non-natural amino acid (3-(1,10-phenanthrol-2-yl)-L-alanine), that takes this structure has demonstrated the ability of this fold to form in the absence of metal ions (Struthers *et al*., 1996a).

Our design methodology consists of an automated side-chain selection algorithm that explicitly and quantitatively considers specific side-chain to backbone and side-chain to side-chain interactions (Dahiyat & Mayo, 1996). The side-chain selection algorithm screens all possible sequences and finds the optimal sequence of amino acid types and side-chain orientations for a given backbone. In order to correctly account for the torsional flexibility of side-chains and the geometric specificity of side-chain placement, we consider a discrete set of all allowed conformers of each side-chain, called rotamers (Ponder & Richards, 1987). The immense search problem presented by rotamer sequence optimization is overcome by application of the Dead-End Elimination (DEE) theorem (Desmet *et al*., 1992; Goldstein 1994; De Maeyer *et al*., 1997). Our implementation of the DEE theorem extends its utility to sequence design and rapidly finds the globally optimal sequence in its optimal conformation.

In previous work we determined the different contributions of core, surface, and boundary residues to the scoring of a sequence arrangement. The core of a coiled coil and of the streptococcal protein G β1 domain were successfully redesigned using a van der Waals potential to account for steric constraints and an atomic solvation potential favoring the burial and penalizing the exposure of non-polar surface area (Dahiyat & Mayo, 1996, 1997b). Effective solvation parameters and the appropriate balance between packing and solvation terms were found by systematic analysis of experimental data and feedback into the simulation. Solvent exposed residues on the surface of a protein are designed using a hydrogen-bond potential and secondary structure propensities in addition to a van der Waals potential (Dahiyat & Mayo, 1997a). Coiled coils designed with such a scoring function were 10 to 12°C more thermally stable than the naturally occurring analog. Residues that form the boundary between the core and surface require a combination of the core and the surface scoring functions. The algorithm considers both hydrophobic and hydrophilic amino acids at boundary positions, while core positions are restricted to hydrophobic amino acids and surface positions are restricted to hydrophilic amino acids. We use these scoring functions without modification here in order to provide a rigorous test of the generality of our current algorithm.

Sequence design

The sequence selection algorithm requires structure coordinates that define the target motif's backbone. The Brookhaven Protein Data Bank (PDB) (Bernstein *et al*., 1977) was examined for high resolution structures of the ββα motif, and the second zinc finger module of the DNA binding protein Zif268 (PDB code 1zaa) was selected as our design template (Pavletich & Pabo, 1991). The backbone of the second module aligns very closely with the other two zinc fingers in Zif268 and with zinc fingers in other proteins and is therefore representative of this fold class. 28 residues were taken from the crystal structure starting at lysine 33 in the numbering of PDB entry 1zaa which corresponds to our position 1. The first 12 residues comprise the β sheet with a tight turn at the sixth and seventh positions. Two residues connect the sheet to the helix, which extends through position 26 and is capped by the last two residues.

In order to assign the residue positions in the template structure into core, surface or boundary classes, the extent of side-chain burial in Zif268 and the direction of the C^{α}-C^{β} vectors were examined. The small size of this motif limits to one (position 5) the number of residues that can be assigned unambiguously to the core while six residues (positions 3, 12, 18, 21, 22, and 25) were classified as boundary. Three of these residues are from the sheet (positions 3, 5, and 12) and four are from the helix (positions 18, 21, 22, and 25). One of the zinc binding residues of Zif268 is in the core and two are in the boundary, but the fourth, position 8, has a C^{α}-C^{β} vector directed away from the protein's geometric center and is therefore classified as a surface position. The other surface positions considered by the design algorithm are 4, 9, and 11 from the sheet, 15, 16, 17, 19, 20, and 23 from the helix and 14, 27, and 28 which cap the helix ends. The remaining exposed positions, which either were in turns, had irregular backbone dihedrals or were partially buried, were not included in the sequence selection for this initial study. As in our previous studies, the amino acids considered at the core positions during sequence selection were A, V, L, I, F, Y, and W; the amino acids considered at the surface positions were A, S, T, H, D, N, E, Q, K, and R; and the combined core and surface amino acid sets (16 amino acids) were considered at the boundary positions. The scoring functions used were identical to our previous work (Figure 1 legend).

In total, 20 out of 28 positions of the template were optimized during sequence selection. The algorithm first selects Gly for all positions with ϕ angles greater than 0° in order to minimize backbone strain (residues 9 and 27). The 18 remaining residues were split into two sets and optimized separately to speed the calculation. One set con-

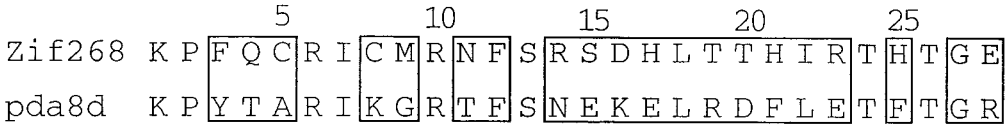

Figure 1. Sequence of pda8d aligned with the second zinc finger of Zif268. The boxed positions were designed using the sequence selection algorithm. The coordinates of PDB record 1zaa (Pavletich & Pabo, 1991) from residues 33 to 60 were used as the structure template. In our numbering, position 1 corresponds to 1zaa position 33. The program BIOGRAF (Molecular Simulations Incorporated, San Diego, CA) was used to generate explicit hydrogen atoms on the structure which was then conjugate gradient minimized for 50 steps using the Dreiding force field (Mayo *et al.*, 1990). As in our previous work (Dahiyat & Mayo, 1997a), a backbone-dependent rotamer library was used (Dunbrack & Karplus, 1993). χ_1 and χ_2 angle values of rotamers for all aromatic amino acids, and χ_1 angle values for all other hydrophobic amino acids were expanded ±1 standard deviation about the mean value reported in the Dunbrack and Karplus library. χ_3 angles that were undetermined from the database statistics were assigned the following values: Arg, −60°, 60°, and 180°; Gln, −120°, -60°, 0°, 60°, 120°, and 180°; Glu, 0°, 60°, and 120°; Lys, −60°, 60°, and 180°. χ_4 angles that were undetermined from the database statistics were assigned the following values: Arg, −120°, −60°, 60°, 120°, and 180°; Lys, −60°, 60°, and 180°. Rotamers with combinations of χ_3 and χ_4 that resulted in sequential g^+/g^- or g^-/g^+ angles were eliminated. All rotamers contained explicit hydrogen atoms and were built with bond lengths and angles from the Dreiding force field. All His rotamers were protonated on both N^δ and N^ε. A Lennard-Jones 12-6 potential with van der Waals radii scaled by 0.9 (Dahiyat & Mayo, 1997b) was used for van der Waals interactions for all residues. An atomic solvation parameter of 23 cal/mol/Å^2 was used to favor hydrophobic burial and to penalize solvent exposure for core and boundary residues (Dahiyat & Mayo, 1996, 1997b). To calculate side-chain non-polar exposure in our optimization framework, we first consider the total hydrophobic area exposed by a rotamer in isolation. This exposure is decreased by the area buried in rotamer/template contacts, and the sum of the areas buried in rotamer/rotamer contacts, quantities that are calculated as pairwise interactions between rotamers as required for DEE. The remaining exposed area is then converted to a penalty energy using a solvation parameter with the same magnitude as for hydrophobic burial but with opposite sign. The Richards definition of solvent-accessible surface area (Lee & Richards, 1971) was used and areas were calculated with the Connolly algorithm (Connolly, 1983). All residues with hydrogen bond donor or acceptors used a hydrogen bond potential based on the potential used in Dreiding but with more restrictive angle-dependent terms to limit the occurrence of unfavorable hydrogen bond geometries (Dahiyat & Mayo, 1997a). A secondary structure propensity potential was used for surface β sheet positions (residues 4 and 11) (Dahiyat & Mayo, 1997a). Propensity values from Serrano and coworkers were used (Munoz & Serrano, 1994). Sequence optimization was performed with a modified version of DEE (Dahiyat & Mayo, 1996). The set consisting of positions 3, 5, 8, 12, 18, 21, 22, and 25 contained 1.2×10^9 possible amino acid sequences and 4.3×10^{19} rotamer sequences. The set consisting of positions 4, 11, 14, 15, 16, 17, 19, 20, 23, and 28 contained 10^{10} possible amino acid sequences and 4.1×10^{23} rotamer sequences. The energy calculations and sequence optimizations took a total of 281 CPU minutes. All calculations were performed on a Silicon Graphics Power Challenge server with 10 R10000 processors running in parallel.

tained the one core, the six boundary positions and position 8 which resulted in 1.2×10^9 possible amino acid sequences corresponding to 4.3×10^{19} rotamer sequences. The other set contained the remaining ten surface residues which had 10^{10} possible amino acid sequences and 4.1×10^{23} rotamer sequences. The two groups do not interact strongly with each other making their sequence optimizations mutually independent, though there are strong interactions within each group. Each optimization was carried out with the non-optimized positions in the template set to the crystallographic coordinates.

The optimal sequences found from the two calculations were combined and are shown in Figure 1 aligned with the sequence from the second zinc finger of Zif268. Even though all of the hydrophilic amino acids were considered at each of the boundary positions, only non-polar amino acids were selected. The calculated seven core and boundary positions form a well-packed buried cluster. The Phe side-chains selected by the algorithm at the zinc binding His positions, 21 and 25, are 80% buried and the Ala at 5 is 100% buried while the Lys at 8 is greater than 60% exposed to solvent (Figure 2). The other boundary positions demonstrate the strong steric constraints on buried residues by packing similar side-chains in an arrangement similar to Zif268 (Figure 2). The calculated optimal configuration buried ~830 Å^2 of nonpolar surface area, with Phe12 (96% buried) and Leu18 (88% buried) anchoring the cluster. On the helix surface, the algorithm positions Asn14 as a helix N-cap with a hydrogen bond between its side-chain carbonyl oxygen and the backbone amide proton of residue 16. The six charged residues on the helix form three pairs of hydrogen bonds, though in our coiled coil designs helical surface hydrogen bonds appeared to be less important than the overall helix propensity of the sequence. Positions 4 and 11 on the exposed sheet surface were selected to be Thr, one of the best β-sheet forming residues (Kim & Berg, 1993; Minor & Kim, 1994; Smith *et al.*, 1994).

Combining the 20 designed positions with the Zif268 amino acids at the remaining eight sites results in a peptide with overall 39% (11/28) homology to Zif268, which reduces to 15% (3/20) homology when only the designed positions are considered. A BLAST (Altschul *et al.*, 1990) search of the non-redundant protein sequence database of the National Center for Biotechnology Information

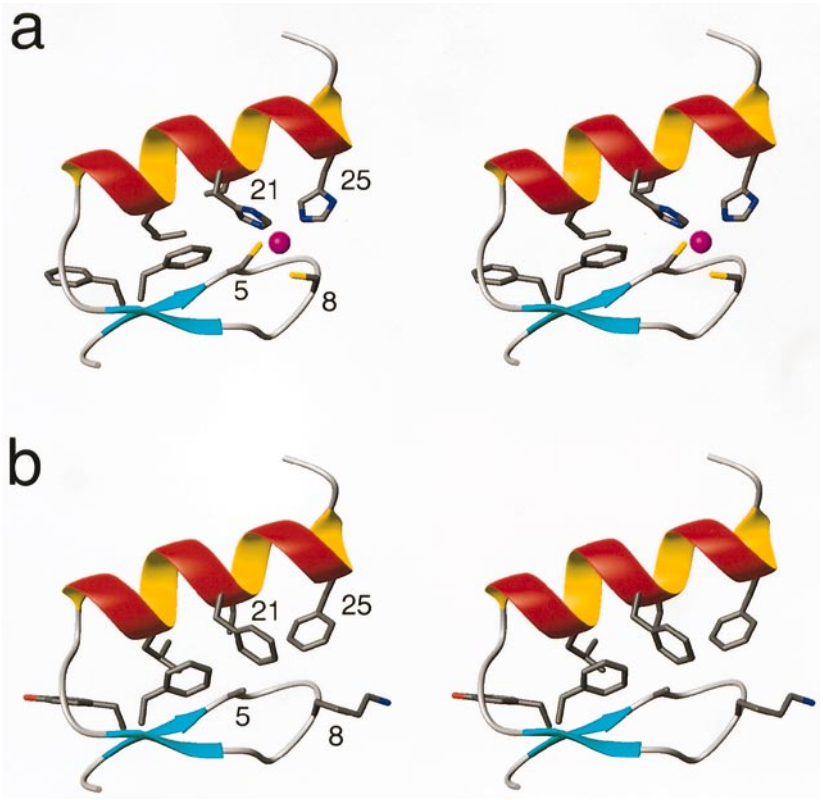

Figure 2. Comparison of Zif268 and calculated pda8d structures. For clarity, only side-chains from residues 3, 5, 8, 12, 18, 21, 22, and 25 are shown. a, Stereo diagram of Zif268 showing its buried residues and zinc binding site. b, Stereo diagram of the calculated pda8d side-chain orientations showing the same residue positions as in a. Diagrams were made with MOLMOL (Koradi *et al.*, 1996).

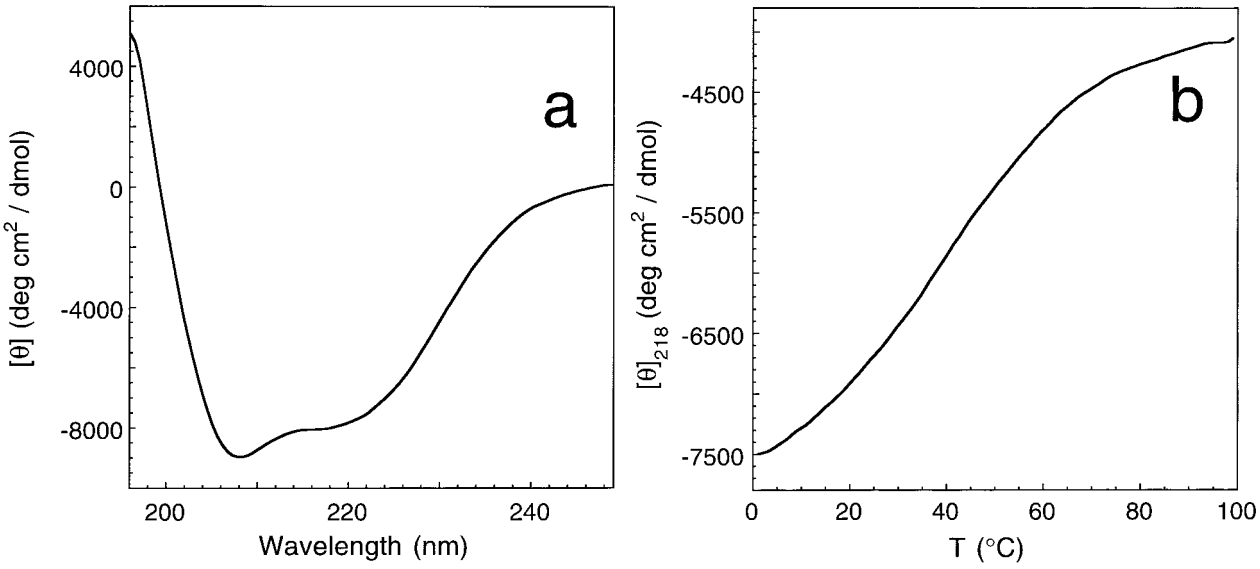

Figure 3. CD measurements of pda8d. a, Far UV CD spectrum of pda8d. Protein concentration was 43 μM in 50 mM sodium phosphate at pH 5.0. The spectrum was acquired at 1°C in a 1 mm cuvette and was baseline corrected with a buffer blank. The spectrum is the average of three scans using a one second integration time and 1 nm increments. All CD data were acquired on an Aviv 62DS spectrometer equipped with a thermoelectric temperature control unit. b, Thermal unfolding of pda8d monitored by CD. Protein concentration was 115 μM in 50 mM sodium phosphate at pH 5.0. Unfolding was monitored at 218 nm in a 1 mm cuvette using 2deg. increments with an averaging time of 40 seconds and an equilibration time of 120 seconds per increment. Reversibility was confirmed by comparing 1°C CD spectra from before and after heating to 99°C. Peptide concentrations were determined by UV spectrophotometry. Pda8d was synthesized using standard solid phase FMOC chemistry on an Applied Biosystems 433A automated peptide synthesizer. The peptide was cleaved from the resin with TFA and purified by reversed phase high performance liquid chromatography on a Vydac C8 column (25 cm × 10 mm) with a linear acetonitrile-water gradient containing 0.1% TFA. Peptide was lyophilized and stored at −20°C. Matrix assisted laser desorption mass spectroscopy yielded a molecular weight of 3363 daltons (3362.8 calculated).

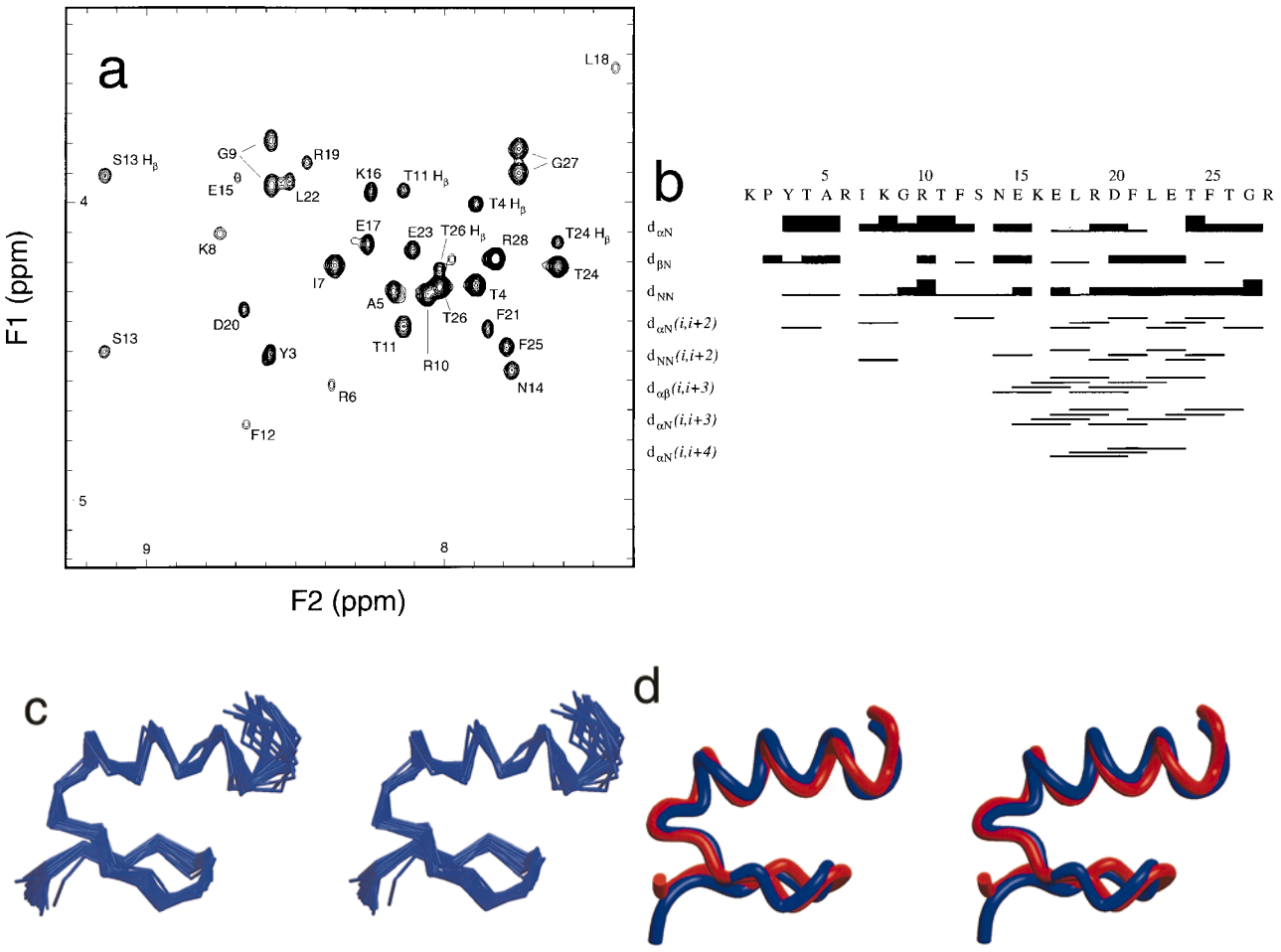

Figure 4. NMR spectra and solution structure of pda8d. a, TOCSY H^{α}-HN fingerprint region of pda8d. NMR data were collected on a Varian Unityplus 600 MHz spectrometer equipped with a Nalorac inverse probe with a self-shielded *z*-gradient. NMR samples were prepared in 90/10 $H_2O/^2H_2O$ or 99.9% ^2H_2O with 50 mM sodium phosphate at pH 5.0. Sample pH was adjusted using a glass electrode with no correction for the effect of ^2H_2O on measured pH. All spectra for assignments were collected at 7°C. Sample concentration was approximately 2 mM. NMR assignments were based on standard homonuclear methods using DQF-COSY, NOESY and TOCSY spectra (Wuthrich, 1986). NOESY and TOCSY spectra were acquired with 2048 points in F_2 and 512 increments in F_1 and DQF-COSY spectra were acquired with 4096 points in F_2 and 1024 increments in F_1. All spectra were acquired with a spectral width of 7500 Hz and 32 transients. NOESY spectra were recorded with mixing times of 100 and 200 ms and TOCSY spectra were recorded with an isotropic mixing time of 80 ms. In TOCSY and DQF-COSY spectra water suppression was achieved by presaturation during a relaxation delay of 1.5 and 2.0 seconds, respectively. Water suppression in the NOESY spectra was accomplished with the WATERGATE pulse sequence (Piotto *et al.*, 1992). Chemical shifts were referenced to the HO^2H resonance. Spectra were zero-filled in both F_2 and F_1 and apodized with a shifted gaussian in F_2 and a cosine bell in F_1 (NOESY and TOCSY) or a 30° shifted sine bell in F_2 and a shifted gaussian in F_1 (DQF-COSY). Water-sLED experiments (Altieri *et al.*, 1995) were run at 25 °C at 1.5 mM, 400 µM and 100 µM in 99.9% ^2H_2O with 50 mM sodium phosphate at pH 5.0. Axial gradient field strength was varied from 3.26 to 53.1 G/cm and a diffusion time of 50 ms was used. Spectra were processed with 2 Hz line broadening and integrals of the aromatic and high field aliphatic protons were calculated and fit to an equation relating resonance amplitude to gradient strength in order to extract diffusion coefficients (Altieri *et al.*, 1995). Diffusion coefficients were 1.48×10^{-7}, 1.62×10^{-7} and 1.73×10^{-7} cm^2/s at 1.5 mM, 400 µM and 100 µM, respectively. The diffusion coefficient for the zinc finger monomer control was 1.72×10^{-7} cm^2/s and for protein G β1 was 1.49×10^{-7} cm^2/s. b, NMR assignments summary and NOE connectivities of pda8d. Bars represent unambiguous connectivities and the bar thickness of the sequential connections is indexed to the intensity of the resonance. c, Solution structure of pda8d. Stereoview showing the best fit superposition of the 32 converged simulated annealing structures from X-PLOR (Brünger, 1992). The backbone C^{α} trace is shown in blue. The amino terminus is at the lower left of the Figure and the carboxy terminus is at the upper right of the figure. The structure consists of two antiparallel strands from positions 3 to 6 (back strand) and 9 to 12 (front strand), with a hairpin turn at residues 7 and 8, followed by a helix from positions 15 to 26. The termini, residues 1, 2, 27, and 28, have very few NOE restraints and are disordered. NOEs were classified into three distance-bound ranges based on cross-peak intensity: strong (1.8 to 2.7 Å), medium (1.8 to 3.3 Å) and weak (1.8 to 5.0 Å). Upper bounds for restraints involving methyl protons were increased by 0.5 Å to account for the increased intensity of methyl resonances. All partially overlapped NOEs were set to weak restraints. Standard hybrid distance geometry-simulated annealing protocols were followed (Nilges *et al.*, 1988, 1991; Kuszewski *et al.*, 1992). 98 distance geometry structures were generated and, following regularization and refinement, resulted in an ensemble of 32 structures with no restraint violations greater than 0.3 Å, rms deviations from idealized bond lengths less than 0.01 Å and rms deviations from idealized bond angles and impropers less than 1°. Coordinates will be deposited with the Brookhaven Protein Data Bank and are available from the authors on request until processed and released. d, Comparison of pda8d solution structure and the design target. Stereoview of the best fit superposition of the average NMR structure of pda8d (blue) and the backbone of Zif268 (red). Residues 3 to 26 were used in the fit.

Table 1. NMR structure determination of pda8d: distance restraints, structural statistics, atomic root-mean-square (rms) deviations, and comparison to the design target.

Distance restraints	
Intraresidue	148
Sequential	94
Short range (|i-j| = 2-5 residues)	78
Long range (|i-j| > 5 residues)	34
Total	354
Structural statistics	
	⟨SA⟩ ±SD
Rms deviation from distance restraints (Å)	0.049 ± .004
Rms deviation from idealized geometry (Å)	
Bonds (Å)	0.0051 ± 0.0004
Angles (degrees)	0.76 ± 0.04
Impropers (degrees)	0.56 ± 0.04
Atomic rms deviations (Å)[a]	
	⟨SA⟩ *versus* SA ±SD
Backbone	0.55 ±0.03
Backbone + non-polar side-chains	1.05 ±0.06
Heavy atoms	1.25 ±0.04
Atomic rms deviations between pda8d and the design target (Å)[a]	
	SA *versus* target
Backbone	1.04
Heavy atoms	2.15

⟨SA⟩ are the 32 simulated annealing structures, SA is the average structure and SD is the standard deviation. The design target is the backbone of Zif268.

[a] Atomic rms deviations are for residues 3 to 26, inclusive. The termini, residues 1, 2, 27, and 28, were highly disordered and had very few non-sequential or non-intraresidue contacts.

finds weak homology, less than 40%, to several zinc finger proteins and fragments of other unrelated proteins. None of the alignments had significance values less than 0.26. By objectively selecting 20 out of 28 residues on the Zif268 template, a peptide with little homology to known proteins and no zinc binding site was designed.

Experimental characterization

The far UV circular dichroism (CD) spectrum of the designed molecule, pda8d, shows a maximum at 195 nm and minima at 218 nm and 208 nm, which is indicative of a folded structure (Figure 3a). The thermal melt is weakly cooperative, with an inflection point at 39°C, and is completely reversible (Figure 3b). The broad melt is consistent with a low enthalpy of folding which is expected for a motif with a small hydrophobic core. This behavior contrasts the uncooperative transitions observed for other short peptides (Weiss & Keutmann, 1990; Scholtz *et al.*, 1991; Struthers *et al.*, 1996b).

Sedimentation equilibrium studies at 100 μM and both 7°C and 25°C give a molecular mass of 3490, in good agreement with the calculated mass of 3362, indicating the peptide is monomeric. At concentrations greater than 500 μM, however, the data do not fit well to an ideal single species model. When the data were fit to a monomer-dimer-tetramer model, dissociation constants of 0.5 to 1.5 mM for monomer-to-dimer and greater than 4 mM for dimer-to-tetramer were found, though the interaction was too weak to accurately measure these values. Diffusion coefficient measurements using the water-sLED pulse sequence (Altieri *et al.*, 1995) agreed with the sedimentation results: at 100 μM pda8d has a diffusion coefficient close to that of a monomeric zinc finger control, while at 1.5 mM the diffusion coefficient is similar to that of protein G β1, a 56 residue protein. The CD spectrum of pda8d is concentration independent from 10 μM to 2.6 mM. NMR COSY spectra taken at 2.1 mM and 100 μM were almost identical with five of the H^{α}-HN cross-peaks shifted no more than 0.1 ppm and the rest of the cross-peaks remaining unchanged. These data indicate that pda8d undergoes a weak association at high concentration, but this association has essentially no effect on the peptide's structure.

The NMR chemical shifts of pda8d are well dispersed, suggesting that the protein is folded and well-ordered. The H^{α}-HN fingerprint region of the TOCSY spectrum is well-resolved with no overlapping resonances (Figure 4a) and all of the H^{α} and HN resonances have been assigned. All unambiguous sequential and medium-range NOEs are shown in Figure 4b. H^{α}-HN and/or HN-HN NOEs were found for all pairs of residues except R6-I7 and K16-E17, both of which have degenerate HN chemical shifts, and P2-Y3 which have degenerate H^{α} chemical shifts. An NOE is present, however, from a P2 H^{δ} to the Y3 HN analogous to sequential HN-HN connections. Also, strong K1 H^{α} to P2 H^{δ} NOEs are present and allowed completion of the resonance assignments (see Supplementary Material).

The structure of pda8d was determined using 354 NOE restraints (12.6 restraints per residue) that were non-redundant with covalent structure. An

ensemble of 32 structures (Figure 4c) was obtained using X-PLOR (Brünger, 1992) with standard protocols for hybrid distance geometry-simulated annealing. The structures in the ensemble had good covalent geometry and no NOE restraint violations greater than 0.3 Å. As shown in Table 1, the backbone was well defined with a root-mean-square (rms) deviation from the mean of 0.55 Å when the disordered termini (residues 1, 2, 27, and 28) were excluded. The rms deviation for the backbone (3 to 26) plus the buried side-chains (residues 3, 5, 7, 12, 18, 21, 22, and 25) was 1.05 Å.

The NMR solution structure of pda8d shows that it folds into a ββα motif with well-defined secondary structure elements and tertiary organization which match the design target. A direct comparison of the design template, the backbone of the second zinc finger of Zif268, to the pda8d solution structure highlights their similarity (Figure 4d). Alignment of the pda8d backbone to the design target is excellent, with an atomic rms deviation of 1.04 Å (Table 1). Pda8d and the design target correspond throughout their entire structures, including the turns connecting the secondary structure elements.

In conclusion, the experimental characterization of pda8d shows that it is folded and well-ordered with a weakly cooperative thermal transition, and that its structure is an excellent match to the design target. To our knowledge, pda8d is the shortest sequence of naturally occurring amino acids that folds to a unique structure without metal binding, oligomerization or disulfide bond formation (McKnight *et al*., 1996). The successful design of pda8d supports the use of objective, quantitative sequence selection algorithms for protein design. Also, this work is an important step towards the goal of the successful automated design of a complete protein sequence. Though our algorithm requires a template backbone as input, recent work indicates that it is not sensitive to even fairly large perturbations in backbone geometry (Su & Mayo, 1997). This robustness suggests that the algorithm can be used to design sequences for *de novo* backbones.

Acknowledgments

We thank Scott Ross for assistance with NMR studies, Pak Poon of the UCLA Molecular Biology Institute for sedimentation equilibrium studies, and Gary Hathaway of the Caltech Protein and Peptide Microanalytical Laboratory for mass spectra. We acknowledge financial support from the Rita Allen Foundation, the David and Lucile Packard Foundation and the Searle Scholars Program/The Chicago Community Trust. B.I.D. is partially supported by NIH Training Grant GM 08346.

References

Altieri, A. S., Hinton, D. P. & Byrd, R. A. (1995). Association of biomolecular systems *via* pulsed field gradient NMR self-diffusion measurements. *J. Am. Chem. Soc.* **117**, 7566–7567.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr, Brice, M. D. & Rodgers, J. R., *et al*. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Betz, S. F. & Degrado, W. F. (1996). Controlling topology and native-like behavior of de novo-designed peptides: design and characterization of antiparallel 4–stranded coiled coils. *Biochemistry,* **35**, 6955–6962.

Brünger, A. T. (1992). *X-PLOR Version 3.1 A system for X-ray Crystallography and NMR,* Yale University Press, New Haven.

Connolly, M. L. (1983). Solvent accessible surfaces of proteins and nucleic acids. *Science,* **221**, 709–713.

Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci.* **5**, 895–903.

Dahiyat, B. I. & Mayo, S. L. (1997a). Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333–1337.

Dahiyat, B. I. & Mayo, S. L. (1997b). Probing the role of packing specificity in protein design. *Proc. Natl Acad. Sci. USA,* in the press.

De Maeyer, M., Desmet, J. & Laster, I. (1997). All in one: a highly detailed rotamer library improves both accuracy and speed in the modeling of sidechains by dead-end elimination. *Folding Design,* **2**, 53–66.

Desjarlais, J. R. & Handel, T. M. (1995). De novo design of the hydrophobic cores of proteins. *Protein Sci.* **4**, 2006–2018.

Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature,* **356**, 539–542.

Dunbrack, R. L. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins: an application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574.

Goldstein, R. F. (1994). Efficient rotamer elimination applied to protein side-chains and related spin-glasses. *Biophys. J.* **66**, 1335–1340.

Harbury, P. B., Tidor, B. & Kim, P. S. (1995). Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc. Natl Acad. Sci. USA,* **92**, 8408–8412.

Hellinga, H. W., Caradonna, J. P. & Richards, F. M. (1991). Construction of new ligand-binding sites in proteins of known structure 2. Grafting of buried transition-metal binding site into *Escherichia coli* thioredoxin. *J. Mol. Biol.* **222**, 787–803.

Hurley, J. H., Baase, W. A. & Matthews, B. W. (1992). Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J. Mol. Biol.* **224**, 1143–1154.

Kim, C. W. A. & Berg, J. M. (1993). Thermodynamic β-sheet forming propensities measured using a zinc finger host peptide. *Nature,* **362**, 267–270.

Klemba, M., Gardner, K. H., Marino, S., Clarke, N. D. & Regan, L. (1995). Novel metal-binding proteins by design. *Nature Struct. Biol.* **2**, 368–373.

Koradi, R., Billeter, M. & Wuthrich, K. (1996). Molmol: a program for the display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.

Kuszewski, J., Nilges, M. & Brunger, A. T. (1992). Sampling and efficiency of metrix matrix distance

geometry: a novel ''partial'' metrization algorithm. *J. Biomol. NMR,* **2**, 33–56.
Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
Mayo, S. L., Olafson, B. D. & Goddard, W. A., III (1990). Dreiding: a generic forcefield for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909.
McKnight, C. J., Doering, D. S., Matsudaira, P. T. & Kim, P. S. (1996). A thermostable 35-residue subdomain within villin headpiece. *J. Mol. Biol.* **260**, 126–134.
Minor, D. L. & Kim, P. S. (1994). Measurement of the β-sheet-forming propensities of amino acids. *Nature,* **367**, 660–663.
Munoz, V. & Serrano, L. (1994). Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. *Proteins: Struct. Funct. Genet.* **20**, 301–311.
Nilges, M., Clore, G. M. & Gronenborn, A. M. (1988). Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Letters,* **229**, 317–324.
Nilges, M., Kuszewski, J. & Brünger, A. T. (1991). Sampling properties of simulated annealing and distance geometry. In *Computational Aspects of the Study of Biological Macromolecules by NMR* (Hoch, J. C., Poulsen, F. M. & Redfield, C., eds), pp. 451–457, Plenum Press, New York.
Pavletich, N. P. & Pabo, C. O. (1991). Zinc finger DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science,* **252**, 809–817.
Piotto, M., Saudek, V. & Sklenar, V. (1992). Gradient tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR,* **2**, 661–665.
Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins-use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775–791.
Scholtz, J. M., Marqusee, S., Baldwin, R. L., York, E. J., Stewart, J. M. & Santoro, M., *et al.* (1991). Calorimetric determination of the enthalpy change for the alpha- helix to coil transition of an alanine peptide in water. *Proc. Natl Acad. Sci. USA,* **88**, 2854–2858.
Smith, C. K., Withka, J. M. & Regan, L. (1994). A thermodynamic scale for the β-sheet-forming tendencies of amino acids. *Biochemistry,* **33**, 5510–5517.
Struthers, M. D., Cheng, R. P. & Imperiali, B. (1996a). Design of a monomeric 23-residue polypeptide with defined tertiary structure. *Science,* **271**, 342–345.
Struthers, M. D., Cheng, R. P. & Imperiali, B. (1996b). Economy in protein design: evolution of a metal-independent ββα motif based on the zinc finger domains. *J. Am. Chem. Soc.* **118**, 3073–3081.
Su, A. & Mayo, S. L. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.* **6**, 1701–1707.
Weiss, M. A. & Keutmann, H. T. (1990). Alternating zinc finger motifs in the male-associated protein ZFY: defining architectural rules by mutagenesis and design of an aromatic swap second-site revertant. *Biochemistry,* **29**, 9808–9813.
Wuthrich, K. (1986). *NMR of Proteins and Nucleic Acids,* John Wiley and Sons, New York.

Edited by P. E. Wright

(Received 22 April 1997; received in revised form 7 August 1997; accepted 7 August 1997)

http://www.hbuk.co.uk/jmb

Supplementary material for this paper is available from JMB Online.