De Novo Protein Design: Fully Automated Sequence Selection

Bassil I. Dahiyat† and Stephen L. Mayo*

The first fully automated design and experimental validation of a novel sequence for an entire protein is described. A computational design algorithm based on physical chemical potential functions and stereochemical constraints was used to screen a combinatorial library of 1.9×10^{27} possible amino acid sequences for compatibility with the design target, a ββα protein motif based on the polypeptide backbone structure of a zinc finger domain. A BLAST search shows that the designed sequence, full sequence design 1 (FSD-1), has very low identity to any known protein sequence. The solution structure of FSD-1 was solved by nuclear magnetic resonance spectroscopy and indicates that FSD-1 forms a compact well-ordered structure, which is in excellent agreement with the design target structure. This result demonstrates that computational methods can perform the immense combinatorial search required for protein design, and it suggests that an unbiased and quantitative algorithm can be used in various structural contexts.

Significant advances have been made toward the design of stable, well-folded proteins with novel sequences (*1*). These efforts have generated insight into the factors that control protein folding and have suggested new approaches to biotechnology (*2*). In order to broaden the scope and power of protein design techniques, several groups have developed and experimentally tested systematic quantitative methods for protein design directed toward developing general design algorithms (*3*, *4*). These techniques, which have been used to screen possible sequences for compatibility with the desired protein fold, have been focused mostly on the redesign of protein cores.

We have sought to expand the range of computational protein design to residues of all parts of a protein: the buried core, the solvent-exposed surface, and the boundary between core and surface (*4*–*6*). Our goal is an unbiased, quantitative design algorithm that is based on the physical properties that determine protein structure and stability and that is not limited to specific folds or motifs. Such a method should escape the lack of generality of design approaches based on system-specific heuristics or subjective considerations or both. We have developed our algorithm by combining theory, computation, and experiment in a cycle that has improved our understanding of the physical chemistry governing protein design (*4*). We now report the successful design by the algorithm of an original sequence for an entire protein and the experimental validation of the protein's structure.

Sequence selection. Our design methodology begins with a backbone fold and we attempt to select an amino acid sequence that will stabilize this target structure. The method consists of an automated side-chain selection algorithm that explicitly and quantitatively considers specific interactions between (i) side chain and backbone and (ii) side chain and side chain (*4*). The side chain selection algorithm screens all possible amino acid sequences and finds the optimal sequence 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 (*7*). The sizable search problem presented by rotamer sequence optimization is overcome by application of the dead-end elimination (DEE) theorem (*8*). Our implementation of the DEE theorem extends its utility to sequence design and rapidly finds the globally optimal sequence in its optimal conformation (*4*).

Previously we determined the different contributions of core, surface, and boundary residues to the scoring of a sequence arrangement. The sequence predictions of a scoring function, or a combination of scoring functions, were experimentally tested in order to assess the accuracy of the algorithm and to derive improvements to it. We successfully redesigned the core of a coiled coil and of the streptococcal protein G β1 (Gβ1) domain using a van der Waals potential to account for steric constraints and an atomic solvation potential favoring the burial and penalizing the exposure of nonpolar surface area (*4*, *6*). 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 were designed with the use of a hydrogen-bond potential and secondary structure propensities in addition to a van der Waals potential. Coiled coils designed with such a scoring function were 10° to 12°C more thermally stable than the naturally occurring analog (*5*). 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, whereas core positions are restricted to hydrophobic amino acids and surface positions are restricted to hydrophilic amino acids.

In order to assess the capability of our design algorithm, we have computed the entire amino acid sequence for a small protein motif. We sought a protein fold 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. We chose the ββα motif typified by the zinc finger DNA binding module (*9*). Although this motif consists of fewer than 30 residues, it does contain sheet, helix, and turn structures. The ability of this fold to form in the absence of metal ions or disulfide bonds has been demonstrated by Imperiali and co-workers, who designed a 23-residue peptide, containing an unusual amino acid (D-proline) and a nonnatural amino acid [3-(1,10-phenanthrol-2-yl)-L-alanine], which achieved this fold (*10*); our initial characterization of a partially computed sequence indicated that it also forms this fold (*11*). In computing the full sequence for this target fold, we use the scoring functions from our previous work without modification (*12*). The ββα motif was not used in any of our prior work to develop the design methodology and therefore provides a test of the algorithm's generality.

The sequence selection algorithm requires structure coordinates that define the target motif's backbone (N, Cα, C, and O atoms and Cα-Cβ vectors). The Brookhaven Protein Data Bank (PDB) (*13*) was examined for high-resolution structures of the ββα motif, and the second zinc finger module of the DNA binding protein Zif268 was selected as our design template (*9*, *14*). In order to assign the residue positions in the template structure into core, surface, or boundary classes, the orientation of the Cα-Cβ vectors was assessed relative to a solvent-accessible surface computed with only the template Cα atoms (*15*). The small size of this motif limits to one

B I. Dahiyat, Division of Chemistry and Chemical Engineering, California Institute of Technology, mail code 147-75, Pasadena, CA 91125, USA.
S. L. Mayo, Howard Hughes Medical institute and Division of Biology, California Institute of Technology, mail code 147-75, Pasadena, CA 91125, USA.
*To whom correspondence should be addressed. E-mail: steve@mayo.caltech.edu
†Present address: Xencor, Pasadena, CA 91106, USA.

(position 5) the number of residues that can be assigned unambiguously to the core, whereas seven residues (positions 3, 7, 12, 18, 21, 22, and 25) were classified as boundary and the remaining 20 residues were assigned to the surface. Whereas three of the zinc binding positions of Zif268 are in the boundary or core, one residue, position 8, has a Cα-Cβ vector directed away from the geometric center of the protein and is classified as a surface position. As in our previous studies, the amino acids considered at the core positions during sequence selection were Ala, Val, Leu, Ile, Phe, Tyr, and Trp; the amino acids considered at the surface positions were Ala, Ser, Thr, His, Asp, Asn, Glu, Gln, Lys, and Arg; and the combined core and surface amino acid sets (16 amino acids) were considered at the boundary positions. Two of the residue positions (9 and 27) have φ angles greater than 0° and are set to Gly by the sequence selection algorithm to minimize backbone strain.

The total number of amino acid sequences that must be considered by the design algorithm is the product of the number of possible amino acids at each residue position. The ββα motif residue classification described above results in a virtual combinatorial library of 1.9×10^{27} possible amino acid sequences (*16*). This library size is 15 orders of magnitude larger than that accessible by experimental random library approaches. A corresponding peptide library consisting of only a single molecule for each 28-residue sequence would have a mass of 11.6 metric tons (*17*). In order to accurately model the geometric specificity of side-chain placement, we explicitly consider the torsional flexibility of amino acid side chains in our sequence scoring by representing each amino acid with a discrete set of allowed conformations, called rotamers (*18*). As a result, the design algorithm must consider all rotamers for each possible amino acid at each residue position. The total size of the search space for the ββα motif is therefore 1.1×10^{62} possible rotamer sequences. We use a search algorithm based on an extension of the DEE theorem to solve the rotamer sequence optimization problem (*4*, *8*). Efficient implementation of the DEE theorem has made complete protein sequence design tractable for about 50 residues on current parallel computers in a single calculation. The rotamer optimization problem for the ββα motif required 90 CPU hours to find the optimal sequence (*19*, *20*).

The optimal sequence (Fig. 1) is called full sequence design (FSD-1). Even though all of the hydrophilic amino acids were considered at each of the boundary positions, the algorithm selected only nonpolar amino acids. The eight core and boundary positions are predicted to form a well-packed buried cluster. The Phe side chains selected by the algorithm at positions 21 and 25, the zinc-binding His positions of Zif268, are more than 80 percent buried, and the Ala at position 5 is 100 percent buried but the Lys at position 8 is more than 60 percent exposed to solvent (Fig. 2). The other boundary positions demonstrate the steric constraints on buried residues by packing similar side chains in an arrangement similar to that of Zif268 (Fig. 2). The calculated optimal configuration for core and boundary residues buries ~1150 Å^2 of nonpolar surface area. On the helix surface, the algorithm places Asn^{14} with a hydrogen bond between its side-chain carbonyl oxygen and the backbone amide proton of residue 16. The eight charged residues on the helix form three pairs of hydrogen bonds, although in our coiled-coil designs, helical surface hydrogen bonds appeared to be less important than the overall helix propensity of the sequence (*5*). Positions 4 and 11 on the exposed sheet surface were selected by the program to be Thr, one of the best β-sheet forming residues (*21*).

Alignment of the sequences for FSD-1 and Zif268 (Fig. 1) indicates that only 6 of the 28 residues (21 percent) are identical and only 11 (39 percent) are similar. Four of the identities are in the buried cluster, which is consistent with the expectation that buried residues are more conserved than solvent-exposed residues for a given motif (*22*). A BLAST (*23*) search of the FSD-1 sequence against the nonredundant protein sequence database of the National Center for Biotechnology Information did not reveal any zinc finger protein sequences. Fur-

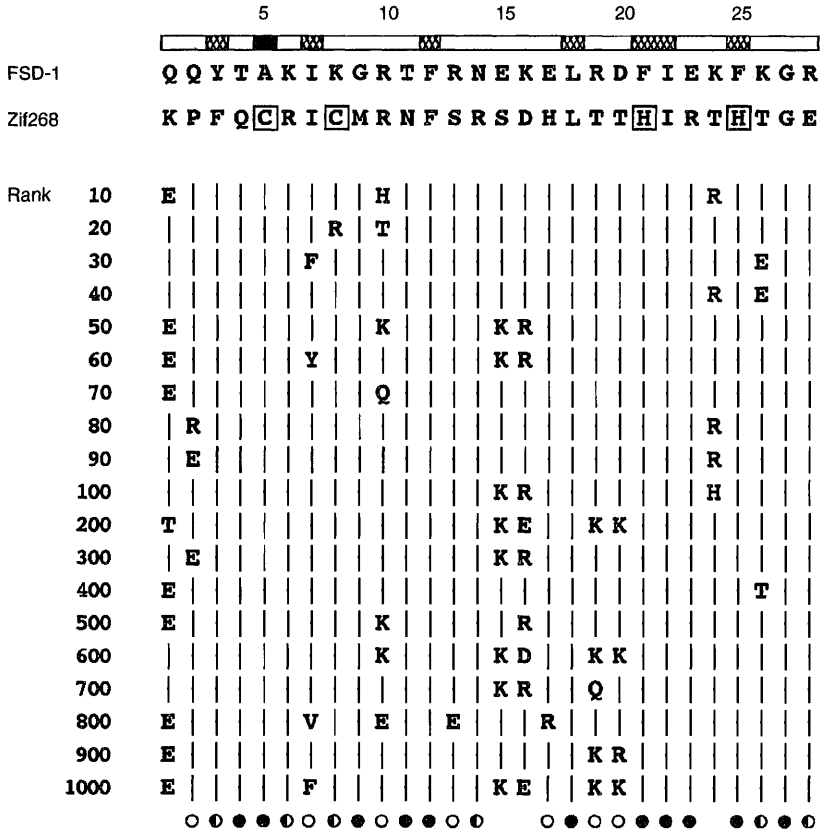

Fig. 1. Sequence of FSD-1 aligned with the second zinc finger of Zif268. The bar at the top of the figure shows the residue position classifications: the solid bar indicates the single core position, the hatched bars indicate the seven boundary positions and the open bars indicate the 20 surface positions. The alignment matches positions of FSD-1 to the corresponding backbone template positions of Zif268. Of the six identical positions (21 percent) between FSD-1 and Zif268, four are buried (Ile^7, Phe^{12}, Leu^{18}, and Ile^{22}). The zinc binding residues of Zif268 are boxed. Representative nonoptimal sequence solutions determined by means of a Monte Carlo simulated annealing protocol are shown with their rank. Vertical lines indicate identity with FSD-1. The symbols at the bottom of the figure show the degree of sequence conservation for each residue position computed across the top 1000 sequences: filled circles indicate more than 99 percent conservation, half-filled circles indicate conservation between 90 and 99 percent, open circles indicate conservation between 50 and 90 percent, and the absence of a symbol indicates less than 50% conservation. The consensus sequence determined by choosing the amino acid with the highest occurrence at each position is identical to the sequence of FSD-1. Single-letter abbreviations for amino acid residues as follows: 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.

ther, the BLAST search found only low identity matches of weak statistical significance to fragments of various unrelated proteins. The highest identity matches were 10 residues (36 percent) with P values ranging from 0.63 to 1.0, where P is the probability of a match being a chance occurrence. Random 28-residue sequences that consist of amino acids allowed in the ββα position classification described above produced similar BLAST search results, with 10- or 11-residue identities (36 to 39 percent) and P values ranging from 0.35 to 1.0, further suggesting that the matches for FSD-1 are statistically insignificant. The low identity with any known protein sequence demonstrates the novelty of the FSD-1 sequence and underscores that no sequence information from any protein motif was used in our sequence scoring function.

In order to examine the robustness of the computed sequence, we used the sequence of FSD-1 as the starting point of a Monte Carlo simulated annealing run. The Monte Carlo search revealed high scoring, suboptimal sequences in the neighborhood of the optimal solution (*4*). The energy spread from the ground-state solution to the 1000th most stable sequence is about 5 kcal/mol, an indication that the density of states is high. The amino acids comprising the core of the molecule, with the exception of position 7, are essentially invariant (Fig. 1). Almost all of the sequence variation occurs at surface positions, and typically involves conservative changes. Asn[14], which is predicted to form a stabilizing hydrogen bond to the helix back-

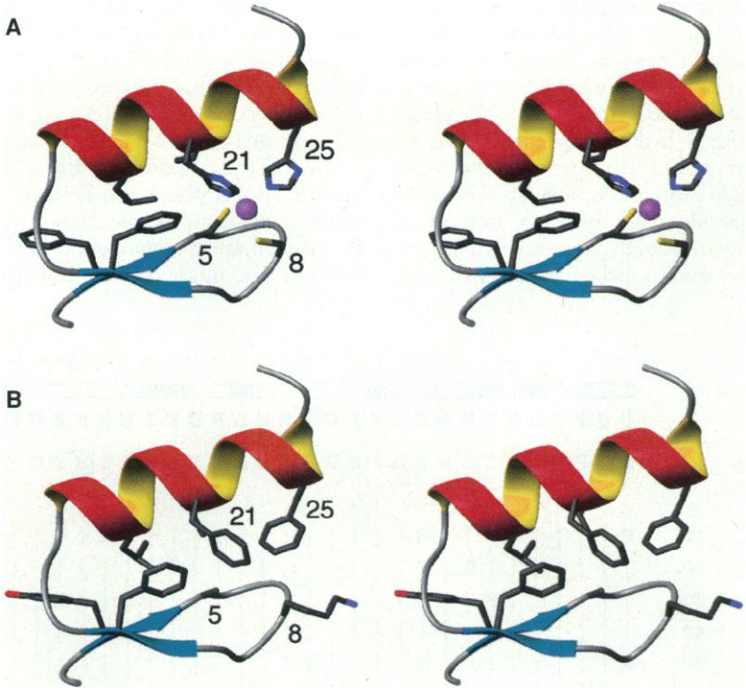

Fig. 2. Comparison of Zif268 (*9*) and computed FSD-1 structures. (**A**) Stereoview of the second zinc finger module of Zif268 showing its buried residues and zinc binding site. (**B**) Stereoview of the computed orientations of buried side chains in FSD-1. For clarity, only side chains from residues 3, 5, 8, 12, 18, 21, 22, and 25 are shown. Color figures were created with MOLMOL (*38*).

Table 1. NMR structure determination: distance restraints, structural statistics, and atomic root-mean-square (rms) deviations. ⟨*SA*⟩ are the 41 simulated annealing structures, *SA* is the average structure before energy minimization, $(SA)_r$ is the restrained energy minimized average structure, and SD is the standard deviation.

Distance restraints		
Intraresidue	97	
Sequential	83	
Short range ($\|i-j\|$ = 2 to 5 residues)	59	
Long range ($\|i-j\|$ > 5 residues)	35	
Hydrogen bond	10	
Total	284	
Structural statistics		
rms deviations	⟨*SA*⟩ ± SD	$(SA)_r$
Distance restraints (Å)	0.043 ± 0.003	0.038
Idealized geometry		
Bonds (Å)	0.0041 ± 0.0002	0.0037
Angles (degrees)	0.67 ± 0.02	0.65
Impropers (degrees)	0.53 ± 0.05	0.51
*Atomic rms deviations (Å)**		
	⟨*SA*⟩ versus *SA* ± SD	⟨*SA*⟩ versus $(SA)_r$ ± SD
Backbone	0.54 ± 0.15	0.69 ± 0.16
Backbone + nonpolar side chains†	0.99 ± 0.17	1.16 ± 0.18
Heavy atoms	1.43 ± 0.20	1.90 ± 0.29

*Atomic rms deviations are for residues 3 to 26, inclusive. Residues 1, 2, 27, and 28 were disordered [ϕ, ψ, angular order parameters (*34*) < 0.78] and had only sequential and $|i-j| = 2$ NOEs. †Nonpolar side chains are from residues Tyr^3, Ala^5, Ile^7, Phe^{12}, Leu^{18}, Phe^{21}, Ile^{22}, and Phe^{25}, which constitute the core of the protein.

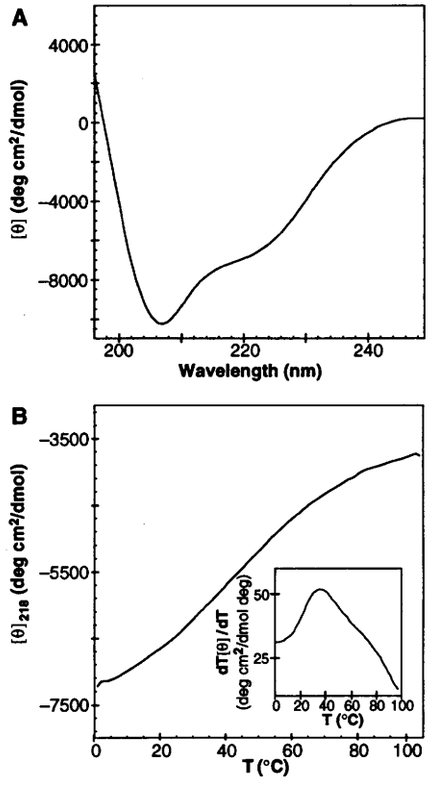

Fig. 3. Circular dichroism (CD) measurements of FSD-1. (**A**) Far-UV CD spectrum of FSD-1 at 1°C. The minima at 220 and 207 nm indicate a folded structure. (**B**) Thermal unfolding of FSD-1 monitored by CD. The melting curve has an inflection point at 39°C. To illustrate the cooperativity of the thermal transition, the melting curve was fit to a two-state model [(*39*) and the derivative of the fit is shown (inset)]. The melting temperature determined from this fit is 42°C.

bone, is among the most conserved surface positions. The strong sequence conservation observed for critical areas of the molecule suggests that, if a representative sequence folds into the design target structure, then many sequences whose variations do not disrupt the critical interactions may be equally competent. Even if billions of sequences would successfully achieve the target fold, they would represent only a very small proportion of the 10^{27} possible sequences.

Experimental validation. FSD-1 was synthesized in order to allow us to characterize its structure and assess the performance of the design algorithm (*24*). The far-ultraviolet (UV) circular dichroism (CD) spectrum of FSD-1 shows minima at 220 nm and 207 nm, which is indicative of a folded structure (Fig. 3A) (*25*). The thermal melt is weakly cooperative, with an inflection point at 39°C (Fig. 3B), and is completely reversible. 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 thermal unfolding transitions observed for other folded short peptides (*26*). FSD-1 is highly soluble (greater than 3 mM), and equilibrium sedimentation studies at 100 μM, 500 μM, and 1 mM show the protein to be monomeric (*27*). The sedimentation data fit well to a single species, monomer model with a molecular mass of 3630 at 1 mM, in good agreement with the calculated monomer mass of 3488. Also, far UV-CD spectra showed no concentration dependence from 50 μM to 2 mM, and nuclear magnetic resonance (NMR) COSY spectra taken at 100 μM and 2 mM were essentially identical.

The solution structure of FSD-1 was solved by means of homonuclear 2D ^{1}H NMR spectroscopy (*28*). NMR spectra were well dispersed, indicating an ordered protein structure and easing resonance assignments. Proton chemical shift assignments were determined with standard homonuclear methods (*29*). Unambiguous sequential and short-range NOEs (Fig. 4) indicate helical secondary structure from residues 15 to 26 in agreement with the design target. Representative long-range NOEs from the helix to Ile^7 and Phe^{12} indicate a hydrophobic core consistent with the desired tertiary structure (Fig. 4B).

The structure of FSD-1 was determined from 284 experimental restraints (10.1 restraints per residue) that were nonredundant with covalent structure including 274 NOE distance restraints and 10 hydrogen bond restraints involving slowly exchanging amide protons (*30*). Structure calculations were performed with X-PLOR (*31*) with the use of standard protocols for hybrid distance geometry-simulated annealing (*32*). An ensemble of 41 structures converged with good covalent geometry and no distance restraint violations greater than 0.3 Å (Fig. 5 and Table 1). The backbone of FSD-1 is well defined with a root-mean-square (rms) deviation from the mean of 0.54 Å (residues 3 to 26). Consideration of the buried side chains (Tyr^3, Ala^5, Ile^7, Phe^{12}, Leu^{18}, Phe^{21}, Ile^{22}, and Phe^{25}) along with the backbone gives an rms deviation of 0.99 Å, indicating that the core of the molecule is well ordered. The stereochemical quality of the ensemble of structures was examined with PROCHECK (*33*). Apart from the disordered termini and the glycine residues, 87 percent of the residues fall in the most favored region and the remainder in the allowed region of ϕ,ψ space. Modest heterogeneity is evident in the first strand (residues 3 to 6), which has an average backbone angular order parameter, $\langle S \rangle$ (*34*), of 0.96 ± 0.04 compared to the second strand (residues 9 to 12) with an $\langle S \rangle$ = 0.98 ± 0.02 and the helix (residues 15 to 26) with an $\langle S \rangle$ = 0.99 ± 0.01. Overall, FSD-1 is notably well ordered and, to our knowledge, is the shortest sequence consist-

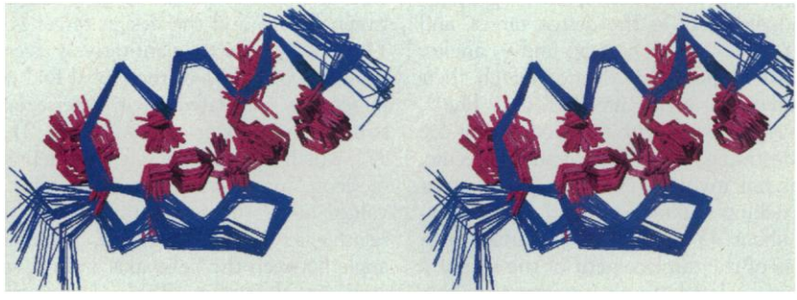

Fig. 5. Solution structure of FSD-1. Stereoview showing the best-fit superposition of the 41 converged simulated annealing structures from X-PLOR (*31*). The backbone Cα trace is shown in blue and the side-chain heavy atoms of the hydrophobic residues (Tyr^3, Ala^5, Ile^7, Phe^{12}, Leu^{18}, Phe^{21}, Ile^{22}, and Phe^{25}) are shown in magenta. The amino terminus is at the lower left of the figure and the carboxyl 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.

Fig. 4. NOE contacts for FSD-1. (**A**) Sequential and short-range NOE connectivities. The d denotes a contact between the indicated protons. All adjacent residues are connected by Hα-HN, HN-HN, or Hβ-HN NOE cross-peaks. The helix (residues 15 to 26) is well defined by short-range connections, as is the hairpin turn at residues 7 and 8. (**B**) Representative NOE contacts from aromatic to methyl protons. Several long-range NOEs from Ile^7 and Phe^{12} to the helix help define the fold of the protein. The starred peak has an ambiguous F1 assignment, Ile^{22} Hd1 or Leu^{18} Hd2.

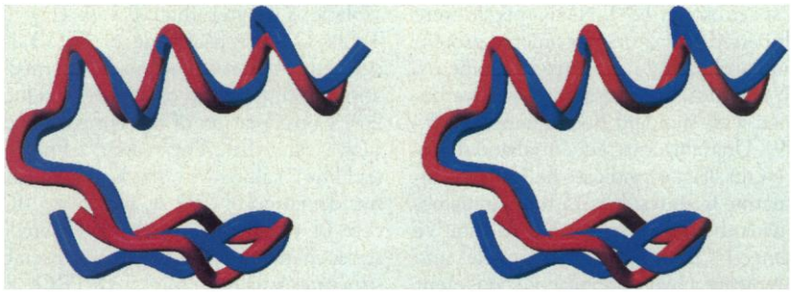

Fig. 6. Comparison of the FSD-1 structure (blue) and the design target (red). Stereoview of the best-fit superposition of the restrained energy minimized average NMR structure of FSD-1 and the backbone of Zif268. Residues 3 to 26 are shown.

Table 2. Comparison of the FSD-1 experimentally determined structure and the design target structure. The FSD-1 structure is the restrained energy minimized average from the NMR structure determination. The design target structure is the second DNA binding module of the zinc finger Zif268 (*9*)

Atomic rms deviations (Å)	
Backbone, residues 3 to 26	1.98
Backbone, residues 8 to 26	0.98

*Super-secondary structure parameters**	FSD-1	Design target
h (Å)	9.9	8.9
θ (degrees)	14.2	16.5
Ω (degrees)	13.1	13.5

**h*, θ, and Ω are calculated as described (*36, 37*). *h* is the distance between the centroid of the helix Cα coordinates (residues 15 to 26) and the least-squares plane fit to the Cα coordinates of the sheet (residues 3 to 12); θ is the angle of inclination of the principal moment of the helix Cα atoms with the plane of the sheet; Ω is the angle between the projection of the principal moment of the helix onto the sheet and the projection of the average least-squares fit line to the strand Cα coordinates (residues 3 to 6 and 9 to 12) onto the sheet.

ing entirely of naturally occurring amino acids that folds to a well-ordered structure without metal binding, oligomerization, or disulfide bond formation (*35*).

The packing pattern of the hydrophobic core of the NMR structure ensemble of FSD-1 (Tyr^3, Ile^7, Phe^{12}, Leu^{18}, Phe^{21}, Ile^{22}, and Phe^{25}) is similar to the computed packing arrangement. Five of the seven residues have χ_1 angles in the same gauche$^+$, gauche$^-$ or trans category as the design target, and three residues match both χ_1 and χ_2 angles. The two residues that do not match their computed χ_1 angles are Ile^7 and Phe^{25}, which is consistent with their location at the less constrained open end of the molecule. Ala^5 is not involved in its expected extensive packing interactions and instead exposes about 45 percent of its surface area because of the displacement of the strand 1 backbone relative to the design template. Conversely, Lys^8 behaves as predicted by the algorithm with its solvent exposure (60 percent) and χ_1 and χ_2 angles matching the computed structure. Because there are few NOEs involving solvent-exposed side chains, most of these side chains are disordered in the solution structure, a state that precludes examination of the predicted surface residue hydrogen bonds. However, Asn^{14} forms a hydrogen bond from its side chain carbonyl oxygen as predicted, but to the amide of Glu^{17}, not Lys^{16} as expected from the design. This hydrogen bond is present in 95 percent of the structure ensemble and has a donor-acceptor distance of 2.6 ± 0.06 Å. In general, the side chains of FSD-1 correspond well with the design algorithm predictions, but further refinement of the scoring function and rotamer library should improve sequence selection and side chain placement and improve the correlation between the predicted and observed structures.

We compared the average restrained minimized structure of FSD-1 and the design target (Fig. 6). The overall backbone rms deviation of FSD-1 from the design target is 1.98 Å for residues 3 to 26 and only 0.98 Å for residues 8 to 26 (Table 2). The largest difference between FSD-1 and the target structure occurs from residues 4 to 7, with a displacement of 3.0 to 3.5 Å of the backbone atom positions of strand 1. The agreement for strand 2, the strand-to-helix turn, and the helix is remarkable, with the differences nearly within the accuracy of the structure determination. For this region of the structure, the rms difference of φ, ψ angles between FSD-1 and the design target is only 14 ± 9°. In order to quantitatively assess the similarity of FSD-1 to the global fold of the target, we calculated their supersecondary structure parameter values (Table 2) (*36, 37*), which describe the relative orientations of secondary structure units in proteins. The values of θ, the inclination of the helix relative to the sheet, and Ω, the dihedral angle between the helix axis and the strand axes (see legend to Table 2), are nearly identical. The height of the helix above the sheet, *h*, is only 1 Å greater in FSD-1. A study of protein core design as a function of helix height for Gβ1 variants demonstrated that up to 1.5 Å variation in helix height has little effect on sequence selection (*37*). The comparison of supersecondary structure parameter values and backbone coordinates highlights the excellent agreement between the experimentally determined structure of FSD-1 and the design target, and demonstrates the success of our algorithm at computing a sequence for this ββα motif.

The quality of the match between FSD-1 and the design target demonstrates the ability of our algorithm to design a sequence for a fold that contains the three major secondary structure elements of proteins: sheet, helix, and turn. Since the ββα fold is different from those used to develop the sequence-selection methodology, the design of FSD-1 represents a successful transfer of our algorithm to a new motif. Further tests of the performance of the algorithm on several different motifs are necessary, although its basis in physical chemistry and the absence of heuristics and subjective considerations should allow the algorithm to be used in many different structural contexts. Also, the generation of various kinds of backbone templates for use as input to our fully automated sequence selection algorithm could enable the design of new protein folds. Recent results indicate that the sequence selection algorithm is not sensitive to even fairly large perturbations in backbone geometry and should be robust enough to accommodate computer-generated backbones (*37*).

The key to using a quantitative method for the FSD-1 design, and for the continued development of the methodology, is the tight coupling of theory, computation, and experiment used to improve the accuracy of the physical chemical potential functions in our algorithm. When combined with these potential functions, computational optimization methods such as DEE can rapidly find sequences for structures too large for experimental library screening or too complex for subjective approaches. Given that the FSD-1 sequence was computed with only a 4-Giga-FLOPS computer (*19*), and that TeraFLOPS computers are now available with PetaFLOPS computers on the drawing board, the prospect for pursuing even larger and more complex designs is excellent.

REFERENCES AND NOTES

1. M. H. J. Cordes, A. R. Davidson, R. T. Sauer, *Curr. Opinion Struct. Biol.* **6**, 3 (1996).
2. D. Y. Jackson *et al.*, *Science* **266**, 243 (1994); B. Li *et al.*, *ibid.* **270**, 1657 (1995); J. S. Marvin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 4366 (1997).
3. H. W. Hellinga, J. P. Caradonna, F. M. Richards, *J. Mol. Biol.* **222**, 787 (1991); J. H. Hurley, W. A. Baase, B. W. Matthews, *ibid.* **224**, 1143 (1992); J. R. Desjarlais and T. M. Handel, *Protein Sci.* **4**, 2006 (1995); P. B. Harbury, B. Tidor, P. S. Kim, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 8408 (1995); M. Klemba, K. H. Gard-

ner, S. Marino, N. D. Clarke, L. Regan, *Nature Struc. Biol.* **2**, 368 (1995); S. F. Betz and W. F. Degrado, *Biochemistry* **35**, 6955 (1996).
4. B. I. Dahiyat and S. L. Mayo, *Protein Sci.* **5**, 895 (1996).
5. _____, *ibid.* **6**, 1333 (1997).
6. _____, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 10172 (1997).
7. J. W. Ponder and F. M. Richards, *J. Mol. Biol.* **193**, 775 (1987).
8. [See J. Desmet, M. De Maeyer, B. Hazes, I. Lasters, *Nature* **356**, 539 (1992); R. F. Goldstein, *Biophys. J.* **66**, 1335 (1994); M. De Maeyer, J. Desmet, I. Laster, *Folding Design* **2**, 53 (1997)] DEE finds and eliminates rotamers that are mathematically provable to be inconsistent (or dead-ending) with the global minimum energy solution of the system. A rotamer r at some residue position i will be dead-ending if, when compared with some other rotamer t, at the same residue position the following inequality is satisfied:

$$E(i_r) - E(i_t) + \sum_j \min_s [E(i_r j_s) - E(i_t j_s)] > 0$$

where $E(i_r)$ and $E(i_t)$ are rotamer-template energies, $E(i_r j_s)$ and $E(i_t j_s)$ are rotamer-rotamer energies for rotamers on residues i and j, and the function min_s selects the rotamer s on residue j that minimizes the argument of the function. Iterative application of the elimination criterion results in a rapid and substantial reduction in the combinatorial size of the problem and application of similar but higher-order elimination criteria are required to find the ground-state solution.
9. N. P. Pavletich and C. O. Pabo, *Science* **252**, 809 (1991).
10. M. D. Struthers, R. P. Cheng, B. Imperiali, *ibid.* **271**, 342 (1996).
11. B. I. Dahiyat and S. L. Mayo, unpublished results.
12. Potential functions and parameters for van der Waals interactions, solvation, hydrogen bonding, and secondary structure propensity are described in our previous work (*4–6*). A secondary structure propensity potential was used for surface β-sheet positions where the $i - 1$ and $i + 1$ residues were also in β-sheet conformations (*5*). Propensity values from Serrano and co-workers were used [V. Munoz and L. Serrano, *Proteins Struct. Funct. Genet.* **20**, 301 (1994)].
13. F. C. Bernstein *et al.*, *J. Mol. Biol.* **112**, 535 (1977).
14. The coordinates of PDB record 1zaa (*9, 13*) 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, Inc., San Diego, CA) was used to generate explicit hydrogens on the structure which was then conjugate-gradient minimized for 50 steps by means of the Dreiding force field (*40*).
15. A solvent-accessible surface was generated using the Connolly algorithm (*41*) with a probe radius of 8.0 Å, a dot density of 10 $Å^{-2}$, and a Cα radius of 1.95 Å. A residue was classified as a core position if the distance from its Cα, along its Cα-Cβ vector, to the solvent-accessible surface was greater than 5.0 Å, and if the distance from its Cβ to the nearest surface point was greater than 2.0 Å. The remaining residues were classified as surface positions if the sum of the distances from their Cα, along their Cα-Cβ vector, to the solvent-accessible surface plus the distance from their Cβ to the closest surface point was less than 2.7 Å. All remaining residues were classified as boundary positions. The classifications for Zif268 were used as computed except that positions 1, 17, and 23 were converted from the boundary to the surface class to account for end effects from the proximity of chain termini to these residues in the tertiary structure and inaccuracies in the assignment.
16. One core position (7 possible amino acids), 7 boundary positions (16 possible amino acids), 18 surface positions (10 possible amino acids), and 2 positions with φ greater than 0° (1 possible amino acid) result in $7 * 16^7 * 10^{18} * 1^2 = 1.88 \times 10^{27}$ possible amino acid sequences.
17. 1.88×10^{27} peptide molecules, with an average mass of 3712 daltons for the possible compositions allowed by the residue position classification, would weigh $(1.88 \times 10^{27} * 3712$ daltons) = 1.159×10^7 g = 11.6 metric tons.
18. As in our previous work (*5*), a backbone-dependent rotamer library was used [R. L. Dunbrack and M. Karplus, *J. Mol. Biol.* **230**, 543 (1993)]. All His rotamers were protonated on both Nδ and Nε.
19. All calculations were performed on a Silicon Graphics Power Challenge server with 10 R10000 processors running in parallel. Peak performance is 3.9 GigaFLOPS (FLOPS = floating point operations per second).
20. The sequence optimization consists of two phases: pairwise rotamer energy calculations and DEE searching. The DEE optimization was initially run with control parameters set for optimal speed followed by a DEE-based, residue-pairwise, round-robin optimization. The energy calculations took 53 CPU (central processing unit) hours and sequence optimizations took 37 CPU hours.
21. C. W. A. Kim and J. M. Berg, *Nature* **362**, 267 (1993); D. L. Minor and P. S. Kim, *ibid.* **367**, 660 (1994); C. K. Smith, J. M. Withka, L. Regan, *Biochemistry* **33**, 5510 (1994).
22. J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, R. T. Sauer, *Science* **247**, 1306 (1990).
23. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
24. FSD-1 was synthesized by means of standard solid-phase Fmoc chemistry. The peptide was cleaved from the resin with trifluoroacetic acid and purified by reversed-phase high-performance liquid chromatography. Peptide was lyophilized and stored at −20°C. Matrix-assisted laser desorption mass spectrometry yielded a molecular weight of 3489.7 daltons (3489.0 calculated).
25. Protein concentration was 50 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 3 scans with a 1-s integration time and 1-nm increments. All CD data were acquired on an Aviv 62DS spectrometer equipped with a thermoelectric temperature control unit. Thermal unfolding was monitored at 218 nm in a 1-mm cuvette with 2° increments and an averaging time of 40 s and an equilibration time of 120 s per increment. Reversibility was confirmed by comparison of 1°C CD spectra before and after heating to 99°C. Peptide concentrations were determined by UV spectrophotometry.
26. J. M. Scholtz *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 2854 (1991); M. A. Weiss and H. T. Keutmann, *Biochemistry* **29**, 9808 (1990); M. D. Struthers, R. P. Cheng, B. Imperiali, *J. Am. Chem. Soc.* **118**, 3073 (1996).
27. Sedimentation equilibrium studies were performed on a Beckman XL-A ultracentrifuge equipped with an An-60 Ti analytical rotor at a speed of 40,000 rpm. Protein concentration was 100 μM, 500 μM, or 1 mM in 50 mM sodium phosphate at pH 5.0 and 7°C. Absorption was monitored at 286 nm (500 μM and 1 mM) or 234 nm (100 μM). Concentration profiles were fit to an ideal single species model which resulted in randomly distributed residuals.
28. 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 (~2 mM) were prepared in H_2O-D_2O (90:10) or in 99.9 percent D_2O with 50 mM sodium phosphate at pH 5.0 (uncorrected glass electrode). All spectra were collected at 7°C. DQF-COSY [U. Piantini, O. W. Sorensen, R. R. Ernst, *J. Am. Chem. Soc.* **104**, 6800 (1982)], TOCSY [A. Bax and D. G. Davis, *J. Magnetic Reson.* **65**, 355 (1985)], and NOESY [J. Jeener, B. H. Meier, P. Bachmann, R. R. Ernst, *J. Chem. Phys.* **71**, 4546 (1979)] spectra were acquired to accomplish resonance assignments and structure determination. NOESY spectra were recorded with mixing times of 200 ms for use during resonance assignments and 100 ms to derive distance restraints. Water suppression was accomplished either with presaturation during the relaxation delay or pulsed field gradients [M. Piotto, V. Saudek, V. Sklenar, *J. Biomol. NMR* **2**, 661 (1992)]. Spectra were processed with VNMR (Varian Associates, Palo Alto, CA), and spectra were assigned with ANSIG [P. J. Kraulis, *J. Magnetic Reson.* **24**, 627 (1989)].
29. K. Wuthrich, *NMR of Proteins and Nucleic Acids* (Wiley, New York, 1986).
30. NOEs were classified into three distance-bound ranges based on cross-peak intensity calibrated to the Tyr^3 Hδ-Hε crosspeak: 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. Hydrogen bond restraints were derived from hydrogen deuterium-exchange kinetics measurements followed by one dimensional ^1H spectroscopy. Unambiguously assigned amide peaks for Tyr^3, Phe^{12}, Leu^{18}, Phe^{21}, and Phe^{25} were protected from exchange at 7°C, pH 5.0. Hydrogen bond restraints (two per hydrogen bond) were only included at the late stages of structure refinement when initial calculations indicated the donor-acceptor pairings.
31. A. T. Brünger, *X-PLOR, version 3.1, A System for X-ray Crystallography and NMR* (Yale Univ. Press, New Haven, CT, 1992).
32. Standard hybrid distance geometry-simulated annealing protocols were followed [M. Nilges, G. M. Clore, A. M. Gronenborn, *FEBS Lett.* **229**, 317 (1988); M. Nilges, J. Kuszewski, A. T. Brünger, in *Computational Aspects of the Study of Biological Macromolecules by NMR* J. C. Hoch, Ed. (Plenum, New York, 1991); J. Kuszewski, M. Nilges, A. T. Brünger, *J. Biomol. NMR* **2**, 33 (1992)]. Distance geometry structures (100) were generated, regularized, and refined, resulting in an ensemble, called ⟨*SA*⟩, of 41 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°. An average structure was generated by superimposing and then averaging the coordinates of the ensemble, followed by refinement and restrained minimization.
33. R. A. Laskowski, M. W. Macarthur, D. S. Moss, J. M. Thornton, *J. Appl. Crystallogr.* **26**, 283 (1993).
34. S. G. Hyberts, M. S. Goldberg, T. F. Havel, G. Wagner, *Protein Sci.* **1**, 736 (1992).
35. C. J. McKnight, P. T. Matsudaira, P. S. Kim, *Nature Struct. Biol.* **4**, 180 (1997).
36. J. Janin and C. Chothia, *J. Mol. Biol.* **143**, 95 (1980); F. E. Cohen, M. J. E. Sternberg, W. R. Taylor, *ibid.* **56**, 821 (1982).
37. A. Su and S. L. Mayo, *Protein Sci.* **6**, 1701 (1997).
38. R. Koradi, M. Billeter, K. Wuthrich, *J. Mol. Graph.* **14**, 51 (1996).
39. W. J. Becktel and J. A. Schellman, *Biopolymers* **26**, 1859 (1987).
40. S. L. Mayo, B. D. Olafson, W. A. Goddard III, *J. Phys. Chem.* **94**, 8897 (1990).
41. M. L. Connolly, *Science* **221**, 709 (1983).
42. We thank P. Poon and T. Laue for sedimentation equilibrium measurements and discussions, A. Su for assistance calculating super-secondary structure parameters, S. Ross for assistance with NMR measurements, G. Hathaway for mass spectrometry, J. Abelson and P. Bjorkman for critical reading of the manuscript, and R. A. Olofson for helpful discussions. Supported by the Howard Hughes Medical Institute (S.L.M.), the Rita Allen Foundation, the Chandler Family Trust, the Booth Ferris Foundation, the David and Lucile Packard Foundation, the Searle Scholars Program and The Chicago Community Trust, and grant GM08346 from the National Institutes of Health (B.I.D.). Coordinates and NMR restraints have been deposited in the Brookhaven Protein Data Bank with accession numbers 1FSD and R1FSDMR, respectively.

16 June 1997; accepted 8 September 1997