Structure *14*, 847–856, May 2006 ©2006 Elsevier Ltd All rights reserved DOI 10.1016/j.str.2006.02.011

High-Resolution Structural Validation of the Computational Redesign of Human U1A Protein

Neil Dobson,[2] Gautam Dantas,[1] David Baker,[1] and Gabriele Varani[1,2,*]
[1] Department of Biochemistry
[2] Department of Chemistry
University of Washington
Seattle, Washington 98195

Summary

Achieving atomic-level resolution in the computational design of a protein structure remains a challenging problem despite recent progress. Rigorous experimental tests are needed to improve protein design algorithms, yet studies of the structure and dynamics of computationally designed proteins are very few. The NMR structure and backbone dynamics of a redesigned protein of 96 amino acids are compared here with the design target, human U1A protein. We demonstrate that the redesigned protein reproduces the target structure to within the uncertainty of the NMR coordinates, even as 65 out of 96 amino acids were simultaneously changed by purely computational methods. The dynamics of the backbone of the redesigned protein also mirror those of human U1A, suggesting that the protein design algorithm captures the shape of the potential energy landscape in addition to the local energy minimum.

Introduction

Protein design represents one of the great challenges of computational structural biology. The ability to successfully design new proteins would allow us to generate new reagents or enzymes and, at the same time, provide us with an understanding of the principles of protein folding and stability. Remarkable progress has recently been made after a decade of steady progress in this field (Dahiyat and Mayo, 1997; Harbury et al., 1998; Havranek and Harbury, 2003; Kaplan and DeGrado, 2004; Mooers et al., 2003; Pokala and Handel, 2001; Shifman and Mayo, 2003), as demonstrated, for example, by the design of a completely new fold not previously observed in nature (Kuhlman et al., 2003) and of new enzymatic activities into existing protein scaffolds (Dwyer et al., 2004; Looger et al., 2003). Despite these successes, the problem of designing a protein with a desired structure remains far from being solved. Therefore, it is essential that existing computational methods be further improved after their experimental verification through prediction and design challenges.

Studies of the structure of proteins designed (or redesigned) by computational methods are very few (Johnson et al., 1999; Kuhlman et al., 2003; Walsh et al., 1999, 2001), although such studies provide essential validations of design protocols and insight into their limitations (Shifman and Mayo, 2003). Equally significant is the opportunity to assess the role of motion in protein stabilization (Johnson et al., 1999; Lee and Wand, 2001; Wand, 2001), because dynamics has the potential to provide information on the shape of the protein's energy landscape. The systematic investigation of the structure and dynamics of proteins designed to fold into new or already existing structures, and the comparison with studies of native proteins, would also provide a unique opportunity to explore the evolution of protein families.

The *Rosetta* protein design algorithm was recently used to completely redesign computationally the hydrophobic core of 9 proteins of 60–100 amino acids (Dantas et al., 2003). In this test, new protein sequences were sought that would fold into a predefined structure, as determined by X-ray crystallography. Three out of nine proteins formed thermodynamically stable structures, as evidenced by circular dichroism (CD) melting studies and by 1D NMR experiments. In the present work, we study URNdesign, a protein designed to reproduce the structure of a well-known RNA binding protein, human U1A (Dantas et al., 2003). The structure and dynamics of the native protein have been studied by us in the past because of its role as a model for RNA recognition by the largest RNA binding protein family, the RNA recognition motif or RRM (Allain et al., 1996; Avis et al., 1996; Mittermaier et al., 1999; Nagai et al., 1995; Oubridge et al., 1994). In order to establish the structural accuracy of the redesign, we used NMR spectroscopy to determine the solution structure of URNdesign and to study the dynamics of the protein's backbone. The results presented here show that the design algorithm has generated a protein that is remarkably close in structure and backbone dynamics to the design target, even if 70% of all amino acids (65/96) were simultaneously changed by the design algorithm.

Results

Protein Design

The URNdesign sequence was generated by completely redesigning human U1A protein by using the *Rosetta Design* algorithm (Dantas et al., 2003). The protein structure generated by X-ray crystallography (Oubridge et al., 1994) was first minimized to start the design from an energetically favored conformation, and it was then kept fixed during the actual redesign. This protocol generated a protein sequence in which 65 out of the 96 amino acids were changed simultaneously (Figure 1A); thus, nearly 70% of all amino acids were simultaneously changed computationally in generating the URNdesign sequence from the U1A 3D coordinates. However, the two sequences remain similar within the core, since only 11 out of 21 highly buried residues were changed. Because of this similarity, the redesigned protein can be identified by sequence alignment as a member of the RRM superfamily. Polar and nonpolar residues were changed to a comparable extent: about 40% of all hydrophobic amino acids and 37% of polar residues were strictly conserved. However, the number of

*Correspondence: varani@chem.washington.edu

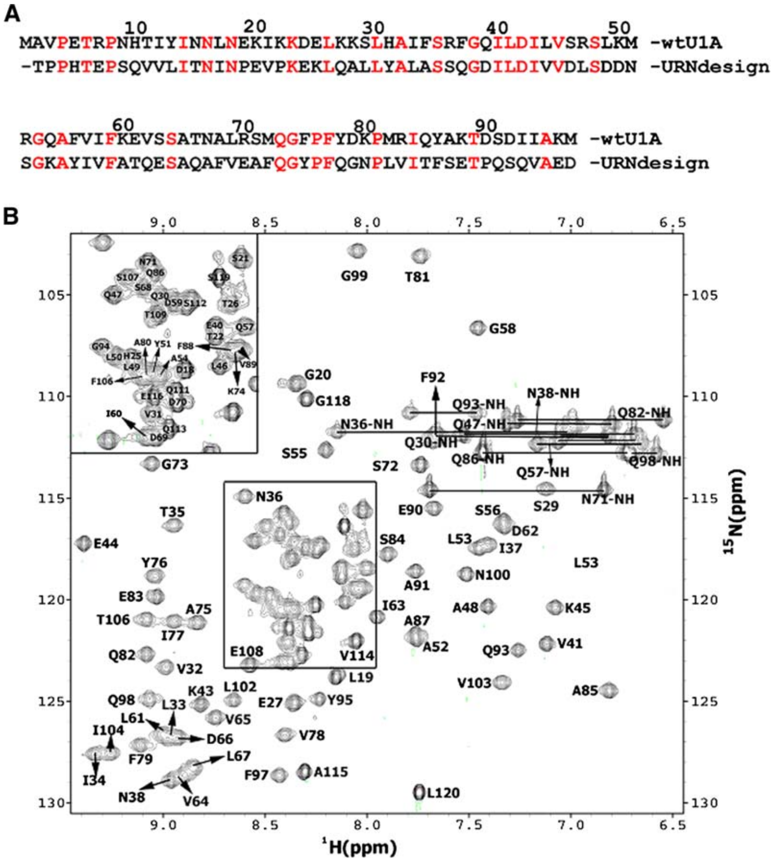

Figure 1. Sequence Alignment for U1A and URNdesign Proteins and NMR Spectra of URNdesign

(A) The sequence of the RNA binding domain of human U1A protein (top) is aligned to that of its computational redesign (bottom); conserved residues are in red.

(B) ^{1}H-^{15}N HSQC spectrum of URNdesign recorded at 298K with spectral assignments indicated; the inset represents an expanded view of the most crowded region of the spectrum.

charged residues was markedly different, as was the overall charge of the protein. Only 18% of all residues are charged in URNdesign compared to 29% in U1A. Polar, noncharged or hydrophobic residues replace all but 3 of the many Lys and Arg residues in U1A (16 in total, many involved in RNA binding [Allain et al., 1996; Oubridge et al., 1994]).

NMR Analysis and Structure Determination

URNdesign was found to have sharp lines and good dispersion in 1D NMR experiments. It also had well-defined profiles in both thermal and chemical denaturation experiments, and it had enhanced thermodynamic stability compared to U1A (Dantas et al., 2003). The HSQC spectrum of URNdesign is well dispersed as well, comparable to what was reported for the U1A protein (Avis et al., 1996): 95 out of 100 backbone amides in our construct are well resolved in the ^{1}H-^{15}N spectrum (Figure 1B). Designer proteins are often found to aggregate or form molten globule structures. In a case we have recently studied, we found that a protein designed to be monomeric was instead an exceptionally stable dimer (G.D. et al., submitted). In this case, the very high quality of the HSQC spectra, the T2 values (about 100 ms), the correlation time obtained from analysis of relaxation data (see below), the size exclusion chromatography profile, and quantitative analytical ultracentrifugation all demonstrate that URNdesign is a well-folded monomeric protein at NMR concentration, and there is no evidence whatsoever for aggregation or dimerization.

Assignments of the protein resonances were obtained as described in the Experimental Procedures. Over 94% of the backbone NH, CO, C_{α}, and C_{β} nuclei, >96% of all side chain ^{1}H and ^{13}C resonances, and >86% of Gln/Asn NH_2 residues were thus assigned; Arg N_{ε} and guanidinium groups as well as Lys NH_3 residues remain unassigned. Aromatic Tyr/Phe/His and Trp resonances are completely assigned, although not stereospecifically. The missing assignments are all in the regions of the protein outside of the central folded domain.

A list of all NOE interactions resolved in NOESY spectra was generated from 3D ^{15}N- and ^{13}C-edited NOESY, 2D NOESY in water, and 2D NOESY in D_2O, and interactions were introduced as unassigned distance constraints into CYANA (Güntert, 2003). The macro CANDID automatically assigned these NOE crosspeaks based on the chemical shift list generated from triple resonance data. Dihedral constraints generated from TALOS (Cornilescu et al., 1999) were also added at this stage. This partially assigned NOE list was refined by examining each constraint against the original data during successive rounds of calculation and refinement. By using CYANA2.0, we quickly obtained converged structures that had only five distance violations >0.2 Å; these initial structures were refined to the final statistics presented in Tables 1 and 2. Figure 2A shows a backbone superposition of the 20 lowest-energy structures obtained in the final round of calculations.

The total number of assigned NOEs (1602 in total) is comparable to the number obtained for the U1A protein

Table 1. Experimental Restraints

NOE distance restraints (first round/final round)[a]	
Total	2754/1602
Intraresidue and sequential ($[i - j] \leq 1$)	1847/819
Medium range ($1 \leq [i - j] \leq 5$)	729/296
Long range ($[i - j] \geq 5$)	178/487
Dihedral angle constraints[b]	91
Hydrogen bond constraints	40 (20 H bonds)
Total number of constraints	1733
Number of constraints per residue	18.6
Long-range constraints per residue	5.2
Residual constraint violations[a]	
Distance violations > 0.2 Å	0
Van der Waals violations[a]	
0.2–0.5 Å	9
>0.5 Å	0
Dihedral angle violations[a] > 1°	0
CYANA target function (first round/final round)	
CANDID run	135 A^2/7.5 A^2
Final CALC run	—/2.05 A^2

[a] First and final round refer to statistics generated from the CANDID macro in CYANA2.0.

[b] Dihedral angle constraints were generated from TALOS (Cornilescu et al., 1999).

(Avis et al., 1996) (1995 in total for a larger construct of 117 amino acids). A total of 323 NOEs remained unassigned, representing only 6.2% of all unique automatically selected NOEs. The number of short-, medium-, and long-range interactions is also comparable to those observed for U1A, as a percentage of the total number of NOEs (52%, 17%, and 31%, respectively, for U1A; 51%, 18%, and 30%, respectively, for URNdesign). The number of hydrogen bond constraints is also essentially the same—20 for URNdesign compared to 19 for U1A—and the number of dihedral angle constraints was in excess of those obtained for U1A—91 compared to 12—because of TALOS (Cornilescu et al., 1999). The final ensemble has an rmsd from the average structure of 0.26 Å for backbone atoms and 0.67 Å for all heavy atoms over the ordered region (residues 9–99). When the structure was analyzed with Procheck, 97.9% of dihedral angles were found in allowed regions of the Ramachandran plot, leaving just 0.5% in disallowed regions. These all belong to residue G74 as well as residues G99 and N100 at the very C terminus of the protein.

Table 2. Structural Statistics

Rmsd values (residues 8–88)[a]	
Backbone atoms (Å)	0.26
All heavy atoms (Å)	0.67
Rmsd values (residues 1–96)[a]	
Backbone atoms (Å)	1.73
All heavy-atoms (Å)	2.01
PROCHECK analysis	
Most favored regions (%)	79.1
Additionally allowed (%)	19.8
Generously allowed (%)	1.6
Disallowed (%)	0.5
G-score (ϕ, ψ)	−0.67
G-score (all dihedrals)	−0.88

[a] Structural statistics reported are based on analysis of the best 20 conformers of 100 generated by CYANA2.0.

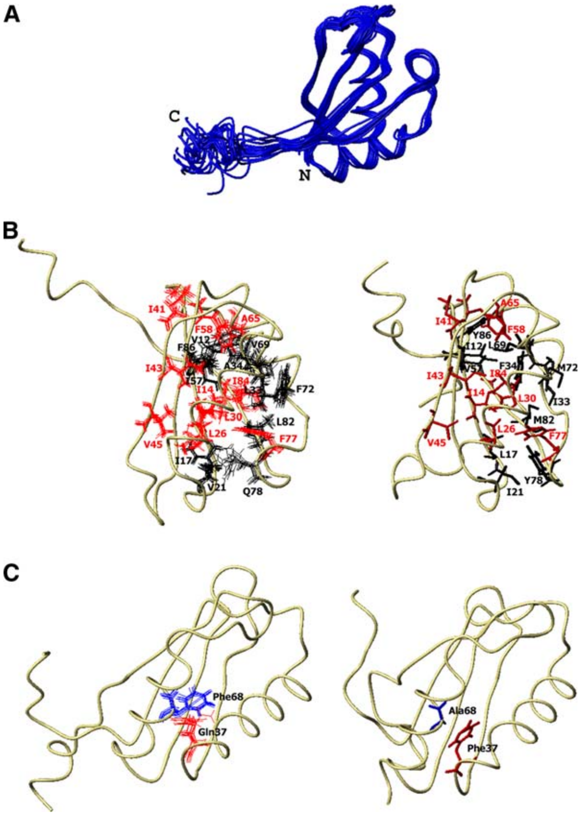

Figure 2. Structure of URNdesign

(A) The 20 NMR structures of URNdesign with the lowest energy are superposed over the ordered region of the protein (residues 8–96) and displayed in a ribbon representation. Structural figures are generated with MolMol (Koradi et al., 1996).

(B) Superposition of core residues for the ten best URNdesign structures (left) and comparison with the hydrophobic core of U1A (right). Core residues conserved between the two sequences are shown in red; nonconserved residues are shown in black.

(C) Close-up view of the interaction between Gln37 and Phe68 in URNdesign (left) and between Phe37 and Ala68 in U1A (right).

Structure of the Redesigned U1A Protein

URNdesign retains the canonical RRM fold consisting of a four-stranded antiparallel β sheet and two α helices arranged as the split $\alpha\beta$ fold. At the end of the domain is a small section of helix, α_3, as observed in both free (Avis et al., 1996) and RNA bound U1A protein (Allain et al., 1996; Oubridge et al., 1994). However, this C-terminal helix is disordered compared to the rest of the protein: it is present in only 13 of the 20 best structures. The design was based on the crystal structure of U1A bound to RNA, where helix α_3 is held in place by extensive interactions with the RNA (Allain et al., 1996; Oubridge et al., 1994). In the free protein, the helix lies almost parallel to β_4 across the RNA binding surface (Avis et al., 1996) and experiences conformational mobility (Mittermaier et al., 1999). In URNdesign, helix C is swung away from the β sheet, as in the design target, even if neither RNA nor the hydrophobic residues present in U1A protein (I92, I93, and M96) that pack the helix against the β sheet are present in URNdesign.

Submission of URNdesign to the DALI server identifies U1A and other RRMs as the structures closest to

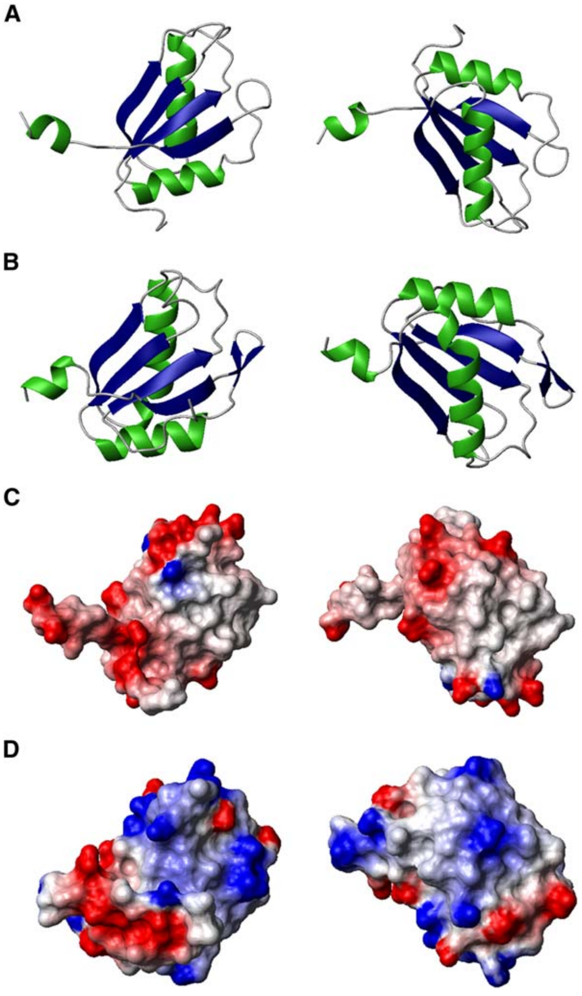

Figure 3. Comparison of the U1A and URNdesign Protein Structures

(A) Structure of URNdesign in two orientations rotated by ~180.

(B) Structure of the U1A protein (Oubridge et al., 1994). The U1A and URNdesign structures were superposed then shifted side by side to emphasize their remarkable similarity in overall fold and also in regards to many of the structural details.

(C) Surface electrostatic potential of URNdesign; the protein surface is largely acidic.

(D) Surface electrostatic potential of the U1A protein. The protein surface is very basic; the RNA binding surface is the basic patch in the upper right corner of the figure on the left.

URNdesign, together with other proteins (for example, phosphoribosyl-aminoimidazole synthetase) with the same split αβ fold. Backbone superposition of the URNdesign core structure with that of the U1A X-ray crystal structure on which the design was based confirms the remarkable similarity in the backbone conformation between the two structures (rmsd = 0.99 Å) and therefore that the design was very successful (Figures 3A and 3B). This difference increases when the partially disordered C-terminal helix is included, but it remains very low at only 1.45 Å. The major difference between the two structures is the conformation of the loop connecting the second and third strands of the β sheet. However, this difference is not structurally significant; in both the free U1A protein (Mittermaier et al., 1999) and URNdesign (see below), the loop is in conformational exchange, and only in the protein-RNA complex is it held in place rigidly through extensive interaction with the RNA.

Core residues that hold the structure together show similar contacts in the two proteins, even when their identity is changed (Figures 2B and 2C). For example, Phe37 stabilizes the position of helices α1 and α2 against the β sheet in U1A; it is replaced by Gln in URNdesign. At the same time, Ala68 (a highly conserved residue among RRM proteins) is replaced by Phe, with the aromatic ring occupying almost the same spatial position as Phe37. It is noticeable that there are no salt bridges in URNdesign. While the total number of polar versus hydrophobic residues has not been altered drastically between the two structures, the number of hydrophobic residues in the structured regions of the protein (α helices and β sheets) has increased from 22 to 29, while the number of polar residues was reduced from 28 to 17.

Sequence Comparison between U1A and URNdesign

The identity of residues that are conserved or changed between the two structures is highly informative (Figure 1A). For example, all three Glys in U1A were retained in URNdesign to stabilize tight turns or pack hydrophobic amino acids; by keeping the backbone rigid in the design, apart from its initial regularization, it is difficult (perhaps impossible) to change these Gly residues.

The RRM motif is identified at the sequence level by two highly conserved segments of 8 and 6 amino acids, referred to as RNP1 and RNP2 (Nagai et al., 1995; Varani and Nagai, 1998). These segments reside in the two central strands of the β sheet (Nagai et al., 1990) and play a key role in promoting RNA recognition (Allain et al., 1996; Oubridge et al., 1994). Residues within RNP1 and RNP2 that face the protein core are generally retained in URNdesign, while surface-exposed residues from these motives are not. Most RRM proteins contain either Tyr or Phe at the third (residue 54 in U1A) and fifth positions (residue 56) of RNP1 and in the second position (residue 13) of RNP2. These residues make intermolecular interactions with RNA, although U1A contains a Gln at position 54. In URNdesign, these residues are replaced by a Lys at position 54 and Leu at position 56. Within the RNP1 sequence (covering β3), residue 53 is Gly in nearly all RRMs (Birney et al., 1993) and is retained. Phe59 and the very highly conserved Ala55 (Ala or Gly in most RRMs) are not changed, and Val57 is conservatively substituted to Ile. In contrast, Arg52 (a key RNA binding residue in U1A) is changed to Gln. Within RNP2 (covering β1), the RNA-facing Tyr13 is replaced with Ile, while Ile12 is conservatively substituted. However, Ile14, Asn16, and Leu17, all highly conserved in RRM proteins, are replaced, while Asn15, an RNA-facing, surface-exposed residue, is retained.

Outside of RNP1 and RNP2, residues involved in packing the RRM hydrophobic core are generally conserved or conservatively substituted, but with some interesting exceptions. Within helix α1, Leu30 is retained, but in the loop following α1, Phe37 (almost always Phe/Tyr) is changed to Gln, while Gly38 (just as highly conserved) is retained. Within β2, both Ile40 and Val45 are retained. Ala68 in the helix immediately following β3 is nearly universally conserved as Ala, yet it is changed

Table 3. Rotamer Recovery

	All			Buried			Middle			Surface		
	#Cor	#Tot	Frac	#Cor	#Tot	Frac	#Cor	#Tot	Frac	#Cor	#Tot	Frac
CHI_1	48	83	0.58	12	17	0.71	17	30	0.57	19	36	0.53
CHI_2	23	33	0.70	7	8	0.88	12	13	0.92	4	12	0.33
CHI_3	3	6	0.50	0	0	N/A	2	4	0.50	1	2	0.50

"Buried" refers to amino acids with more than 19 neighboring amino acids; "middle" refers to those with more than 13 but less than 19 neighboring amino acids; "surface" refers to those with less than 13 neighboring amino acids. "#Cor" represents the number of rotamers with correct χ, "#Tot" represents the total number of rotamers in each category, and "Frac" represents the fraction of correctly predicted χ rotamers. "*CHI_1*" refers to statistics for all $\chi1$ rotamers, "*CHI_2*" refers to statistics for $\chi2$ rotamers of amino acids for which $\chi1$ was correct, and "*CHI_3*" refers to statistics for $\chi3$ rotamers of amino acids for which $\chi1$ and $\chi2$ were correctly predicted.

to Phe in URNdesign to interact with the residue that has substituted Phe37 (Figure 2C), although Ala65 (nearly as well conserved) is retained and Ile69 is conservatively substituted to Val. The Gly at position 74 in the loop connecting $\alpha2$ and $\beta4$ remains as Gly, while Asp79 in U1A (most often a Gly in other RRMs) is changed to Gly, bringing URNdesign closer to the RRM consensus than U1A. Concerning the hydrophobic residues on the inner face of $\beta4$, Val82 is conservatively substituted, while Ile84 is retained.

Much lower levels of conservation are observed in residues that are divergent in RRMs yet known to be important for RNA binding. Charged residues in the U1A structure are concentrated in the loop between β_1 and helix A (Lys20, Lys22, and Lys23) and the loop between β_2 and β_3 (Arg47, Lys50, and Arg52). Of these residues, only Lys23 is conserved in URNdesign, and the side chain is oriented 180° away from the position seen in the target structure. The very basic character of the $\beta2$-$\beta3$ loop as well as two functionally important Lys residues within $\alpha3$ (Lys96 and Lys98) have been switched completely by the inclusion of 4 Asp residues in place of those basic residues in URNdesign. The surface charge distribution reflects this marked difference between the two proteins. The U1A surface has large basic areas involved in RNA binding that are completely missing in URNdesign, which is instead acidic (Figures 3C and 3D). When we tested if URNdesign would bind RNA by electrophoretic band shift experiments, we saw no evidence for an interaction at protein concentrations as high as 10 μM; U1A binds its cognate RNA with a subnanomolar dissociation constant (not shown).

Evaluation of the Design

Analysis of the amino acid side chain conformations in the NMR structure of URNdesign shows good recovery of the rotamers identified in the redesign, with core residues showing the highest recovery (Table 3). About 58% of all $\chi1$ angles were found experimentally to be in the same rotameric state predicted computationally, and these include 71% of residues in the protein core. For residues with the correctly predicted $\chi1$, recovery of $\chi2$ is 70% overall and 88% for core residues. While surface residues show lower rotamer recovery, at least 50% or more of surface rotamers are still recovered in each category. These results strongly suggest that the design process has identified a global energy minimum.

The experimentally determined structures of U1A and URNdesign were then reexamined in the context of the current *RosettaDesign* energy function. The native U1A crystal structure (Oubridge et al., 1994) has a better *Rosetta* energy compared to the URNdesign NMR structure, but both have better energies than the NMR structure of U1A (Avis et al., 1996). Most likely, this difference reflects the higher quality of crystallographic structures as well as the improvement of NMR structures over the last 10 years. We then redesigned the URNdesign NMR backbone with the most current *RosettaDesign* energy function, allowing only the amino acid present in native human U1A or the designed residue from URNdesign at each position (a binary choice only). In this test, all of the core positions return the URNdesign sequence, suggesting that the protein core cannot be further optimized by simple single amino acid reversions. Although about 25% of all surface residues switch back to wild-type sequence, there are only marginal improvements in energy. However, when the URNdesign backbone was completely redesigned by allowing all 19 amino acids at each position (except Cys, which is not yet modeled), as in the original design of the URN design sequence, 82% (75% for the core) of the sequence was changed. Of the amino acids that do not change in this second redesign exercise, 40% were unchanged in the original design (i.e., they are the same as in U1A). Therefore, *RosettaDesign* predicts that the URNdesign sequence is not the global energy minimum for the backbone observed in the URNdesign NMR structure: in other words, the designed NMR structure is completely redesignable. The implications of this result are discussed below.

Dynamics of the Redesigned Protein

Relaxation times (^{15}N T$_1$, ^{15}N T$_2$, and ^{1}H-^{15}N NOEs) were measured for URNdesign by the standard techniques described in the Experimental Procedures and can be compared to those obtained for a construct of free U1A protein including amino acids 2–102 (Mittermaier et al., 1999). In both U1A and URNdesign, ^{15}N T$_1$ values are between 500 and 550 ms, with the exception of unfolded residues at the N and C termini (Figure 4). The T$_2$ values are also very similar between the two proteins. However, the data were collected at two different fields (600 Mhz for U1A; 500 Mhz for URNdesign), suggesting that URNdesign would have a slightly increased correlation time compared to U1A. Indeed, the overall correlation time for URNdesign is larger than for U1A and was calculated by ModelFree (Lipari and Szabo, 1982a, 1982b) to be 7.4 ns compared to the reported value of 6.1 ns for U1A. The most likely explanation for this

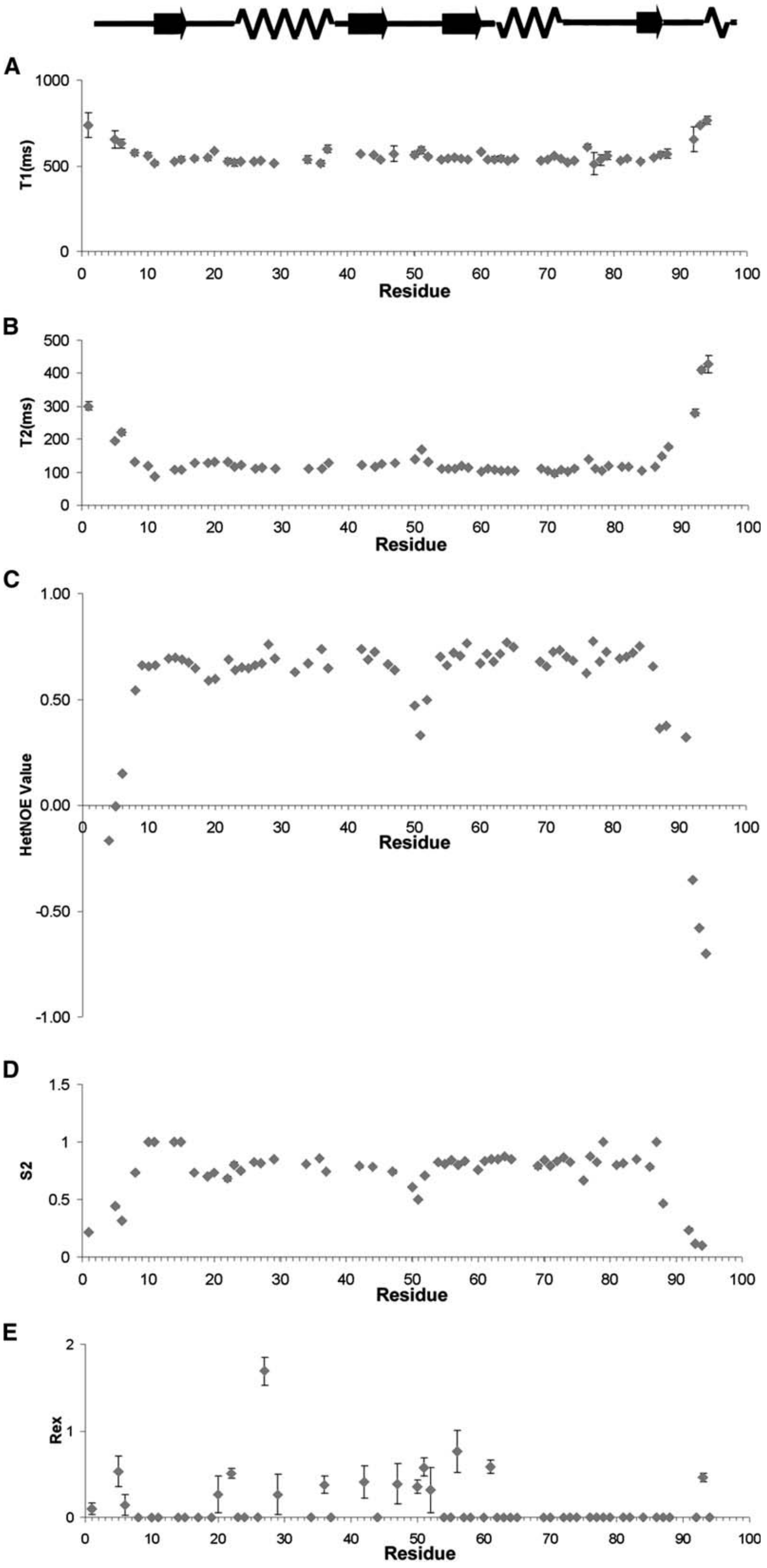

Figure 4. Backbone Dynamics of the URN design Protein

(A) ^{15}N T$_1$; the error estimate for the T$_1$ and T$_2$ rate constants reflects the likely error of the best fit obtained for a perfect exponential decay.

(B) ^{15}N T$_2$.

(C) ^{15}N heteronuclear NOEs; uncertainties in these measurements were estimated from the base plane noise in 2D ^{1}H-^{15}N HSQC spectra recorded with and without proton saturation.

(D) Order parameters (S^2) calculated assuming an axially symmetric diffusion tensor with $D_{iso} = 1.858 \times 10^7\ s^{-1}$ and $D_{par}/D_{perp} = 0.69$.

(E) Exchange contribution R_{ex} to T$_2$ obtained by ModelFree; values less than ~1 are not considered significant.

difference is partial folding of the C-terminal helix, which creates hydrodynamic drag for the entire protein.

The ^{15}N T$_2$ values are longer than average in U1A for residues in the two central strands of the β sheet (β1 and β3), loop 3 (connecting β2 and β3), helix C, and the loop connecting β4 with helix C. When the relaxation data were analyzed with ModelFree, residues in loop 3 could not be fit well without including significant R_{ex} values and had decreased order parameters as well, validating the observation of conformational exchange.

These observations reflected conformational exchange on the U1A protein RNA binding surface that is not present in URNdesign. If anything, increased nanosecond-picosecond motion in the β2-β3 loop of URNdesign is revealed by lower NOEs and slightly increased T2 values, without any evidence for slower conformational dynamics in this or other regions of the protein (R_{ex} values are all insignificant, with one exception; Figure 4E).

Discussion

A key goal of protein design is to reliably predict the sequence of amino acids capable of folding into a predefined 3D structure. In a demanding test of the *Rosetta* algorithm, 9 proteins of about 100 amino acids chosen to represent distinct folds were redesigned computationally to fold into the native structures (Dantas et al., 2003). In three out of nine cases, unique sequences with the characteristics of native proteins were identified: clear melting transitions were observed between folded and unfolded structures, and well-defined 1D NMR spectra were obtained (Dantas et al., 2003). In order to establish the accuracy of the design, we determined the structure and dynamics of the redesigned U1A protein, a protein studied by us extensively in the past (Allain et al., 1996; Avis et al., 1996; Mittermaier et al., 1999; Oubridge et al., 1994; Varani et al., 2000).

U1A Protein Was Redesigned with Atomic-Level Accuracy

The URNdesign protein reproduces the target structure to within less than 1 Å for the core αβ region of the domain (amino acids 8–88) (Figures 3A and 3B); this difference is comparable to the uncertainty in the coordinates observed over the 20 best NMR-derived structures (Figure 2A). This difference is much smaller than the structural divergence observed among different RRM proteins. In the superfamily, the orientation of the α helices with respect to the β sheet, the length of the β strands and of the loops connecting the secondary structural elements, and even the presence of additional structural elements C-terminal to the domain provide for considerably greater structural diversity than the small difference observed between U1A and URNdesign (Figures 3A and 3B). A short C-terminal helix was observed as well, as designed, although, in the native protein, this addition to the canonical RRM fold is held in place primarily by interactions with RNA (Allain et al., 1996; Oubridge et al., 1994). Even including this less well-ordered α helix, the redesigned structure differs from the target by only ~1.4 Å.

The design of the URNdesign sequence was executed while keeping the native backbone coordinates fixed after an initial regularization. As a consequence, sequence memory was retained during the sequence selection, although the overall computation was in itself unbiased, because the fixed backbone imposes significant steric restraints in the protein core. It has been previously observed that the fixed backbone approximation can lead to rejection of sequences of lower energies that may adopt very similar (and for many design purposes effectively the same) folds (Pokala and Handel, 2001). Thus, the divergence between the U1A and URNdesign sequences is much smaller than the sequence space observed for RRM proteins or, more generally, for proteins belonging to the split αβ fold of which the RRM is an example. In the absence of functional constraints, as in the URNdesign computational redesign, one might expect an even higher tolerance for sequence changes.

These results demonstrate that the computational redesign of U1A protein has yielded a structure remarkably similar to the target, even while simultaneously changing 70% of all amino acids (65/96 amino acids). Three recent examples of successful designs of complex protein folds concerned a new three-helix bundle (Walsh et al., 1999); the new αβ fold of Top7, a structure not yet observed in nature (Kuhlman et al., 2003); and the redesign of the hydrophobic core of ubiquitin (Johnson et al., 1999). Together, these studies demonstrate that structural specificity and atomic-level accuracy can be obtained for proteins of at least 100 amino acids by computational design methods.

Origin of the Increased Thermodynamic Stability of the Redesigned Protein

Although U1A is a very stable protein in its own right, URNdesign was found to be more stable by more than 2 kcal/mol by guanidinium-induced and thermal denaturation experiments (Dantas et al., 2003). It is likely that the additional thermodynamic stability of URN design arises from improved packing interactions within the protein core. The sequence optimization used in the design focuses solely on the stability of the native state through the maximization of pair-additive interactions and emphasizes hydrophobic contacts. As a consequence, the number of buried hydrophobic residues increases from 22 to 29, but no salt bridge is retained. Natural selection obviously has additional requirements: proteins must be not just thermodynamically stable, but they must fold to the native structure without aggregating and must perform their function. RNA binding may not be possible if an RRM protein is structurally too rigid (Crowder et al., 2001). As demonstrated most clearly by ultracentrifugation studies, URNdesign does not aggregate or dimerize appreciably at millimolar concentrations, and it has the folding characteristics associated with native proteins (Scalley-Kim and Baker, 2004). However, the absence of salt bridges in the design (that confer structural specificity) identifies a potential limitation of the scoring function noted in other design tests as well (G.D. et al., submitted).

A related question is whether the design algorithm has captured just a local energy minimum in an otherwise ragged energy landscape, or has instead reproduced the smooth energy landscapes of small natural proteins. The rigid monomeric structure and the highly cooperative folding transition observed for URNdesign (Scalley-Kim and Baker, 2004), together with the results of protein dynamics studies, suggest that a global energy minimum has indeed been found. The ^{15}N relaxation times and ModelFree-derived order parameters were similar to those observed for U1A (Mittermaier et al., 1999) (Figure 4). Conformational flexibility was also observed in the same loop regions for U1A and URN design, namely, for the loops connecting β1 with α1 and β2 with β3. These are the "jaws" of the protein (Nagai et al., 1990) and represent primary sites for specific recognition of RNA (Kenan et al., 1991; Varani and Nagai, 1998). URNdesign was designed without any consideration of function, and many surface-exposed amino acids involved in RNA binding were mutated to acidic residues (Figures 3C and 3D), resulting in complete loss of RNA binding. However, the functionally important conformational flexibility of these loops (Allain et al., 1996; Mittermaier et al., 1999) was retained. The greatest difference in dynamics between native and redesigned proteins was the disappearance of conformational exchange at the interface between helix C and the β sheet. However, helix C was designed to occupy the position observed in the U1A-RNA complex and not in the free protein, away from the β sheet surface.

Together, these observations suggest that not just the location of the global energy minimum was unchanged by the design, but also that the shape of the energy landscape was not altered, at least in regards to the protein backbone. It will be interesting to extend studies of dynamics to backbone carbonyls and to the side chains to establish whether the native-like rigidity observed for the backbone is retained in the hydrophobic core.

Conclusions

The present study demonstrate that computational design algorithms can redesign thermodynamically stable proteins of at least 100 amino acids while exquisitely maintaining the fidelity of the structure and dynamics observed in the native protein. The design based on the URNdesign coordinates indicates that even the subtle backbone changes observed between URNdesign and the native protein structure allows the former sequence to be redesigned to a divergent new sequence. In other words, small changes in the backbone coordinates (1 Å rmsd) are sufficient to expand significantly the sequence diversity within the hydrophobic core. This result indicates that truly dramatic increases in sequence diversity and perhaps protein stability could be obtained by incorporating even limited backbone flexibility into the design algorithm. An interesting future challenge would be the design of proteins that retain the characteristic RRM fold but have completely lost sequence homology with the superfamily.

Experimental Procedures

Protein Expression and Purification

Two versions of the URNdesign protein were prepared, with an S-peptide tag at the N terminus or a His tag at the C terminus. Most experiments were conducted on this second construct with the sequence: MDSPDLGSTPPHTEPSQVVLITNINPEVPKEKLQAL LYALASSQGDILDIVVDLSDDNSGKAYIVFATQESAQAFVEAFQGYPF QGNPLVITFSETPQSQVAEDGSL. Sequence alignments were performed by using NCBI BLAST2 (Basic Local Alignment Search Tool) online from the European Bioinformatics Institute homepage (www.ebi.ac.uk/services).

Proteins were expressed in *E. coli* BL21 (DE3) from a pET29b vector. Uniformly ^{15}N- and ^{15}N/^{13}C-labeled samples were prepared by growing bacteria in M9 minimal media supplemented with 0.5 gl^{-1} ^{15}N-NH$_4$Cl and 2 gl^{-1} ^{13}C-glucose (Spectra Isotope). The protein was isolated by ion-exchange chromatography on a DEAE fast flow Sepharose column. Further purification steps included ion-exchange chromatography on a Resource Q column followed by size exclusion fractionation on Superdex 75 (all by Amersham). The final yield was about 20 mg/L of culture. Sample purity and molecular mass were verified by SDS-PAGE and matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectroscopy.

NMR Data Collection and Analysis

All URNdesign samples were prepared for NMR experiments in Shigemi susceptibility-matched NMR tubes, at 0.7–1.0 mM concentration in H_2O solution containing 5% or 100% ^2H_2O, 50 mM sodium phosphate, and 0.01 mM EDTA (pH 6.8). All experiments were recorded at 25ºC unless otherwise specified. Triple resonance NMR experiments were collected on a Bruker Avance 500 MHz spectrometer equipped with a TXI HCN triple resonance probe with triple axis gradients. 3D ^{15}N-edited NOESY spectra and 2D NOESY and TOCSY data sets were recorded on a Bruker Avance 750 MHz spectrometer equipped with a TXI HCN triple resonance probe with a z axis gradient. 3D ^{13}C-edited NOESY spectra were recorded at Environmental Molecular Sciences Laboratory (EMSL) at PNNL in Richland, WA by using a Varian 600 MHz spectrometer equipped with a cryoprobe. Data were processed with NMRPipe (Delaglio et al., 1995) and analyzed with Sparky (Goddard and Kneller, 2006).

Backbone amide ^{1}H and ^{15}N, C_α, C=O, and side chain C_β resonances were assigned by using HNCO, HNCACB, CBCA(CO)NH, HBHA(CO)NH, HN(CO)CA, and 3D ^{15}N-edited TOCSY experiments (Sattler et al., 1999). Side chain assignments were obtained by analysis of 3D HCCH-TOCSY and 3D ^{13}C-edited NOESY experiments. Aromatic side chain assignments were obtained from 2D NOESY and TOCSY spectra recorded in D_2O and through the analysis of ^{13}C-edited spectra optimized for detection of aromatic resonances. The spectra used in deriving distance constraints included 3D ^{15}N-edited NOESY and 3D ^{13}C-edited NOESY, 2D NOESY in H_2O, and 2D NOESY in ^2H_2O recorded at 750 MHz with mixing times of 100 ms.

Structure Determination

Protein structure determination was conducted in a semiautomated iterative manner by using CYANA2.0 (Güntert, 2003). The NOESY peak lists used as input for automated analysis were generated automatically with Sparky (Goddard and Kneller, 2006) based on the chemical shift list generated in the assignment process. After the first few rounds of calculations, the spectra were analyzed again to identify additional crosspeaks consistent with the structural model and to remove misidentified NOEs. Slowly exchanging amides were identified by lyophilizing the protein from H_2O then dissolving it in D_2O; hydrogen bond donors were identified by the presence of an amide peak in the HSQC recorded after 30 min. The corresponding acceptors were attributed by visualizing coordinates obtained from CYANA without any hydrogen bonding constraint to identify carbonyl groups that were at a distance of approximately 2.0 Å from slow exchanging amides. Hydrogen bonding constraints were then added at this stage of the refinement. TALOS (Cornilescu et al., 1999) was used to generate ϕ and ψ dihedral angle constraints. Residues for which the prediction was deemed to be "good" (9 out of 10 best-fit residues clustered together within allowed parts of the Ramachandran plot) were used to generate a dihedral constraint list.

The CYANA2.0 program immediately gave target functions in the correct range for a "good" structure, 135 A^2 and 7.5 A^2. After several rounds of refinement, the final run gave a final target function of 2.05 A^2 with no upper distance and angle violations greater than 0.2 Å and 5º, respectively. However, CYANA2.0 confers greater weight to van der Waals contacts, and nine close atom contacts remained violated after the final CYANA2.0 round. The quality of the structure was evaluated with Procheck (http://rcsb-deposit.rutgers.edu). Experimental statistics are reported in Table 1, and structural statistics are reported in Table 2.

Protein Dynamics

Standard pulse sequences were used to measure the ^{15}N T_1, ^{15}N T_2, and heteronuclear NOEs (Farrow et al., 1994) essentially as we described recently (Deka et al., 2005). Spectra were recorded with 112 complex points in the indirect dimension and with delays of 0.010, 0.050, 0.100, 0.150, 0.200, 0.250, 0.300, 0.350, 0.400, 0.500, and 0.600 s for the T_1 experiments and 0.008, 0.016, 0.024, 0.032, 0.040, 0.048, 0.064, 0.080, 0.096, 0.112, and 0.120 s for T_2. The relaxation delay was 1.9 s. For the heteronuclear NOE measurements, a pair of spectra was recorded with and without proton saturation, which was achieved by application of ^{1}H 120º pulses every 5 ms. Spectra recorded with proton saturation utilized a 2 s recycle delay followed by a 3 s period of saturation, while spectra recorded in the absence of saturation employed a recycle delay of 5 s. Linear prediction was applied in the indirect dimension to increase the number of complex points to 224. Peak heights were calculated for every assigned peak in the T_1 and T_2 spectra and fitted into an exponential curve by using Sparky (Goddard and Kneller, 2006). The error estimates for the rate constants reflect the likely error of the best fit from the parameters obtained for a perfect exponential decay. Heteronuclear NOE values were calculated from the ratio of peak heights for spectra recorded with and without proton saturation. Errors in these measurements were estimated from the plane base noise in the spectra. Analysis of the relaxation data was conducted by using the ModelFree model (Lipari and Szabo, 1982a, 1982b) exactly as we recently reported (Deka et al., 2005).

Acknowledgments

We are grateful to the National Magnetic Resonance Facility at Madison (Madison, WI, supported by the National Institutes of Health [NIH]) for access to instrumentation and Pacific Northwest National Laboratory, Richland, WA (supported by the Department of Energy). We would also like to thank Dr. Jack Schonbrun for help with calculating rotamer-recovery statistics. This work was supported by grants from NIH-National Institutes of General Medical Sciences to G.V. and to D.B.

Received: July 21, 2005
Revised: December 22, 2005
Accepted: February 21, 2006
Published: May 16, 2006

References

Allain, F.-H.T., Gubser, C.C., Howe, P.W.A., Nagai, K., Neuhaus, D., and Varani, G. (1996). Specificity of ribonucleoprotein interaction determined by RNA folding during complex formation. Nature *380*, 646–650.

Avis, J., Allain, F.H.-T., Howe, P.W.A., Varani, G., Neuhaus, D., and Nagai, K. (1996). Solution structure of the N-terminal RNP domain of U1A protein: the role of C-terminal residues in structure stability and RNA binding. J. Mol. Biol. *257*, 398–411.

Birney, E., Kumar, S., and Krainer, A.R. (1993). Analysis of the RNA-recognition motif and RS and RGG domains: conservation in Metazoan pre-mRNA splicing factors. Nucleic Acids Res. *21*, 5803–5816.

Cornilescu, G., Delaglio, F., and Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J. Biomol. NMR *13*, 289–302.

Crowder, S., Holton, J., and Alber, T. (2001). Covariance analysis of RNA recognition motifs identifies functionally linked amino acids. J. Mol. Biol. *310*, 793–800.

Dahiyat, B.I., and Mayo, S.L. (1997). Probing the role of packing specificity in protein design. Proc. Natl. Acad. Sci. USA *94*, 10172–10177.

Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J. Mol. Biol. *332*, 449–460.

Deka, P., Paranj, P.K., Perez-Canadillas, J.M., and Varani, G. (2005). Protein and RNA dynamics play key roles in determining the specific recognition of GU-rich polydenylation regulatory elements by human Cstf-64 protein. J. Mol. Biol. *347*, 719–733.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J., and Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J. Biomol. NMR *6*, 277–293.

Dwyer, M.A., Looger, L.L., and Hellinga, H.W. (2004). Computational design of a biologically active enzyme. Science *304*, 1967–1971.

Farrow, N.A., Muhandiram, R., Singer, A.U., Pascal, S.M., Kay, C.M., Gish, G., Shoelson, S.E., Pawson, T., Forman-Kay, J.D., and Kay, L.E. (1994). Backbone dynamics of a free and a phosphopeptide complexed Src homology 2 domain studied by ^{15}N NMR relaxation. Biochemistry *33*, 5984–6003.

Goddard, T.D., and Kneller, D.G. (2006). Sparky 3 (http://www.cgl.ucsf.edu/home/sparky/).

Güntert, P. (2003). Automated NMR protein structure calculation. Prog. Nuclear Magn. Res Spectrosc. *43*, 105–125.

Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T., and Kim, P.S. (1998). High-resolution protein design with backbone freedom. Science *282*, 1462–1467.

Havranek, J.J., and Harbury, P.B. (2003). Automated design of specificity in molecular recognition. Nat. Struct. Biol. *10*, 45–52.

Johnson, E.C., Lazar, G.A., Desjarlais, J.R., and Handel, T.M. (1999). Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. Structure *7*, 967–976.

Kaplan, J., and DeGrado, W.F. (2004). *De novo* design of catalytic proteins. Proc. Natl. Acad. Sci. USA *101*, 11566–11570.

Kenan, D.J., Query, C.C., and Keene, J.D. (1991). RNA recognition: towards identifying determinants of specificity. Trends Biochem. Sci. *16*, 214–220.

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

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. Science *302*, 1364–1368.

Lee, A.L., and Wand, A.J. (2001). Microscopic origin of entropy, heat capacity and the glass transition in proteins. Nature *411*, 501–504.

Lipari, G., and Szabo, A. (1982a). Model-free approach to the interpretation of nuclear magnetic relaxation in macromolecules. 1. Theory and range of validity. J. Am. Chem. Soc. *104*, 4546–4559.

Lipari, G., and Szabo, A. (1982b). Model-free approach to the interpretation of nuclear magnetic relaxation in macromolecules. 2. Analysis of experimental results. J. Am. Chem. Soc. *104*, 4559–4570.

Looger, L.L., Dwyer, M.A., Smith, J.J., and Hellinga, H.W. (2003). Computational design of receptor and sensor proteins with novel functions. Nature *423*, 132–133.

Mittermaier, A., Varani, L., Muhandiram, D.R., Kay, L.E., and Varani, G. (1999). Changes in side chain and backbone dynamics identify determinants of specificity in RNA recognition by human U1A protein. J. Mol. Biol. *294*, 967–979.

Mooers, B.H.M., Datta, D., Baase, W.A., Zollars, E.S., Mayo, S.L., and Matthews, B.W. (2003). Repacking the core of T4 lysozyme by automated design. J. Mol. Biol. *332*, 741–756.

Nagai, K., Oubridge, C., Jessen, T.H., Li, J., and Evans, P.R. (1990). Structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A. Nature *348*, 515–520.

Nagai, K., Oubridge, C., Ito, N., Avis, J., and Evans, P. (1995). The RNP domain: a sequence-specific RNA-binding domain involved in processing and transport of RNA. Trends Biochem. Sci. *20*, 235–240.

Oubridge, C., Ito, N., Evans, P.R., Teo, C.-H., and Nagai, K. (1994). Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. Nature *372*, 432–438.

Pokala, N., and Handel, T.M. (2001). Protein design—where we were, where we are, where we're going. J. Struct. Biol. *134*, 269–281.

Sattler, M., Schleucher, J., and Griesinger, C. (1999). Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog. NMR Spectrosc. *34*, 93–158.

Scalley-Kim, M., and Baker, D. (2004). Characterization of the folding energy landscapes of computer generated proteins suggests high folding free energy barriers and cooperativity may be consequences of natural selection. J. Mol. Biol. *338*, 573–583.

Shifman, J.M., and Mayo, S.L. (2003). Exploring the origins of binding specificity through the computational redesign of calmodulin. Proc. Natl. Acad. Sci. USA *100*, 13274–13279.

Varani, G., and Nagai, K. (1998). RNA recognition by RNP proteins during RNA processing and maturation. Annu. Rev. Biophys. Biomol. Struct. *27*, 407–445.

Varani, L., Gunderson, S., Kay, L.E., Neuhaus, D., Mattaj, I., and Varani, G. (2000). The NMR structure of the 38 kDa RNA-protein complex reveals the basis for cooperativity in inhibition of polyadenylation by human U1A protein. Nat. Struct. Biol. *7*, 329–335.

Walsh, S.T.R., Cheng, H., Bryson, J.W., Roder, H., and DeGrado, W.F. (1999). Solution structure and dynamics of a de novo designed three-helix bundle protein. Proc. Natl. Acad. Sci. USA *96*, 5486–5491.

Walsh, S.T.R., Lee, A.L., DeGrado, W.F., and Wand, A.J. (2001). Dynamics of a de novo deisgned three-helix bundle protein studied by 15N, 13C and 2D NMR relaxation methods. Biochemistry *40*, 9560–9569.

Wand, A.J. (2001). Dynamic activation of protein function: a view emerging from NMR spectroscopy. Nat. Struct. Biol. *8*, 926–931.

Accession Numbers

The coordinates for 20 NMR-derived URNdesign structures and NMR constraint files have been deposited with the RCSB Protein Data Bank under the identifier code 2A3J. The chemical shift list corresponding to this structure determination has been deposited in the BioMagRes Database (http://www.bmrb.wisc.edu) under accession number 6493.