doi:10.1016/S0022-2836(03)00888-X J. Mol. Biol. (2003) 332, 449–460

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins

Gautam Dantas[1]†, Brian Kuhlman[1]†, David Callender[1] Michelle Wong[1] and David Baker[1,2]*

[1]*Department of Biochemistry University of Washington Seattle, WA 98195, USA*

[2]*Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA*

**Corresponding author*

A previously developed computer program for protein design, RosettaDesign, was used to predict low free energy sequences for nine naturally occurring protein backbones. RosettaDesign had no knowledge of the naturally occurring sequences and on average 65% of the residues in the designed sequences differ from wild-type. Synthetic genes for ten completely redesigned proteins were generated, and the proteins were expressed, purified, and then characterized using circular dichroism, chemical and temperature denaturation and NMR experiments. Although high-resolution structures have not yet been determined, eight of these proteins appear to be folded and their circular dichroism spectra are similar to those of their wild-type counterparts. Six of the proteins have stabilities equal to or up to 7 kcal/mol greater than their wild-type counterparts, and four of the proteins have NMR spectra consistent with a well-packed, rigid structure. These encouraging results indicate that the computational protein design methods can, with significant reliability, identify amino acid sequences compatible with a target protein backbone.

Keywords: computational protein design; protein engineering; protein stability; protein thermodynamics

Introduction

The ultimate goal of protein design is the creation of novel proteins that perform specified tasks. A necessary requirement for meeting this goal is the ability to identify sequences that fold with sufficient stability into a target structure. Towards this end several laboratories have developed computer programs for identifying amino acid sequences compatible with a given protein.[1–8] A rigorous test for these models is the complete redesign of naturally occurring proteins. In such a test, the only information given to the method is the backbone coordinates of the protein to be redesigned. Although there has been considerable recent success in the field of computational protein design,[9–12] the pioneering zinc finger redesign by Mayo and co-workers is the only published report in which automated procedures have been used to completely redesign a naturally occurring protein backbone.[1]

Previously we demonstrated that our method for protein design, RosettaDesign, produces native-like sequences when run on a large test set of naturally occurring protein backbones.[13] On average, 30% of the residues were identical with their wild-type counterpart, and in the core the level of identity was 50%. These results suggested that RosettaDesign was performing well, but they did not indicate whether the design sequences would actually fold into the target structures. The RosettaDesign method has been applied successfully to the redesign of protein folding pathways,[14] backbone conformations,[15] and oligomerization states.[16] Here, to more rigorously test RosettaDesign and, more generally, to assess the consistency with which modern computational protein design methodology can completely redesign the sequences of small proteins, we make and characterize complete redesigns of nine globular proteins.

Like all automated procedures for protein

† G. D. and B. K. contributed equally to this work.

Present address: B. Kuhlman, Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599, USA.

E-mail address of the corresponding author: dabaker@u.washington.edu

design, RosettaDesign has two main components: an energy function that ranks the relative fitness of various amino sequences for a given protein structure and a search function for rapidly scanning sequence space. The energy function used by RosettaDesign is dominated by Lennard–Jones interactions, an orientation dependent hydrogen bonding potential,[17] and an implicit solvation model.[18] The Lennard–Jones term favors atoms being closely packed, but not too close to each other and therefore provides the steric information needed to correctly pack a protein core. The implicit solvation model penalizes the burial of polar atoms and therefore favors hydrophobic residues in the core of the protein and hydrophilic residues on the protein surface. The hydrogen bond term offsets the implicit solvation model by rewarding buried polar groups that form good hydrogen bonds. Amino acid specific reference energies approximate the average free energy of each of the amino acids in the denatured state.

Even for a small protein the size of sequence space is enormous and therefore it is not feasible to explicitly calculate the energy of every possible sequence. RosettaDesign uses a simple Monte Carlo optimization to identify low energy sequences. Starting from a completely random sequence, single amino acid substitutions are accepted or rejected using the Metropolis criterion. To make the search discrete, side-chain conformations are restricted to the backbone torsion angle dependent rotamer conformations in Dunbrack's library.[19] This optimization procedure converges to very similar sequences (70–80% identity) when multiple runs are started with different random sequences. Unlike the commonly used dead-end elimination algorithm, the Monte Carlo protocol does not guarantee that the final sequence will be at the global energy minimum, but the convergence observed from multiple runs strongly suggests that the search is not getting trapped in local minima. An advantage of the Monte Carlo protocol is that it is very fast, a typical search for a 100 residue protein takes approximately five minutes on a desktop computer.

Results

RosettaDesign was used to design sequences for nine globular proteins: the src SH3 domain, lambda repressor, U1A, protein L, tenascin, procarboxypeptidase, acylphosphatase, S6, and FKBP12 (Figure 1). For protein L, two sequences were chosen for experimental study. On average, the redesigned protein sequences are 35% identical with the wild-type sequence over all residues and 50% identical for the core residues (Table 1). In general, the overall amino acid composition in the redesigns is similar to that of the wild-type proteins, although a few of the redesigns are more hydrophobic than the wild-type protein. The redesign of S6 has the most dramatic change, 59% of the residues are non-polar in the redesign while

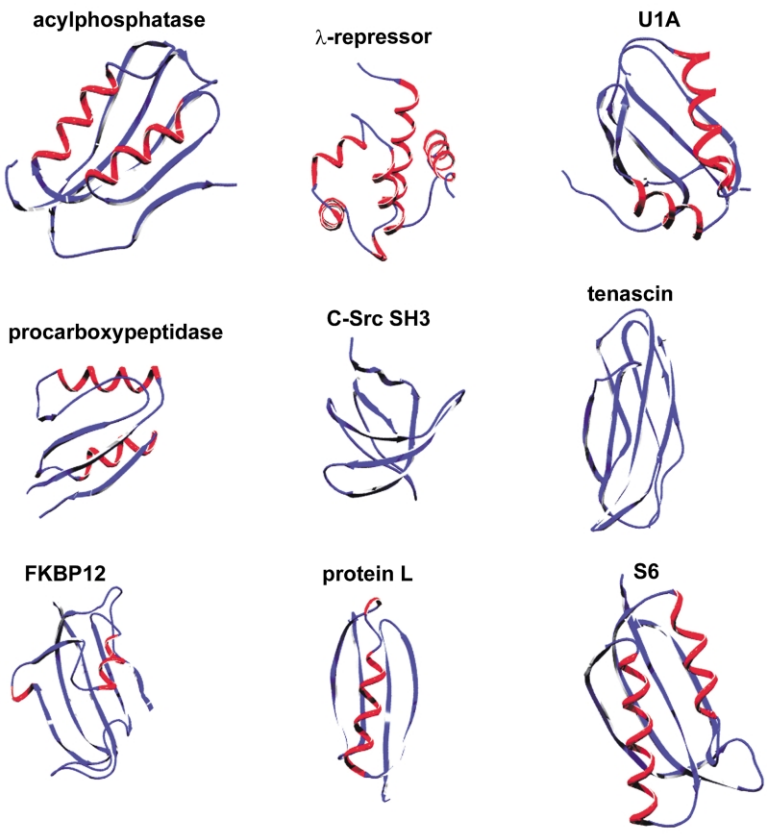

Figure 1. Ribbon diagrams of the nine redesigned structures. The PDB codes and respective residue numbers are: acylphosphatase (2acy, 1–98), lambda repressor (1lmb, 6–92), U1A (1urn, 2–97), procarboxypeptidase (1aye, 10–79), C-Src SH3 (1fmk, 83–142), tenascin (1ten, 803–890), FKB12 (1fkb, 1–107), protein L (1hz5, 1–62), S6 (1ris, 1–94).

Table 1. Sequence alignments comparing the wild-type sequences (WT) to the design sequences (D)

```
                  10        20        30        40        50
ACY-WT    AEGDTLISVDYEIFGKVQGVFFRKYTQAEGKKLGLVGWVQNTDQGTVQGQLQGPASKVRH
ACY-D     PTGDSYIQVKWQVKGDVTGNNFRKMVAEFAEALGLVGKVTYTDNGTVSGQVEGPAEQVLK

                  70        80        90
ACY-WT    MQEWLETKGSPKSHIDRASFHNEKVIVKLDYTDFQIVK
ACY-D     FLEWLARSGSPNADIKQTVFTNMTRIDRLTMETFKIDE

                  10        20        30        40        50        60
AYE-WT    DQVLEIVPSNEEQIKNLLQLEAQEHLQLDFWKSPTTPGETAHVRVPFVNVQAVKVFLESQGIAYSIMIED
AYE-D     KTIFVIVPTNEEQVAFLEALAKQDELNFDWQNPPTEPGQPVVILIPSDMVEWFLEMLKAKGIPFTVYVEE

                  10        20        30        40        50
FKB-WT    GVQVETISPGDGRTFPKRGQTCVVHYTGMLEDGKKFDSSRDRNKPFKFMLGKQEVIRGWE
FKB-D     GVTVVTQESGDGNNRPKPGELVIIFYTWMHKDGPPISSSADQGTPYRFYLGQNQVPEGLQ

                  70        80        90        100
FKB-WT    EGVAQMSVGQRAKLTISPDYAYGATGHPGIIPPHATLVFDVELLKLE
FKB-D     EAVANLSQGERVTIVIDSSKTYGETGLPGVVPPGTVLIFDVLLVQLV

                  10        20        30        40        50
FMK-WT    VTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYVAPSDS
FMK-D     TTLFVATSPYESTTDNDLPFRKGDKIWIEDNAPGDYWKAVSSTTGKTGYIPADKIRPAGA

                  10        20        30        40        50
LMB-WT    PLTQEQLEDARRLKAIYEKKKNELGLSQESVADKMGMGQSGVGALFNGINALNAYNAALL
LMB-D     GNSETEQAIAKRLQAIFEELAEELNLSQEKVATLIGGSKEEFEKQLKGQQSPNLERAKRF

                  70        80
LMB-WT    AKILKVSVEEFSPSIAREIYEMYEAVS
LMB-D     AEIFNVSISDFSEYLYRLYLEQLKERF

                  10        20        30        40        50        60
PL-WT     EVTIKANLIFANGSTQTAEFKGTFEKATSEAYAYADTLKKDNGEWTVDVADKGYTLNIKFAG
PL1-D     EKTVEANFIFADGKTTTIRFTGSEEEAKKRVLAYAEELKDTYGEYSVDYKNGGEQINIKFKG
PL2-D     DTTVRVIFIFADGKTTTIEFTGSEEAAKKQAQEYAQSLRDNYGDYSIDYQNGGELIKIVFSG

                  10        20        30        40        50
S6-WT     MRRYEVNIVLNPNLDQSQLALEKEIIQRALENYGARVEKVEELGLRRLAYPIAKDPQGYF
S6-D      YRVFIIIIYLDPTLSDEELKKLFEMILELLQKYGFDITAIYFQGETELDAPINGTKKAFL

                  70        80        90
S6-WT     LWYQVEMPEDRVNDLARELRIRDNVRRVMVVKSQ
S6-D      IVIIVVGPPDTVEEFRRALQSLPYVLQVEIVPYE

                  10        20        30        40        50
TEN-WT    LDAPSQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDENQYSIGN
TEN-D     LPPPYNITVTNIGPTTAVLVYVRSESPSDGYNITFGTKNDDSDRVTVTLPSENTSYVITN

                  70        80
TEN-WT    LKPDTEYEVSLISRRGDMSSNPAKETFTT
TEN-D     LKPNTTFQITIRSQNGDKSSPPVSTYFTL

                  10        20        30        40        50
U1A-WT    AVPETRPNHTIYINNLNEKIKKDELKKSLHAIFSRFGQILDILVSRSLKMRGQAFVIFKE
U1A-D     TPPHTEPSQVVLITNINPEVPKEKLQALLYALASSQGDILDIVVDLSDDNSGKAYIVFAT

                  70        80        90
U1A-WT    VSSATNALRSMQGFPFYDKPMRIQYAKTDSDIIAKM
U1A-D     QESAQAFVEAFQGYPFQGNPLVITFSETPQSQVAED
```

only 49% of the residues are non-polar in the wild-type protein.

Synthetic genes which place each of the ten protein sequences under the control of the T7 promoter, with a C-terminal 6× His tag, and a codon usage optimal for *Escherichia coli* were obtained from BlueHeron Biotechnologies. Following induction in *E. coli*, each of the proteins was clearly visible on Coomassie-stained SDS/polyacrylamide gels, and it was possible to purify all ten proteins to reasonable homogeneity using nickel affinity chromatography.

The folding and stability of each of the redesigned proteins was assessed using a battery of biophysical techniques. The extent of secondary structure in the completely redesigned proteins was assessed by circular dichroism spectroscopy (Figure 2). Size-exclusion chromatography was

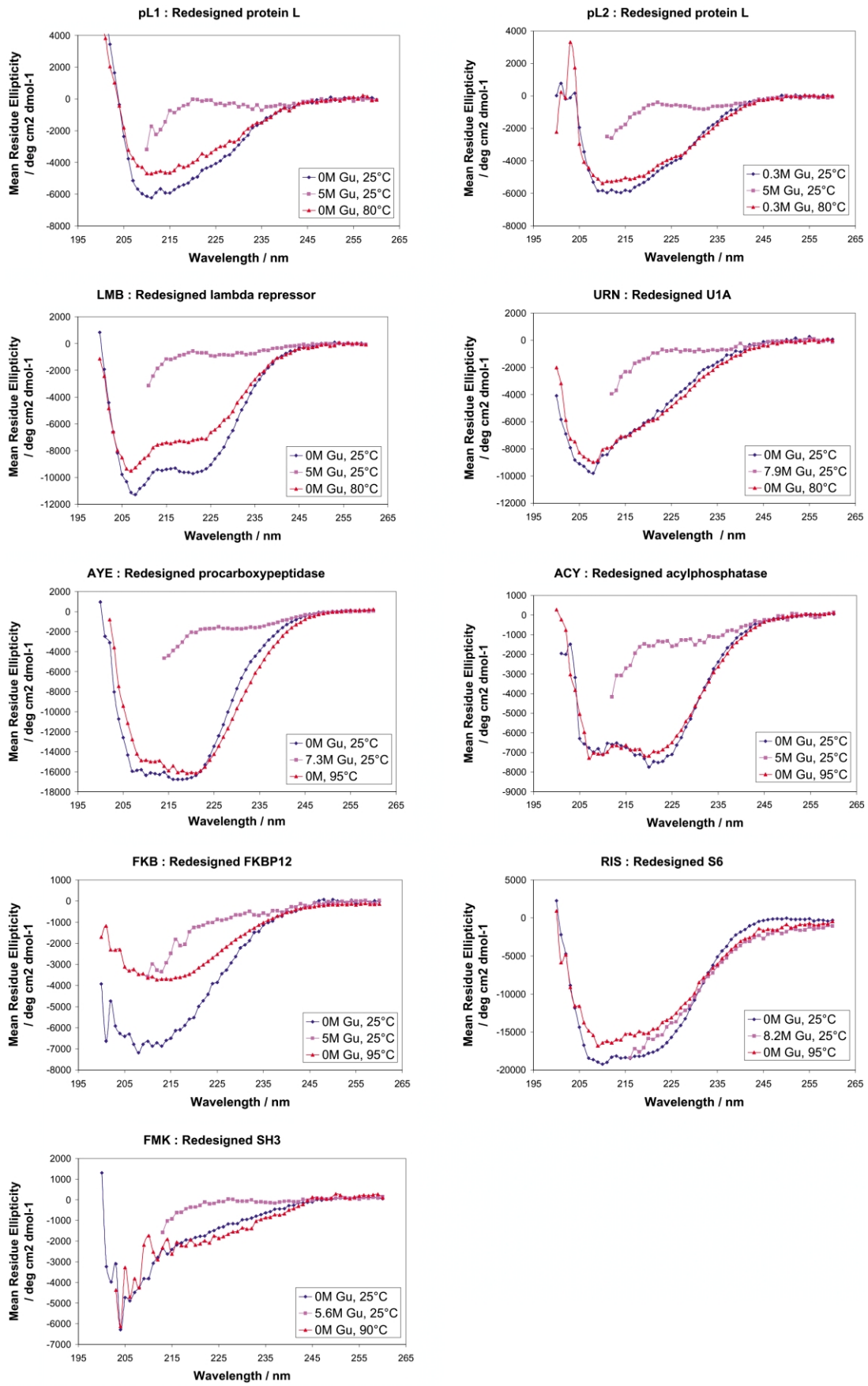

Figure 2. Circular dichroism spectra of the redesigned proteins. The CD spectra of eight of the redesigned proteins (pL1, pL2, LMB, URN, AYE, ACY, RIS) show the expected WT-like secondary structure content. The spectrum of redesigned src-SH3 resembles a random-coil. Far-UV CD spectra were collected on 15–25 μM protein samples in 50 mM sodium phosphate (pH 7.0) at 25 °C (blue), at 95–98 °C (red) or in 5–8 M Gu-HCl at 25 °C (pink).

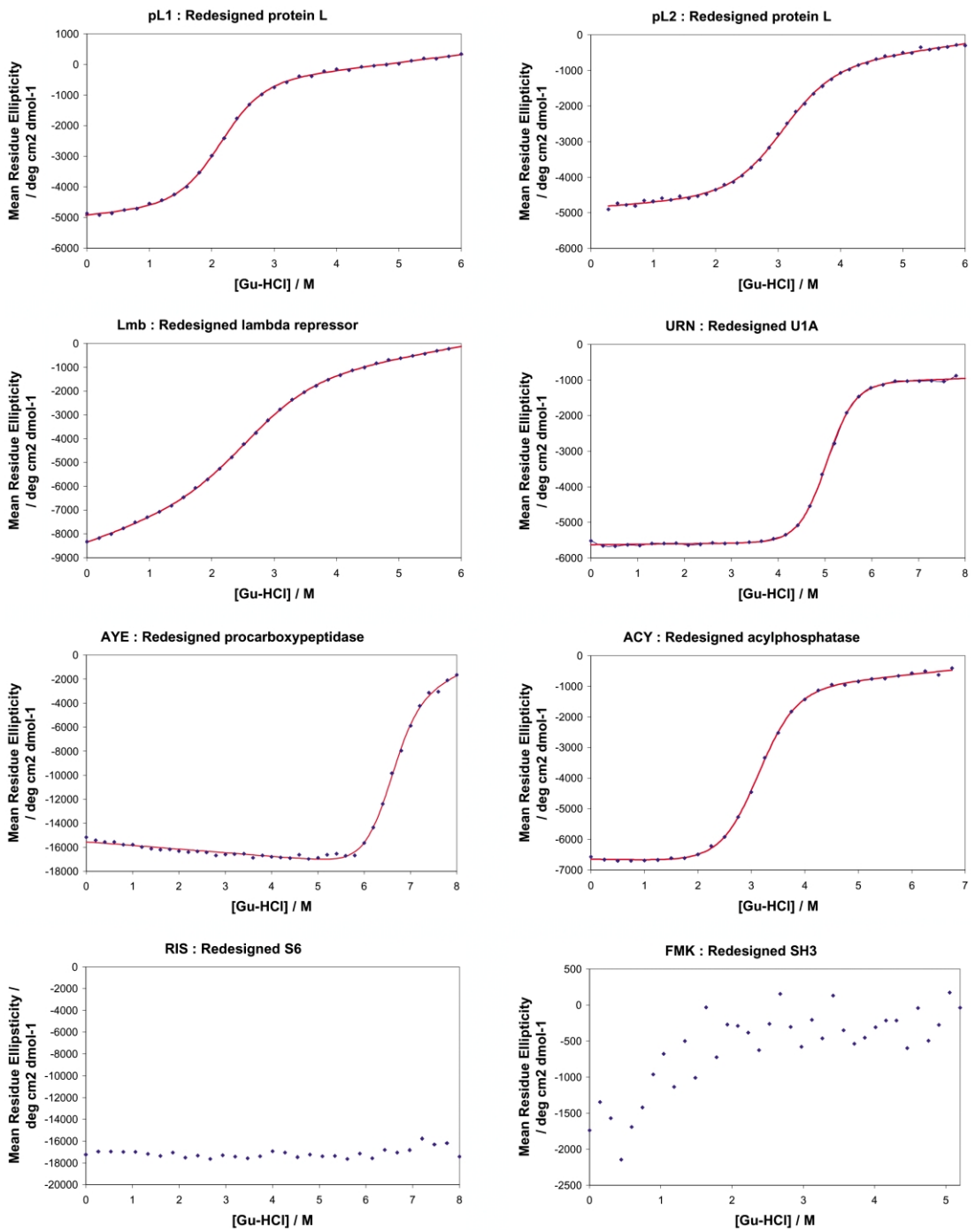

Figure 3. Chemical denaturation of the redesigned proteins. The Gu-HCl-induced denaturation profiles of seven of the redesigned proteins (pL1, pL2, LMB, URN, AYE, ACY) are two-state and co-operative. Redesigned S6 does not denature at any Gu-HCl concentration. The erratic melt of redesigned SH3 suggests that the protein adopts a random coil structure. Ellipticity at 220 nm was monitored as a function of Gu-HCl concentration for ~5 μM protein in 50 mM sodium phosphate, pH 7.0, 25 °C, in a 1 cm cuvette. The data were fit using a two-state model with a linear dependence of the free energy of unfolding (ΔG_U^{H2O}) on denaturant concentration. ΔG_U^{H2O} values are given in Table 3. The data sets used are averages of duplicate experiments with 30 separate denaturant concentrations.

used to determine if the proteins were monomeric (data not shown). Chemical (Figure 3) and thermal (Figure 4) denaturation experiments were used to confirm that the proteins were folded and to determine their stabilities. One-dimensional ^{1}H NMR experiments (Figure 5) were used to further confirm that the proteins were folded and to probe the rigidity of their structures. On the basis of the results from these experiments we were able to place the proteins into different categories: unfolded *versus* folded, lower or higher than WT stability, and more or less rigid (Table 2).

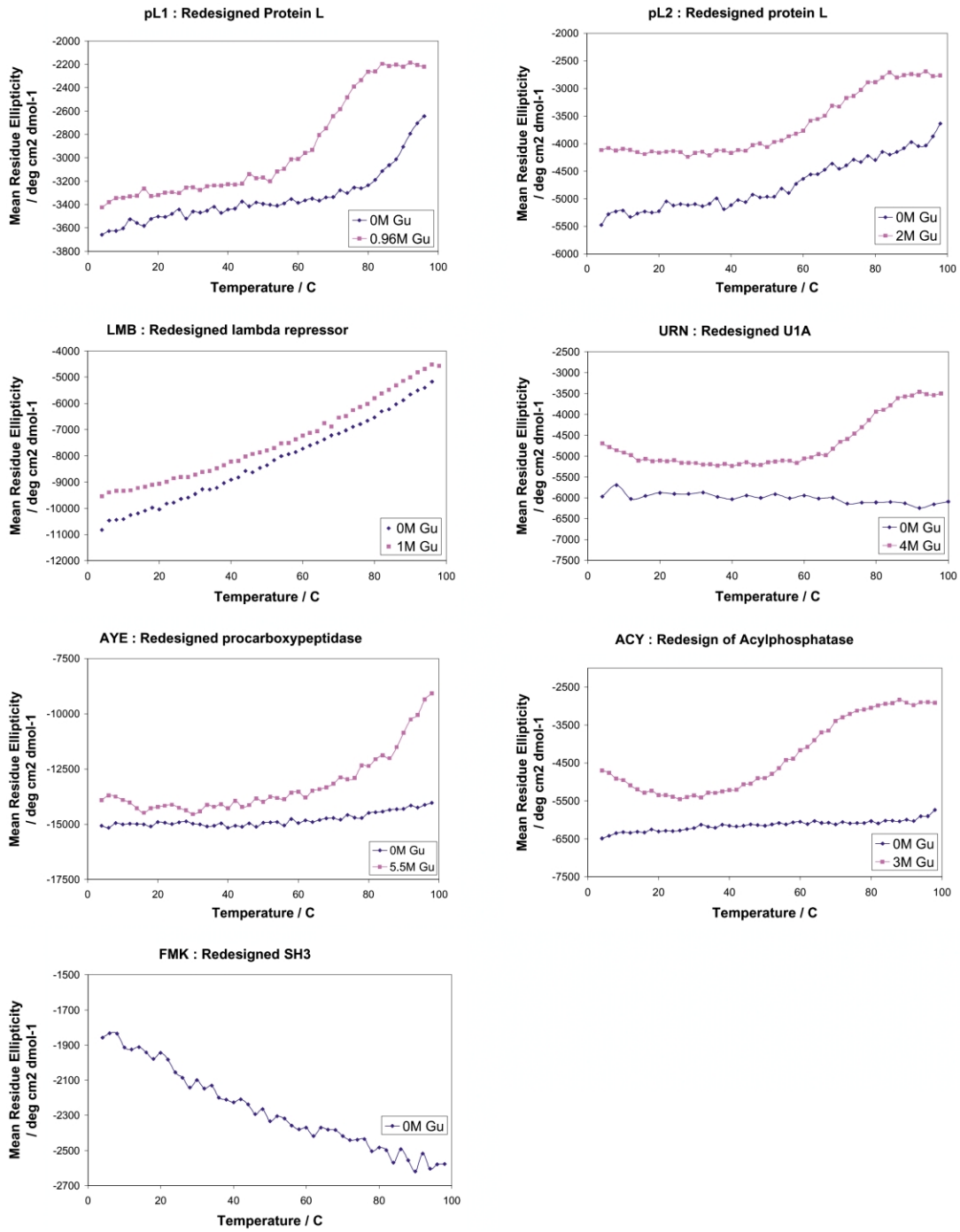

Figure 4. Thermal denaturation of the redesigned proteins. The temperature-induced denaturation profiles of five of the redesigned proteins (pL1, pL2, URN, AYE, ACY) are two-state and co-operative. Redesigned lambda repressor exhibits a non-cooperative temperature melt, and redesigned SH3 is unfolded at all temperatures. Ellipticity at 220 nm was monitored as a function of temperature for ~10 μM protein in 50 mM sodium phosphate, pH 7.0 in a 2 mm cuvette (blue curves). Pink curves are temperature melts performed for each protein at a Gu-HCl concentration where each protein was still folded at 25 °C (as ascertained from Figure 3).

Only one of the proteins, the SH3 redesign, is clearly unfolded. The CD spectra of redesigned SH3 (Figure 2) is typical of a random coil and the 1D ^{1}H NMR spectrum (Figure 5) shows sharp lines and very little dispersion, strongly indicative of an unfolded protein.

Three of the proteins were multimeric even at low concentration. The tenascin redesign visibly aggregates at low concentrations and could not be further characterized. Size-exclusion chromatography of the FKBP12 and S6 redesigns suggest they form oligomers, but the two proteins do not form extensive aggregates and their CD spectra are similar to their naturally occurring

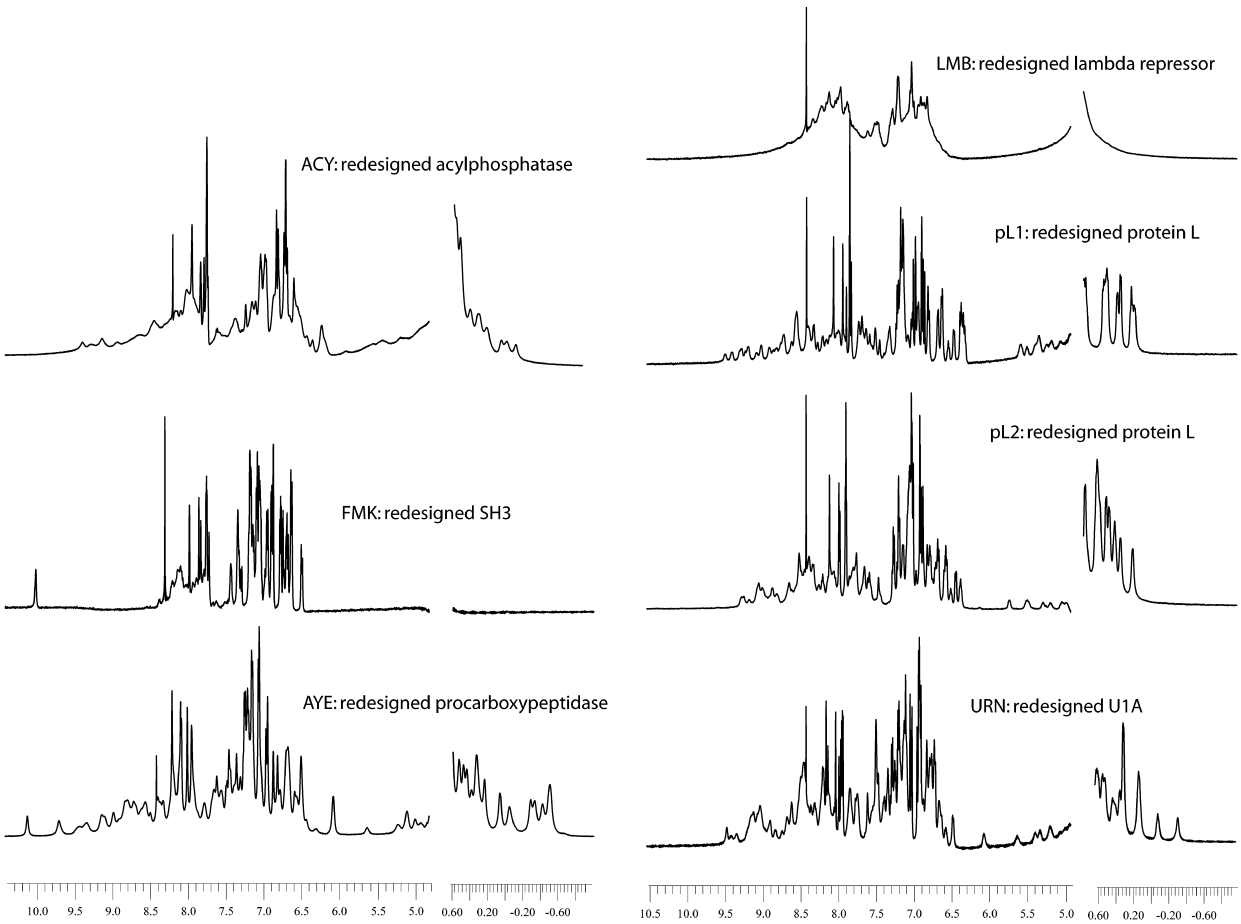

Figure 5. One-dimensional ^{1}H NMR spectra of the redesigned proteins. The sharp lines and strong dispersion in the spectra of pL1, pL2, URN and AYE suggest that these proteins are well folded in a unique conformation. Additionally, the peaks between 5ppm and 5.5ppm suggest that these proteins have residues in a β-sheet, which is consistent with the target structure for these designs. Spectra were obtained at 27 °C in 50 mM sodium phosphate (pH 7). Protein concentrations were between 600 μM and 1.2 mM.

Table 2. Summary of experimental results

Redesigned proteins	CD spectra	Gu-HCl melt	Temperature melt	1D ^{1}H NMR	Verdict
src SH3	Random-coil	Non-cooperative	Non-cooperative	Sharp-lines; no dispersion	Unfolded
Tenascin	β-Sheet like WT	Aggregated unable to determine (UTD)			Aggregated
λ-Repressor	α-Helical like WT	Cooperative stability < WT	Non-cooperative	Broad-lines; weak dispersion	Destabilised less rigid
Acylphosphatase	α/β-Like WT	Cooperative stability = WT	Cooperative T_m > WT	Broad-lines; strong dispersion	Stable less rigid
Immunophillin FKBP12	α/β-Like WT	Cooperative stability = WT	UTD	UTD	Stable multimeric
Ribosomal S6	α/β-Like WT	Does not denature	UTD	UTD	Stabilised multimeric
Protein L 1	α/β-Like WT	Cooperative stability < WT	Cooperative T_m > WT	Sharp lines; strong dispersion	Destabilised well-folded
Protein L 2	α/β-Like WT	Cooperative stability = WT	Cooperative T_m > WT	Sharp lines; strong dispersion	Stable well-folded
RNA-binding U1A	α/β-Like WT	Cooperative stability > WT	Cooperative T_m > WT	Sharp lines; strong dispersion	Stabilised well-folded
Procarboxypeptidase	α/β-Like WT	Cooperative stability > WT	Cooperative T_m > WT	Sharp lines; strong dispersion	Stabilised well-folded

Table 3. Thermodynamic stability of the designed and wild-type proteins

Protein	ΔG_U^{H2O} (WT, (kcal mol^{-1})	ΔG_U^{H2O} (design, kcal mol^{-1})	*m*-GuHCl (WT, kcal mol^{-1} M^{-1})	*m*-GuHCl (design, kcal mol^{-1} M^{-1})	T_m (WT, °C)	T_m (design, °C)
Lambda repressor[27]	4.8	2.8	2.4	1.1	56	–
U1A[28]	8.1	9.9	1.8	2.0	[ ]	>100
Src SH3[29]	3.8	–	1.6	–	[ ]	–
S6[30]	11.6	–	[ ]	–	99	–
Acylphosphatase[31]	4.8	5.3	[ ]	1.7	54	>100
Procarboxypeptidase[32]	4.1	11.9	[ ]	2.0	70–77	>100
FKBP12[33]	4.6	4.8–7.1[a]	5.4	–	[ ]	–
Protein L (1)[34,35]	4.6	3.7	1.9	1.4	70	~100
Protein L (2)[34,35]	4.6	4.4	1.9	1.8	70	>100

–, Unable to determine. [ ] not found in the literature.
[a] Due to a strongly sloping "folded" baseline, slightly different baseline estimates yield significantly different ΔG estimates for redesigned FKBP12, with very similar fitting errors. This high variability may be due to the guanidine-induced solubilization of aggregates of this protein at low guanidine concentrations.

counterparts, suggesting that they may adopt the target structures. While the FKBP12 redesign denatured at high guanidine concentration, as evidenced by the change in the CD spectrum (Figure 2), the CD spectrum of redesigned S6 was remarkably resistant to both temperature and chemical denaturant (Figure 2); hence redesigned S6 may be stabilized by intermolecular as well as intramolecular interactions. Clearly the computational design method, in addition to optimizing the stability of a given structure, needs to take into account solubility issues as well. This may perhaps be achieved by negative design against possible intermolecular interactions, for example by placing inwardly pointing charged amino acids in edge beta strands as suggested by Richardson & Richardson.[20]

The six remaining redesigned proteins appear monomeric and folded as evidenced by size-exclusion chromatography, CD spectra, and chemical denaturation experiments. The proteins chromatographed as monomers by gel-filtration chromatography, and comparison of their CD spectra (Figure 2) to previously published CD spectra of their naturally occurring counterparts suggested a very similar distribution of secondary structures. Chemical denaturation data fit well to a simple two-state folding model (Figure 3), and for the designed proteins with buried tryptophan residues, unfolding transitions monitored by CD and by intrinsic fluorescence (data not shown) were coincident, further supporting the two-state model typical for small naturally occurring proteins. The free energies of unfolding and their denaturant dependencies (*m* values) for the redesigned proteins were estimated from the fits of the chemical denaturation data (Figure 3) to the two-state model. The *m* values of the designed proteins are in the range of those of naturally occurring small proteins; of the four cases where direct comparisons are possible, two of the designed proteins have smaller *m* values than their wild-type counterparts, one has a larger value, and one a very similar value (Table 3). Of the six proteins, two, the redesigned lambda repressor and one of the two protein L redesigns (pL1), are clearly less stable than their naturally occurring counterparts (Figure 3 and Table 3). The redesigned acylphosphatase and the second protein L redesign (pL2) have roughly the same stability as their naturally occurring counterparts (Table 3). In contrast, the U1A and procarboxypeptidase redesigns were significantly more stable than the naturally occurring proteins, redesigned U1A by ~2 kcal/mol and redesigned procarboxypeptidase by a striking ~7 kcal/mol.

Two common features of many naturally occurring proteins are cooperative thermal denaturation transitions and NMR spectra with strong dispersion and sharp lines. Both of these features appear to be linked to the rigidity of the protein structure. In a rigid protein each atom is located in a well-defined environment and therefore only one sharp NMR peak is observed for each resonance. In contrast, if the structure is more molten then the atom may be in multiple different environments on a time-scale relevant to the NMR measurement and therefore broad NMR lines are observed. A highly cooperative thermal transition indicates a large change in enthalpy upon unfolding and is consistent with a change from a rigid folded protein to a dynamic unfolded protein.

To assess the rigidity of the redesigned proteins, 1D NMR spectra and temperature melts were obtained. Redesigned lambda repressor does not have a cooperative thermal melt or an NMR spectrum with sharp lines, suggesting that it may be more flexible than the other redesigns. Redesigned acylphosphatase has a cooperative thermal melt, but the NMR spectrum has broad lines, which may in this case reflect some intermolecular association at high concentration (and hence slower tumbling times and broader NMR lines).

Remarkably, four of the beta sheet-containing protein redesigns appear to be as rigid as most naturally occurring proteins. The two protein L redesigns and the redesigns of U1A and procarboxypeptidase have cooperative thermal melts

(Figure 4) and NMR spectra with relatively sharp lines and good dispersion (Figure 5). In addition, the NMR spectra for these proteins have small peaks just downfield of the water (5–5.5ppm) that are probably from C^{α} protons on the backbone and are strongly indicative of a β-sheet.[21]

Another hallmark of naturally occurring proteins is a large change in heat capacity upon folding. Two of the proteins, redesigned U1A and acylphosphatase, cold-denature at intermediate concentrations of guanidine and estimates of ΔC°_{p} could be obtained from fits of temperature denaturation experiments to the Gibbs–Hemholtz equation. For both proteins the ΔC°_{p} per residue is approximately $10\,\mathrm{cal\,deg^{-1}\,mol^{-1}}$, which falls within the range of ΔC°_{p} per residue values reported for natural proteins of this size.[22]

Discussion

Here we have shown that RosettaDesign can reliably predict sequences that fold to stable structures, and that the designed proteins often have features typical of naturally occurring proteins. Half of the folded designs have NMR spectra and temperature melts typical of tightly packed proteins. These findings significantly extend the pioneering successful complete redesign of the 25 residue zinc finger Zif268[1] to a broad range of considerably larger proteins.

Since so many mutations were made to each protein it is difficult to determine why some designs were more successful than others, but there are some trends. Three redesigns were significantly more stable than their wild-type versions (the redesigns of S6, U1A, and procarboxypeptidase) and two redesigns were less stable (one of the protein L redesigns and the redesign of lambda repressor). In each of the cases where the designs were more stable, the design sequence had a greater fraction of hydrophobic amino acids than the wild-type protein. In the two cases where the design was less stable, the redesigned sequences were less hydrophobic than those of the wild-type protein. These results are consistent with the notion that the burial of hydrophobic groups is one of the driving forces of protein folding.[23] More detailed comparison of the wild-type and redesigned proteins must await high-resolution determination of the structures of the redesigned proteins.

Six of the ten design sequences were soluble and monomeric at NMR concentrations (1mM) as judged by gel-filtration, while one was unfolded. Why do the remaining three proteins self associate? Redesigned tenascin was visibly aggregated even at low concentrations, and therefore it was not possible to determine if it was folded. One possibility is that it aggregates to such a high degree because it is unfolded and therefore its hydrophobic core residues are exposed to interactions with other molecules. Redesigned S6 and FKBP12 do not visibly aggregate at low (CD) or high (NMR) concentrations, but are multimeric. This lower degree of association, when compared to redesigned tenascin, is probably due to association of folded monomers. Indeed, the redesigns of S6 and FKBP12 have an increase in the fraction of non-polar accessible (i.e. "sticky") surface area compared to the wild-type counterparts; this criterion could be used as a filter to ensure that future designs are soluble. To test this we are currently constructing variants of the S6 and FKBP12 redesigns that have fewer hydrophobic residues on their surface. Additionally, redesigned tenascin and FKBP12 seem to lack any of the "aggregation preventors" that native proteins employ with their edge beta strands, namely strand kinks or inward pointing charged residues.[20] Since designing a kink in any strand would change the redesign backbone from its native target, we are testing the feasibility of the second anti-aggregation strategy by constructing variants of redesigned tenascin and FKBP12 that replace edge-strand partially surface-exposed hydrophobic residues with charged residues.

In only one case, the redesign of the src SH3 domain, was the designed protein clearly unfolded. To examine why this designed sequence was so unstable, we used the program Probe[24] to look for clashes in our model of the designed protein. Probe identified a large clash between Ile26 and Ala39. An examination of the multiple sequence alignment for SH3 domains showed that these amino acid residues are often seen at these positions, but typically not together.[25] There is a strong preference for Ile26 to be paired with Gly39, and Leu26 to be paired with Ala39. The atomic radii used in our simulations are scaled by 0.95 relative to CHARMM 19 radii in order to compensate for the use of fixed rotamers. If the radii are increased to their full size, then RosettaDesign shows a strong preference for a Leu-Ala pair. Currently we are testing these findings by mutating Ile26 to Leu or Ala39 to Gly. This is a case where using reduced radii can be costly, and suggests the need for more realistic radii coupled with a better sampling of side-chain conformational space.

The large-scale test described here establishes that RosettaDesign can redesign naturally occurring proteins with a reasonable chance of success. These encouraging results suggest that the program is ready to attack the next big challenge in the field of protein design, the creation of proteins with novel structures.

Materials and Methods

Computational procedure

Our computational model for protein design, RosettaDesign, is largely unchanged from that described.[13] RosettaDesign contains two main components: an energy function that ranks the relative fitness of amino

sequences for a given protein structure and a Monte Carlo optimization procedure for rapidly searching sequence space. The energy function is a linear combination of a 12–6 Lennard–Jones potential, the Lararidis–Karplus implicit solvation model,[18] an empirical hydrogen bonding potential,[17] backbone dependent rotamer probabilities,[19] amino acid probabilities for particular regions of phi,psi space, and a simple electrostatics term derived from the probability that two types of polar amino acid residues are found near each other in the PDB.[26] In addition, each amino acid has a unique reference energy that controls the frequency that it is placed during design. Except for the hydrogen bonding term, all of these energies were computed as described previously (see the Supplementary Material for a detailed description). The new hydrogen bonding potential was derived from hydrogen bonding geometries in high-resolution protein structures,[17] and is consistent with quantum mechanics calculations on formamide and acetamide dimers (A. Morozov, & D.B., unpublished results). Environment dependent hydrogen bond weights were used to roughly account for the reduced dielectric in the protein interior and the loss of side-chain entropy upon formation of side-chain–side-chain hydrogen bonds on the surface.[17]

The Monte Carlo optimization procedure used to scan sequence space started with a random sequence. The side-chain conformations of each amino acid were modeled using Dunbrack's backbone dependent rotamer library.[19] Only rotamers observed more than 3% of the time were considered. Each round of Monte Carlo consisted of replacing one rotamer, evaluating the energy change, and accepting the change if it passed the Metropolis criterion. A rotamer replacement may or may not involve changing amino acid identity. A typical run consisted of a few hundred thousand rotamer replacements, at which point the energy had typically plateaued.

Two rounds of optimization were used for each protein that was redesigned. The first round consisted of 100 independent runs in which all 20 amino acid residues were allowed at each position. Dunbrack's standard rotamer library was used for this round. During the second round the amino acid residues considered at each sequence position were restricted to those observed at that position in the results from the first round. Typically between one and five amino acid residues were considered at each position in the second round. Because there were fewer amino acid residues being considered in the second round it was possible to use an expanded rotamer library. In addition to the standard Dunbrack rotamers, new rotamers were constructed with chi angles plus one and minus one standard deviation away from the most commonly observed chi angles. These new rotamers were given a small energy penalty to account for the fact that they are sub-optimal. As in the first round, 100 independent runs were performed for each protein in the second round. From these runs, the lowest energy sequence was chosen for experimental study. In general, it is not clear if using a second round of design with more rotamers was helpful. The average identity between the design sequences and the native sequence did not increase from round one to round two.

Protein expression and purification

Genes corresponding to the computationally selected protein sequences were purchased from BlueHeron Biotechnologies. The gene constructs were cloned in plasmid pet29b(+) (Novagen) and expressed in the BL21(DE3)pLysS strain of *E. coli*. A 6× histidine tag at the C terminus of each construct allowed for the single-step purification of the expressed proteins on a Ni^{+} affinity column (Pharmacia Biotech). Column-purified protein was dialysed 10^4-fold against 50 mM sodium phosphate (pH 7.0), which is the buffer used in all subsequent experiments. Protein identity and purity was determined by SDS-PAGE and ESI-MALDI mass spectroscopy. Protein concentrations were determined by UV absorbance at 280 nm with extinction coefficients calculated using the ExPASy Protparam tool†.

Circular dichroism (CD)

CD data were collected on an Aviv 62A DS spectrometer. Far-UV CD wavelength scans (260–200 nm) at varying protein concentrations (15–25 μM), guanidinium hydrochloride (Gu-HCl) concentrations (0–8.3 M), and temperatures (0–98 °C) were collected in a 1 mm pathlength cuvette. Gu-HCl-induced protein denaturation was followed by change in ellipticity at 220 nm in a 1 cm pathlength cuvette, using a Microlab titrator (Hamilton) for denaturant mixing. Temperature was maintained at 25 °C with a Peltier device. All CD data were converted to mean residue ellipticity. Temperature-induced protein denaturation was followed by the change in ellipticity at 220 nm in a 2 mm pathlength cuvette. To obtain a value for ΔG_U^{H2O}, chemical denaturation curves were fit by non-linear least-squares analysis using the linear extrapolation model as applied by Santoro & Bolen. To obtain a value for ΔC°_{p}, thermal denaturation curves were fit using the Gibbs–Hemholtz equation in the form:

$$\phi = \phi_f + \frac{(\phi_u - \phi_f)}{1 + e^{\frac{-\Delta G^{\circ}}{RT}}}$$

$$-\Delta G^{\circ} = \Delta H^{\circ}\left(1 - \frac{T}{T_m}\right) + \Delta C^{\circ}_{p}\left\{T - T_m - T\ln\left(\frac{T}{T_m}\right)\right\}$$

where ϕ is CD signal, ϕ_f and ϕ_u are the estimated CD signal for the folded and unfolded states, respectively, R is the gas constant, T is temperature, T_m is the temperature where 50% of the protein is folded, ΔG° is the change in the Gibbs free energy for the unfolding reaction, ΔH° is the change in enthalpy, and ΔC°_{p} is the change in heat capacity.

Size-exclusion (gel-filtration) chromatography

Size-exclusion chromatography was carried out using an analytical Superdex-75 column (Amersham Pharmacia) with the Pharmacia FPLC system (GP-250 gradient programmer, P-500 Pump). Protein samples at NMR concentrations (600 μM–1.2mM) and CD concentrations (10–40 μM) were equilibrated in 20 mM EDTA, 50 mM sodium phosphate (pH 7.0 at 25 °C) and run on the Superdex-750 column at 1 ml/minute.

Nuclear magnetic resonance

One-dimensional spectra were obtained on a Bruker AMX500 using water presaturation. Spectra were obtained at 27 °C in 50mM sodium phosphate (pH 7).

† http://us.expasy.org/tools/protparam.html

Protein concentrations were between 600 μM and 1.2mM.

Solvent-accessible surface area

Solvent-accessible surface area of non-polar atoms was calculated using the program Whatif†.

Acknowledgements

We thank Lynne R. Spencer, Peter Brzovic, Ponni Rajagopol, Jennifer Keefe and Rachel Klevitt for aid in obtaining the NMR spectra. B.K. was partially supported by a Damon Runyon–Walter Winchell Foundation fellowship. This work was supported by a grant from the NIH.

References

1. Dahiyat, B. I. & Mayo, S. L. (1997). *De novo* protein design: fully automated sequence selection. *Science*, **278**, 82–87.
2. Desjarlais, J. R. & Handel, T. M. (1995). *De novo* design of the hydrophobic cores of proteins. *Protein Sci.* **4**, 2006–2018.
3. Havranek, J. J. & Harbury, P. B. (2003). Automated design of specificity in molecular recognition. *Nature Struct. Biol.* **10**, 45–52.
4. Looger, L. L. & Hellinga, H. W. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J. Mol. Biol.* **307**, 429–445.
5. Koehl, P. & Levitt, M. (1999). *De novo* protein design. I. In search of stability and specificity. *J. Mol. Biol.* **293**, 1161–1181.
6. Summa, C. M., Rosenblatt, M. M., Hong, J. K., Lear, J. D. & DeGrado, W. F. (2002). Computational *de novo* design, and characterization of an A(2)B(2) diiron protein. *J. Mol. Biol.* **321**, 923–938.
7. Reina, J., Lacroix, E., Hobson, S. D., Fernandez-Ballester, G., Rybin, V., Schwab, M. S. *et al.* (2002). Computer-aided design of a PDZ domain to recognize new target sequences. *Nature Struct. Biol.* **9**, 621–627.
8. Wernisch, L., Hery, S. & Wodak, S. J. (2000). Automatic protein design with all atom force-fields by exact and heuristic optimization. *J. Mol. Biol.* **301**, 713–736.
9. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science*, **282**, 1462–1467.
10. Benson, D. E., Conrad, D. W., de Lorimier, R. M., Trammell, S. A. & Hellinga, H. W. (2001). Design of bioelectronic interfaces by exploiting hinge-bending motions in proteins. *Science*, **293**, 1641–1644.
11. Shimaoka, M., Shifman, J. M., Jing, H., Takagi, J., Mayo, S. L. & Springer, T. A. (2000). Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nature Struct. Biol.* **7**, 674–678.
12. Offredi, F., Dubail, F., Kischel, P., Sarinski, K., Stern, A. S., Van de Weerdt, C. *et al.* (2003). *De novo* backbone and sequence design of an idealized alpha/beta-barrel protein: evidence of stable tertiary structure. *J. Mol. Biol.* **325**, 163–174.
13. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
14. Nauli, S., Kuhlman, B. & Baker, D. (2001). Computer-based redesign of a protein folding pathway. *Nature Struct. Biol.* **8**, 602–605.
15. Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y. & Baker, D. (2002). Accurate computer-based design of a new backbone conformation in the second turn of protein L. *J. Mol. Biol.* **315**, 471–477.
16. Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y. & Baker, D. (2001). Conversion of monomeric protein L to an obligate dimer by computational protein design. *Proc. Natl Acad. Sci. USA*, **98**, 10687–10691.
17. Kortemme, T., Morozov, A. V. & Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. *J. Mol. Biol.* **326**, 1239–1259.
18. Lazaridis, T. & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins: Struct. Funct. Genet.* **35**, 133–152.
19. Dunbrack, R. L., Jr & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661–1681.
20. Richardson, J. S. & Richardson, D. C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc. Natl Acad. Sci. USA*, **99**, 2754–2759.
21. Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*. Wiley, New York.
22. Myers, J. K., Pace, C. N. & Scholtz, J. M. (1995). Denaturant *m* values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* **4**, 2138–2148.
23. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**, 7133–7155.
24. Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K. *et al.* (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J. Mol. Biol.* **285**, 1711–1733.
25. Larson, S. M., Di Nardo, A. A. & Davidson, A. R. (2000). Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J. Mol. Biol.* **303**, 433–446.
26. Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C. & Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins: Struct. Funct. Genet.* **34**, 82–95.
27. Lim, W. A., Farruggio, D. C. & Sauer, R. T. (1992). Structural and energetic consequences of disruptive mutations in a protein core. *Biochemistry*, **31**, 4324–4333.
28. Kranz, J. K., Lu, J. & Hall, K. B. (1996). Contribution of the tyrosines to the structure and function of the human U1A N-terminal RNA binding domain. *Protein Sci.* **5**, 1567–1583.
29. Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for

† http://www.cmbi.kun.nl/gv/servers/WIWWWI/

folding of the src SH3 domain. *Nature Struct. Biol.* **5**, 714–720.
30. Uversky, V. N., Abdullaev, Z. Kh., Arseniev, A. S., Bocharov, E. V., Dolgikh, D. A., Latypov, R. F. *et al.* (1999). Structure and stability of recombinant protein depend on the extra N-terminal methionine residue: S6 permutein from direct and fusion expression systems. *Biochim. Biophys. Acta*, **1432**, 324–332.
31. Taddei, N., Chiti, F., Paoli, P., Fiaschi, T., Bucciantini, M., Stefani, M. *et al.* (1999). Thermodynamics and kinetics of folding of common-type acylphosphatase: comparison to the highly homologous muscle isoenzyme. *Biochemistry*, **38**, 2135–2142.
32. Villegas, V., Azuaga, A., Catasus, L., Reverter, D., Mateo, P. L., Aviles, F. X. & Serrano, L. (1995). Evidence for a two-state transition in the folding process of the activation domain of human procarboxypeptidase A2. *Biochemistry*, **34**, 15105–15110.
33. Veeraraghavan, S., Holzman, T. F. & Nall, B. T. (1996). Autocatalyzed protein folding. *Biochemistry*, **35**, 10601–10607.
34. Scalley, M. L., Yi, Q., Gu, H., McCormack, A., Yates, J. R., III & Baker, D. (1997). Kinetics of folding of the IgG binding domain of peptostreptococcal protein L. *Biochemistry*, **36**, 3373–3382.
35. Yi, Q., Scalley, M. L., Simons, K. T., Gladwin, S. T. & Baker, D. (1997). Characterization of the free energy spectrum of peptostreptococcal protein L. *Fold. Des.* **2**, 271–280.

Edited by B. Honig

(Received 29 January 2003; received in revised form 3 July 2003; accepted 8 July 2003)

Supplementary Material comprising text and one Table is available on Science Direct