doi:10.1016/j.jmb.2006.07.092 J. Mol. Biol. (2006) 362, 1004–1024

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Mis-translation of a Computationally Designed Protein Yields an Exceptionally Stable Homodimer: Implications for Protein Engineering and Evolution

Gautam Dantas[1], Alexander L. Watters[2], Bradley M. Lunde[3] Ziad M. Eletr[7], Nancy G. Isern[8], Toby Roseman[4], Jan Lipfert[9] Sebastian Doniach[9], Martin Tompa[4], Brian Kuhlman[7] Barry L. Stoddard[10], Gabriele Varani[1,5] and David Baker[1,6]*

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

[2]*Department of Molecular and Cellular Biology, University of Washington, Seattle 98195 USA*

[3]*Bio-Molecular Structure and Design Program, University of Washington, Seattle 98195 USA*

[4]*Department of Computer Science and Engineering University of Washington Seattle 98195, USA*

[5]*Department of Chemistry University of Washington Seattle 98195, USA*

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

[7]*Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill NC 27599, USA*

**Corresponding author*

We recently used computational protein design to create an extremely stable, globular protein, Top7, with a sequence and fold not observed previously in nature. Since Top7 was created in the absence of genetic selection, it provides a rare opportunity to investigate aspects of the cellular protein production and surveillance machinery that are subject to natural selection. Here we show that a portion of the Top7 protein corresponding to the final 49 C-terminal residues is efficiently mis-translated and accumulates at high levels in *Escherichia coli*. We used circular dichroism, size-exclusion chromatography, small-angle X-ray scattering, analytical ultra-centrifugation, and NMR spectroscopy to show that the resulting C-terminal fragment (CFr) protein adopts a compact, extremely stable, homo-dimeric structure. Based on the solution structure, we engineered an even more stable variant of CFr by disulfide-induced covalent circularisation that should be an excellent platform for design of novel functions. The accumulation of high levels of CFr exposes the high error rate of the protein translation machinery. The rarity of correspondingly stable fragments in natural proteins coupled with the observation that high quality ribosome binding sites are found to occur within *E. coli* protein-coding regions significantly less often than expected by random chance implies a stringent evolutionary pressure against protein sub-fragments that can independently fold into stable structures. The symmetric self-association between two identical mis-translated CFr sub-domains to generate an extremely stable structure parallels a mechanism for natural protein-fold evolution by modular recombination of protein sub-structures.

Keywords: mistranslation; protein-fold evolution; protein sub-fragments; NMR structure; protein engineering

Present address: G. Dantas, Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA.

Abbreviations used: AUC, analytical ultra-centrifugation; D_2O, deuterium oxide; ESI, electrospray-ionization; MS, mass spectroscopy; CFr, C-terminal fragment; GuHCl, guanidinium hydrochloride; HSQC, heteronuclear single-quantum coherence; NaPi, Sodium phosphate; NOE(SY), nuclear Overhauser effect (spectroscopy); MALDI-TOF, matrix-assisted laser desorption ionization - time of flight; *Rg*, radius of gyration; RMSD, root-mean-squared deviation; SASA, solvent accessible surface area; SAXS, small-Angle X-ray Scattering; SD, Shine–Dalgarno; TOCSY, total correlation spectroscopy.

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

[8]EMSL High Field Molecular Resonance Facility, PNNL Richland, WA 99352, USA

[9]Department of Physics Stanford University, Stanford CA 94305, USA

[10]Division of Basic Sciences Fred Hutchison Cancer Research Center, 1100 Fairview Ave N Seattle, WA 98109, USA

Introduction

The last decade has seen tremendous advances in the field of computational protein design. *In silico* protein sequence and structure optimisation algorithms have been successfully applied to completely redesign and thermodynamically stabilise naturally occurring protein structures,[1,2] to create novel[3] and thermodynamically stabilised enzymes,[4] to redesign protein–protein[5,6] and protein–ligand[7] interactions and to create extremely stable new protein structures.[8,9] Structural validation in many cases has confirmed the high-resolution accuracy of the design.[1,4–6,8–10] The accurate identification of extremely low energy regions of the protein sequence structure landscape is further validated by the finding that these designed proteins often achieve thermodynamic stabilities greater than those reported for any naturally occurring proteins.[2,9]

An obvious application of these exceptionally stable proteins is the generation of longer-lasting designer proteins and therapeutics.[11] However, while exceptional protein stability would have advantages in resistance to proteolysis and unfolding, there may also be biological costs once these proteins are expressed or delivered in the cell. It is therefore of considerable interest to investigate how computationally designed proteins are handled by the cellular protein production and surveillance machinery.

Translation processes often lead to faulty protein products, due to inappropriate translation initiation, ribosomal processivity errors, or missense errors where the mRNA transcript is erroneously decoded.[12–15] The overwhelming majority of these mis-translated proteins fail to assume native-like conformations, and are cleared from the cell by post-translational processes that involve a functional cooperation between molecular chaperones assisting in folding and the proteasome system.[15–17] Aberrant protein translation products that fold into stable substructures can evade cellular surveillance mechanisms and their subsequent accumulation can significantly damage or kill cells.[18–21] These phenomena are implicated in the pathology of a large number of diseases, including diabetes, cancer, and many neurodegenerative disorders.[22–24] Since exceptionally stable computationally designed proteins are created in the absence of specific evolu-

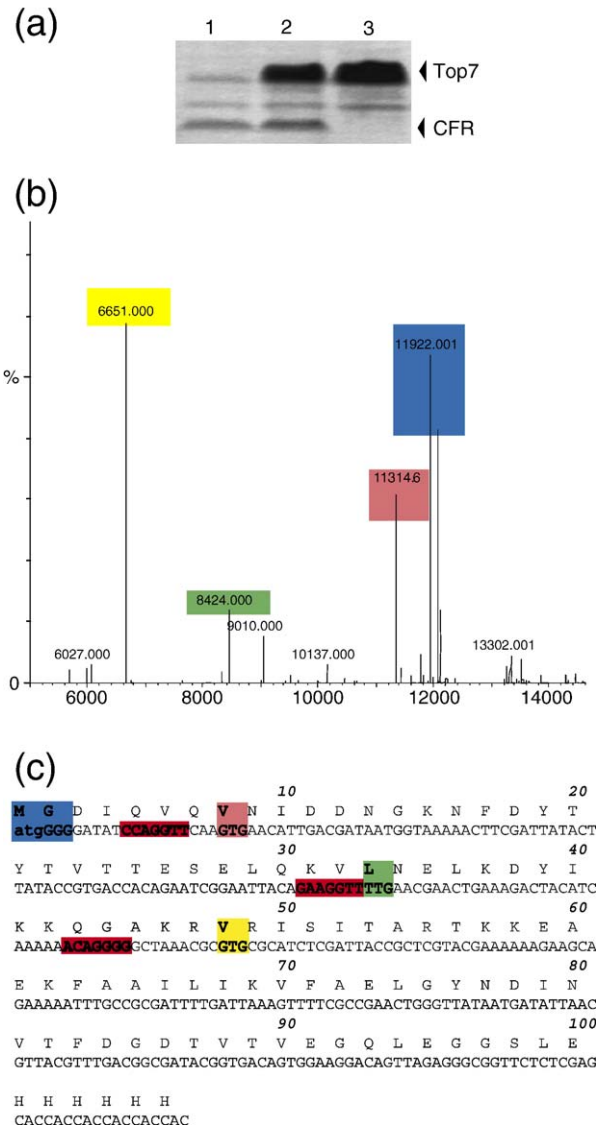

Figure 1. Mis-translation of Top7. (a) Coomassie-stained SDS-PAGE gel of Top7 protein variants ATG1ATT, wild-type and GTG48GTT (lanes 1-3, respectively). (b) ESI-MS spectrum of Top7_ATG1ATT. (c) Top7 protein (top lines) and DNA (bottom lines) sequence, with primary and alternate initiation codons highlighted (colours match the peaks from (b)). Degenerate Shine–Dalgarno sequences are highlighted in red.

tionary pressure, they provide a rare opportunity to reveal aspects of the cellular protein production and surveillance machinery that are subject to natural selection.

We recently generated an extremely stable, small, globular protein, called Top7, with a sequence and fold not observed previously in nature, using purely computational techniques.[9] Biophysical and structural analysis of Top7 demonstrated the high-resolution accuracy of our design. Here we show that a portion of the Top7 protein corresponding to the final 49 C-terminal residues is efficiently mis-translated in *Escherichia coli*. The solution structure of the resulting C-terminal fragment (CFr) protein reveals a compact, stable, homo-dimeric structure. Further stabilisation of CFr by disulfide-induced covalent circularisation yields a super-stable miniature protein that can serve as a robust scaffold for further protein engineering. The rarity of correspondingly stable fragments in natural proteins suggests evolution selects against protein fragments than can form stably folded structures.

Results

During the purification of the computationally designed Top7 protein, a strong band corresponding to a molecular mass of ~6.5 kDa was consistently observed on SDS-PAGE gels. This band was observed in addition to the Top7 band (~12.5 kDa) and remained even after Ni^+ affinity chromatography (Figure 1(a), lane 2). A subsequent anion-exchange purification step, however, was sufficient to isolate only the full-length Top7 as observed on SDS-PAGE and further confirmed by electrospray-ionization mass spectroscopy (ESI-MS), thereby allowing complete biophysical and structural characterisation of the pure Top7 protein.[9] In order to study the kinetic folding landscape of Top7, it nonetheless became clear that many mutant variants of the protein would need to be generated, and hence a practical interest arose in identifying and removing the lower molecular mass band. Since this smaller protein was retained in high yield following the Ni^+ affinity purification step, it was most likely a fragment of full-length Top7 that contained the C-terminal 6xHis tag and was either a product of proteolytic cleavage or of mis-translation.

Proteolysis or mis-translation?

To investigate the possibility that the Top7 sub-fragment was a proteolytic product, Top7 bacterial cell lysates were incubated at room temperature for up to three days in the presence and absence of protease inhibitors. Full-length Top7 was observed by SDS-PAGE in the supernatant fraction at relatively equal concentrations at all incubation times. Surprisingly, the ~6.5 kDa Top7 sub-fragment band was also observed in all supernatant fractions, also at relatively equal concentrations at all incubation times (data not shown). Since no appreciable degradation of Top7 was observed *in vitro* under conditions where many natural proteins show significant degradation,[25] and no enrichment of the sub-fragment was observed with increasing incubation time, it seemed unlikely that the sub-fragment was generated by Top7 proteolysis.

Matrix-assisted laser desorption ionization - time of flight (MALDI-TOF)-MS analysis of Ni^+-affinity purified Top7 confirmed that a species of ~6613 Da was present in addition to full-length protein (data not shown). The predicted molecular mass corresponded to a product ~30 Da larger than a polypeptide starting at Val48 and ~120 Da smaller than a polypeptide starting at Arg47. The sub-fragment was subsequently isolated from full-length Top7 by anion-exchange chromatography and analysed by N-terminal MS sequencing. The first six residues were found to be Met-Arg-Ile-Ser-Ile-Thr, corresponding to a Met followed by the sequence Arg49 to Thr53 of Top7. Methionine is ~30 Da larger than valine and hence a Top7 fragment starting with a Val48Met mutation matches the MALDI-TOF-MS predicted molecular mass. Since the plasmid coding for full-length Top7 did not contain this internal mutation, these results

Table 1. Statistics of ribosome binding sites in *E. coli* protein-coding regions

Threshold[a]	S[b]	C[c]	μ[d]	Σ[d]	z-score[e]
1000^{th}	3.18×10^{-4}	900	1238	36	−9.4
1500^{th}	1.13×10^{-4}	2507	3311	59	−13.6
2000^{th}	4.29×10^{-5}	5589	7144	88	−17.7
2500^{th}	2.02×10^{-5}	9830	11,944	112	−18.9
CCAGGTTcaa**GTG**	2.96×10^{-6}	26,790	29,931	189	−16.6
GAAGGTT**TTG**	2.99×10^{-7}	51,583	56,234	267	−17.4
ACAGGGGgctaaacgc**GTG**	2.68×10^{-5}	8069	9972	98	−19.3

[a] The first four rows correspond to the 1000th, 1500th, 2000th, and 2500th best-scoring *E. coli* upstream ribosome binding sites, respectively. The last three rows correspond to the alternative translation initiation sites within Top7 (Figure 1(c)), as they appear in 5′ to 3′ order. The Shine–Dalgarno sequence is shown in upper case and the start codon in bold.

[b] The score, S, is a product of the probability at each position in a putative ribosomal binding site, using the model described in Supplementary Data, Table S1.

[c] C is the number of sites within real protein-coding regions with scores at least S.

[d] The mean, μ, and standard deviation, σ, of the number of sites found in 300 random shufflings of the protein-coding regions, respectively, with scores at least S.

[e] The z-score is defined as the number of standard deviations separating C and μ.

suggested that the sub-fragment might be a product of mis-translation of the Top7 mRNA starting at amino acid position 48.

In prokaryotes, two key sequence features guide the ribosome to initiate translation from a specific location on mRNA–an AUG or AUG-cognate initiation codon, and the five to nine nucleotide ribosomal binding sequence (Shine–Dalgarno (SD) sequence) found three to 13 nucleotides upstream of the initiation codon.[13] The Val48 codon in the Top7 gene sequence is GTG. While >90% of *E. coli* translation is initiated at ATG, a small fraction of translation initiation occurs at GTG (8%), TTG (1%), and in one known case at ATT.[26] Could the Val(GTG)48 be the site (and cause) of mis-translation? To test this idea, we generated two single point mutants of the Top7 gene: a silent codon change from GTG to GTT at Val48 (GTG48GTT), and an N-terminal codon change from ATG to ATT to substitute the N-terminal Met with Ile (ATG1ATT). Since GTT has never been observed as a translation initiation codon, mis-translation from Val48 should be abrogated in this context, allowing translation of only the full-length product. The ATT variant at position one should disrupt translation of full-length Top7, but should not affect translation of the sub-fragment. Each of these variants were expressed, Ni^{+} affinity purified, and visualised with SDS-PAGE (Figure 1(a)). The GTG48GTT variant shows no observable expression of the ~6.5 kDa sub-fragment band (lane 3). The ATG1ATT variant shows significant reduction of the full-length Top7, but expression of the sub-fragment was essentially unaffected (lane 1). These variants were further analysed by ESI-MS, which confirmed the SDS-PAGE results (Figure 1(b)). However, the MS results for ATG1ATT also suggested that at least two other minor species of intermediate molecular mass between full-length Top7 and the ~6.5 kDa sub-fragment were present in the preparation. The predicted molecular masses for these two species match well to Top7 fragments beginning at Val8 (GTG) and Leu33 (TTG), both of which are coded for by potential alternate initiation codons (Figure 1(c)). In fact, zooming in on the 6–15 kDa region in the SDS-PAGE gels after increased protein staining also showed the presence of faint bands between Top7 and the ~6.5 kDa fragment. Analysis of the Top7 gene sequence revealed that degenerate versions of the *E. coli* ribosomal binding site (Shine–Dalgarno, SD) sequence are also present just upstream of all three identified Top7 mis-translation sites, and might also contribute to mis-translation (Figure 1(c)). To test whether the SD sequence was critical for mis-translation of the sub-fragment starting at Val48, we generated another point mutant of the wild-type gene that changes codon 44 from GGG to TCT, which should disrupt the putative SD sequence (ACAG**GGG** to ACAG**TCT**) without changing the Val(GTG) 48 initiation codon. The GGG44TCT variant was identical to the GTG48GTT variant in the observed ablation of the sub-fragment and no observed effect on translation of the full-length Top7 (SDS-PAGE and ESI-MS, data not shown). This result indicates that both translation initiation features are critical for the efficient mis-translation of the ~6.5 kDa fragment of Top7.

If evolution has selected against corresponding mis-translations in natural genes, one would expect to observe a reduced frequency of translation

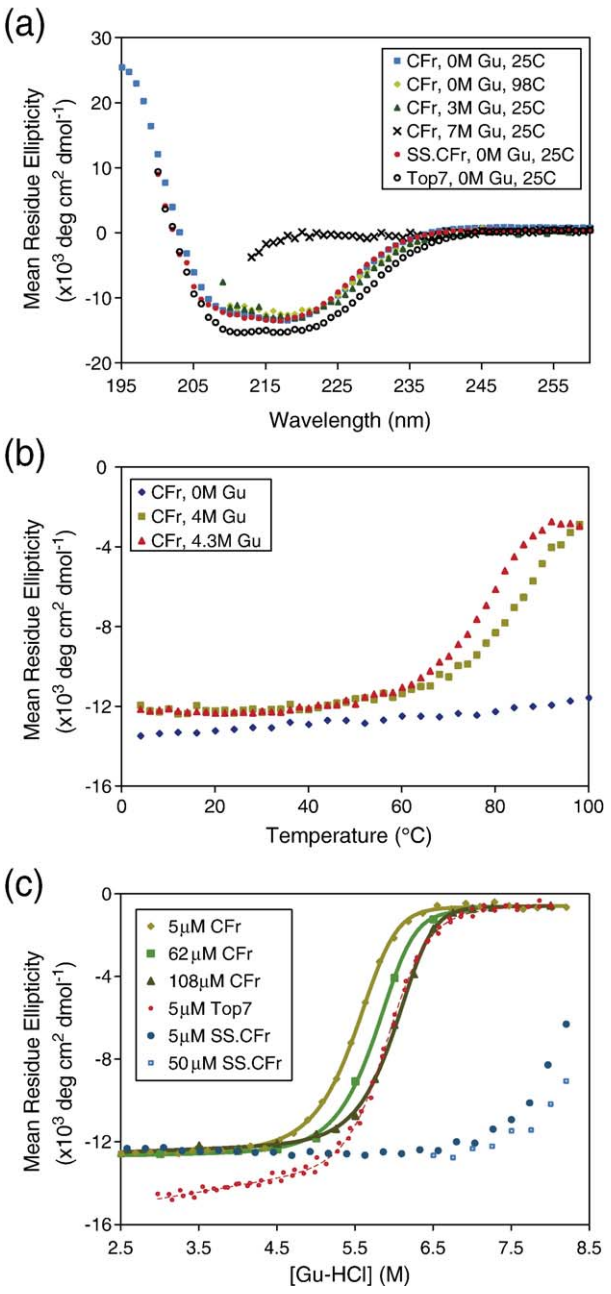

Figure 2. Biophysical characterisation of CFr and SS.CFr. (a) The far-ultraviolet (UV) CD spectrum of 25 μM CFr, 25 μM SS.CFr and 20 μM Top7 in 25 mM Tris–HCl (pH 8.0) at varying temperatures and concentrations of GuHCl. (b) CD signal at 220 nm as a function of temperature and GuHCl concentration for 12 μM CFr in 25 mM Tris–HCl (pH 8.0). (c) CD Signal at 220 nm as a function of GuHCl concentration for multiple concentrations of CFr, SS.CFr, and Top7 in 25 mM Tris–HCl (pH 8.0) at 25 °C.

initiation sequence features within the coding region of natural genes when compared to the frequency expected by random chance. Accordingly, we have computed the frequency of initiation codons in the context of an SD sequence within the 4237 annotated protein-coding regions of the *E. coli* genome, and compared them to the expected frequency if the codons were randomly permuted. Table 1 shows the results of this comparison for seven different thresholds that might reasonably be used to define what constitutes a "high-scoring" ribosomal binding site. The first four of these thresholds correspond to the 1000th, 1500th, 2000th, and 2500th best-scoring upstream ribosome binding sites, respectively, from the 2912 annotated *E. coli* genes which have at least 20 bp of non-coding DNA upstream of their start codons. The last three rows correspond to the scores of the alternative translation initiation sites within the Top7 gene (Figure 1(c)), as they appear in 5′ to 3′ order. For all seven thresholds *S* of Table 1, the number of observed instances within the real *E. coli* protein-coding regions with scores at least *S* (shown in column 3) is far below its expectation in randomly shuffled coding regions (shown in column 4). A standard measure of this difference is the *z*-score (shown in column 6), which is the number of standard deviations by which the observed and expected values differ. These results support the theory that evolution has selected against genetic features that would allow for mis-translation of protein sub-fragments.

Biophysical characterisation of CFr

The sequence of the ~6.5 kDa fragment of Top7 begins at a boundary between secondary structure elements in the Top7 structure and includes strands 3, 4 and 5, as well as helix 2 of Top7. This fragment is translated at high levels, is expressed in the soluble fraction, does not aggregate significantly, and is as resistant to cellular proteases as Top7. These results strongly suggest that this fragment has intrinsic stability and structure. For further analysis, a separate gene construct that codes for the ~6.5 kDa C-terminal fragment (CFr) of Top7 was made as described in Materials and Methods. Like Top7, the CFr protein can be obtained with high yield (25 mg/l) and purity (>99%) from the soluble fraction of the bacterial lysate. ESI-MS confirmed that a full-length protein of 7036 Da was isolated; this mass is within 0.1 Da of its theoretical molecular mass (Supplementary Data, Figure S1A).

Circular dichroism spectra strongly suggest that CFr is folded with α/β secondary structure, comparable in relative composition to Top7 (Figure 2(a)). CFr secondary structure appears unchanged at 98 °C or in 3 M guanidine-hydrochloride (GuHCl), but the CD spectrum of the protein is consistent with an unfolded polypeptide at 7 M GuHCl. In the presence of intermediate GuHCl concentrations (4.3 M), CFr unfolds cooperatively with temperature (Figure 2(b)), displaying remarkably high thermal stability, comparable to Top7. CFr also displays co-operative unfolding by GuHCl-induced chemical denaturation (Figure 2(c)). However, unlike Top7, CFr appears to be more stable with increasing protein concentration. These concentration dependent effects are generally indicative of the presence of quaternary structure during the unfolding transition. This was confirmed by gel filtration analysis of CFr at 25 μM and 1.2 mM; the protein resolves as a single peak with a molecular mass corresponding to a CFr dimer (data not

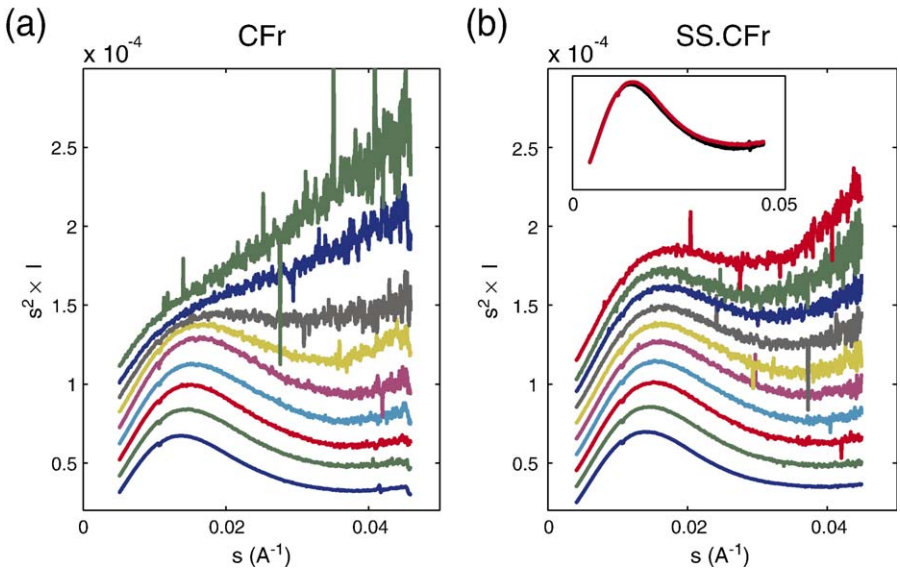

Figure 3. Small-angle X-ray scattering (SAXS) profiles of CFr and SS.CFr. Kratky plots ($s^2 \cdot I$ *versus* s) for (a) CFr and (b) SS.CFr as a function of GuHCl concentration are from bottom to top 1 M (blue), 2 M (green), 3 M (red), 4 M (light blue), 5 M (purple), 6 M (light brown), 6.5 M (black), and 7 M (blue) GuHCl for both CFr and SS.Cfr. The last profile for CFr is at 8 M (green) GuHCl and the last two profiles for SS.CFr are at 7.5 M (green) and 8 M (red) GuHCl. Profiles are vertically offset for clarity. ((b), inset) Superimposed profiles for CFr (black) and SS.CFr (red) at 1 M GuHCl.

shown). For a more robust characterisation of its unfolding behaviour and oligomeric state, CFr was analysed by small-angle X-ray scattering (SAXS) and analytical ultra-centrifugation (AUC). SAXS profiles of 2 mM CFr exhibit a single peak characteristic of a folded protein up to 5 M GuHCl, whereas the profiles at 7 M and 8 M GuHCl are indicative of a completely unfolded protein (Figure 3(a)). AUC scans of 35 μM −97 μM CFr show the protein to be dimeric at 0 M and 4 M GuHCl (where it appears folded by CD and SAXS), and monomeric at 7 M GuHCl (where it appears unfolded by CD and SAXS) (Figure 4(a) and (c)). These results suggest that CFr is an obligate dimer; the folded monomer is essentially never populated and the denaturation may be represented as an equilibrium transition between folded dimer and unfolded monomer. If this model is correct, the analysis of unfolding curves at different protein concentrations should result in similar values for $\Delta G°$ or K_d (see Materials and Methods for a description of this fitting procedure). Indeed, the $\Delta G°$ fit values are the same within experimental error for the different folding experiments: 26.4 kcal/mol (108 μM CFr), 25.5 kcal/mol (62 μM CFr), and 25.5 kcal/mol (5 μM CFr), confirming that CFr exists as an obligate dimer. A $\Delta G°$ value of 25.5 kcal/mol corresponds to a dissociation constant (K_d) of ~200 *zeptoM* (10^{-21} M).

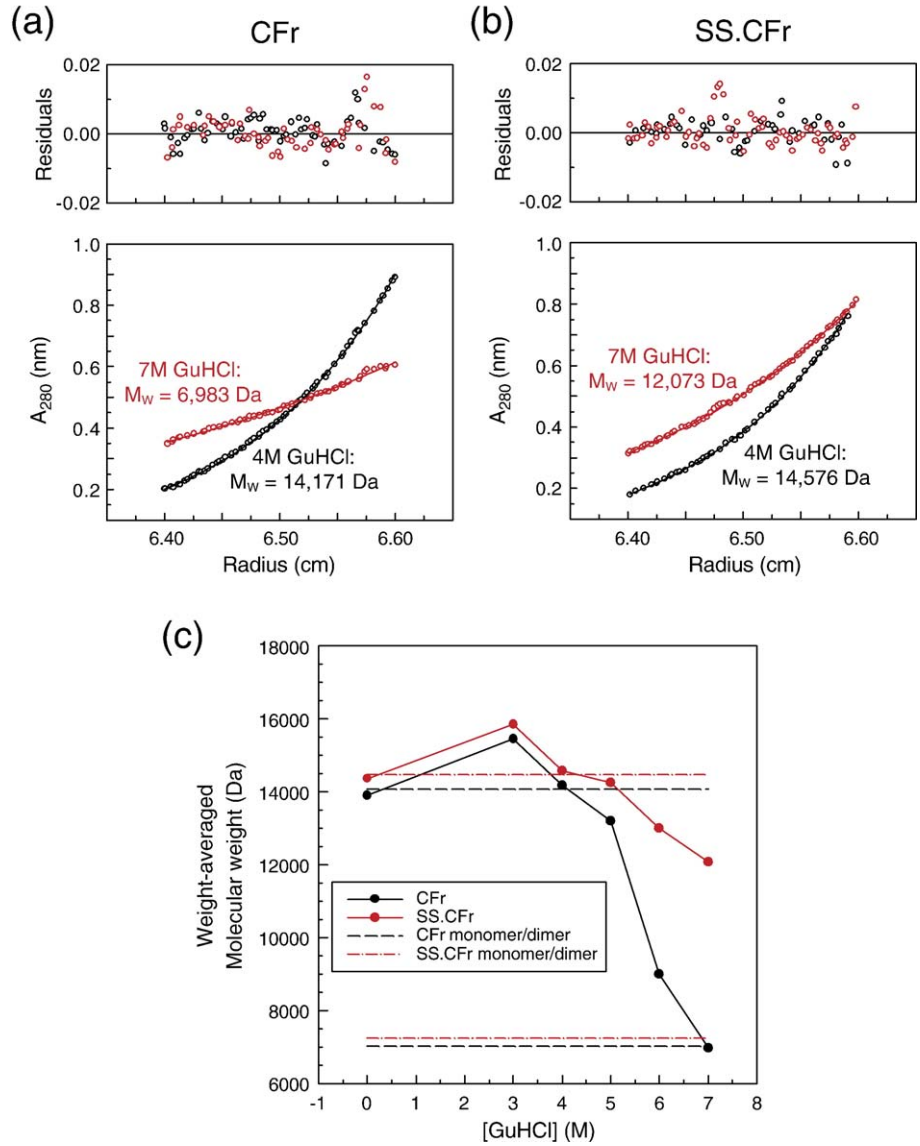

Figure 4. Analytical ultra-centrifugation (AUC) studies of CFr and SS.CFr. Selected equilibrium sedimentation profiles for (a) CFr and (b) SS.CFr collected at 30,000 rpm, 20 °C at protein concentrations of 59 μM −66 μM in solvent containing 4 M (black circles) or 7 M (red circles) GuHCl. The fitted weight-averaged molecular mass (M_r) was determined using a global fit to nine equilibrium scans collected at three protein concentrations and three speeds (see Materials and Methods). (c) Fitted M_r*versus* concentration of GuHCl plot. Fitted M_r values were determined as described above for CFr (black circles) and SS.CFr (red circles) at varying concentrations of denaturant. Horizontal lines represent predicted monomer/dimer molecular massesfor CFr, 7,037/14,074 (black, broken), and SS.CFr, 7,241/14,482 (red, dotted-dashed).

1D ^{1}H spectra and 2D ^{1}H-^{15}N heteronuclear single-quantum coherence (HSQC) spectra of CFr exhibit the features of a rigid well-folded protein (Figure 5), with well-dispersed and sharp peaks. Notably, the HSQC spectrum contains a single set of cross-peaks for each NH in the protein. Since CFr is a dimer, this result implies fully symmetric association. Solution structures of symmetric protein dimers are difficult to determine using conventional nuclear-Overhauser effect (NOE)-guided NMR techniques, because it is very difficult to distinguish between intra and inter-subunit NOEs. We employed asymmetric isotope labelling of the protein, in combination with isotope editing techniques[27,28] to resolve intra-subunit NOEs from inter-subunit NOEs in CFr and determine the symmetric homo-dimer solution structure.

Determination of the NMR structure of CFr

Protein backbone and side-chain assignments were obtained as described in the Materials and Methods. Structure determination was conducted in a two-step process, a fully automated iterative step dominated by NOE-derived distance constraints for generating models of a single subunit of CFr (CFrA), followed by a partly automated iterative step for building the symmetric homo-dimer model using manually assigned interfacial-NOE constraints. In the final calculation 100 structures were generated, of which the top 20 (Figure 6) had an average target function of 1.20(±0.11) Å^2 (Table 2) and an ensemble RMSD value of 0.33(±0.10) Å over backbone atoms and 0.75(±0.09) Å over heavy-atoms in residues 3 through 51 in both subunits (Table 3). There were no distance constraints violated by more than 0.1 Å and no angle constraints violated by more than 1°. When the ensemble was analysed with ProcheckNMR,[29] 99.2% of all dihedral angles were found in the allowed regions of the Ramachandran plot (Table 3). The small number of disallowed dihedral angles are all found for residues in the linker region (Glu2 and Gly52–His58).

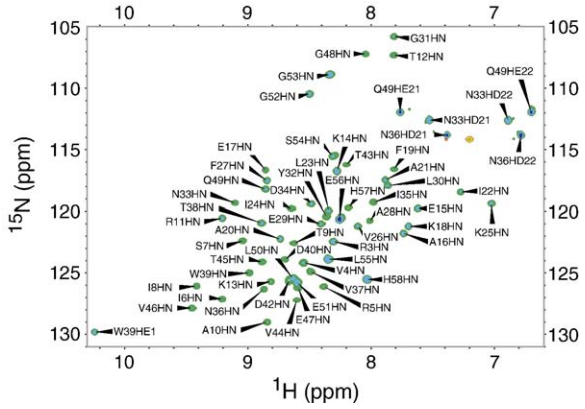

Figure 5. ^{1}H-^{15}N HSQC spectrum of CFr. The HSQC spectrum of ~1 mM ^{15}N-CFr in 50 mM phosphate buffer (pH 7.0), recorded at 298 K and 500 MHz.[81] Peaks are labelled with the one-letter amino acid code and sequence number.

CFr structure

Each of the two subunits of the CFr dimer adopts the same fold observed for the corresponding sequence in Top7, one helix packed on a three-stranded, antiparallel β-sheet (Figure 7; Top7 in purple, CFrA in green). The subunits form a symmetric antiparallel dimer, with all interfacial residues contributed by the first strand of the β-sheet and by the helix (Figure 8). The two subunits have virtually identical structures with an RMSD value of 0.41 Å over backbone atoms and 0.81 Å over all atoms (best NMR model, residues 3–51). Each subunit is also extremely similar to the corresponding portion of the Top7 crystal structure with an average backbone RMSD value of 1.12 Å (Figure 7). These deviations are as likely to reflect inaccuracies in the models as genuine structural differences. The largest deviation is in the hairpin between the second and third strand of the β-sheet (Asp40-Gly41-Asp42 in CFr); ignoring these residues improves the Top7 to CFr backbone RMSD value to 0.91 Å. The backbone NH of Gly41 is the only amide not observed in the HSQC spectrum, suggesting this loop is flexible in solution. Significantly, it is also not visible in the HSQC spectrum of the Top7 protein (data not shown).

The CFr dimer interface buries a total of 1457 Å^2 of solvent-accessible surface area (SASA), which accounts for about 19% of the surface of each subunit (Figure 8(a); interface carbon atoms in green and yellow). Ten residues on the β-sheet and ten on the helix (Figure 8(b); green or yellow cartoons and sticks) contribute to the CFr interface, and interestingly, these residues are buried to a very similar extent in the Top7 structure (data not shown). The CFr dimer interface is an extension of the individual CFr subunit hydrophobic cores; the strands of the two subunits form an extended six-stranded antiparallel β-sheet, stabilised by backbone hydrogen bonds across the interface between the first strands of both subunits. Of particular note is a pair of strong symmetric inter-subunit hydrogen bonds formed between the backbone NH of Ser7 on one subunit and backbone carbonyl of Ser7 on the other subunit: this NH remains very strongly protected after prolonged D_2O exchange. The tight packing observed between buried β-sheet residues interacting across the dimer interface (Val4, Ile6, Ile8, and Ala10 in both subunits) appears identical to the inter-strand side-chain packing observed within each subunit (this "continuous sheet core" is illustrated in Figure 8(c); *AB_00_SHEET*). Similar tight packing is also observed between helical side-chains interacting across the interface (Figure 8(c); *AB_00_HELIX*). Two symmetric aromatic clusters are formed between Phe19 on one subunit and Phe27 and Tyr32 on the other subunit, where the edge of the Phe19 aromatic ring stacks against the faces of the other two aromatics. Another strong interaction is a set of symmetric hydrogen bonds between the hydroxyl of Tyr32 on one subunit and the carboxyl moiety of Glu15 on the other subunit, which form an interfacial stitch at the helical caps.

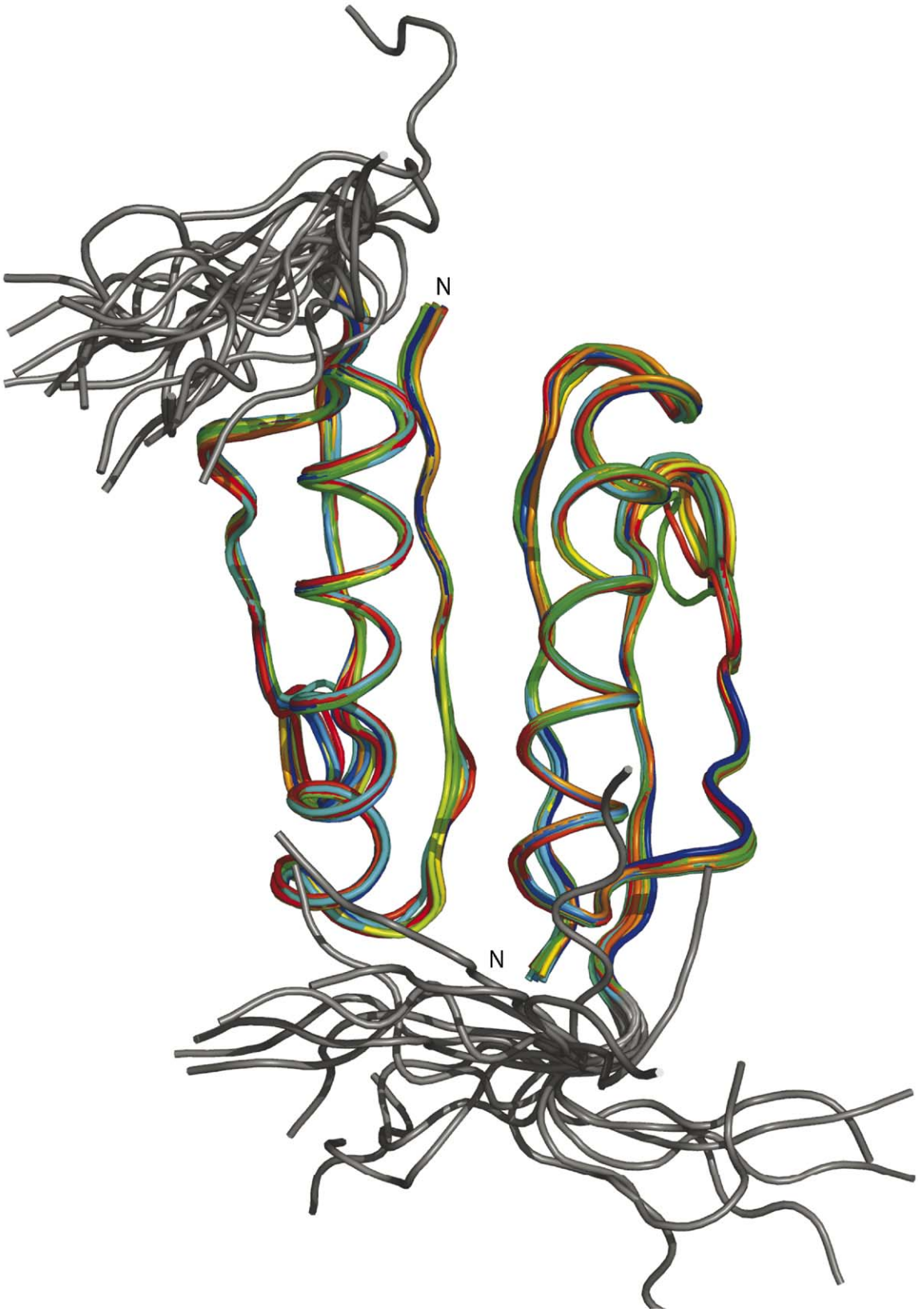

Figure 6. NMR-generated structures of CFr. The top 20 NMR models from the final CFr structure calculation are shown as ribbons. Each model is superimposed on the average backbone coordinates for residues 3–51 (structured region, separate colour for each model) in both chains from the entire ensemble. The structured regions have an ensemble RMSD value of 0.33 Å over the backbone atoms and 0.75 Å over all heavy atoms. Residues from the unstructured tails (52–58) are coloured in grey.

Table 2. NMR experimental constraints for CFr (residues 2–58)

A. *Monomer calculation*	
Unique NOE distance constraints (first round/final round)[a]	
Total	1915/1116
Intra-residue and Sequential ($[i–j] \leq 1$)	1312/645
Medium-range ($1 \leq [i–j] \leq 5$)	513/198
Long-range ($[i–j] \geq 5$)	90/273
Dihedral angle constraints[b]	76
Hydrogen bond constraints	16 (8 H-bonds)
Total number of constraints	1208
Number of constraints per residue	21.2
Long-range constraints per residue	4.8
B. *Dimer calculation*[c]	
Intra-subunit NOE constraints	2232
Inter-subunit NOE constraints	46
Dihedral angle constraints[b]	130
Hydrogen bond constraints	36 (18 H-bonds)
Total number of constraints	2444
Residual constraint violations[b]	
Distance violations	
(0.2-0.5) (Å)	0
(>0.5) (Å)	0
Van der Waals violations	
(0.2-0.5) (Å)	2
(>0.5) (Å)	0
Max. violation (Å)	0.33
Dihedral angle violations	
(1-10°)	0
(>10°)	0
CYANA target function (first round/final round)*	
NOEASSIGN monomer calculation	107.7 Å^2/5.2 Å^2
Final dimer calculation	————/1.2 Å^2

[a] First and final round refer to statistics from the NOEASSIGN macro in Cyana2.0.
[b] Dihedral angle constraints were generated from TALOS.[66]
[c] All dimer restraints and violations are twofold redundant due to the symmetric nature of the structure (see Materials and Methods for details).

Backbone dynamics

Further evidence for the structural stability of the CFr protein is provided by the measurements of ^{15}N T_1, T_2 and ^{1}H-^{15}N heteronuclear nuclear Overhauser enhancements (NOEs) that were measured by standard techniques described in the Materials and Methods. The results (Figure 9) show relatively uniform and featureless values with only relatively small variations across the sequence, with the exception of the unfolded C-terminal tail. The heteronuclear NOE is high, as expected for a domain rigid on the ns–ps time scale of motion, and the average value of T_2 and T_1 are consistent with a protein of about 10–15 kDa, the size of the CFr dimer. Even in loops, the values of T_2 are nearly constant, with the exception of residue Gly41 that appears to be exchange broadened.

Further stabilisation by disulfide circularisation of CFr

The high thermodynamic stability of the CFr structure makes it an ideal candidate as a scaffold for further design of novel or improved functions. Since functional design often involves making amino acid mutations that sacrifice thermodynamic stability, design on an extremely stable template should allow, at least in principle, for a larger number of "functionalising" mutations. We investigated the possibility of further stabilising CFr by the simple method of disulfide-induced protein circularisation. Since the NMR structure shows that the N and C termini of each subunit are next to each other, we chose positions at the end of both termini to add single cysteine residues such that their thiol groups could be within disulfide-forming distance. Formation of a disulfide bond between these two terminal cysteine residues should yield a covalently circularised form of each CFr subunit. The corresponding SS.CFr clone was generated and the protein purified as described in Materials and Methods.

ESI-MS showed that SS.CFr was isolated as a 7241 Da species, which corresponds to a single completely oxidised intra-molecular disulfide bond per subunit (within 0.1 Da of predicted M_r; Supplementary Data, Figure S1B). The CD wavelength scan of SS.CFr appears identical to CFr (Figure 2(a)), and the SAXS scattering profiles of the two proteins are indistinguishable at low denaturant concentrations (Figure 3(b), inset), suggesting the disulfide has not perturbed protein secondary or tertiary structure. The CD chemical denaturation profile of SS.CFr (Figure 2(c)) shows it to be dramatically stabilised over CFr, the protein begins to unfold only at 6.5 M GuHCl and appears to still be in the unfolding transition at 8.2 M GuHCl. The SAXS profile of 2 mM SS.CFr also indicates that the protein is still in the unfolding transition at 8 M GuHCl (Figure 3(b)). In comparison, both CFr and Top7 are almost completely unfolded by 6.5 M GuHCl (Figure 2(c)). Like CFr, SS.CFr also shows protein concentration dependence in its chemical denaturation, suggesting that it too exists as an obligate dimer. AUC scans confirm that SS.CFr (33 μM −105 μM) is predominantly dimeric from 0 M to 5 M GuHCl, but a small fraction of the monomeric form appears as the protein begins to unfold between 6 M and 7 M GuHCl (Figure 4(b) and (c)). These results indicate that SS.CFr is stabilised over CFr, and it is likely to be one of the most stable proteins reported regardless of class or size.

Table 3. Structural statistics for CFr dimer

A. *RMSD from averaged structure (Å)*[a]	
(Structured region, residues 3–51 in both chains)	
Backbone atoms	0.33
All heavy atoms	0.75
B. *PROCHECK-NMR analysis*[a]	
(All residues in both chains)	
Most favoured regions (%)	80.3
Additionally allowed (%)	17.3
Generously allowed (%)	1.6
Disallowed (%)	0.8

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

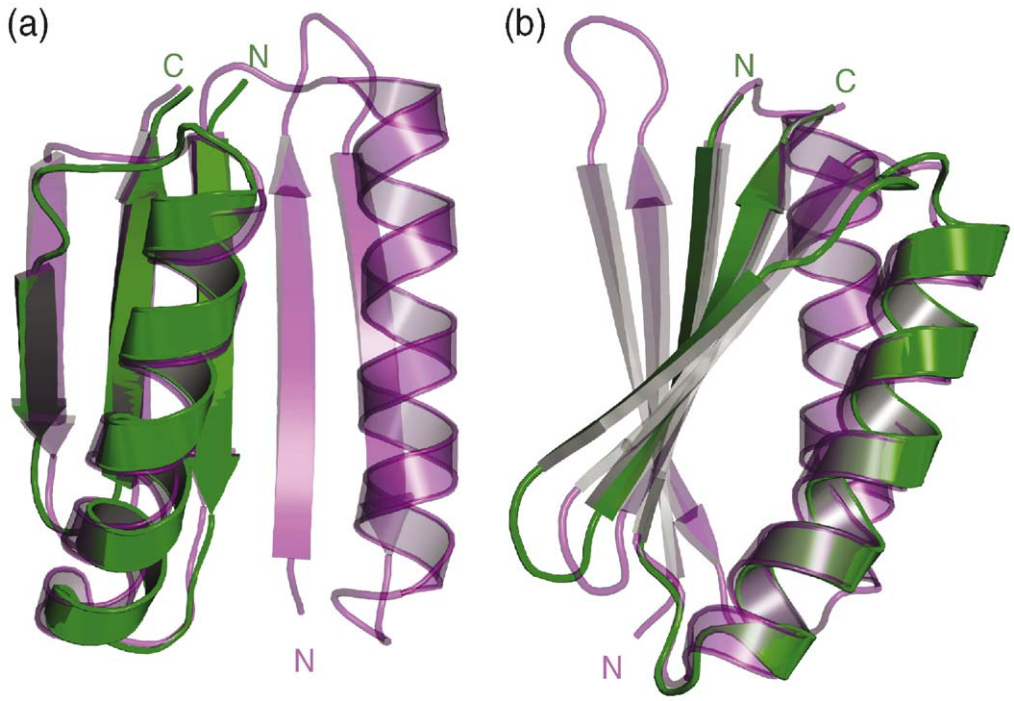

Figure 7. Comparison of the Top7 and CFr structures. (a) and (b) Ribbon diagrams of residues 3–51 from one subunit of the CFr NMR structure (green) superimposed on the corresponding region of the Top7 X-ray structure (purple). The backbone RMSD value over these residues is 1.12 Å. The two diagrams are related by a 90° rotation around the vertical axis in the plane of the page.

The SS.CFr construct was crystallized in an attempt to determine a higher resolution structure than that which was achieved for CFr by NMR spectroscopy. However, extensive crystallization trials and subsequent screening of specimens at the Advanced Light Source (ALS) yielded crystals that diffract to only 3.6 Å resolution, which provides no higher structural resolution than the NMR structure reported above. A strong molecular replacement solution to the phase problem was found, which generated models displaying relative subunit orientations and packing that agrees well with the NMR-derived structure. Additionally, difference maps calculated after molecular replacement demonstrate the presence of a disulfide bond bridging the N and C termini of the engineered construct, confirming this additional aspect of the design cycle.

Discussion

Initiation is usually the rate-limiting step of translation under normal conditions,[21,26] and ample evidence exists for regulation of protein synthesis at this step.[13,14] The significant bias in nucleotide frequencies observed in the translation initiation region of natural genes[30–32] suggests a stringent evolutionary selection for strong translation initiation signals at the sequence level. In an analysis of 30 complete prokaryotic genomes, a significant positive correlation was observed between the strength of the SD sequence and predicted expression level of a gene, such that highly expressed genes were much more likely to have a strong SD sequence than average genes.[33] Mutational analysis of translation initiation regions of a variety of genes have confirmed that disruption of the start codon or the SD sequence adversely affects translation efficiency and accuracy of initiation at the proper start codon.[34–36] Since appropriate initiation sequences are clearly important for efficient translation of normal genes, it should follow that similar sequences are avoided within the coding regions of genes to prevent mis-translation of sub-gene fragments. We have shown that the CFr fragment can be efficiently translated from within the Top7 gene due to the fortuitous presence of an initiation codon and a degenerate SD sequence at appropriate positions within the coding region of the Top7 gene, and that removal of either sequence feature is sufficient to completely abrogate CFr mis-translation, without affecting translation of the full-length Top7 protein. If evolution has selected against corresponding mis-translations in natural genes, one would expect to observe a reduced frequency of translation initiation sequence features within the coding region of natural genes when compared to the frequency expected by random chance. Saito & Tomita have shown conclusively that in both eukaryotes and prokaryotes, the frequencies of AUG triplets just upstream and downstream of the natural initiation codon are significantly lower than expected by random chance, which "is likely due to negative selectional pressure, since protein mis-translation is evolutionarily disadvantageous."[37] We extended this analysis to the complete coding regions of genes, and observed that high quality

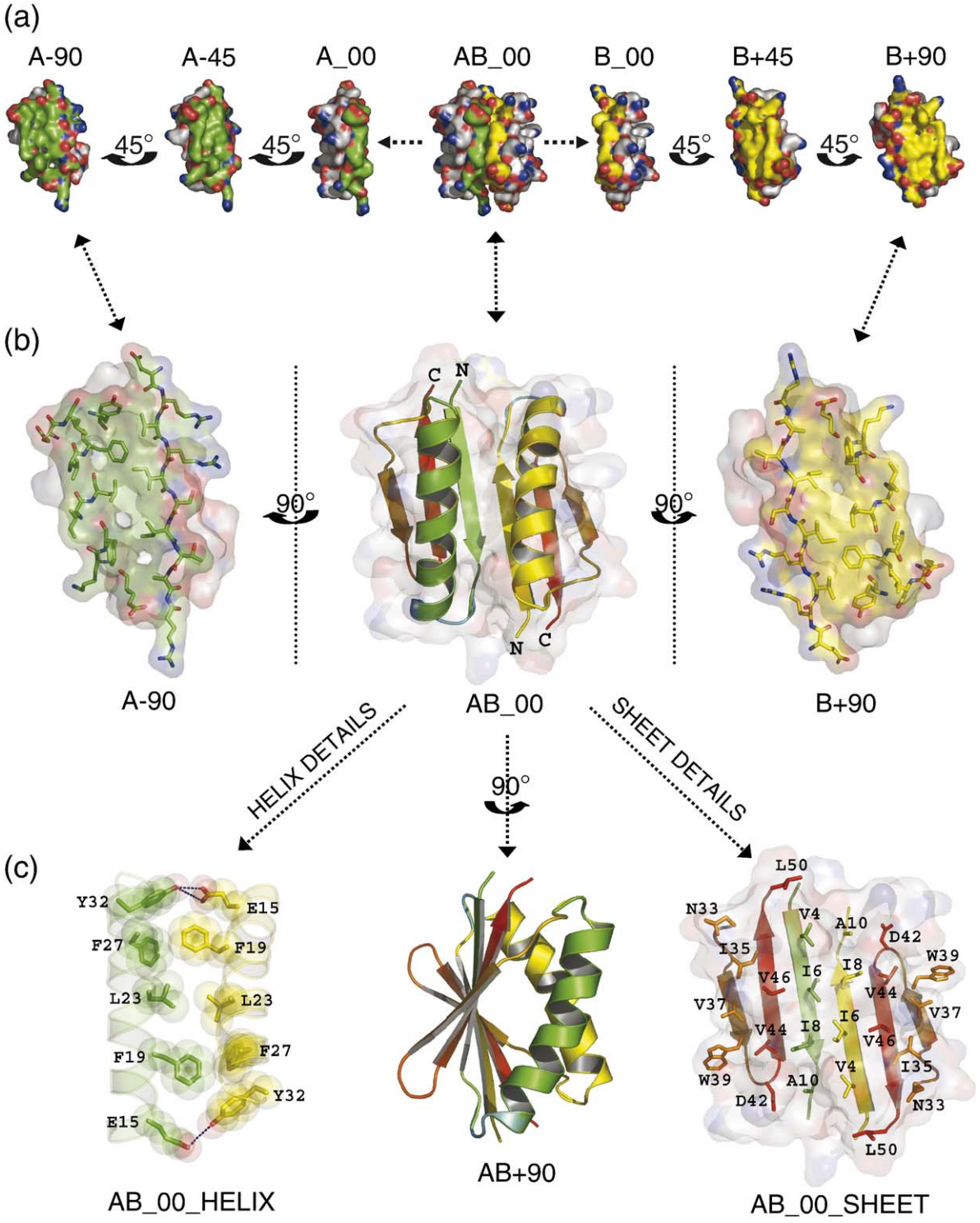

Figure 8. Details of the CFr NMR structure. (a) Seven views of the two subunits of CFr shown in surface representation. Interfacial carbon atoms are coloured green in subunit A and yellow in subunit B, and all other atoms are in CPK colour. Starting with the centre model of the dimer, the three models to the left (subunit A) and to the right (subunit B) show the dimer opening like a book. (b) Three views of CFr subunits, with the dimer model in the centre opened like a book (left: subunit A in green, right: subunit B in yellow). The centre model shows a ribbon representation of the two subunits with interfacial regions coloured in green and yellow. The flanking models show the interfacial side-chains as green or yellow sticks. Surface representations are overlaid with 80% transparency to show orientation relative to (a). (c) Specific interactions between the subunit interfaces are highlighted in the right (helices) and left (sheet) panels. Backbone secondary structure is represented as ribbons and side-chains are represented as sticks. The model in the centre of the panel is another ribbon representation of the dimer. The numerical suffix in each model label represents the degree of rotation from the centre model (in (a) and (b)) around the vertical axis in the plane of the page (e.g. B+90 is subunit B rotated 90° from the orientation of the dimer). All straight dotted arrows between models represent translations in the plane of the page. All curved arrows between models represent rotations around the vertical axis in the plane of the page.

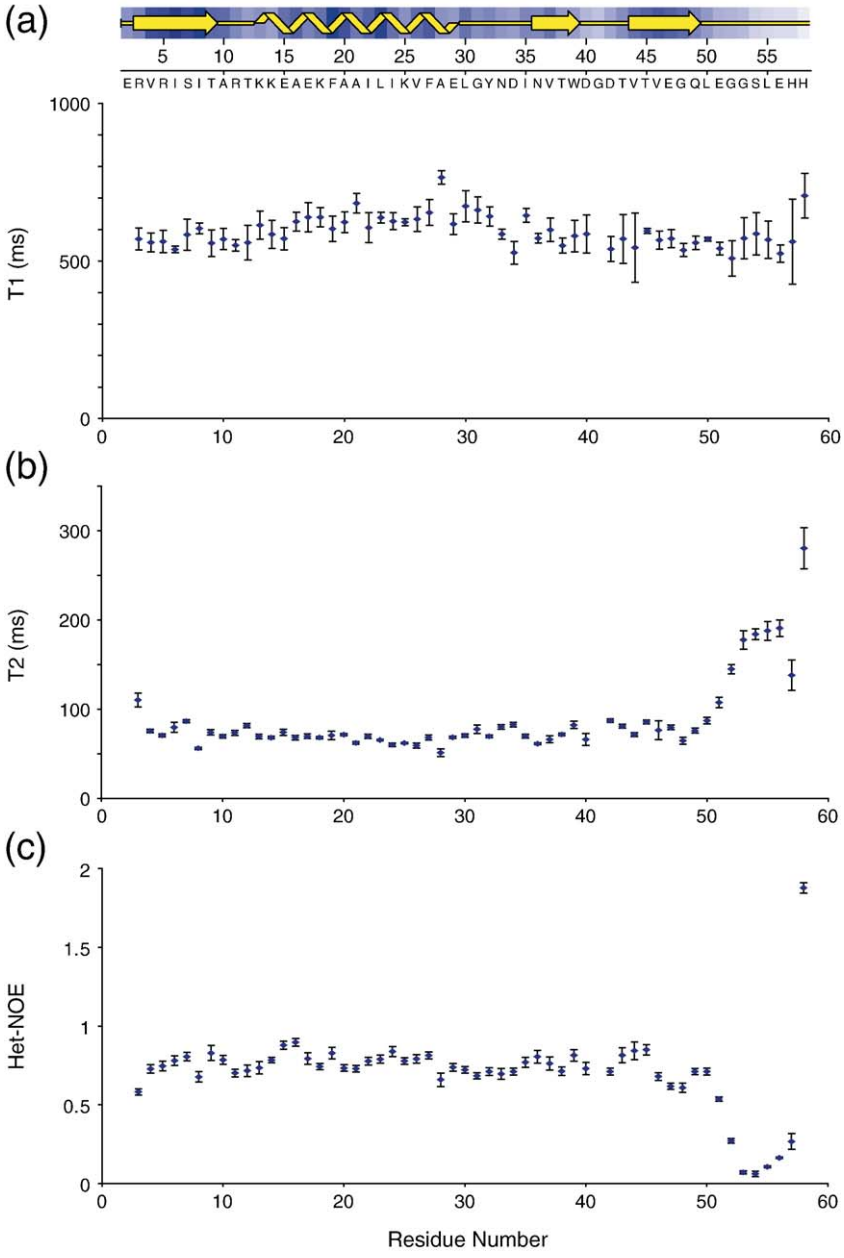

Figure 9. Backbone Dynamics of CFr. (a) ^{15}N T_1 measurements; (b) ^{15}N T_2 measurements; (c) ^{15}N Het-NOE measurements.

ribosome binding sites, defined as a start codon in the context of a strong SD sequence, are found to occur significantly less often within *E. coli* coding sequences than expected by random chance. These new results provide quantitative evidence to support the theory that evolution has selected against genetic features that would allow for mis-translation of protein sub-fragments. Additionally, the probabilistic model we implemented to quantify the *E. coli* translation initiation motifs correctly identifies the three sites of mis-translation we observed experimentally in the Top7 gene, with the dominant observed mis-translation product (CFr) scoring the highest with our model. The model also correctly predicts that either of our two independent sets of experimentally observed CFr ablating mutations (initiation codon or SD sequence) would reduce the probability of CFr mis-translation to zero. The model may be useful for identifying and removing translation initiation motifs from within any gene to be expressed in *E. coli*.

Despite the observed genetic evolutionary selection against mis-translation of internal protein fragments, many newly synthesized natural polypeptides are products of aberrant translation reactions.[12–15] This is because in addition to inappropriate translation initiation, an aberrant protein product may be produced by ribosomal processivity errors (such as ribosomal slipping, hopping, or drop-off) or missense errors where the mRNA transcript is erroneously decoded.[14] The overwhelming majority of these mis-translated proteins fail to assume native-like conformations and are cleared from the cell by post-translational processes that involve a functional cooperation between molecular chaperones assisting in folding and the proteasome system.[15–17] Super-stable protein fragments like CFr that can fold with native-like tertiary structure

would challenge this cellular surveillance machinery, and hence would be expected to be under negative evolutionary selection. When alternatively translated proteins are stable enough to evade the cellular surveillance machinery, they can compete with the natural isoform to function in a dominant-negative fashion, as in the case of the HIV-1 Gag protein.[38] A significant number of human disease pathologies involve mechanisms that implicate protein fragments that are a result of an error in translation of the native protein, including fragments of C/EBPα in acute myeloid leukemia,[39] GATA1 in Down syndrome-related leukemia,[40] c-myc in Burkitt's lymphoma,[41] and lyl-1 in T cell acute lymphoblastic leukemia.[42] The evidence for the rarity of super-stable protein sub-fragments also comes from the large body of work on limited proteolysis of natural proteins which has revealed that, with a few notable exceptions where independently folded stable native-like fragments are observed,[43,44] most proteolytic fragments are either completely unfolded or mis-folded, or adopt only partially folded states that require complementation with fragments corresponding to the rest of the protein to adopt rigid native-like protein structure.[45–47] Due to their low stability and/or conformational flexibility, most if not all of these fragments would be expected to be cleared by molecular chaperones and the proteasome before they could challenge their full-length counterparts.[15–17,48] Cellular homeostasis would be challenged only if the fragments were too stable or were being selectively overproduced (as in the case of cellular immortalization leading to cancer). This latter mechanism was also demonstrated in experiments where protein fragments of Ile-tRNA synthetase were overexpressed *in vivo*, causing dominant lethality to host cells, presumably due to fragment-induced mis-folding of the full-length protein.[49] Since stable protein sub-fragments clearly stand to disrupt homeostasis by challenging the cellular surveillance system, we propose that evolution has selected against protein structures that can yield stable sub-fragments that can adopt native-like conformations.

The simplest level of evolutionary selection, perhaps, is against extreme thermodynamic stability of any protein. We have shown that both Top7[9] and CFr display thermodynamic stability profiles significantly higher than most, if not all, natural proteins of similar shape and size. In the design of the novel sequence and topology of Top7, every amino acid was selected to stabilize the final folded structure, in the absence of any functional constraints. By contrast, nature selects proteins to fulfill very specific functions in a time-dependent fashion, and hence natural proteins need only be just stable enough to fulfill their function, after which they are cleared away by the proteasomal degradation machinery. It is reasonable to expect that extremely stable proteins (like Top7) have a higher probability of containing independently stable sub-structures than proteins of lower stability (most natural proteins). In addition to this intrinsic probability (which nature can select against), however, we show that the Top7 protein contains specific sequence and structural features that increase the ability of its sub-fragment, CFr, to achieve a stable, rigid, native-like structure, and hence suggest aspects of protein structure that may be under evolutionary control. First, the Top7 topology has a low contact order; the primary sequence separation between most structural amino acid neighbours is low. This allows Top7 to be stabilized by largely local interactions, significantly increasing the probability that contiguous sequence fragments can adopt independently stable tertiary structures. Second, the buried hydrophobic residues in the Top7 core have a high-level of sequence symmetry. Of the residues in close contact between the two helices, the first helix contributes three leucine residues and an isoleucine, while the second helix contributes two leucine residues, an isoleucine, and a valine. In the β-sheet, two core isoleucine residues on the third strand in Top7 (Ile6 and Ile8 in CFr) interact with two valine residues from the first strand in Top7. This high-level of sequence symmetry allows the CFr fragment to effectively mimic the packing of the Top7 core by self-associating into a symmetric homodimer. This mechanism has been previously observed in the proteolytically derived C-terminal fragment 255–316 of thermolysin, which also adopts a symmetric homodimeric structure, with the dimer interface effectively mimicking interactions from the core of the parent protein.[43] Finally, the interacting surfaces on the two helices of Top7 have no large protrusions or intrusions, no interdigitation of side-chains, allowing the self-interaction in the CFr dimer to be as viable as the heterologous interaction with the N-terminal portion of Top7. In addition to highlighting protein structural features that might be under evolutionary selection, our observations provide guidelines for synthetic protein engineers who either wish to avoid super-stable protein progeny or conversely wish to create protein folds that can yield stable sub-fragments for the purpose of functional regulation of the full-length protein. This balance between the danger and utility of protein sub-fragments leads us to the final evolutionary implication of our analysis.

It has been suggested that many natural single domain protein structures that have a high internal sequence and structural symmetry (such as ribonuclease inhibitor and proteins containing ankyrin or HEAT repeats) may have arisen by duplication of a single ancestral gene-product that initially formed homo-multimers of identical chains, which were gradually replaced by single polypeptide chains encoding multiple repeats.[50,51] The formation of the CFr dimer from a fragment of Top7 may parallel this natural protein-fold evolution by modular recombination of stable protein sub-structures. On the surface, this might seem to contradict the theory that evolution has selected against stable protein sub-structures. However, analyses of most modern repeat-containing proteins show that the internal interaction surfaces of the repeats have

evolved to be inter-dependent, such that in isolation, a single repeat unit cannot fold into a independent stable structure.[50–52] In fact, these observations suggest that autonomously folded ancient peptides evolved to associate interdependently into modern larger monomeric proteins with diverse functions, but in turn the ancestral peptide components of modern proteins were selected to lose their ability to fold autonomously to prevent protein fragments from interfering with the structure and function of parent domains. Evidence for the delicate nature of this evolutionary balance is clearly implied by the numerous aforementioned disease states caused by the selective stabilization of fragment isoforms of natural proteins.[39–42] Submission of the CFr structure to the DALI server[53] finds 122 natural protein domains with significant structural homology (*Z*-score > 2.0) to the CFr template. In many of these cases, the CFr subunits are found to be homologous to multiple non-overlapping parts of the same protein (e.g. the *E. coli* acriflavine resistance protein pump[54] has four distinct regions of homology to CFr), suggesting that a CFr-like module could have played a role in natural protein-fold evolution.

In addition to the evolutionary implications of the mis-translation and subsequent structural characterisation of CFr and SS.CFr, these extremely stable proteins also serve a potentially significant practical utility as novel scaffolds for further protein design. Their extremely high thermodynamic stability should allow, in principle, for their employment in industrial applications where most proteins would be rapidly degraded, such as at 100 °C or at extremely high denaturant concentrations.[55,56] Polypeptides of this length (~50 amino acids) can also routinely and cheaply be produced in high yield and purity by chemical synthesis (as opposed to bacterial expression).[57–59] Chemical synthesis has the distinct advantage over bacterial expression of allowing for the efficient and selective covalent modification of amino acids and/or the covalent addition of non-amino acid functional groups to the polypeptide chain, allowing for the potential design of extremely chemically diverse nano-scale protein machines.[59,60] The symmetric homo-dimeric nature of CFr and SS.CFr can provide an additional benefit as a scaffold, in that a singly functionalised monomer will yield a doubly functionalised macromolecular unit. Interestingly, the scorpion toxin fold family[61] has a similar overall architecture to a CFr monomer (one helix packed on a three-stranded antiparallel sheet), and has been successfully employed as a protein engineering scaffold.[56,61] However, all scorpion toxin fold proteins have six cysteine residues that participate in three specific internal disulfide bonds which are required for the protein to fold accurately, whereas CFr (with no disulfides) and SS.CFr (with only one disulfide) fold into extremely stable structures bereft of these extra internal covalent constraints. Our current efforts using CFr and SS.CFr as scaffolds include their design for the presentation of epitope-peptides for production of antibodies against HIV, and their functionalisation with peroxide-activating catalysts for bioremediation.

Materials and Methods

Protein expression and purification

The gene coding for the CFr protein sequence (amino acid residues Val48 through Gly95 in Top7) was PCR amplified from the Top7 gene sequence and cloned into plasmid pet29b(+) (Novagen). The CFr protein has the sequence: MERVRISITARTKKEAEKFAAILIKVFAELGYNDINVTWDGDTVTVEGQLEGGSLEHHHHHH. The SS.CFr gene construct was generated by PCR amplifying the CFr construct using oligonucleotide primers that add a Cys-Glu sequence at position 3 and change Glu51 to Cys, and sub-cloning this fragment back into pet29b(+). The SS.CFr protein has the sequence: MECERVRISITARTKKEAEKFAAILIKVFAELGYNDINVTWDGDTVTVEGQLCGGSLEHHHHHH. Point mutants of Top7 (ATG1ATT, GTG48GTT, and GGG44TCT) were generated using the Quick Change Site-Directed mutagenesis kit (Stratagene).

The 6× histidine-tagged proteins were expressed in the BL21(DE3)pLysS strain of *E. coli*. Cells were grown in LB media at 37 °C to an A_{600} of 0.6, induced with 1 mM isopropyl-thio-β-D-galactosidase (IPTG), and cells were harvested after another 4–5 h of growth at 37 °C. Harvested cells were lysed by sonication, and soluble protein collected after centrifugation of cellular debris. Soluble protein was purified on a Ni^+ affinity column (Pharmacia) followed by 10^4-fold dialysis against 25 mM Tris–HCl (pH 8.0). The protein was further purified on a QFF anion exchange column (Pharmacia) with a 50 mM to 600 mM NaCl gradient in 25 mM Tris–HCl (pH 8.0), followed by a final 10^4-fold dialysis against 25 mM Tris–HCl (pH 8.0) (or 50 mM sodium phosphate (pH 7.0) for NMR). To ensure complete disulfide formation, anion-exchange purified SS.CFr was oxidised in the presence of 20 mM potassium ferricyanide [$K_3Fe(CN_6)$] for 10 min at room temperature, prior to the final dialysis steps. Protein identity and purity were 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†).

For NMR studies, uniformly ^{15}N and $^{15}N/^{13}C$ labelled samples were prepared by growing bacteria in M9 minimal media supplemented with 0.5 g/l of [^{15}N] NH_4Cl and 2 g/l of [^{13}C]glucose (Spectra Isotope). Purification was identical to that executed for the unlabelled samples. For $^{12}C/^{13}C$ filtered NOESY experiments, equimolar amounts of $^{15}N^{12}C$ and $^{15}N^{13}C$ samples were mixed, the protein was then denatured in 8 M GuHCl with overnight mixing to ensure complete monomerisation, dialysed back into 50 mM NaPi (pH 7.0) to allow refolding and dimerisation, lyophilised, and brought up in 100% D_2O.

Limited proteolysis

Bacterial cells containing over-expressed Top7 were lysed by three freeze-thaw cycles in the presence or

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

absence of protease inhibitors (1 mM PMSF, 1 mM benzamidine). These two lysates were then divided into four equal fractions, which were incubated at room temperature for 2, 4, 24, and 72 h, respectively. After the incubation period, the lysates were centrifuged and separated into supernatant and pellet, which were subsequently visualised by SDS-PAGE.

Statistical analysis of *E. coli* ribsome binding site motifs

There are three steps in the measurement of ribosome binding sites within *E. coli* protein-coding regions: (1) infer a probabilistic model M of the true upstream ribosome binding sites; (2) use the model M to measure the observed number of high-scoring ribosome binding sites within real protein-coding regions; and (3) use the model M to measure the expected number μ of high-scoring ribosome binding sites and its standard deviation σ in randomly generated coding regions.

For the first step, we adopted the simple iterative approach of Kibler & Hampson.[62] The *E. coli* genome contains 2912 annotated genes each with at least 20 bp of non-coding DNA upstream of their start codons. From this training set T of 2912 upstream sequences each of length 20 bp, we extracted all 7-mers that differ in at most one position from TAAGGAG, known to be the optimal Shine–Dalgarno sequence,[63] since it is the reverse complement of the 3′ end of *E. coli*'s 16 S rRNA. The initial Shine–Dalgarno profile P was formed from this collection of 7-mers. This profile is a 4×7 matrix whose columns give the probability distributions for each of the seven positions of the Shine–Dalgarno sequence. The profile P can then be used to score any 7-mer. We then iterated the following process until convergence. (1) Extract from the training set T the 2000 7-mers that score highest according to the profile P. (2) Use these 2000 7-mers to compute a new profile P. When this process converges, P should be a good approximation to the distribution of true Shine–Dalgarno sequences.

Using the 2000 highest scoring matches of this converged profile P in the training set T, we computed the 4×3 profile of the start codons and the probability distribution of the distance separating the Shine–Dalgarno 7-mer from the start codon. These two profiles and the distance distribution are shown in Supplementary Data, Table S1. Together, the three distributions given in Table S1 comprise the probabilistic model M described above, and can be used to score any ribosome binding site.

The next step was to use M to measure the observed number of high-scoring ribosome binding sites within *E. coli*'s 4237 annotated protein-coding regions. For any score threshold S (such as the thresholds shown in Table 1), the model M of Supplementary Data, Table S1 can be used to count the number of sequences internal to protein-coding regions with scores at least S. We insisted that the internal "start codon" that matches Supplementary Data, Table S1(c) occur in the correct reading frame, so that translation could proceed in that open reading frame.

Finally, we repeated this counting process in "random" coding regions. We simulated random coding regions by randomly permuting the codons of all of *E. coli*'s real protein-coding regions. We kept each codon intact so as to preserve *E. coli*'s natural codon biases. We repeated this random codon shuffling process 300 times, computing the means and standard deviations in Table 1 over these 300 trials.

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 concentrations used for NMR (600 μM–1.2 mM) or CD (5-100 μM) were equilibrated in 20 mM EDTA, 25 mM Tris (pH 8.0) at 25 °C, and run on the Superdex-750 column at 1 ml/min.

Small-angle X-ray scattering (SAXS)

SAXS measurements were carried out at the BESSRC-CAT beamline 12-ID at the Advanced Photon Source (Argonne, IL). Immediately before data collection, the samples were centrifuged for 10 min at 11,000***g***. The measurements were performed at 25(±1) °C in a custom-made, thermostated flow cell[64] at a flow rate of ~1 ml/min and a photon energy of 12 keV. For each condition, a total of 40 measurements of 1.0 s integration time each were taken. All data were image-corrected and circularly averaged after data taking. The 40 profiles for each condition were averaged, and appropriate buffer scattering profiles were subtracted for background correction. There were no signs of radiation damage. Measurements were performed at varying concentrations of GuHCl in 25 mM Tris (pH 8.0), at a protein concentration of 14.5 mg/ml, unless otherwise noted.

Changes in protein conformation monitored by SAXS are represented as Kratky plots, which are graphs of $s^2 \cdot I(s)$ as a function of s, where s is the momentum transfer vector ($s = 2\sin(\theta)/\lambda$, where $\lambda = 1$ Å is the X-ray wavelength and 2θ is the scattering angle). Porod's Law states that for large s the scattering from an object with a well defined surface falls approximately as s^{-4},[65] which leads to decrease as s^{-2} in the Kratky representation for large s. Well-folded proteins therefore have a characteristic peak in the Kratky plot. Scattering from a random polymer falls like s^{-1}, which leads to a linear rise at high s in the Kratky plot for unfolded proteins.[66]

Analytical ultra-centrifugation (AUC)

Sedimentation equilibrium studies on CFr and SS.CFr were conducted in a Beckman XL-A analytical ultracentrifuge using 12 mm Epon charcoal-filled centerpieces containing six channels. Studies on each protein were conducted at three concentrations in 25 mM Tris (pH 8.0) in the presence of 0, 3, 4, 5, 6, and 7 M guanidine hydrochloride (GuHCl). Centerpiece sample channels were filled with 110 μl of protein sample and reference channels were filled with 120 μl of matched solvent. All scans were conducted at 20 °C at an absorbance wavelength of 280 nm. Protein concentrations were determined using scans conducted at 3000 rpm and low, intermediate, and high protein concentrations that fell in the range of 33–39 μM, 59–66 μM, and 89–105 μM. Scans were collected at three rotor speeds, 25,000, 30,000, and 45,000 rpm, using equilibration times of 10 h for each speed. This equilibration time was deemed sufficient by identical absorbance scans collected after 8 and 10 h at each speed.

Solvent densities were determined at 20 °C using an Anton Paar DMA 5000 densitometer. Triplicate measurements were collected and averaged for 25 mM Tris (pH 8.0) in the presence of varying concentrations of GuHCl (Supplementary Data, Table S2). The partial specific volumes of CFr and SS.CFr (0.733 and 0.730 ml/g, respectively) were determined at 25 °C from amino acid composition[67] and adjusted to 20 °C.[68]

Data analysis was performed using Beckman XL-A Data Analysis Software.version 4.0. Individual equilibrium scans were fit to a single ideal species model using non-linear least-squares analysis to determine a weight-averaged molecular mass, M_r. During this analysis, the baseline offset was allowed to float; if it was found to be >±0.08, it was fixed to zero so that the goodness of fit could be assessed for each case. Analysis of residuals to the fit allowed for detection of aggregation or non-ideal behavior in a few scans. Following this analysis, global fits were performed across the three protein concentrations and three speeds (nine scans) to re-determine M_r. The residuals, typically small (<±0.02) and random, and baseline offsets (typically <±0.04) were most often improved during the global data analysis.

Circular dichroism

CD data were collected on an Aviv 62A DS spectrometer. Far-UV CD wavelength scans (260 nm–195 nm) at varying protein concentrations (10 μM −25 μM), guanidinium hydrochloride (GuHCl) concentrations (0–8.3 M), and temperatures (0–98 °C) were collected in a 1 mm path length cuvette. GuHCl induced protein denaturation was followed by the change in ellipticity at 220 nm in a 1 cm path length cuvette, using a Microlab titrator (Hamilton) for denaturant mixing. Temperature was maintained at 25 °C with a Peltier device. Temperature-induced protein denaturation was followed by the change in ellipticity at 220 nm in a 2 mm path length cuvette. All CD data were converted to mean residue ellipticity. The dimer dissociation constant (K_d) and the free energy of unfolding (ΔG_U^{H2O}) were calculated according to the procedure described by Kuhlman and co-workers,[69] where chemical denaturation curves were fit to an equilibrium model between unfolded monomer (U) and folded dimer (F):

$$F_2 \xleftrightarrow{K_d} 2U$$

where:

$$\exp\left(\frac{-\Delta G^\circ}{RT}\right) = K_\mathrm{d} = \left[U\right]^2 \Big/ \left[F_2\right] = 2P_\mathrm{t}\left[f_\mathrm{u}^2/(1-f_\mathrm{u})\right]$$

where P_t is the total protein concentration, f_u is fraction of unfolded protein, R is the gas constant and T is the temperature. The final equation used to fit the circular dichroism data (θ) takes the form:

$$\theta([Gu]) = (\theta_\mathrm{U} - \theta_\mathrm{F})\cdot f_\mathrm{u} + \theta_\mathrm{F}$$

where:

$$f_u = 0.5\left(-\alpha + \sqrt{(\alpha^2 + 4\alpha)}\right),$$

$$\alpha = \left(\frac{\exp\left(\frac{-\Delta G^\circ}{RT}\right)}{2P_\mathrm{t}}\right)$$

and $\Delta G°$ and the circular dichroism signal of folded (θ_F) and unfolded (θ_U) protein are assumed to vary linearly with denaturant concentration:

$$\begin{aligned}\Delta G^\circ([GuHCl]) &= \Delta G^\circ(0M\ Gu - HCl) + m\cdot[GuHCl]\\ \theta_U([GuHCl]) &= \theta_U(0M\ GuHCl) + a\cdot[GuHCl]\\ \theta_F([GuHCl]) &= \theta_F(0M\ GuHCl) + b\cdot[GuHCl].\end{aligned}$$

Nuclear magnetic resonance spectroscopy

All CFr samples were prepared for NMR experiments in Shigemi susceptibility-matched NMR tubes, at 0.7 mM–1.0 mM concentration in H_2O solution containing 10% D_2O or in 100% D_2O, 50 mM sodium phosphate (pH 7.0). All experiments were recorded at 298K 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. Three-dimensional ^{15}N-edited NOESY spectra and 2-dimensional NOESY and TOCSY datasets were recorded on a Bruker Avance 750 MHz spectrometer equipped with a TXI HCN triple resonance probe with z-axis gradient. Three-dimensional ^{13}C-edited NOESY and two and three-dimensional ^{12}C/^{13}C-filtered NOESY spectra were recorded at Environmental Molecular Sciences Laboratory (EMSL) at PNNL in Richland, WA using a Varian 600MHz spectrometer equipped with a cryoprobe. Data were processed with NMRPipe[70] and analyzed with SPARKY‡.

Backbone amide ^{1}H and ^{15}N, Cα, C=O and side chain Cβ resonances were assigned using ^{1}H-^{15}N HSQC, HNCO, HNCACB, CBCA(CO)NH, HBHA(CO)NH, HN(CO)CA and 3D ^{15}N edited TOCSY experiments.[71] Over 98% of the backbone N, (N)H, C(O), Cα and Cβ nuclei for residues 2–58 could be assigned (no assignments were possible for the N-terminal methionine and for the last four histidine residues at the C terminus). 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 two-dimensional NOESY and TOCSY spectra recorded in D_2O buffers. Side-chain ^{1}H and ^{13}C resonances were >92% assigned, whereas the aromatic side-chains (Phe, Tyr, Trp) were >68% assigned. Gln/Asn NH_2 were 100% assigned while Arg Nε and guanidinium groups and Lys NH_3 remain unassigned. The spectra used in deriving distance constraints included 3D ^{15}N-edited NOESY and 3D ^{13}C-edited NOESY, 2D NOESY in H_2O (80 ms and 120 ms mixing) and 2D NOESY in D_2O (120 ms mixing) recorded at 750MHz. Additionally, inter-subunit distance constraints were derived from 2D and 3D ^{12}C/^{13}C-filtered NOESY spectra.[27,28]

Protein structure determination by NMR

Structure determination was conducted in a two-step process using the program CYANA 2.0,[72] a fully automated iterative step for generating models of the monomeric unit of CFr, followed by a partly automated iterative step for building the symmetric homo-dimer model with manually assigned interfacial constraints. Fully automated structure determination of the CFr dimer was not possible in CYANA because the symmetric nature of the dimer made it impossible for the program to distinguish between inter-subunit and intra-subunit NOEs. The experimental NMR data used for structural analysis included the NOESY peak lists derived from the 3D ^{15}N- and ^{13}C-edited NOESY data together with the 2D NOESY data collected in both H_2O and D_2O. In addition, the 2D and 3D ^{12}C/^{13}C-filtered NOESY peak

‡ Goddard, T. D. & Kneller, D. G. (2005). SPARKY 3.111. University of California, San Fransisco) on Windows or Linux workstations.

lists were added prior to the second step. Hydrogen bonding constraints derived from slow amide exchange data (as described below), and Φ–Ψ angle constraints generated from chemical shift data using the program TALOS[73] were also used. The NOESY peak lists used as input for automated analysis with CYANA were generated automatically using the program SPARKY based on the chemical shift list generated in the assignment process. Peaks volumes were calculated using SPARKY's Gaussian integration tool. Slowly exchanging amides were identified by lyophilizing the protein from H_2O, then dissolving it in D_2O and acquiring 2D ^1H-^{15}N HSQC spectra at 30 min and 50 h after dissolving in D_2O. Hydrogen bond donors were identified by the presence of an amide peak in the 2D ^1H-^{15}N HSQC spectrum recorded at 30 min. The corresponding acceptors were identified by visualizing PDB files obtained from CYANA in Rasmol 2.7.1[74] to identify carbonyl groups that were at a distance of approximately 2.0 Å from slow exchanging hydrogen atoms. Each step of structural refinement in CYANA was performed with and without these hydrogen-bonding constraints.

For the structure determination of a single subunit of CFr (i.e. one chain from the symmetric homo-dimer or CFrA), 3873 NOE peaks (many of them repetitions of the same peak observed in different spectra) were semi-automatically generated from 3D ^{15}N and ^{13}C-edited NOESY and 2D NOESY spectra in H_2O and D_2O, using the program SPARKY. In addition, 76 dihedral constraints were generated with the program TALOS and 32 hydrogen bond constraints were generated by analysis of D_2O protection experiments. The NOEASSIGN macro in CYANA was used to automatically assign >92% of the NOE input peaks. Together with the dihedral and hydrogen bond constraints, the 3783 initial cross-peaks yielded 1116 unique distance constraints that were used in the final CFrA structure calculation. In the final calculation, 100 structures were generated, of which the top 20 structures had an average target function value of 5.24(±0.08) Å^2 and an ensemble RMSD value of 0.24(±0.08) Å over backbone atoms and 0.76(±0.14) Å over heavy atoms in residues 3 through 51. There were 16 distance constraint violations between 0.1 Å −0.25 Å and two angle constraint violations of between 33°–36°.

In the next step of refinement, results from the CFrA structure calculation were combined with inter-subunit NOE data from the 2D and 3D $^{12}C/^{13}C$ filtered NOESYs to determine the CFr dimer structure. The CFrA sequence list, chemical shift list and the intra-subunit distance constraints derived from the last round of CYANA were duplicated to generate an equivalent copy of data for a second chain labelled CFrB. A flexible 60 Å tether was introduced between the C terminus of CfrA and the N terminus of CFrB to allow each monomer to refine separately during the calculation while also allowing a generous range of motion for relative inter-subunit re-orientation. A total of 23 inter-subunit NOEs were assigned by manually inspecting the 2D and 3D ^{12}C-^{13}C filtered NOESYs in SPARKY, based on the earlier intra-subunit backbone and side-chain assignments and the intra-subunit NOEs assignments from the CYANA runs. All inter-subunit NOE assignments were made as double assignments between equivalent pairs of interacting nuclei between CFrA and CFrB (i.e. an interaction assigned between nucleus X on CFrA with nucleus Y on CFrB automatically implied the same interaction between nucleus Y on CFrA with nucleus X on CFrB). Peak volumes calculated by SPARKY's Gaussian integration tool were converted into upper distance constraints in CYANA by setting the ratio of volumes to upper distance constraints equal to that obtained in the automated intra-subunit NOE assignment step. This inter-subunit upper distance constraint list was then used in combination with the CFrA and CFrB chemical shift lists and intra-subunit distance constraint lists as input for a single round of structure calculation that consisted of 100 separate simulated annealing runs using torsion angle dynamics. Similar structure calculations were also run with CFrA duplicated hydrogen-bonding constraints and TALOS-derived dihedral angle constraints, including hydrogen bonds that were observed across the interface. All violated constraints were investigated and were removed or modified only if it appeared that they had been mis-assigned (intra-subunit instead of inter-subunit) or poorly integrated. Unassigned NOEs from the CFrA automated structure calculation were also investigated at this stage to assign them, if possible, as inter-subunit NOEs. Two cycles of this type of refinement were sufficient to obtain structures with appropriate target function values, tight ensemble convergence and no distance or dihedral violations. The only violation after the final CYANA run was the same single intra-residue close atom contact in each monomer (Ile35 CG2 to C(O) violated by 0.33 Å). The quality of the final structure was evaluated with ProcheckNMR.[29] Experimental constraints and structural statistics are reported in Tables 2 and 3, respectively.

Solvent accessible surface area (SASA)

SASA was calculated using the program NACCESS§ SASA buried in the dimer interface (D_{SASA}) was calculated as:

$$D_{SASA} = (CFrA_{SASA} + CFrB_{SASA}) - CfrAB_{SASA}$$

where $CFrA_{SASA}$ and $CFrB_{SASA}$ are the SASA for each subunit treated separately, and $CFrAB_{SASA}$ is the SASA for the dimer structure. Interfacial residues are defined as any amino acid that loses >1 Å^2 SASA when the dimer is compared to the individual subunits.

Measurements of ^{15}N nuclear relaxation rates and ^{15}N-^1H heteronuclear NOEs

Standard pulse sequences were used to measure the ^{15}N T_1, T_2 and heteronuclear NOEs.[75,76] All experiments utilize pulsed-field gradients for coherence selection, reduction of artefacts and sensitivity enhancement. In the CPMG sequence of the T_2 experiment, ^1H 180° pulses were applied for elimination of cross-correlation between ^1H-^{15}N dipolar and ^{15}N CSA relaxation mechanisms.[77] A delay of 0.75 ms was inserted between successive applications of ^{15}N 180° with ^1H 180° pulses applied every 4 ms in the CPMG pulse train. Spectra were recorded with 112 complex points in the indirect dimension and with spectral widths of 1822.49 and 6009.6 in the ^{15}N and ^1H dimensions, respectively. Delays of 0.030, 0.060, 0.100, 0.150, 0.220, 0.310, 0.420, and 0.550 s were used for the T_1 experiments. T_2 spectra were measured from spectra recorded with delays of 0.008, 0.016, 0.024, 0.032, 0.048, 0.064, 0.080, 0.096, and 0.120 s. The relaxation delay was 1.9 s for each experimental set. For the

§ Hubbard, S. J. & Thornton, J. M. (1993). NACCESS. Department of Biochemistry and Molecular Biology, University College London).

heteronuclear NOE measurements, a pair of spectra was recorded with and without proton saturation that was achieved by application of ^{1}H 120° pulses every 5 ms. Spectra recorded with proton saturation utilized a 2 s recycle delay followed by a 3 s period of saturation, while those recorded in the absence of saturation employed a recycle delay of 5 s.

All spectra were processed using NMRPipe/NMRDraw software with polynomial baseline correction after multiplication with cosine-bell window functions. Linear prediction was applied in the indirect dimension to increase the number of complex points in that dimension to 224 in the T_1/T_2 heteronuclear NOE experiments, followed by zero filling to generate 512 points. Peak heights were calculated for every assigned peak in the T_1 and T_2 spectra and fitted into an exponential curve using the SPARKY relaxation fit software‖ T_1 and T_2 values were determined from the decay curves using the equation:

$$I(t) = I(o)\exp(-\tau/T_{1,2})$$

Where $I(o)$ is the initial peak intensity and τ is the delay time. The error estimates for the rate constants reflects the likely error of the best fit from the parameters obtained for a perfect exponential decay. Average values and errors are reported in Results.

Heteronuclear NOE values were calculated from the ratio of peak heights with and without proton saturation. Errors in these measurements were estimated from the plane base noise in 2D ^{1}H-^{15}N-HSQC spectra recorded with and without proton saturation.

X-ray crystallography

SS.CFr was crystallized in hanging drops (1 μl of protein solution at 20 mg/ml with 1 μl of well solution). The well solutions ranged from 30%–40% (v/v) methyl-2,4-pentanediol (MPD), 6% PEG-4K and 0.1 M of Na-Hepes (pH 6.9). The protein crystals grew within two to six days and were between 50 μm −200 μm on a side. Since MPD is a cryoprotectant at 30–40%, crystals were dunked in fresh well solution and directly flash frozen in liquid nitrogen. With this treatment, the crystals diffracted in a tetragonal space group ($P4_32_12$) with unit cell dimensions $a=58.3$ Å, $b=58.3$ Å, $c=96.7$ Å. A single wavelength (0.9793 Å) native data set was collected to 3.6 Å resolution on beam-line 5.4.1 at the ALS (Advanced Light Source, Lawrence Berkeley Laboratory, Berkeley) using a four panel ADSC CCD area detector. Data were processed and scaled using HKL2000.[78]

The phases for the SS.CFr dataset were solved by molecular replacement (MR) with the program EPMR¶. Residues Glu2–Leu50 in both subunits of the CFr NMR structure (best NMR model) were used as the search model. The two subunits were input as separate chains to allow for relative rigid-body re-orientation. The correlation coefficient for the initial MR search, using data to 4.0 Å resolution, was 0.58, *versus* background of 0.36. Further structural refinement against the model-derived MR phases was attempted with model building in simulated annealing composite-omit maps in XtalView,[79] along with rigid-body refinement, torsion-angle based simulated annealing, and conjugate-gradient based minimization in CNS.[80]

Protein Data Bank and BioMagRes database accession numbers

The coordinates and corresponding NMR constraint files for 20 NMR-derived CFr structures have been deposited with the RCSB Protein Data Bank[a] under the identifier code 2GJH, and the chemical shift list corresponding to this structure determination has been deposited in the BioMagRes Database[b] under the accession code 7101.

Acknowledgements

We acknowledge the expert assistance of Steve Reichow, Tom Leeper, and Kate Godin in NMR data collection and processing, and modelling and refinement of the CFr structure; Priti Deka for help with NMR dynamics analysis of CFr; Juan Pizarro and Django Sussman for help with crystallographic data collection and processing; Soenke Seifert for help with SAXS data collection; Mark DePristo for insightful comments about mechanisms of protein evolution; the facilities at NMRFAM (Madison, WI, supported by NIH) and PNNL (Richland, WA, supported by DOE) for access to NMR instrumentation, and the facilities at the Advanced Light Source (Berkeley, CA, supported by DOE); and the Advanced Photon Source (Argonne, IL, supported by DOE) for access to their synchrotron-source X-ray beamlines. This work is supported in part by grants from NIH-NIGMS (to G.V.) and NIH and HHMI (to D.B.).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.07.092

References

1. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science*, **278**, 82–87.
2. Dantas, G., Kuhlman, B., Callender, D., Wong, M. & 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.
3. Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2004). Computational design of a biologically active enzyme. *Science*, **304**, 1967–1971.
4. Korkegian, A., Black, M. E., Baker, D. & Stoddard,

‖ Goddard, T. D. & Kneller, D. G. (2005). SPARKY 3.111. University of California, San Fransisco).

¶ Kissinger, C. R. & Gehlhaar, D. K. (1997). EPMR: a program for crystallographic molecular replacement by evolutionary search. Agouron Pharmaceuticals, La Jolla, CA).

[a] http://www.rcsb.org/pdb/
[b] http://www.bmrb.wisc.edu

B. L. (2005). Computational thermostabilization of an enzyme. *Science*, **308**, 857–860.
5. Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L. & Baker, D. (2004). Computational redesign of protein-protein interaction specificity. *Nature Struct. Mol. Biol.* **11**, 371–379.
6. Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D., Monnat, R. J. & Stoddard, B. L. (2002). Design, activity, and structure of a highly specific artificial endonuclease. *Mol. Cell.* **10**, 895–905.
7. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature*, **423**, 185–190.
8. 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.
9. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.
10. Dobson, N., Dantas, G., Baker, D. & Varani, G. (2006). High-resolution structural validation of the computational redesign of human U1A protein. *Structure*, **14**, 847–856.
11. Dahiyat, B. I. (1999). In silico design for protein stabilization. *Curr. Opin. Biotechnol.* **10**, 387–390.
12. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Rev. Genet.* **6**, 678–687.
13. Kozak, M. (1999). Initiation of translation in prokaryotes and eukaryotes. *Gene*, **234**, 187–208.
14. Kurland, C. G. (1992). Translational accuracy and the fitness of bacteria. *Annu. Rev. Genet.* **26**, 29–50.
15. McClellan, A. J., Tam, S., Kaganovich, D. & Frydman, J. (2005). Protein quality control: chaperones culling corrupt conformations. *Nature Cell. Biol.* **7**, 736–741.
16. Vabulas, R. M. & Hartl, F. U. (2005). Protein synthesis upon acute nutrient restriction relies on proteasome function. *Science*, **310**, 1960–1963.
17. McClellan, A. J., Scott, M. D. & Frydman, J. (2005). Folding and quality control of the VHL tumor suppressor proceed through distinct chaperone pathways. *Cell*, **121**, 739–748.
18. Cazzola, M. & Skoda, R. C. (2000). Translational pathophysiology: a novel molecular mechanism of human disease. *Blood*, **95**, 3280–3288.
19. Bence, N. F., Sampat, R. M. & Kopito, R. R. (2001). Impairment of the ubiquitin-proteasome system by protein aggregation. *Science*, **292**, 1552–1555.
20. Horwich, A. (2002). Protein aggregation in disease: a role for folding intermediates forming specific multimeric interactions. *J. Clin. Invest.* **110**, 1221–1232.
21. Kozak, M. (2002). Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, **299**, 1–34.
22. Dobson, C. M. (1999). Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24**, 329–332.
23. Cohen, F. E. & Kelly, J. W. (2003). Therapeutic approaches to protein-misfolding diseases. *Nature*, **426**, 905–909.
24. Selkoe, D. J. (2003). Folding proteins in fatal ways. *Nature*, **426**, 900–904.
25. Maurizi, M. R. (1992). Proteases and protein degradation in *Escherichia coli*. *Experientia*, **48**, 178–201.
26. Gualerzi, C. O. & Pon, C. L. (1990). Initiation of mRNA translation in prokaryotes. *Biochemistry*, **29**, 5881–5889.
27. Folkers, P. J. M., Folmer, R. H. A., Konings, R. N. H. & Hilbers, C. W. (1993). Overcoming the ambiguity problem encountered in the analysis of nuclear overhauser magnetic-resonance spectra of symmetrical dimer proteins. *J. Amer. Chem. Soc.* **115**, 3798–3799.
28. Zwahlen, C., Legault, P., Vincent, S. J. F., Greenblatt, J., Konrat, R. & Kay, L. E. (1997). Methods for measurement of intermolecular NOEs by multinuclear NMR spectroscopy: application to a bacteriophage lambda N-peptide/boxB RNA complex. *J. Amer. Chem. Soc.* **119**, 6711–6721.
29. Laskowski, R. J., Macarthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystalogl.* **26**, 283–291.
30. Sakai, H., Imamura, C., Osada, Y., Saito, R., Washio, T. & Tomita, M. (2001). Correlation between Shine–Dalgarno sequence conservation and codon usage of bacterial genes. *J. Mol. Evol.* **52**, 164–170.
31. Stenstrom, C. M., Holmgren, E. & Isaksson, L. A. (2001). Cooperative effects by the initiation codon and its flanking regions on translation initiation. *Gene*, **273**, 259–265.
32. Yamagishi, K., Oshima, T., Masuda, Y., Ara, T., Kanaya, S. & Mori, H. (2002). Conservation of translation initiation sites based on dinucleotide frequency and codon usage in *Escherichia coli* K-12 (W3110): non-random distribution of A/T-rich sequences immediately upstream of the translation initiation codon. *DNA Res.* **9**, 19–24.
33. Ma, J., Campbell, A. & Karlin, S. (2002). Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J. Bacteriol.* **184**, 5733–5745.
34. Kozak, M. (1984). Selection of initiation sites by eucaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin. *Nucl. Acids Res.* **12**, 3873–3893.
35. Spanjaard, R. A. & van Duin, J. (1989). Translational reinitiation in the presence and absence of a Shine and Dalgarno sequence. *Nucl. Acids Res.* **17**, 5501–5507.
36. de Smit, M. H. & van Duin, J. (1990). Control of prokaryotic translational initiation by mRNA secondary structure. *Prog. Nucl. Acid Res. Mol. Biol.* **38**, 1–35.
37. Saito, R. & Tomita, M. (1999). On negative selection against ATG triplets near start codons in eukaryotic and prokaryotic genomes. *J. Mol. Evol.* **48**, 213–217.
38. Schubert, U., Ott, D. E., Chertova, E. N., Welker, R., Tessmer, U., Princiotta, M. F. *et al.* (2000). Proteasome inhibition interferes with gag polyprotein processing, release, and maturation of HIV-1 and HIV-2. *Proc. Natl Acad. Sci. USA*, **97**, 13057–13062.
39. Pabst, T., Mueller, B. U., Zhang, P., Radomska, H. S., Narravula, S., Schnittger, S. *et al.* (2001). Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-alpha (C/EBPalpha), in acute myeloid leukemia. *Nature Genet.* **27**, 263–270.
40. Wechsler, J., Greene, M., McDevitt, M. A., Anastasi, J., Karp, J. E., Le Beau, M. M. & Crispino, J. D. (2002). Acquired mutations in GATA1 in the megakaryoblastic leukemia of Down syndrome. *Nature Genet.* **32**, 148–152.
41. Hann, S. R., King, M. W., Bentley, D. L., Anderson, C. W. & Eisenman, R. N. (1988). A non-AUG translational initiation in c-myc exon 1 generates an N-terminally distinct protein whose synthesis is disrupted in Burkitt's lymphomas. *Cell*, **52**, 185–195.

42. Mellentin, J. D., Smith, S. D. & Cleary, M. L. (1989). lyl-1, a novel gene altered by chromosomal translocation in T cell leukemia, codes for a protein with a helix-loop-helix DNA binding motif. *Cell*, **58**, 77–83.
43. Rico, M., Jimenez, M. A., Gonzalez, C., De Filippis, V. & Fontana, A. (1994). NMR solution structure of the C-terminal fragment 255-316 of thermolysin: a dimer formed by subunits having the native structure. *Biochemistry*, **33**, 14834–14847.
44. Tasayco, M. L. & Carey, J. (1992). Ordered self-assembly of polypeptide fragments to form nativelike dimeric trp repressor. *Science*, **255**, 594–597.
45. Fontana, A., de Laureto, P. P., Spolaore, B., Frare, E., Picotti, P. & Zambonin, M. (2004). Probing protein structure by limited proteolysis. *Acta Biochim. Pol.* **51**, 299–321.
46. Wu, L. C., Grandori, R. & Carey, J. (1994). Autonomous subdomains in protein folding. *Protein Sci.* **3**, 369–371.
47. Philipp, S., Kim, Y. M., Durr, I., Wenzl, G., Vogt, M. & Flecker, P. (1998). Mutational analysis of disulfide bonds in the trypsin-reactive subdomain of a Bowman-Birk-type inhibitor of trypsin and chymotrypsin–cooperative versus autonomous refolding of subdomains. *Eur. J. Biochem.* **251**, 854–862.
48. Goldberg, A. L. (2003). Protein degradation and protection against misfolded or damaged proteins. *Nature*, **426**, 895–899.
49. Michaels, J. E., Schimmel, P., Shiba, K. & Miller, W. T. (1996). Dominant negative inhibition by fragments of a monomeric enzyme. *Proc. Natl Acad. Sci. USA*, **93**, 14452–14455.
50. Lupas, A. N., Ponting, C. P. & Russell, R. B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134**, 191–203.
51. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134**, 117–131.
52. Grishin, N. V. (2001). Fold change in evolution of protein structures. *J. Struct. Biol.* **134**, 167–185.
53. Holm, L. & Sander, C. (1995). Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.* **20**, 478–480.
54. Yu, E. W., McDermott, G., Zgurskaya, H. I., Nikaido, H. & Koshland, D. E., Jr (2003). Structural basis of multiple drug-binding capacity of the AcrB multidrug efflux pump. *Science*, **300**, 976–980.
55. Bloom, J. D., Silberg, J. J., Wilke, C. O., Drummond, D. A., Adami, C. & Arnold, F. H. (2005). Thermodynamic prediction of protein neutrality. *Proc. Natl Acad. Sci. USA*, **102**, 606–611.
56. Martin, L. & Vita, C. (2000). Engineering novel bioactive mini-proteins from small size natural and de novo designed scaffolds. *Curr. Protein Pept. Sci.* **1**, 403–430.
57. Schnolzer, M. & Kent, S. B. (1992). Constructing proteins by dovetailing unprotected synthetic peptides: backbone-engineered HIV protease. *Science*, **256**, 221–225.
58. Dawson, P. E., Muir, T. W., Clark-Lewis, I. & Kent, S. B. (1994). Synthesis of proteins by native chemical ligation. *Science*, **266**, 776–779.
59. Kochendoerfer, G. G. (2001). Chemical protein synthesis methods in drug discovery. *Curr. Opin. Drug Discov. Devel.* **4**, 205–214.
60. Kochendoerfer, G. G., Chen, S. Y., Mao, F., Cressman, S., Traviglia, S., Shao, H. *et al.* (2003). Design and chemical synthesis of a homogeneous polymer-modified erythropoiesis protein. *Science*, **299**, 884–887.
61. Vita, C., Roumestand, C., Toma, F. & Menez, A. (1995). Scorpion toxins as natural scaffolds for protein engineering. *Proc. Natl Acad. Sci. USA*, **92**, 6404–6408.
62. Kibler, D. & Hampson, S. (2002). Characterizing the *E. coli* shine-dalgarno site: probability matrices and weight matrices. *International Conference on Mathematical and Engineering Techniques in Medicine and Biological Science - METMBS 2002*. CSREA Press, Las Vegas, NV, USA.
63. Shine, J. & Dalgarno, L. (1974). The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl Acad. Sci. USA*, **71**, 1342–1346.
64. Lipfert, J., Millett, I. S., Seifert, S. & Doniach, S. (2006). Sample holder for small-angle X-ray scattering static and flow cell measurements. *Rev. of Sci. Instrum.* **77**, 046108-1–046108-3.
65. Glatter, O. & Kratky, O. (1982). *Small Angle X-ray Scattering*. Academic Press, London.
66. Doniach, S. (2001). Changes in biomolecular conformation seen by small angle X-ray scattering. *Chem. Rev.* **101**, 1763–1778.
67. Cohn, E. J. & Edsall, J. T. (1943). *Proteins, Amino Acids and Peptides as Ions and Dipolar Ions.* Reinhold Publishing Corporation, New York.
68. Laue, T. M. (1992). Short column sedimentation equilibrium analysis for characterization of macromolecules in solution. Spinco Business Unit, Palo Alto, CA.
69. 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.
70. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 277–293.
71. Sattler, M., Schleucher, J. & Griesinger, C. (1999). Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectr.* **34**, 93–158.
72. Guntert, P. (2003). Automated NMR protein structure calculation. *Prog. Nucl. Magn. Reson. Spectr.* **43**, 105–125.
73. Cornilescu, G., Delaglio, F. & Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, **13**, 289–302.
74. Sayle, R. A. & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374.
75. Farrow, N. A., Muhandiram, R., Singer, A. U., Pascal, S. M., Kay, C. M., Gish, G. *et al.* (1994). Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by ^{15}N NMR relaxation. *Biochemistry*, **33**, 5984–6003.
76. Deka, P., Rajan, P. K., Perez-Canadillas, J. M. & Varani, G. (2005). Protein and RNA dynamics play key roles in determining the specific recognition of GU_rich polyadenylation regulatory elements by human Cstf-64 protein. *J. Mol. Biol.* **347**, 719–733.
77. Boyd, J., Hommel, U. & Campbell, I. D. (1990). Influence of cross-correlation between dipolar and anisotropic chemical-shift relaxation mechanisms

upon longitudinal relaxation rates of N-15 in macromolecules. *Chem. Phys. Letters*, **175**, 477–482.
78. Otwinowski, Z. & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326.
79. McRee, D. E. (1999). A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156–165.
80. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W. *et al.* (1998). Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Crystallog. sect. D*, **54**, 905–921.
81. Mori, S., Abeygunawardana, C., Johnson, M. O. & van Zijl, P. C. (1995). Improved sensitivity of HSQC spectra of exchanging protons at short interscan delays using a new fast HSQC (FHSQC) detection scheme that avoids water saturation. *J. Magn. Reson. ser. B*, **108**, 94–98.
82. Kohn, J. E., Millett, I. S., Jacob, J., Zagrovic, B., Dillon, T. M., Cingel, N. *et al.* (2004). Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl Acad. Sci. USA*, **101**, 12491–12496.

Edited by F. Schmid

(Received 17 May 2006; received in revised form 21 July 2006; accepted 29 July 2006)
Available online 4 August 2006