In Silico Protein Evolution by Intelligent Design: Creating New and Improved Protein Structures

Gautam Dantas

A dissertation submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

University of Washington

2005

Program Authorised to Offer Degree: Biochemistry

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Gautam Dantas

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

David Baker

Reading Committee:

David Baker

Barry L. Stoddard

Gabriele Varani

Date: 29th July 2005

In presenting this dissertation in partial fulfilment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of the dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, An Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature 

Date 29th July 2005

University of Washington

Abstract

In Silico Protein Evolution by Intelligent Design:

Creating New and Improved Protein Structures

Gautam Dantas

Chair of the Supervisory Committee:
Professor David Baker
Department of Biochemistry

Natural proteins perform a startling diversity of biological functions, but comprise a miniscule fraction of the theoretical sequence-structure space that polypeptides might occupy. The goal of protein design is to identify new free-energy minima in this sequence-structure landscape so as to expand the functional repertoire of polypeptides beyond that observed in nature. The accurate design of new proteins requires an exacting understanding of the forces that govern protein structure and folding and should allow for the creation of novel molecular machines and therapeutics. This dissertation details the development of a computational protein design method, RosettaDesign, its application to design new and improved protein structures, and the rigorous experimental characterisation and analyses of the designed proteins to evaluate and improve the design process.

First, we applied RosettaDesign to computationally redesign the sequence of nine, natural, globular proteins. Experimental characterisation revealed that eight of the

redesigned proteins were folded with similar secondary structure to their wild-type counterparts, and six had stabilities equal to or up to 7 kcal / mol greater than the wild-type counterparts. High resolution structures of the two most dramatically stabilised redesigned proteins (human procarboxypeptidase and U1A) showed them to be virtually identical to the template natural counterparts.

Second, we extended the capabilities of RosettaDesign to create a protein topology not observed in nature, by iterating between sequence design and structure prediction. We applied this general computational strategy to create a 93-residue α/β protein called Top7 with a novel sequence and topology. We showed that the Top7 protein is folded and extremely stable, and the striking similarity between the x-ray crystal structure and the designed model demonstrated the unprecedented high-resolution accuracy of the design.

Third, we showed that the final 49 C-terminal residues of Top7 (named CFr) can be efficiently mistranslated in *E. coli*. While an overwhelming majority of naturally mistranslated polypeptides are unfolded, the CFr protein folds into an independently stable, obligate, symmetric homo-dimer, with a novel, high-affinity interface. We further stabilised CFr by disulfide-induced covalent circularisation to create an ideal scaffold for novel functional protein design.

TABLE OF CONTENTS

Page

LIST OF ABBREVIATIONS

CD	Circular dichroism
GuHCl	Guanidinium hydrochloride
Tris-HCl	Tris(hydroxymethyl)aminomethane hydrochloride
NaPi	Sodium phosphate
HEPES	4-(2-Hydroxyethyl)piperazine-1-ethanesulfonic acid
PEG	Poly-ethylene glycol
D_2O	Deuterium oxide
SDS-PAGE	Sodium dodecyl sulphate - polyacrylamide gel electrophoresis
NMR	Nuclear magnetic resonance
NOE(SY)	Nuclear overhauser effect (spectroscopy)
HSQC	Heteronuclear single-quantum coherence
TOCSY	Total correlation spectroscopy
SAD	Single-anomalous diffraction
ESI	Electrospray-ionization
MALDI-TOF	Matrix-assisted laser desorption ionization - time of flight
MS	Mass spectrometry
AUC	Analytical ultra-centrifugation
PDB	Protein data bank
RMSD	Root mean squared deviation

LIST OF FIGURES

Number Page

LIST OF TABLES

Number Page

ACKNOWLEDGMENTS

Chapter Two is reprinted with permission from *Journal of Molecular Biology*, and is written in collaboration with Brian Kuhlman, David Callender, Michelle Wong and David Baker (Dantas, Kuhlman et al. 2003). I would like to thank Lynne R. Spencer, Peter Brzovic, Ponni Rajagopol, Jennifer Keefe and Rachel Klevitt for aid in obtaining the NMR spectra.

Chapter Three is reprinted with permission from *Science*, and is written in collaboration with Brian Kuhlman, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard and David Baker (Kuhlman, Dantas et al. 2003). I would like to acknowledge the expert assistance of Betty Shen in crystallographic phasing, modelling and refinement of the TOP7 structure, Carol Rohl for aiding in the incorporation of RosettaDesign into Rosetta, Charlie Strauss for helping to generate the initial models of Top7, Tom Leeper for help with 2D NMR studies, the Klevitt laboratory for help with preliminary NMR studies, and the facilities at the Advanced Light Source (Berkeley, CA, supported by DOE) for access to their synchrotron-source x-ray beamlines.

Chapter Four is written in collaboration with Alexander L. Watters, Brad Lunde, Ziad Eletr, Nancy Isern, Brian Kuhlman, Barry L. Stoddard, Gabriele Varani and David Baker. I would like to 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, 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) for access to their synchrotron-source x-ray beamlines.

I would like to thank David Baker for his respectful trust in affording me immense freedom in the choice and execution of scientific projects and ideas. I feel privileged to have received such thoughtful guidance and expert mentorship from so brilliant a scientist.

I would like to thank Brian Kuhlman for developing RosettaDesign and for patiently introducing me to many concepts of computational biochemistry.

I would like to thank Barry Stoddard and Gabriel Varani for welcoming me into their labs and for sharing their stellar expertise in protein structure determination and analysis. I would also like to thank the many members of the Stoddard and Varani labs who selflessly shared their intellectual and instrumental resources with me.

I would like to thank the members of my PhD committee, David Baker, Gary Christian, Larry Loeb, Barry Stoddard, and Gabriele Varani, for their thoughtful and critical scientific advice throughout my PhD.

I would like to thank all the members of the Baker lab (2000-2005) for their friendship and for providing such a rich intellectual environment to work in. I would especially like to thank my fellow experimentalists, Michelle Scalley-Kim, Alex Watters, Sehat Nauli, David Callender, Colin Corrent, Lukasz Joachimiak, Vanita Sood, Jim Havranek, Eric "Alsdorf" Althof, Daniela Roethlisberger, and John Karanicolas. I would also like to specifically acknowledge the computational assistance of Brian Kuhlman, Jens Meiler, Phil Bradley, Jack Schonbrun, Chu Wang, and Lin Jiang.

I would like to thank my wife Laurel and the rest of my family in India for their unwavering encouragement and support of my academic endeavours, and for the immense sacrifices they have made to allow me to pursue those academic endeavours.

DEDICATION

To Laurel, Ammum, Daddy, Rohit, Poompa, Aai, and Baba.

CHAPTER 1

Engineering and Characterising New and Improved Protein Structures

Significant strides have been taken towards improving protein design methods in the last 10-15 years. While the specific goals of research groups developing these methods have been varied, the general drive behind protein design can be considered to be two-fold. Firstly, there is a tremendous value in being able to design proteins to fulfil particular functions, either new or improved, in a variety of environmental conditions. This value is particularly appreciated in the medical field where certain enzymes or hormones may need to be modified or newly created to aid in combating various disease states at the molecular level (DeGrado, Summa et al. 1999). Secondly, the results of protein design efforts can contribute a wealth of information towards understanding the basic forces that govern protein stability and folding. This understanding improves our ability to accurately predict the tertiary structure of proteins given their primary sequence, and can be used to better interpret the wealth of genomic information that we are amassing (Street and Mayo 1999). Consequently, an improved understanding of protein folding can further enhance our ability to rationally design proteins with new or improved functions for therapeutic purposes (Regan 1999).

Many attempts at protein design are focused on redesigning naturally occurring proteins with known structure. These studies attempt to characterize and understand the

functional and structural effects of changing all or part of the native sequence of the natural protein. On the other hand, the aim of true *de novo* protein design is to engineer a protein that will fold into an explicitly defined 3-dimensional structure, with a sequence that is not related to any known naturally occurring protein (DeGrado, Summa et al. 1999). While a variety of techniques have been employed for both types of design, particularly significant improvements have been made in the development of computational protein design methods in recent years.

Current computational techniques for protein design involve two key elements: the selection of an optimal solution from the tremendous range of sequence and structure possibilities, and the description of the various interactions in the protein in terms of an explicit mathematical energy function (Pokala and Handel 2001). A commonly used technique in protein design for greatly reducing the sequence-structure space, developed by Ponder and Richards (1987), is to i) assume a fixed protein backbone and to ii) restrict side chains to only adopt statistically preferred conformations or *rotamers*. This reduces the design problem to determining the lowest energy combination of side-chain rotamers for a fixed backbone template. However, the resultant sequence-structure search space is still too large to calculate the energy of every possible solution. For instance, allowing for three rotamers for each of the 20 amino acids at each sequence position, there are still more than 10^{177} sequence solutions for a 100 amino acid protein. A variety of search algorithms have therefore been employed to selectively sample these solutions to arrive at the lowest energy solutions. These algorithms (Desjarlais and Clarke 1998; Voigt,

Gordon et al. 2000) fall into two broad categories. Stochastic algorithms (e.g. Metropolis Monte Carlo), the faster of the two categories, work by semi-randomly sampling the search space to reach lower energy solutions, but suffer from the disadvantage of not being guaranteed to converge to the global minimum energy solution or even the same solution on different runs. In comparison, deterministic algorithms (e.g. Dead End Elimination), which do semi-exhaustive searches, always converge on the same (often global minimum) solution, but are fairly slow and get extremely complex and expensive in terms of search time as the search space grows (Pokala and Handel 2001). Once the search method has been chosen, the last major step in the computational protein design process involves the definition and appropriate parameterisation of a suitable energy function that models the various interactions in a 3-dimensional protein structure. Successful energy functions that have been employed are usually made up of a linear combination of terms that represent atomic van der Waals interactions, electrostatic interactions, hydrogen bonding interactions, and solvation energies (Bryson, Betz et al. 1995; Dahiyat and Mayo 1997). The contributions of each of these terms, as well as the inclusion of other optional constraints such as secondary structure propensities, have been varied and tested by many groups, but no absolute standards had surfaced at the beginning of this project. The parameterisation of these terms is often determined by tweaking them to improve the correlation between computed and experimentally determined properties of the sequences (Pokala and Handel 2001). It should be noted that while the fixed backbone assumption of Ponder and Richards (Ponder and Richards 1987) is effective in reducing the sequence-structure search space for rational protein

design, it can lead to rejection of sequences of lower energies that may adopt a very similar (and for many design purposes, effectively the same) fold (Pokala and Handel 2001). There are ample examples in the literature which show that natural proteins are quite tolerant to mutations that would be nonpermissable if the backbone was rigid because they can adjust to relieve strain (Alber, Bell et al. 1988; Baldwin, Hajiseyedjavadi et al. 1993; Lim, Hodel et al. 1994). Hence, explicit backbone flexibility during the design process is another significant parameter that can be varied in computational design to better model the natural environment of protein sequence-structure space.

In the late 90's, Street and Mayo (1999) formalised the concept of a "protein design cycle", which involves iteratively cycling between theory and experiment, towards developing a reliable technique for rational protein design. In the first step of the cycle, computational methods are used to predict sequences that will yield the most stable (lowest energy) structure for the intended protein fold. Next, the designed proteins are tested experimentally to evaluate how well the observed fold matches the predicted one, with particular attention given to protein stability and the mechanism of folding. The experimental data can then be fed back into the computational prediction program to either validate or correct the predictive power of the program. Further iterations of this cycle should lead to computational protein design methods that become progressively better in their ability to accurately predict experimental behaviour.

At the onset of the research described in this dissertation, our in-house computational protein design program, RosettaDesign (Kuhlman and Baker 2000), had been successfully applied to redesign protein folding pathways (Nauli, Kuhlman et al. 2001), backbone conformations (Kuhlman, O'Neill et al. 2002), and oligomerisation states (Kuhlman, O'Neill et al. 2001). The lessons learned from the experimental evaluations of these earlier design studies had been incorporated into the design protocol, and after these first successful rounds in the design cycle, we were eager to apply RosettaDesign to tackle some of the newer challenges in the protein design field. This dissertation describes some of these challenges, our attempts to apply and develop the RosettaDesign methodology to specifically address these different challenges, and the rigorous experimental characterisation and analyses of the designer proteins we created in each of these attempts.

The pioneering redesign of the 25 residue zinc finger Zif268 by Mayo and co-workers (Dahiyat and Mayo 1997) had demonstrated that automated procedures could be used to completely redesign a naturally occurring backbone. Despite the ground-breaking nature of that work, it stood as the only experimentally validated complete protein redesign attempt at the beginning of the new millennium. To extend these results to a larger and more diverse set of protein structures, we applied RosettaDesign to redesign nine globular proteins, and we then experimentally characterised these redesigned proteins with a battery of biophysical and structural techniques (Chapter 2) (Dantas, Kuhlman et al. 2003). The nine naturally occurring proteins had experimentally

determined three-dimensional structures, stabilities and folding mechanisms, were in the 80-120 residue length range, and represented folds with varying compositions of secondary structure. The general workings of RosettaDesign (2000) followed the paradigm for computational protein design described above – the program assumed a fixed protein backbone, restricted amino acid usage to a backbone-dependent rotamer library (1997), employed a Metropolis Monte Carlo (Metropolis, Rosenbluth et al. 1953) search procedure and an energy function composed of a linear combination of terms that model the forces that are important for protein structure and stability. Experimental characterisation of the redesign proteins demonstrated that RosettaDesign could reliably predict sequences that fold to stable structures, and that the redesigned proteins often have features typical of naturally occurring proteins. Eight of the proteins appeared folded, with similar secondary structure to their wild-type counterparts, and six had stabilities equal to or up to 7 kcal / mol greater than the wild-type counterparts. We have subsequently discovered (in collaboration with Ethan Merritt, and Gabriele Varani and co-workers) that the high-resolution structures of the two most dramatically stabilised redesigned proteins (redesigned human procarboxypeptidase and U1A) are essentially identical to their wild-type counterparts (< 1Å RMSD over backbone atoms). These encouraging results from the first large scale test of computational protein design suggested that we were ready to apply RosettaDesign to attack the next big challenge in the protein design field, the creation of proteins with novel structures.

The observation that a number of globular protein folds have not yet been observed in nature begs the question: are these folds not physically realisable or have they simply not yet been sampled by the evolutionary process or characterised by a structural biologist? Methods for *de novo* design of novel protein structures provide a route to resolving this question and perhaps, more importantly, a possible route to novel protein machines and therapeutics. At the turn of the 21st century, the landmark design by Harbury and colleagues of coiled-coil oligomers with a right handed superhelical twist was the only example of an experimentally validated attempt at computational *de novo* design of a novel protein fold (Harbury, Plecs et al. 1998). Since the three-dimensional backbone geometry (and hence flexibility) of coiled-coils can be explicitly represented with parametric equations, these authors were able to refine the backbone conformation for a large number of fixed amino acid sequences. Since this explicit geometric parameterisation is only possible for a limited number of internally-symmetric protein structures, a method that could be applied to the design of any general novel target structure was still needed. Accordingly, we developed a general computational strategy that iterates between sequence design and structure prediction, and applied it to design a 93-residue α/β protein called Top7 with a novel sequence and topology (Chapter 3) (Kuhlman, Dantas et al. 2003). We could utilise RosettaDesign, essentially as before, for the sequence optimisation steps. Since the goal of the backbone optimisation step is to identify the lowest free energy backbone conformation for a fixed amino acid sequence, it is formally analogous to the high-resolution structure prediction problem. The *ab initio* structure prediction module of our Rosetta computational algorithm was therefore

combined with RosettaDesign to allow for simultaneous optimisation of protein sequence and structure under the framework of the same energy function and search protocol. We biophysically and structurally characterised the Top7 protein, and found it to be folded and extremely stable, and the striking similarity between the x-ray crystal structure and the designed model (1.17 Å root mean squared deviation over backbone atoms) demonstrated the unprecedented high-resolution accuracy of our design.

During our experimental characterisation of the Top7 protein, we discovered that in addition to the production of full-length Top7, a portion of the original Top7 gene construct corresponding to the final 49 C-terminal residues is efficiently mis-translated in E. coli. Studies on translational error and degradation of natural proteins suggest a significant rate of aberrant translation, caused either by reactions that end prematurely or initiate inappropriately (Kurland 1992; Goldberg 2003), that result in an unfolded protein that is quickly targeted to the proteosome. The mistranslated progeny of Top7, however, appeared to form a folded protein with high stability, which could provide a significant challenge to normal cellular machinery. Since this is generally an unwanted outcome, elucidation of the cause of mistranslation and the biophysical and structural characterisation of the mistranslated fragment should allow us to learn how to control against such occurrences in future designs. Chapter 4 describes these efforts, and details the discovery that this C-terminal fragment of Top7 (CFr) forms a stable, obligate, symmetric homo-dimer with a novel protein-protein interaction with zeptomolar affinity. We use the structural information from this study to further stabilise the CFr protein by

engineering a intra-molecular disulfide bond that covalently circularises the protein. Our current efforts are focused on using these new *super-proteins* as scaffolds for the design of new functions and novel folds.

CHAPTER 2

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

A previously developed computer program for protein design, RosettaDesign, (Kuhlman and Baker 2000) was used to predict low free energy sequences for nine naturally occurring protein backbones. RosettaDesign had no knowledge of the naturally occurring sequences and on average 65% of the residues in the designed sequences differ from wild type. Synthetic genes for ten completely redesigned proteins were generated, and the proteins were expressed, purified, and then characterized using circular dichroism, chemical and temperature denaturation and NMR experiments. Although high resolution structures have not yet been determined, eight of these proteins appear to be folded and their circular dichroism spectra are similar to those of their wild type counterparts. Six of the proteins have stabilities equal to or up to 7kcal/mol greater than their wild type counterparts, and four of the proteins have NMR spectra consistent with a well-packed, rigid structure. These encouraging results indicate that computational protein design methods can, with significant reliability, identify amino acid sequences compatible with a target protein backbone.

Introduction

The ultimate goal of protein design is the creation of novel proteins that perform specified tasks. A necessary requirement for meeting this goal is the ability to identify sequences that fold with sufficient stability into a target structure. Towards this end several laboratories have developed computer programs for identifying amino acid sequences compatible with a given protein.(Desjarlais and Handel 1995; Dahiyat and Mayo 1997; Koehl and Levitt 1999; Wernisch, Hery et al. 2000; Looger and Hellinga 2001; Reina, Lacroix et al. 2002; Summa, Rosenblatt et al. 2002; Havranek and Harbury 2003) A rigorous test for these models is the complete redesign of naturally occurring proteins. In such a test, the only information given to the method is the backbone coordinates of the protein to be redesigned. Although there has been considerable recent success in the field of computational protein design,(Harbury, Plecs et al. 1998; Shimaoka, Shifman et al. 2000; Benson, Conrad et al. 2001; Offredi, Dubail et al. 2003) the pioneering zinc finger redesign by Mayo and coworkers is the only published report in which automated procedures have been used to completely redesign a naturally occurring protein backbone.(Dahiyat and Mayo 1997)

Previously we demonstrated that our method for protein design, RosettaDesign, produces native-like sequences when run on a large test set of naturally occurring protein backbones.(Kuhlman and Baker 2000) On average 30% of the residues were identical to their wild type counterpart, and in the core the level of identity was 50%. These results suggested that RosettaDesign was performing well, but they did not indicate whether the design sequences would actually fold into the target structures. The RosettaDesign

method has been applied successfully to the redesign of protein folding pathways,(Nauli, Kuhlman et al. 2001) backbone conformations,(Kuhlman, O'Neill et al. 2002) and oligomerisation states.(Kuhlman, O'Neill et al. 2001) Here, to more rigorously test RosettaDesign, and more generally, to assess the consistency with which modern computational protein design methodology can completely redesign the sequences of small proteins, we make and characterize complete redesigns of nine globular proteins.

Like all automated procedures for protein design, RosettaDesign has two main components: an energy function that ranks the relative fitness of various amino sequences for a given protein structure and a search function for rapidly scanning sequence space. The energy function used by RosettaDesign is dominated by Lennard-Jones interactions, an orientation dependent hydrogen bonding potential(Kortemme, Morozov et al. 2003), and an implicit solvation model.(Lazaridis and Karplus 1999) The Lennard-Jones term favours atoms being closely packed, but not too close to each other and therefore provides the steric information needed to correctly pack a protein core. The implicit solvation model penalizes the burial of polar atoms and therefore favours hydrophobic residues in the core of the protein and hydrophilic residues on the protein surface. The hydrogen bond term offsets the implicit solvation model by rewarding buried polar groups that form good hydrogen bonds. Amino acid specific reference energies approximate the average free energy of each of the amino acids in the denatured state.

Even for a small protein the size of sequence space is enormous and therefore it is not feasible to explicitly calculate the energy of every possible sequence. RosettaDesign uses a simple Monte Carlo optimization to identify low energy sequences. Starting from a completely random sequence, single amino acid substitutions are accepted or rejected using the Metropolis criterion. To make the search discrete, side chain conformations are restricted to the backbone torsion angle dependent rotamer conformations in Dunbrack's library.(Dunbrack and Cohen 1997) This optimization procedure converges to very similar sequences (70-80% identity) when multiple runs are started with different random sequences. Unlike the commonly used dead-end elimination algorithm, the Monte Carlo protocol does not guarantee that the final sequence will be at the global energy minimum, but the convergence observed from multiple runs strongly suggests that the search is not getting trapped in local minima. An advantage of the Monte Carlo protocol is that it is very fast, a typical search for a 100 residue protein takes approximately 5 minutes on a desktop computer.

Results

RosettaDesign was used to design sequences for nine globular proteins: the src SH3 domain, lambda repressor, U1A, protein L, tenascin, procarboxypeptidase, acylphosphatase, S6, and FKBP12 (Figure 2.1). For protein L two sequences were chosen for experimental study. On average the redesigned protein sequences are 35% identical to the wild type sequence over all residues and 50% identical for the core residues (Table 2.1). In general the overall amino acid composition in the redesigns is

similar to that of the wild type proteins, although a few of the redesigns are more hydrophobic than the wild type protein. The redesign of S6 has the most dramatic change, 59% of the residues are non-polar in the redesign while only 49% of the residues are non-polar in the wild type protein.

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

The folding and stability of each of the redesigned proteins was assessed using a battery of biophysical techniques. The extent of secondary structure in the completely redesigned proteins was assessed by circular dichroism spectroscopy (Figure 2.2). Size exclusion chromatography was used to determine if the proteins were monomeric (data not shown). Chemical (Figure 2.3) and thermal (Figure 2.4) denaturation experiments were used to confirm that the proteins were folded and to determine their stabilities. One-dimensional ^{1}H-NMR experiments (Figure 2.5) were used to further confirm that the proteins were folded and to probe the rigidity of their structures. Based on the results from these experiments we were able to place the proteins into different categories: unfolded versus folded, lower or higher than WT stability, and more or less rigid (Table 2.2).

Only one of the proteins, the SH3 redesign, is clearly unfolded. The CD spectra of redesigned SH3 (Figure 2.2) is typical of a random coil and the 1D ^{1}H-NMR spectrum (Figure 2.5) shows sharp lines and very little dispersion, strongly indicative of an unfolded protein.

Three of the proteins were multimeric even at low concentration. The tenascin redesign visibly aggregates at low concentrations and could not be further characterized. Size exclusion chromatography of the FKBP12 and S6 redesigns suggest they form oligomers, but the two proteins do not form extensive aggregates and their CD spectra are similar to their naturally occurring counterparts, suggesting that they may adopt the target structures. While the FKBP12 redesign denatured at high guanidine concentration, as evidenced by the change in the CD spectrum (Figure 2.2), the CD spectrum of redesigned S6 was remarkably resistant to both temperature and chemical denaturant (Figure 2.2); hence redesigned S6 may be stabilized by intermolecular as well as intra-molecular interactions. Clearly the computational design method, in addition to optimizing the stability of a given structure, needs to take into account solubility issues as well. This may perhaps be achieved by negative design against possible intermolecular interactions, for example by placing inwardly pointing charged amino acids in edge beta strands as suggested by the Richardsons.(Richardson and Richardson 2002)

The six remaining redesigned proteins appear monomeric and folded as evidenced by size exclusion chromatography, CD spectra, and chemical denaturation experiments. The proteins chromatographed as monomers by gel filtration chromatography, and comparison of their CD spectra (Figure 2.2) to previously published CD spectra of their naturally occurring counterparts suggested a very similar distribution of secondary structures. Chemical denaturation data fit well to a simple two state folding model (Figure 2.3), and for the designed proteins with buried tryptophan residues, unfolding transitions monitored by CD and by intrinsic fluorescence (data not shown) were coincident, further supporting the two state model typical for small naturally occurring proteins. The free energies of unfolding and their denaturant dependencies (m values) for the redesigned proteins were estimated from the fits of the chemical denaturation data (Figure 2.3) to the two state model. The m values of the designed proteins are in the range of those of naturally occurring small proteins; of the four cases where direct comparisons are possible, two of the designed proteins have smaller m values than their wild type counterparts, one has a larger value, and one a very similar value (Table 2.3). Of the six proteins, two—the redesigned lambda repressor and one of the two protein L redesigns (pL1)—are clearly less stable than their naturally occurring counterparts (Figure 2.3 and Table 2.3). The redesigned acylphosphatase and the second protein L redesign (pL2) have roughly the same stability as their naturally occurring counterparts (Table 2.3). In contrast, the U1A and procarboxypeptidase redesigns were significantly more stable than the naturally occurring proteins, redesigned U1A by ~2kcal/mol and redesigned procarboxypeptidase by a striking ~7kcal/mol.

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

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

Remarkably, four of the beta-sheet containing protein redesigns appear to be as rigid as most naturally occurring proteins. The two protein L redesigns and the redesigns of U1A and procarboxypeptidase have cooperative thermal melts (Figure 2.4) and NMR

spectra with relatively sharp lines and good dispersion (Figure 2.5). In addition, the NMR spectra for these proteins have small peaks just downfield of the water (5ppm – 5.5ppm) that are probably from Cα protons on the backbone and are strongly indicative of a β-sheet.(Wüthrich 1986)

Another hallmark of naturally occurring proteins is a large change in heat capacity upon folding. Two of the proteins, redesigned U1A and acylphosphatase, cold-denature at intermediate concentrations of guanidine and estimates of ΔC°_{p} could be obtained from fits of temperature denaturation experiments to the Gibbs-Hemholtz equation. For both proteins the ΔC°_{p} per residue is approximately 10 cal deg^{-1} mol^{-1} which falls within the range of ΔC°_{p} per residue values reported for natural proteins of this size.(Myers, Pace et al. 1995)

Discussion

Here we have shown that RosettaDesign can reliably predict sequences that fold to stable structures, and that the designed proteins often have features typical of naturally occurring proteins. Half of the folded designs have NMR spectra and temperature melts typical of tightly packed proteins. These findings significantly extend the pioneering successful complete redesign of the 25 residue zinc finger Zif268(Dahiyat and Mayo 1997) to a broad range of considerably larger proteins.

Since so many mutations were made to each protein it is difficult to determine why some designs were more successful than others, but there are some trends. Three redesigns were significantly more stable than their wild type versions (the redesigns of S6, U1A, and procarboxypeptidase) and two redesigns were less stable (one of the protein L redesigns and the redesign of lambda repressor). In each of the cases where the designs were more stable, the design sequence had a higher fraction of hydrophobic amino acids than the wild type protein. In the two cases where the design was less stable, the redesigned sequences were less hydrophobic than that of the wild type protein. These results are consistent with the notion that the burial of hydrophobic groups is one of the driving forces of protein folding.(Dill 1990) More detailed comparison of the wild type and redesigned proteins must await high resolution determination of the structures of the redesigned proteins.

Six of the ten design sequences were soluble and monomeric at NMR concentrations (1 mM) as judged by gel filtration, while one was unfolded. Why do the remaining three proteins self associate? Redesigned tenascin was visibly aggregated even at low concentrations, and therefore it was not possible to determine if it was folded. One possibility is that it aggregates to such a high degree because it is unfolded and therefore its hydrophobic core residues are exposed to interactions with other molecules. Redesigned S6 and FKBP12 do not visibly aggregate at low (CD) or high (NMR) concentrations, but are multimeric. This lower degree of association, when compared to redesigned tenascin, is probably due to association of folded monomers. Indeed, the

redesigns of S6 and FKBP12 have an increase in the fraction of non-polar accessible (i.e. "sticky") surface area compared to the wild type counterparts; this criterion could be used as a filter to ensure that future designs are soluble. To test this we are currently constructing variants of the S6 and FKBP12 redesigns that have fewer hydrophobic residues on their surface. Additionally, redesigned tenascin and FKBP12 seem to lack any of the "aggregation preventors" that native proteins employ with their edge beta strands, namely strand kinks or inward pointing charged residues.(Richardson and Richardson 2002) Since designing a kink in any strand would change the redesign backbone from its the native target, we are testing the feasibility of the second anti-aggregation strategy by constructing variants of redesigned tenascin and FKBP12 that replace edge-strand partially surface-exposed hydrophobic residues with charged residues.

In only one case, the redesign of the src SH3 domain, was the designed protein clearly unfolded. To examine why this designed sequence was so unstable, we used the program Probe(Word, Lovell et al. 1999) to look for clashes in our model of the designed protein. Probe identified a large clash between Ile 26 and Ala 39. An examination of the multiple sequence alignment for SH3 domains showed that these amino acids are often seen at these positions, but typically not together.(Larson, Di Nardo et al. 2000) There is a strong preference for Ile 26 to be paired with Gly 39, and Leu 26 to be paired with Ala 39. The atomic radii used in our simulations are scaled by 0.95 relative to CHARMM 19 radii in order to compensate for the use of fixed rotamers. If the radii are increased to

their full size, then RosettaDesign shows a strong preference for a Leu-Ala pair. Currently we are testing these findings by mutating Ile 26 to Leu or Ala 39 to Gly. This is a case where using reduced radii can be costly, and suggests the need for more realistic radii coupled with a better sampling of side chain conformational space.

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

Materials and Methods

Computational Procedure

Our computational model for protein design, RosettaDesign, is largely unchanged from that described previously.(Kuhlman and Baker 2000) RosettaDesign contains two main components: an energy function that ranks the relative fitness of amino sequences for a given protein structure and a Monte Carlo optimization procedure for rapidly searching sequence space. The energy function is a linear combination of a 12-6 Lennard-Jones potential, the Lararidis-Karplus implicit solvation model,(Lazaridis and Karplus 1999) an empirical hydrogen bonding potential,(Kortemme, Morozov et al. 2003) backbone dependent rotamer probabilities,(Dunbrack and Cohen 1997) amino acid probabilities for particular regions of phi,psi space, and a simple electrostatics term derived from the probability that two types of polar amino acids are found near each

other in the PDB.(Simons, Ruczinski et al. 1999) In addition, each amino acid has a unique reference energy that controls the frequency that it is placed during design. The linear sum of the seven energy terms, each with its own weight, can be represented as:

$$E_{protein} = W_{rot}E_{rot} + W_{atr}E_{atr} + W_{rep}E_{rep} + W_{solv}E_{solv} + W_{pair}E_{pair} + W_{hbond}E_{hbond} - E_{ref}$$

$$(1)\ E_{rot} = \sum_{i}^{nres} -\ln(P(rot(i) \mid phi(i), psi(i)) - \ln(\frac{P(aa(i) \mid phi(i), psi(i))}{P(aa(i))})$$

E_{rot} represents the internal energy of a rotamer/amino acid and is derived from the frequency of particular rotamers and amino acids for a given set of backbone torsion angles (phi angle, psi angle). The first term, the probability for a *rotamer* given phi and psi is taken directly from Dunbrack and Cohen,(Dunbrack and Cohen 1997) and the second term, the probability of an *amino acid* given phi and psi was computed from PDB statistics.

$$(2)\ E_{atr} = \sum_{i}^{natom}\sum_{j>i}^{natom}\left[\left(\frac{r_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{r_{ij}}{d_{ij}}\right)^{6}\right]e_{ij} \quad if \quad \frac{r_{ij}}{d_{ij}} < 1.12$$

$$(3)\ E_{rep} = \sum_{i}^{natom}\sum_{j>i}^{natom}\left(10.0 - 11.2\left(\frac{d_{ij}}{r_{ij}}\right)\right) \quad if \quad \frac{r_{ij}}{d_{ij}} > 1.12$$

E_{atr} is the attractive portion of a 12-6 Lennard-Jones potential. d_{ij} is the distance between the two atoms, r_{ij} is the sum of the van der Waals radii and e_{ij} the square root of the product of the well depths. The values for r_{ij} and e_{ij} are taken from the CHARMM19

parameter set,(Neria, Fischer et al. 1996) except that the r_{ij} values were scaled by 0.95 to account for the fixed rotamer approximation. Explicit hydrogens were only used to evaluate hydrogen bonds. E_{rep} represents the repulsive energy between two atoms and is dampened in comparison to the typical 12-6 potential because a fixed backbone and rotamer set is being used in this model. E_{atr} and E_{rep} are not evaluated between atoms in the same amino acid because these energies are already part of the E_{rot} term.

$$(4)\ E_{solv} = \sum_{i}^{natom} \sum_{j>i}^{natom} \left\{ \frac{-2\Delta G_i^{free}}{4\pi\sqrt{\pi}\lambda_i r_{ij}^2} \exp(-d_{ij}^2) V_j - \frac{2\Delta G_j^{free}}{4\pi\sqrt{\pi}\lambda_j r_{ij}^2} \exp(-d_{ji}^2) V_i \right\}$$

E_{solv} is the solvation energy of an atom calculated using the implicit solvent model developed by Lazaridis and Karplus.(Lazaridis and Karplus 1999) d_{ij} and r_{ij} are the same as in E_{atr}, ΔG^{free} is related to the solvation energy of the fully solvated atom, λ_i is a correlation length, and V is atomic volume. The values for the parameters are taken from Lazaridis and Karplus.(Lazaridis and Karplus 1999) We have left out the intrinsic solvation amino acid of each atom because it is completely determined by amino acid identity and can be incorporated into the reference energies.

$$(5)\ E_{pair} = \sum_{i}^{nres} \sum_{j>i}^{nres} \frac{P(aa_i, aa_j \mid d_{ij}, env_i, env_j)}{P(aa_i \mid d_{ij}, env_i) P(aa_j \mid d_{ij}, env_j)}$$

E_{pair} is derived from the probability of seeing two amino acids close together in space in the PDB database after accounting for the intrinsic probabilities of these amino acids to be in that environment. This term primarily reflects electrostatic effects.(Simons, Ruczinski et al. 1999) We only evaluated this term between polar amino acids.

E_{hbond} is the energy of sidechain-backbone and backbone-backbone hydrogen bonds. The hydrogen bonding potential was developed by Kortemme and co-workers and was derived by examining the hydrogen bond geometries in the protein data base.(Kortemme, Morozov et al. 2003) Additionally, the weight placed on the hydrogen bonding term was scaled by the number of neighbours surrounding the hydrogen bond. The exact values for the weights were determined by calculating the frequency that particular hydrogen bonding atoms on particular amino acids were found to form hydrogen bonds in naturally occurring proteins as a function of number of neighbours. For instance, if the side chain oxygen on serines with 15 neighbours were found in hydrogen bonds half the time in naturally occurring proteins than the energy for a serine hydrogen bond with 15 neighbours would be multiplied by 0.5. Neighbours were defined as residues with Cβ atoms within 10 Å. Side chain backbone hydrogen bonds were not allowed in cases where the backbone group was hydrogen bonding with another backbone atom.

Last, every amino acid has a reference energy, E_{ref}.

$$(7)\ E_{ref} = \sum_{i}^{nres} W_{ref}(aa(i))$$

The weights on these terms and the 20 reference energies were determined by maximizing the product of $\exp(-E(aa_{obs}))/(\Sigma\exp(-E(aa_i)))$ over a training set of 30 proteins using a conjugate-gradient-based optimization method, where $E(aa_{obs})$ is the energy of the native amino acid at a position and the partition function in the denominator is over all 20 amino acids at each position. In this process only one residue was changed at a time and all other residues were kept in their native conformation. Subsequently the parameters were refined slightly on the basis of the results of complete redesign calculations on the training-set proteins. The weights used for this study are given in Table 2.4. The reference values are dramatically different than we have used previously because we have removed the intrinsic solvation energy of an atom from the Lazaridis-Karplus solvation energy.

The Monte Carlo optimization procedure used to scan sequence space started with a random sequence. The side chain conformations of each amino acid were modelled using Dunbrack's backbone dependent rotamer library.(Dunbrack and Cohen 1997) Only rotamers observed more than 3% of the time were considered. Each round of Monte Carlo consisted of replacing one rotamer, evaluating the energy change, and accepting the change if it passed the Metropolis criterion. A rotamer replacement may or may not

involve changing amino acid identity. A typical run consisted of a few hundred thousand rotamer replacements, at which point the energy had typically plateaued.

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

Protein Expression and Purification

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

Circular Dichroism (CD)

CD data were collected on an Aviv 62A DS spectrometer. Far-UV CD wavelength scans (260-200nm) at varying protein concentrations (15-25μM), guanidinium hydrochloride (GuHCl) concentrations (0-8.3M), and temperatures (0-98°C) were collected in a 1mm path-length cuvette. GuHCl induced protein denaturation was followed by the change in ellipticity at 220nm in a 1cm path-length cuvette, using a Microlab titrator (Hamilton) for denaturant mixing. Temperature was maintained at 25°C with a Peltier device. All CD data were converted to mean residue ellipticity. Temperature induced protein denaturation was followed by the change in ellipticity at 220nm in a 2mm path-length cuvette. To obtain a value for ΔG_U^{H2O}, chemical denaturation curves were fit by nonlinear least squares analysis using the linear

extrapolation model as applied by Santoro and Bolen. To obtain a value for $\Delta C°_p$, thermal denaturation curves were fit using the Gibbs-Hemholtz equation in the form:

$$\phi = \phi_f + \frac{(\phi_u - \phi_f)}{1 + e^{\frac{-\Delta G°}{R \cdot T}}}$$

$$-\Delta G° = \Delta H°(1 - \frac{T}{T_m}) + \Delta C°_p\{T - T_m - T \cdot \ln(\frac{T}{T_m})\}$$

where ϕ is CD signal, ϕ_f and ϕ_u are the estimated CD signal for the folded and unfolded states, respectively, R is the gas constant, T is temperature, T_m is the temperature where 50% of the protein is folded, $\Delta G°$ is the change in the Gibbs free energy for the unfolding reaction, $\Delta H°$ is the change in enthalpy, and $\Delta C°_p$ is the change in heat capacity.

Size Exclusion (Gel Filtration) Chromatography

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

Nuclear Magnetic Resonance

One dimensional spectra were obtained on a Bruker AMX500 using water presaturation. Spectra were obtained at 27°C in 50 mM sodium phosphate pH 7. Protein concentrations were between 600 μM and 1.2 mM.

Solvent Accessible Surface Area

Solvent accessible surface area of non-polar atoms was calculated using the program Whatif (http://www.cmbi.kun.nl/gv/servers/WIWWWI/).

Table 2.1: Sequence alignments comparing the wild type sequences (WT) to the design sequences (D).

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

                70        80        90
ACY-WT      MQEWLETKGSPKSHIDRASFHNEKVIVKLDYTDFQIVK
ACY-D       FLEWLARSGSPNADIKQTVFTNMTRIDRLTMETFKIDE

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

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

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

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

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

                70        80
LMB-WT      AKILKVSVEEFSPSIAREIYEMYEAVS
LMB-D       AEIFNVSISDFSEYLYRLYLEQLKERF

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

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

                70        80        90
S6-WT       LWYQVEMPEDRVNDLARELRIRDNVRRVMVVKSQ
S6-D        IVIIVVGPPDTVEEFRRALQSLPYVLQVEIVPYE

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

                70        80
TEN-WT      LKPDTEYEVSLISRRGDMSSNPAKETFTT
TEN-D       LKPNTTFQITIRSQNGDKSSPPVSTYFTL

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

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

Table 2.2: Summary of experimental results used to characterize the redesigned proteins

Redesigned Proteins	CD Spectra	GuHCl Melt	Temp Melt	1D ^{1}H NMR	VERDICT
src SH3	Random-coil	Non-cooperative	Non-cooperative	Sharp-lines; No dispersion	**UNFOLDED**
Tenascin	β -sheet Like WT	Aggregated **U**nable **T**o **D**etermine (UTD)			**AGGREGATED**
λ-repressor	α-helical Like WT	Cooperative Stability < WT	Non-cooperative	Broad-lines; Weak dispersion	**DESTABILISED LESS RIGID**
acylphosphatase	α-β Like WT	Cooperative Stability = WT	Cooperative T_m > WT	Broad-lines; Strong dispersion	**STABLE LESS RIGID**
Immunophillin FKBP12	α-β Like WT	Cooperative Stability = WT	UTD	UTD	**STABLE MULTIMERIC**
ribosomal S6	α-β Like WT	Does not denature	UTD	UTD	**STABILISED MULTIMERIC**
protein L 1	α-β Like WT	Cooperative Stability < WT	Cooperative T_m > WT	Sharp lines; Strong dispersion	**DESTABILISED WELL-FOLDED**
protein L 2	α-β Like WT	Cooperative Stability = WT	Cooperative T_m > WT	Sharp lines; Strong dispersion	**STABLE WELL-FOLDED**
RNA-binding U1A	α-β Like WT	Cooperative Stability > WT	Cooperative T_m > WT	Sharp lines; Strong dispersion	**STABILISED WELL-FOLDED**
Procarb-oxypeptidase	α-β Like WT	Cooperative Stability > WT	Cooperative T_m > WT	Sharp lines; Strong dispersion	**STABILISED WELL-FOLDED**

Table 2.3: Thermodynamic stability of the designed and wild type proteins

Protein	ΔG_U^{H2O} (WT) / kcal mol^{-1}	ΔG_U^{H2O} (design) / kcal mol^{-1}	m-GuHCl (WT) / kcal mol^{-1} M^{-1}	m-GuHCl (design) / kcal mol^{-1} M^{-1}	Tm (WT) / °C	Tm (design) / °C
lambda repressor[1]	4.8	2.8	2.4	1.1	56	-
human U1A[2]	8.1	9.9	1.8	2.0	[]	> 100
Src SH3[3]	3.8	-	1.6	-	[]	-
Ribosomal S6[4]	11.6	-	[]	-	99	-
Acylphos-phatase[5]	4.8	5.3	[]	1.7	54	> 100
Procarboxy-peptidase[6]	4.1	11.9	[]	2.0	70-77	> 100
FKBP12[7]	4.6	4.8-7.1*	5.4	-	[]	-
Protein L (1)[8]	4.6	3.7	1.9	1.4	70	~ 100
Protein L (2)[8]	4.6	4.4	1.9	1.8	70	> 100

- unable to determine

[] not found in literature

* due to a strongly sloping "folded" baseline, slightly different baseline estimates yield significantly different ΔG estimates for redesigned FKBP12, with very similar fitting errors. This high variability may be due to the guanidine-induced solubilization of aggregates of this protein at low guanidine concentrations

[1] (Lim, Farruggio et al. 1992)

[2] (Kranz, Lu et al. 1996)

[3] (Grantcharova, Riddle et al. 1998)

[4] (Uversky, Abdullaev et al. 1999)

[5] (Taddei, Chiti et al. 1999)

[6] (Villegas, Azuaga et al. 1995)

[7] (Veeraraghavan, Holzman et al. 1996)

[8] (Scalley, Yi et al. 1997; Yi, Scalley et al. 1997)

Table 2.4: Weights and reference energies used for calculating protein energies (kcal mol^{-1}).

W_{atr}	1.02	W_{ref}(Lys)	1.30
W_{rep}	0.60	W_{ref}(Leu)	-1.62
W_{sol}	1.01	W_{ref}(Met)	0.25
W_{pair}	0.40	W_{ref}(Asn)	1.05
W_{hbond}	3.50	W_{ref}(Pro)	-0.22
W_{rot}	0.86	W_{ref}(Gln)	1.19
W_{ref}(Ala)	-0.03	W_{ref}(Arg)	2.10
W_{ref}(Asp)	2.30	W_{ref}(Ser)	0.92
W_{ref}(Glu)	2.26	W_{ref}(Thr)	0.38
W_{ref}(Phe)	-1.74	W_{ref}(Val)	-1.35
W_{ref}(Gly)	1.6	W_{ref}(Trp)	-2.19
W_{ref}(His)	-1.25	W_{ref}(Tyr)	-1.04
W_{ref}(Ile)	-2.15		

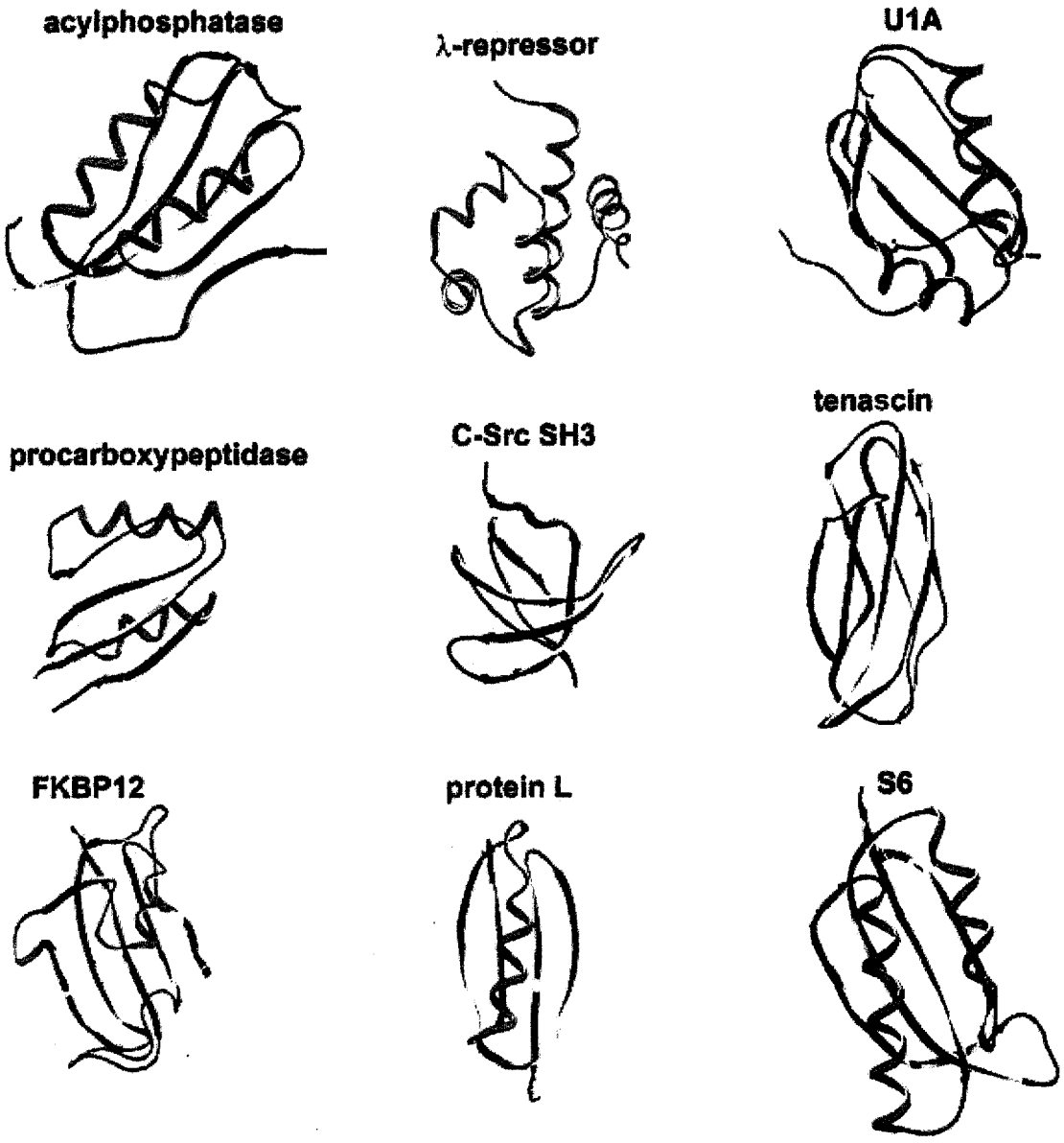


Figure 2.1 Ribbon diagrams of the nine redesigned structures. The PDB codes and respective residue numbers are: acylphosphatase (2acy, 1-98), lambda repressor (1lmb, 6-92), U1A (1urn, 2-97), procarboxypeptidase (1aye, 10-79), C-Src SH3 (1fmk, 83-142), tenascin (1ten, 803-890), FKB12 (1fkb, 1-107), protein L (1hz5, 1-62), S6 (1ris, 1-94).

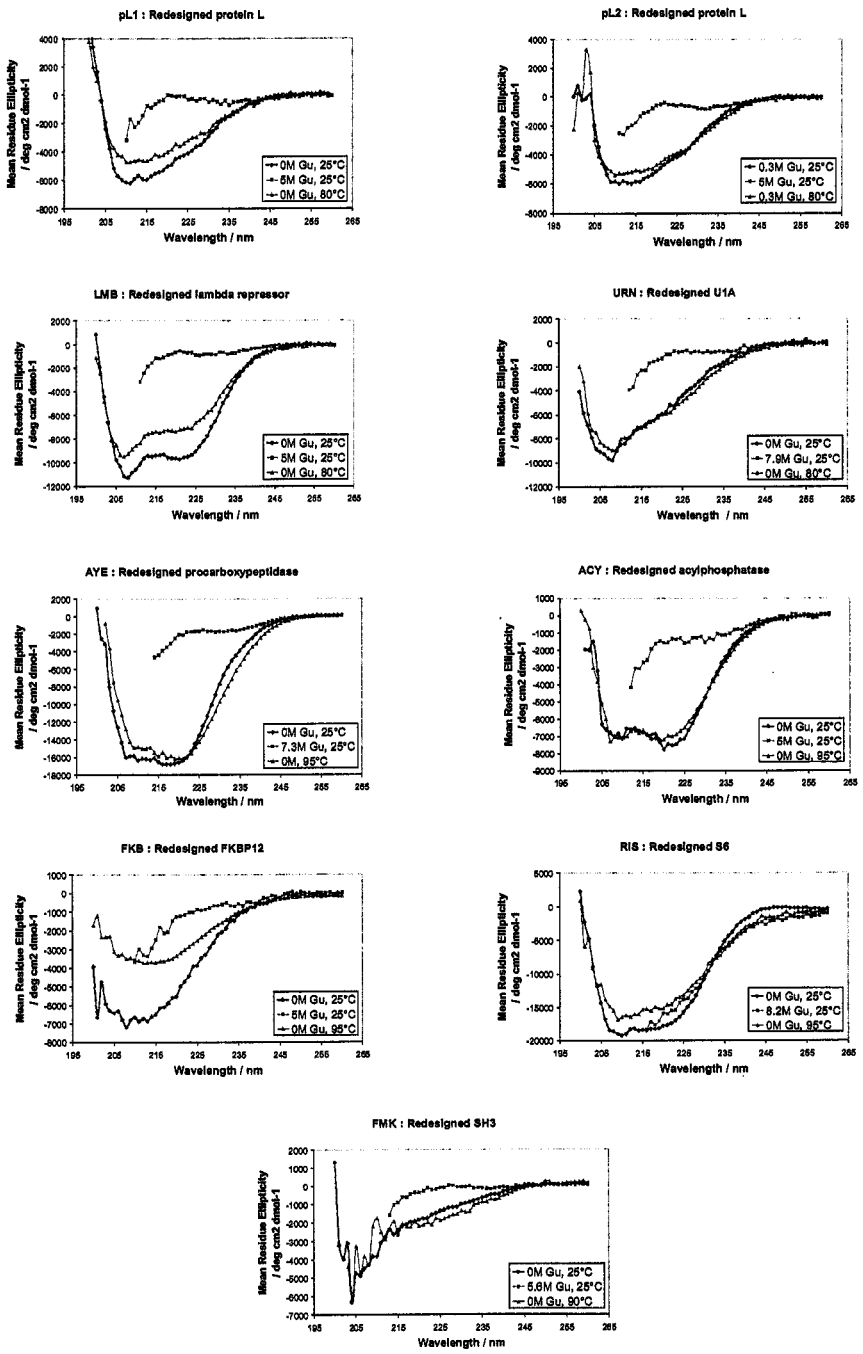


Figure 2.2 Circular dichroism spectra of the redesigned proteins. The CD spectra of eight of the redesigned proteins (pL1, pL2, LMB, URN, AYE, ACY, RIS) show the expected WT-like secondary structure content. The spectrum of redesigned src-SH3 resembles a random-coil. Far-UV CD spectra were collected on 15-25μM protein samples in 50mM sodium phosphate pH 7.0 at 25°C (Blue), at 95-98°C (Yellow) or in 5-8M GuHCl at 25°C (Pink).

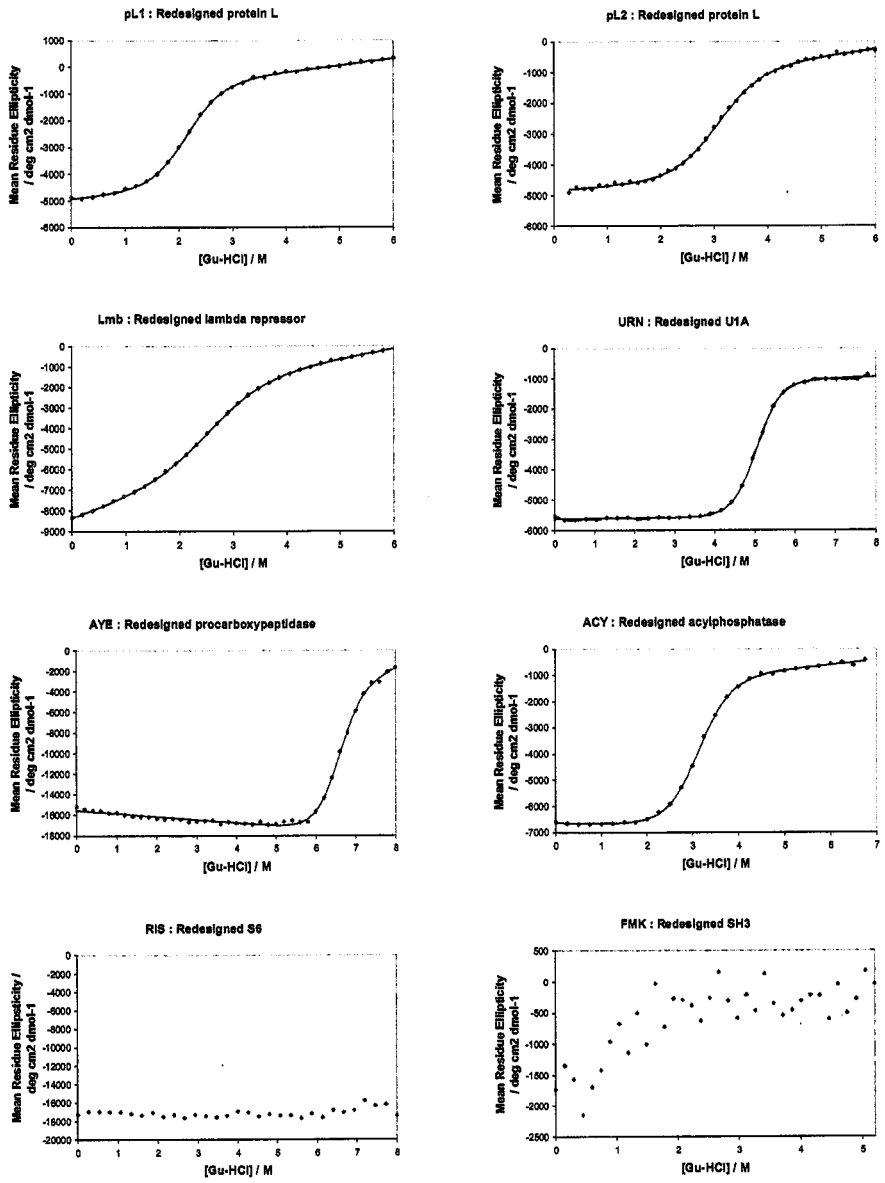


Figure 2.3 Chemical denaturation of the redesigned proteins. The GuHCl induced denaturation profiles of seven of the redesigned proteins (pL1, pL2, LMB, URN, AYE, ACY) are two-state and co-operative. Redesigned S6 does not denature at any GuHCl concentration. The erratic melt of redesigned SH3 suggests that the protein adopts a random coil structure. Ellipticity at 220nm was monitored as a function of GuHCl concentration for ~5μM protein in 50mM sodium phosphate, pH 7.0, 25°C, in a 1cm cuvette. The data were fit using a two-state model with a linear dependence of the free energy of unfolding (ΔG_U^{H2O}) on denaturant concentration. ΔG_U^{H2O} values are tabulated in Table 2.3. The data sets used are averages of duplicate experiments with 30 separate denaturant concentrations.

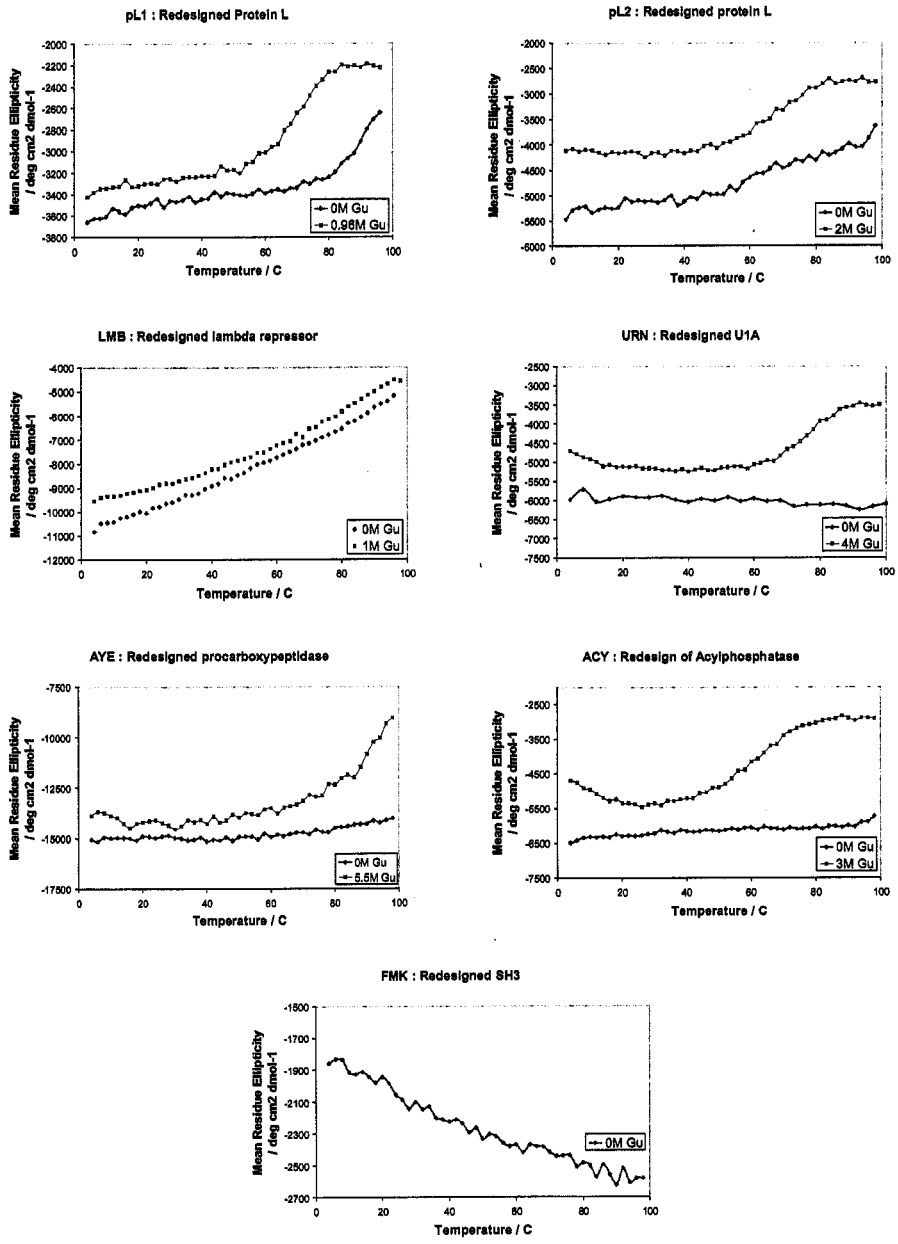


Figure 2.4 Thermal denaturation of the redesigned proteins. The temperature induced denaturation profiles of five of the redesigned proteins (pL1, pL2, URN, AYE, ACY) are two-state and co-operative. Redesigned lambda repressor exhibits a non-cooperative temperature melt, whereas redesigned SH3 exhibits non-cooperative cold denaturation. Ellipticity at 220nm was monitored as a function of temperature for ~10μM protein in 50mM sodium phosphate, pH 7.0 in a 2mm cuvette (blue curves). Pink curves are temperature melts performed for each protein at a GuHCl concentration where each protein was still folded at 25°C (as ascertained from Figure 2.3).

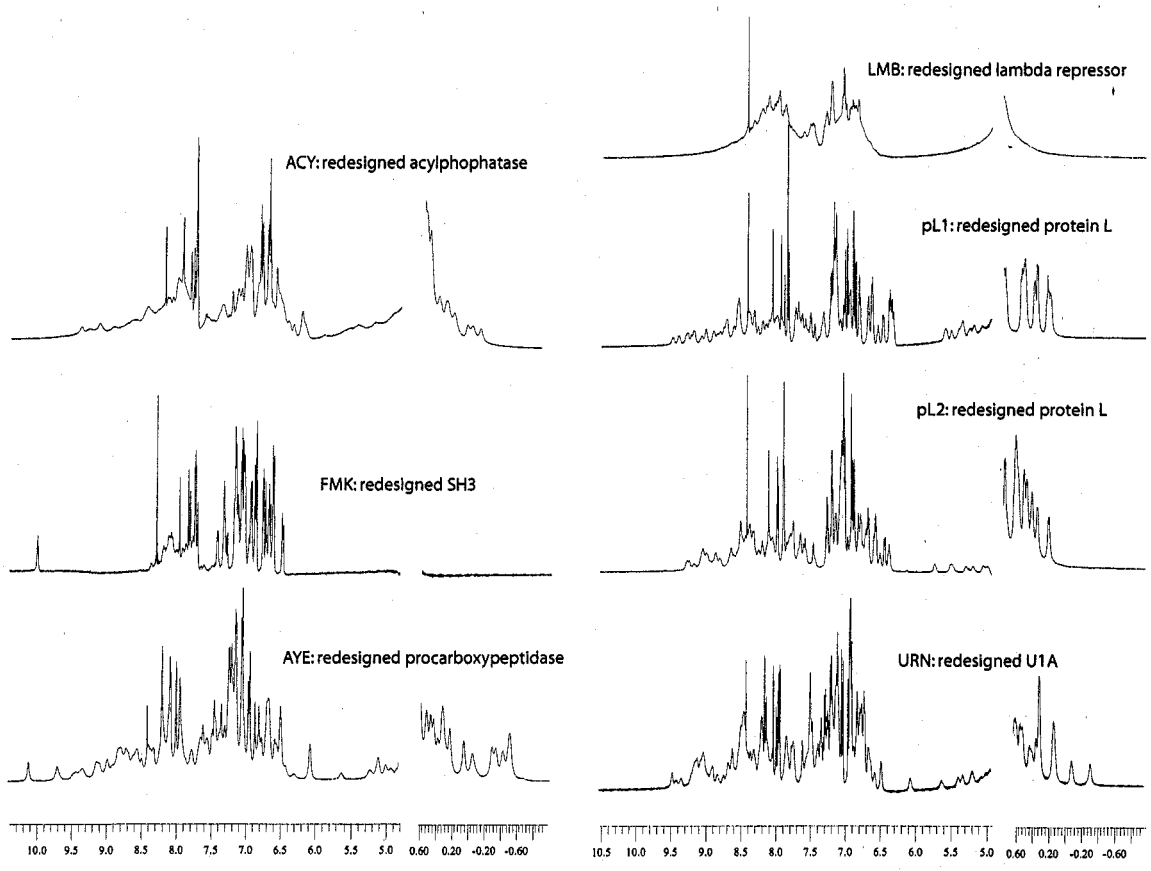


Figure 2.5 One dimensional ^{1}H-NMR spectra of the redesigned proteins. The sharp lines and strong dispersion in the spectra of pL1, pL2, URN and AYE suggest that these proteins are well folded in a unique conformation. Additionally, the peaks between 5 and 5.5 ppm suggest that these proteins have residues in a β-sheet, which is consistent with the target structure for these designs. Spectra were obtained at 27°C in 50 mM sodium phosphate pH 7. Protein concentrations were between 600 μM and 1.2 mM.

CHAPTER 3

Design of a Novel Globular Protein Fold with Atomic Level Accuracy

A major challenge of computational protein design is the creation of novel proteins with arbitrarily chosen three dimensional structures. Here we use a general computational strategy that iterates between sequence design and structure prediction to design a 93 residue α/β protein called Top7 with a novel sequence and topology. Top7 was found experimentally to be folded and extremely stable, and the X-ray crystal structure of Top7 is strikingly similar (RMSD = 1.2 Å) to the design model. The ability to design a new protein fold makes possible the exploration of the large regions of the protein universe not yet observed in nature.

Introduction

There are a large but finite number of protein folds observed thus far in nature, and it is not clear whether the structures not yet observed are physically unrealizable or have simply not yet been sampled by the evolutionary process or characterized by a structural biologist. Methods for de novo design of novel protein structures provide a route to resolving this question, and perhaps more importantly, a possible route to novel protein machines and therapeutics.

There has been considerable progress in the development of computational methods for identifying amino acid sequences compatible with a target structure (Ponder and Richards 1987; Desjarlais and Handel 1995; Dahiyat and Mayo 1997), notably the pioneering complete redesign of a zinc finger protein by Mayo and coworkers (Dahiyat and Mayo 1997). In general, these methods have not been used to create new protein structures, but rather have been used to redesign naturally occurring proteins so that they have enhanced stability or new functionality (Malakauskas and Mayo 1998; Reina, Lacroix et al. 2002; Looger, Dwyer et al. 2003). Because of the strong steric restrictions in the cores of globular proteins, the packing of side chains in redesigned proteins is often quite similar to that in the original native protein (Dahiyat and Mayo 1997; Johnson, Lazar et al. 1999), and hence high resolution protein backbone coordinates contain some memory of the original native sequence (Su and Mayo 1997; Koehl and Levitt 1999; Kuhlman and Baker 2000; Raha, Wollacott et al. 2000; Jaramillo, Wernisch et al. 2002). When creating a new protein from scratch there is no such sequence memory to aid the process, and it is not even known if the target backbone is designable. Thus the computational design of novel protein structures is a more rigorous test of current forcefields and optimization methodology than the redesign of naturally occurring proteins.

Because it is unlikely that any arbitrarily chosen protein backbone will be designable, it is essential that the design procedure include a search of nearby conformational space in addition to sequence space. With the exception of the method

used by Desjarlais and Handel (Desjarlais and Handel 1995) to redesign the hydrophobic core of a small naturally occurring protein, most previous approaches have either optimized the amino acid sequence for a large number of fixed backbone conformations (Su and Mayo 1997; Kuhlman, O'Neill et al. 2002; Larson, England et al. 2002; Reina, Lacroix et al. 2002) or, as in the landmark design by Harbury and colleagues of coiled coil oligomers with a right handed superhelical twist (Harbury, Plecs et al. 1998), refined the backbone conformation for a large number of fixed amino acid sequences (Harbury, Plecs et al. 1998; Keating, Malashkevich et al. 2001). The range of sequence-structure pairs that can be searched using these approaches is restricted by the need to specify, in advance, a limited number of backbone conformations or amino acid sequences.

We have developed a general procedure for identifying very low free energy sequence-structure pairs that iterates between sequence optimization and structure prediction and can be applied to the design of any desired target structure. The same energy function is used to guide the search at all stages, and at each stage only the lowest energy sequence or structure identified in the previous iteration is optimized, thereby avoiding the large scale and computationally expensive enumeration of alternative backbones or alternative sequences. Unlike the genetic algorithm of Desjarlais and Handel (Desjarlais and Handel 1995), in which randomly selected torsion angles and residue identities were simultaneously perturbed, our procedure iterates between full scale optimization of sequence for a fixed backbone conformation and gradient based optimization of the backbone coordinates for a fixed sequence. We use this approach to

create a 93 residue α/β protein with a topology not present in the Protein Structure Database (PDB).

Results

Generation of starting models

The target structure for the de novo design process can range from a detailed backbone model to a back of the envelope sketch. As we aimed to create a novel protein fold, we selected a topology cartoon for a fold not present in the PDB according to the TOPS server (http://www.tops.leeds.ac.uk/). A rough two-dimensional diagram was created of the target structure (Figure 3.1), and constraints were identified that define the topology (Figure 3.1, arrows). Three dimensional models satisfying the constraints were then generated by assembling three and nine residue fragments from the PDB with secondary structures consistent with the diagram using the Rosetta de novo structure prediction methodology (Bowers, Strauss et al. 2000), leading to 172 backbone only models that have the desired topology and secondary structure content and have RMS deviations from each other of 2-3Å.

Generation of starting sequences

A sequence was designed for each model using the RosettaDesign (Kuhlman and Baker 2000) Monte Carlo search protocol and energy function, which is dominated by a 12-6 Lennard-Jones potential, an orientation dependent hydrogen bonding term (Kortemme, Morozov et al. 2003), and an implicit solvation model (Lazaridis and

Karplus 1999). All amino acids except for cysteine were allowed at 71 of the 93 positions (~110 rotamers from Dunbrack's library (Dunbrack and Cohen 1997) per position); the remaining 22 surface beta sheet positions were restricted to polar amino acids (~75 rotamers per position). The search through the 11071 * 7522 ($>10^{186}$) rotamer combinations took ~10 minutes for each model on a Pentium III processor.

Because the starting backbone conformations were generated without regard to side chain packing, it was anticipated that sequences with very low free energies might not exist (i.e., the structures would not be designable). Indeed, the lowest energy sequences selected for the starting structures had energies considerably higher than native proteins of roughly the same size. In particular, the Lennard-Jones interaction energies for core residues were on average 0.8 kcal/mol less favourable than the interaction energies for the same residues in native protein cores. The finding that low energy sequences do not exist for protein backbones generated without regard to side chain packing emphasizes the need to couple sequence design with backbone flexibility for general protein design.

Simultaneous optimization of sequence and structure.

The critical feature of the design protocol is the cycling between sequence design, as described above, and backbone optimization. The goal of the backbone optimization step—to identify the lowest free energy backbone conformation for a fixed amino acid sequence—is formally analogous to the high resolution structure prediction problem and

we used the Rosetta program (Bonneau, Tsai et al. 2001) which we have developed for structure prediction. The backbone torsion angles were optimized using a Monte Carlo minimization protocol (Rohl, Straus et al. 2003) in which each move has the following parts: (1) An initial perturbation consisting of either a small random change in the torsion angles of 1-5 randomly selected residues, or substitution of the backbone torsion angles of one to three consecutive residues with torsion angles from a structure in the PDB. In the latter case the torsion angles of neighbouring residues were varied to minimize the displacement of the downstream portion of the chain. (2) A rapid optimization of side chain conformation for all residue positions that had a higher energy following step 1, by cycling through each rotamer at each position in turn, and replacing the current side chain conformation with the lowest energy rotamer conformation. (3) Optimization of the backbone torsion angles in a ten residue window surrounding the site of insertion by energy minimization using a quasi-Newton method (Press 1992). Moves were accepted or rejected based on the energy difference between the final minimized structure and the starting structure according to the Metropolis criterion. The same energy function was used for backbone optimization and sequence design. Each round of backbone relaxation consisted of several thousand such Monte Carlo minimization moves; a full combinatorial optimization of side chain rotamer conformations was carried out using a Monte Carlo procedure every twenty moves.

For each starting structure 5 independent simulations, each with 15 cycles of sequence design and backbone optimization, were used to obtain low energy structure

sequence pairs. Final energies were comparable to those observed for naturally occurring proteins. Proteins designed using an initial version of the protocol with a damped Lennard-Jones repulsive term and using Monte Carlo optimization without the minimization step were observed experimentally to be quite stable, but appeared to have somewhat molten cores (data not shown). To optimize steric packing, the atomic radii were reparameterized based on the distances of closest approach of atom pairs in high resolution protein structures, explicit protons were included on all atoms, the penalty for atom-atom overlaps was greatly steepened, and the full Monte Carlo minimization protocol was used for varying the backbone, resulting in the generation of much lower energy sequence structure pairs (20% of the final 860 models had more favourable Lennard-Jones energies than an average protein in the PDB). With these improvements, the protocol was used to design a protein sequence called Top7 (Figure 3.1).

The average Lennard-Jones energies for the buried residues in Top7 become favourable during the relaxation process (Table 3.1), and while the structural changes during the iterative refinement process are modest (the final protein backbone model has an RMSD of 1.1Å from the starting model), they bring about dramatic changes in the designed sequence: only 31% of the Top7 residues are identical to those in the starting sequence. Neither the Top7 sequence nor the sequence prior to the iterative sequence-structure refinement process have significant similarity to any naturally occurring protein sequence; the closest match to the Top7 sequence found using PSI-BLAST (Altschul,

Madden et al. 1997) in the protein sequence database (NR) is weaker than would be expected by random chance (E-value = 1.6).

Biophysical and structural characterization of Top7

The folding, stability, and structure of the Top7 protein were analyzed using a variety of biophysical methods. The Top7 protein is highly soluble (at 25-60mg/ml) and is monomeric as determined by gel filtration chromatography. The circular dichroism (CD) spectrum of Top7 is characteristic of α/β proteins (Figure 3.2A) and the protein is remarkably thermally stable: the CD spectrum at 98°C is virtually indistinguishable from the CD spectrum at 25°C. At intermediate concentrations of GuHCl (~ 5 M GuHCl), Top7 unfolds cooperatively with an increase in temperature and exhibits cold denaturation (Figure 3.2B). Fitting these data to the Gibbs-Helmholtz equation gave a change in heat capacity ($\Delta C°p$) per residue associated with unfolding of approximately 10 cal deg^{-1} mol^{-1}, a typical value for well-folded proteins of this size (Myers, Pace et al. 1995), The GuHCl induced chemical denaturation of Top7 is cooperative and the steep transition is characteristic of the two-state unfolding expected for a small, monomeric, single-domain protein (Figure 3.2C). Fitting the chemical denaturation data to a two-state unfolding model yields a free energy of unfolding of 13.2 kcal mol^{-1} at 25°C, indicating that Top7 is more stable than most proteins of similar size (Plaxco, Simons et al. 1998). The NOESY and HSQC spectra of Top7 (Figure 3.2D-E) exhibit features characteristic of a folded protein with significant β-sheet content. The HSQC spectrum contains the expected number of cross-peaks and the dispersion is comparable to that of

α/β proteins of similar size. Strong backbone NH-Hα cross peaks and the observation of Hα resonances downfield of the water signal (to 6 ppm) indicate the presence of a β-sheet, while NH-NH peaks are consistent with a partial helical character for the protein. Crystallization trials with Top7 yielded crystals that diffracted to 2.5 Å. Rather remarkably, a strong molecular replacement (MR) solution to the phase problem was found using the design model. This suggested immediately that the design model was quite close to the true structure—even the small deviations of NMR solution structures from X-ray crystal structures can make molecular replacement searches fail. To obtain unbiased phase information, a selenomethionyl substituted variant of Top7 with a surface lysine at position 37 mutated to methionine was produced and crystallized, and the X-ray crystal structure was solved to 2.5Å by direct rebuilding into an unbiased single-wavelength anomalous difference (SAD) electron density map (Figure 3.3B) and residual difference Fourier maps. The final R_{work} and R_{free} were 0.268 and 0.293 respectively (X-ray data and refinement statistics are provided in Table 3.2).

The high resolution crystal structure reveals that the Top7 protein adopts the designed topology (Figure 3.4A). Indeed, the structure is strikingly similar to the design model at atomic resolution (1.17Å RMSD over all backbone atoms). The overall protein structure is very well ordered, with the exception of two turns (comprising residues 11 to 15 and 24 to 31), each of which exhibit elevated B-factors and poor quality electron density. The first of these two turns, and the immediately adjoining residues from its neighbouring strand, deviate the most significantly from the computational model. However, even in

this region the all-atom RMSD between the two models does not exceed 2.8Å. In contrast, the C-terminal half of the X-ray structure is well ordered and very similar to the computational model; for example the region Asn78-Gly85 has an all-atom RMSD of 0.79Å (Figure 3.4B). Many side-chains in the core of the solved structure are effectively superposable with those of the designed Top7 (Figure 3.4C).

Like the design model, the Top7 crystal structure is judged to be a novel topology by the TOPS server. The strongest structural similarity found in a Dali search of the protein databank (Holm and Sander 1995) is to a discontinuous portion of the 668 residue protein S-adenosylmethionine decarboxylase, which has a large 68 residue insertion between strands 1 and 2, and the third and fourth strands are connected by an unrefined loop instead of a helix. According to Alexey Murzin, the curator of the SCOP database, the Top7 fold is not present in SCOP (Hubbard, Murzin et al. 1997; Murzin 2003).

Discussion

The 1.17Å backbone atom RMSD between the Top7 design model and the crystal structure implies that deep minima in the free energy function used in design correspond to deep minima in the actual free energy landscape and hence is an important validation of the accuracy of current potential functions. This atomic level accuracy contrasts sharply with the low accuracy of *ab initio* structure predictions for naturally occurring sequences—the most accurate structure predictions in the Critical Assessment of Structure Prediction (CASP) experiments for 90-100 residue proteins have RMSDs

greater than 4Å from the experimentally determined structure. Why does the simultaneous optimization of sequence and structure identify the global free energy minimum while the optimization of structure for fixed sequence does not? The answer may involve both of the challenges facing *ab initio* structure prediction—the vast size and ruggedness of the conformational space to be searched, and the limited accuracy of current potential functions. The capability to alter the sequence and hence reconfigure the landscape may greatly facilitate the search for low free energy protein structures compared to standard *ab initio* prediction where the sequence is fixed. In addition, Top7 lacks functional constraints which can lead to locally suboptimal regions in native structures that are particularly challenging for structure prediction, and the more extensive optimization of the folding free energy may partially compensate for inaccuracies in the potential functions. Finally, it should be noted that the design process focused entirely on minimizing the free energy of the folded monomeric structure—attaining a highly stable new structure did not require extensive negative design against possible alternative conformations (Havranek and Harbury 2003; Jin, Kambara et al. 2003) nor consideration of the kinetic process of protein folding (Mirny and Shakhnovich 2001).

The design of Top7 shows that globular protein folds not yet observed in nature not only are physically possible but can be extremely stable. This extends the earlier observation that helical coiled coil geometries not found in nature can be generated by computational protein design (Harbury, Plecs et al. 1998). The protein therapeutics and

molecular machines of the future should thus not be limited to the structures sampled by the biological evolutionary process. The methods used to design Top7 are in principle applicable to any globular protein structure and open the door to the exploration and utilization of a vast new world of protein structures and architectures.

Materials and Methods

Protein Expression and Purification

Synthetic genes which place the computationally selected protein sequences under the control of the T7 promoter, with a C terminal 6X His tag, and a codon usage optimal for *Escherichia coli* (*E. coli*) were obtained from BlueHeron Biotechnologies. The gene constructs were cloned in plasmid pet29b(+) (Novagen) and expressed in the BL21(DE3)pLysS strain of *E. Coli.* Cells were grown in LB media at 37°C to an OD_{600} of 0.6, induced with 1mM isopropyl-thio-β-D-galactosidase (IPTG), and cells were harvested after another 5 hours of growth at 37°C. Harvested cells were lysed by three freeze-thaw cycles, and soluble protein collected after centrifugation of cellular debris. Soluble protein was purified on a Ni^{+} affinity column (Pharmacia Biotech) followed by10^4-fold dialysis against 25mM TRIS-HCl, 30mM NaCl, pH 8.0. Protein was further purified on a QFF anion exchange column (Pharmacia) with a 30mM to 500mM NaCl gradient in 25mM TRIS-HCl, pH 8.0, followed by a final 10^4-fold dialysis against 25mM TRIS-HCl, 30mM NaCl, pH 8.0. Protein identity and purity was determined by SDS-PAGE and ESI-MALDI Mass Spectroscopy. Protein concentrations were determined by

UV absorbance at 280nm with extinction coefficients calculated using the ExPASy Protparam tool (http://us.expasy.org/tools/protparam.html).

The following modifications were made to the above procedure for Top7 crystallography. A Lys^{37} to Met^{37} point mutant of Top7 (Top7_K37M) was generated using the Single Quikchange Mutagenesis kit (Stratagene). Selenomethionine containing Top7_K37M was expressed in minimal media from the *E. coli* strain BL21(DE3) adapted for growth with methionine pathway inhibition (Doublie 1997). Cells were grown in minimal media at 37°C to an OD_{600} of 0.8 and the following amino acids were added to inhibit the methionine biosynthetic pathway: 100 mg/L lysine, threonine, phenylalanine; 75 mg/L selenomethionine; 50mg/L leucine, isoleucine, valine. Following a 15-minute incubation at 37°C, IPTG was added to induce expression and the cultures were harvested after 5 hours of growth at 37°C. Purification was performed as described.

^{15}N-labelled Top7 was prepared by expression in M9 minimal media with ^{15}N-labelled NH_4Cl. Purification was performed as described for unlabelled protein.

Circular Dichroism (CD)

CD data were collected on an Aviv 62A DS spectrometer. Far-UV CD wavelength scans (260-200nm) at varying protein concentrations (15-25μM), guanidinium hydrochloride (GuHCl) concentrations (0-8.3M), and temperatures (0-98°C) were collected in a 1mm path-length cuvette. GuHCl induced protein denaturation was

followed by the change in ellipticity at 220nm in a 1cm path-length cuvette, using a Microlab titrator (Hamilton) for denaturant mixing. Temperature was maintained at 25°C with a Peltier device. All CD data were converted to mean residue ellipticity. Temperature induced protein denaturation was followed by the change in ellipticity at 220nm in a 2mm path-length cuvette. To obtain a value for ΔG_U^{H2O}, chemical denaturation curves were fit by nonlinear least squares analysis using the linear extrapolation model as applied by Santoro and Bolen. To obtain a value for $\Delta C°_p$, thermal denaturation curves were fit using the Gibbs-Hemholtz equation in the form:

$$\phi = \phi_f + \frac{(\phi_u - \phi_f)}{1 + e^{\frac{-\Delta G°}{R \cdot T}}}$$

$$-\Delta G° = \Delta H°(1 - \frac{T}{T_m}) + \Delta C°_p\{T - T_m - T \cdot \ln(\frac{T}{T_m})\}$$

where ϕ is CD signal, ϕ_f and ϕ_u are the estimated CD signal for the folded and unfolded states, respectively, R is the gas constant, T is temperature, T_m is the temperature where 50% of the protein is folded, $\Delta G°$ is the change in the Gibbs free energy for the unfolding reaction, $\Delta H°$ is the change in enthalpy, and $\Delta C°_p$ is the change in heat capacity.

Nuclear Magnetic Resonance (NMR)

The 2D NOESY spectrum of ~1mM Top7 25mM sodium phosphate pH 6.0 was recorded at 298K at 500Mhz and 200ms mixing time using Watergate suppression. The

2D HSQC spectrum of ~1mM 15N-labelled Top7 25mM sodium phosphate pH 6.0 was recorded at 298K at 500Mhz using the fast HSQC scheme of Mori et al. (Mori, Abeygunawardana et al. 1995)

Crystallization

Selenomethionyl substituted Top7_K37M was crystallized in hanging drops (1 µl of protein solution at 25 mg/ml with 1 µl of well solution). The well solutions ranged from 15 - 20% PEG 3350 and 250 mM ammonium formate pH 6.6. The protein crystals grew within a day and were between 50-200 µm on a side. They were initially transferred to a cryo-solution of well solution at 25% PEG 3350 plus 25 % (v/v) glycerol in 4 steps of increasing glycerol and flash frozen in liquid nitrogen. With this treatment the crystals diffracted in a trigonal space group ($P3_221$) with unit cell dimensions a = 35.9 Å, b = 35.9 Å, c = 140.6 Å. A single wavelength (0.9793 Å) anomalous dispersion (SAD) (Hendrickson 1991) data set was collected to 2.5 Å resolution on beam-line 8.2.1 at the ALS (Advanced Light Source, Lawrence Berkeley Laboratory, Berkeley) using a four panel ADSC CCD area detector. Multiwavelength data collection (MAD phasing) was not possible due to significant radiation decay. Data were processed and scaled using HKL2000 (Otwinowski and Minor 1997).

Structure Determination

The structure of Top7_K35M was solved by molecular replacement with the program EPMR (Kissinger and Gehlhaar 1997), and by direct rebuilding into an unbiased

SAD electron density map and residual difference Fourier maps. For molecular replacement, 19 surface large surface residues such as Lys, Arg, and Glx were truncated to Ala in the search model. The correlation coefficient for the initial MR search, using data to 4.0 Å resolution, was 0.52, vs. background of 0.36. For SAD phasing, the position of SeMet 37 was determined from an anomalous difference Patterson map. The initial phasing power and figure of merit for SAD phasing was 1.99 and 0.24 prior to density modification. An interpretable electron density map was obtained after density modification with solvent flipping with a solvent content of 43 % (CNS). An initial model was built using XtalView (McRee 1999) and O (Jones, Zou et al. 1991). The model was refined with CNS using the mlhl target (maximum likelihood, Hendrickson-Lattman coefficients) with 5% of the data excluded for the calculation of the cross-validating free R (Keywegt and Jones 1996). 88% of all the built residues are in the most favourable regions of Ramachandran space and 12% are in the allowed regions (Laskowski, MacArthur et al. 1993). Statistics from phasing and refinement are shown in Table 3.2. The structure has been deposited in the PDB with the accession code 1QYS. Examples of the experimental electron density map were generated with XtalView and Raster 3D (Merritt and Bacon 1997). Ribbon diagrams were generated with SwissPDB Viewer (Guex and Peitsch 1997).

Energy Function

The energy of a protein was computed as a linear sum of the following 11 energy terms.

$$E_{protein} = W_{rot}E_{rot} + W_{ad\_phi,psi}E_{ad\_phi,psi} + W_{rama}E_{rama} + W_{atr}E_{atr} + W_{solv}E_{solv} + W_{pair}E_{pair} + W_{bb\_hbond}E_{bb\_hbond} + W_{sc\_hbond}E_{sc\_hbond} + W_{sc\_bb\_hbond}E_{sc\_bb\_hbond} + W_{pair}E_{pair} - E_{ref}$$

The weights (W) for each term are given in Table 3.4. To calculate the solvation energy (E_{solv}) and the Lennard-Jones energies (E_{atr} and E_{rep}) the various atoms of the 20 amino acids were binned into types (Table 3.3).

Lennard-Jones Potential (E_{atr} and E_{rep})

A standard 12-6 Lennard-Jones potential is used except there is cut-off distance below which the potential is extrapolated linearly. Favourable energies are placed in E_{atr} and unfavourable energies are placed in E_{rep}.

$$E_{atr} = \sum_{i}^{natom}\sum_{j>i}^{natom}\left[\left(\frac{r_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{r_{ij}}{d_{ij}}\right)^{6}\right]e_{ij} \quad if \quad \frac{r_{ij}}{d_{ij}} < 1.12$$

$$E_{rep} = \sum_{i}^{natom}\sum_{j>i}^{natom}\left[\left(\frac{r_{ij}}{d_{ij}}\right)^{12} - 2\left(\frac{r_{ij}}{d_{ij}}\right)^{6}\right]e_{ij} \quad if \quad 1.33 > \frac{r_{ij}}{d_{ij}} > 1.12$$

$$+ \sum_{i}^{natom}\sum_{j>i}^{natom} y_{\text{int}ercept} - d_{ij} * slope \quad if \quad \frac{r_{ij}}{d_{ij}} > 1.33$$

$$slope = -12e_{ij}\left(1.33^{13} - 1.33^{7}\right) * \left(1/r_{ij}\right)$$

$$y_{\text{int}ercept} = -slope * \left(\frac{r_{ij}}{1.33}\right) + e_{ij}\left(1.33^{12} - 2(1.33)^{6}\right)$$

$$r_{ij} = r_i + r_j$$

$$e_{ij} = \sqrt{e_i e_j}$$

Lazaridis-Karplus solvation model (E_{solv})

An implicit solvation model developed by Lazaridis and Karplus is used to evaluate the solvation energy of a protein (Lazaridis and Karplus 1999).

$$E_{solv} = \sum_{i}^{natom} \sum_{j>i}^{natom} \left\{ \frac{-2\Delta G_i^{free}}{4\pi\sqrt{\pi}\lambda_i r_{ij}^2} \exp(-d_{ij}^2) V_j - \frac{2\Delta G_j^{free}}{4\pi\sqrt{\pi}\lambda_j r_{ij}^2} \exp(-d_{ji}^2) V_i \right\}$$

d_{ij} and r_{ij} are the same as in E_{atr}, ΔG^{free} is related to the solvation energy of the fully solvated atom, λ_i is a correlation length, and V is atomic volume. The values for the parameters are taken from Lazaridis and Karplus, except some of the ΔG^{free} values have been perturbed to better reproduce the relative frequencies amino acids are placed in the core versus the surface during design experiments (Table 3.5). We have left out the intrinsic solvation energy of each atom because the sum of these values is a constant for each amino acid and can be incorporated into the reference energies.

Rotamer Self-energy (E_{rot})

$$E_{rot} = \sum_{i}^{nres} -ln(P(rot(i) \mid phi(i), psi(i))$$

E_{rot} is related to the self-energy of a rotamer and was derived from Protein Data Bank statistics by observing the probability of a particular rotamer and amino acid for a

given phi angle and psi angle. These probabilities were taken directly from Dunbrack and Cohen (Dunbrack and Cohen 1997). During the final design simulations we also considered rotamers with chi angles perturbed from the most commonly observed chi angles (+/- 0.5 standard deviation). These sub-rotamers were penalized by assuming a gaussian distribution about the mean using tabulated variances from Dunbrack and Cohen.

Amino acid preferences for particular regions of phi, psi space ($E_{aa|phi,psi}$)

A non-redundant set of PDB files were used to determine the probabilities for observing each of the 20 amino acids within 10°x10° bins in phi,psi space, P(aa,|phi,psi). The energy was calculated by taking the negative log of the probabilities.

Amino acid dependent torsion potential for phi and psi (E_{rama})

For each of the 20 amino-acid types in each of three secondary structure types (helix, strand, and other as defined by DSSP), the frequency of (phi,psi) pairs was determined for 10°x10° bins. Probabilities were calculated using added pseudocounts, and the potential calculated by taking the log of the interpolated probabilities.

Residue pair potential (E_{pair})

$$E_{pair} = \sum_{i}^{nres} \sum_{j>i}^{nres} \frac{P(aa_i, aa_j \mid d_{ij}, env_i, env_j)}{P(aa_i \mid d_{ij}, env_i) P(aa_j \mid d_{ij}, env_j)}$$

E_{pair} is derived from the probability of seeing two amino acids close together in space in the PDB database after accounting for the intrinsic probabilities of these amino acids to be in that environment (Simons, Ruczinski et al. 1999). Two classes of environments are considered, buried and exposed, and five distance bins were used, 0-4.5, 4.5-6.0, 6.0-7.5, 7.5-9.0 and 9.0-10.5. This term was only evaluated between polar amino acids. The distances were measured between the action centres on each residue, e.g. the nitrogen on the lysine sidechain.

Orientation-dependent hydrogen bonding term (E_{bb_hbond}, E_{sc_hbond}, $E_{bb\_sc\_hbond}$)

The energy of backbone-backbone, sidechain-backbone and sidechain-sidechain hydrogen bonds were determined using a function derived from the distances and angles observed for naturally occurring hydrogen bonds in the PDB database. This function is described in detail in the supporting material of Kortemme & Baker (Kortemme, Morozov et al. 2003). In this study we did not weight the strength of the hydrogen bonds according to their degree of burial. We removed this weight to encourage hydrogen bonds at positions that are partially buried.

Energy of the unfolded state (E_{ref})

$$E_{ref} = \sum_{i}^{nres} W_{ref}(aa(i))$$

To approximate the energy of the unfolded state each amino acid is assigned a empirically determined reference energy.

Setting the weights

The weights on these terms and the 20 reference energies were determined by maximizing the product of $\exp(-E(aa_{obs}))/(\Sigma\exp(-E(aa_i)))$ over a training set of 30 proteins using a conjugate-gradient-based optimization method, where $E(aa_{obs})$ is the energy of the native amino acid at a position and the partition function in the denominator is over all 20 amino acids at each position. In this process only one residue was changed at a time and all other residues were kept in their native conformation. Subsequently the parameters were refined slightly on the basis of the results of complete redesign calculations on the training-set proteins. The weights used for this study are given in Table 3.6. The reference values are dramatically different than we have used previously because we have removed the intrinsic solvation energy of an atom from the Lazaridis-Karplus solvation energy.

Table 3.1. Sequences (A) and energies (B) for Top7 before and after iterative cycles of backbone and sequence optimization (kcal mol^{-1}). Expected Lennard-Jones energies are derived from the average Lennard-Jones energy for each of the twenty amino acids for different degrees of burial.

A.

before DIEITVRINNNGEDYDYKKTATTLSEINAHFEELEKHLKEENGEKITISVKLRNEKEAYW

after DIQVQVNIDDNGKNFDYTYTVTTESELQKVLNELKDYIKKQGAKRVRISITARTKKEAEK

before VAAKIKEQALRAGVETIQIDKQSDTMTATLGKQ

after FAAILIKVFAELGYNDINVTFDGDTVTVEGQLE

B.

	Top7 before relaxation	Final Top7 model
Lennard-Jones (LJ) attractive	-370	-385
Lennard-Jones (LJ) repulsive	28	8.6
Hydrogen bonding	-89	-80
Solvation energy	188	175
Total energy	-324	-386
LJ attractive – expected LJ attractive (avg. per buried residue)	0.3	-0.3
LJ repulsive – expected LJ repulsive (avg. per buried residue)	0.2	-0.2

Table 3.2 Top7 Crystal Structure Statistics

DATA COLLECTION	
Resolution	50-2.5Å
Space Group	$P3_221$ [primitive trigonal]
Unit Cell Dimensions	35.9 Å, 35.9 Å, 140.6 Å
Wavelength	0.9793
Asymmetric Unit	Monomer
V_{im}	2.1 $Å^3$/dalton
Total Reflections	144,933
Unique Reflections	6,989
Completeness / (2.59-2.5)	99.1 % / (100.0%)
R_{emerge} / (2.59-2.5)	4.5 / (34.4)
I / σ / (2.59-2.5)	37.8 (5.0)
PHASING	
Phasing Power	1.99
Figure of Merit (before/after DM)	0.24 (0.85)
REFINEMENT	
R_{work}	0.268
R_{free}	0.293
Number of atoms	693
Number of waters	7
Residues in most-favoured regions	75 (88.2%)
Residues in additional allowed regions	7 (8.2 %)
Residues in generously allowed regions	3 (3.5%)
Residues in disallowed regions	0 (0.0%)
r.m.s.d bond length	0.0076
r.m.s.d. bond angles	1.35
Mean B value, mainchain	61.30 $Å^2$
Mean B value, sidechain	66.67 $Å^2$

Table 3.3: Definitions for atom types used in the Rosetta energy function

Atom Type Number	Atom type description
1	carbonyl carbon in sidechain of Asn and Gln, and guanidyl carbon in Arg
2	carboxyl carbon in Asp and Glu
3	aliphatic carbon with one hydrogen
4	aliphatic carbon with two hydrogens
5	aliphatic carbon with three hydrogens
6	aromatic ring carbon
7	nitrogen in Trp sidechain
8	nitrogen in His sidechain
9	nitrogen in Asn and Gln sidechain
10	nitrogen in Lys sidechain
11	nitrogen in Arg sidechain
12	nitrogen in Pro backbone
13	hydroxyl oxygen
14	carbonyl oxygen in Asn and Gln sidechains
15	carboxyl oxygen in Asp and Glu
16	sulphur in Cys and Met
17	backbone nitrogen
18	backbone alpha carbon
19	backbone carbonyl carbon
20	backbone oxygen
21	polar hydrogen
22	nonpolar hydrogen
23	aromatic hydrogen
24	backbone HN

Table 3.4 Well depths and radii used for the Lennard-Jones calculations. The well depths are those used in the CHARMM19 parameter set (Neria, Fischer et al. 1996). The radii were determined by fitting the Lennard-Jones potential to the distribution of distances observed between the atom types in the PDB.

Atom Type	Radii(r)	well depth (e)
1	2.00	0.1200
2	2.00	0.1200
3	2.00	0.0486
4	2.00	0.1142
5	2.00	0.1811
6	2.00	0.1200
7	1.75[1]	0.2384
8	1.75[1]	0.2384
9	1.75[1]	0.2384
10	1.75[1]	0.2384
11	1.75[1]	0.2384
12	1.75[1]	0.2384
13	1.55[1,2]	0.1591
14	1.55[2]	0.1591
15	1.55[2]	0.2100
16	1.90	0.1600
17	1.75	0.2384
18	2.00	0.0486
19	2.00	0.1400
20	1.55	0.1591
21	1.00[3]	0.0500
22	1.20	0.0500
23	1.20	0.0500
24	1.00[3]	0.0500

[1] These atom types are hydrogen bond donors and when paired with atom types that are hydrogen bond acceptors(13,14,15), r_{ij} is set to 2.95, the optimal distance for hydrogen bonding. This is to prevent the repulsive portion of the Lennard-Jones term from disfavouring hydrogen bonds.

[2] These atom types are hydrogen bond acceptors and when paired with atom types that are hydrogen bond donors (7,8,9,10,11,12,13) r_{ij} is set to 2.95.

[3] These are polar hydrogens and when paired with hydrogen bond acceptors (13,14,15), r_{ij} is set to 1.95.

Table 3.5 Parameters for the Lazaridis-Karplus solvation model.

Atom Type	ΔG^{free}	V	λ
1	0.00	14.7	3.5
2	-1.40	8.3	3.5
3	-0.25	23.7	3.5
4	0.52	22.4	3.5
5	1.50	30.0	3.5
6	0.08	18.4	3.5
7	-8.9	4.4	3.5
8	-4.0	4.4	3.5
9	-7.8	11.2	3.5
10	-20.0	11.2	6.0
11	-11.0	11.2	6.0
12	-1.55	0.0	3.5
13	-6.77	10.8	3.5
14	-7.8	10.8	3.5
15	-10.0	10.8	6.0
16	-4.1	14.7	3.5
17	-5.0	4.4	3.5
18	1.00	23.7	3.5
19	1.00	14.7	3.5
20	-5.00	10.8	3.5
21	0.00	0.0	3.5
22	0.00	0.0	3.5
23	0.00	0.0	3.5
24	0.00	0.0	3.5

Table 3.6: Weights used for the Rosetta energy function.

W_{atr}	0.80	W_{ref}(Ile)	-0.45
W_{rep}	0.65	W_{ref}(Lys)	1.30
W_{sol}	0.65	W_{ref}(Leu)	-1.62
W_{pair}	0.65	W_{ref}(Met)	0.25
W_{bb_hbond}	0.80	W_{ref}(Asn)	0.17
W_{sc_hbond}	1.10	W_{ref}(Pro)	-0.30
$W_{sc\_bb\_hbond}$	1.10	W_{ref}(Gln)	0.14
W_{rot}	0.70	W_{ref}(Arg)	0.60
W_{ref}(Ala)	0.05	W_{ref}(Ser)	0. 14
W_{ref}(Asp)	1.43	W_{ref}(Thr)	0.31
W_{ref}(Glu)	1.44	W_{ref}(Val)	-0.25
W_{ref}(Phe)	-1.20	W_{ref}(Trp)	-1.85
W_{ref}(Gly)	0.37	W_{ref}(Tyr)	-1.08
W_{ref}(His)	-1.92		

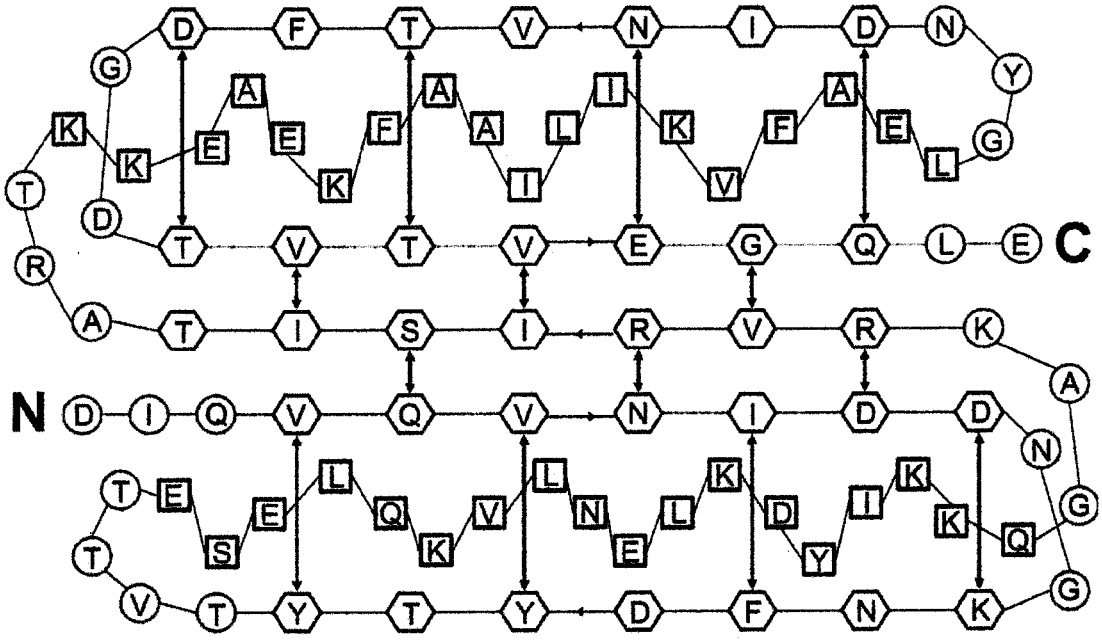


Figure 3.1 A two dimensional schematic of the target fold (hexagon = strand, square = helix, circle = other). Hydrogen bond partners are shown as purple arrows. The amino acids shown are those in the final designed (Top7) sequence.

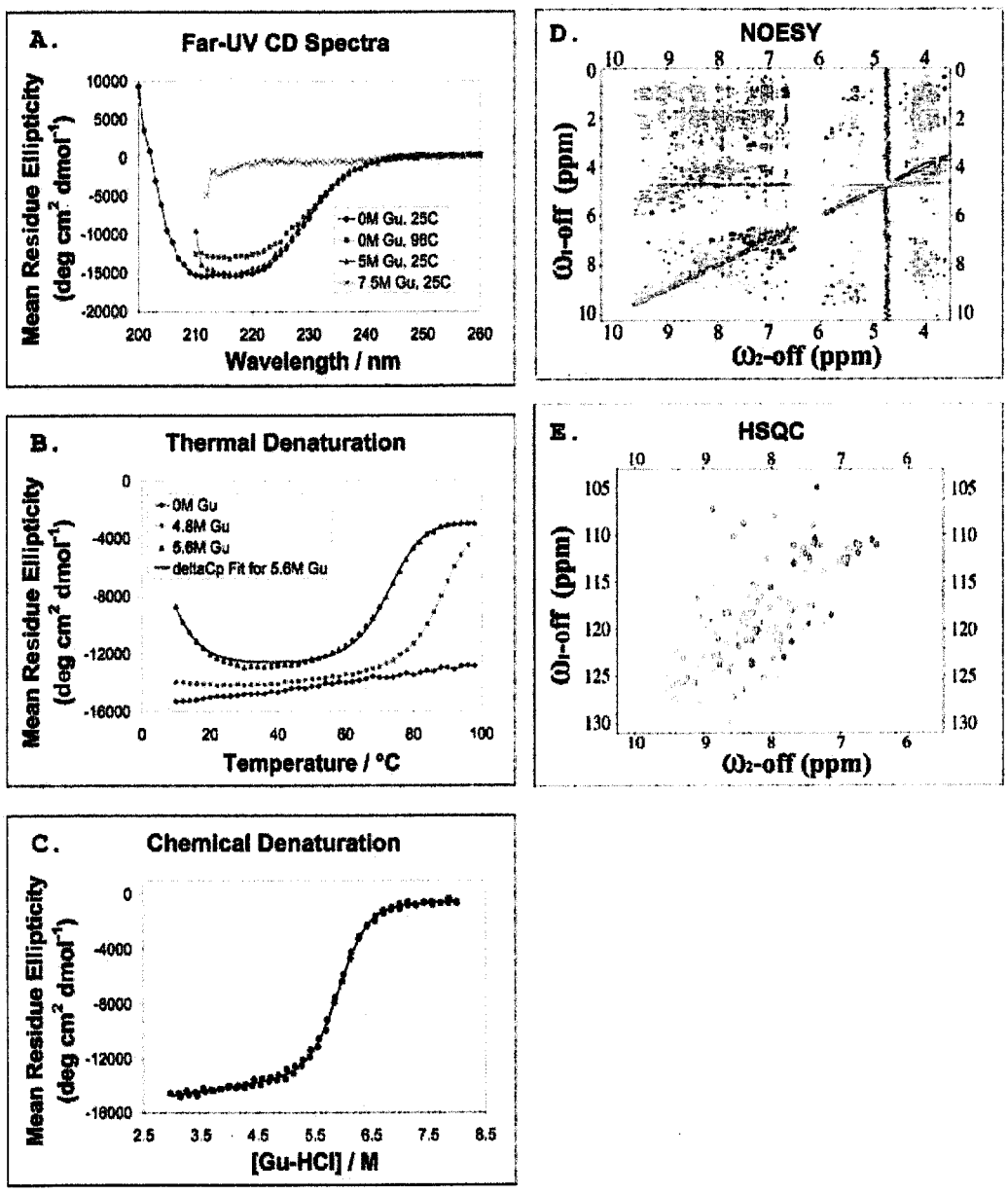


Figure 3.2 Biophysical characterization of Top7. A. The far-UV circular dichroism (CD) spectrum of 20uM Top7 in 25mM TRIS-HCl, 30mM NaCl, pH 8.0 at varying temperatures and concentrations of GuHCl. B. CD signal at 220 nm as a function of temperature and GuHCl for 8μM TOP7 in 25mM TRIS-HCl, 30mM NaCl, pH 8.0, in a 2mm cuvette. C. CD signal at 220nm as a function of GuHCl concentration for 5μM protein in 25mM TRIS-HCl, 30mM NaCl, pH 8.0 at 25°C, in a 1cm cuvette. D. The NOESY spectrum of ~1mM Top7 at pH 6.0 recorded at 298K, 500 MHz, and 200 ms mixing time using Watergate suppression. E. The HSQC spectrum of ~1mM 15N-Top7 at pH 6.0 recorded at 298K and 500Mhz using the fast HSQC scheme of Mori et al. (*36*)

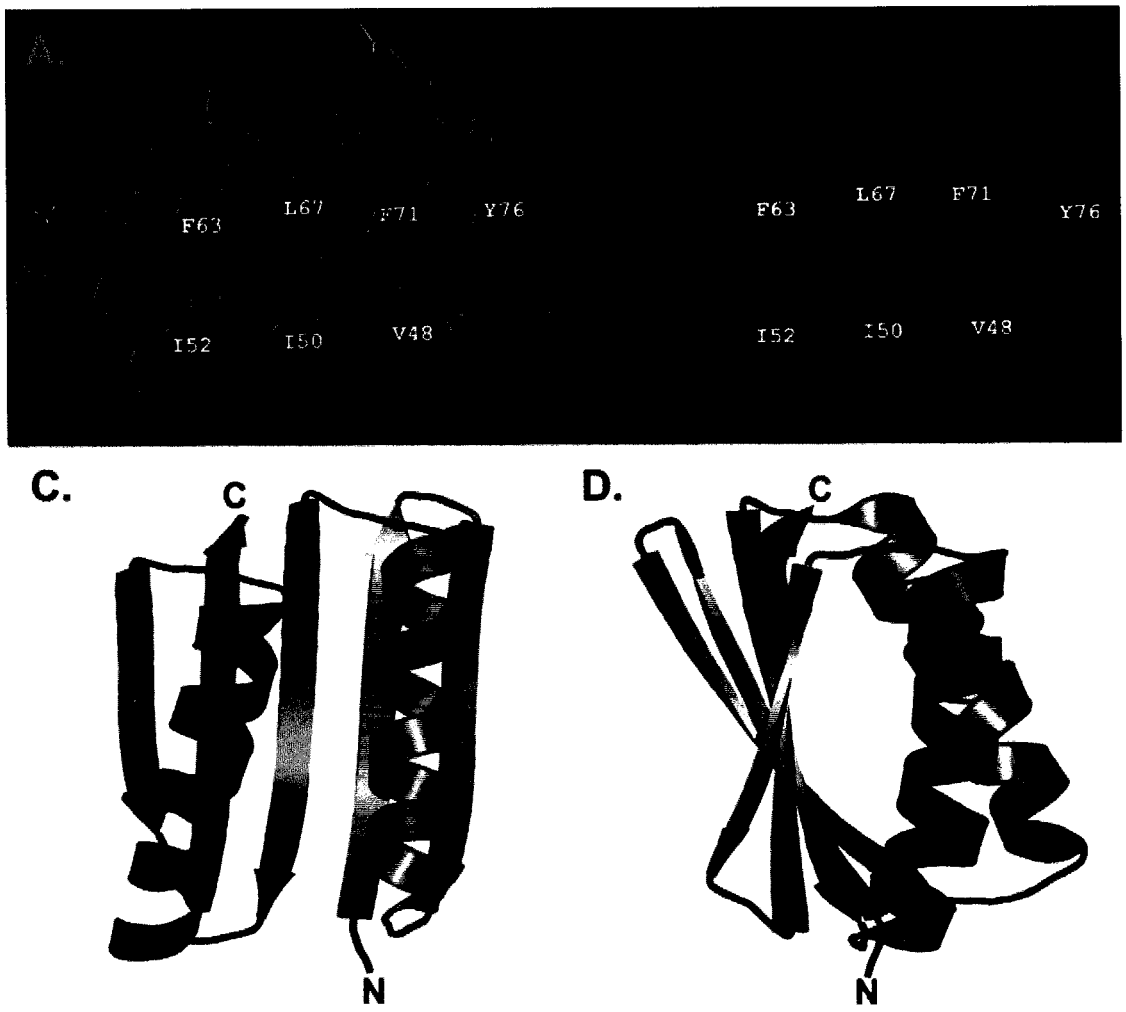


Figure 3.3 Schematic representation of Top7 in unbiased Single-Wavelength Anomalous Diffraction (SAD) density. A. and B. Stick representations of residues 46 through 76 from the computationally designed Top7 (left, green) and from the 2.5Å x-ray structure (right, red) are shown in unbiased density (blue). The map was generated from SAD phasing from a single selenomethionyl substituted variant of Top7, followed by density modification. C. and D. Ribbon diagrams of Top7 with residues 46-76 highlighted in red. The two diagrams are related by a 90° rotation around the vertical axis.

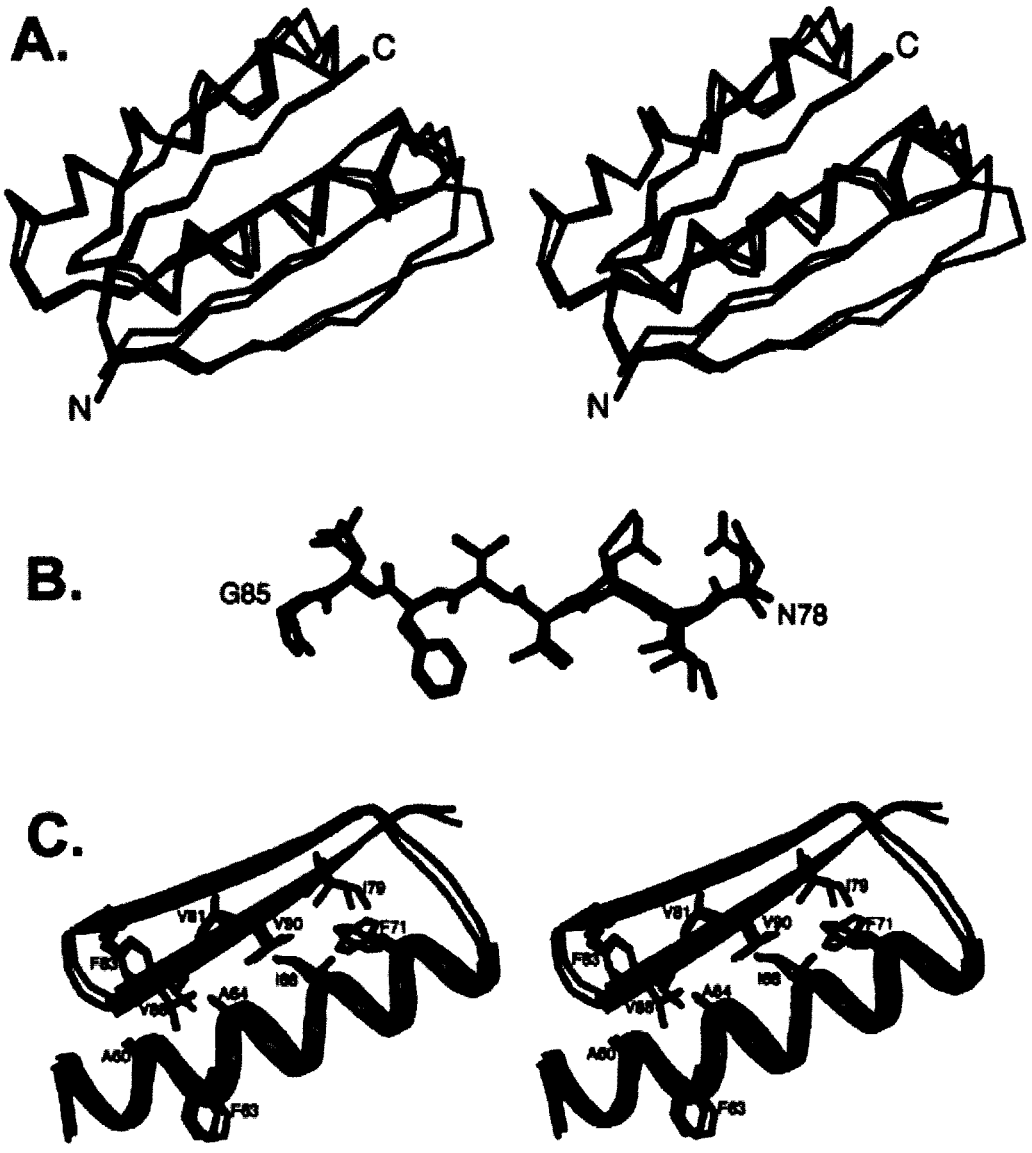


Figure 3.4 Comparison of the computationally designed model (blue) to the solved X-ray structure (red) of Top7. A. C-α overlay of the model and structure in stereo (backbone RMSD = 1.17Å). B. The C-terminal halves of the X-ray structure and model are extraordinarily similar. The representative region shown (Asn78-Gly85) has an all-atom RMSD of 0.79Å and a backbone RMSD of 0.54Å. C. Stereo representation of the effectively superposable side-chains in the cores of the designed model and the solved structure.

CHAPTER 4

A fragment of an *in silico* designed novel-fold protein forms a super-stable symmetric homodimer with a novel high-affinity interface.

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 (Dahiyat and Mayo 1997) and thermodynamically stabilise natural protein folds (Dantas, Kuhlman et al. 2003), to create novel (Dwyer, Looger et al. 2004) and thermodynamically-stabilised enzymes (Korkegian, Black et al. 2005), to redesign protein-protein (Kortemme, Joachimiak et al. 2004) and protein-ligand (Looger, Dwyer et al. 2003) interactions and to create extremely stable novel protein structures (Harbury, Plecs et al. 1998; Kuhlman, Dantas et al. 2003). Structural validation in many cases has confirmed the high-resolution accuracy of the design, lending credo to the assumption that true solution-state global energy minima had been captured by the computational algorithms (Dahiyat and Mayo 1997; Harbury, Plecs et al. 1998; Kuhlman, Dantas et al. 2003; Kortemme, Joachimiak et al. 2004; Korkegian, Black et al. 2005) (Dobson, N., Dantas, G., et al. manuscript in preparation). This 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 natural protein (Dantas,

Kuhlman et al. 2003; Kuhlman, Dantas et al. 2003). The obvious application of this stabilisation is the creation of longer-lasting designer proteins and therapeutics (Dahiyat 1999). An interesting question raised by these studies is whether there is any biological cost for this increased stability, especially in the context of the cellular environment.

Upwards of 30% of newly synthesized natural polypeptides in bacteria and eukaryotes are products of translation reactions that initiate inappropriately or end prematurely, or that erroneously decode the mRNA transcript (Kurland 1992; Schubert, Anton et al. 2000). An overwhelming majority of these mistranslated proteins fail to assume native-like conformations, and are targeted to the proteasome for efficient degradation within minutes of synthesis (Schubert, Anton et al. 2000; Goldberg 2003). However, aberrant protein translation products that fold into stable substructures can evade cellular surveillance mechanisms and their subsequent accumulation can significantly damage or kill cells (Cazzola and Skoda 2000; Bence, Sampat et al. 2001; Horwich 2002; Kozak 2002). These phenomena are implicated in the pathology of a large number of diseases, including diabetes, cancer, and many neurodegenerative disorders (Dobson 1999; Cohen and Kelly 2003; Selkoe 2003). Due to the significant rate of aberrant protein translation, the burden for maintaining cellular homeostasis falls on the highly selective and precisely regulated cellular degradation machinery, implying a stringent evolutionary selection against processes that could challenge this machinery (Goldberg 2003). Evolutionary selection would therefore be expected to be present against proteins that can be mistranslated into fragments that are independently and stably

folded. Computationally designed proteins provide a rare opportunity to reveal aspects of biology that are subject to natural selection, since they were designed in the absence of any specific evolutionary pressure. It is therefore of considerable interest to investigate how computationally designed proteins are handled by the cellular protein production and surveillance machinery.

We recently generated an extremely stable, small, globular protein, Top7, with a sequence and fold not observed previously in nature, using purely computational techniques (Kuhlman, Dantas et al. 2003). Biophysical and structural analysis of Top7 demonstrated the high-resolution accuracy of our design. Here we show that a portion of the original Top7 gene construct corresponding to the final 49 C-terminal residues is efficiently mistranslated in E. coli, and we solve the solution structure of this resulting CFr protein to reveal 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 discovery of CFr exposes the high error rate of the protein translation machinery and the rarity of corresponding stable fragments of natural proteins implies stringent evolutionary pressure against protein sub-fragments that can independently and stably fold. The symmetric self-association between two identical mistranslated CFr sub-units to create an extremely stable structure also suggests a mechanism for protein-fold evolution by modular recombination of stable protein sub-structures.

Results

During the original purification of the computationally designed Top7 protein, a strong band corresponding to a molecular weight of ~6.5kDa was consistently observed on SDS-PAGE gels. This band was observed in addition to the Top7 band (~12.5kDa) and remained even after Ni^+ affinity chromatography (Figure 4.1A, 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 ESI-MS, thereby allowing complete biophysical and structural characterisation of the pure Top7 protein (Kuhlman, Dantas et al. 2003). 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 determinant of the lower molecular weight 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 mistranslation.

Proteolysis or Mistranslation?

To investigate the possibility of the Top7 sub-fragment being a proteolytic product, Top7 bacterial cell lysates were incubated at room temperature for a series of time-intervals (up to 3 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. However, the ~6.5kDa Top7 sub-fragment band

was also observed in all supernatant fractions, also at relatively equal concentrations at all incubation times (data not shown). In effect, no appreciable degradation of Top7 was observed in vitro under conditions where many natural proteins show significant degradation (Maurizi 1992), and certainly no enrichment of this sub-fragment was observed with increasing incubation time. These results strongly argue against Top7 proteolysis yielding this sub-fragment.

MALDI-TOF-MS analysis of Ni^{+}-affinity purified Top7 confirmed that a strong species of ~6613Da was also present in addition to full length protein (data not shown). Assuming that this sub-fragment contained the 6XHis tag, the predicted molecular weight corresponded to a product ~30Da larger than a polypeptide starting at Val48 and ~120Da smaller than a polypeptide starting at Arg47. The sub-fragment was subsequently isolated away 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 ~30Da larger than Valine and hence a Top7 fragment starting with a V48M mutation matches the MALDI-TOF-MS predicted molecular weight. Since the plasmid coding for full-length Top7 does not contain this internal mutation, these results suggested that the sub-fragment might be a product of mistranslation of the Top7 mRNA starting at amino acid position 48.

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 (Gualerzi and Pon 1990). Could the Val48 GTG be the site (and cause) of mistranslation? To test this idea, we generated two single point mutants of the Top7 gene: a silent Val48Val (V48V) from GTG to GTT, and a Met1Ile (M1I) from ATG to ATT. Since GTT is never observed as a translation initiation codon, mistranslation from Val48 should be abrogated in this context, allowing translation of only the full-length product. Similarly, 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 4.1A). The V48V variant shows no observable expression of the ~6.5kDa sub-fragment band (lane 3). The M1I 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 4.1B). However, the MS results for M1I also suggested that at least two other minor species of intermediate molecular weight between full-length Top7 and the ~6.5kDa sub-fragment were present in the preparation. The predicted molecular weights 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 4.1C). In fact, zooming in on the 6-15kDa region in the SDS-PAGE gels after increased protein staining also showed the presence of faint protein bands between Top7 and the ~6.5kDa fragment. Further analysis of the Top7 gene sequence also revealed that degenerate

versions of the *E. coli* ribosomal binding site sequence (called the Shine-Dalgarno (SD) sequence) are present 2-8 bases upstream of all three identified Top7 mistranslation sites, and might aid in the observed mistranslation (Figure 4.1C).

The sequence of the ~6.5kDa fragment begins at a secondary structure break in the Top7 structure and includes strands 3, 4, and 5, and 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. Can this small sub-fragment of a purely computationally designed novel fold protein form a stable, independent structure? If so, does it adopt a structure similar to strands 3 through 5 in Top7, or are there structural rearrangements required to bury the hydrophobic protein core of Top7 that would be solvent exposed when its N-terminal half is ripped away? And what, if any, implication might this have on why nature does not appear to make proteins with the stability and fold observed for Top7?

Biophysical Characterisation of CFr

For further analysis, a separate gene construct that codes for the ~6.5kDa C-terminal Fragment (CFr) of Top7 was made as described in the Methods section. Like Top7, the CFr protein can be obtained with high yield (25mg/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 0.1Da of its theoretical molecular weight (Figure 4.2A).

Circular dichroism (CD) spectra strongly suggest that CFr is folded with α/β secondary structure, comparable in relative composition to Top7 (Figure 4.3A). CFr secondary structure appears unchanged at 98°C or in 3M guanidine-hydrochloride (GuHCl), but the CD spectrum of the protein is consistent with an unfolded polypeptide at 7M GuHCl. In the presence of intermediate GuHCl concentrations (4.3M), CFr unfolds co-operatively with temperature (Figure 4.3B), displaying remarkably high thermal stability, comparable to Top7. CFr also displays co-operative unfolding by GuHCl-induced chemical denaturation (Figure 4.3C). 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. Gel filtration analysis of CFr at 25μM and 1.2mM confirmed this suspicion – the protein resolves as a single peak with a molecular weight corresponding to a CFr dimer (data not shown). For a more robust characterisation of its oligomeric state, CFr was analysed by analytical ultra-centrifugation (AUC). In the 35-97μM concentration range, CFr appears predominantly dimeric at 0M and 4M GuHCl (where it appears folded by CD), and predominantly monomeric at 7M GuHCl (where it appears unfolded by CD) (Figure 4.4A,C). These results suggest that CFr may be 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 K_u or $\Delta G°$ (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 (108μM CFr), 25.5 kcal/mol (62μM CFr), and 25.5 kcal/mol (5μM CFr), confirming that CFr exists as an obligate dimer. A $\Delta G°$ of 25.5 kcal/mol corresponds to a dissociation constant of ~200 *zeptoM* (10^{-21} M).

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 4.5). The spectra are well-dispersed and sharp. Notably, the HSQC spectrum contains a single set of cross-peaks for each NH in the protein. Since CFr is a dimer, the HSQC indicates that every residue NH in one dimeric subunit is in the same magnetic environment as that corresponding NH in the other dimeric subunit, implying 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 (Folkers, Folmer et al. 1993; Zwahlen, Legault et al. 1997) to resolve intra-subunit NOEs from inter-subunit NOEs in CFr, to determine the symmetric homodimer solution structure.

Determination of the NMR Structure of CFr

Protein backbone and sidechain assignments were obtained as described in the Methods section. Over 98% of the backbone N, (N)H, C(O), C_α and C_β nuclei for residues 2-58 could be assigned (no assignments were possible for the N-terminal methionine and for the last four histidines at the C-terminus). Side chain ^1H and ^{13}C resonances were >92% assigned, but the aromatic side chains (Phe, Tyr, Trp) were >68% assigned. Gln/Asn NH_2 were 100% assigned while Arg N_ε and guanidinium groups and Lys NH_3 remain unassigned.

Structure determination was conducted in a 2-step process using the program CYANA 2.0 (Guntert 2003) – a fully automated iterative step 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 constraints. In the first step, 3873 NOE peaks 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, these data 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 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 (by 0.1-0.25 Å) and 2 angle constraint violations (by 33-36°).

In the second step of the calculation, all the intra-subunit distance and dihedral constraints from the final CFrA structure calculation were duplicated to generate constraints for a second subunit, CFrB. Additionally, 23 inter-subunit NOE peaks were manually assigned from 2D and 3D $^{12}C/^{13}C$ filtered NOESY spectra, and converted to 46 inter-subunit distance constraints between CFrA and CFrB. Initial structure calculations with these combined constraints yielded a small number of intra-subunit constraint violations, which upon further analysis could all be re-assigned as inter-subunit constraints. Two rounds of iterative structure calculation and manual refinement were sufficient to yield high-quality dimer structures. In the final calculation 100 structures were generated, of which the top 20 (Figure 4.6) had an average target function of 1.20 (±0.11) $Å^2$ (Table 4.1) and an ensemble RMSD 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 4.2). 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 (Laskowski, Macarthur et al. 1993), 99.2% of all dihedral angles were found in the allowed regions of the Ramachandran plot (Table 4.2). The small number of disallowed dihedrals are all found for residues in the linker region (Glu2 and Gly52-His58).

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, anti-parallel β-sheet (Figure 4.7, Top7 in purple, CFrA in green). The subunits form a symmetric anti-parallel dimer, with all interfacial residues contributed by the first strand of the β-sheet and by the helix (Figure 4.8). The two subunits have virtually identical structures with an RMSD of 0.41Å over backbone atoms and 0.81Å 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 of 1.12Å (Figure 4.7). This deviation corresponds to typical very low differences between NMR and crystallographic coordinates for proteins of identical structures and may reflect inaccuracies in the models as much 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 to 0.91Å. 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 1457Å^2 of Solvent Accessible Surface Area (SASA) (i.e. ~730Å^2 per subunit), which accounts for about 19% of the surface of each subunit (Figure 4.8A, interface carbons in green and yellow). Approximately 20

amino acids contribute to the interface (defined by any amino acids that lose >1Å^2 SASA when the dimer is compared to the individual subunits). Half of these residues are on the first strand of the β-sheet and half are on the helix (Figure 4.8B, green or yellow cartoons and sticks). Notably, they have a similar relative level of burial as the corresponding residues in the Top7 structure (data not shown). The CFr dimer interface appears to be an extension of the individual CFr subunit cores. The strands of the two subunits form an extended six-stranded anti-parallel 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, since this NH remains very strongly protected after prolonged D_2O exchange. The interfacial packing of inward-pointing residues on the first strands of both subunits (Val4, Ile6, Ile8, Ala10) appears identical to the packing between the side-chains on strands within each subunit (this "continuous sheet core" is illustrated in Figure 4.8C, *AB_00_SHEET*). Similar tight packing is also observed between helical side-chains interacting across the interface (Figure 4.8C, *AB_00_HELIX*). Notably, 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.

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 the CFr structure by the simple method of disulfide-induced protein circularisation. Since the NMR structure shows that the N- and C-termini of each CFr 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 cysteines should yield a covalently circularised form of each CFr subunit. The corresponding SS.CFr clone was generated and the protein purified as described in the Methods section.

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 MW, Figure 4.2B). The CD wavelength scan of SS.CFr appears identical to CFr (Figure 4.3A), suggesting the disulfide has not affected protein secondary structure. The chemical denaturation profile of SS.CFr (Figure 4.3C) shows it to be dramatically stabilised over CFr – the protein begins to unfold only at 6.5M GuHCl and appears to still be in the unfolding transition at 8.2M GuHCl. In comparison, both CFr and Top7 are completely unfolded at 6.5M GuHCl. 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 (33-105μM) is predominantly dimeric from 0M to 5M GuHCl, while only a small fraction of the monomeric form appears as the protein begins to unfold between 6M and 7M GuHCl (Figure 4.4B,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.

Since the accuracy of computational protein design likely improves with the increasing resolution of the template scaffold, we attempted to determine the crystal structure of SS.CFr. However, crystallization trials to date have only yielded crystals that diffract to 3.6Å, which can provide no higher structural resolution than the CFr NMR structure. Nevertheless, a 3.6Å data-set was collected at the Advanced Light Source (ALS) synchrotron, and a strong molecular replacement (MR) solution to the phase problem was found with the use of the CFr NMR dimer model. The SS.CFr MR solution yields a repeating crystal lattice with good packing between symmetry mates and no intermolecular clashes. While structural refinement in CNS initially resulted in a noticeable decrease in the Free-R factor, further rounds of refinement did not result in low enough R-factors to establish sufficient confidence in the final structural models. Nonetheless, these results suggest that the structure of SS.CFr is consistent with the CFr NMR dimer structure. Our current efforts are focused on producing SS.CFr crystals that diffract to higher resolution.

Discussion

CFr was generated by unintended yet efficient mistranslation of the Top7 gene sequence. Mass spectrometry analysis identified the primary reason for this and other minor Top7 mistranslation products to be the existence of GTG and TTG codons within the sequence, which E. coli sometimes uses in place of ATG to initiate translation (Gualerzi and Pon 1990). Replacement of the CFr GTG with GTT completely abrogated mistranslation without affecting Top7 expression. Biophysical characterisation of the CFr protein showed that is expressed as a highly stable and obligate dimer. The solution NMR structure further revealed that the CFr dimer is composed of an extremely stable novel symmetric interface formed between two identical CFr subunits, and analysis of NMR backbone dynamics further confirmed the rigidity of the structure. The CFr structure was further stabilised by disulfide-induced covalent circularisation of the individual CFr subunits.

The mistranslation leaves the residues on Top7's third β-strand and second helix (that were designed to be buried in the complete protein) with the challenge of how to remain buried. The answer chosen by those residues in CFr is simple – to pack symmetrically against the same hydrophobic patch from a twin copy of the monomeric subunit. Since the CFr dimer interaction was not intentionally designed, what features of the original Top7 design may have favoured the formation of a stable sub-fragment that dimerises to form a novel and surprisingly well-packed interface? One possibility is the low contact-order in the Top7 structure. The topology of Top7 dictates that the protein is

stabilised largely by local interactions, making it more likely for contiguous sequence fragments to adopt stable tertiary structures. Indeed, comparison of either of the subunits of the CFr NMR structure with the corresponding region in the Top7 structure shows that the backbone is virtually unchanged and that residues in the CFr subunit cores are essentially in the same environment and conformation as they were in Top7. While the sequence-local stabilisation allows the CFr sequence to adopt the same tertiary arrangement within each subunit, it might also partly explain why the subunits can easily self-associate. The two helices of Top7 have a predominantly "flat" hydrophobic interaction, i.e. the interacting surfaces on the helices have no large protrusions or intrusions to offer more than a general grease-on-grease surface complementarity. This implies no potential clashes between the two surfaces, but also implies that the surfaces are not dependent on each other to fold well. Additionally, the hydrophobic residues in close contact between the helices have a high level of sequence symmetry – the first helix contributes three leucines and an isoleucine and the second helix contributes two leucines, an isoleucine and a valine. This sequence symmetry is partially present in the interactions observed in the core of the β-sheet of Top7 as well. Two core isoleucines on the third strand in Top7 (Ile6 and Ile8 in CFr) interact with two valines from the first strand in Top7. In CFr, these isoleucines interact across the dimer interface with each other (Ile6 from one subunit with Ile8 of the other subunit). Since Val is a subset of Ile, the Van der Waals interactions between these residues are quite similar in CFr and Top7. These features allow CFr to form a self-interaction that closely approximates the overall nature of the Top7 core. These results suggest a few features we could specifically

incorporate into the design process in the future, to avoid super-stable progeny – (1) higher sequence contact-order to promote more non-local structural stabilisation, (2) design of knob-into-hole like packing between secondary structure elements to make the stability of these elements inter-dependent, and (3) design of specific atomic interactions in the protein core, such as buried salt-bridges, that would stabilise the native conformation but would destabilise non-native (i.e. not-designed) conformations.

There are some stabilising features of the CFr dimer interaction, however, that cannot be explained by any aspect of the Top7 design. In particular, the two symmetric aromatic clusters between the CFr dimer helices (Phe19 on one subunit and Phe27 and Tyr32 on the other subunit) do not mimic any comparable Top7 interaction. Another new interaction in CFr is the set of symmetric hydrogen bonds between the hydroxyl of Tyr32 on one subunit and the carboxyl moiety of Glu15 on the other subunit, which appear to form an interfacial stitch at the helical caps. The corresponding Tyr and Glu residues from the second helix in Top7 are involved in hydrogen bonds with residues on the first helix, but both interactions are water-mediated. Ultimately, these new interactions and those mimicked from the Top7 protein-core are likely synergistically important in making CFr such a strong obligate dimer.

Given the high rate of aberrant translation of natural proteins (Goldberg 2003), the accumulation of stable protein sub-fragments like CFr could provide a significant challenge to cellular homeostasis. For instance, stable sub-fragments that mimic

elements of native protein structure can act in a dominant-negative manner to interfere with functionally important interactions of the full-length "parent" protein with other cellular proteins, as seems to occur with the HIV-1 Gag protein (Schubert, Anton et al. 2000). Mistranslated natural proteins rarely yield stable substructures like CFr, and are instead generally unfolded and targeted to the proteasome for efficient degradation within minutes of synthesis (Schubert, Anton et al. 2000; Goldberg 2003). Taken with the high rate of aberrant protein translation, this suggests that evolution may have specifically selected against natural proteins containing subfragments that can fold autonomously. On the other hand, since the mistranslated Top7 sub-fragment, CFr, adopts a stable structure by forming an obligate symmetric self-association between two identical sub-units, it may reveal a mechanism for protein-fold evolution by modular recombination of stable protein sub-structures. Indeed, it has been previously proposed 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 homomultimers of identical chains, which were gradually replaced by single polypeptide chains encoding multiple repeats (Andrade, Perez-Iratxeta et al. 2001; Lupas, Ponting et al. 2001). This theory is supported by proteolytic dissection experiments of proteins such as Trp repressor or cytochrome c that yield small fragments capable of undergoing spontaneous noncovalent association to form subdomains with native-like secondary and tertiary structural features (Wu, Grandori et al. 1994). Submission of the CFr structure to the DALI server (Holm and Sander 1995) finds 122 natural protein

domains with significant structural homology (Z-score ≥ 2.0) to the CFr template. In many of these cases CFr subunits are found to be homologous to multiple non-overlapping parts of the same protein (eg. the *E. coli* acriflavine resistance protein pump (Yu, McDermott et al. 2003) has four distinct regions of homology to CFr), suggesting that the CFr fold might even be a natural protein-structure building block.

In addition to the evolutionary implications of the mistranslation 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 and extremely high denaturant concentrations. 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) (Schnolzer and Kent 1992; Dawson, Muir et al. 1994; Kochendoerfer 2001). 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 (Kochendoerfer 2001; Kochendoerfer, Chen et al. 2003). 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. 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 acids 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: MERVRISITARTKKEAEKFAAILIKVFAELGYNDINVTWDGDTVTVEGQLEGGSL EHHHHHH. 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 subcloning this fragment back into pet29b(+). The SS.CFr protein has the sequence: MECERVRISITARTKKEAEKFAAILIKVFAELGYNDINVTWDGDTVTVEGQLCGG SLEHHHHHH. Point mutants of Top7 (M1I and V48V) were generated using the Quick Change Site-Directed mutagenesis kit (Stratagene).

The 6X 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 OD_{600} of 0.6, induced with 1mM isopropyl-thio-β-D-galactosidase (IPTG), and cells were harvested after another 4-5 hours 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 Biotech) followed by 10^4-fold dialysis against 25mM TRIS-HCl, pH 8.0. The protein was further purified on a QFF anion exchange column (Pharmacia) with a 50mM to 600mM NaCl gradient in 25mM TRIS-HCl, pH 8.0, followed by a final 10^4-fold dialysis against 25mM TRIS-HCl, pH 8.0, (or 50mM sodium phosphate, pH 7.0 for NMR). To ensure complete disulfide formation, anion-exchange purified SS.CFr was oxidised in the presence of 20mM potassium ferricyanide [$K_3Fe(CN_6)$] for ten minutes at room temperature, prior to the final dialysis steps. Protein identity and purity was determined by SDS-PAGE and ESI-MALDI Mass Spectroscopy. Protein concentrations were determined by UV absorbance at 280nm with extinction coefficients calculated using the ExPASy Protparam tool (http://us.expasy.org/tools/protparam.html).

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 gl^{-1} of ^{15}N-NH_4Cl and 2gl^{-1} 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 8M GuHCl with overnight mixing to ensure dimer formation, dialysed back into 50mM NaPi pH 7.0, 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 absence of protease inhibitors (1mM PMSF, 1mM benzamidine). These two lysates were then divided into 4 equal fractions, which were incubated at room temperature for 2, 4, 24, and 72 hours respectively. After the incubation period, the lysates were centrifuged and separated into supernatant and pellet, which were subsequently visualised by SDS-PAGE.

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

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 3,000 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 hours for each speed. This equilibration time was deemed sufficient by identical absorbance scans collected after eight and ten hours 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 varying concentrations of GuHCl (Table 4.3). 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 (Cohn and Edsall 1943) and adjusted to 20°C (Laue 1992).

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 nonlinear least squares analysis to determine a weight-averaged molecular weight, M_W. During this analysis, the baseline offset was allowed to float, and if found to be > ±0.08, 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 (9 scans) to re-determine M_W. 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)

CD data were collected on an Aviv 62A DS spectrometer. Far-UV CD wavelength scans (260-195nm) at varying protein concentrations (10-25μM), guanidinium hydrochloride (*GuHCl*) concentrations (0-8.3M), and temperatures (0-98°C) were collected in a 1mm path length cuvette. GuHCl induced protein denaturation was followed by the change in ellipticity at 220nm in a 1cm 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 220nm in a 2mm path length cuvette. All CD data were converted to mean residue ellipticity. To obtain a value for the dimer association constant (K_d) and the free energy of unfolding (ΔG_U^{H2O}), chemical denaturation curves were fit to an equilibrium model between unfolded monomer (U) and folded dimer (F):

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

where

$$\exp\left(\frac{-\Delta G^\circ}{RT}\right) = K_d = [U]^2/[F_2] = 2P_t\left[f_u^2/(1-f_u)\right],$$

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:

$$\theta([Gu]) = (\theta_U - \theta_F) \bullet f_u + \theta_F,$$

where

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

$$\alpha = \left(\frac{\exp\left(\frac{-\Delta G^\circ}{RT}\right)}{2P_t} \right),$$

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

$$\Delta G^\circ([GuHCl]) = \Delta G^\circ(0M\ Gu - HCl) + m \bullet [GuHCl]$$
$$\theta_U([GuHCl]) = \theta_U(0M\ GuHCl) + a \bullet [GuHCl]$$
$$\theta_F([GuHCl]) = \theta_F(0M\ GuHCl) + b \bullet [GuHCl].$$

Nuclear Magnetic Resonance (NMR) spectroscopy

All CFr samples were prepared for NMR experiments in Shigemi susceptibility-matched NMR tubes, at 0.7-1.0mM concentration in H_2O solution containing 10% D_2O or in 100% D_2O, 50mM 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 (Delaglio, Grzesiek et al. 1995) and analyzed with SPARKY (Goddard and Kneller 2005) on Windows or Linux workstations.

Backbone amide ^{1}H and ^{15}N, C_α, C=O and side chain C_β 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 (Sattler, Schleucher et al. 1999). Side chain assignments were obtained by analysis of 3D HCCH-TOCSY and 3D ^{13}C-edited NOESY experiments. Aromatic side chain assignments were obtained from 2-dimensional NOESY and TOCSY spectra recorded in D_2O buffers. The spectra used in deriving distance constraints included 3D ^{15}N-edited NOESY and 3D ^{13}C-edited NOESY, 2D NOESY in H_2O (80ms and 120ms mixing) and 2D NOESY in D_2O (120ms mixing) recorded at 750MHz. Additionally, inter-subunit distance constraints were derived from 2D and 3D ^{12}C/^{13}C-filtered NOESY spectra (Folkers, Folmer et al. 1993; Zwahlen, Legault et al. 1997).

Protein structure determination by NMR

Structure determination was conducted in a 2-step process using the program CYANA 2.0 (Guntert 2003) – 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-edited 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 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 (Cornilescu, Delaglio et al. 1999) 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 ^{1}H-^{15}N HSQC spectra at 30

minutes and 50 hours after dissolving in D_2O. Hydrogen bond donors were identified by the presence of an amide peak in the 2D ^{1}H-^{15}N HSQC recorded at 30 minutes. The corresponding acceptors were identified by visualizing PDB files obtained from CYANA in Rasmol 2.7.1 (Sayle and Milner-White 1995) to identify carbonyl groups that were at a distance of approximately 2.0 Å from slow exchanging hydrogens. Each step of structural refinement in CYANA was performed with and without these hydrogen-bonding constraints.

Structure determination for a single subunit of CFr (i.e. one chain from the symmetric homo-dimer or CFrA) was performed by inputting unassigned peak lists together with full backbone assignments into CYANA. Initial NOESY cross-peak assignments derived from matching chemical shifts are subsequently refined in several cycles consisting of structure calculations using an error-tolerant target function followed by an assessment of the possible peak assignments in the light of the (preliminary) three-dimensional structures obtained (Guntert 2003). Successive complete CYANA calculations were used to identify additional cross-peaks consistent with the structural models and to remove mis-identified NOEs. After each CYANA calculation, any unassigned peaks were examined to verify that CYANA had not discarded the peak information in error. TALOS-derived dihedral angle constraints and hydrogen-bonding constraints were added during later CYANA runs, and the manual refinement process was continued until the CFrA structural ensembles had reasonable target functions ($<5Å^2$), low ensemble RMSDs, and minimal violations.

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 60Å tether was defined in CYANA 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 rigid body re-orientations. 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 sidechain assignments and the intra-subunit NOEs assignments from the CYANA runs. All inter-subunit NOE assignments were made as effectively 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 were calculated by SPARKY's Gaussian integration tool, and these 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 0.33Å). The quality of the final structure was evaluated with ProcheckNMR (Laskowski, Macarthur et al. 1993). Experimental constraints and structural statistics are reported in Table 4.1 and 4.2, respectively.

Solvent Accessible Surface Area (SASA)

SASA was calculated using the program NACCESS (Hubbard and Thornton 1993). 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.

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

Standard pulse sequences were used to measure the ^{15}N T_1, T_2 and heteronuclear NOEs (Farrow, Muhandiram et al. 1994, Deka, 2005 #129). All experiments utilize pulsed-field gradients for coherence selection, reduction of artifacts and sensitivity enhancement. In the CPMG sequence of the T_2 experiment, ^{1}H 180° pulses were applied for elimination of cross-correlation between ^{1}H-^{15}N dipolar and ^{15}N CSA relaxation mechanisms (Boyd, Hommel et al. 1990). A delay of 0.75 ms was inserted between successive applications of ^{15}N 180° with ^{1}H 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 ^{1}H dimensions, respectively. Delays of 0.030, 0.060, 0.100, 0.150, 0.220, 0.310, 0.420, and 0.550 seconds 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 seconds. The relaxation delay was 1.9 seconds for each experimental set. For the heteronuclear NOE measurements, a pair of spectra was recorded with and without proton saturation. Proton saturation was achieved by application of ^{1}H 120° pulses every 5 ms. Spectra recorded with proton saturation utilized a 2 second recycle delay followed by a 3 second period of saturation, while spectra recorded in the absence of saturation employed a recycle delay of 5 seconds.

All spectra were processed using NMRPipe/NMR-Draw 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 (Goddard and Kneller 2005). 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. 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% MPD, 6% PEG-4K and 0.1M Na-HEPES pH 6.9. The protein crystals grew within 2-6 days and were between 50-200 µm on a side. Since MPD is a cryo-protectant 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 (Otwinowski and Minor 1997).

The phases for the SS.CFr dataset were solved by molecular replacement (MR) with the program EPMR (Kissinger and Gehlhaar 1997). 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, vs. 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 (McRee 1999), along with rigid-body refinement, torsion-angle based simulated annealing, and conjugate-gradient based minimization in CNS (Brunger, Adams et al. 1998).

Table 4.1: NMR experimental constraints for CFr *(residues 2-58)*

Monomer Calculation	
Unique NOE distance constraints (first round/final round)*	
Total	1915/1116
Intraresidue & Sequential ([i - j] ≤1)	1312/645
Medium-range (1≤[i - j]≤5)	513/198
Long-range ([i - j]≥5)	90/273
Dihedral angle constraints†	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
Dimer Calculation‡	
Intra-subunit NOE constraints	2232
Inter-subunit NOE constraints	46
Dihedral angle constraints†	130
Hydrogen bond constraints	36 (18 H-bonds)
Total number of constraints	2444
Residual constraint violations‡	
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 A^2/5.2 A^2
Final dimer calculation	----------/1.2 A^2

* First and final round refer to statistics from the NOEASSIGN macro in Cyana2.0.
† Dihedral angle constraints were generated from TALOS (Cornilescu, Delaglio et al. 1999)
‡ All dimer restraints and violations are two-fold redundant due to the symmetric nature of the structure (see *Methods* for details).

Table 4.2: Structural statistics for CFr dimer

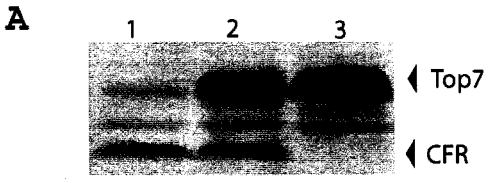
RMSD from averaged structure (Å)[††]	
(Structured region, residues 3-51 in both chains)	
Backbone atoms	0.33
All heavy-atoms	0.75
PROCHECK-NMR analysis[††]	
(All residues in both chains)	
Most favoured regions (%)	80.3
Additionally allowed (%)	17.3
Generously allowed (%)	1.6
Disallowed (%)	0.8

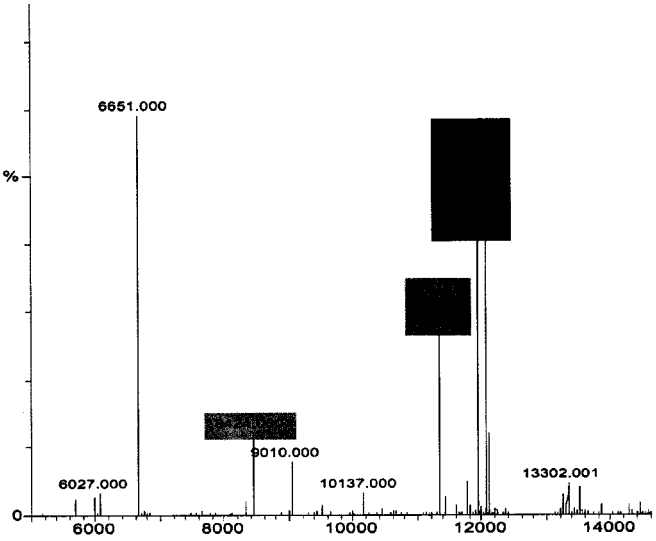
[††] Structural statistics reported are based on analysis of the best 20 conformers of 100 generated by CYANA.

Table 4.3: Measured solvent densities, ρ, of 25mM tris-HCl, pH 8.0 at various concentrations of guanidine hydrochloride

GuHCl	ρ [a] (g/mL)	Standard Deviation
0	0.99927	0.00002
3	1.07807	0.00008
4	1.10153	0.00046
5	1.2521	0.00014
6	1.14766	0.00038
7	1.16931	0.00021

[a] Values represent an average of three measurements





C

```
[redacted]D I Q V Q [redacted]N I D D N G K N F D Y T
[redacted]GATATCC[redacted]TTCAA[redacted]AACATTGACGATAATGGTAAAAACTTCGATTATACT
Y T V T T E S E L Q K V [redacted]N E L K D Y I
TATACCGTGACCACAGAATCGGAATTAC[redacted]TT[redacted]AACGAACTGAAAGACTACATC
K K Q G A K R V R I S I T A R T K K E A
AAAAAAC[redacted]CTAAACGCGTGCGCATCTCGATTACCGCTCGTACGAAAAAAGAAGCA
E K F A A I L I K V F A E L G Y N D I N
GAAAAATTTGCCGCGATTTTGATTAAAGTTTTCGCCGAACTGGGTTATAATGATATTAAC
V T F D G D T V T V E G Q L E G G S L E
GTTACGTTTGACGGCGATACGGTGACAGTGGAAGGACAGTTAGAGGGCGGTTCTCTCGAG
H H H H H H
CACCACCACCACCACCAC
```

Figure 4.1 Mistranslation of Top7. (A) Coomassie-stained SDS-PAGE gel of Top7 protein variants M1I, wild-type, and V48V (lanes 1-3, respectively). (B) ESI-MS spectrum of Top7_M1I. (C) Top7 protein (top lines) and DNA (bottom lines) sequence, with primary and alternate initiation codons highlighted (colours match the peaks from panel B). Degenerate Shine-Dalgarno sequences are highlighted in red.

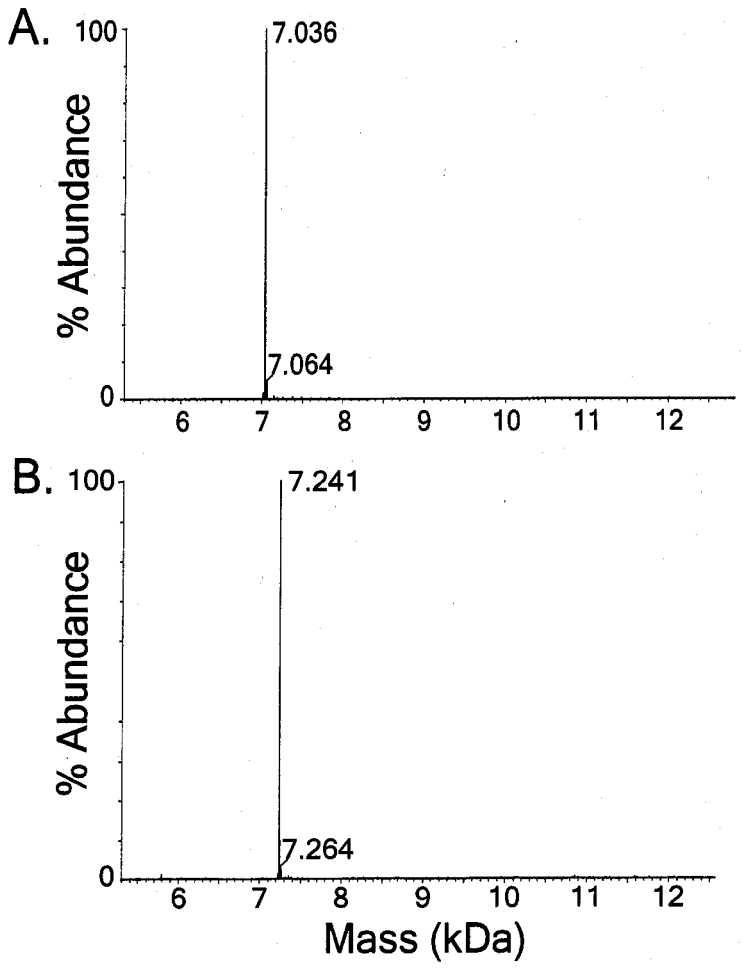


Figure 4.2 ESI-MS spectra of CFr (A) and SS.CFr (B). Major peaks are labelled with observed molecular weight.

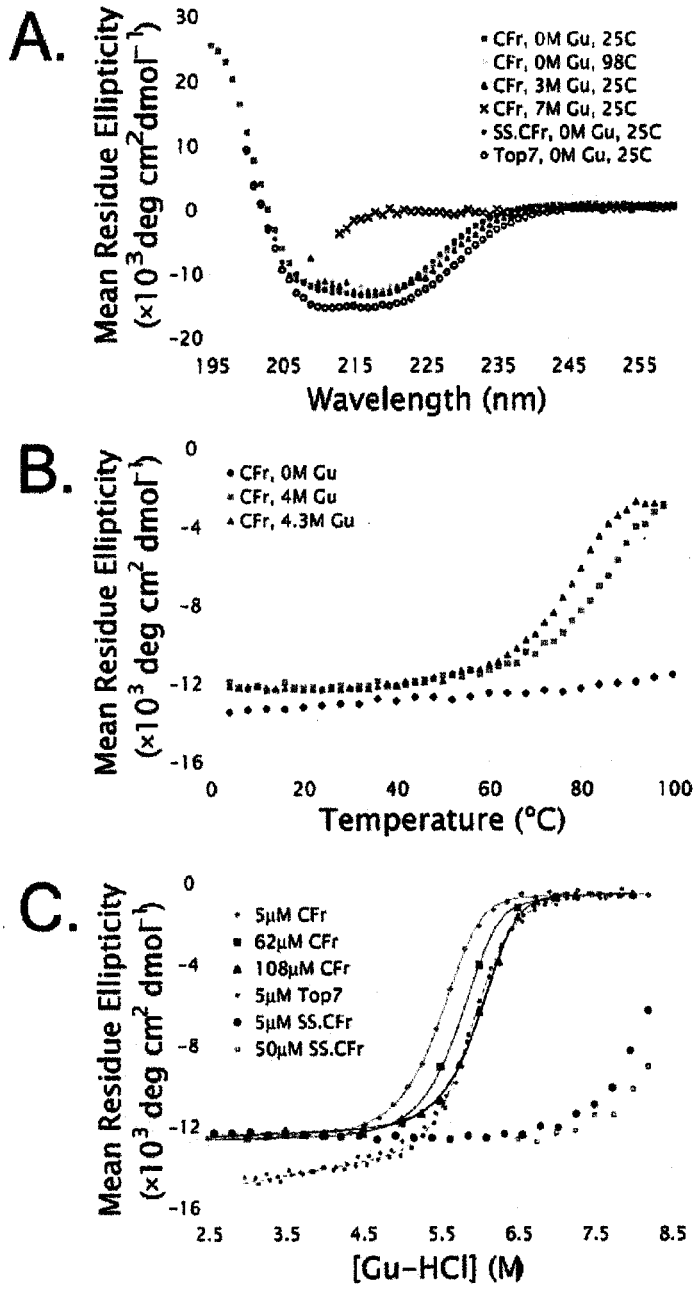


Figure 4.3 Biophysical characterisation of CFr and SS.CFr. (A) The far-ultraviolet (UV) CD spectrum of 25μM CFr, 25μM SS.CFr and 20μM Top7 in 25mM tris-HCl, pH 8.0 at varying temperatures and GuHCl concentrations. (B) CD signal at 220nm as a function of temperature and GuHCl for 12μM CFr in 25mM tris-HCl, pH 8.0, in a 2-mm cuvette. (C) CD Signal at 220nm as a function of GuHCl concentration for multiple concentrations of CFr, SS.CFr, and Top7 in 25mM tris-HCl, pH 8.0, at 25°C in a 1-cm cuvette.

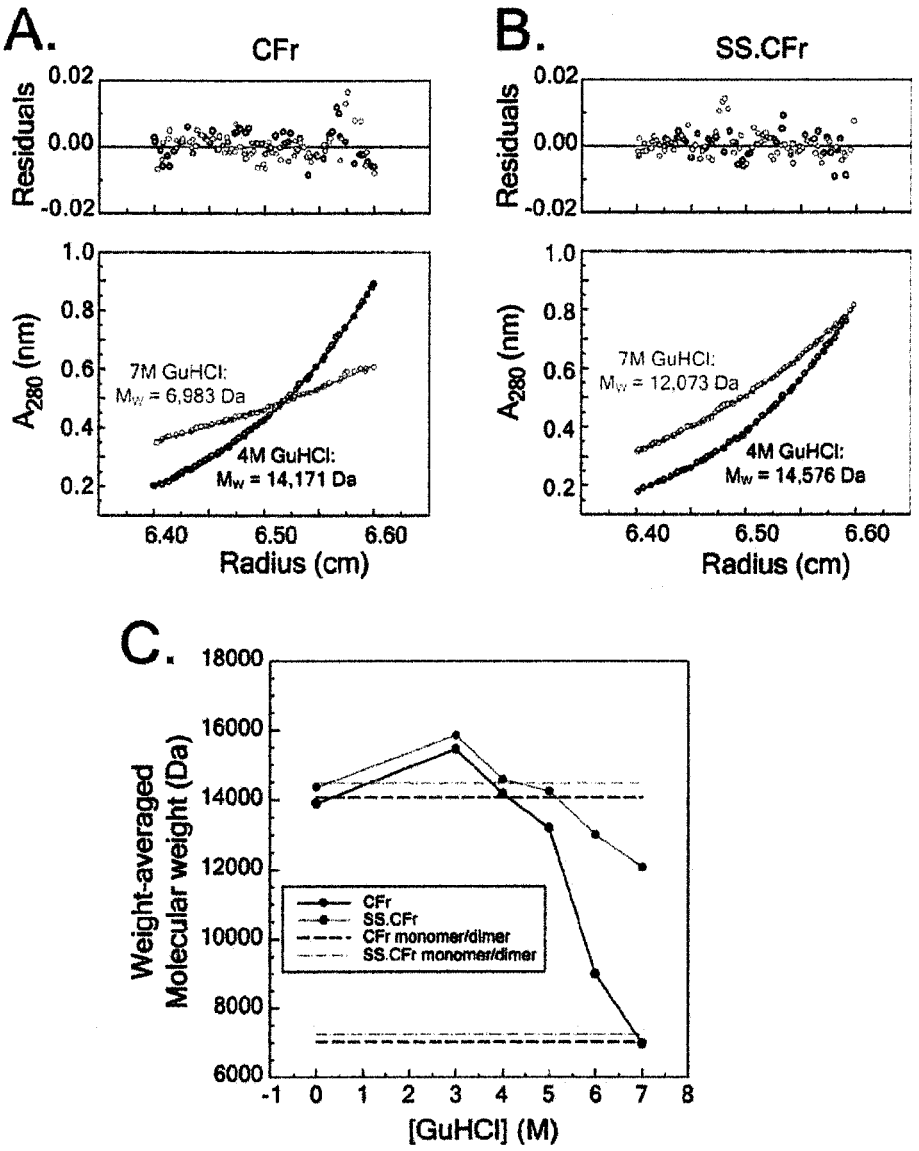


Figure 4.4 Analytical Ultra-Centrifugation (AUC) studies of CFr and SS.CFr. Selected equilibrium profiles collected for (A) CFr and (B) SS.CFr collected at 30,000 rpm, 20°C at protein concentrations of 59-66 μM in solvent containing 4 M (black circles) or 7 M (red circles) GuHCl. The fitted weight-averaged molecular weight (M_W) was determined using a global fit to 9 equilibrium scans collected at three protein concentrations and three speeds (see Methods). (C) Fitted M_W vs. concentration of GuHCl plot. Fitted M_W 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 weights for CFr, 7,037/14,074 (black, dashed), and SS.CFr, 7,241/14,482 (red, dotted-dashed).

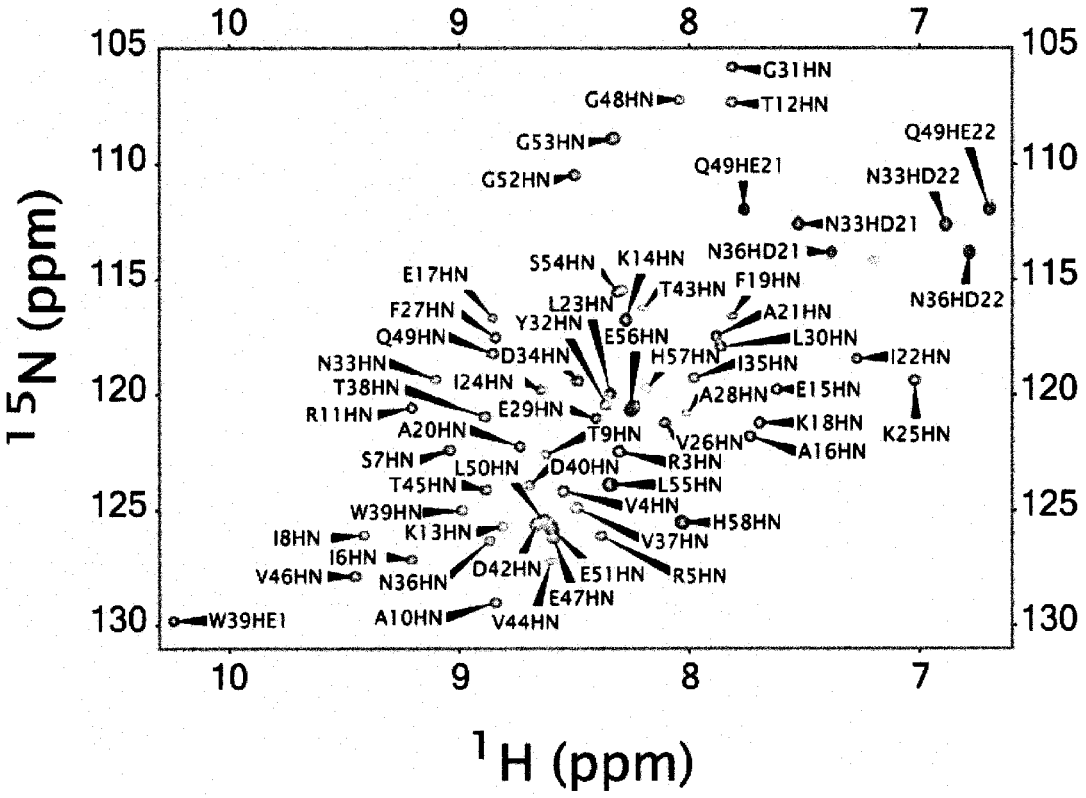


Figure 4.5 ^{1}H-^{15}N HSQC spectrum of CFr. The HSQC spectrum of ~1mM ^{15}N-CFr in 50mM NaPi, pH 7.0, recorded at 298 K and 500 MHz with the use of the fast HSQC scheme of Mori *et al.* (Mori, Abeygunawardana et al. 1995). Peaks are labelled with one-letter amino-acid code and sequence number.

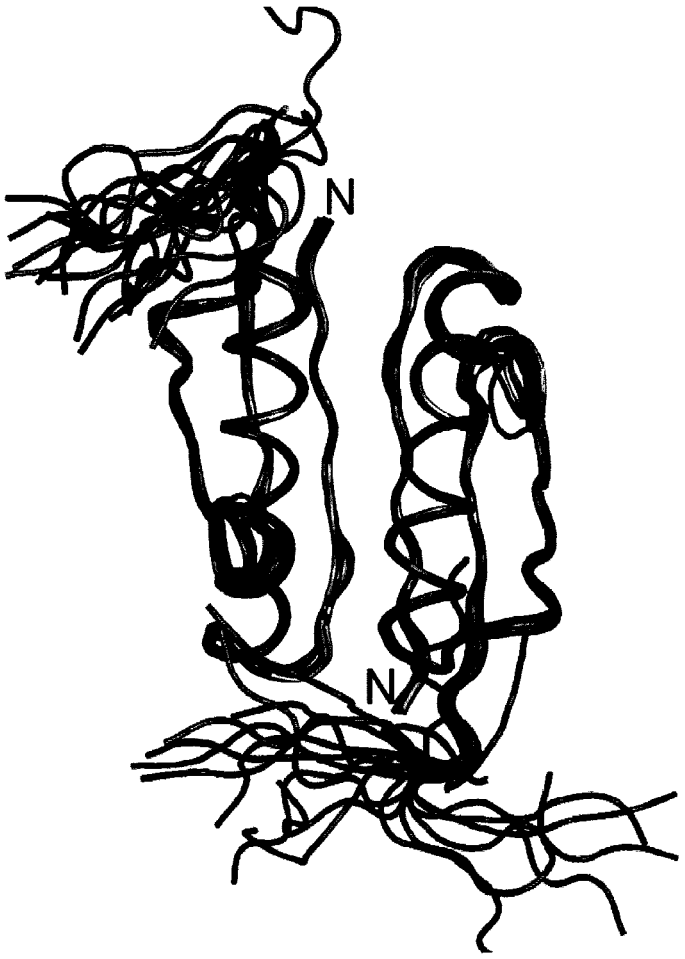


Figure 4.6 Schematic representation of the NMR-generated structures of CFr. Top 20 NMR models from the final CFr structure calculation are shown as ribbons. Each model is superimposed on the average backbone co-ordinates for residues 3-51 (structured region, separate colour for each model) in both chains from the entire ensemble. The structured regions have an ensemble RMSD of 0.33Å over backbone atoms and 0.75Å over all heavy-atoms. Unstructured tail residues (52-58) are coloured in grey.

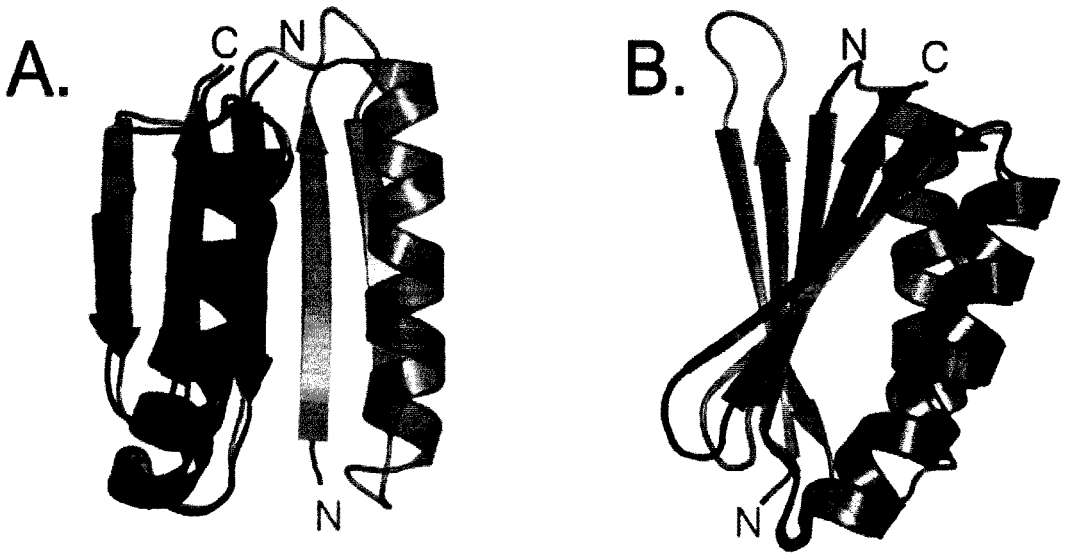


Figure 4.7 Comparison of the Top7 and CFr structures. (A and B) Ribbon diagrams of residues 3-51 from one subunit of the CFr NMR structure (green) superimposed on the corresponding region of the Top7 x-ray structure (purple). The backbone RMSD over these residues is 1.12Å. The two diagrams are related by a 90° rotation around the vertical axis in the plane of the page.

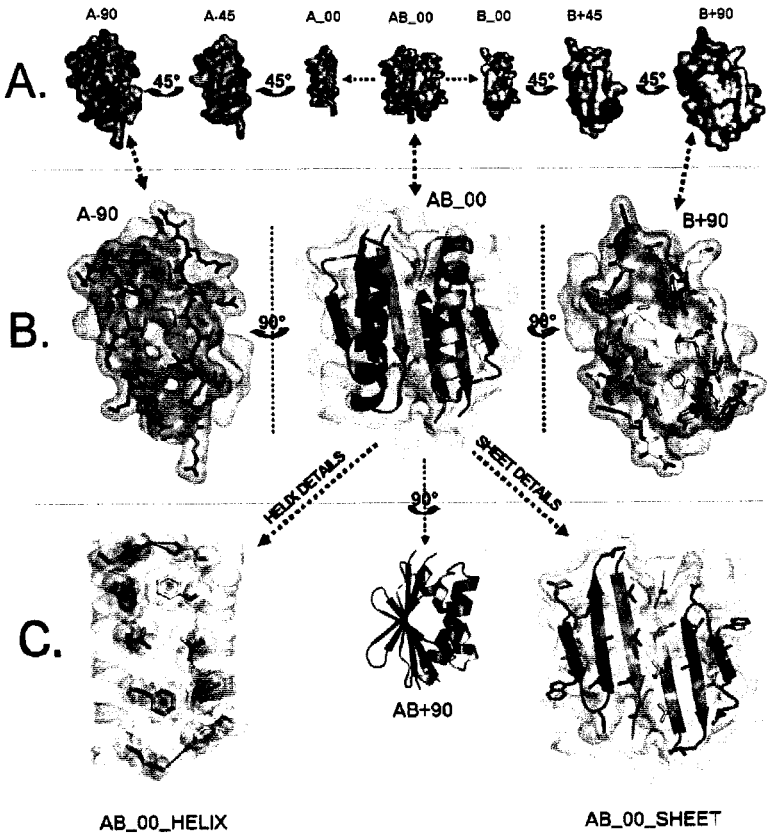


Figure 4.8 Details of the CFr NMR structure. (A) Seven views of the two subunits of CFr shown in surface representation. Interfacial carbon atoms are coloured green in subunit A and yellow in subunit B, and all other atoms are in CPK colour. Starting with the centre model of the dimer, the three models 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 sidechains as green or yellow sticks. Surface representations are overlaid with 80% transparency to show orientation relative to panel 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 sidechains 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 panels 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.

BIBLIOGRAPHY

Alber, T., J. A. Bell, et al. (1988). "Replacements of Pro86 in phage T4 lysozyme extend an alpha-helix but do not alter protein stability." Science **239**(4840): 631-5.

Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-402.

Andrade, M. A., C. Perez-Iratxeta, et al. (2001). "Protein repeats: structures, functions, and evolution." J Struct Biol **134**(2-3): 117-31.

Baldwin, E. P., O. Hajiseyedjavadi, et al. (1993). "The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme." Science **262**(5140): 1715-8.

Bence, N. F., R. M. Sampat, et al. (2001). "Impairment of the ubiquitin-proteasome system by protein aggregation." Science **292**(5521): 1552-5.

Benson, D. E., D. W. Conrad, et al. (2001). "Design of bioelectronic interfaces by exploiting hinge-bending motions in proteins." Science **293**(5535): 1641-4.

Bonneau, R., J. Tsai, et al. (2001). "Rosetta in CASP4: progress in ab initio protein structure prediction." Proteins **Suppl 5**: 119-26.

Bowers, P. M., C. E. Strauss, et al. (2000). "De novo protein structure determination using sparse NMR data." J Biomol NMR **18**(4): 311-8.

Boyd, J., U. Hommel, et al. (1990). "Influence of Cross-Correlation between Dipolar and Anisotropic Chemical-Shift Relaxation Mechanisms Upon Longitudinal Relaxation Rates of N-15 in Macromolecules." Chemical Physics Letters **175**(5): 477-482.

Brunger, A. T., P. D. Adams, et al. (1998). "Crystallography & NMR system: A new software suite for macromolecular structure determination." Acta Crystallogr D Biol Crystallogr **54**(Pt 5): 905-21.

Bryson, J. W., S. F. Betz, et al. (1995). "Protein design: a hierarchic approach." Science **270**(5238): 935-41.

Cazzola, M. and R. C. Skoda (2000). "Translational pathophysiology: a novel molecular mechanism of human disease." Blood **95**(11): 3280-8.

Cohen, F. E. and J. W. Kelly (2003). "Therapeutic approaches to protein-misfolding diseases." Nature **426**(6968): 905-9.

Cohn, E. J. and J. T. Edsall (1943). Proteins, amino acids and peptides as ions and dipolar ions. New York, Reinhold Publishing Corporation.

Cornilescu, G., F. Delaglio, et al. (1999). "Protein backbone angle restraints from searching a database for chemical shift and sequence homology." J Biomol NMR **13**(3): 289-302.

Dahiyat, B. I. (1999). "In silico design for protein stabilization." Curr Opin Biotechnol **10**(4): 387-90.

Dahiyat, B. I. and S. L. Mayo (1997). "De novo protein design: fully automated sequence selection." Science **278**(5335): 82-7.

Dantas, G., B. Kuhlman, et al. (2003). "A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins." J Mol Biol **332**(2): 449-60.

Dawson, P. E., T. W. Muir, et al. (1994). "Synthesis of proteins by native chemical ligation." Science **266**(5186): 776-9.

DeGrado, W. F., C. M. Summa, et al. (1999). "De novo design and structural characterization of proteins and metalloproteins." Annu Rev Biochem **68**: 779-819.

Delaglio, F., S. Grzesiek, et al. (1995). "NMRPipe: a multidimensional spectral processing system based on UNIX pipes." J Biomol NMR **6**(3): 277-93.

Desjarlais, J. R. and N. D. Clarke (1998). "Computer search algorithms in protein modification and design." Curr Opin Struct Biol **8**(4): 471-5.

Desjarlais, J. R. and T. M. Handel (1995). "De novo design of the hydrophobic cores of proteins." Protein Sci **4**(10): 2006-18.

Dill, K. A. (1990). "Dominant forces in protein folding." Biochemistry **29**(31): 7133-55.

Dobson, C. M. (1999). "Protein misfolding, evolution and disease." Trends Biochem Sci **24**(9): 329-32.

Doublie, S. (1997). "Preparation of selenomethionyl proteins for phase determination." Methods in Enzymology **276**: 523 - 530.

Dunbrack, R. L., Jr. and F. E. Cohen (1997). "Bayesian statistical analysis of protein side-chain rotamer preferences." Protein Sci **6**(8): 1661-81.

Dwyer, M. A., L. L. Looger, et al. (2004). "Computational design of a biologically active enzyme." Science **304**(5679): 1967-71.

Farrow, N. A., R. Muhandiram, et al. (1994). "Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation." Biochemistry **33**(19): 5984-6003.

Folkers, P. J. M., R. H. A. Folmer, et al. (1993). "Overcoming The Ambiguity Problem Encountered In The Analysis Of Nuclear Overhauser Magnetic-Resonance Spectra Of Symmetrical Dimer Proteins." Journal Of The American Chemical Society **115**(9): 3798-3799.

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

Goldberg, A. L. (2003). "Protein degradation and protection against misfolded or damaged proteins." Nature **426**(6968): 895-9.

Grantcharova, V. P., D. S. Riddle, et al. (1998). "Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain." Nat Struct Biol **5**(8): 714-20.

Gualerzi, C. O. and C. L. Pon (1990). "Initiation of mRNA translation in prokaryotes." Biochemistry **29**(25): 5881-9.

Guex, N. and M. C. Peitsch (1997). "SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling." Electrophoresis **18**: 2714 - 2723.

Guntert, P. (2003). "Automated NMR protein structure calculation." Progress In Nuclear Magnetic Resonance Spectroscopy **43**(3-4): 105-125.

Harbury, P. B., J. J. Plecs, et al. (1998). "High-resolution protein design with backbone freedom." Science **282**(5393): 1462-7.

Havranek, J. J. and P. B. Harbury (2003). "Automated design of specificity in molecular recognition." Nat Struct Biol **10**(1): 45-52.

Hendrickson, W. A. (1991). "Determination of macromolecular structures from anomalous diffraction of synchrotron radiation." Science **254**(5028): 51-8.

Holm, L. and C. Sander (1995). "Dali: a network tool for protein structure comparison." Trends Biochem Sci **20**(11): 478-80.

Horwich, A. (2002). "Protein aggregation in disease: a role for folding intermediates forming specific multimeric interactions." J Clin Invest **110**(9): 1221-32.

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

Hubbard, T. J., A. G. Murzin, et al. (1997). "SCOP: a structural classification of proteins database." Nucleic Acids Res **25**(1): 236-9.

Jaramillo, A., L. Wernisch, et al. (2002). "Folding free energy function selects native-like protein sequences in the core but not on the surface." Proc Natl Acad Sci U S A **99**(21): 13554-9.

Jin, W., O. Kambara, et al. (2003). "De novo design of foldable proteins with smooth folding funnel. Automated negative design and experimental verification." Structure (Camb) **11**(5): 581-90.

Johnson, E. C., G. A. Lazar, et al. (1999). "Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin." Structure Fold Des **7**(8): 967-76.

Jones, T. A., J.-Y. Zou, et al. (1991). "Improved methods for building protein models in electron density maps and the location of errors in these models." Acta Cryst A **47**: 110 - 119.

Keating, A. E., V. N. Malashkevich, et al. (2001). "Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils." Proc Natl Acad Sci U S A **98**(26): 14825-30.

Keywegt, G. J. and T. A. Jones (1996). Halloween.masks and bones. From First Map to Final Model. S. Bailey, R. Hubbard and D. Waller. Warrington, U.K., SERC Daresbury Laboratory: 59 -66.

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

Kochendoerfer, G. G. (2001). "Chemical protein synthesis methods in drug discovery." Curr Opin Drug Discov Devel **4**(2): 205-14.

Kochendoerfer, G. G., S. Y. Chen, et al. (2003). "Design and chemical synthesis of a homogeneous polymer-modified erythropoiesis protein." Science **299**(5608): 884-7.

Koehl, P. and M. Levitt (1999). "De novo protein design. I. In search of stability and specificity." J Mol Biol **293**(5): 1161-81.

Korkegian, A., M. E. Black, et al. (2005). "Computational thermostabilization of an enzyme." Science **308**(5723): 857-60.

Kortemme, T., L. A. Joachimiak, et al. (2004). "Computational redesign of protein-protein interaction specificity." Nat Struct Mol Biol **11**(4): 371-9.

Kortemme, T., A. V. Morozov, et al. (2003). "An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes." J Mol Biol **326**(4): 1239-59.

Kozak, M. (2002). "Pushing the limits of the scanning mechanism for initiation of translation." Gene **299**(1-2): 1-34.

Kranz, J. K., J. Lu, et al. (1996). "Contribution of the tyrosines to the structure and function of the human U1A N-terminal RNA binding domain." Protein Sci **5**(8): 1567-83.

Kuhlman, B. and D. Baker (2000). "Native protein sequences are close to optimal for their structures." Proc Natl Acad Sci U S A **97**(19): 10383-8.

Kuhlman, B., G. Dantas, et al. (2003). "Design of a novel globular protein fold with atomic-level accuracy." Science **302**(5649): 1364-8.

Kuhlman, B., J. W. O'Neill, et al. (2001). "Conversion of monomeric protein L to an obligate dimer by computational protein design." Proc Natl Acad Sci U S A **98**(19): 10687-91.

Kuhlman, B., J. W. O'Neill, et al. (2002). "Accurate computer-based design of a new backbone conformation in the second turn of protein L." J Mol Biol **315**(3): 471-7.

Kurland, C. G. (1992). "Translational accuracy and the fitness of bacteria." Annu Rev Genet **26**: 29-50.

Larson, S. M., A. A. Di Nardo, et al. (2000). "Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions." J Mol Biol **303**(3): 433-46.

Larson, S. M., J. L. England, et al. (2002). "Thoroughly sampling sequence space: large-scale protein design of structural ensembles." Protein Sci **11**(12): 2804-13.

Laskowski, R. A., M. W. MacArthur, et al. (1993). "PROCHECK: a program to check the stereochemical quality of protein structures." J. Appl. Cryst. **26**: 283-291.

Laskowski, R. J., M. W. Macarthur, et al. (1993). "PROCHECK: a program to check the stereochemical quality of protein structures." J. Appl. Crystall. **26**: 283 - 291.

Laue, T. M. (1992). Short column sedimentation equilibrium analysis for characterization of macromolecules in solution. Technical Information DS-835. Palo Alto, CA, Spinco Business Unit.

Lazaridis, T. and M. Karplus (1999). "Effective energy function for proteins in solution." Proteins **35**(2): 133-52.

Lim, W. A., D. C. Farruggio, et al. (1992). "Structural and energetic consequences of disruptive mutations in a protein core." Biochemistry **31**(17): 4324-33.

Lim, W. A., A. Hodel, et al. (1994). "The crystal structure of a mutant protein with altered but improved hydrophobic core packing." Proc Natl Acad Sci U S A **91**(1): 423-7.

Looger, L. L., M. A. Dwyer, et al. (2003). "Computational design of receptor and sensor proteins with novel functions." Nature **423**(6936): 185-90.

Looger, L. L. and H. W. Hellinga (2001). "Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics." J Mol Biol **307**(1): 429-45.

Lupas, A. N., C. P. Ponting, et al. (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**(2-3): 191-203.

Malakauskas, S. M. and S. L. Mayo (1998). "Design, structure and stability of a hyperthermophilic protein variant." Nat Struct Biol **5**(6): 470-5.

Maurizi, M. R. (1992). "Proteases and protein degradation in Escherichia coli." Experientia **48**(2): 178-201.

McRee, D. E. (1999). "A Versatile Program for Manipulating Atomic Coordinates and Electron Density." J. Structural Biology **125**: 156-165.

Merritt, E. A. and D. J. Bacon (1997). "Raster3D: Photorealistic molecular graphics." Macromolecular Crystallography, Pt B **277**: 505-524.

Metropolis, N., A. Rosenbluth, et al. (1953). "Equations of state calculations by fast computing machines." J Chem Phys **21**: 1087-1092.

Mirny, L. and E. Shakhnovich (2001). "Evolutionary conservation of the folding nucleus." J Mol Biol **308**(2): 123-9.

Mori, S., C. Abeygunawardana, et al. (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 B **108**(1): 94-8.

Murzin, A. G. (2003). "Top7 represents a novel SCOP class." D. Baker.

Myers, J. K., C. N. Pace, et al. (1995). "Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding." Protein Sci **4**(10): 2138-48.

Nauli, S., B. Kuhlman, et al. (2001). "Computer-based redesign of a protein folding pathway." Nat Struct Biol **8**(7): 602-5.

Neria, E., S. Fischer, et al. (1996). "Simulation of activation free energies in molecular systems." Journal of Chemical Physics **105**(5): 1902-1921.

Offredi, F., F. Dubail, et al. (2003). "De novo Backbone and Sequence Design of an Idealized alpha/beta-barrel Protein: Evidence of Stable Tertiary Structure." J Mol Biol **325**(1): 163-74.

Otwinowski, Z. and W. Minor (1997). "Processing of X-ray diffraction data collected in oscillation mode." Methods in Enzymology **276**: 307 - 326.

Plaxco, K. W., K. T. Simons, et al. (1998). "Contact order, transition state placement, and the refolding rates of single domain proteins." J. Mol. Biol. **277**: 985-994.

Pokala, N. and T. M. Handel (2001). "Review: protein design--where we were, where we are, where we're going." J Struct Biol **134**(2-3): 269-81.

Ponder, J. W. and F. M. Richards (1987). "Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes." J Mol Biol **193**(4): 775-91.

Press, W. H. (1992). Numerical recipes in C: the art of scientific computing. Cambridge; New York, Cambridge University Press.

Raha, K., A. M. Wollacott, et al. (2000). "Prediction of amino acid sequence from structure." Protein Sci **9**(6): 1106-19.

Regan, L. (1999). "Protein redesign." Curr Opin Struct Biol **9**(4): 494-9.

Reina, J., E. Lacroix, et al. (2002). "Computer-aided design of a PDZ domain to recognize new target sequences." Nat Struct Biol **9**(8): 621-7.

Richardson, J. S. and D. C. Richardson (2002). "Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation." Proc Natl Acad Sci U S A **99**(5): 2754-9.

Rohl, C., C. E. M. Straus, et al. (2003). "Protein structure prediction using Rosetta." Methods in Enzymology.

Sattler, M., J. Schleucher, et al. (1999). "Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients." Progress In Nuclear Magnetic Resonance Spectroscopy **34**(2): 93-158.

Sayle, R. A. and E. J. Milner-White (1995). "RASMOL: biomolecular graphics for all." Trends Biochem Sci **20**(9): 374.

Scalley, M. L., Q. Yi, et al. (1997). "Kinetics of folding of the IgG binding domain of peptostreptococcal protein L." Biochemistry **36**(11): 3373-82.

Schnolzer, M. and S. B. Kent (1992). "Constructing proteins by dovetailing unprotected synthetic peptides: backbone-engineered HIV protease." Science **256**(5054): 221-5.

Schubert, U., L. C. Anton, et al. (2000). "Rapid degradation of a large fraction of newly synthesized proteins by proteasomes." Nature **404**(6779): 770-4.

Selkoe, D. J. (2003). "Folding proteins in fatal ways." Nature **426**(6968): 900-4.

Shimaoka, M., J. M. Shifman, et al. (2000). "Computational design of an integrin I domain stabilized in the open high affinity conformation." Nat Struct Biol **7**(8): 674-8.

Simons, K. T., I. Ruczinski, et al. (1999). "Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins." Proteins **34**(1): 82-95.

Street, A. G. and S. L. Mayo (1999). "Computational protein design." Structure Fold Des **7**(5): R105-9.

Su, A. and S. L. Mayo (1997). "Coupling backbone flexibility and amino acid sequence selection in protein design." Protein Sci **6**(8): 1701-7.

Summa, C. M., M. M. Rosenblatt, et al. (2002). "Computational de novo design, and characterization of an A(2)B(2) diiron protein." J Mol Biol **321**(5): 923-38.

Taddei, N., F. Chiti, et al. (1999). "Thermodynamics and kinetics of folding of common-type acylphosphatase: comparison to the highly homologous muscle isoenzyme." Biochemistry **38**(7): 2135-42.

Uversky, V. N., Z. K. Abdullaev, et al. (1999). "Structure and stability of recombinant protein depend on the extra N-terminal methionine residue: S6 permutein from direct and fusion expression systems." Biochim Biophys Acta **1432**(2): 324-32.

Veeraraghavan, S., T. F. Holzman, et al. (1996). "Autocatalyzed protein folding." Biochemistry **35**(33): 10601-7.

Villegas, V., A. Azuaga, et al. (1995). "Evidence for a two-state transition in the folding process of the activation domain of human procarboxypeptidase A2." Biochemistry **34**(46): 15105-10.

Voigt, C. A., D. B. Gordon, et al. (2000). "Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design." J Mol Biol **299**(3): 789-803.

Wernisch, L., S. Hery, et al. (2000). "Automatic protein design with all atom force-fields by exact and heuristic optimization." J Mol Biol **301**(3): 713-36.

Word, J. M., S. C. Lovell, et al. (1999). "Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms." J Mol Biol **285**(4): 1711-33.

Wu, L. C., R. Grandori, et al. (1994). "Autonomous subdomains in protein folding." Protein Sci **3**(3): 369-71.

Wüthrich, K. (1986). NMR of Proteins and Nucleic Acids. New York, John Wiley & Sons, Inc.

Yi, Q., M. L. Scalley, et al. (1997). "Characterization of the free energy spectrum of peptostreptococcal protein L." Fold Des **2**(5): 271-80.

Yu, E. W., G. McDermott, et al. (2003). "Structural basis of multiple drug-binding capacity of the AcrB multidrug efflux pump." Science **300**(5621): 976-80.

Zwahlen, C., P. Legault, et al. (1997). "Methods for measurement of intermolecular NOEs by multinuclear NMR spectroscopy: Application to a bacteriophage lambda N-peptide/boxB RNA complex." Journal Of The American Chemical Society **119**(29): 6711-6721.

VITA

GAUTAM DANTAS

EDUCATION

Ph.D., Biochemistry 2000 – 2005
University of Washington, Seattle, WA
Advisor: Prof. David Baker

B.A., Chemistry and Biology with Biochemistry Emphasis 1996 – 2000
Macalester College, St. Paul, MN
Advisor: Prof. Rebecca Hoye

WORK EXPERIENCE AND SKILLS

Research Associate, Biochemistry. University of Washington. 2000 – 2005
Advisor: Prof. David Baker

- Experimentally characterised the folding and stability of nine computationally redesigned globular proteins.
- Experimentally characterised the folding and stability of Top7, a protein computationally designed to have a fold not found in nature. Solved the x-ray crystal structure of the Top7 protein to illustrate the atomic-level accuracy of the design process.
- Experimentally characterised the folding and stability of an ultra-stable sub-fragment of the Top7 protein. Solved the NMR structure of this sub-fragment.
- Computationally designed novel interactions and specificities between the human siderocalin protein and a group of small ligands. Experimental characterisation in progress.
- Proficient in the use and application of computational protein structure prediction and design algorithms.
- Extensive molecular biology experience (cloning, PCR, site-directed mutagenesis, gene assembly, phage display, gel electrophoresis).
- Extensive protein purification experience (ion-exchange chromatography, affinity chromatography, size-exclusion chromatography, FPLC, HPLC, and standard electrophoretic techniques).

- Highly skilled in protein biophysical and structural characterisation techniques (circular dichroism, UV/VIS and fluorescence spectroscopy, MALDI mass spectrometry, 1D, 2D, and 3D homo- and hetero-nuclear NMR spectroscopy, x-ray crystallography).

Teaching Assistant, Biochemistry. University of Washington. 2001 – 2002

- Lectured, supervised, and graded a senior undergraduate biochemistry laboratory course
- Led discussion sections and wrote and graded exams for introductory and advanced undergraduate biochemistry theory courses.

Research Assistant, Biochemistry/Toxicology. Macalester College. 1999 – 2000

Advisor: Prof. James Straka

- Developed and characterised an assay for the detection of organophosphorous toxins in watersheds.
- Developed skills in the collection and cataloguing of fish, amphibian, and water samples in the field, and animal care in the laboratory.
- Extensive experience with chemical and biochemical analysis of specimens in the laboratory setting (blood extraction from animals, enzyme preparation from animal tissues, fluorescence-based spectrophotometric kinetics assay, LC/GC mass spectrometry).

Research Assistant, Organic Chemistry. Macalester College. 1998 – 1999

Advisor: Prof. Rebecca Hoye

- Synthesized a novel macrocyclic polyamine with high yield and purity, using a new amino-protection strategy.
- Developed skills in the design and execution of multi-step organic synthesis protocols.
- Developed skills in organic compound characterisation techniques (1D NMR, LC/GC mass spectrometry, FTIR spectroscopy, UV/VIS spectroscopy)

HONORS AND AWARDS

Newcomb Cleveland Prize (American Association of the Advancement of Science) for Outstanding Publication of the Year	2004
Biochemistry Department Representative, Dreyfus Grant Committee	2002 – 2004
Member, Phi Lambda Upsilon National Chemistry Honor Society	2000 – Present
Graduated with Honours in Chemistry and Biology, Macalester College	2000
Violet O. Beltmann Competitive Undergraduate Research Scholarship	1998 – 2000
International Baccalaureate Diploma	1996

PUBLICATIONS AND PRESENTATIONS

1. **Dantas, G.**, Watters, A.L., Lunde, B., Eletr, Z., Isern, N., Kuhlman, B., Stoddard, B.L., Varani, G., & Baker, D. (2005). A fragment of an *in silico* designed novel-fold protein forms a super-stable symmetric homodimer with a novel high-affinity interface. *Manuscript in preparation.*

2. Dobson, N., **Dantas, G.**, Baker, D., & Varani, G. (2005). Complete computational redesign of human U1A protein: Thermodynamic stabilization with high-resolution validation of protein structure and dynamics. *Structure, submitted.*

3. Dobson, N., **Dantas, G.**, & Varani, G. (2005). ^{1}H, ^{13}C and ^{15}N resonance assignments of URNDesign, a computationally redesigned RRM protein. *Journal of Biomolecular NMR, submitted.*

4. **Dantas, G.***, Kuhlman, B.*, Ireton G.C., Varani G., Stoddard B.L. & Baker D. (2004). Design of a novel globular protein fold with atomic level accuracy. Northwest Crystallography Conference, 2004. Oral presentation.

5. 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.

6. **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.

7. Hoye, R.C., Richman, J.E., **Dantas, G.**, Lightbourne, M.F. & Shinneman L.S. (2001). Synthesis of polyazamacrocyclic compounds via modified Richman-Atkins cyclizations of β-trimethylsilylethanesulfonamides. *J. Org. Chem.*, 66:2722-2725.

8. **Dantas, G.**, Harmes, D.C., & Straka J.G. (1999). Toxicity of organopesticides and heavy metals: standardization of procedures for evaluating watershed quality. *PEW Midstates Undergraduate Symposium in the Physical Sciences*, Univ. of Chicago. Poster and oral presentations.

* these authors contributed equally to this work