Exploring Protein Backbone Designability:
The Computational Redesign and *de novo* Design of Helix Bundle Proteins

Grant Sterling Murphy

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry.

Chapel Hill
2011

Approved by

Brian Kuhlman

Gary Pielak

Marcey Waters

Charlie Carter

Mischa Machius

ABSTRACT

Grant Sterling Murphy : Exploring Protein Backbone Designability:
The Computational Redesign and *de novo* Design of Helix Bundle Proteins

(Under the direction of Brian Kuhlman)

Protein design rigorously tests our mastery of protein folding, stability and function. Protein design can be separated into redesign and *de novo* design by the issue of designability, which states that not all protein backbones will lead to viable sequences. The goal of redesign is to find favorable sequences for proteins with known structures, using their experimental coordinates as design models. *De novo* design requires design model coordinates to be created from scratch and then finds favorable sequences. Nature provides designable backbones in the case of fixed backbone redesign. In flexible redesign and *de novo* design, however, we have no guarantee of designability.

This work develops computational methods for flexible redesign and *de novo* design of diverse protein folds, probing questions of designability. We successfully used flexible redesign on several helix-bundle proteins and solved X-ray and NMR structures for one redesigned protein. The design model and the experimental structures are highly similar, < 1.0 Å backbone rmsd. Our success in *de novo* design has been modest. We have not succeeded in the *de novo* design of an all β-fold and continue to pursue this challenge. We have succeeded in the *de novo* design of a four helix-bundle protein. Preliminary NMR data suggests our design model and the experimental structure are the same fold and are similar at a global level with a backbone rmsd of ≤ 3Å.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

$\Delta C_p°$	Change in Specific Heat
$\Delta G°$	Free Energy of Unfolding
$\Delta H°$	Enthalpy of Unfolding
Å	Angstrom
m-value	Slope of ΔG vs. Denaturant Concentration
M	Molarity
K	Kelvin
°C	Degrees Celsius
T_m	Melting Temperature
CD	Circular Dichroism
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
RMSD	Root Mean Square Deviation
WT	Wild Type
E. Coli	*Escherichia coli*
IPTG	isopropyl β–D–thiogalactoside
GDT	Global Distance Test

Chapter 1

Introduction

The History of the Molecular Revolution and the Birth of Protein Design

Our understanding of the physical and chemical forces that control the structure and function of macromolecular systems has advanced rapidly in the last century, developing into the fields of molecular and structural biology. It is inspiring to consider that the identities of all twenty proteinogenic amino acids were not known until the discovery of Threonine in 1935 by William C. Rose(Rose 1968). Perhaps even more amazing is that the structures of the twenty proteinogenic amino acids were not known by crystallography until after this point and the chemical synthesis or purification from natural sources of each amino acid was an arduous process. Even with this limited knowledge physicists, chemists, and biologist were actively pursuing the fundamental questions in protein folding, protein structure, and genetics. Less than thirty years later, in 1951, Pauling had (correctly) predicted the structures of the α-helix and the β-sheets(Pauling and Corey 1951; Pauling, Corey et al. 1951). Followed quickly by the solution of the structure of DNA by X-ray crystallography in 1953 and the first X-ray structures of a protein by Perutz and Kendrew in 1959(Watson and Crick 1953; Kendrew and Perutz 1957; Perutz, Rossmann et al. 1960). In 1963, Merrifield published a watershed paper detailing the rapid and efficient synthesis of long polypeptide chains, 4 amino acids, by solid phase peptide synthesis, his technique is still the standard method of peptide/protein synthesis and has been extended to over 70 amino acids(Merrifield 1965). The 1970's and early 80's saw the development of advanced cloning technology, recombinant protein expression, protein purification, and the ability to solve protein structures by X-ray crystallography and NMR(Regnier 1983;

Williamson, Havel et al. 1985; Mullis, Faloona et al. 1986). With these technologies in place, the field of protein design began to coalesce as researchers who had previously engineered small molecules and/or manipulated proteins with single mutations imagined the possibilities of creating designer proteins.

The History of Protein Design

The first attempt at the design of an entire protein came from Gutte et al in 1979(Gutte, Daumigen et al. 1979). Prior to 1979, Gutte, Kaiser, and others attempted the design of organic biological mimics and small peptides with great success. In 1969, Gutte published the first successful chemical synthesis of an entire enzyme, Ribonuclease A, and showed it to be functional and indistinguishable from Ribonuclease A purified from natural sources(Gutte and Merrifield 1969). With the knowledge and ability to explore new protein sequences and permutations of existing protein sequences, two design methodologies became prevalent and are the major approaches used in protein design problems today. The first method begins with a protein sequence or structure known to be viable. Mutations are made to this sequence for a desired goal. This problem is known as the protein redesign problem and has been highly successful in the last 15 years. The second method requires the creation of protein sequences and structures from scratch. This problem is known as *de novo* protein design. The *de novo* design process is more challenging, more rewarding and if successful indicates a true understanding of the fundamental rules that govern protein folding, stability and function.

It is possible to perform redesign and *de novo* design at various levels of protein structure and knowledge. The most common incarnation of protein design focuses on the accurate representation of atomic level interactions that define protein structure and function. This incredibly detailed representation is not required and several protein design challenges have been successfully solved using primary sequence information, evolutionary techniques, and/or hierarchical methods.

While evolutionary methods have been successful in protein design and facilitate the testing of great numbers of possible sequences, they often lack the phenomenological relationship that is necessary to understand the roots of success, failure, and the ability to explicitly improve our knowledge(Hecht, Das et al. 2004). In contrast, hierarchical methods and model systems attempt to make the problems of protein design tractable and often directly lead to an improved understanding of protein folding and design. Unfortunately, the testing of hierarchical methods is inherently methodical, slow, and prone to generalization and simplification.

The most common and arguably the most successful technique in protein design has been the modeling of explicit atomic level interactions using techniques that in large part derive from the concepts and equations for free energy from physical chemistry and statistical mechanics. The inherent limitations with this method are the vast computational resources required (although a few examples of detailed model building by hand do exist), the challenging process of codifying the complex and often approximated rules of physics, chemistry, and biology, and the difficulties in improving computational methods from a limited number of successful attempts.

A Timeline of Protein Design

Gutte et al. published the first example of a designed protein in 1979 and was a true driving force for the field of protein design. The goal of their research was to create a designed sequence that would mimic the activity of Ribonuclease A (Figure 1A). This attempt was successful and is surprising for several reasons. First, they created their sequence *de novo* using simple rules about secondary structure preferences and modeled the secondary structure elements based on a non-continuous $\beta\beta\alpha$ motif from Ribonuclease A. Second, the sequence bound the intended ligand. Third, the sequence showed catalytic activity(Gutte, Daumigen et al. 1979). These are goals that are still challenging today. The entire history of protein design is so large in scope that it cannot be covered fully here, but several highlights and some necessary background will be described.

After Gutte's initial success, the next major breakthroughs in protein design came from Bill DeGrado while at DuPont's Central Research and Development lab, and from Dave and Jane Richardson at Duke University.

In a series of papers beginning in 1986, DeGrado details the hierarchal process of creating a protein by first assembling independent helices into a tetramer coiled coil. He then describes linking the helices as helical hairpins. Finally, he links the helical hairpins to create a single chain protein. Initially, DeGrado and Eisenberg designed a single α-helical sequence from first principles using intuition, a healthy dose of computer graphics modeling and by some by hand CPK modeling to guide the selection of a simple sequence, α_1 (Figure 1B), that would associate as a tetramer coiled coil.

Ac-Glu-Leu-Leu-Lys-Lys-Leu-Leu-Glu-Glu-Leu-Lys-Gly

This sequence was shown to be helical by circular dichroism and indirectly to be a tetramer by titration experiments. The sequence crystallized and diffracted to 2.7Å but the structure was not solved (Eisenberg, Wilcox et al. 1986). In a second publication in

1987, two very similar sequences to the initial sequence were shown to be helical and associated as a dimer of helical hairpins, known as α_2 (Figure 1B). These designs were validated in the same manner as α_1 (Ho 1987). In 1988, Regan and DeGrado published the final paper in this series, where they created a single chain version of α_1 called α_4 (Figure 1B). The designed protein was shown to be helical, stable, and a highly cooperative folder. The structure of this protein was not determined (Regan and DeGrado 1988).

At the same time that DeGrado was approaching the creation of a helix bundle protein in a hierarchical method, the Richardson lab was attempting to create a four helix bundle protein, Felix (Figure 1C), *de novo* in a single step. The sequence was designed as part of a graduate student and faculty seminar in 1985. Felix was shown to be helical and contained a designed disulfide bond but was ultimately determined to be a molten globule(Hecht, Richardson et al. 1990). The *de novo* design, at an atomic level, of a single chain α-helical bundle is currently an unsolved challenge in protein design.

Concurrently with the design effort of Felix, the Richardson's were also creating a *de novo* β-sandwich fold, β-bellin, a dimer of four stranded β-sheets. The designed sequence went thru several iterations and was eventually shown to be soluble, primarily of β-strand secondary structure by CD, and to contain a designed disulfide bond between the two sheets of the dimer. However the designed structure resembled a molten globule in the initial stages and after several rounds of iteration required chemical linkers and modifications to obtain a structure that resembled native proteins. In a second attempt at the creation of an all β-fold, the Richardson's designed β-doublet, another four stranded dimeric β-sandwich. Again, the designed protein showed many features of native proteins but ultimately NMR data indicated the designed protein was more molten globule than native protein(Quinn, Tweedy et al. 1994). The *de novo* design of a single chain β-fold is currently an unsolved challenge in protein design.

In a landmark paper in 1997, Dahiyat and Mayo published the first example of computationally designing a protein sequence in an entirely automated fashion. The NMR solution structure of the designed protein, pda8d, was solved. The overall similarity of the design model and the solution structure is striking, with a backbone RMSD of 1.04 Å(Dahiyat and Mayo 1997). From this point forward in the history of protein design, the use of computational methods to generate and evaluate the favorability of protein sequences became more and more commonplace. Today it is almost assumed that designed sequences are generated computationally, and the use of design *manus deus*, by the hand of god (the protein designer), is often taboo or frowned upon.

In 1998, Harbury published a paper that included the use of flexible backbone protein design to create *de novo,* a novel protein fold. Almost all protein design attempts until this point used the approximation of a fixed backbone during design. Harbury attempted the design of novel undecatad coiled coils with a right-handed super helical twist. The amino acids allowed to design at the core positions (a, d and h) of the undecatad bundle were restricted to the amino acids alanine, valine, leucine, isoleucine, alloisoleucine, and norvaline. This restriction gives a total of 3993 unique core sequences for which dimer, trimer, and tetramer models were built using parametric equations that describe the bundle geometry. The calculation took eight days. Ultimately, the structure of a designed tetramer was solved by X-ray crystallography and matches the design model with 0.2 Å RMSD over the side-chain and backbone atoms which define the core of the bundle(Harbury, Plecs et al. 1998).

DeGrado continued his work in the design of helix bundle proteins by solving the crystal structure of a three-helix bundle protein, called Coiled-Ser. The structure was a trimer-coiled coil with an antiparallel arrangement between helix 1 and 2, and parallel arrangement between helix 3 and helix 1. This crystal structure was used as the design template for DeGrado's next design challenge, the creation of a single chain three-helix bundle, α_3D, using a genetic algorithm to select the core amino acid sequence. This structure was solved by NMR and was similar to the design model with 1.9 Å RMSD backbone deviation between the design model and the lowest energy member of the NMR ensemble(Walsh, Cheng et al. 1999). This result was a great success in protein design. This result is presented as a *de novo* design, but it begins with the backbone coordinates from a solved X-ray crystal structure of the protein Coiled-Ser. Coiled-Ser was created during a set of experiments to calculate the helical propensity for each amino acid to form as part of a model helix peptide. All twenty amino acids were tested at the solvent exposed F position of the helix heptad repeat in the model helix peptide. From these experiments the crystal structure of Coiled-Ser was solved(Lovejoy, Choe et al. 1993). Even though the Coiled-Ser protein does not exist in nature and the designed sequence of α_3D is not homologous to any known protein, by beginning the design process with the backbone coordinates from the Coiled-Ser crystal structure the design of α_3D is not truly *de novo*. The protein constitutes something between a redesign and a *de novo* design.

The next great success in protein design came in 2003, with the creation of another novel protein fold, Top7(Figure 1F). Kuhlman and Baker developed the protein design algorithm RosettaDesign and had previously tested it on the large scale redesign of several protein folds, the redesign of a monomer into a domain swapped dimer, and to alter the folding pathway of a protein thru loop redesign(Kuhlman, O'Neill et al. 2001; Nauli, Kuhlman et al. 2001; Kuhlman, O'Neill et al. 2002; Dantas, Kuhlman et al. 2003). RosettaDesign was built on the framework of Rosetta, a protein structure prediction software that had been successful in CASP competition(Bonneau, Tsai et al. 2001). In a method that coupled fixed backbone protein design with high-resolution structure refinement, Kuhlman created sequences in an automated fashion for a novel mixed α/β fold. Top7 was shown to be well folded, highly stable, to fold cooperatively, and ultimately its structure was solved by X-ray crystallography. The X-ray crystal structure and the design model were extremely similar with a backbone RMSD of 1.17 Å.(Kuhlman, Dantas et al. 2003)

Protein design has matured to a point where we are beginning to develop useful technologies and appear to be on the cusp of protein design being able to dramatically impact the landscape of medicinal therapeutics, industrial and chemical manufacturing, energy research, and the creation of orthogonal biochemical systems. Already several pseudo-enzymes, biosensors, and other functional proteins have been created(Jha, Leaver-Fay et al. ; Zanghellini, Jiang et al. 2006; Rothlisberger, Khersonsky et al. 2008; Murphy, Bolduc et al. 2009). However, in order for protein design to realize this future, we must develop systems that consistently produce successful designs. Computational methods have shown great promise towards increasing the success rate of protein design.

Macromolecular Modeling with Rosetta

There are several high quality protein design and modeling packages. One of the most successful of these packages is the macromolecular modeling software Rosetta. Rosetta began as a protein structure prediction method. Over the last 15 years Rosetta has grown to be a leading software for protein structure prediction, docking, and design. Rosetta is a collaborative project with greater than ten research labs contributing to its development with new methods and features constantly being developed.

The heart of Rosetta's success, especially in protein design, is the Rosetta full-atom energy function. The Rosetta energy function is a linear combination of physically based terms and statistically derived knowledge based terms from experimentally determined high-resolution protein structures(Leaver-Fay, Tyka et al.).

Rosetta Energy Terms

van der Waals Forces

The Rosetta energy function separates the attraction and repulsion of atoms due to van der Waals forces. The functional form of the energy term is a modified 12-6 Lennard-Jones potential.

Hydrogen Bonding Forces

The Rosetta energy function has terms to describe backbone – backbone hydrogen bonds, backbone – side-chain hydrogen bonds, and side-chain – side-chain hydrogen bonds. The terms are separated to give precedence to backbone – backbone hydrogen bonds to ensure that the formation of secondary structure is not disturbed during design or refinement. The functional form of the hydrogen bond energy term is a summation of the deviation of the distance between the acceptor atom and the donated hydrogen, the angle between the acceptor, the hydrogen and the donator, and the angle defined by the hydrogen the acceptor and its connected carbon neighbor.

Solvation

Rosetta uses a semiemperical solvation model developed by Lazaridis and Karplus(Lazaridis and Karplus 2000). The method requires the distance between two atoms, the volume of the atoms, and a set of reference energies for all atom types parameterized from known protein structures. The method is useful for protein design because it avoids the computationally expensive step of calculating the interaction of a protein's surface with explicit solvent.

Electrostatics

Rosetta treats electrostatics in a pair-wise fashion from the observed probability of two polar amino acids to be near each other in the protein databank.

Backbone Torsion Term

The preference for certain amino acids for particular phi and psi angles is well known. Rosetta models these preferences by the probability of observing an amino acid at a particular phi and psi value with a particular type of secondary structure, helix, strand or loop, within the protein databank.

Side-chain Torsion Term

The side-chain torsion preferences of amino acids are also well known and particular conformations of chi angles are called rotamers. Rosetta models the favorability of rotamers for each amino acid based on the probability of that amino acid being observed in the protein databank in a particular phi and psi bin. The rotamer probabilities are taken from the rotamer library created by Dunbrack(Dunbrack 2002).

Probability of an Amino Acid

Rosetta also incorporates a unique term that attempts to capture the secondary structure and neighbor preferences for each amino acid, amounting to an environmental term. The statistics for this term were determined from the protein databank.

Reference energy

To mimic the natural distribution of amino acids observed in the protein databank, each amino acid is given a reference energy. This term is basically a constant for each amino acid that modifies the rate at which amino acids are accepted during design.

Rosetta also incorporates many other energy terms and has different score functions for coarse-grained modeling. These terms will be discussed as needed.

Design Challenges Attempted

Examination of the history of protein design shows that there are two main model systems, the α-helical bundle and the β-sandwich. Each system has a tendency to fail in a different manner. Helix bundles are prone to form molten globules or to fold/associate in an undesired target state. β-sandwiches are more likely to aggregate, making it nearly impossible to rescue a failed design. We begin our research in protein design by first focusing on helix bundles and then expanding to a diverse set of protein folds including β-sandwiches. The scope of this research ultimately attempts to address questions concerning the designability of protein backbones, for both protein redesign and *de novo* design.

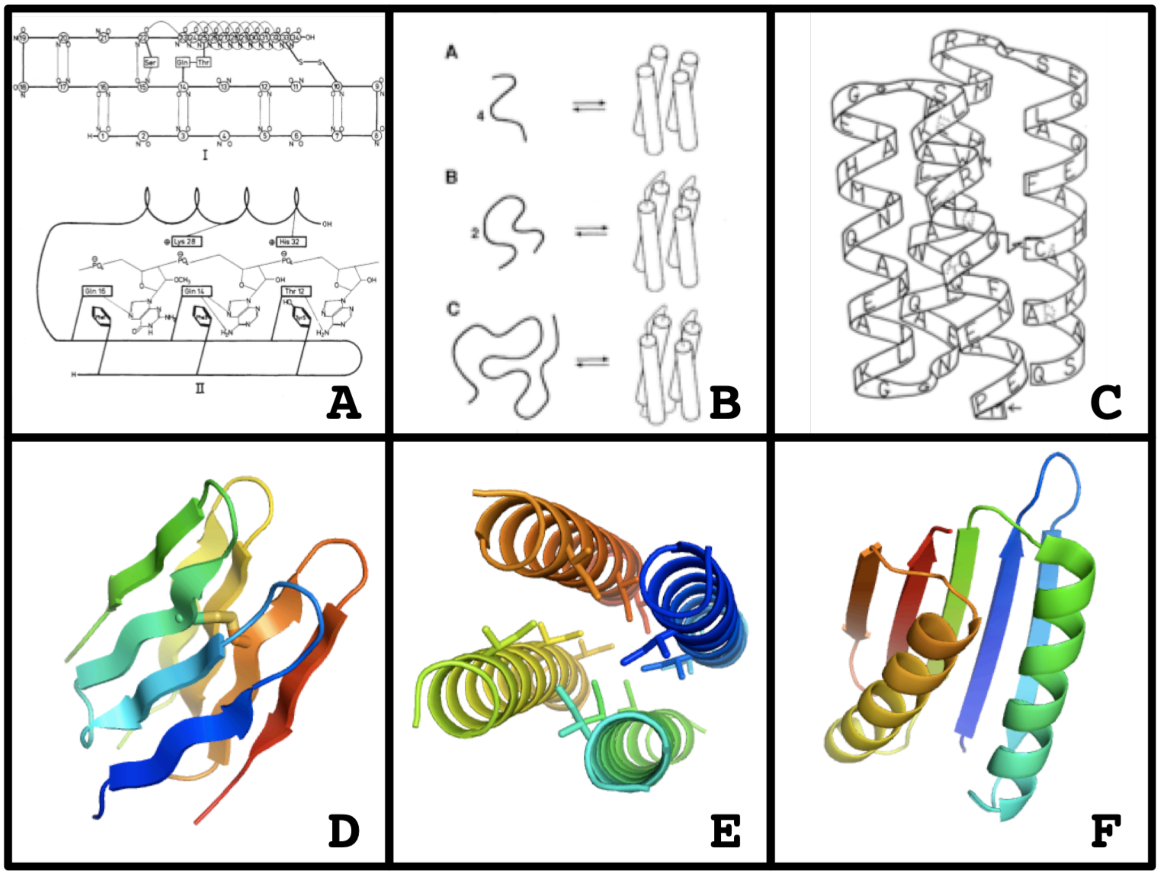
We develop new computational methods that attempt to increase the success rate of protein design by focusing on the bottlenecks that inhibit the rapid generation of high-quality designed protein sequences, whether they are redesigned from nature or they are completely *de novo*. We test these computational methods experimentally and give biophysical and structural evidence for the success of our computational methods. Our computational and experimental techniques are divided into methods that address the protein redesign problem and the *de novo* design problem. Ultimately, we used similar techniques in both challenges.

We first investigate the available sequence space during the redesign of naturally occurring proteins in Chapter Two, by incorporating different levels of backbone flexibility during design and by explicitly eliminating the possibility of designing the native sequence. In Chapter Three, we tackle the challenge of complete *de novo* design for a diverse set of protein folds, both novel and naturally occurring, in an automated fashion. The techniques developed here begin to blur the distinction between protein redesign and *de novo* design. In Chapter Four, we discuss the future of protein design and suggest possible routes forward to solve current and future design challenges.

Figures

Figure 1.1: Important moments in the history of protein design.

In ‘79 Gutte attempted what is considered the first protein design, successfully creating a minimal Ribonuclease A (A). Over the course of several years DeGrado et al. successfully created a single chain *de novo* four-helix bundle in a hierarchal method (B). The Richardson’s attempted the *de novo* design in a single step of a four-helix bundle and a dimer four-stranded β-sandwich, their results were a partial success (C & D are design models). Harbury successfully used flexible backbone design to create a novel undecatad tetramer coiled coil with a right-handed super helical twist (E is an experimental structure). Kuhlman successfully *de novo* designed a novel fold called Top7 (F is an experimental structure).



References

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

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.

Dunbrack, R. L., Jr. (2002). "Rotamer libraries in the 21st century." Curr Opin Struct Biol **12**(4): 431-40.

Eisenberg, D., W. Wilcox, et al. (1986). "The design, synthesis, and crystallization of an alpha-helical peptide." Proteins **1**(1): 16-22.

Gutte, B., M. Daumigen, et al. (1979). "Design, synthesis and characterisation of a 34-residue polypeptide that interacts with nucleic acids." Nature **281**(5733): 650-5.

Gutte, B. and R. B. Merrifield (1969). "The total synthesis of an enzyme with ribonuclease A activity." J Am Chem Soc **91**(2): 501-2.

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

Hecht, M. H., A. Das, et al. (2004). "De novo proteins from designed combinatorial libraries." Protein Sci **13**(7): 1711-23.

Hecht, M. H., J. S. Richardson, et al. (1990). "De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence." Science **249**(4971): 884-91.

Ho, S. P. a. D., W. F. (1987). "Design of a 4-Helix Bundle Protein: Synthesis of Peptides Which Self-Associate into a Helical Protein." J Am Chem Soc: 6751-6758.

Jha, R. K., A. Leaver-Fay, et al. "Computational design of a PAK1 binding protein." J Mol Biol **400**(2): 257-70.

Kendrew, J. C. and M. F. Perutz (1957). "X-ray studies of compounds of biological interest." Annu Rev Biochem **26**: 327-72.

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.

Lazaridis, T. and M. Karplus (2000). "Effective energy functions for protein structure prediction." Curr Opin Struct Biol **10**(2): 139-45.

Leaver-Fay, A., M. Tyka, et al. "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules." Methods Enzymol **487**: 545-74.

Lovejoy, B., S. Choe, et al. (1993). "Crystal structure of a synthetic triple-stranded alpha-helical bundle." Science **259**(5099): 1288-93.

Merrifield, R. B. (1965). "Solid-Phase Peptide Syntheses." Endeavour **24**: 3-7.

Mullis, K., F. Faloona, et al. (1986). "Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction." Cold Spring Harb Symp Quant Biol **51 Pt 1**: 263-73.

Murphy, P. M., J. M. Bolduc, et al. (2009). "Alteration of enzyme specificity by computational loop remodeling and design." Proc Natl Acad Sci U S A **106**(23): 9215-20.

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

Pauling, L. and R. B. Corey (1951). "Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets." Proc Natl Acad Sci U S A **37**(11): 729-40.

Pauling, L., R. B. Corey, et al. (1951). "The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain." Proc Natl Acad Sci U S A **37**(4): 205-11.

Perutz, M. F., M. G. Rossmann, et al. (1960). "Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis." Nature **185**(4711): 416-22.

Quinn, T. P., N. B. Tweedy, et al. (1994). "Betadoublet: de novo design, synthesis, and characterization of a beta-sandwich protein." Proc Natl Acad Sci U S A **91**(19): 8747-51.

Regan, L. and W. F. DeGrado (1988). "Characterization of a helical protein designed from first principles." Science **241**(4868): 976-8.

Regnier, F. E. (1983). "High-performance liquid chromatography of biopolymers." Science **222**(4621): 245-52.

Rose, W. C. (1968). "II. The sequence of events leading to the establishment of the amino acid needs of man." Am J Public Health Nations Health **58**(11): 2020-7.

Rothlisberger, D., O. Khersonsky, et al. (2008). "Kemp elimination catalysts by computational enzyme design." Nature **453**(7192): 190-5.

Walsh, S. T., H. Cheng, et al. (1999). "Solution structure and dynamics of a de novo designed three-helix bundle protein." Proc Natl Acad Sci U S A **96**(10): 5486-91.

Watson, J. D. and F. H. Crick (1953). "Genetical implications of the structure of deoxyribonucleic acid." Nature **171**(4361): 964-7.

Williamson, M. P., T. F. Havel, et al. (1985). "Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry." J Mol Biol **182**(2): 295-315.

Zanghellini, A., L. Jiang, et al. (2006). "New algorithms and an in silico benchmark for computational enzyme design." Protein Sci **15**(12): 2785-94.

Chapter 2

Increasing Sequence Diversity with Flexible Backbone Protein Design: The Complete Redesign of a Protein Hydrophobic Core

Grant S. Murphy[1], Jeffery L. Mills[2,3], Michael J. Miley[4], Mischa Machius[4], Thomas Szyperski[2,3] and Brian Kuhlman[5*]

[1]Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-3290, USA

[2]Department of Chemistry, State University of New York at Buffalo, Buffalo, NY, 14260, USA

[3]Northeast Structural Genomics Consortium

[4]Center for Structural Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA

[5]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-7260, USA

[*]corresponding author. Email: bkuhlman@email.unc.edu

Phone: 919-843-0188

Fax: 919-966-2852

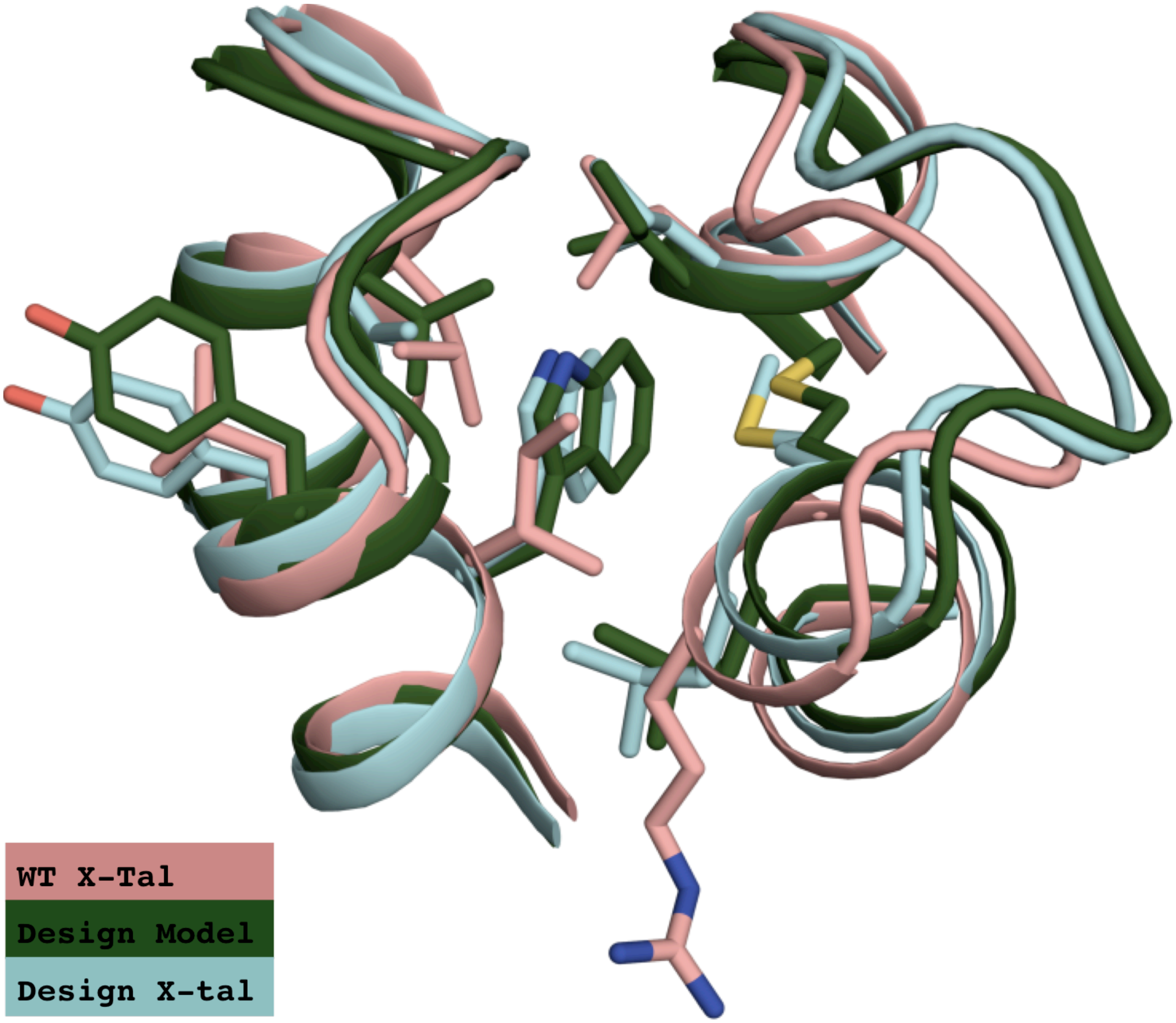
Running title: Complete Redesign of a Protein Hydrophobic Core

Summary

Protein design is a rigorous test of our understanding of protein stability and structure. Successful design methods should allow us to explore regions of sequence space not found in nature. However, when redesigning naturally occurring protein structures most fixed backbone design algorithms return amino acid sequences that share strong sequence identity with wild-type sequences, especially in the protein core. This places a restriction on function space that can be explored and is not consistent with observations from nature, where sequences of very low identity have similar structures. Here, we allow backbone flexibility during design to mutate every position in the core of a four-helix bundle protein (38 residues). In general, only small perturbations to the backbone, 1-2 Å, were needed to dramatically repack the core. The redesigned protein is exceptionally stable (melting point > 140°C) and an NMR structure and an X-ray crystal structure show the side chains and backbone were accurately modeled (all-atom RMSD = 1.3 Å).

Keywords: Computational Protein Design, *de novo* Protein Design, Flexible Backbone Design

Graphical Abstract



Highlights

- Flexible backbone design has been used to mutate every position in a protein core

- The redesign is a hyperthermophile (melting temperature > 140°C).

- An NMR structure and an X-Ray structure closely match the design model.

- Designed backbone perturbations are accurately recapitulated in the structures.

Introduction

A primary goal of protein design is to create proteins that have sequences, structures and functions not found in nature. This can be accomplished by designing new protein structures from scratch or by modifying sequences and structures of proteins found in nature. The second approach is appealing because in many cases it should be more likely to succeed, and it is the approach nature typically uses to evolve new functional proteins. There are many examples of naturally occurring protein pairs that are structurally homologous(have the same fold), but have different functions and low sequence identity (< 15%). Recapitulating or expanding on this sequence diversity by design, however, is not straightforward. Most computational methods for protein design are built on side chain optimization algorithms that work most efficiently with a fixed protein backbone (Gordon, Marshall et al. 1999). When redesigning naturally occurring proteins with these methods, the computationally optimized sequences often closely resemble the native sequence, especially in the protein core, where >60% sequence identity is common(Desjarlais and Handel 1999; Kuhlman and Baker 2000; Pokala and Handel 2001). It is clear from these studies and from the structural analysis of naturally occurring homologs that to expand sequence diversity it is necessary to allow perturbations to the protein backbone. Even small changes to the backbone (2 Å), can open large regions of sequence space (Yin, Ding et al. 2007). The challenge for protein designers is identifying backbone and sequence perturbations that are energetically favorable.

A variety of strategies have been developed for performing protein design with backbone flexibility (Grigoryan and Degrado ; Su and Mayo 1997; Desjarlais and Handel 1999; Dantas, Corrent et al. 2007; Georgiev and Donald 2007; Friedland, Linares et al.

2008; Fung, Floudas et al. 2008; Apgar, Hahn et al. 2009; Davis, Raha et al. 2009; Havranek and Baker 2009; Mandell and Kortemme 2009), however, few have been experimentally validated with high-resolution structures of the design model(Correia, Ban et al. ; Sammond, Bosch et al. ; Harbury, Tidor et al. 1995; Harbury, Plecs et al. 1998; Kuhlman, O'Neill et al. 2002; Kuhlman, Dantas et al. 2003; Hu, Wang et al. 2007; Murphy, Bolduc et al. 2009). Perhaps the most tested approach has been iterative rounds of sequence optimization and backbone refinement with the molecular modeling program Rosetta. Sequence optimization is performed using a simulated annealing protocol that searches for low energy combinations of side-chain rotamers. Structure refinement uses Monte Carlo sampling of small backbone torsion angle perturbations coupled with gradient-based minimization of dihedral angles. Both stages of optimization use an energy function that rewards tight packing, commonly observed side chain and backbone torsion angles, favorable hydrogen bond geometries and low energies of desolvation. This approach has been used to design a protein from scratch, design a protein-binding peptide and design new protein loop conformations (Kuhlman, Dantas et al. 2003; Dantas, Corrent et al. 2007; Hu, Wang et al. 2007). In this study, we explore if iterative optimization of sequence and structure with Rosetta can be used to aggressively redesign an entire protein core.

Our specific goal was to mutate every residue in the core of the four-helix bundle protein, CheA phosphotransferase, while maintaining the overall fold and stability of the protein (Figure 2.1A). Several *de novo* design and redesign projects have focused on helix bundle proteins(Hecht, Richardson et al. 1990). From these studies, it is evident that many sequences will adopt collapsed helical structures as long as the amphipathic nature of the helices is preserved and the sequence has significant helical propensity (Kamtekar, Schiffer et al. 1993; DeGrado and Nilsson 1997). What is more challenging to design are sequences that adopt a specific pre-determined structure and show characteristics of natural helix bundle proteins, such as cooperative thermal unfolding. Many previously reported helical bundle designs formed an ensemble of collapsed conformations, called a molten globule. In cases where the structure for a design was experimentally determined, it often did not agree with the initial design model (Lovejoy, Choe et al. 1993; Hill and DeGrado 2000; Willis, Bishop et al. 2000). One striking success story is the accurate *de novo* design of a four-helix coiled-coil with a right-handed super-helical twist(Harbury, Plecs et al. 1998). A key component of this work was optimization of packing energies via backbone refinement as well as sequence design. Here, we show that flexible backbone design can be used to perturb the structure and sequence of a pre-existing protein with atomic-level accuracy.

Results

Core Redesign of the CheA Four Helix Bundle

The four-helix bundle CheA phosphotransferase was chosen as the design template (pdbcode: 1tqg), because of its simple up-down helix bundle topology and its moderate size of 105 amino acids. Thirty-eight positions from the CheA X-ray crystal structure were identified as being completely or partially buried and were targeted for redesign (Figure 2.1A and Figure 2.2). Our initial hypothesis, based on previous protein redesign experiments, was that the protein backbone would need to be perturbed in order to completely redesign the protein core. To test this hypothesis, four different computational procedures were used to generate designed sequences: (1) fixed backbone design with all amino acids allowed at each design position (FBAA), (2) fixed backbone design with the native amino acid disallowed at each design position (FBNN), (3) flexible backbone design with all amino acids allowed at each position (DRAA) and (4) flexible backbone design with the native amino acid disallowed at each design position (DRNN). In the naming scheme, FB stands for fixed backbone, DR stands for the flexible backbone design strategy of design followed by refinement of the backbone, AA states that all amino acids were allowed during design and NN indicates that only non native amino acids were allowed during design.

The fixed backbone protocol used Rosetta's standard rotamer optimization protocol, which uses Monte Carlo sampling of backbone dependent side chain rotamers to search for low energy sequences. The flexible backbone protocol used the same sequence optimization algorithm, but iterated sequence optimization with high-resolution backbone refinement using Monte Carlo sampling and gradient-based minimization of backbone torsion angles. Backbone perturbations with this protocol are generally modest, 1-2 Å. 25,000 independent trajectories were run for each protocol. As anticipated, in the two cases where all amino acids were allowed, FBAA and DRAA, the flexible backbone procedure generated sequences with lower sequence identity, ranging from 10 to 50%, compared to traditional fixed backbone design experiments at 30 to 60%. To evaluate packing density in the redesigns the RosettaHoles algorithm was used (Sheffler and Baker 2009). RosettaHoles explicitly searches for small voids in the protein that are inaccessible to water, and assigns a score to each residue between 0 and 1 that reflects packing around that residue. RosettaHoles scores closer to 1 indicate fewer voids. Residues in high-resolution crystal structures generally have scores between 0.5 and 1.0 for the entire protein. Models generated with the FBAA and FBNN protocols had RosettaHoles scores between 0.2 and 0.3 for the core residues, while the DRNN and DRAA models had scores between 0.4 and 0.5.

For each of the four protocols, a single sequence was selected for experimental validation (Figure 2.2). Sequences were selected for experimental testing based on their total Rosetta energy, the quality of packing, correct predicted secondary structure, performance in *ab initio* folding experiments and deviation from the wild-type sequence (see methods for more details). In the case of FBAA, most of the sequences were not considered because they had >40% sequence identity with the wild-type sequence. For comparison, Figure 2.2 also show the lowest scoring sequence generated with the FBAA protocol, labeled TRAD. It has 61% identity with the wild-type sequence in the core of the protein.

The computational experiments that incorporated flexible backbone design show subtle but important backbone movement (Figure 2.1B, 2.1C, 2.1D, Figure 2.3 and Supplemental Figures 3, 4, and 5). The designed sequence, DRNN, and the DRNN design model are the most varied from the native sequence and CheA crystal structure (Figure 2.1B, 2.1C and 2.1D) and will be used to illustrate the types of backbone changes due to flexible backbone design. The final DRNN design model has a backbone rmsd of 1.57 Å compared to the CheA crystal structure (conformation A). The largest backbone deviations between the design model and the crystal structure are seen in loop 3, helix 1, and helix 4. Although its sequence was not varied, loop 3 is pushed away from the center of the helix bundle because of the incorporation of a tryptophan at position 39, previously an isoleucine. Using a global alignment, the backbone rmsd of loop 3 to the wild-type protein is 1.88 Å and the all atom rmsd is 2.88 Å. Helix 1 is perturbed by 1.85 Å and helix 4 is perturbed by 2.08 Å (Figure 2.1C and 2.1D). The sequence identity of the 38 designed core residues is 0% compared to the native CheA and the total sequence identity is 57.14%. A diverse set of mutations were predicted for the 38 core design positions, 27 mutations were hydrophobic/aromatic residues mutated to different hydrophobic/aromatic residues, 6 mutations were hydrophobic/aromatic residues mutated to polar residues, 3 mutations were polar residues mutated to hydrophobic/aromatic residues, and 2 mutations were polar amino acids mutated to polar amino acids. The definition of core used in this study is broader than typically seen, but all of the positions chosen for design were greater than 50% buried in the wild-type template, and made significant contacts with residues, that were completely buried.

Protein Expression and Behavior

Three of the designed proteins, FBAA, DRAA and DRNN expressed in the soluble cell fraction at a variety of induction temperatures, 16°C-37°C, and produced greater than 50 mgs of cleaved purified protein per 1.5 liters of culture. The proteins eluted as single peaks from a gel filtration purification step with molecular weights consistent with the expected monomer weights, ~14KD. In contrast, FBNN was only found in the insoluble fraction of the cell pellet. This behavior was seen at all tested temperatures and IPTG induction concentrations.

Biophysical Characterization of Redesigned CheA

Far-UV circular dichroism experiments confirmed that the designed proteins are primarily α-helical, with strong minima present at 220 nm and 208 nm (Figure 2.4A and Supplemental Figures 2.1 and 2.2). It was not possible to unfold two of the designed proteins (FBAA, DRNN) during standard thermal denaturations from 4°C to 97°C (Figure 2.4B and Supplemental Figure 2.1). Chemical denaturation with guanidine chloride (GuCl) shows that the designed proteins undergo highly cooperative unfolding events (Figure 2.4C and Supplemental Figures 2.1 and 2.2). To determine accurate values for m, Tm, ΔH°, ΔCp°, and ΔG°, a Gibbs-Helmholtz surface was constructed by fitting several thermal denaturations with varying amounts of GuCl to the Gibbs-Helmholtz equation modified to consider the effect of denaturant concentration (Table 1, Figure 2.4D, 2.4E and Supplemental Figures 2.1 and 2.2). The designed proteins are hyperthermophiles with Tm values between 96°C and 142°C and ΔG° unfolding values between 5.5 and 16.2 kcal/mol. The most ambitious design, DRNN, was the most stable. For comparison, the wild-type protein has a ΔG° of unfolding of 3.5 kcal/mol and a Tm of 91°C. The designed proteins have elevated values for ΔCp° ranging from 0.83 to 1.1 kcal/mol•deg, the expected value based on changes in solvent accessible surface area is ~0.5 kcal/mol•deg and the wild-type value is 0.61 kcal/mol•deg(Myers, Pace et al. 1995). The ΔH° values range from 63 to 128 kcal/mol and the m values range from 1.9 to 3.4 kcal/(mol•M), the wild-type protein has values of 41 kcal/mol and 1.4 kcal/(mol•M).

X-ray Crystal Structure of DRNN

The structure of the designed protein DRNN was determined to 1.85 Å by X-ray crystallography, with R_{free} 0.23 and R_{work} 0.19. A strong molecular replacement solution was found using the design model with all side-chain atoms removed. The initial round of refinement was very encouraging. The $2F_o$-$2F_c$ difference density map clearly identified several of the designed amino acids, such as Tryptophan 39 (Figure 2.5A). During refinement, an amino acid side-chain was only built into the model if the electron density strongly indicated which rotamer(s) were present in the crystal. It was possible to place rotamers for all of the 38 core design positions in this manner, many of the surface and loop residues were also assigned in this manner. Finally, after all amino acids with complete backbone and side-chain density were built, the auto-fit and refine methods in RefMac (Cowtan, Emsley et al.) and Coot (Emsley, Lohkamp et al.) were used to build the handful of remaining side-chains. The final stages of refinement included TLS parameters as well as building waters and relaxing geometric constraints. The solved structure scores well in the metrics tested by the molprobity server and also ranks in the ~95th percentile for RosettaHoles packing score, 0.64, in the 1.0-2.0 resolution range (Figure 2.5B-F).

There is very good agreement between the DRNN design model and the experimental structure (Figure 2.6). The all atom rmsd between the design model and chain A and chain B of the experimental structure are 1.5 Å and1.3 Å respectively. The 38 core design positions were predicted with good accuracy, 34 of 35 positions were observed in the correct rotamer state, only valine 29 was observed in a different rotamer from the design model for reasons that could not be explained by crystal contacts. Three design positions were observed in different rotamer states, Y37, K90 and K92. These differences are explained by crystallographic contacts (K90), or hydrogen bonding with crystallographic waters (Y37 and K92) that were not included in the design model. The prediction of the backbone of loop 3, which was extensively remodeled is also highly accurate, 0.32 Å and 0.38 Å over backbone atoms for chains A and B. The Rosetta high-resolution refinement protocol has been shown to accurately model the atomic level interactions present in protein crystal structures(Raman, Vernon et al. 2009). The high degree of accuracy between the design model and the experimental structure is likely a result of the high-resolution refinement protocol. Consider for instance if the

RosettaDesign algorithm generated a poor sequence, the high-resolution refinement protocol would be unlikely to find a low energy backbone conformation in nearby conformation space without destabilizing the fold. If the fold were destabilized to accommodate a poor sequence, this would be reflected by a decrease in the total energy, in the quality of packing, or in the number of unsatisfied polar atoms. Failure to pass these metrics would be used to eliminate the poor sequence and its conformation from possible experimental selection.

NMR Structure of DRNN

The 2D ^{15}N-HSQC of DRNN is consistent with a well-folded protein and the fingerprint region of the spectrum has good dispersion despite being an all-helical protein (Figure 2.7A). The NMR structure of DRNN was solved using 1484 constraints and is in good agreement with the design model. The backbone rmsd between the DRNN design model and the first member of the NMR ensemble is 2.28 Å. The largest backbone deviations between the design model and the NMR ensemble are the N-termini of helix 1 and C-termini of helix 4 (Figure 2.7B). The backbone rmsd over residues 15-105 is 1.2 Å, between the design model and the first member of the NMR ensemble. Comparing the side-chain rotamers of the NMR ensemble to the design model, 31 of the core designed positions agree with the design model in at least one of the twenty members of the NMR ensemble. The designed rotamer was not observed in the NMR ensemble for the remaining seven core design positions L15, L18, T53, I68, L71, L76, and L94. The similarity of the NMR ensemble compared to the design model on a global level and for the 31 positions where the design rotamer was observed is striking, the region surrounding tryptophan 39 is an excellent example, the all atom rmsd of the 19 amino acids which are neighbors of tryptophan 39 is 1.35 Å (Figure 2.7C and Supplemental Figure 2.6). While, the NMR solution structure has a larger RMS deviation from the design model than the X-ray crystal structure, the design model satisfies >93% of the NMR constraints. This suggests that differences between the NMR solution structure, the design model and the X-ray crystal structure may be due to either the NMR structure being underdetermined or perhaps highlights differences in the conformational space searched and the potentials used during NMR refinement vs. X-ray refinement and computational refinement.

Comparison of the DRNN NMR Structure, DRNN X-ray Crystal Structure and the DRNN Design Model

The structural agreement between the design model, the NMR ensemble and the X-ray crystal structure is amazing. It is interesting to note that valine 29, the only design position where the predicted rotamer was not observed in the crystal structure, is consistent between the NMR ensemble and the DRNN design model. The design model and the x-ray crystal structure both satisfy > 93% of the NMR constraints used to solve the NMR structure. CS-Rosetta was also used to solve the NMR structure of DRNN(Shen, Vernon et al. 2009). The CS-Rosetta structure is a much closer match to the design model and to the crystal structure with a backbone rmsd to the design model of 1.0 Å.

Another metric to compare the design model, the NMR structure and the X-ray crystal structure is a Global Distance Test (GDT). This metric can show how structures deviate or become similar with increasing bounds of rms/distance cutoffs. The comparison of the template structure to the design model and the NMR and X-ray structures is of particular interest (Figure 2.8). The design model is a closer match in GDT space to the X-ray crystal structure than either the NMR structure or the wild-type template.

Discussion

Previously, in a large-scale test of protein redesign using Rosetta, an all α-fold and an all β-fold were redesigned, experimentally tested and shown to be less stable and/or less folded than their wild-type templates(Dantas, Kuhlman et al. 2003). Hu et al. successfully redesigned an all β-fold using Rosetta by focusing on reproducing natural β-sandwich sequence propensities. The X-ray crystal structure and the biophysical data showed that redesign to be well folded and more stable than the wild-type β-sandwich template(Hu, Wang et al. 2008). Here we show that the redesign of stable well-folded all α-folds has been achieved using Rosetta, specifically the redesign of an up-down four-helix bundle. We show that it is possible to dramatically redesign the hydrophobic core of a protein to various degrees of sequence identity using both fixed backbone and flexible backbone design strategies (Figure 2.2). This did not require any modifications to the Rosetta energy function. We also demonstrate that flexible backbone design is necessary to generate high quality backbones for sequences that are highly dissimilar from the template protein's sequence.

As evidence for the necessity of flexible backbone design to explore highly divergent sequences with accurate backbone modeling, we present the fact that FBNN designed sequences were either not expressible, highly unstable, or prone to expression only in inclusion bodies. This was somewhat expected, FBNN was indented as a control to show and test that (1) sequences generated with a fixed backbone under stringent sequence restrictions would have less favorable Rosetta energy (FBNN has the worst energy of the designs tested) and (2) that experimentally we would observe FBNN to be the least stable design and (3) that this stringent sequence restriction could be rescued by allowing flexible backbone design. In comparison, our other design experiments which were either not as stringent in sequence space or incorporated backbone flexibility resulted in stable well-folded sequences.

Our most extreme redesign, DRNN, was predicted to an amazing level of atomic accuracy, with a backbone rmsd of the 0.8 Å and an all-atom rmsd of 1.3 Å (Chain B) (Figure 2.6 and Figure 2.9). The accuracy of the model is even more striking when considering the 38 core designed residues, which have an all-atom rmsd of 0.95 Å, combined with the fact that 34 designed rotamers were correctly predicted, 3 positions are in different rotamers but are involved in crystal contacts not in the design model, and only a single position, valine 29, is observed in a rotamer not predicted by the design model.

It is interesting to note however that in the NMR structure ensemble valine 29 occupies the same rotamer as the design model. The region around Valine 29 is predicted with high accuracy compared to the design (Figure 2.7C and Supplemental Figure 2.6).

Conclusions

These results are compelling evidence that the Rosetta all-atom energy function and conformational search methods capture a large portion of the physical chemistry and physics responsible for protein stability and for the perturbations observed in backbone and side-chain reorganizations caused by mutations observed in wild-type proteins. As additional evidence for this, one may consider that the design of DRNN in some respects is highly similar to the process taken for the structure prediction of structural homologs with low sequence identity. This is a problem that Rosetta has been successful at in CASP events.

We present here that it is possible to accurately predict the backbone and side-chain conformations for aggressively redesign proteins. Here we mutated only the core residues of an already thermophilic protein and achieved an increase in thermostability of > 50°C and an increase in ΔG° of 14 kcals/mol.

The core redesign strategy maybe especially useful for the stabilization of enzymes, ligand binding proteins, and protein-protein interface partners where preservation of a functional surface or pocket is important. Designing only core residues, it need not be as aggressive as DRNN considering that FBAA was equally stable, can improve thermal and chemical stability with the possibility of retaining or modulating biological activity.

Materials and Methods

Computational Methods

Fixed Backbone Protein Design Protocol

The fixed backbone protein design protocol used here is the standard fixed backbone design protocol released with Rosetta3.3. The design protocol consists of a side-chain packing algorithm, which uses simulated annealing to search rotamer space, using rotamers from the Dunbrack rotamer library and the Rosetta energy function to evaluate the fitness of sequences(Leaver-Fay, Tyka et al.).

Flexible Backbone Protein Design Protocol

The redesign sequences were generated using a new protocol within the Rosetta framework. The protocol has two stages, fixed backbone sequence design and fixed sequence backbone and side-chain dihedral optimization. The protocol iterates between these two stages until the energy difference between cycle i and cycle i-1 is less than 1.0 Rosetta Energy Units (REU), in practice this is ~5 redesign simulations for proteins between 100 and 200 residues. The fixed backbone sequence design step is the standard Rosetta side-chain packing algorithm described above and elsewhere. The fixed sequence backbone and side-chain dihedral optimization is the Rosetta structure optimization protocol used in structure prediction and refinement.

Computational Protein Design Experiments

Four different types of computational experiments were performed: (1) fixed backbone design where all amino acids were allowed at design positions (FBAA), (2) fixed backbone design where the native amino acid was not allowed at design positions (FBNN), (3) flexible backbone where all amino acids were allowed at design positions (DRAA) and

(4) flexible backbone design where the native amino acid was not allowed at design positions (DRNN).

Core Redesign of CheA Four-Helix Bundle

To redesign the core residues of the CheA four-helix bundle, 38 positions were identified as buried or partially buried positions. These positions have at least 15 neighbors within 10 Å, where a neighbor is defined by the distance between Cβ atoms on residues i and j. Positions identified as core residues were visually inspected to remove any non-buried surface positions with a high number of neighbors. During this visual inspection, all attempts were made to include all partially buried side-chain positions, excluding positions identified as being in a loop by the DSSP algorithm(Kabsch and Sander 1983). During the design stage, the 38 designable core positions were allowed to change amino acid identity as described for each type of protein design experiment. An additional seven surface positions were allowed to design and mutate to any amino acid identity. The remaining 60 positions were not allowed to change amino acid identity but were free to change rotamer state. The possible rotamer states for each amino acid type are taken from the Dunbrack backbone dependant rotamer library(Dunbrack 2002). The 38 core designable positions were given more rotamer freedom, allowing additional sampling of rotamer states, the side-chain chi angles where given 12 extra rotamer states at ± 0.25, 0.50, 0.75, 1.00, 1.25, and 1.50 standard deviations from the most favorable dihedral angles for each rotamer. The seven designable and 60 surface positions were given extra rotamer states at ± 0.5 and 1.0 standard deviations from the most favorable rotamer states. All positions were free to sample phi, psi, omega, and all dihedral chi angles during backbone and side-chain perturbation and minimization. A total of 25,000 design simulations were performed for each computational protein design experiment.

Selection of Designed Sequence for Experimental Characterization

The 25,000 designed sequences were ranked by their quality of core packing, as measured by RosettaHoles, sequences with scores less than 0.5 (0.4 for FBAA and FBNN) were pruned(Sheffler and Baker 2009). Sequences where the core design positions were predominately of a single amino acid type, greater than 50%, were pruned. This filter eliminates sequences where the protein core is composed primarily of only a few amino acids types, mostly alanine and leucine. The 50 lowest scoring models, based on total Rosetta energy, were evaluated for their secondary structure propensities using the secondary structure prediction server JPRED3(Cole, Barber et al. 2008). All 50 design models were predicted to have similar secondary structures compared to the design model and the native CheA. The ten lowest energy models were subjected to structure prediction using Rosetta's structure prediction method. This filter evaluates if the designed sequence is predicted to adopt the desired fold, all designed sequences recovered the desired fold. The ten lowest energy sequences for each experiment were evaluated by eye and one sequence from each experiment was chosen for experimental characterization. It is interesting to note that the sequence chosen from the DRNN experiment was also the lowest scoring sequence out of the 25,000 designed sequences generated in that experiment.

Experimental Methods

Protein Expression and Purification

A codon optimized gene for each designed sequence, and a modified version of the wild-type CheA was purchased from Genscript, lowercase letters are due to cloning and capital letters are the designed sequences. Each gene was supplied as 4 μg of lyophilized DNA in puc57 vector. The gene of interest was pcr amplified out of the parent vector, purified using a pcr clean up kit from Fermentas, double digested with NdeI and XhoI from NEB, and purified again using a pcr clean up kit, and finally ligated into Pet21b vector from Novagen which had been prepared by double digesting with NdeI and XhoI and using a Fermentas gel extraction clean up kit. The ligation reaction was transformed into XL-10 Gold cells from Stratagene.

```
> 1TQG_MOD_WT
mGSHQEYLQQFVDETKEYLQNLNDTLDELEKNPEDMELINEAFRALHTLKEMAETMGFSSMAKL
CHTLENILDKARNSEIKITSDLLDKIKDGVDMITRMVDKIVS
gsylvprgslehhhhhh*

>FBAA
mGSHQEYLQKFADEAKELLQNINDFLKELEKNPEDMEMINKVLRAFHTLKELAETMGFSSMAKM
AHTAANLADKAANSEIKITSDLLDKLKDMADMLTRFVDKLVS
gsylvprgslehhhhhh*

>FBNN
mGSHQEYIQKVADELKEHFQNINDFIKEMEKNPEDMEKVNKIQREFHTAKEIFETMGFSSAAKI
AHTAHNLADKSSNSEIKITSDLIDKLKDYADMLTRFMDKLVS
gsylvprgslehhhhhh*

>DRAA
mGSHDEYRKKAADELKELLQNINDVLDELEKNPEDMEKINKAQRLFHTIKDKAQTMGFSSAAKY
AHTGENIADKAANSEIKITSDLLDKLKDYADMITRELDKYVS
gsylvprgslehhhhhh*

>DRNN
mGSHQEYIKKVTDELKELIQNVNDDIKEVEKNPEDMEYWNKIYRLVHTMKEITETMGFSSVAKV
LHTIMNLVDKMLNSEIKITSDLIDKVKKKLDMVTRELDKKVS
gsylvprgslehhhhhh*
```

Each protein was expressed in BL21 (DE3) pLysS cells from Stratagene. Cells were grown in LB media with 100 mg/ml ampicilin at 37°C to an OD_{600} of 0.6 and induced with

0.5 mM IPTG for 12 hours at 16°C. Cells were centrifuged at 4500 x g for 30 minutes and cell pellets were resuspended in 0.5 M NaCl, 0.2 M NaK pH 7.0, 10% glycerol, 1% triton, dtt, and treated with dnase, rnase, benzamidine, and pmsf after three rounds of sonication using a sonicator set to 70% power for 45 seconds. The cell lysate was cleared twice by centrifugation at 18,000 x g for 30 minutes. The supernatants were then filtered using a 0.22 μM filter from Millipore. The supernatant was purified using a HisTRAP from GE Healthcare. The elution was concentrated to 2 mls and further purified on a Superdex S75 gel filtration column.

Circular Dichroism

CD data were collected on a Jasco J-815 CD spectrometer. Far-UV CD scans were collected using a 1 mm cuvette at concentrations between 10-20 μM protein in 50 μM sodium phosphate at pH 7.4 and 20°C. Thermal denaturation of samples was conducted between 4°C and 97°C while measuring CD signal at 208 nm and 222 nm.

Chemical denaturation by guanidine chloride (GuCl) was done by titrating a sample of 15 uM designed protein in 0M GuCl into a sample of 15 uM designed protein with 7.8 M GuCl. Great care was taken to ensure the concentration of designed protein in each sample was the same. The GuCl concentration was monitored by the change in refractive index. Thermodynamic parameters were calculated assuming that the folding of the designed protein was a two state process and by fitting both the thermal and chemical denaturations to the Gibbs-Helmholtz equation.

Nuclear Magnetic Resonance Spectroscopy

The designed proteins were concentrated to 1 mM in 20 mM Sodium Phosphate pH 6.5 with 10% D_2O. ^{1}H NMR spectra were collected on a Varian Inova 600 MHz spectrometer at 25°C. NMR data and figures were processed using NMRPipe and NMRDraw. For double labeled, ^{15}N and ^{13}C, NMR experiments the designed protein was grown and purified in the same manner except that minimal media with ^{13}C glucose and ^{15}N ammonium chloride were added to the cell medium during induction. 2D and 3D NMR experiments for structure calculations were performed on Inova 750 MHz and 600 MHz spectrometers at 25°C. The series of experiments performed were ^{15}N-HSQC, ^{13}C-HSQC, CBCACONH, HBHACONH, HCCH-COSY, HCCH-TOCSY, HNCACB, HNCACO, HNCO, and NOESY. The raw and processed data are available as BMRB accession number 17612 and the final NMR ensemble from the PDB as code 2LCH.

Protein Crystallization and X-ray Crystallography

The designed protein was crystallized in 0.2 M Mg Acetate and 20% v/v PEG 3350. Crystallization experiments were performed using the hanging drop method, 0.5 μl protein at 20 mg/ml and 0.5 μl of the crystallization buffer. The diffraction data was collected at the APS Argonne National Laboratory GM/CA-CAT beam line. The crystal structure was solved by molecular replacement using the design model with all side-chain atoms removed except Cβ atoms. The diffraction data was indexed using HKL2000(Otwinowski and Minor 1997). The crystallography suite CPP4(Winn, Ballard et al.) and the refinement software COOT(Emsley, Lohkamp et al.) were used to solve the structure.

Figures

Figure 2.1: Global comparison of the wild-type template and DRNN design model.

Thirty-eight design positions shown as grey sticks were identified in the wild-type template (A). The final design model for DRNN with the designed positions shown as green sticks (B). DRNN's backbone and helix crossing angles have been subtly changed by the flexible backbone design procedure (C and D). The helices are labeled H1-H4 in panel C. Panels A, B, and D are in the same orientation and panel C is a top down view of the bundle.

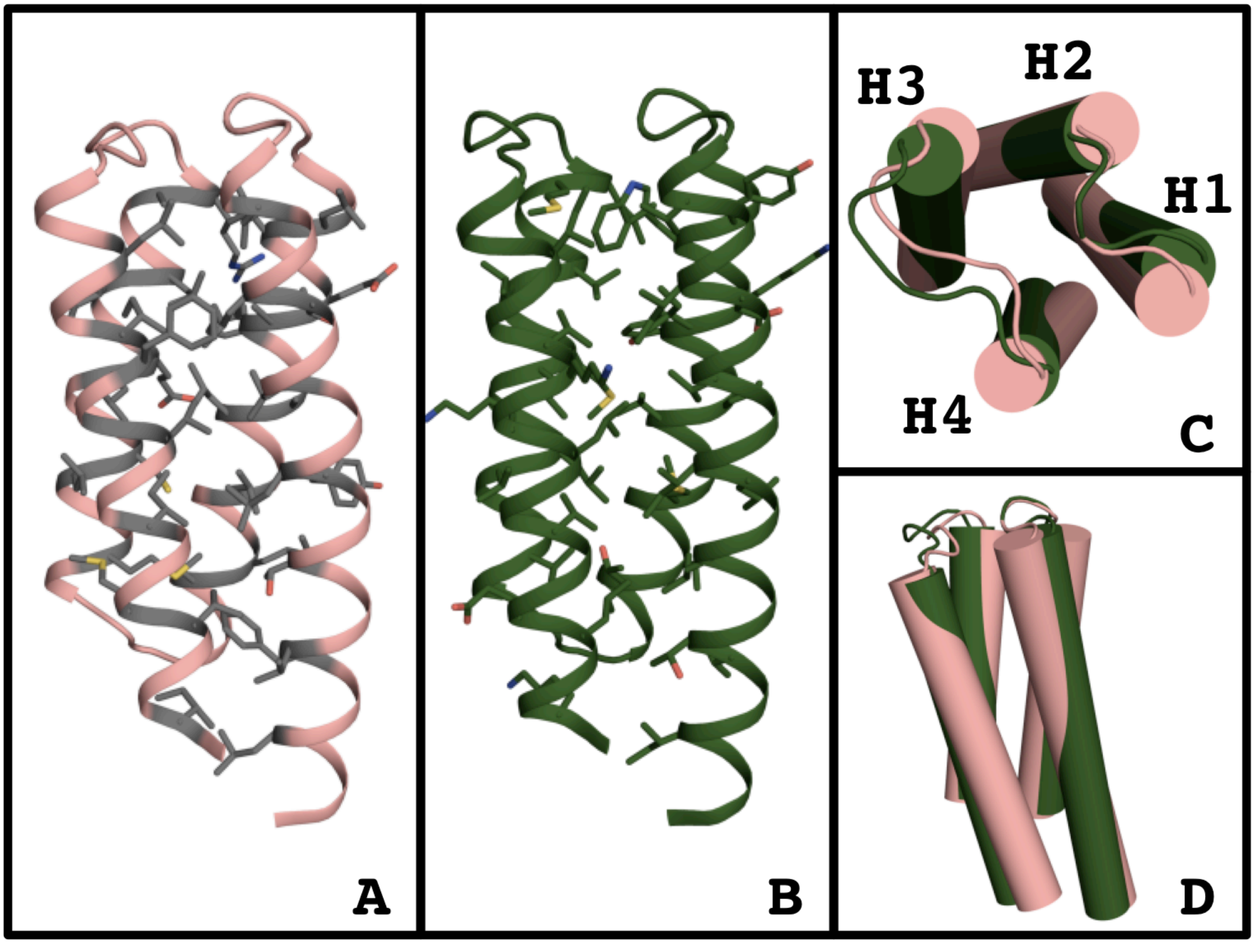
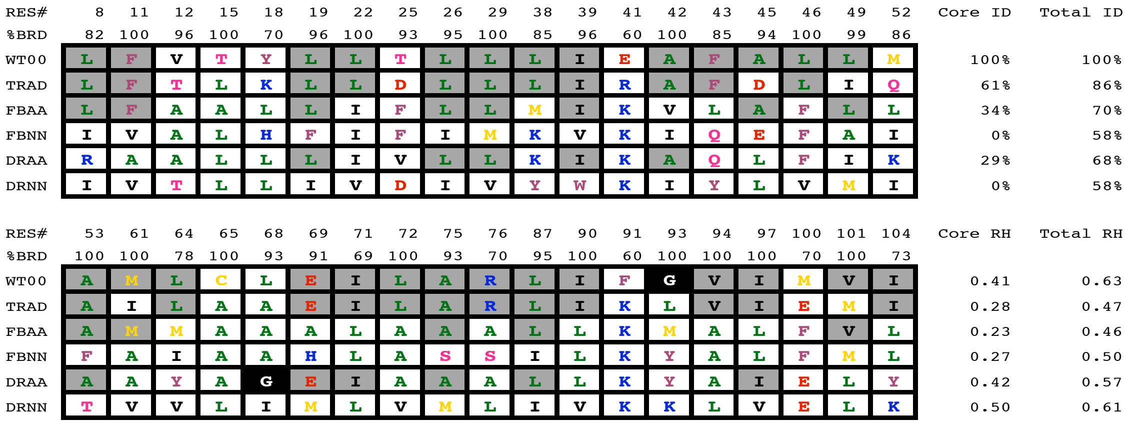


Figure 2.2: Comparison of wild-type and designed sequences.

The core sequences for wild-type(WT00), the traditional output from RosettaDesign (TRAD), and the four design experiments FBAA, FBNN, DRAA, and DRNN are shown. The core and total sequence identity and the core and total RosettaHoles scores are given for each sequence. The percent of burial for each core positions is shown as %BRD. Residue number is listed as RES#. Gray boxes indicate that a position is conserved between the wild-type sequence and one or more of the designed sequences. The one letter amino acids codes are colored red (E,D), orange (M,C), green (L,A), blue (K,R,H), black (I,V), pink (N,Q,S,T), plum (F,W,Y), and glycine is shown white on a black background



RES#	8	11	12	15	18	19	22	25	26	29	38	39	41	42	43	45	46	49	52	Core ID	Total ID
%BRD	82	100	96	100	70	96	100	93	95	100	85	96	60	100	85	94	100	99	86		
WT00	L	F	V	T	Y	L	L	T	L	L	L	I	E	A	F	A	L	L	M	100%	100%
TRAD	L	F	T	L	K	L	L	D	L	L	L	I	R	A	F	D	L	I	Q	61%	86%
FBAA	L	F	A	A	L	L	I	F	L	L	M	I	K	V	L	A	F	L	L	34%	70%
FBNN	I	V	A	L	H	F	I	F	I	M	K	V	K	I	Q	E	F	A	I	0%	58%
DRAA	R	A	A	L	L	L	I	V	L	L	K	I	K	A	Q	L	F	I	K	29%	68%
DRNN	I	V	T	L	L	I	V	D	I	V	Y	W	K	I	Y	L	V	M	I	0%	58%

RES#	53	61	64	65	68	69	71	72	75	76	87	90	91	93	94	97	100	101	104	Core RH	Total RH
%BRD	100	100	78	100	93	91	69	100	93	70	95	100	60	100	100	100	70	100	73		
WT00	A	M	L	C	L	E	I	L	A	R	L	I	F	G	V	I	M	V	I	0.41	0.63
TRAD	A	I	L	A	A	E	I	L	A	R	L	I	K	L	V	I	E	M	I	0.28	0.47
FBAA	A	M	M	A	A	A	L	A	A	A	L	L	K	M	A	L	F	V	L	0.23	0.46
FBNN	F	A	I	A	A	H	L	A	S	S	I	L	K	Y	A	L	F	M	L	0.27	0.50
DRAA	A	A	Y	A	G	E	I	A	A	A	L	L	K	Y	A	I	E	L	Y	0.42	0.57
DRNN	T	V	V	L	I	M	L	V	M	L	I	V	K	K	L	V	E	L	K	0.50	0.61

Figure 2.3: Comparison of wild-type template and DRNN design model.

The design and the wild-type bundle can be divided into five layers of interacting side-chains. Panel A shows the global view of the side-chain layers. Panels B-F show the layers with wild-type in salmon and DRNN in green, positions that were not designed are shown in grey.

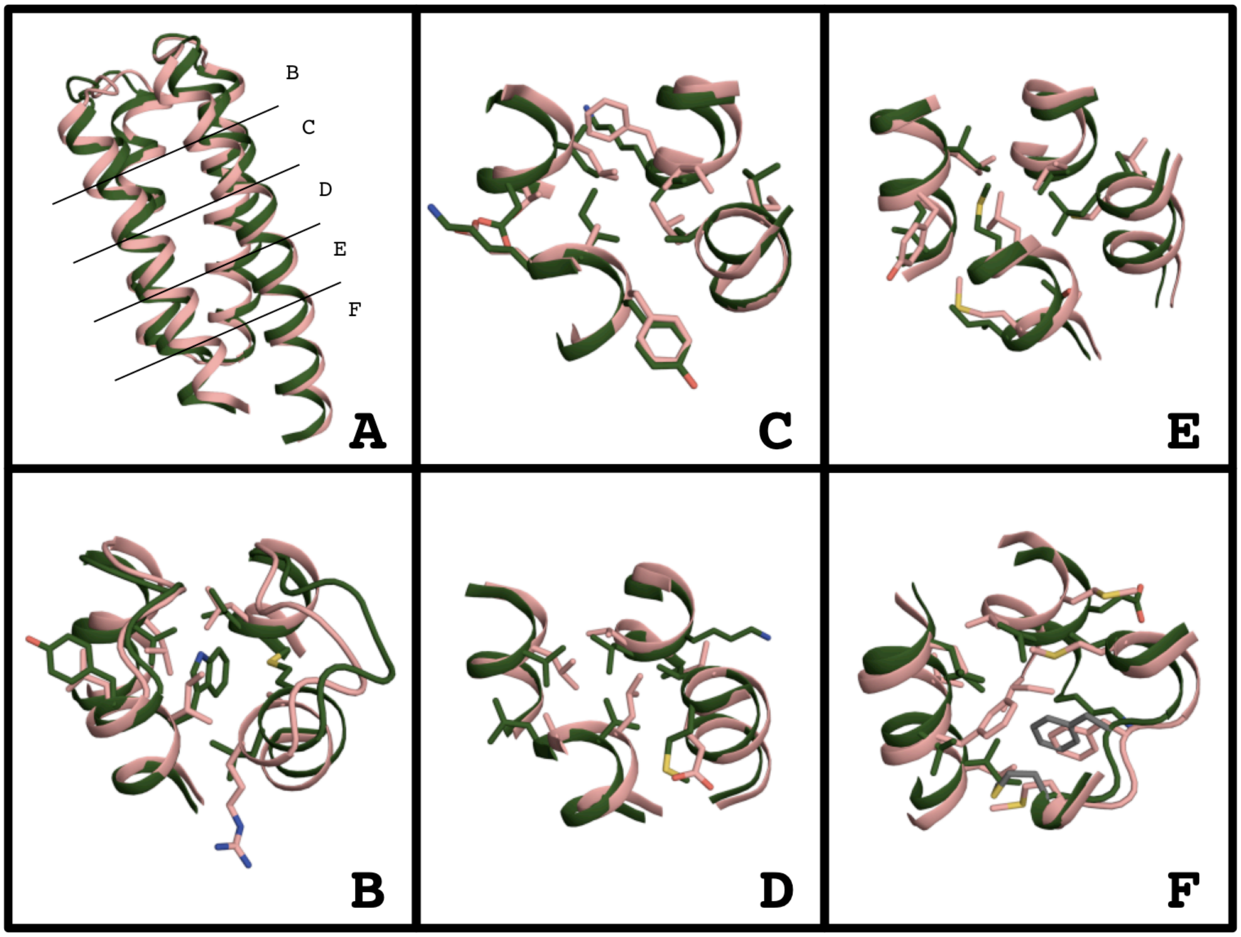


Figure 2.4: Biophysical characterization of DRNN and wild-type template.

Far UV Circular Dichroism (A), Thermal Denaturation (B), and Chemical Denaturation (C) of DRNN (green) and wild-type (salmon). Global fits (mesh) of thermal and chemical denaturation data for wild-type(D) and DRNN(E) using in the fitting of the Gibbs-Helmholtz equation. All experiments were done at 10-20 uM protein concentration in 50 μM sodium phosphate at pH 7.4 and 20°C .

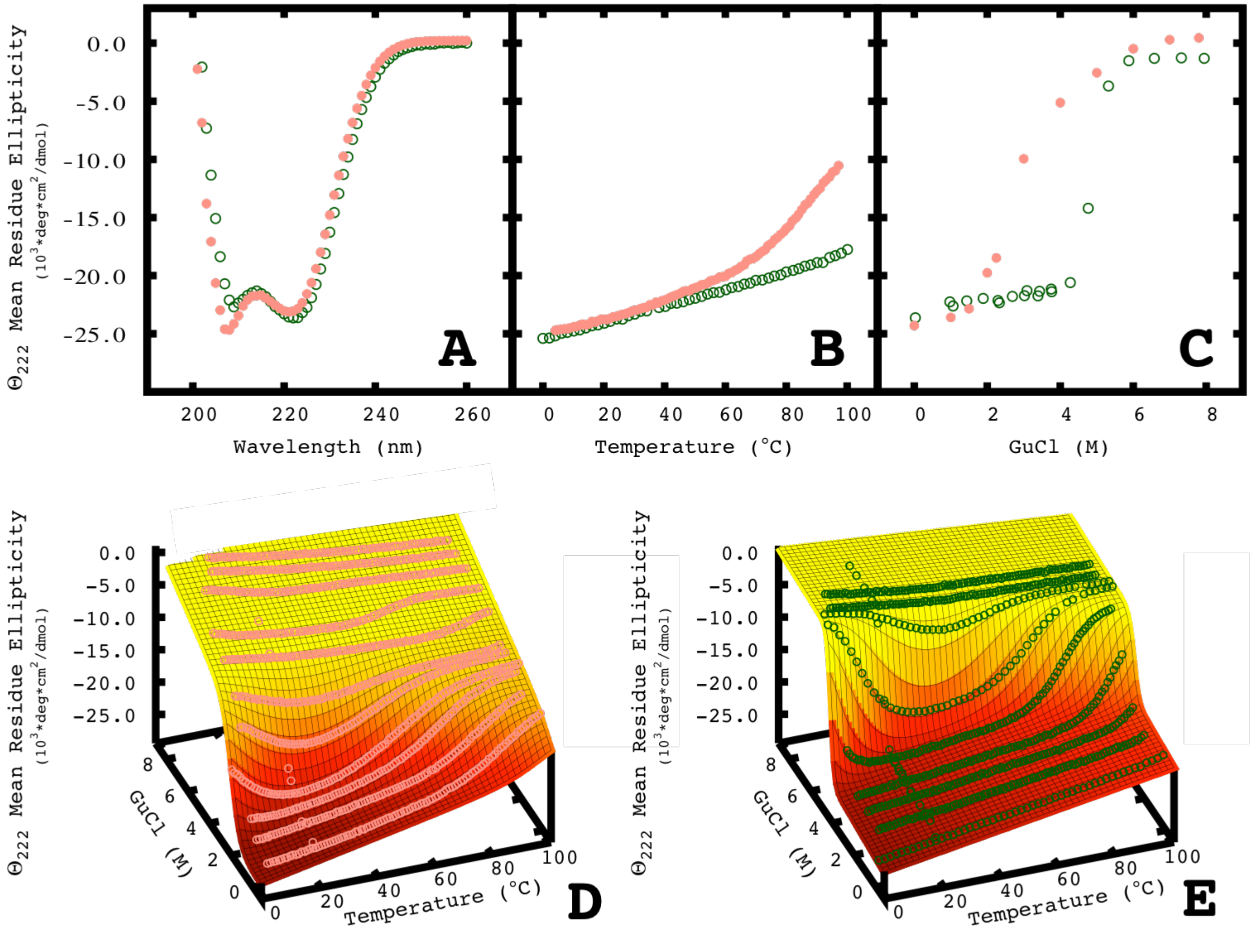


Figure 2.5: X-ray crystal structure of DRNN at 1.85Å.

The Fo-Fc electron density (green) around residue W39 during the 1st round of refinement with a backbone only model of DRNN as the molecular replacement solution (A), DRNN backbone in cyan cartoon, a tryptophan residue is shown for clarity. The final 2Fo-Fc density (purple) for chain A of the DRNN X-ray crystal structure in the five layers used to describe the wild-type and design model, sticks shown for all design positions and residues 56M and 58F.

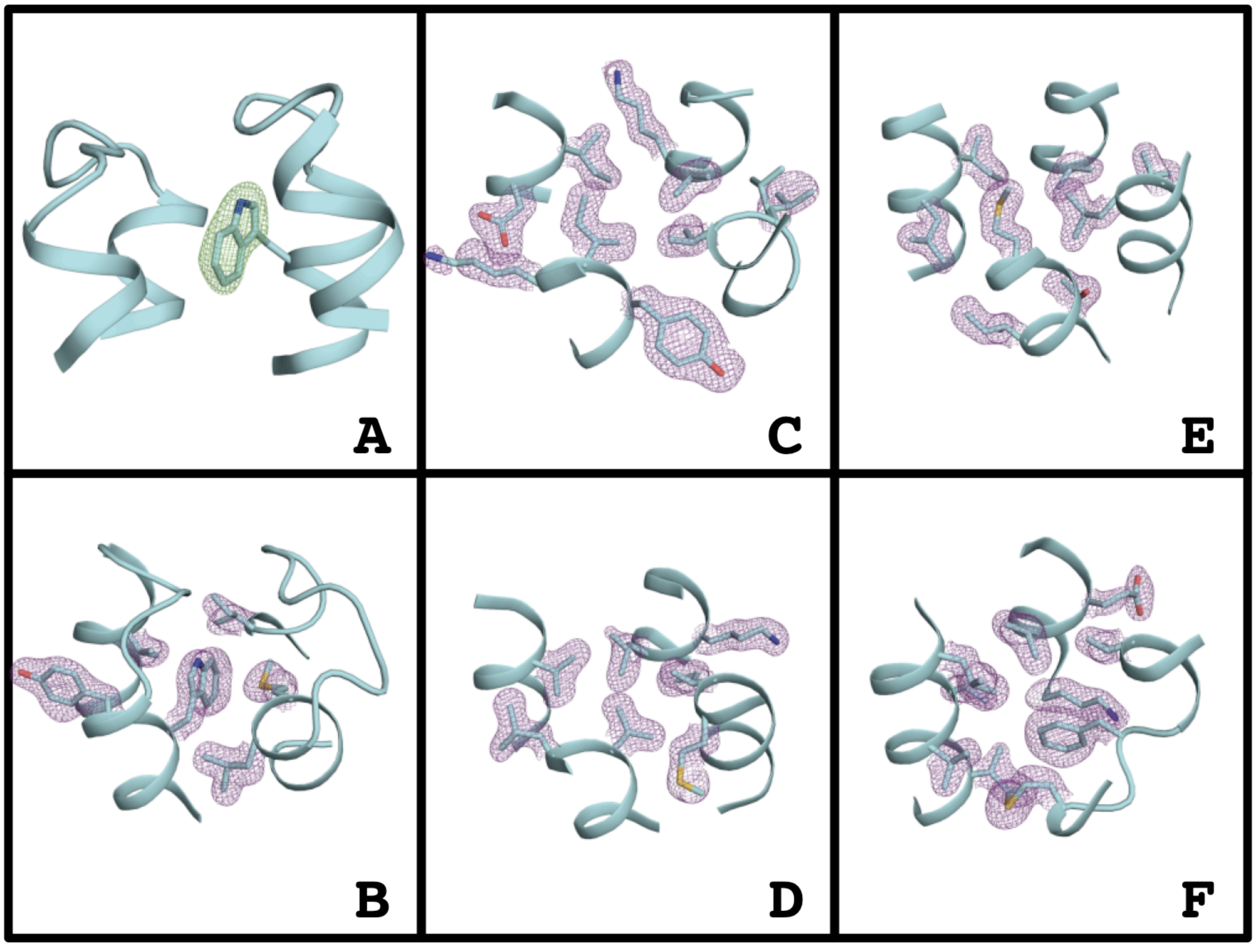


Figure 2.6: Comparison of DRNN design model and DRNN X-ray crystal structure.

The DRNN design model (green) and chain B of the X-ray crystal structure (cyan) shown in a global view (A) and as the five layers that make the bundle core (B-F) , positions that were not designed are shown in grey.

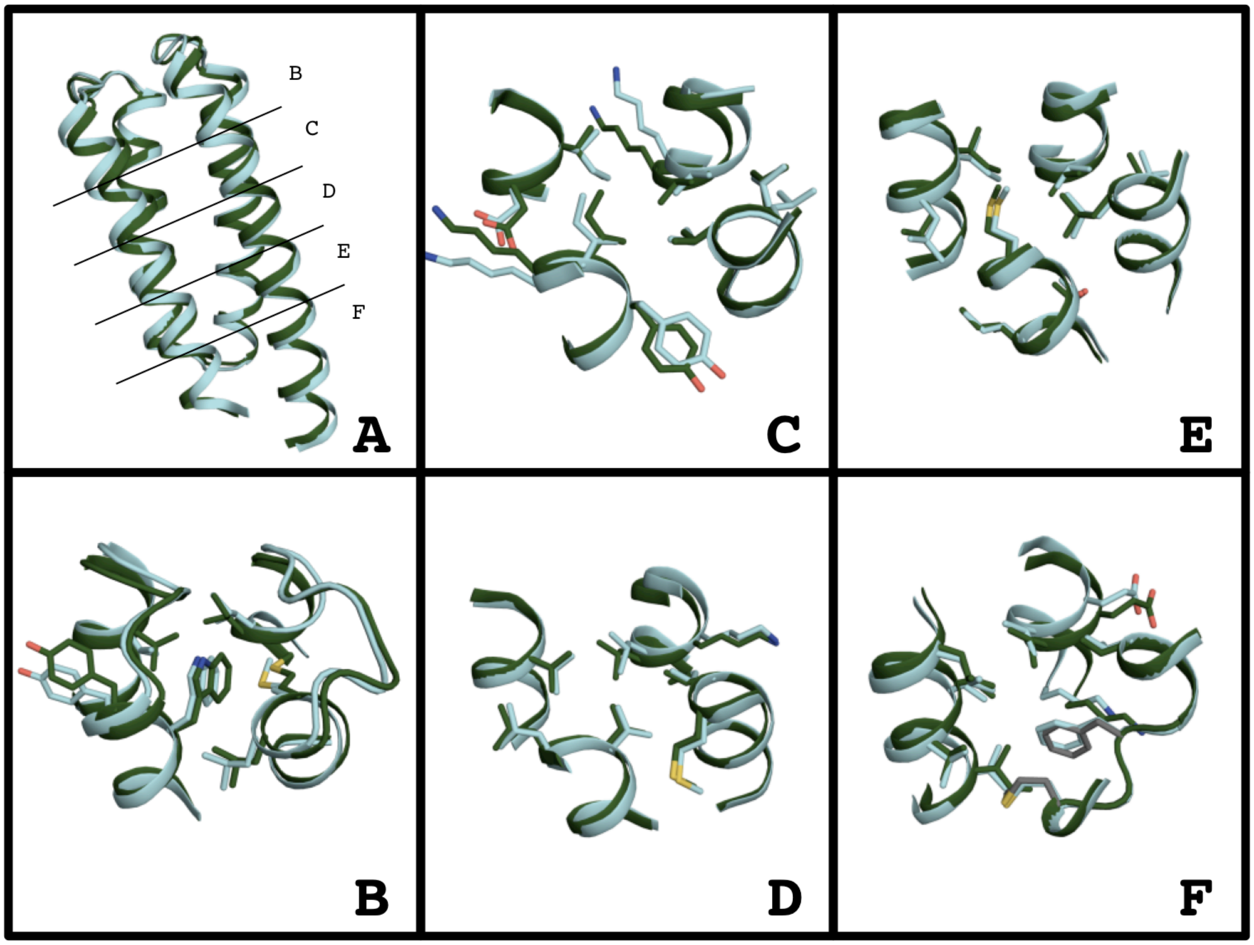


Figure 2.7: ^{15}N HSQC and NMR solution structure of DRNN.

The ^{1}H-^{15}N HSQC of DRNN at 750 uM in 50 uM NaPO$_4$ pH 6.5 (A). A global comparison of DRNN model and the DRNN solution structure (B). A zoom in on the region around W39 between DRNN and the solution structure (C) (layer B in figures 5 and 6)

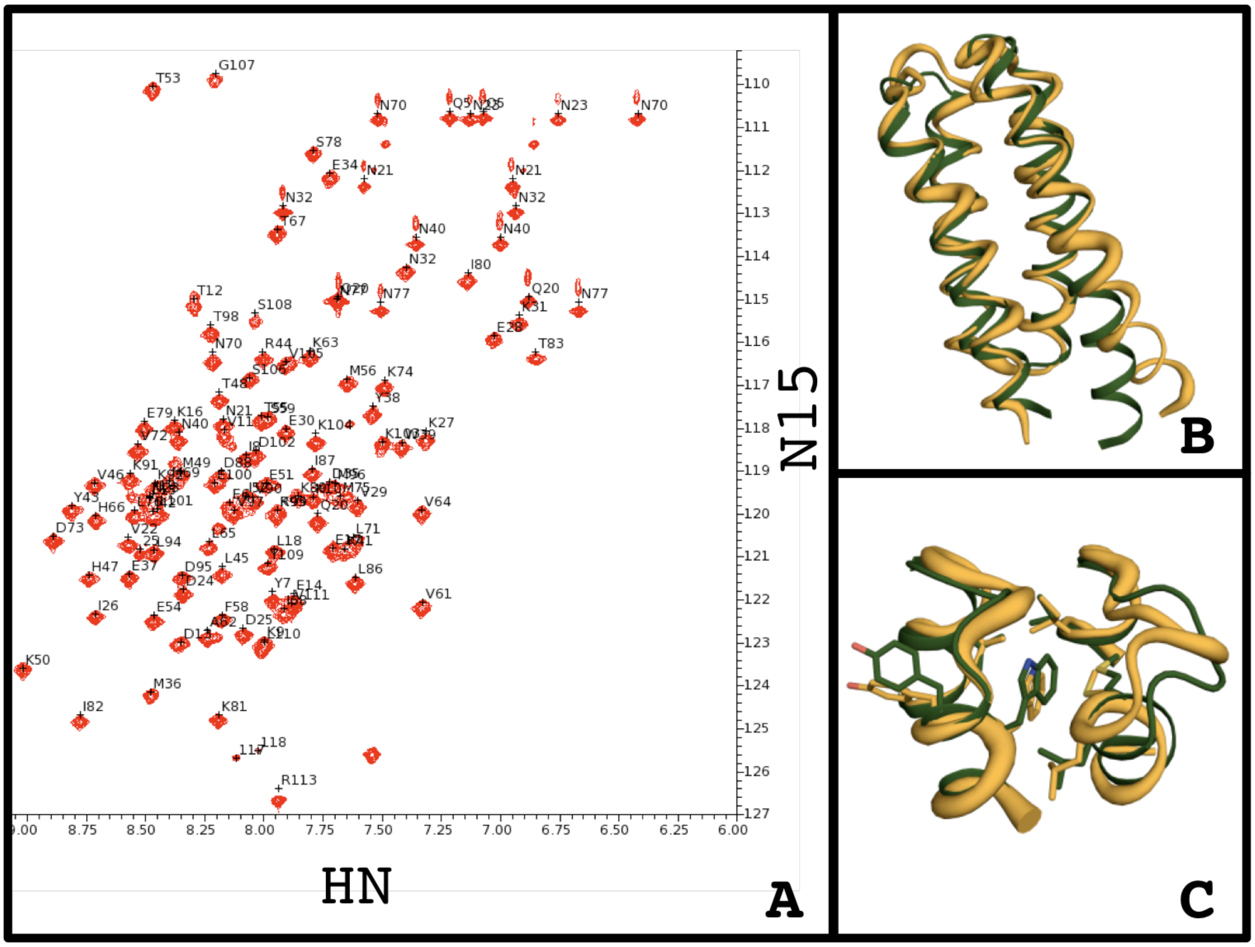


Figure 2.8: GDT plot of DRNN design model, NMR, X-ray, and wild-type template structures. GDT comparison of wild-type template (salmon), DRNN NMR structure (orange), and DRNN X-ray crystal structure (cyan) versus the DRNN design model (A). GDT comparison of wild-type template (salmon), DRNN NMR structure (orange), and DRNN design model (green) versus the DRNN X-ray crystal structure (B).

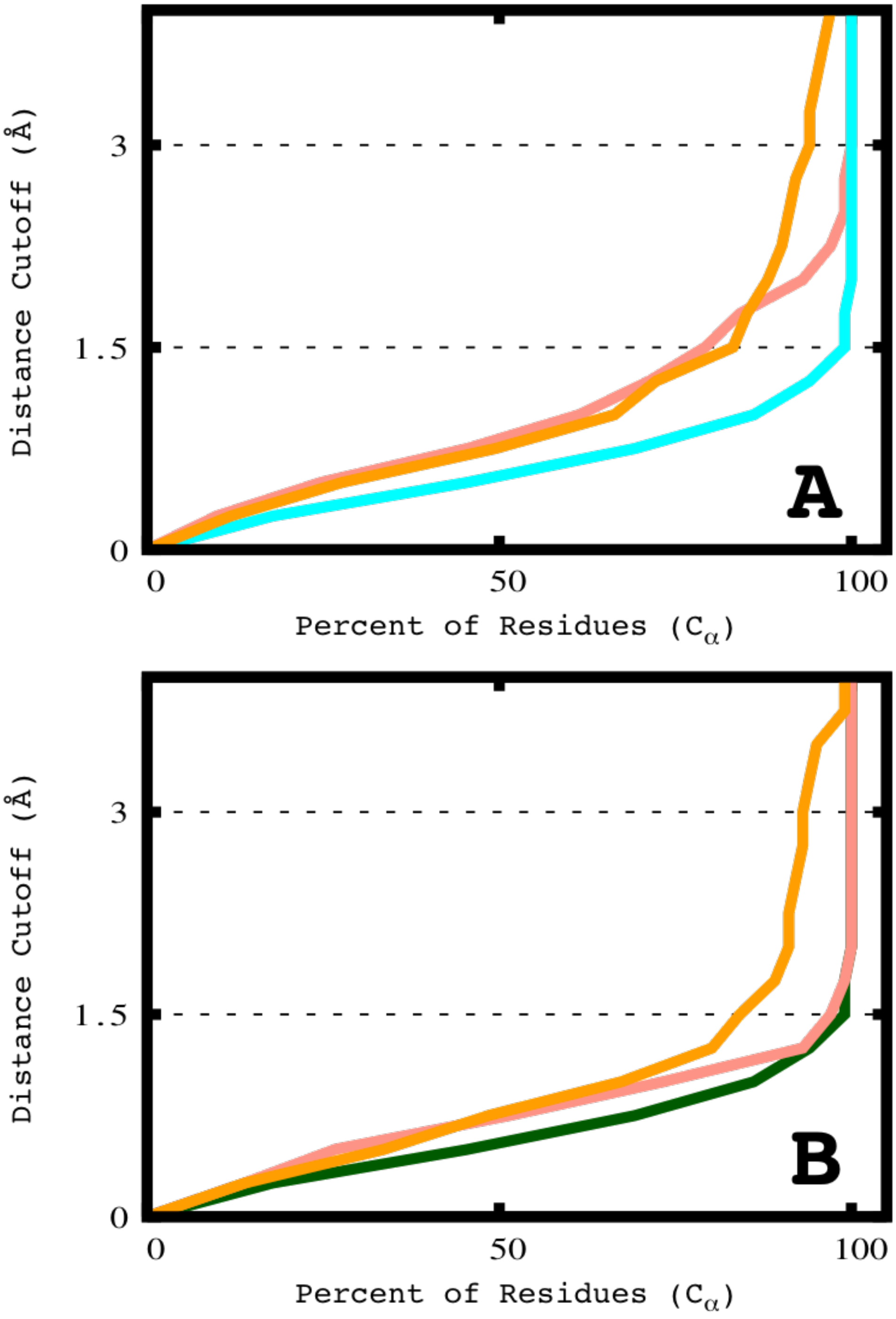
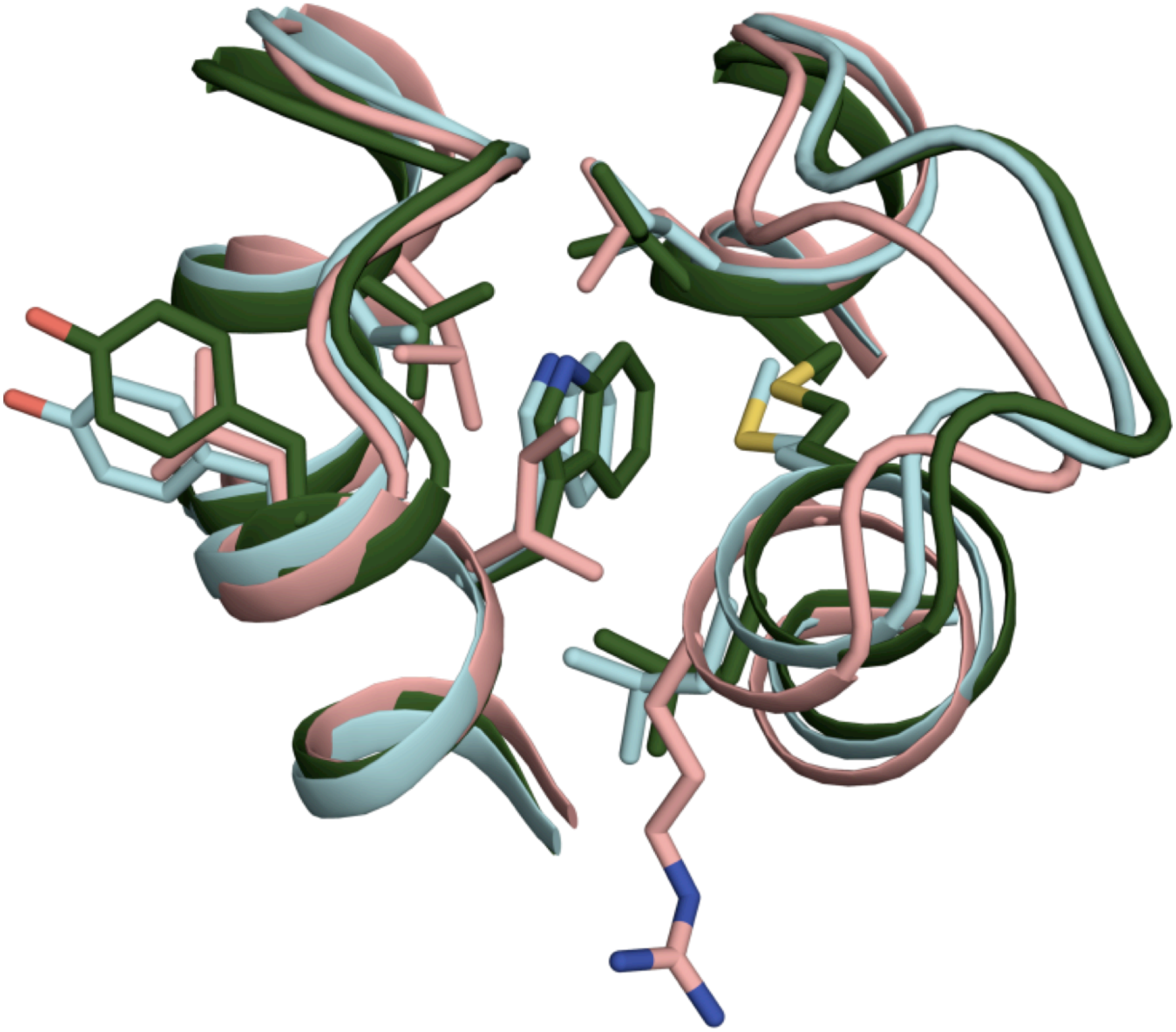
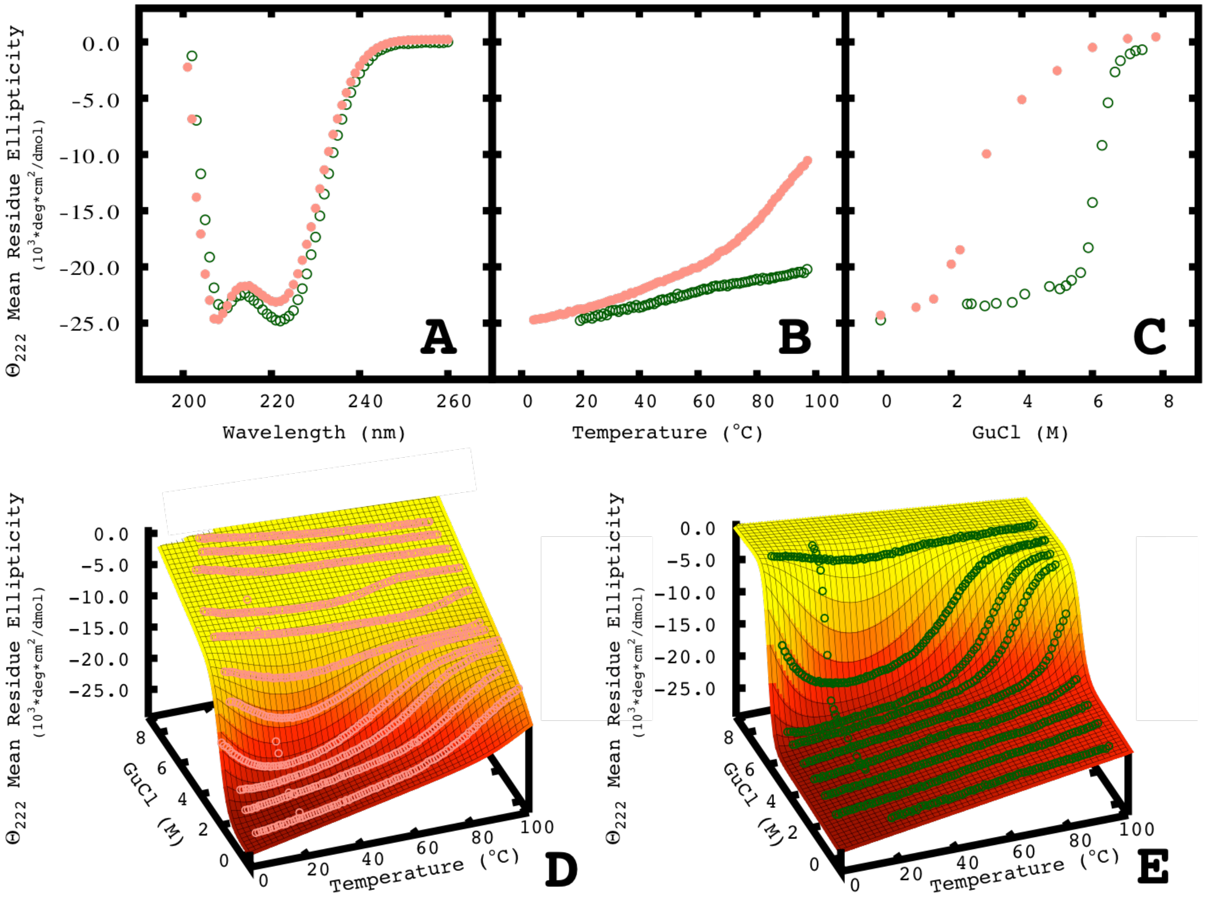


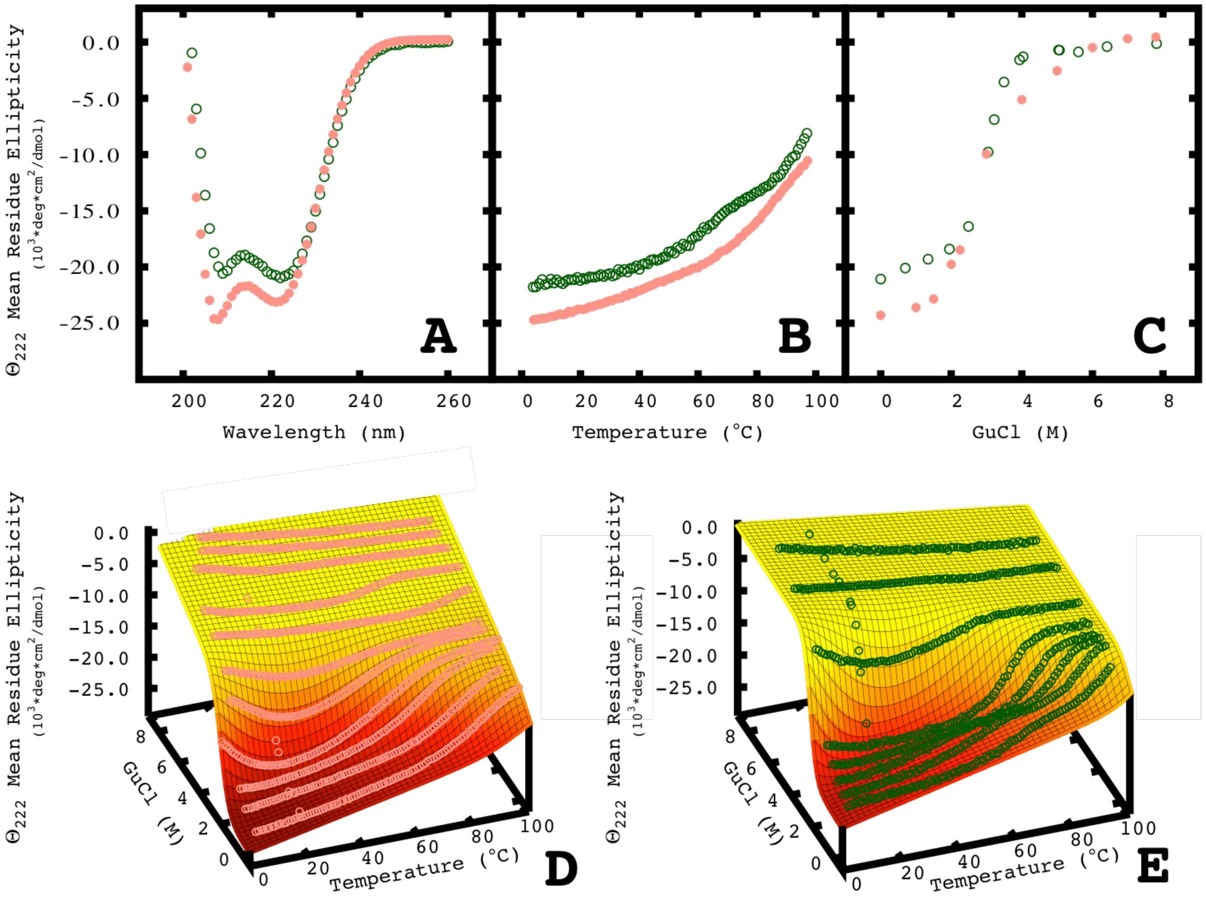
Figure 2.9: Comparison of wild-type template, DRNN design model and crystal structure. The wild-type template (salmon), DRNN design model (green), and the DRNN X-ray crystal structure (cyan) compared in the region of W39 (helix layer B shown in figures 2.3B, 2.5B, and 2.6B)



Supplemental Figure 2.1: Biophysical characterization of FBAA and wild-type template. Far UV Circular Dichroism (A), Thermal Denaturation (B), and Chemical Denaturation (C) of DRNN (green) and wild-type (salmon). Global fits (mesh) of thermal and chemical denaturation data for wild-type(D) and FBAA(E) used in the fitting of the Gibbs-Helmholtz equation. All experiments were done at 10-20 uM protein concentration in 50 μM sodium phosphate at pH 7.4 and 20°C

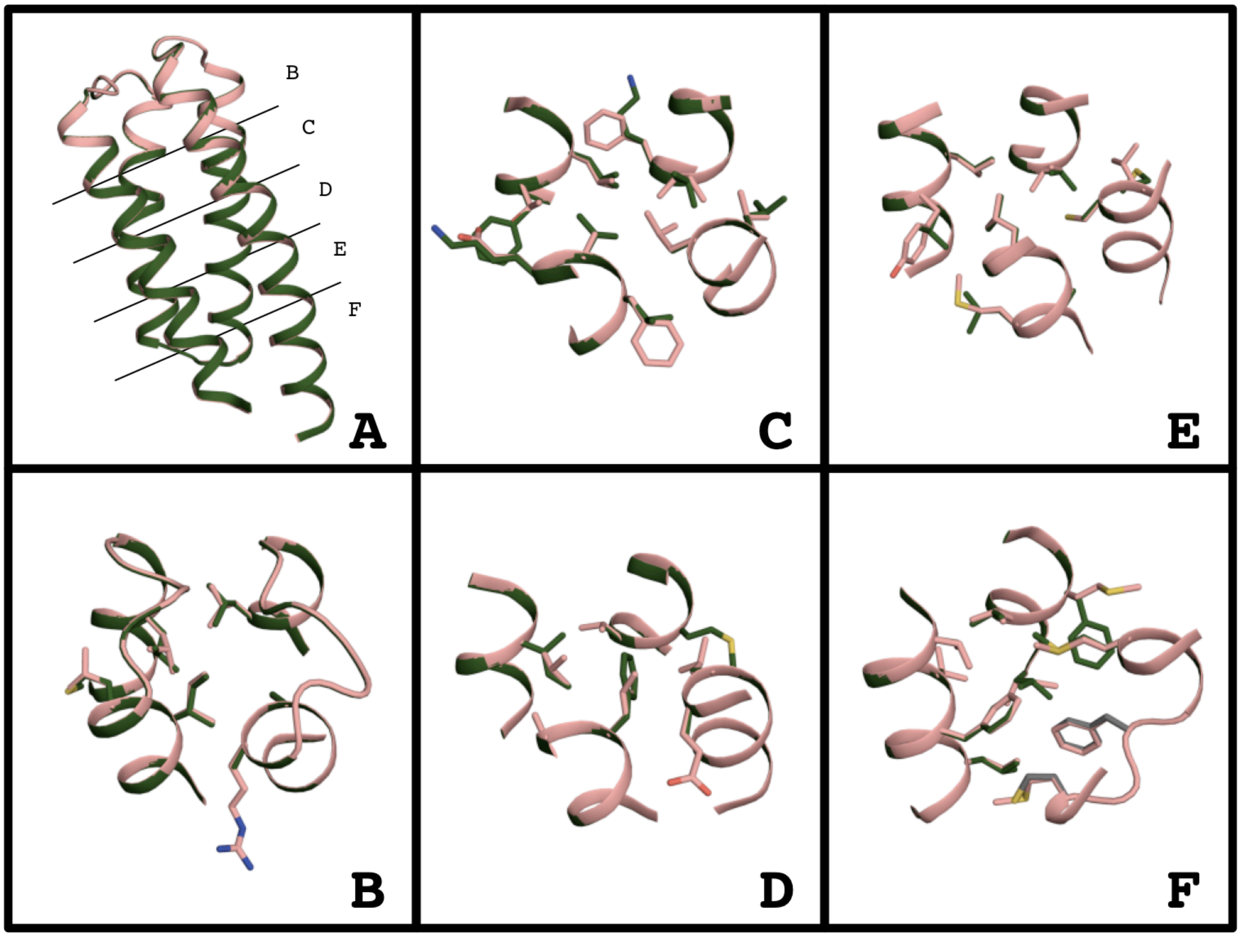


Supplemental Figure 2.2: Biophysical characterization of DRAA and wild-type template. Far UV Circular Dichroism (A), Thermal Denaturation (B), and Chemical Denaturation (C) of DRNN (green) and wild-type (salmon). Global fits (mesh) of thermal and chemical denaturation data for wild-type(D) and DRAA(E) used in the fitting of the Gibbs-Helmholtz equation. All experiments were done at 10-20 uM protein concentration in 50 μM sodium phosphate at pH 7.4 and 20°C



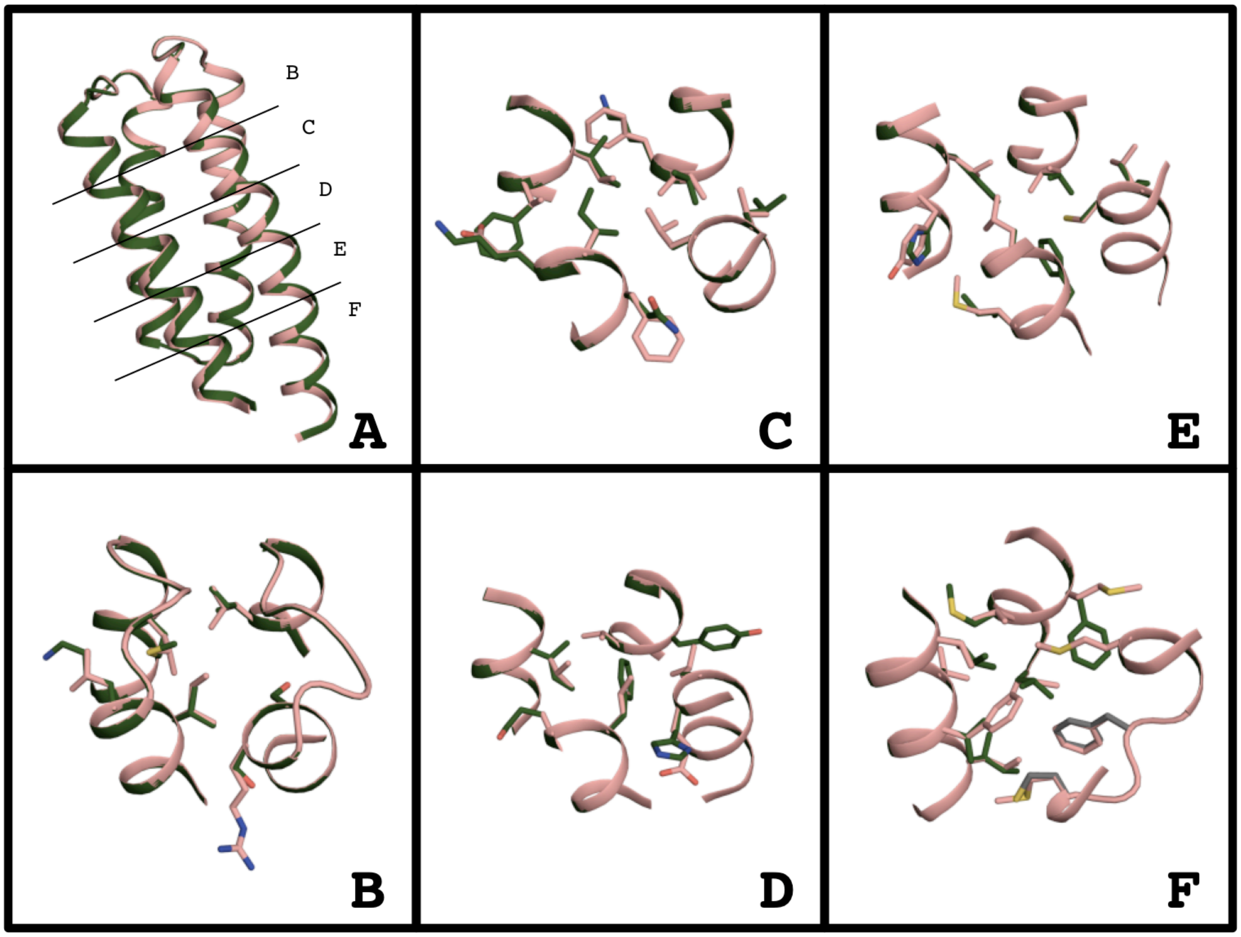
Supplemental Figure 2.3: Comparison of wild-type and FBAA design model.

The design and the wild-type bundle can be divided into 5 layers of interacting side-chains. Panel A shows the global view of the side-chain layers. Panels B-F show the layers with wild-type in salmon and FBAA in green, positions that were not designed are shown in grey.



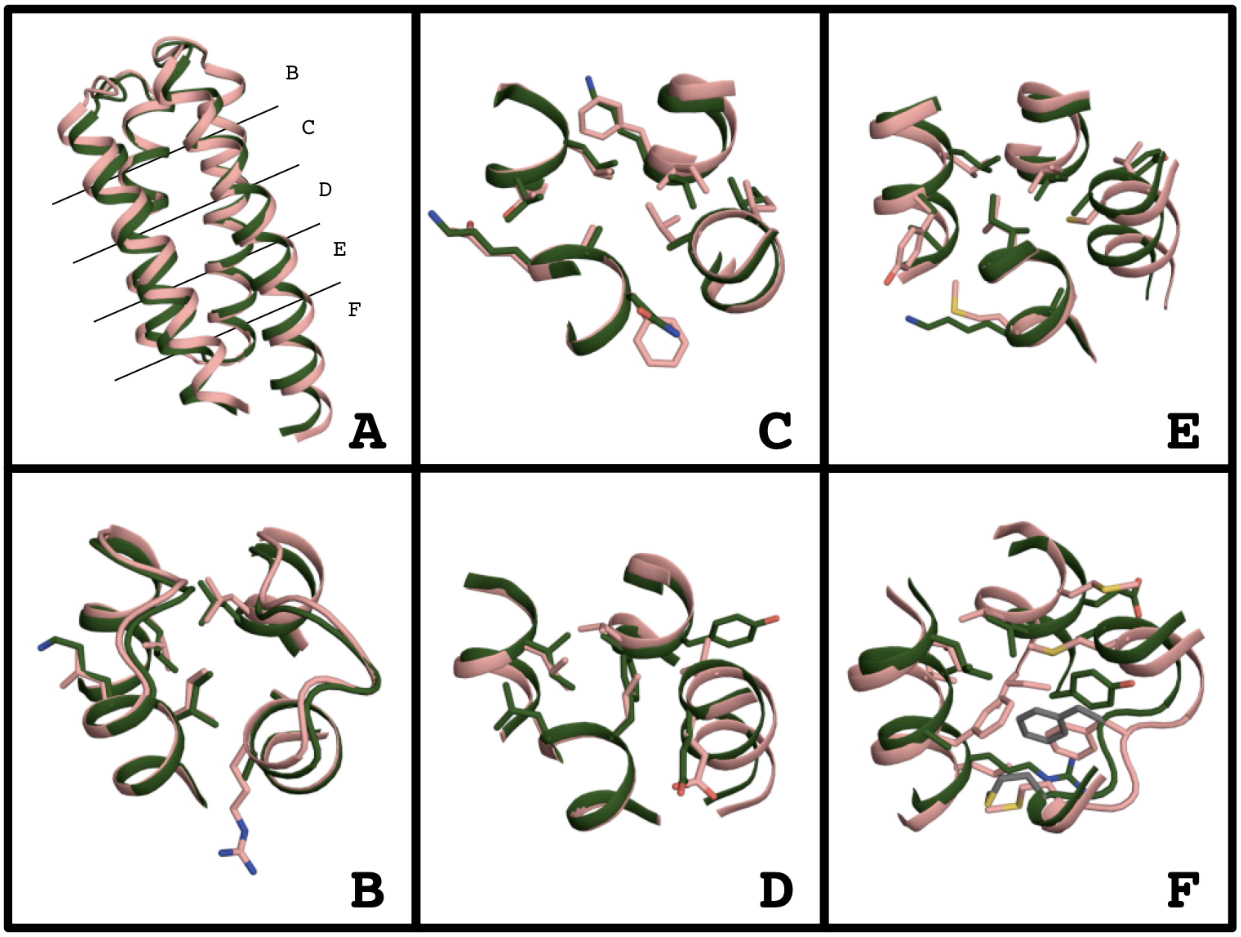
Supplemental Figure 2.4: Comparison of wild-type and FBNN design model.

The design and the wild-type bundle can be divided into 5 layers of interacting side-chains. Panel A shows the global view of the side-chain layers. Panels B-F show the layers with wild-type in salmon and FBNN in green, positions that were not designed are shown in grey.



Supplemental Figure 2.5: Comparison of wild-type and DRAA design model.

The design and the wild-type bundle can be divided into 5 layers of interacting side-chains. Panel A shows the global view of the side-chain layers. Panels B-F show the layers with wild-type in salmon and DRAA in green, positions that were not designed are shown in grey.



Supplemental Figure 2.6: Comparison of DRNN design model and NMR solution structure. The DRNN design model (green) and NMR solution structure (orange) shown in a global view (A) and as the five layers that make the bundle core (B-F), positions that were not designed are shown in grey.

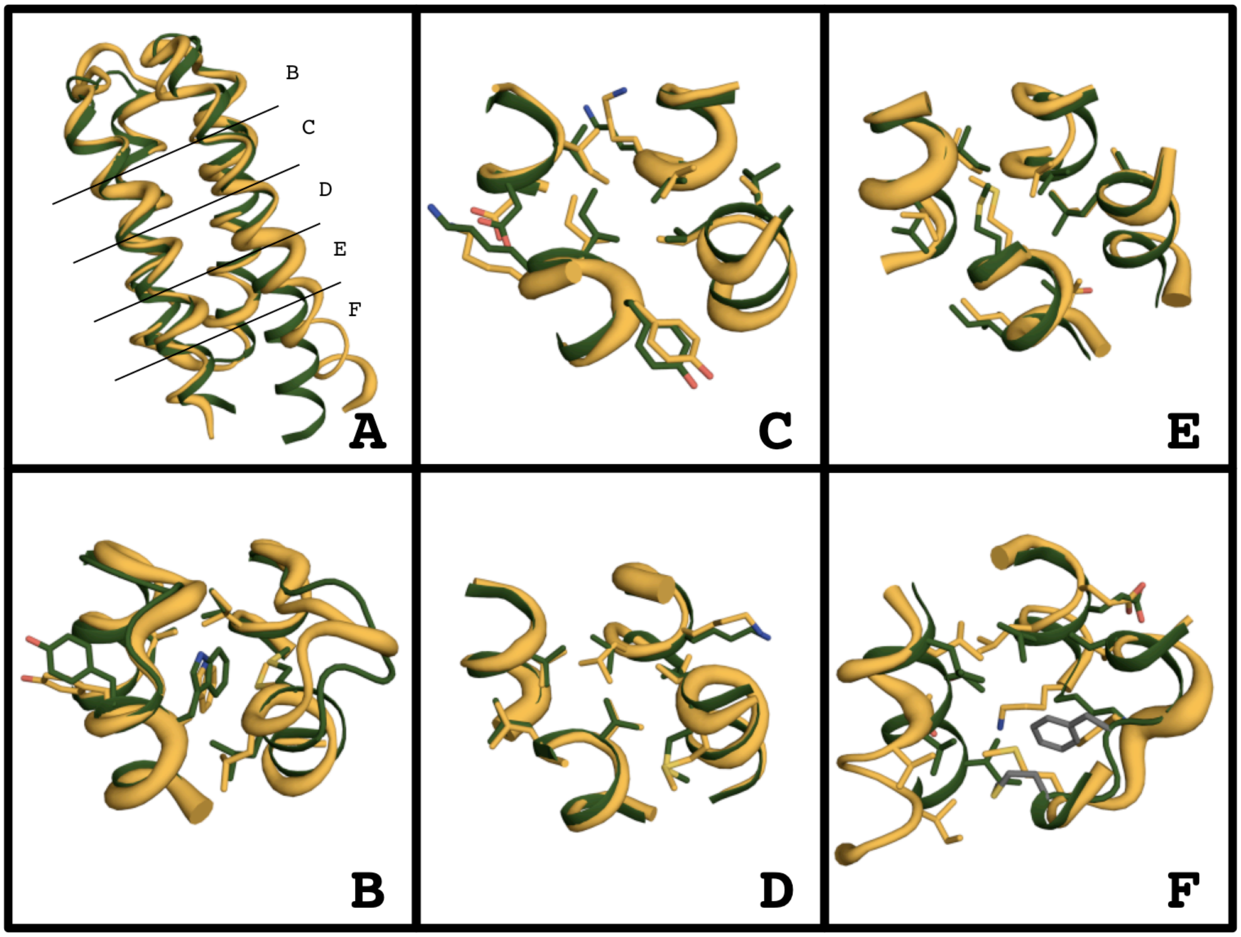
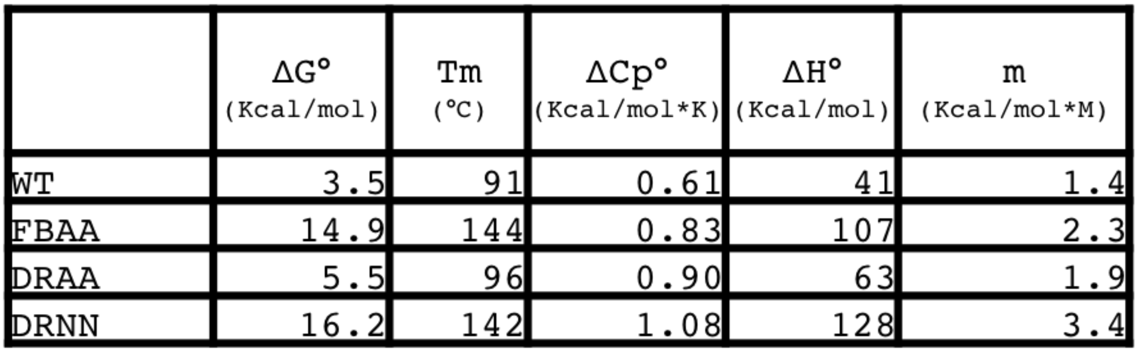


Table 2.1: Thermodynamic parameters for wild-type and designed sequences. **Values for ΔG°, Tm, ΔCp°, ΔH°, and m were calculated by globally fitting a surface of chemical and thermal melts using the Gibbs-Helmholtz equations.**



	ΔG° (Kcal/mol)	Tm (°C)	ΔCp° (Kcal/mol*K)	ΔH° (Kcal/mol)	m (Kcal/mol*M)
WT	3.5	91	0.61	41	1.4
FBAA	14.9	144	0.83	107	2.3
DRAA	5.5	96	0.90	63	1.9
DRNN	16.2	142	1.08	128	3.4

Supplementary Table 2.1:Statistics of X-ray Crystallography

```
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : REFMAC 5.5.0109
REMARK   3   AUTHORS     : MURSHUDOV,VAGIN,DODSON
REMARK   3
REMARK   3    REFINEMENT TARGET : MAXIMUM LIKELIHOOD
REMARK   3
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) :   1.85
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) :  42.21
REMARK   3   DATA CUTOFF            (SIGMA(F)) : NONE
REMARK   3   COMPLETENESS FOR RANGE        (%) :  96.32
REMARK   3   NUMBER OF REFLECTIONS             :   15344
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD          : THROUGHOUT
REMARK   3   FREE R VALUE TEST SET SELECTION  : RANDOM
REMARK   3   R VALUE     (WORKING + TEST SET) : 0.18693
REMARK   3   R VALUE            (WORKING SET) :  0.18391
REMARK   3   FREE R VALUE                     :  0.24417
REMARK   3   FREE R VALUE TEST SET SIZE   (%) :  5.1
REMARK   3   FREE R VALUE TEST SET COUNT      :   823
REMARK   3
REMARK   3  FIT IN THE HIGHEST RESOLUTION BIN.
REMARK   3   TOTAL NUMBER OF BINS USED           :      20
REMARK   3   BIN RESOLUTION RANGE HIGH           :    1.854
REMARK   3   BIN RESOLUTION RANGE LOW            :    1.902
REMARK   3   REFLECTION IN BIN     (WORKING SET) :     1043
REMARK   3   BIN COMPLETENESS (WORKING+TEST) (%) :    88.94
REMARK   3   BIN R VALUE           (WORKING SET) :    0.121
REMARK   3   BIN FREE R VALUE SET COUNT          :       51
REMARK   3   BIN FREE R VALUE                    :    0.162
REMARK   3
REMARK   3  NUMBER OF NON-HYDROGEN ATOMS USED IN
REFINEMENT.
REMARK   3   ALL ATOMS                :     1779
```

Rosetta Command Lines

For DRAA and DRNN experiments the following command lines were used, and extra rotamers were assigned automatically as described in the methods

```
~/DesignRelaxApp.macosgccrelease -database ~/database/ -s *.pdb -core_design –DRNN
~/DesignRelaxApp.macosgccrelease -database ~/database/ -s *.pdb -core_design –DRAA
```

For FBAA experiments the standard Rosetta fixed backbone design protocol was used

```
~/fixbb.macosgccrelease -database ~/database/ -s *.pdb -resfile  fbaa_resfile
```

the fbaa_resfile contained the following information

```
for fixed positions          RES# A NATAA USE_INPUT_SC EX 1 LEVEL 4 EX 2 LEVEL 4
For designable positions     RES# A ALLAA EX 1 LEVEL 6 EX 2 LEVEL 6
```

For FBNN experiments the standard Rosetta fixed backbone design protocol was used

```
~/fixbb.macosgccrelease -database ~/database/ -s *.pdb -resfile  fbnn_resfile
```

the fbnn_resfile contained the following information

```
for fixed positions          RES# A NATAA USE_INPUT_SC EX 1 LEVEL 4 EX 2 LEVEL 4
For designable positions     RES# A NOTAA “NATIVE_RES” EX 1 LEVEL 6 EX 2 LEVEL 6
```

Depending on the computational resources available the flag –lin_mem_ig 10 may be need to use large number of rotamers for any of these experiments

References

Apgar, J. R., S. Hahn, et al. (2009). "Cluster expansion models for flexible-backbone protein energetics." J Comput Chem **30**(15): 2402-13.

Cole, C., J. D. Barber, et al. (2008). "The Jpred 3 secondary structure prediction server." Nucleic Acids Res **36**(Web Server issue): W197-201.

Correia, B. E., Y. E. Ban, et al. "Computational protein design using flexible backbone remodeling and resurfacing: case studies in structure-based antigen design." J Mol Biol **405**(1): 284-97.

Cowtan, K., P. Emsley, et al. "From crystal to structure with CCP4." Acta Crystallogr D Biol Crystallogr **67**(Pt 4): 233-4.

Dantas, G., C. Corrent, et al. (2007). "High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design." J Mol Biol **366**(4): 1209-21.

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.

Davis, I. W., K. Raha, et al. (2009). "Blind docking of pharmaceutically relevant compounds using RosettaLigand." Protein Sci **18**(9): 1998-2002.

DeGrado, W. F. and B. O. Nilsson (1997). "Engineering and design Screening, selection and design: standing at the crossroads in three dimensions." Curr Opin Struct Biol **7**(4): 455-6.

Desjarlais, J. R. and T. M. Handel (1999). "Side-chain and backbone flexibility in protein core design." J Mol Biol **290**(1): 305-18.

Dunbrack, R. L., Jr. (2002). "Rotamer libraries in the 21st century." Curr Opin Struct Biol **12**(4): 431-40.

Emsley, P., B. Lohkamp, et al. "Features and development of Coot." Acta Crystallogr D Biol Crystallogr **66**(Pt 4): 486-501.

Friedland, G. D., A. J. Linares, et al. (2008). "A simple model of backbone flexibility improves modeling of side-chain conformational variability." J Mol Biol **380**(4): 757-74.

Fung, H. K., C. A. Floudas, et al. (2008). "Toward full-sequence de novo protein design with flexible templates for human beta-defensin-2." Biophys J **94**(2): 584-99.

Georgiev, I. and B. R. Donald (2007). "Dead-end elimination with backbone flexibility." Bioinformatics **23**(13): i185-94.

Gordon, D. B., S. A. Marshall, et al. (1999). "Energy functions for protein design." Curr Opin Struct Biol **9**(4): 509-13.

Grigoryan, G. and W. F. Degrado "Probing designability via a generalized model of helical bundle geometry." J Mol Biol **405**(4): 1079-100.

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

Harbury, P. B., B. Tidor, et al. (1995). "Repacking protein cores with backbone freedom: structure prediction for coiled coils." Proc Natl Acad Sci U S A **92**(18): 8408-12.

Havranek, J. J. and D. Baker (2009). "Motif-directed flexible backbone design of functional interactions." Protein Sci **18**(6): 1293-305.

Hecht, M. H., J. S. Richardson, et al. (1990). "De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence." Science **249**(4971): 884-91.

Hill, R. B. and W. F. DeGrado (2000). "A polar, solvent-exposed residue can be essential for native protein structure." Structure **8**(5): 471-9.

Hu, X., H. Wang, et al. (2007). "High-resolution design of a protein loop." Proc Natl Acad Sci U S A **104**(45): 17668-73.

Hu, X., H. Wang, et al. (2008). "Computer-based redesign of a beta sandwich protein suggests that extensive negative design is not required for de novo beta sheet design." Structure **16**(12): 1799-805.

Kabsch, W. and C. Sander (1983). "How good are predictions of protein secondary structure?" FEBS Lett **155**(2): 179-82.

Kamtekar, S., J. M. Schiffer, et al. (1993). "Protein design by binary patterning of polar and nonpolar amino acids." Science **262**(5140): 1680-5.

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. (2002). "Accurate computer-based design of a new backbone conformation in the second turn of protein L." J Mol Biol **315**(3): 471-7.

Leaver-Fay, A., M. Tyka, et al. "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules." Methods Enzymol **487**: 545-74.

Lovejoy, B., S. Choe, et al. (1993). "Crystal structure of a synthetic triple-stranded alpha-helical bundle." Science **259**(5099): 1288-93.

Mandell, D. J. and T. Kortemme (2009). "Backbone flexibility in computational protein design." Curr Opin Biotechnol **20**(4): 420-8.

Murphy, P. M., J. M. Bolduc, et al. (2009). "Alteration of enzyme specificity by computational loop remodeling and design." Proc Natl Acad Sci U S A **106**(23): 9215-20.

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.

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

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.

Raman, S., R. Vernon, et al. (2009). "Structure prediction for CASP8 with all-atom refinement using Rosetta." Proteins **77 Suppl 9**: 89-99.

Sammond, D. W., D. E. Bosch, et al. "Computational design of the sequence and structure of a protein-binding peptide." J Am Chem Soc **133**(12): 4190-2.

Sheffler, W. and D. Baker (2009). "RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design, and validation." Protein Sci **18**(1): 229-39.

Shen, Y., R. Vernon, et al. (2009). "De novo protein structure generation from incomplete chemical shift assignments." J Biomol NMR **43**(2): 63-78.

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

Willis, M. A., B. Bishop, et al. (2000). "Dramatic structural and thermodynamic consequences of repacking a protein's hydrophobic core." Structure **8**(12): 1319-28.

Winn, M. D., C. C. Ballard, et al. "Overview of the CCP4 suite and current developments." Acta Crystallogr D Biol Crystallogr **67**(Pt 4): 235-42.

Yin, S., F. Ding, et al. (2007). "Modeling backbone flexibility improves protein stability estimation." Structure **15**(12): 1567-76.

Chapter 3

The Automated *de novo* Design of Diverse Protein Folds

Grant S. Murphy[1], Carrie Purbeck[3], Mischa Machius[2], and Brian Kuhlman[3*]

[1]Department of Chemistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-3290, USA

[2]Center for Structural Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA

[3]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-7260, USA

[*]corresponding author. Email: bkuhlman@email.unc.edu

Phone: 919-843-0188

Fax: 919-966-2852

Running title: *de novo* design of diverse protein folds

Abstract

The creation of proteins *de novo*, where the desired three-dimensional structure used in simulations is not from an experimental structure but is created from scratch, is a major goal in protein design. We have developed a computational framework for the rapid and efficient generation of *de novo* sequences for existing and novel protein folds. This computational method is robust and easily used by non-expert users, opening the doors of *de novo* protein design to a broader biochemical audience. We have experimentally tested *de novo* designed sequences for four-helix bundle proteins, helical SAM domain proteins, six stranded β-sandwich folds, and novel six stranded β-sandwich folds. We have crystallized a *de novo* designed four-helix bundle, and solved its structure using NMR chemical shift information, and report the results for the other folds.

Keywords: Computational Protein Design, *de novo* Protein Design, Flexible Backbone Design

Introduction

The creation of proteins *de novo*, from new, is a major goal of computational protein design. Here we use the term *de novo* in the strictest sense possible, to describe design projects that do not begin with coordinates from an existing X-ray or NMR structure. This definition requires the backbone coordinates and the protein sequence of a design model to be created from scratch. With this definition in place, there have only been a handful of attempts at *de novo* design and even fewer success stories(Gutte, Daumigen et al. 1979; Regan and DeGrado 1988; Hecht, Richardson et al. 1990; Quinn, Tweedy et al. 1994). A brief review of the history of *de novo* protein design will provide a foundation for the computational and experimental techniques that will be the focus of this research.

The history of protein design begins with Gutte in 1979. Gutte et al. *de novo* designed a minimalist 34 residue version of Ribonuclease A. At the time and in retrospect the design was an amazing success. It appeared to be folded, stable, and possessed modest catalytic activity(Gutte, Daumigen et al. 1979). While there are only a handful of examples from the literature that of *de novo* design, we cannot cover the full history of *de novo* protein design here, instead we provide a few seminal examples of *de novo* protein design. After Gutte's initial success, a series of papers were published by DeGrado, which describe the creation of a *de novo* helix bundle protein in a hierarchal method. DeGrado began first by creating a single helix that associated as a tetramer. He then created small linker connections between helix 1 and 2 and helix 3 and 4, which associated as a dimer of helical hairpins. Finally, he connected helix 2 and helix 3 with a small linker to create a single chain protein(Regan and DeGrado 1988). DeGrado's designs were helical, stable and appeared folded. The structures of these designed proteins were not solved and so the success of the designs at a global fold or detailed atomic level could not be verified.

The next major *de novo* designs were a helix bundle (Felix) by Hecht in 1990 and the *de novo* design of a dimer β-sandwich (β-doublet) by Quinn in 1994, both done in the Richardson's lab(Hecht, Richardson et al. 1990; Quinn, Tweedy et al. 1994), the *de novo* design of a helix bundle with a novel right handed super helical twist arising from an undecatad hydrophobic polar pattern by Harbury in 1998 (3)(Harbury, Plecs et al. 1998), and the *de novo* design of a novel mixed α/β, Top7, fold by Kuhlman in 2003 while in Baker's lab (4)(Kuhlman, Dantas et al. 2003). Of these four attempts, Felix was designed primarily by hand and β-doublet was designed using a mixture of by hand and computational methods. Both were shown to be molten globules(Hecht, Richardson et al. 1990; Quinn, Tweedy et al. 1994). Harbury used a system of parametric equations to create *de novo* backbone models and to incorporate backbone flexibility into his designs. His sequences were chosen computationally, and he validated one of his design models with an X-ray crystal structure(Harbury, Plecs et al. 1998). Kuhlman used a flexible backbone design method that coupled sequence design followed with backbone and side-chain dihedral optimization. The design was validated with an X-ray crystal structure(Kuhlman, Dantas et al. 2003). Somewhat surprisingly, both Kuhlman and Harbury incorporated backbone flexibility, designed novel topologies and were successful. In contrast, the *de novo* design of an natural folds, helix bundles and β-sandwiches, modeled at the atomic level were not successful.

The creation of novel protein folds is a grand challenge. However, it will not be necessary to create novel protein folds to solve every protein design challenge. In fact, nature typically does not create novel protein folds when creating new functions, as evidenced by single folds having many functions. In some cases however, a novel protein fold may prove to be a more desirable or the only solution for a protein design challenge. It is important that we develop protein design techniques for the creation of both novel and existing protein folds. The *de novo* design of existing and novel protein folds must become more reliable and consistent before protein design can mature into a tool that will be used by the greater biochemistry community.

The *de novo* design of proteins can be reduced to two separate but equally important tasks: (1) the creation of a starting structure and (2) the selection of a low energy sequence for a given structure. The creation of high quality, native like, starting structures is a serious challenge, and is the computational bottleneck in *de novo* protein design. This issue is known as the protein backbone designability problem. In contrast, the selection of low energy sequences for a given structure is known to be highly successful in the protein redesign problem. In protein redesign, the starting structure is most often an X-ray crystal structure that nature has already proven to be designable(Dahiyat and Mayo 1997; Dantas, Kuhlman et al. 2003). Ultimately then, *de novo* protein design is reduced to how efficiently designable backbones can be identified. When creating starting structures for novel or existing protein folds, we have no guarantee that the protein backbone will be designable until we attempt to design it. We hope that our energy functions will identify backbones that are not designable by assigning them high-energy sequences or sequences that do not resemble native sequences.

Previous attempts at *de novo* design have taken one of three approaches to create starting structures: (1) the use of parametric equations to generate a continuum of backbones(Harbury, Plecs et al. 1998), (2) the use of idealized backbones motifs and a hierarchical assembly based on geometric methods(Bryson, Betz et al. 1995; Offredi, Dubail et al. 2003), and (3) creation of backbones using protein structure prediction(Kuhlman, Dantas et al. 2003). The method we describe here attempts to incorporate the most powerful features of all three previous techniques, by creating and exploiting a geometric definition of protein tertiary structure, using a set of intuitive rules that resemble the hierarchical processes previously used, and the incorporation of this information to guide our protein structure prediction method towards the desired topology.

Our specific goal was to create a computational framework that would be general enough to create high quality starting structures for most protein folds in an automated fashion, and to validate these computationally designed sequences experimentally.

Results

Computational Results

Flexible Backbone Protein Design

In this work, we use a flexible backbone protein design protocol that has been highly successful in the redesign of helix bundle proteins and similar to other flexible backbone design protocols that have also been successful. Briefly, the protocol iterates between fixed backbone protein design and high-resolution structure refinement until an energy threshold is reached. This method has been shown to rapidly decrease the energy of designed sequences and quickly abandon those designs that will never have low energy due to poor backbone designability or other sequence defects.

Benchmarking High-Resolution Structure Refinement, Flexible Backbone Protein Design, and *de novo* Structure Creation and Design

A series of benchmarking experiments were performed to understand the range of Rosetta energies produced by and from different types of computational protocols and initial starting structures. These experiments were performed to identify the range of Rosetta energies that computationally designed proteins should have prior to attempting experimental characterization. We used the TOP500 database of proteins as an example of high quality naturally occurring protein backbones and sequences. The Rosetta high-resolution structure prediction method was applied ≥ 1000 times for 300 members of this database. Additionally, we applied our flexible backbone protein design method ≥ 1000 times to these 300 structures.

The set of models generated using Rosetta's high-resolution structure prediction method on this data set had an average Rosetta energy per residue of -2.5. The average Rosetta energy per residue for designs generated using the flexible backbone design method for this data set was -2.8 (Figure 3.1A). In addition to benchmarking the energies of models generated using high-resolution structure prediction and flexible backbone design methods, we were also interested in the quality of packing between the original crystal structure, the high-resolution structure prediction models, and the flexible backbone design models, where the general trend is that Rosetta manipulated structures are less well packed then experimentally determined X-ray structures, especially for those determined at high resolution (Figure 3.1B). With this knowledge, we decided that an average Rosetta energy per residue of -2.5 would be the minimum energy a *de novo* design should have before being considered for experimental characterization (a few exceptions have been made), and that designs near or better than -2.8 would be exceptional.

In the next several paragraphs we describe a new computational method for the creation of *de novo* starting structures. The *de novo* structure creation method developed here (*de novo* tertiary topology creator) consists of several components divided into three stages: a definition of the desired topology, preparation for fragment assembly and biased fragment assembly. The definition of the desired topology is a set of geometric constraints that can be derived from a single known protein structure, a family of homologous known protein structures, or defined explicitly in a simple file format. The definition also consists of a list of the number of residues in all secondary structure elements (helix, strand and loop), a list that describes the strand-strand pairings that form β-sheets, and a list of optional constraints that are available as part of Rosetta. All features of the topology definition can be explicitly specified or defined automatically or a mixture of each can be used. This system creates an extremely flexible interface for the creation of *de novo* starting structures.

Three steps occur while defining the desired topology, a helix/strand/loop pattern is chosen randomly from the available helix/strand/loop length ranges provided. This *de novo* helix/strand/loop pattern is used to create geometric constraints and strand-strand pairings that define the topology. At this point a crude model of the topology can be built consisting of only helix and strand elements by doing a rigid body minimization of idealized strand and helix elements onto the *de novo* geometric constraints. This broken topology, because it does not have loops, can be used to specify which positions are likely to be buried, if a specific set of positions are desired to be buried this can be specified and will supersede the above implementation. However, the selection of a singular and specific hydrophobic polar pattern should be done with great care. When creating a hydrophobic polar pattern and a helix, strand, loop pattern by hand, there are no guarantees that a physically real pattern will be created. This is one of the powerful features of the method presented here, all patterns tested will either be physically realizable or will quickly be blacklisted. After a broken topology has been created and the geometric definition of the desired topology is complete, fragments of protein secondary structure can be culled from the protein data bank for use in a biased fragment assembly protocol.

Pieces of secondary structure, fragments of length three and nine residues, which match the desired secondary structure pattern are pulled from a non-redundant set of high-resolution protein structures. These fragments are similar to the fragments used in Rosetta's highly successful *ab initio* structure prediction method where the fragments' backbone dihedrals are used to collapse an extended chain during a Metropolis Monte Carlo folding procedure. In addition to these traditional fragments, a second set of fragments called bridge fragments and edge fragments are created. Bridge fragments span the gap between two pieces of helical or strand secondary structure. Bridge fragments can be of any length, and will be at least the length of the loop between two pieces of helical or strand secondary structure plus one to five residues into each pieces of helical or strand secondary structure. Bridge fragments are chosen based on how well they match the desired secondary structure pattern and the geometric constraints of the broken topology. Edge fragments are β-strand fragments at the edge of a β-sheet. These fragments are the entire length of the edge strand and are enriched for features like β-bulges, prolines, glycines, and disruptions in the strand binary hydrophobic polar pattern.

The use of bridge fragments is two fold, first the conformational search done in the fragment assembly step is highly biased by the bridge fragments for the desired fold, and second bridge and edge fragments combined with a new type of fragment insertion retain features such has helix capping, β-hairpin motifs, and edge strand features. After fragments have been created the protocol proceeds to the fragment assembly step.

The fragment assembly step is a modified version of the Rosetta *ab initio* structure prediction method. Two important modifications have been made. First, a new type of fragment insertion called FragmentSequence insertion is performed. Second, the geometric constraints that define the desired topology are used in the early stages of the assembly protocol to guide the assembly towards the desired region of conformational space. In a traditional fragment assembly protocol a vast majority of computational time is wasted searching undesired conformational space.

FragmentSequence insertions differ from traditional fragment insertions because they modify both the backbone dihedrals of the protein chain and the sequence of the protein, in effect doing design during the middle of creating the starting structure, we call this type of design “design in media res”. It is hoped that this new type of fragment insertion combined with edge and bridge fragments will capture features of natural proteins which provide elements of negative design, and also to further optimize the sequence and structure relationship in regions critical for protein folding, such as long loops and edge strands (further description in Materials and Methods).

The *de novo* structure creation method was benchmarked on a set of naturally occurring protein folds: helix bundles, helical Sam domains, existing β-sandwich folds, β-grasps, and a novel β-sandwich fold (Figure 3.2). The average Rosetta energy per residue and the quality of packing was compared to the TOP500 dataset and to the wild type representatives of the same fold type for the non-novel folds. The lowest energy sequences created by this protocol score similarly or better than refined wild-type proteins, primarily Rosetta total score less than -2.5 per residue and RosettaHoles Score better than 0.6 (Figure 3.3). This is a somewhat artificial result because we have filters that cull designs with poor energies and poor packing, in effect we never see sequences worse than our threshold.

An additional metric tested was the ability of a designed sequence to recover the intended fold using Rosetta's *ab initio* structure prediction method. All of the designs discussed here were successfully predicted by Rosetta's ab initio structure prediction method to ≤ 2.5 Å backbone rmsd(Figure 3.4).

Sequences Chosen for Experimental Characterization

After the initial round of benchmarking of the starting structure creation protocol, a series of experiments where performed to generate sequences that would ultimately be experimentally tested. Designed sequences where created for four different folds: an up down four-helix bundle (dnd_4hb), a helical SAM domain (dnd_sam), an existing β-sandwich fold (dnd_xbs), and a novel β-sandwich fold (dnd_nbs) that we are currently testing. At least 10,000 designed sequences were generated for each fold and were pruned to a small subset based on the following criteria. All sequences were ranked based on their average Rosetta total energy per residue, quality of packing, and the number of buried polar unsatisfied atoms. Designed sequences were pruned if the Rosetta total energy was not less than -2.5 REU/res, if the quality of packing was less than 0.5 as measured by RosettaHoles, and if their were any unsatisfied buried polar atoms. After this pruning, the 10 best scoring sequences were evaluated by their predicted secondary structure from JPRED3, and the ability of Rosetta's structure prediction method to identify the target fold as the lowest energy conformation for the designed sequence.

Experimental Results

Designed Protein Expression and Purification

The designed proteins dnd_4hb and dnd_sam both expressed in high yield and were easily purified by IMAC and gel filtration. Currently, all β-sandwich designs tested have either been inexpressible or have expressed so poorly that the amount of protein yielded was not enough for experimental characterization. The gene for designed protein dnd_nbs has recently been ordered and its results are currently unknown.

Circular Dichroism

The secondary structure of designed protein, dnd_sam, was characterized using Far-UV circular dichroism. The stability of dnd_sam was characterized with thermal and chemical denaturation by circular dichroism. dnd_sam showed moderate helical secondary structure and modest cooperative unfolding both thermally and chemically.

The addition of TMAO, a folding promoter, increased the CD signal at 208 and 222, indicative of further formation of helices. In a combined TMAO/GuCl titration it was apparent that dnd_sam was only partially fold at 20°C with 0 M TMAO and 0 M GuCl and folded in 2 M TMAO. The designed protein dnd_4hb was characterized in the same manner. dnd_4hb was thermally and chemical stable with a Tm of 96°C and a midpoint of unfolding from a GuCl denaturation at 2.6 M GuCl. The ΔG° of unfolding was calculated to be 4.9 kcal/mol by fitting he Gibbs-Helmholtz equation to the surface of several chemical and thermal denaturations. Additionally, parameters for ΔH°=52 kcal/mol, ΔCp°=0.7 kcal/mol•deg, and m=1.9 kcal/(mol•M) were calculated from the fit of the Gibbs-Helmholtz surface (Figure 3.5).

Crystallography

X-ray crystallography for dnd_4hb was pursued in an attempt to solve its structure for comparison against the design model. The designed protein dnd_4hb crystallized readily in many conditions, all of which contained a small organic similar in chemical nature to 2-methyl-2,4-pentanediol. The best diffraction observed was 3.8 Å, in 0.2 M ammonium acetate, 0.1 M tri-sodium citrate and 30% w/v 2-methyl-2,4-pentanediol, in a P6 space group. Unfortunately using the design model as a molecular replacement solution to solve the phase problem was not successful.

Proton NMR

To evaluate the foldedness of designed proteins, proton NMR was performed on dnd_sam and dnd_4hb. The proton NMR spectra for dnd_4hb showed disperse peaks in the methyl and aromatic regions, where as the proton spectra for dnd_sam indicated a partially folded or molten globule structure.

2D/3D NMR experiments

^{15}N-HSQC, ^{13}C-HSQC, HNCACB, CBCA(CO)NH, and HNCO experiments were performed for dnd_4hb. Protein backbone peaks were assigned for the ^{15}N-HSQC for greater than 90% of dnd_4hb (Figure 3.6A, Table 3.1).

NMR Structure of dnd_4hb using CHESHIRE and Chemical Shift Data

The protein backbone chemical shift assignments for dnd_4hb were used in the secondary structure prediction method of Talos+ and to solve the NMR structure using the methods CHESHIRE and CS-Rosetta. Talos+ indicates that the helix and loop pattern predicted from the chemical shifts is nearly identical to the design model. The structure given by CS-Rosetta is nearly identical to the design model, all-atom RMSD for all members of the ensemble less than ~2.0 A, and backbone rmsd of less than ~1.5 A.

As a control to account for the fact that Rosetta may be positively biased towards structures, which it has designed, the software CHESHIRE was also used to solve the structure using only the sequence and chemical shift information for dnd_4hb. The structural ensemble generated by CHESHIRE is similar to the design model. The structure predicted by CHESHIRE is the intended left-handed four-helix bundle fold. All ten members of the ensemble have all atom rmsd less than 3.8Å, and backbone rmsd less than 3.0Å. The first member of the ensemble has all atom rmsd 3.518Å and backbone rmsd 2.643(Figure 3.6B, 3.6C).

Discussion

De novo protein design is still a challenging problem. Here we have presented the successful design of a completely *de novo* helix bundle with NMR structural evidence that suggests we have designed the correct fold and possibly many atomic level interactions. We also report the somewhat successful design of a helix SAM domain, and the failure of three β-sandwich proteins.

It is clear that we understand the rules of protein design for helix bundle proteins more clearly than for other proteins. There are a number of redesign, heuristic, partially *de novo*, evolutionary, and now completely *de novo* success stories for helix bundle and coiled coil proteins(Hecht, Richardson et al. 1990; Wei, Kim et al. 2003; Butterfoss and Kuhlman 2006; Kuhlman and DeGrado 2009). Why is it that protein design succeeds for helix bundle folds and two novel topologies but in general fails elsewhere? Can we divine out from this work and previous work any information that will help future protein design efforts?

Three culprits are typically blamed when designed proteins fail: the designed sequence is not sufficiently thermodynamically stable resulting in a molten globule (1), the sequence has bad folding kinetics leading to aggregation (2), or the sequence has favorable energy for one or many alternative conformations (3).

If designed proteins were failing because of poor free energy, then we would not expect Rosetta, and other methods, to have such a successful track record in the redesign and partial *de novo* design of a broad set of protein folds(Dantas, Kuhlman et al. 2003). The Rosetta energy function has been shown to be an excellent predictor of low energy sequences for monomer proteins(Kuhlman and Baker 2000), and it seems unlikely that this is the primary reason *de novo* designs fail.

It is nearly impossible to know if a design has failed due to poor folding kinetics. Protein design algorithms explicitly optimize thermodynamic favorability but in general do not explicitly consider folding kinetics. For up-down helix bundle proteins, folding kinetics are unlikely to be a limiting step since most helix bundle proteins are quick two state folders. Perhaps this feature is one reason why the design of helical folds has been more successful than other protein folds. However, folding kinetics for other folds could be of great importance, especially in folds with long loops that have been optimized by evolution to favor the native state and actively disfavor alternative states. In the case of Top7, which is known to be a slow folding protein, perhaps it is the incorporation of two fast folding helices and fast forming β-hairpins which stops the protein from aggregating, while the remainder of the β-sheet and fold forms(Zhang and Chan). Further experiments focusing on the redesign of natural proteins for either folding kinetics or for the destabilization of alternative states will help to address these issues.

The destabilization of alternative states, known as negative design, has been addressed in several design projects(Regan and DeGrado 1988; Hecht, Richardson et al. 1990; Harbury, Plecs et al. 1998; Hu, Wang et al. 2008). Hu et al. showed that explicit negative design wasn't necessary for β-sandwich redesign, however that sequence may intrinsically contain negative design features from the wild type protein. Here we attempted to address the favorability of our designed sequences for alternative states by using structure prediction methods to identify the most favorable conformations of each sequence. For all α-folds, we used both Rosetta and iTasser to make these predictions. Our sequences were predicted to be most stable in the desired conformations. For all β-folds, we used Rosetta's structure prediction method supplemented with constraints, to guide the fold toward possible topologies, including the target topology and known alternative folds. In these experiments it appears as if positive design was effective, at least computationally, at eliminating alternative states. However, experimentally we can only draw conclusions about all α-folds.

Conclusions

Clearly, the *de novo* design of any fold is still a huge challenge. The *de novo* design of helix bundle proteins has been attempted and has been successful in a number of examples. Yet when using the same principles to design a different all α-fold, we have limited success. This could be an effect of the small number of sequences tested or it could highlight features missing from our computational methods.

One great challenge with the *de novo* design of all β-folds is the fact that almost no information can be gained when a design fails by aggregation. A powerful experiment for the future *de novo* design of all β-folds will be the systematic redesign of a naturally occurring β-sandwich, taking the opposite approach used in our previous study redesigning helix bundle protein cores. It would be informative to redesign the loops and surface of a natural β-sandwich, leaving the core sequence intact, to probe the role of protein folding and protein solubility in the failure of *de novo* protein design of β-sandwich folds. Initial work by Hu, on a single loop of an FNIII domain has shown this to be a viable experimental system(Hu, Wang et al. 2007).

Protein design is still very early in its development. The small number of attempts at *de novo* design is not a result of disinterest or a fear of failure, but is a reflection of the human hours required to design and test even a single sequence. One route forward to test large numbers of *de novo* is to approach *de novo* design in a high-throughput fashion. With the use of liquid handling robots, pcr assembly of genes, cell free protein expression, and high-throughput biophysical techniques and X-ray crystallography it is possible to test literally hundreds to thousands of designed sequences in the same time frame a single researcher might test a handful of designed sequences. The information gained from experiments of this type would be invaluable for improving our understanding of protein folding, stability and function.

Materials and Methods

Computational Methods

Geometric Representation of Protein Secondary Structure

There are many techniques to describe protein secondary structure and protein tertiary structure ranging from highly detailed (full atomic coordinates) to minimalist (a simple string representation of H/E/L) and methods in between. The complexity of secondary structure and tertiary structure is computationally expensive to represent explicitly and a reduced representation is used here to aid computational efficiency. It also useful to represent protein secondary and tertiary structure in a reduced representation when attempting to remodel protein structure such that a diversity of similar backbone conformations are generated but not exactly the same as the starting conformation.

We develop a simple but effective reduced representation of protein tertiary structure based on the concept that protein topology easily defined and constructed by the helix or strand axis of each piece of regular secondary structure. We represent protein secondary structure as a series of vectors that correspond to the α-helix or β-strand axis. This method was chosen because it is can account for the curvature seen in both helix and strand, as well as for the subtle differences in helix and/or strand lengths and curvature between two homologous proteins. If the heads and tails of the vectors are considered as points, this method provides an easy framework for a simple file based definition of existing or novel protein topologies (Figure 3.7A,7B). To define the helix or strand axis as a series of vectors, the explicit helix and strand axis are first defined. The helix axis is found by identifying the center of a circle that circumscribes the triangle defined by the backbone nitrogen atoms at amino acids i(N), i+1(N), i+2(N) for all amino acids in the helix, the explicit strand axis is defined by the average point of all backbone atoms for the amino acids i,i+1, i+2. We will call these points that define the secondary structure axis, axis points. With this series of axis points defining the helix and strand axis, we attempt to define two vectors that describe the helix or strand. The first vector begins at the N-terminal axis point. The second vector ends at the C-terminal axis point. We now define the true midpoint of the N and C terminal axis points, as M_T, and the apparent midpoint M_A of the helix or strand as the axis point most closely associated with the middle three residues of the helix or strand. Other methods for defining the helix and strand axis exist and are likely as robust. This method has limitations especially for helix or strand elements that have dramatic curvature at one terminus. In addition to the representation of helix and strands as vectors, it is possible to represent loops in a similar fashion. Beginning with the C terminal axis point of helix/strand i and the N terminal axis point of helix/strand i+1, the point M_{TL}, the true midpoint between C and N, can be defined and then the vector from M_{TL} to the average backbone position of the middle three residues of the loop spanning helix/strand i and helix/strand i+1, can be defined as M_{AL}, the apparent loop midpoint. We treat the protein n and c terminus as loops in the representation, assuming that the first and last residue of a protein always fill this role. There are obvious edge cases such as small loops, strands, and helices where this method will be limited, in these cases the vectors are defined in a similar fashion but with less accurate information. With the above framework in place it is now possible to define any protein topology as a series of 2 vectors, or 3 points, for every piece of helix/strand secondary structure, and 1 vector or

2 points for every loop or termini, this has the relationship of TotalPoints = Number HELIX/STRAND Elements*5 +2 or TotalVectors = Number of Helix/Strand Elements*3+1,so for a 6 stranded β- fold we have 6 strands, giving 32 points or 19 vectors. The above framework gives us a reduced representation of protein tertiary structure that can be used to guide the creation of backbone starting structure for protein design without providing explicit information either from human intuition or from atomic details of known protein structures (Figure 3.7C-3.7F).

Defining Protein Tertiary Topology for *de novo* Protein Design

This computational method allows three methods to import the desired protein topology. The first two methods require that the protein fold either exists in nature or that a pdb file of the desired fold exists. If this is the case then a single pdb file can be used and the required geometric information is taken from the given pdb file and used to define the desired topology. The user is then free to define additional bounds, i.e. how much the length of each helix, strand, or loop may change as well as how much the vectors that define the topology may change. In addition, it is also possible to place constraints on the protein sequence that will be generated during design. If the user has more than one pdb which represents the desired topology, then all of these pdb files will be used when creating the geometric definition of the desired topology, and the loop, stand, and helix lengths and vectors will be within these bounds during the simulations. The most exciting method is the generation of a topology from scratch. In this case, a simple file, which describes the desired tertiary structure as a secondary structure string of H,E, and L, along with the organization of strands into sheets, given by a poker hand(E4, E3, E2, E5), the pairing of helices, and the layering of helices and sheets is used in to create initial models and extract the required geometric information. The protocol attempts to collapse the secondary structure elements into a folded tertiary structure. It is possible to interact with this output in a pymol session to arrange the secondary structure elements to create desired tertiary topology. This new pdb can be fed into the protocol using the previously described input methods and the geometry of the desired topology with constructed as before. Additionally, any constraint that is available as part of the Rosetta macromolecular modeling software, i.e., for enzyme design, can also be used during the creation of starting structures (Figure 3.8 Input).

De novo Tertiary Topology Generator

Once the geometry of the desired topology has been defined, the next step is to prepare for a fragment based computational folding protocol (Figure 3.8 Preparation for Fragment Assembly). The computational folding protocol used here is a modification of the highly successful Rosetta ab initio structure prediction method. The Rosetta Ab Initio structure prediction method uses pieces, fragments, of high-resolution protein structures from the PDB. These fragments are three or nine long and are assembles into collapsed models.

A single helix(H), strand(E), loop(L) pattern maybe defined or a range of helix, strand, and loop lengths may be given, and a "random" pattern will be chosen. The selection of the H/E/L pattern in an automated fashion is desirable for several reasons. The number of possible H/E/L patterns that can be searched is much large in an automated fashion and more importantly the design will not be biased either positively or negative by the intuition of the protein designer, giving a more accurate test of the computational method. Fragments can be created using the Robetta Server, the released version of Rosetta FragmentPicking, or on the fly during the protocol. On the fly fragment creation offers the greatest flexibility when testing a diverse set of helix, strand, and loop length conditions for one or many topologies. In addition to the fragments that will be read in from any provided fragment files, another set of fragments will be created on the fly, these fragments are called bridge fragments. Bridge fragments are fragments that connect a helix/strand to another helix/strand, and so consist of (H/E)L(H/E), they can be of any length and are chosen such that they will span the desired helix/strand to helix/strand gap as measured by distance, dihedral angle, and rmsd of the axis points and vectors that define the topology and the potential fragment. This method produces fragments that are much more tightly focused towards the desired topology. An extra step of fragment creation is also done for edge strands in β- sheets. These fragments are evaluated for features such as β-bulges, disruption of the binary hydrophobic polar pattern, and the presence of glycines and prolines. These edge strand fragments are enriched in an attempt to harness the naturally occurring negative design present in edge β-strands. The protocol evaluates the suitability of all other fragment candidate for the secondary structure element that it will be used to rebuild during the folding protocol based on satisfaction of the geometric definition of the desired topology. Prior to the folding simulation a detailed report of the quality of fragments found for the desired topology is written. This information can be used in future simulations to create improved starting structures.

The next step in the creation of a starting structure is the assembly of protein fragments into a single protein backbone (Figure 3.8 Assembly). The method used here is similar in spirit to the structure prediction routine used for Rosetta structure prediction in CASP, with some notable changes. The largest of these changes is the inclusion of a new type of fragment insertion, called a FragmentSequence insertion. The FragmentSequence insertion changes the phi psi and omega dihedrals of a region of the protein chain to match those of the fragment. It also changes the protein sequence of the model to match the fragment for residues, which are proline, preprolines, glycines, and any positions that match the intended hydrophobic/polar pattern of the desired topology, if one was specified. If a desired hydrophobic polar pattern was not specified then a pattern will be created during the definition of the topology. This pattern is chosen based on the burial of side-chains, and only specifies positions as hydrophobic in completely buried, and polar if the position is ≥ 80% solvent exposed. These constraints on the sequence identity do not propagate to the design. FragmentSequnce insertions do not lead to a loop sequence of all prolines or glycines. If a new fragment is inserted overtop of a previously inserted proline or glycines either the new amino acid is accepted or the original amino acid (not the proline) is returned. This method is intended to capture some of the positive and negative design elements seen in the native sequences of loops. For the same reasons we also use FragmentSequence insertions for edge strands in an attempt to preserve edge strand features. While it is true that Rosetta often designs a proline or a glycine into positions where those residues were naturally in a fragment, we use FragmentSequence insertions in the hope to capture more information.

The folding protocol also incorporates constraints that are defined by the geometric definition of the topology and once the protein chain has collapsed the constraints are relaxed and eventually removed entirely.

We call these two features, design "*in medias res*", because we are designing the protein sequence "in the middle" of creating the starting structure backbone. When the protein folding protocol is finished if the generated backbone meets the geometric definition of the topology the structure is passed forward to the flexible backbone design stage. Otherwise the protocol begins again with the same H/E/L loop pattern chosen initially. If a particular H/E/L loop pattern fails to meet the geometric definition of the topology for 10 consecutive simulations, that H/E/L pattern is blacklisted and will not be further sampled.

Flexible Backbone Protein Design

Here we use a simple strategy that has worked in the past. We couple the high-resolution structure refinement method of Rosetta with Rosetta's successful design algorithm. The protocol consists of iterative cycles of protein design followed by high-resolution structure refinement until the difference in energy between cycle i and i+1 is less than 1.0 Rosetta energy unit. If the final design that is created is not less than -2.5 Rosetta Energy units, the average energy for refined high-resolution crystal structures, the structure is rejected and the protocol begins again at the assembly stage. If the design model is less than -2.5 REU the structure is accepted, saved, and the protocol begins again at the assembly stage. If an H/E/L pattern leads to 10 successive rejections, that pattern is black listed (Figure 3.8 Design).

Selection of Sequences for experimental characterization

Several sequences were chosen for experimental characterization. The sequences chosen all scored in the best 10% for Rosetta total score and have favorable packing in the protein core (packstats), and do not have any buried polar unsatisfied atoms. The sequences are also evaluated for their predicted secondary structure using the JPRED3 server. All of the sequences chosen for experimental characterization have similar secondary structure profiles to the desired topology. In addition to these metrics, sequences were "folded" using Rosetta's structure prediction method to evaluate if the sequences were predicted to adopt the desired fold and to identify any low energy alternative folds. For the helical designs, the desired fold was recovered, for the β-sandwich folds a clever computational technique was used to evaluate the probability of adopting the desired folded state. All of the β-folds tested here are six-stranded β-sandwich folds with three strands in each sheet. All possible combinations of right handed physically possible topologies, 12 in all(Woolfson, Evans et al. 1993), were defined using the *de novo* tertiary topology creator. Using the geometric constraints and focused fragments from the *de novo* tertiary topology creator a standard *ab initio* run was performed. The models created for each different possible topology were highly biased towards that topology. Additionally, a standard an initio run was conducted to sample the remaining landscape. This technique overcomes the fact that β-sandwich topologies are notoriously hard to fold using Rosetta structure prediction methods because of the complicated topology and the large lever effects due to small strand movements. While this method for predicting the preference of a designed β-sandwich sequence for the target fold doesn't explore all possible alternative states, it clearly out performs traditional *ab initio*.

CS-Rosetta

The CS-Rosetta protocol was followed as specified in Shen et al. Briefly the protocol follows the traditional Rosetta structure prediction method of fragment assembly but fragments are chosen based on the similarity of sequence, secondary structure, and chemical shift information compared to a set of known chemical shifts. The method has been shown to be reliable for a diverse set of protein folds.

CHESHIRE

Cheshire is a structure prediction method that uses chemical shifts to solve NMR structures; it is part of the modeling software Almost. The method uses fragment assembly and chooses fragments based on the similarity between assigned chemical shifts and known chemical shifts. Unlike CS-Rosetta, this method has not been parameterized against X-ray crystal structure and the energy function and conformational search methods originate from molecular mechanics. The protocol used here follows the procedure outlined in Cavalli et al.

Experimental Methods

Cloning, Expression, and Purification

A codon optimized gene for each *de novo* sequence was purchased from Genscript,

```
>dnd_4hb
mQEERKKLLEKLEKILDEVTDGAPDEARERIEKLAKDVKDELEEGDAKNMIEKFRDEMEQMY
KDAPNAVMEQLLEEIEKLLKKAgsylvprgslehhhhhh*


>dnd_sam
mDEDQMKKRLEKGDKDELKDWLEKTGNGSWEELERGNEAPMIERLGLPPEDKKKMEQHIREI
NEDQRKNDgsylvprgslehhhhhh*
```

```
>dnd_nbs 1
REIEIETNGVKVRVRGCQVTVTYDNAGKTTIHAGTVEVRVHGGDVTITSRCS

>dnd_nbs 2
NTFKFRRGGVDVEVDGCQWTADTRDGARAQWHGDGVTVRVRNGDADVQSDCG

>dnd_nbs 3
RRTTVKRGGVKVTVYNGKVDVDVEQGARARIHIGTVEVDADGTDVDIQKR

>dnd_xbs
mgsylvprgslehhhhhh*
```

lowercase letters are due to cloning and capital letters are the designed sequence. Each gene was supplied as 4 μg of lyophilized DNA in puc57 vector. The gene of interest was pcr amplified out of the parent vector, purified using a pcr clean up kit from Fermentas, double digested with NdeI and XhoI from NEB, and purified again using a pcr clean up kit, and finally ligated into Pet21b vector from Novagen which had been prepared by double digesting with NdeI and XhoI and using a Fermentas gel extraction clean up kit. The ligation reaction was transformed into XL-10 Gold cells from Stratagene.

Each protein was expressed in BL21(DE3) pLysS cells from Stratagene. Cells were grown in LB media with 100 mg/ml ampicilin at 37°C to an OD_{600} of 0.6 and induced with 0.5 mM IPTG for 12 hours at 16°C. Cells were centrifuged at 4500 x g for 30 minutes and cell pellets were resuspended in 0.5 M NaCl, 0.2 M NaK pH 7.0, 10% glycerol, 1% triton, dtt, and treated with dnase, rnase, benzamidine, and pmsf after three rounds of sonication at 70% power for 45 seconds. The cell lysate was cleared twice by centrifugation at 18,000 x g for 30 minutes. The supernatants were then filtered using a 0.22 μM filter from Millipore. The supernatant was purified using a

HisTRAP from GE Healthcare. The elution was concentrated to 2 mls and further purified on a Superdex S75 gel filtration column.

Circular Dichroism

CD data were collected on a Jasco J-815 CD spectrometer. Far-UV CD scans were collected using a 1 mm cuvette at concentrations between 30-40 μM protein in 50 μM sodium phosphate at pH 7.4 and 20°C. Thermal denaturation of samples was conducted between 4°C and 97°C while measuring CD signal at 208 nm and 222 nm. Chemical denaturation by guanidine chloride (GuCl) was done by titrating a sample of 30 uM designed protein in 0M GuCl into a sample of 30 uM designed protein with 7.8 M GuCl. The GuCl concentration was monitored by refractive index. Thermodynamic parameters were calculated assuming that the folding of the designed protein was a two state process and by fitting both the thermal and chemical denaturations to the Gibbs-Helmholtz equation.

Nuclear Magnetic Resonance Spectroscopy

The designed proteins were concentrated to ~1 mM in 20 mM Sodium Phosphate pH 6.5 with 10% D_2O. H1 NMR spectra were collected on Varian Inova 600 and 700 MHz spectrometer at 25°C. NMR data and figures were processed using NMRPipe and NMRDraw. For double labeled, ^{15}N and ^{13}C, NMR experiments, designed protein was grown and purified in the same manner except that minimal media with ^{13}C glucose and ^{15}N ammonium chloride were was the cell medium during induction. A series of experiments ^{15}N-HSQC, ^{13}C-HSQC, CBCACONH, HBHACONH, HNCACB, and HNCACO were performed to assign the protein backbone atoms for dnd_4hb.

Protein Crystallization and X-ray Crystallography

The dnd_4HB designed protein was crystallized in 0.2 M ammonium acetate, 0.1 M tri-sodium citrate and 30% w/v 2-methyl-2,4-pentanediol, by the sitting drop method at the nano-scale using a Rigaku Phoenix liquid handling robot. Diffraction data was collected at the APS GM/CA-CAT beam line.

Figures

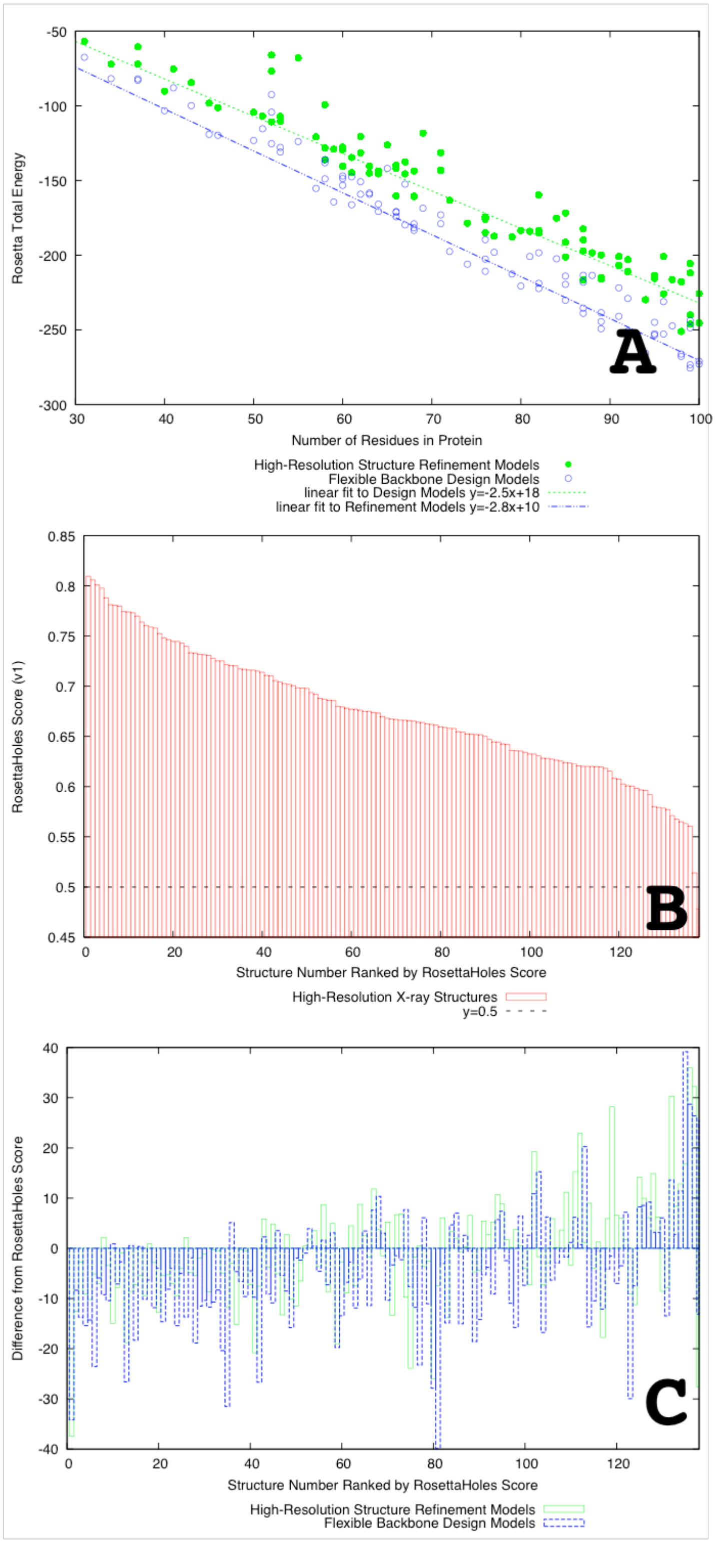


Figure 3.1: Comparison of native proteins, high-resolution refined native proteins, and flexible backbone designed native proteins.

Plotting number of residues in a protein versus the Rosetta energy shows the relationship between chain length and decreasing energy. The average energy per residue of a Rosetta high-resolution refined native protein is -2.5 (green). The average energy per residue of a flexible backbone designed native protein is -2.8 (blue). The slopes of the lines of best fit give the average energy per residue (A). The quality of packing of natives (red) (B), the difference in packing quality between natives, refined natives (green), and designed natives (blue) shows that natives have the best packing but refined and resigned Natives are similar to each other (C).

Figure 3.2: Example Output from *de novo* Design Protocol.

The *de novo* design protocol was computationally benchmarked on a diverse set of proteins, a four helix bundle (dnd_4hb) (A), a helical SAM domain (dnd_sam) (C), a novel β-sandwich fold (dnd_nbs) (B), a β-grasp fold (dnd_grp) (D), and an existing β-sandwich fold (dnd_xbs) (E). Folds A, B, C, and E were tested experimentally.

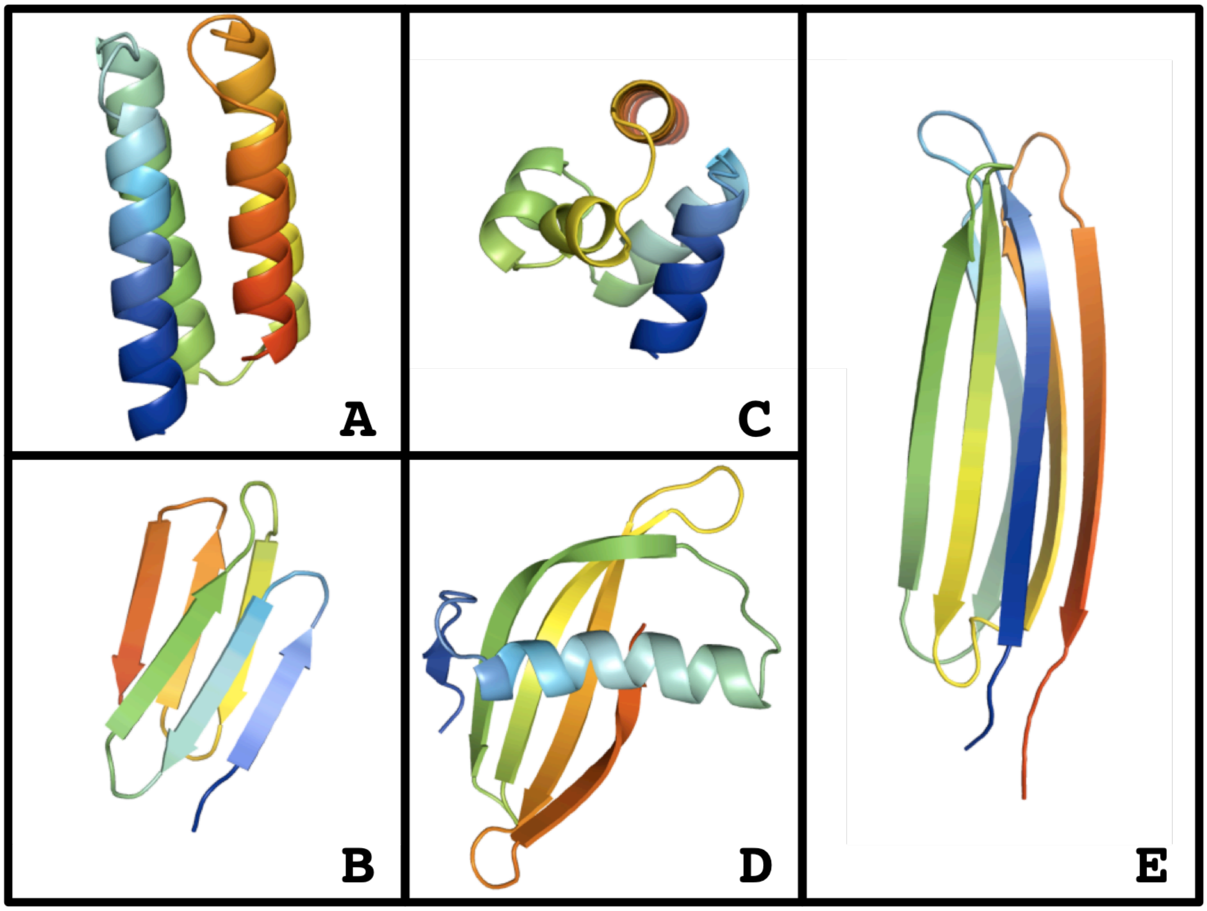
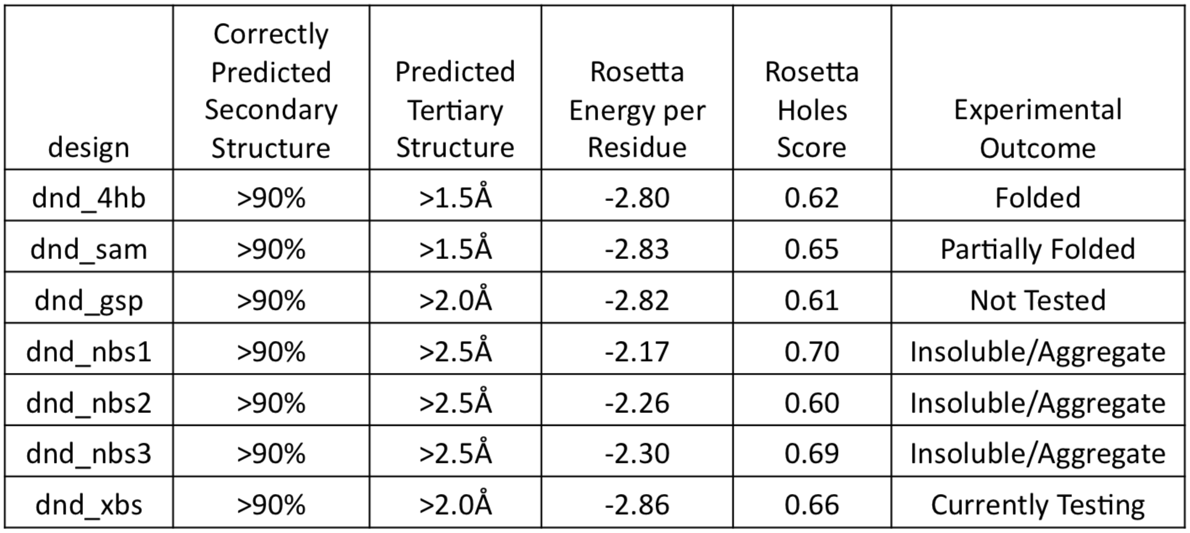


Figure 3.3. Comparison of designed sequences chosen for experimental validation.

A set of seven sequences was chosen for experimental characterization, five have been tested, one is currently being tested (dnd_xbs), and one has not been tested (dnd_grp). Computationally, all seven designs are indistinguishable from native proteins in several metrics. The JPRED3 server was used to predict the designed sequences secondary structure, all matched the model at > 90%. The Rosetta structure prediction method was used to predict each designs tertiary structure and all designs were predicted with 2.5 Å rmsd of the design model. All designs had favorable Rosetta energies, and RosettaHoles scores, however dnd_xbs sequences had significantly lower Rosetta Energy per Residue.



design	Correctly Predicted Secondary Structure	Predicted Tertiary Structure	Rosetta Energy per Residue	Rosetta Holes Score	Experimental Outcome
dnd_4hb	>90%	>1.5Å	-2.80	0.62	Folded
dnd_sam	>90%	>1.5Å	-2.83	0.65	Partially Folded
dnd_gsp	>90%	>2.0Å	-2.82	0.61	Not Tested
dnd_nbs1	>90%	>2.5Å	-2.17	0.70	Insoluble/Aggregate
dnd_nbs2	>90%	>2.5Å	-2.26	0.60	Insoluble/Aggregate
dnd_nbs3	>90%	>2.5Å	-2.30	0.69	Insoluble/Aggregate
dnd_xbs	>90%	>2.0Å	-2.86	0.66	Currently Testing

Figure 3.4 Predicted tertiary structures of designed sequences.

As a metric for the favorability of a designed sequence for the desired fold, designed sequences were folded using Rosetta's *ab initio* structure prediction method. Three examples are shown where the fold was correctly predicted, dnd_sam (A), dnd_4hb (B), and dnd_nbs(C).

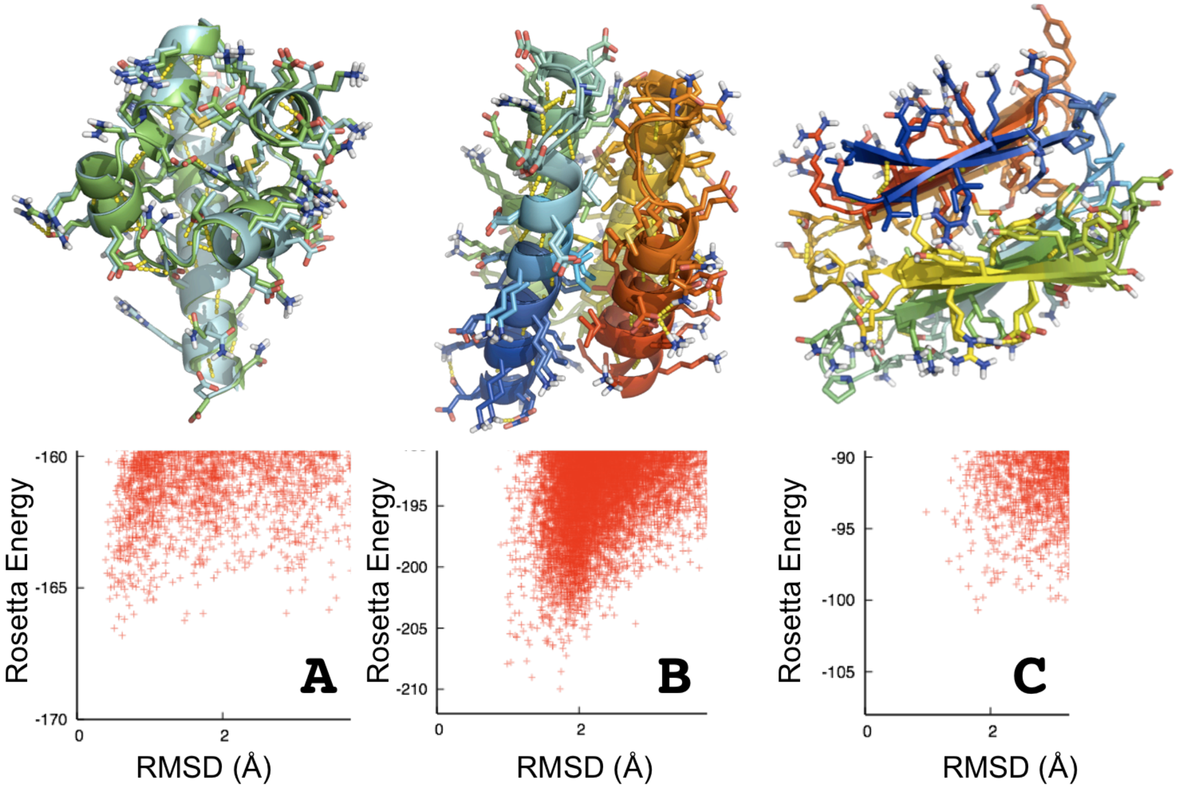
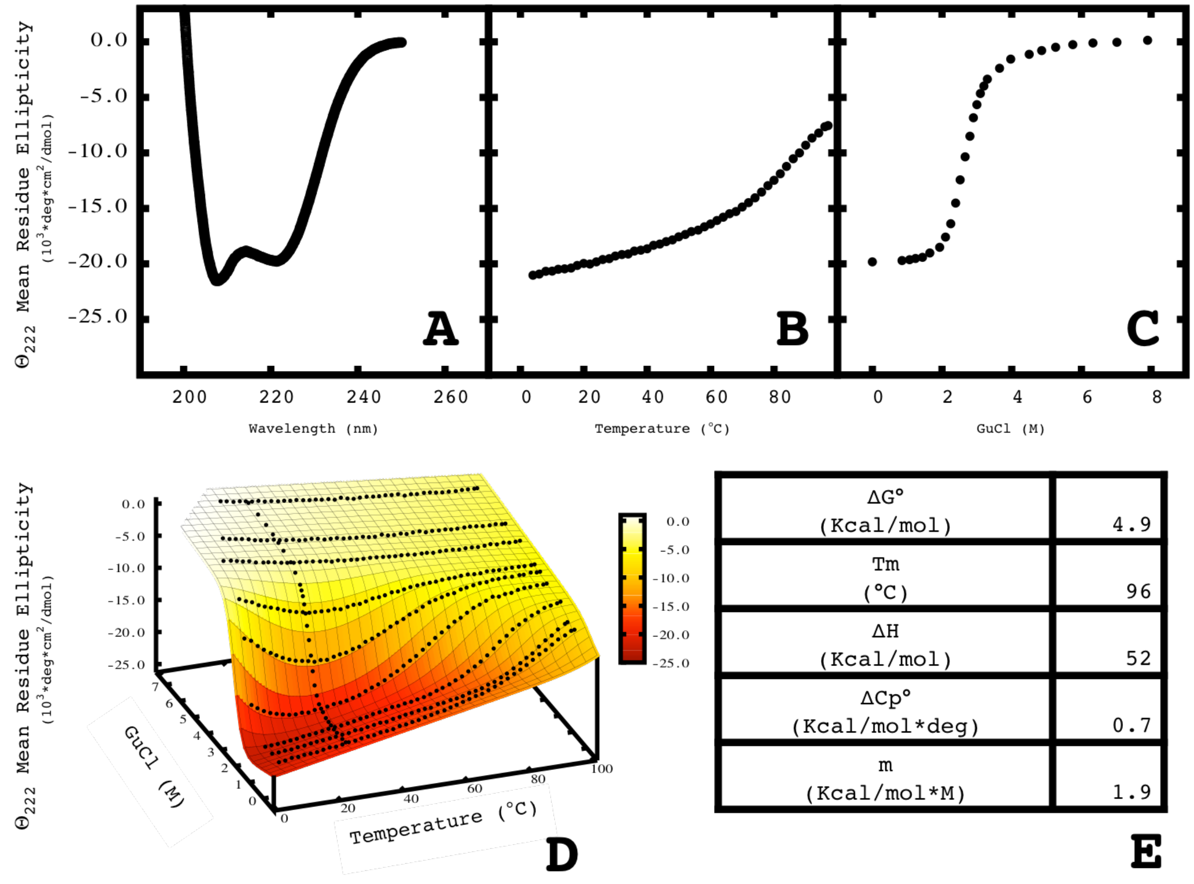


Figure 3.5: Biophysical characterization of dnd_4hb.

Far UV Circular Dichroism (A), Thermal Denaturation (B), and Chemical Denaturation (C) of dnd_4hb. Global fits (mesh) of the Gibbs-Helmholtz equation to thermal and chemical denaturation data for dnd_4hb (D). Thermodynamic parameters calculated by fitting the Gibbs-Helmholtz equation to thermal and chemical denaturation data for dnd_4hb(E). All experiments were done at 30-40 uM protein concentration in 50 μM sodium phosphate at pH 7.4.



ΔG° (Kcal/mol)	4.9
Tm (°C)	96
ΔH (Kcal/mol)	52
ΔCp° (Kcal/mol*deg)	0.7
m (Kcal/mol*M)	1.9

Figure 3.6: Predicted structure of NMR chemical shift data.

The assigned backbone peaks (A) were used in CHESHIRE to predict the structure of dnd_4hb. The global agreement between the design model (green) and the CHESHIRE prediction (cyan) is good, with backbone RMSD of ~2.5Å (B). Many of the atomic level details are also accurately predicted (C).

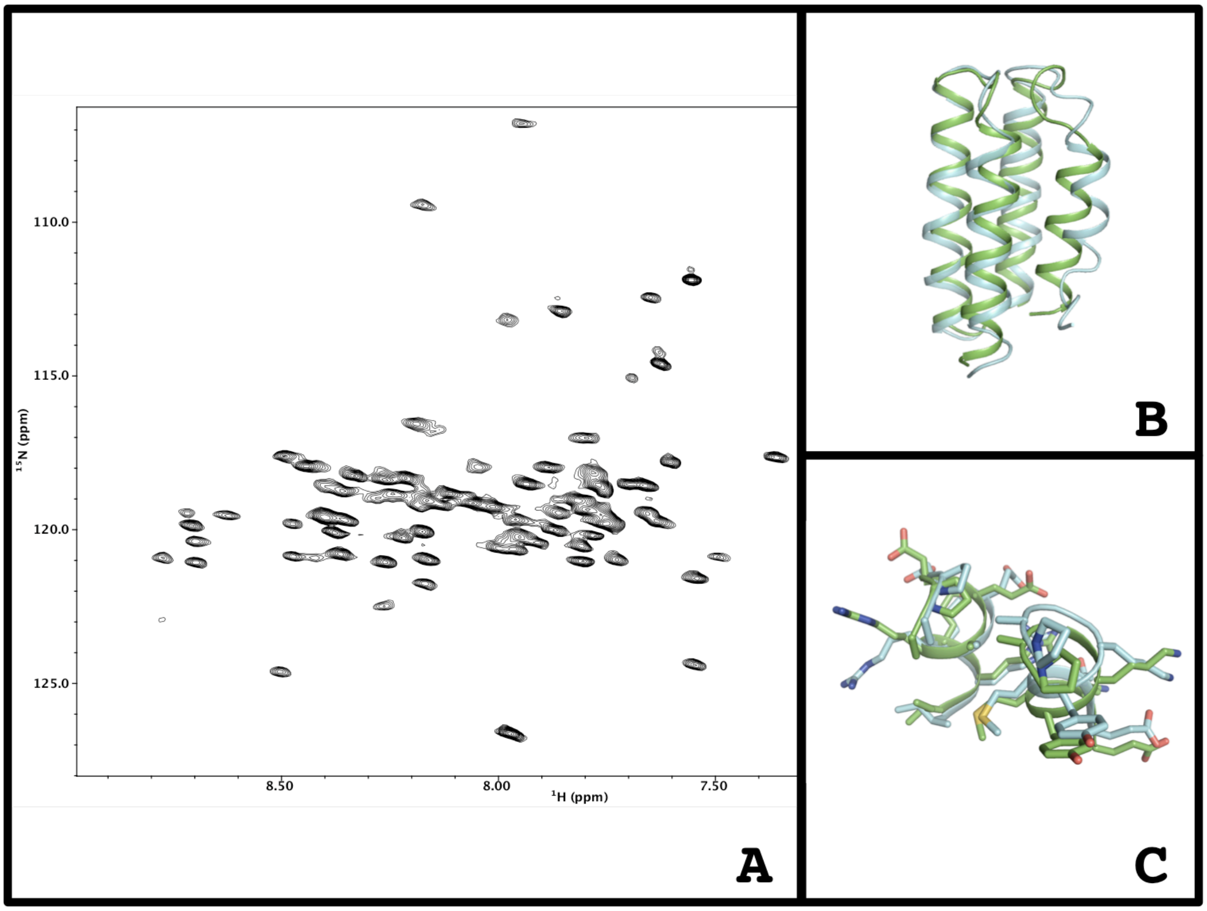
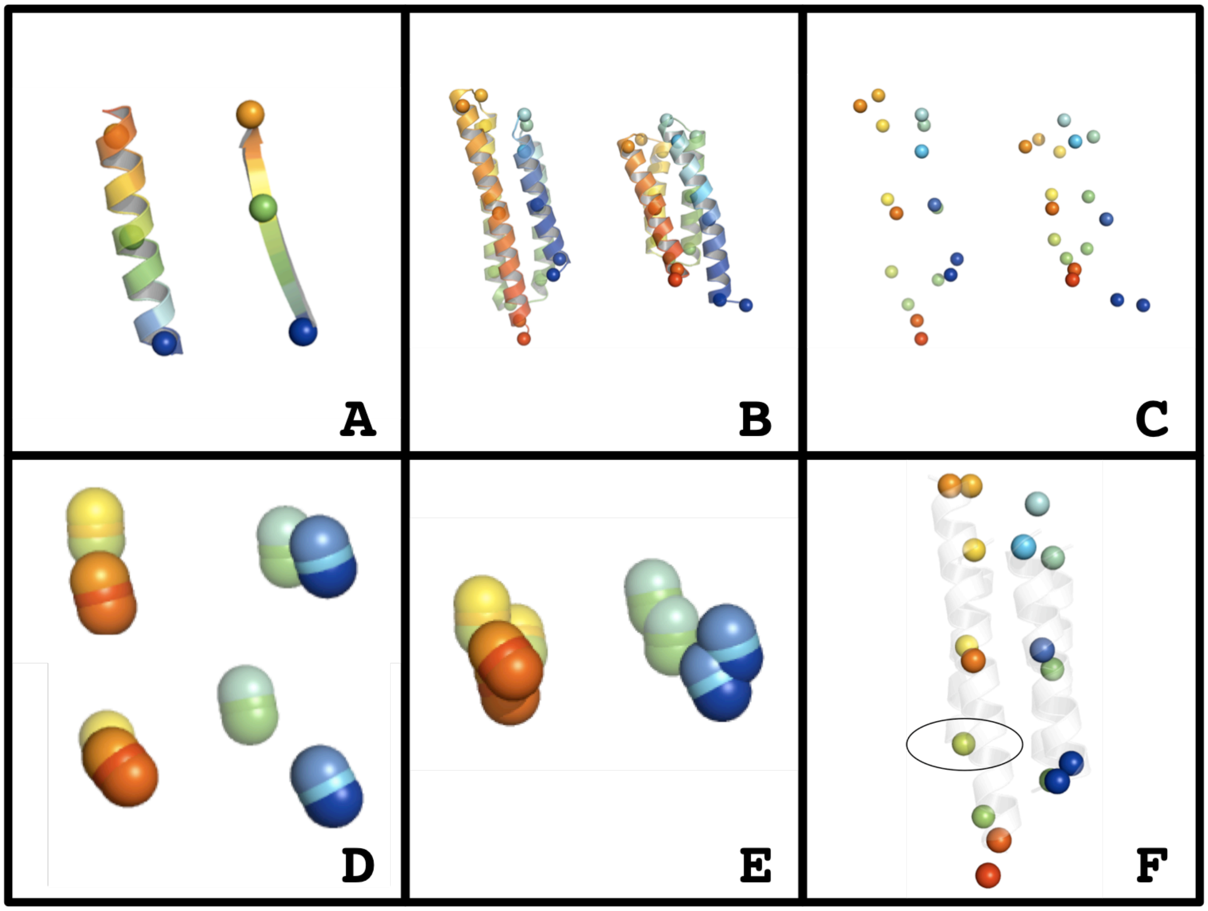


Figure 3.7: Reduced representation of protein structure and deriving geometric constraints for *de novo* design structure creation.

Defining helix and strand elements in a reduced representation as three points (beads) or two vectors is a convenient description of the helix or stand length, direction, and orientation (A). Using this reduced representation of an entire fold leads to simplification of tertiary structure, from an all atom model or Cα trace (B), to only the beads that define the secondary structure element (C). It is possible to normalize the length of each vector in the reduced representation to compare homologous proteins (D), and it is possible to align the normalized reduced representations (E) to define average positions and vectors, and the deviations (F, circle) about those positions and vectors to create geometric bounds defining the desired topology. This information is used to create *de novo* vector definitions of the topology by varying the length of helix, strand and loop elements (F, a possible solution is given in gray cartoon).



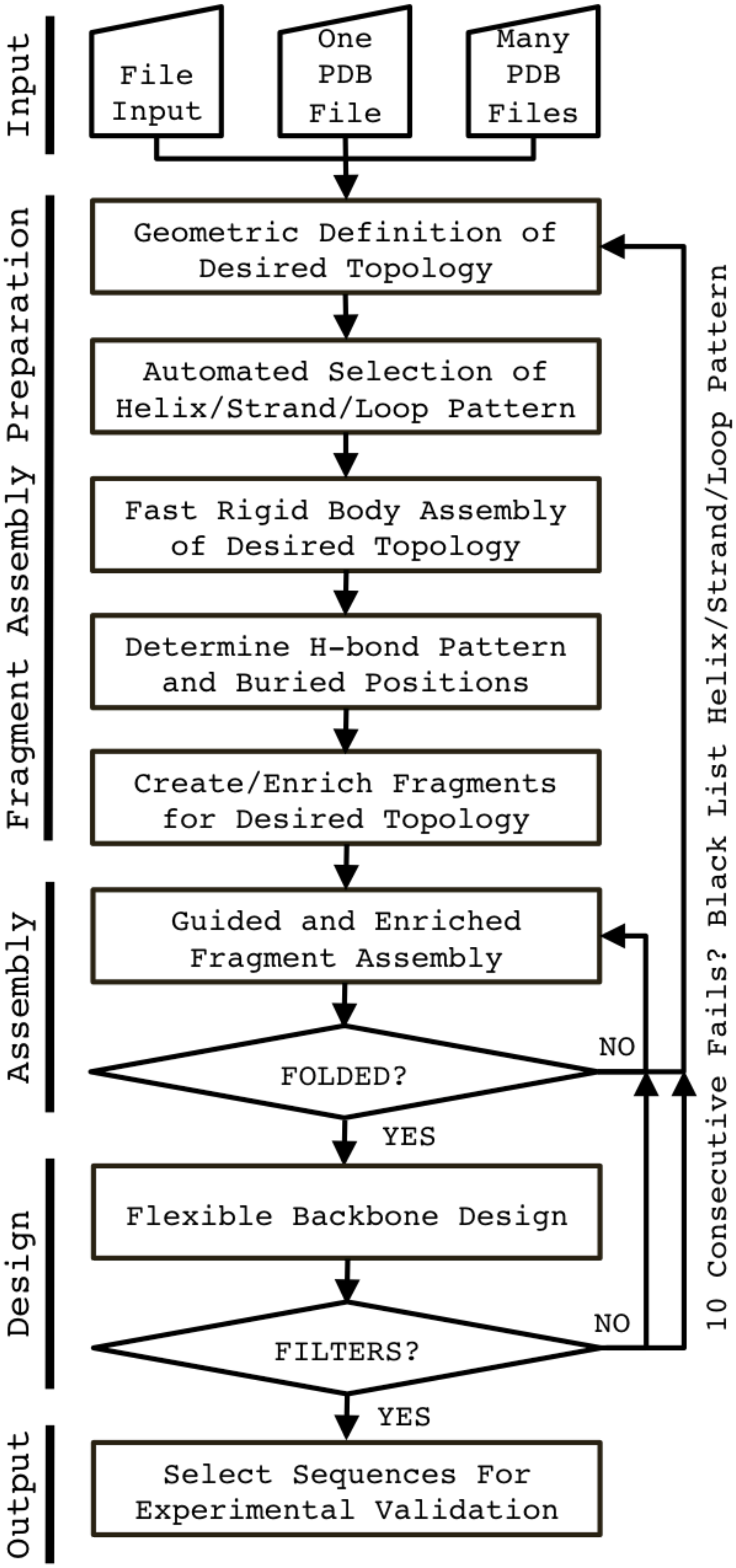
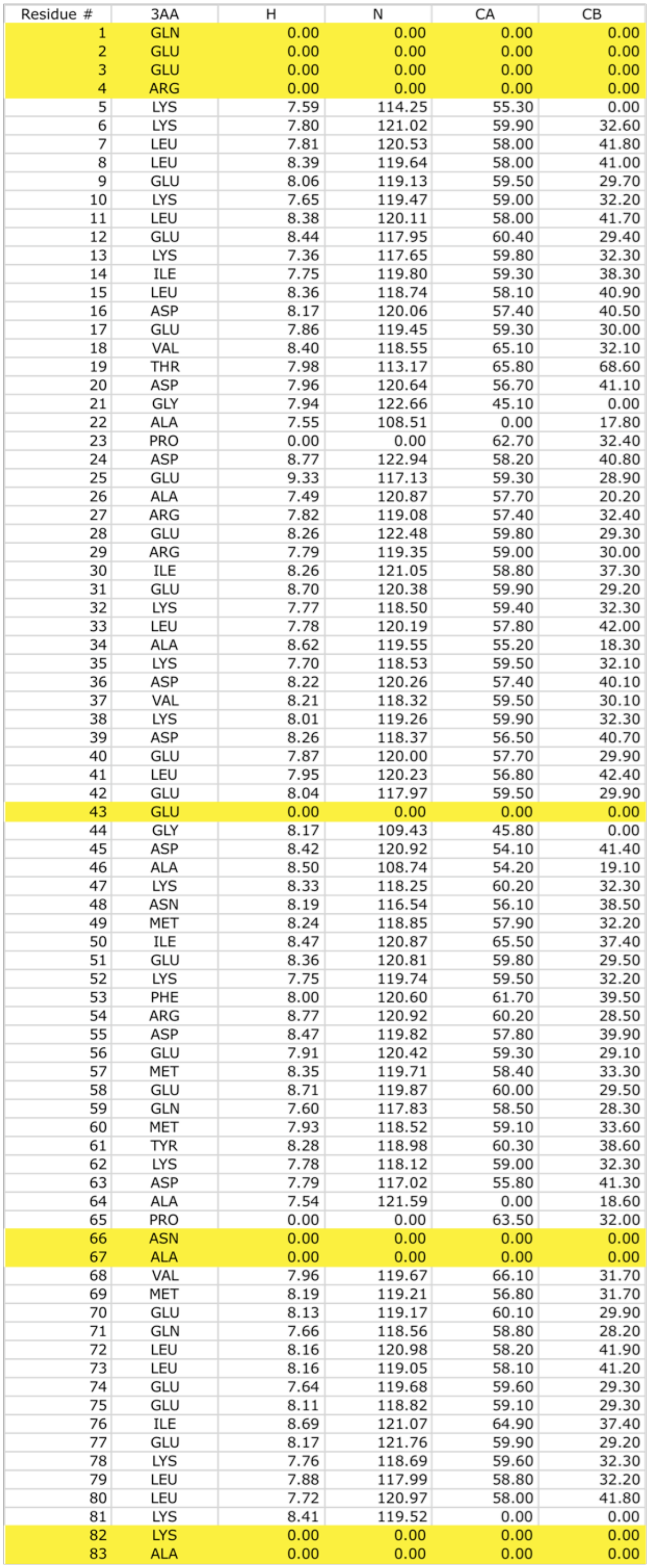

Input
File Input
One PDB File
Many PDB Files
Fragment Assembly Preparation
Geometric Definition of Desired Topology
Automated Selection of Helix/Strand/Loop Pattern
Fast Rigid Body Assembly of Desired Topology
Determine H-bond Pattern and Buried Positions
Create/Enrich Fragments for Desired Topology
Assembly
Guided and Enriched Fragment Assembly
FOLDED?
NO
YES
Design
Flexible Backbone Design
FILTERS?
NO
YES
Output
Select Sequences For Experimental Validation
10 Consecutive Fails? Black List Helix/Strand/Loop Pattern

Figure 3.8: *De novo* Design Protocol Flow Chart

The *de novo* design protocol can be divided into five conceptual steps: Input(1), Fragment Assembly Preparation(2), Assembly(3), Design(4), and Output(5). The details of the method are described in the Results and the Materials and Methods.

Table 3.1 Chemical shift Assignments for dnd_4hb.

The chemical shifts for the H, N, Cα, and Cβ atoms for dnd_4hb were assigned from the ^{15}N-HSQC, ^{13}C-HSQC, CBCACONH, HBHACONH, HNCACB, and HNCACO spectra. Positions highlighted in yellow could not be reliably assigned due to peak overlap.



Residue #	3AA	H	N	CA	CB
1	GLN	0.00	0.00	0.00	0.00
2	GLU	0.00	0.00	0.00	0.00
3	GLU	0.00	0.00	0.00	0.00
4	ARG	0.00	0.00	0.00	0.00
5	LYS	7.59	114.25	55.30	0.00
6	LYS	7.80	121.02	59.90	32.60
7	LEU	7.81	120.53	58.00	41.80
8	LEU	8.39	119.64	58.00	41.00
9	GLU	8.06	119.13	59.50	29.70
10	LYS	7.65	119.47	59.00	32.20
11	LEU	8.38	120.11	58.00	41.70
12	GLU	8.44	117.95	60.40	29.40
13	LYS	7.36	117.65	59.80	32.30
14	ILE	7.75	119.80	59.30	38.30
15	LEU	8.36	118.74	58.10	40.90
16	ASP	8.17	120.06	57.40	40.50
17	GLU	7.86	119.45	59.30	30.00
18	VAL	8.40	118.55	65.10	32.10
19	THR	7.98	113.17	65.80	68.60
20	ASP	7.96	120.64	56.70	41.10
21	GLY	7.94	122.66	45.10	0.00
22	ALA	7.55	108.51	0.00	17.80
23	PRO	0.00	0.00	62.70	32.40
24	ASP	8.77	122.94	58.20	40.80
25	GLU	9.33	117.13	59.30	28.90
26	ALA	7.49	120.87	57.70	20.20
27	ARG	7.82	119.08	57.40	32.40
28	GLU	8.26	122.48	59.80	29.30
29	ARG	7.79	119.35	59.00	30.00
30	ILE	8.26	121.05	58.80	37.30
31	GLU	8.70	120.38	59.90	29.20
32	LYS	7.77	118.50	59.40	32.30
33	LEU	7.78	120.19	57.80	42.00
34	ALA	8.62	119.55	55.20	18.30
35	LYS	7.70	118.53	59.50	32.10
36	ASP	8.22	120.26	57.40	40.10
37	VAL	8.21	118.32	59.50	30.10
38	LYS	8.01	119.26	59.90	32.30
39	ASP	8.26	118.37	56.50	40.70
40	GLU	7.87	120.00	57.70	29.90
41	LEU	7.95	120.23	56.80	42.40
42	GLU	8.04	117.97	59.50	29.90
43	GLU	0.00	0.00	0.00	0.00
44	GLY	8.17	109.43	45.80	0.00
45	ASP	8.42	120.92	54.10	41.40
46	ALA	8.50	108.74	54.20	19.10
47	LYS	8.33	118.25	60.20	32.30
48	ASN	8.19	116.54	56.10	38.50
49	MET	8.24	118.85	57.90	32.20
50	ILE	8.47	120.87	65.50	37.40
51	GLU	8.36	120.81	59.80	29.50
52	LYS	7.75	119.74	59.50	32.20
53	PHE	8.00	120.60	61.70	39.50
54	ARG	8.77	120.92	60.20	28.50
55	ASP	8.47	119.82	57.80	39.90
56	GLU	7.91	120.42	59.30	29.10
57	MET	8.35	119.71	58.40	33.30
58	GLU	8.71	119.87	60.00	29.50
59	GLN	7.60	117.83	58.50	28.30
60	MET	7.93	118.52	59.10	33.60
61	TYR	8.28	118.98	60.30	38.60
62	LYS	7.78	118.12	59.00	32.30
63	ASP	7.79	117.02	55.80	41.30
64	ALA	7.54	121.59	0.00	18.60
65	PRO	0.00	0.00	63.50	32.00
66	ASN	0.00	0.00	0.00	0.00
67	ALA	0.00	0.00	0.00	0.00
68	VAL	7.96	119.67	66.10	31.70
69	MET	8.19	119.21	56.80	31.70
70	GLU	8.13	119.17	60.10	29.90
71	GLN	7.66	118.56	58.80	28.20
72	LEU	8.16	120.98	58.20	41.90
73	LEU	8.16	119.05	58.10	41.20
74	GLU	7.64	119.68	59.60	29.30
75	GLU	8.11	118.82	59.10	29.30
76	ILE	8.69	121.07	64.90	37.40
77	GLU	8.17	121.76	59.90	29.20
78	LYS	7.76	118.69	59.60	32.30
79	LEU	7.88	117.99	58.80	32.20
80	LEU	7.72	120.97	58.00	41.80
81	LYS	8.41	119.52	0.00	0.00
82	LYS	0.00	0.00	0.00	0.00
83	ALA	0.00	0.00	0.00	0.00

Rosetta Command Lines

The following Rosetta command lines were used in this research.

To generate *de novo* designs from a single pdb file in an automated fashion

~/DeNovoTertiaryTopologyCreatorApp.macosgccrelease –database ~/database/ -s *.pdb -DNTTCF DNTTC_file.txt –use_pdb_based_info true –create_hbond_pattern true

To generate *de novo* designs from a family of pdb files in an automated fashion

~/DeNovoTertiaryTopologyCreatorApp.macosgccrelease –database ~/database/ -l pdb_list_file -DNTTCF DNTTC_file.txt –use_pdb_based_info true –create_hbond_pattern true

pdb_list_file is a file containing a list of pdb file names, one per line

To output the broken topology for interaction in pymol use the command line flag,

-dump_broken_pose

To view the quality of fragments selected versus the input structures and output structures use the flag –dump_frags_as_pdb

To generate *de novo* design of novel folds without a pdb file use only a DNTTC file, which can have the following information

```
SEQUENCE    Any of the 20 amino acid codes as caps or lowercase a for any
SECSTRUCT   SECTRUCT PATTERN OF H/E/L
HP_PATTERN  Hydrophobic Polar Pattern of H/P/A(any)
BURIAL            Pattern of – for unknown, B for buried, S for surface
POSE_SIZE   Min-Max Range
STRAND_STRAND_PAIRING 4 2 A - indicates strand 4 and 2 are paired
antiparallel
STRAND_STRAND_PAIRING 2 8 P – indicates strand 2 and 8 are paired parallel
STRAND_STRAND_PAIRING 10 14 X – indicates strand 10 and 14 are in unknown
pairing
STRAND_STRAND_LAYER 2 10 A – indicates strands 2 and 10 are in different
sheets but have side chains in contact, this controls how two sheets
interact with each other
SHEET_LAYER 4 2 8 – indicates that the sheet architecture is 4 2 8
HELIX_HELIX_LAYER 3 5 A – indicates helices 3 and 5 are in antiparallel
contact
HELIX_STRAND_LAYER 4 5 A –helix 5 and strand 4 are antiparallel contact
The ELEMENT keyword can be used to modify the state of a single strand,
helix or loop element
ELEMENT      6  SIZE  16 20 H ..----BBB----.. ..----FWG----..
States that element 6 is between 16 and 20 residues long composed of helix
and has a buried sequence of FWG in the middle of the helix, the .. states
that the pattern maybe shifted
```

All of the above information can be describe, but only the SECSTRUCT is required,

unless full descriptions with ELEMENT are given for all helices, strands, and loops.

References

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

Butterfoss, G. L. and B. Kuhlman (2006). "Computer-based design of novel protein structures." Annu Rev Biophys Biomol Struct **35**: 49-65.

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.

Gutte, B., M. Daumigen, et al. (1979). "Design, synthesis and characterisation of a 34-residue polypeptide that interacts with nucleic acids." Nature **281**(5733): 650-5.

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

Hecht, M. H., J. S. Richardson, et al. (1990). "De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence." Science **249**(4971): 884-91.

Hu, X., H. Wang, et al. (2007). "High-resolution design of a protein loop." Proc Natl Acad Sci U S A **104**(45): 17668-73.

Hu, X., H. Wang, et al. (2008). "Computer-based redesign of a beta sandwich protein suggests that extensive negative design is not required for de novo beta sheet design." Structure **16**(12): 1799-805.

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. and W. F. DeGrado (2009). "Engineering and design: editorial overview." Curr Opin Struct Biol **19**(4): 440-1.

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.

Quinn, T. P., N. B. Tweedy, et al. (1994). "Betadoublet: de novo design, synthesis, and characterization of a beta-sandwich protein." Proc Natl Acad Sci U S A **91**(19): 8747-51.

Regan, L. and W. F. DeGrado (1988). "Characterization of a helical protein designed from first principles." Science **241**(4868): 976-8.

Wei, Y., S. Kim, et al. (2003). "Solution structure of a de novo protein from a designed combinatorial library." Proc Natl Acad Sci U S A **100**(23): 13270-3.

Woolfson, D. N., P. A. Evans, et al. (1993). "Topological and stereochemical restrictions in beta-sandwich protein structures." Protein Eng **6**(5): 461-70.

Zhang, Z. and H. S. Chan "Competition between native topology and nonnative interactions in simple and complex folding kinetics of natural and designed proteins." Proc Natl Acad Sci U S A **107**(7): 2920-5.

Chapter 4

Conclusions, Future Directions and the Future of Protein Design

In the previous two chapters, we have presented computational methods and experimental results indicating that we are rapidly advancing our understanding of the rules that control protein folding and stability. Our primary goals have been the creation of computational techniques that increase the success rate of protein design. We believe the creation of designable protein backbones to be the limiting step in many protein design challenges. Towards this goal we explored our ability to perturb naturally occurring protein backbones, which are known to be designable, and asked the question if we could maintain designability while dramatically remodeling the protein sequence. Next we probed our ability create designable protein backbones *de novo* for a diverse set of protein folds.

The creation of redesigned proteins that no longer resemble the wild type protein in sequence space, are now fully in the grasp of protein design. The redesign problem has been tested by a number of researchers at various levels of design and covering a large portion of protein fold space. We reported in Chapter Two, the redesign of a protein core where the designed sequence has 0% sequence identity to the wild-type protein over the designed positions. To achieve this, we implemented a flexible backbone protein design protocol that attempted to improve the fitness of backbone conformation for the designed sequence in an iterative fashion. This computational technique appears to be robust. We used it here to successfully design eight four-helix bundle proteins, Kuhlman et al. used it in the *de novo* design of Top7, and recently several other design groups have used similar methods with success(Correia, Ban et al. ; Correia, Ban et al. ; Kuhlman, Dantas et al. 2003). Experimentally, we solved the X-ray and NMR structures for our most aggressively redesign helix bundle protein. The experimental structures are in close agreement with our design model. The all-atom rmsd of the X-ray crystal structure is 1.3 Å compared to the design model. The advantages of beginning a design project with a high-resolution crystal structure are undeniable but until this work we were limited to sequences highly similar to the wild type sequence. The ability to accurately predict the backbone conformation for highly dissimilar sequences has profound implications for protein-ligand design and enzyme design. We believe that if given a designable backbone our energy function, conformational search procedures and sequence search procedures can maintain protein backbone designability.

The success of protein redesign is in sharp contrast to the challenges still facing *de novo* protein design. We believe the great challenge in *de novo* protein design is the initial creation of a designable backbone. With this in mind, we developed a flexible and general framework for the creation of starting structures for *de novo* protein design. Computationally, the backbones and sequences generated by this protocol appear to be better than or comparable to native backbones and sequences as measured by the Rosetta total energy and the quality of core packing. However, our experimental results in *de novo* protein design indicate that our ability to create designable backbones and successful sequences for helix-bundle proteins is more advanced than our abilities for other protein folds. There are many aspects of protein folding and stability that have not yet been explicitly incorporated into the design of *de novo* proteins, and perhaps some of these features will be required for the *de novo* design of all β-folds. One feature that is lacking in current protein design algorithms is an explicit consideration of protein folding kinetics, obviously an important feature for all β-folds. This feature is not lacking due to ignorance but due to the inherent challenges of predicting the folding pathway for a hypothetical protein sequence. This challenge is further compounded by the fact that it would be necessary to evaluate the folding kinetics for several hypothetical sequences and then choose the best for experimental characterization. Hopefully, these limitations can be remedied using more advanced computational resources and new representations of protein folding pathways and kinetics.

Future protein design challenges focused on *de novo* protein design, especially of non helical protein folds, would benefit from exploring permutations of the core redesign strategy presented in Chapter Two. Flexible backbone protein redesign performed in a hierarchal method beginning initially with the redesign of only loops, only buried positions, or other functional important sites provides a backdoor to *de novo* design. While these designs would not be truly *de novo*, they would allow us to incrementally test our ability to design different aspects of protein folds with a high chance of success and the ability to rescue failed designs.

We still have a long way to go before we can reliably *de novo* design proteins with minimal time and effort as a step in a larger research endeavor. However, the future is bright for protein design. High performance research computing resources are increasingly becoming accessible to researchers with little or no computational background and the computational techniques and methods at the forefront of protein design are rapidly being passed to end users. This will facilitate the use of protein design as a tool in new fields, to answer new questions.

Waiting in the distance for protein design is the related field of synthetic biology. Synthetic biology opens a new horizon for protein design. Eventually, protein design will create not only *de novo* proteins but *de novo* proteins composed entirely of non-canonical amino acids, and to perform reactions that biology is not capable of performing, possibly leading to the creation of orthogonal macromolecular systems.

References

Correia, B. E., Y. E. Ban, et al. "Computational protein design using flexible backbone remodeling and resurfacing: case studies in structure-based antigen design." J Mol Biol 405(1): 284-97.

Correia, B. E., Y. E. Ban, et al. "Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope." Structure 18(9): 1116-26.

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