Computational design of β sheet proteins

Xiaozhen Hu

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 Biochemistry and Biophysics (Program in Molecular and Cellular Biophysics).

Chapel Hill
2008

Approved by

Advisor: Brian Kuhlman

Reader: Nikolay Dokholyan

Reader: Marshall Edgell

Reader: Jan Hermans

Reader: Andrew Lee

ABSTRACT

Xiaozhen Hu : Computational design of β sheet proteins

(Under the direction of Brian Kuhlman)

Computational protein design has become a very powerful approach to test our understanding of the forces and energetics of macromolecular systems. The ability to design proteins that have specific structures and functions will be very valuable to future protein drug discovery. Protein design technology has been successfully applied to stabilize proteins, increase protein-protein binding affinity and create new protein structures. However, *de novo* design remains very challenging, especially for β-sheet proteins. Most *de novo* designed β-sheet proteins tested to date either misfold or aggregate. In this thesis, we use a hierarchical approach to search for the bottleneck in β-sheet design. First, we tested our ability to redesign the sequence of a naturally occurring β-sheet protein. The molecular modeling program Rosetta was used to design new sequences for the β-sheet protein tenascin. The redesigned proteins are well-folded and have thermal melting temperatures that are 40 °C higher than the wild type. These results indicate that given a designable backbone we can create a well-folded β-sheet protein.

To move towards complete *de novo* design we next asked if we could design a portion of a β-sheet protein from scratch. We tested our ability to design loops by removing a ten-residue loop from tenascin and rebuilding it to have a new but specific conformation. These studies involved the simultaneous search of conformational and sequence space. Two of the designed loops were crystallized, and one of them adopts a structure that is very similar to the design model.

Lastly, we have explored designing whole β-sheet proteins from scratch. Four generations of designs have been tested to date, and unfortunately, none of the designs appear to be well folded. To lay the groundwork for future success, we have been comparing the design models to naturally occurring β-sheet proteins to identify structural features that may be missing from the designs. We find that naturally occurring proteins include fewer voids accessible to small probes (~ 0.7 Å) than our design models. It remains to be seen if more conformational sampling is need to remove these voids or if the energy function requires changes.

ACKNOWLEDGEMENTS

I would like to thank, first and foremost, Brian for giving me the great opportunity to pursue my doctoral dissertation under his direction. Brian has a beautiful mind and a great sense of scientific insight and intuition. He was always able to find a way to overcome the challenges that seemed to prevent my goals. When I encountered obstacles, he was always patient, understanding, and willing to provide guidance when needed. He also gave me a great degree of freedom to try my own ideas which taught me how to think and work independently. I thankful for all of those years working in his laboratory and none of my work would have been possible without his guidance. I would like to thank the members of the Kuhlman laboratory; we shared challenges, experiences and successes. We had a very comfortable and productive environment to work in — all of which have become very fond memories of UNC. I would like to thank my collaborators, Huancheng Wang and Hengming Ke, for their suggestions and help in the crystal structures. I would like to thank my thesis committee members Nikolay Dokholyan, Marshall Edgell, Jan Hermans and Andrew Lee for their great advice and suggestions. I would like to thank Barry Lents for the opportunity to join the Molecular and Cellular Biophysics Training Program. I would like to thank Ashutosh Tripathy from MacInFac, Greg Young from the NMR facility and Laurie Betts from the X-ray crystallography facility for their kind help and support. Last but certainly not least, I would like to thank my parents, Zhenming and Linai, for believing in me and loving me for whoever I am. They have been always there, supporting and encouraging me. Nothing would have been possible without their support. I would like to thank my little girl Audrey, who let me experience the joy of a being a mother and understand how wonderful life is. I would also like to thank my husband, Yibing, for his understanding and constant support .

TALBE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ΔG: free energy of unfolding

CD: circular dichroism

C_m: midpoint of guanidine chloride unfolding transition

E. Coli: Escherichia coli

FNIII: fibronectin type III domain

GuHCl: guanidine chloride

IPTG: isopropyl β-D-thiogalactoside

m value: slope of ΔG versus denaturant concentration

NMR: nuclear magnetic resonance

PDB: protein data bank

RMSD: root mean square distance

SASA: solvent accessible surface area

SDS-PAGE: sodium dodecyl sulfate-polyacrylamide gel electrophoresis

ss: secondary structure

T_m: melting temperature

WT: wild type

CHAPTER 1

INTRODUCTION

EARLY DEVELOPMENT OF COMPUTATIONAL PROTEIN DESIGN

Computational protein design has become a very powerful approach for testing our understanding of the forces and energetics of macromolecular systems[1,2]. Designed proteins with desired structures or functions have been of special interest to pharmaceutical companies and this computational protein design method will make a significant impact on the development of new biotechnological therapeutics[3,4]. Thirty years ago, computational protein design may have sounded like science fiction but recently there has been great progress in the development of protein design methodologies and applications[5-9].

The earliest attempts at computational protein design focused on redesigning naturally occurring proteins while assuming a fixed, native backbone. One remarkable example was the complete redesign of a zinc finger by Dahiyat *et al.* with the backbone fixed[5]. The fixed backbone assumption greatly reduces computational time and works well when an appropriate backbone scaffold exists; however, it is incompatible with *de novo* design because there is no design template available. Studies have shown that incorporating flexibility can improve sequence prediction and is therefore better for novel design[10]. Harbury *et al.*[6] demonstrated the advantage of backbone flexibility by designing novel right-handed coiled-coil bundles. They created a set of right-handed coiled-coils and experimentally validated the structures. Another breakthrough in protein design was the creation of a fold that is not seen in nature so far. Kuhlman *et al*. used the Rosetta program to iteratively optimize both sequence and structure and created a protein with a novel fold (TOP7)[11]. The experimental results showed that TOP7 is very stable and the crystal structure matched the design model very well (root mean square deviation RMSD=1.2 Å). This striking result opened a new window to the exploration of novel proteins.

Protein design technology has many applications: so far it has been used to stabilize proteins, increase

protein-protein binding affinity, alter binding specificity and redesign a folding pathway[12-14]. Grand challenges in the form of *de novo* design problems, such as creating novel enzymes and biosensors, have been addressed[11,13,15-19]. Despite many successes, *de novo* design remains a very challenging problem, because it requires the design of a novel sequence that is unrelated to any naturally occurring protein that can fold into a pre-defined 3-dimensional structure[20]. In order to fold into a well-defined structure, the designed sequence should energetically stabilize the desired fold as well as destabilize the alternative conformations. *De novo* design is a rigorous test of our understanding of protein folding energetics, but the underlying principles are not yet understood well enough to ensure the success of all *de novo* designs[21].

It is fair to say though that we understand relatively well how to make a predominantly helical protein because some successful *de novo* designs for α helix bundles and α/β mixed proteins have been reported[11,22-24]. However, the *de novo* design of purely β-sheet proteins has proven to be more complicated, as these designed sequences usually misfold or aggregate. One possible reason is that the relatively slow folding of β-sheet proteins involves long range interactions, and unlike α-helix, β-sheet formation is determined mostly by tertiary context instead of intrinsic secondary structure preferences[25]. Compared with α-helix proteins, β-sheet proteins are less modular and inherently more difficult to design due to the fundamental difference in the hydrogen bonding patterns of these two different secondary structures[26]. The backbone hydrogen bonds within α-helix will be satisfied by nearby residues within the same secondary structure element. In contrast, β-sheet proteins require a β-strand to interact with a neighboring strand, possibly distant in primary structure, to satisfy backbone hydrogen bonds[27]. Side chains will point alternately above and below the β-sheet to interact with the neighboring residues. The need of a β-sheet to form so many interactions underlies β-sheet proteins' great tendency to aggregate. Additionally, residues with good β-sheet-forming propensities are also intrinsically prone to aggregation[28,29]. Nevertheless, the field of β-sheet design

is of great interest.

PREVIOUS WORK IN *DE NOVO* β-SHEET PROTEIN DESIGN

Recent work includes several groups' successful designs of some small water-soluble peptides, for instance β-hairpins[30], a 20 residue three-stranded antiparallel β-sheet[31] and even a four-stranded β-sheet[32]. Since 1981, the Richardson group has been trying to design several generations of betabellins but the solubility seems a big issue in these designs[33]. Yan *et al.* designed a series of betabellins [34,35] but the best one (betabellin 14D) only folded in the presence of a fold-stabilizing interchain disulfide. It consists of two 32-residue β-sheet packed against each other by a disulfide linkage. The sequence of each half has a pattern of alternating polar/nonpolar for β strands and statistically favored residues for β turns as shown in **Figure 1.1**. D-amino acid residues were used for the turn positions to favor formation of β turns. The single chain of betabellin 14S is not folded. To make it more globular, a disulfide bond was introduced to link the two identical subunit. The double-chain form 14D, is folded into a β-sheet liked structure in the presence of the disulfide linkage suggesting that the folding is induced by the disulfide bond formation. This disulfide bond strategy was also applied in the design of betadoublet[23](**Figure 1.2**), which is water soluble only at low pH; however NMR data suggests that betadoublet does not adopt a single unique conformation, implying that it adopts a molten globular structure.

Sollazzo *et al*. designed a small all β-sheet protein that can bind metal zinc upon folding, but again solubility limited the detailed structural analysis[36]. Recently, Nanda *et al.* designed a mimic of the redox protein rubredoxin, which seems to adopt the target fold[37]. This is one example of a functional *de novo* designed β-sheet protein. To date, attempts at the *de novo* design of β-sheet proteins are very impressive and encouraging, however, the success is still very limited and no design has been

validated with a high-resolution structure. *De novo* design of globular β-sheet proteins still remains an unsolved problem.

In nature, approximately one quarter of all protein domains are β-sheet folds[38]. β-sheet proteins form relatively rigid structures that can serve as good scaffolds for designing molecules with new functions. β-sheet proteins have also been a focus of considerable attention for medical biologists and the pharmaceutical industry[3,39] because they have proven to be good targets for disrupting unwanted protein-protein interactions. Protein-protein interactions are essential to many biological processes; however, uncontrolled interactions may lead to protein misfolding and aggregation which contribute to many different diseases including Alzheimer's disease, Huntington's disease, others[40]. Protein misfolding in these diseases involves protein aggregation into β-sheet rich oligomeric structures. Considerable evidence has shown that these aggregate structures play important roles in disease pathogenesis[41]. One strategy to develop therapies for these diseases is to address protein misfolding and aggregation with rationally designed inhibitors[4,42]. The computational protein design approach provides a valuable way for us to better understand how nature "designs" β-sheet proteins that can avoid misfolding or aggregation, which will be highly useful in future protein therapeutics.

MOLECULAR MODELING PROGRAM - ROSETTA

Rosetta is a molecular modeling software package developed by several research groups. Initially used for *de novo* structure prediction, it has since been expanded to contain protocols for high-resolution structure refinement, loop modeling, molecular docking and protein design[43-46]. The goal of protein design, also known as inverse folding, is to identify a compatible low free energy sequence for an existing protein backbone[12,47]. To solve this problem, we need two things: an energy function

for describing the interactions in proteins and ranking the fitness of a particular sequence for a given backbone structure, and a search algorithm for sampling sequence space[11].

Energy function

Being able to describe the interactions in proteins accurately is the most difficult problem in protein design. The energy function used in Rosetta contains physical potentials and knowledge based potentials derived statistically from the many structures in the Protein Data Bank(PDB)[12]. The potentials are combined in the Rosetta energy function as a linear sum of the following main terms[11,47]:

$$E_{total} = w_{atr}E_{ljatr} + w_{rep}E_{ljrep} + w_{sol}E_{sol} + w_{hbond}E_{hbond} + w_{pair}E_{pair} + w_{rot}E_{rot} + w_{rama}E_{rama} - E_{ref}$$

The main components in the energy function are:

Lennard-Jones potential: A 12-6 Lennard-Jones potential represents van der Waals interactions. This potential is slightly modified from the standard form with the introduction of a distance cutoff below which the potential is extrapolated linearly. The attractive and repulsive energies are split into separate terms, E_{ljatr} and E_{ljrep}, which gives greater flexibility in weighting the terms and improves sequence recovery. To compensate for the fixed backbone assumption, usually the repulsive term is softened to implicitly allow for some level of backbone flexibility.

Solvation energy (E_{sol}): The implicit solvation model developed by Lazaridis and Karplus is used to evaluate the solvation energy for a protein. This is a semiempirical model that is parameterized with experimental data and does not require surface area calculations[48]. This term penalizes surface exposure of hydrophobic residues and favors exposure of hydrophilic residues.

Hydrogen bonding potential (E_{hbond}): Hydrogen bonding is very important to stability and protein-protein interactions. Rosetta uses an orientation-dependent hydrogen bonding term, which is derived from the distribution of three parameters[14] (distance between the hydrogen and acceptor atoms, angle at the hydrogen atom and angle at the acceptor atom) from PDB database. This term allows buried polar atoms if they can form hydrogen bond. Together with the solvation energy term, these two terms balance how many polar residues are placed in the core during a design simulation.

Residue pair potential (E_{pair}): Electrostatic interactions such as salt bridges are very important for protein function; however, these interactions are very dependent on the local environment, which makes it very difficult to model. Rosetta uses a knowledge-based term to model electrostatics. This term is derived from the probability of a pair of polar residues being seen near each other in the PDB database[49].

Rotamer self-energy (E_{rot}): Internal energy of a rotamer is calculated based on the probability of seeing a particular rotamer for a given phi and psi angle in the PDB database. These probabilities are taken from Dunbrack library directly[50] and their negative log values were used as the rotamer internal energy as shown in the following equation

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

Torsion potential (E_{rama}): Rosetta uses ideal bond lengths and bond angles for bonded interactions. The torsion potential is associated with backbone bond torsion angles and is related to Ramachandran torsion preferences. It is derived from PDB statistics by measuring the probabilities of seeing a particular amino acid in a secondary structure type (helix, strand and loop) for a particular phi, psi angle[11].

$$E_{rama} = \sum_{i}^{residue} -\ln[prob(phi(i), psi(i) \mid aa_i, ss_i)]$$

aa = amino acid type

ss = secondary structure type

Reference energy (E_{ref}): Calculation of folded state stability requires a reference to the energy of an unfolded state. Rosetta uses parameterized energies for each residue to represent the free energy of unfolded state. The reference values and the weights for each energy term are calculated to best reproduce native sequences for known structures[11].

Search function

One major challenge in protein design is determining how to scan through sequence space and identify the optimum effectively. The size of sequence space is astronomical: for a 50-residue protein, in which all 20 standard amino acids are allowed at each position, 20^{50} (10^{65}) sequences are possible. To make the search computationally feasible, one simplification is to make the search space discrete. Currently, most computational protein design methods use a discrete set of side chain conformations (rotamers).

Computational protein design requires an efficient search algorithm that is able to scan an enormous search space[51]. The choice of algorithms will influence the accuracy of side chain predictions and the speed of design simulations. There are two categories of search algorithms, stochastic searches and deterministic search algorithms[51]. Deterministic algorithms include self consistent mean field optimization and dead end elimination; they are semiexhaustive search as which will always converge to the same solution (if they are able to converge). Stochastic algorithms include Monte Carlo simulated annealing and genetic algorithms. These methods are based on a random search, the

advantage of these methods is that they can handle complicated problems because they do not require an exhaustive search, the disadvantage is that they are not guaranteed to find the global energy minimum.

Rosetta uses a Monte Carlo search algorithm with simulated annealing to identify low energy sequences for a given structure. This is a simple, fast and widely used stochastic search method. In the Rosetta search algorithm, the initial conformation is generated randomly. This conformation is then perturbed by a single rotamer substitution(sequence could be changed). The substitution may or may not change the sequence identity (**Figure 1.3**). If the substitution lowers the energy ($E_{new} - E_{old} = \Delta E < 0$), it is accepted. Otherwise, the substitution is accepted if $e^{-\Delta E / kT} > R(0 \le R \le 1)$, where k is the Boltzman constant, T is the temperature and R is a random probability. This condition, called the Metropolis criterion, prevents the simulation from getting trapped in local energy minima. A trajectory may consist of a few hundred thousand rotamer substitutions, which is typically for convergence between trajectories.

TARGET FOLD MODEL SYSTEM

Our strategy for *de novo* design is to use a natural protein fold and design a new sequence that is not related to any natural protein sequence, but that will fold into the desired structure. To simplify the design process, we want the template to be just large enough to present true tertiary structure without requiring disulfide bonds or metal binding sites. One common β-sheet tertiary structure is the Fibronectin type III domain (FNIII). This domain occurs in many proteins with very diverse sequences, including cell surface receptors and cell adhesion molecules, which indicates that it is highly designable and could easily be modified to generate new functions. This small domain (about 90 residues) contains seven β-strands arranged into two sheets connected by flexible loops. These

exposed loops may be modified to generate novel functions for molecular recognition. The surface topology is very similar to the immunoglobulin VH structure[52] (**Figure 1.4**, left panel, pdbcode 1ol0), and this FNIII domain has become one popular scaffold for "monobody" design to date[53,54]. Monobodies are antibody-like proteins that bind to specific target proteins. Unlike antibodies, monobodies are usually easy to express and purify in large quantities and are ideal for inhibiting protein-protein interactions. Besides, they are small (~10 kDa), monomeric and lack disulfide bonds so that they are stable in reducing environments. Because of all these excellent characteristics, Huang *et al.* used monobodies to generate the affinity resin that binds to a specific conformation of the target protein so as to purify the desired conformation based on the target protein surface properties[55]. The template we used in this study is the third Fibronectin type III domain from tenascin[56] (pdbcode : 1ten, **Figure 1.4**, right panel). It is small, cysteine-free and monomeric. It is easy to purify and well characterized[57]. Tenascin has been shown to undergo a two-state, thermally reversible unfolding transition. These properties make it an ideal model system for our study.

De novo designed β-sheet proteins often aggregate in solution. One possible reason is that the designed sequence favors no folded structure or equally favors many folded structures. Another reason is that kinetically β-sheet proteins fold slowly so that they easily form aggregated. To design a well-folded structure, is it enough to only search for a sequence that has low free energy for the target structure (positive design)? Should we also include some elements that can destabilize the alternative fold states explicitly in the sequence design process (negative design)[58]? In this thesis, we address the importance of positive design and negative design by pursing different design problems.

DESIGN APPROACH

Experience shows that *de novo* design of all β-sheet proteins is an extremely challenging problem; hence, we decided to break up the process into small steps.

The Rosetta energy function is a linear combination of Lennard-Jones interactions, implicit solvation potential, hydrogen bonding energy and additional knowledge-based energy terms. However, the program does not explicitly factor in side chain entropy, which is also very important to thermodynamics. Residues with more degrees of freedom (Lys, Arg, Met, etc) lose more conformational entropy upon folding and these amino acids are less likely to be buried[59]. The correct placement of a given amino acid is likely to partially depend on how much entropy is lost when the side chain is locally constrained in a folded protein. However, the influence of the side chain entropy on protein design simulations in Rosetta was not clear. In order to investigate how side chain entropy influences protein design simulations, in chapter 2 we will explicitly incorporate side chain entropy into Rosetta and test if it improves recovery of native sequence in design simulations.

To decipher the importance of positive design and negative design, we will pursue a set of different design problems. In chapter 3, we will try to redesign a naturally occurring all β-sheet protein (tenascin) with only positive design. The use of tenascin guarantees that the backbone is designable. This test will determine whether we can use Rosetta to design an all β-sheet protein that is well-folded and stable using only positive design. In this test case, the backbone is fixed, which will bias sequence selection. In chapter 4, we will incorporate backbone flexibility into the design simulation by redesigning a 10 residue loop in tenascin. Being able to explore the backbone degrees of freedom will increase conformational sampling and design complexity. By only designing part of the backbone, we will be able to test if we can design a well-structured loop conformation in the context of a stably folded β-sheet protein.

In chapter 5, we will try to design a whole protein from scratch. From multiple generations of design and experimental characterization of the results, we will have feedback that may be used to improve the design algorithm. We will learn about how proteins fold from both failed attempts and successes.

De novo design of β-sheet protein is a rigorous test of our protein design software and our understanding of the relationship between sequences and structures. By designing a sequence *de novo* we would expect to learn new things that we would not have learned by examining or redesigning naturally occurring proteins. The results of these experiments will give us feedback on how we should improve the design program.

FIGURES

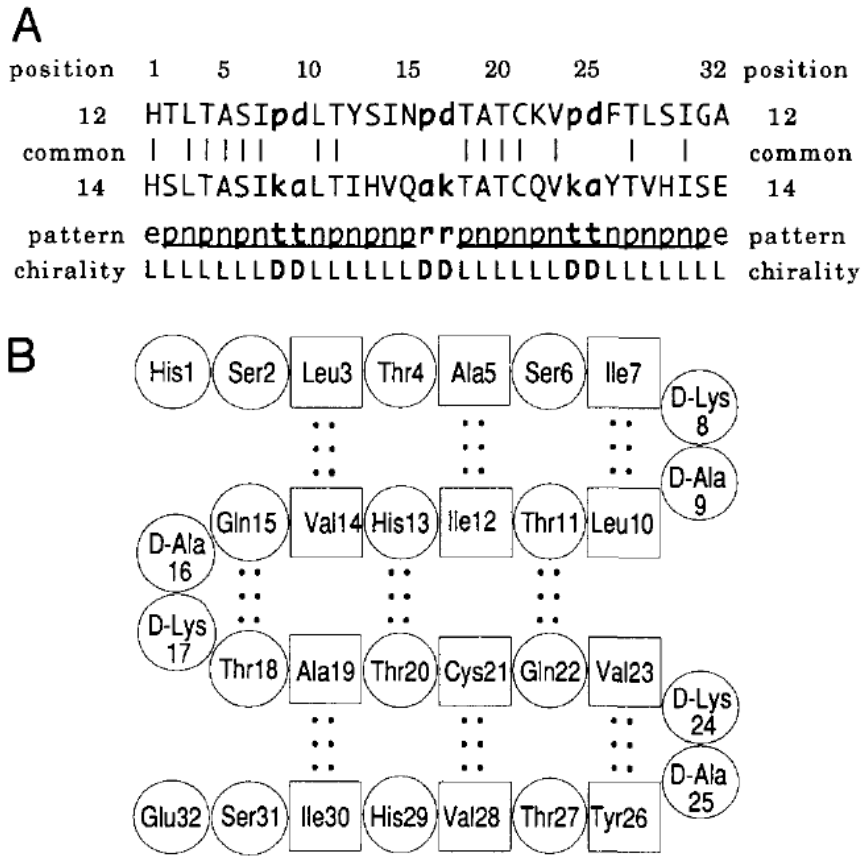


Figure 1.1 Betabellin 14S
Amino acid sequence assignment of the betabellin 14 single chain. A: Pattern of polar(p), nonpolar(n) and turn(t) residues. B: Betabellin target structure[34]

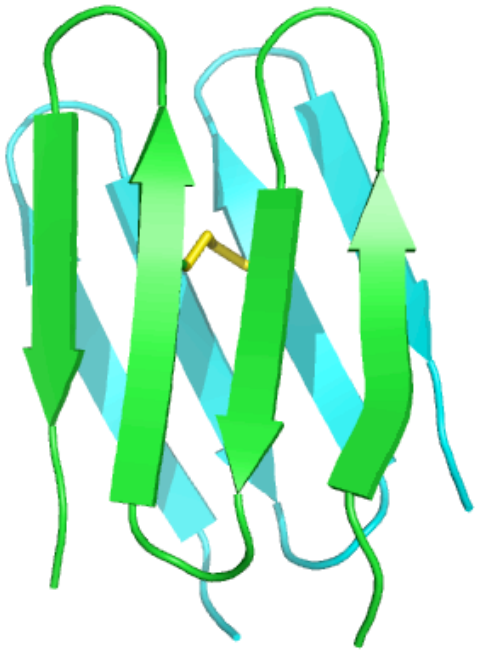


Figure 1.2 Betadoublet
Cartoon diagram of the model of betadoublet(pdb code 1btd). Green and cyan are for the two identical subunit, the disulfide bond linkage is colored by yellow.

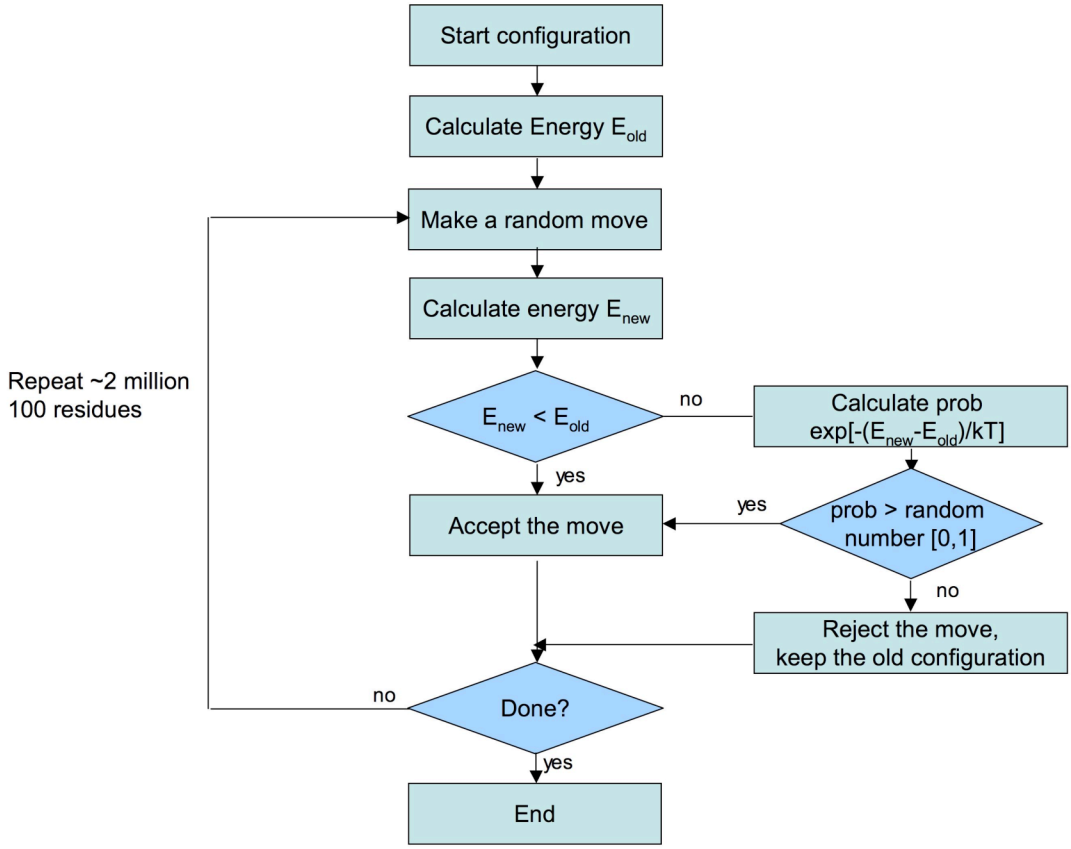


Figure 1.3 The Metropolis sampling algorithm used in Rosetta

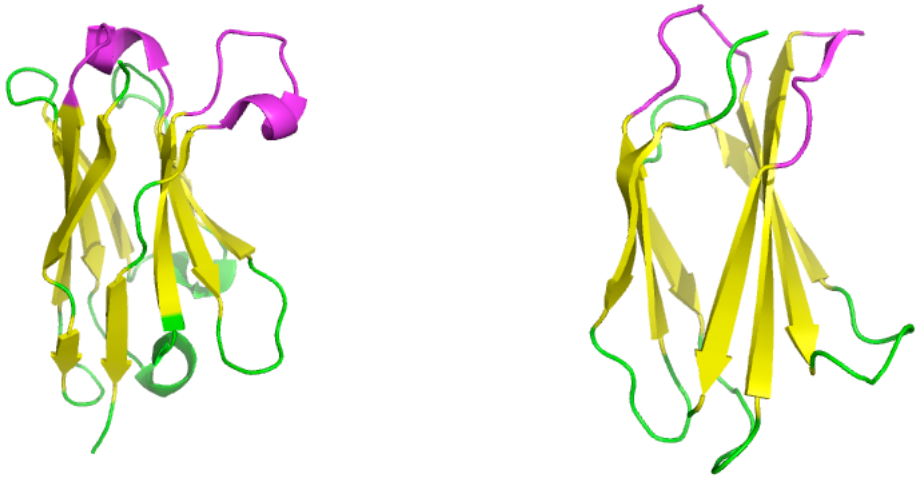


Figure 1.4 Comparison of a monomeric immunoglobulin VH domain(left) with FN3(right). The binding loops are colored in magenta on the top.

REFERENCES

1. Pokala N, Handel TM. Review: protein design--where we were, where we are, where we're going. J Struct Biol 2001;134(2-3):269-281.

2. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A. De novo design and structural characterization of proteins and metalloproteins. Annu Rev Biochem 1999;68:779-819.

3. Rosenberg M, Goldblum A. Computational protein design: a novel path to future protein drugs. Curr Pharm Des 2006;12(31):3973-3997.

4. Estrada LD, Soto C. Inhibition of protein misfolding and aggregation by small rationally-designed peptides. Curr Pharm Des 2006;12(20):2557-2567.

5. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278(5335):82-87.

6. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;282(5393):1462-1467.

7. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66-93.

8. Rothlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D. Kemp elimination catalysts by computational enzyme design. Nature 2008;453(7192):190-195.

9. Lippow SM, Tidor B. Progress in computational protein design. Curr Opin Biotechnol 2007;18(4):305-311.

10. Desjarlais JR, Handel TM. Side-chain and backbone flexibility in protein core design. J Mol Biol 1999;290(1):305-318.

11. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302(5649):1364-1368.

12. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J Mol Biol 2003;332(2):449-460.

13. Sammond DW, Eletr ZM, Purbeck C, Kimple RJ, Siderovski DP, Kuhlman B. Structure-based protocol for identifying mutations that enhance protein-protein binding affinities. J Mol Biol 2007;371(5):1392-1404.

14. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 2003;326(4):1239-1259.

15. Hellinga HW, Marvin JS. Protein engineering and the development of generic biosensors. Trends Biotechnol 1998;16(4):183-189.

16. Joachimiak LA, Kortemme T, Stoddard BL, Baker D. Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. J Mol Biol 2006;361(1):195-208.

17. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, 3rd, Hilvert D, Houk KN, Stoddard BL, Baker D. De novo computational design of retro-aldol enzymes. Science 2008;319(5868):1387-1391.

18. Nauli S, Kuhlman B, Baker D. Computer-based redesign of a protein folding pathway. Nat Struct Biol 2001;8(7):602-605.

19. Lippow SM, Wittrup KD, Tidor B. Computational design of antibody-affinity improvement beyond in vivo maturation. Nat Biotechnol 2007;25(10):1171-1176.

20. Butterfoss GL, Kuhlman B. Computer-based design of novel protein structures. Annu Rev Biophys Biomol Struct 2006;35:49-65.

21. Baltzer L, Nilsson H, Nilsson J. De novo design of proteins--what are the rules? Chem Rev 2001;101(10):3153-3163.

22. Regan L, DeGrado WF. Characterization of a helical protein designed from first principles. Science 1988;241(4868):976-978.

23. Quinn TP, Tweedy NB, Williams RW, Richardson JS, Richardson DC. Betadoublet: de novo design, synthesis, and characterization of a beta-sandwich protein. Proc Natl Acad Sci U S A 1994;91(19):8747-8751.

24. Hecht MH, Das A, Go A, Bradley LH, Wei Y. De novo proteins from designed combinatorial libraries. Protein Sci 2004;13(7):1711-1723.

25. Minor DL, Jr., Kim PS. Context is a major determinant of beta-sheet propensity. Nature 1994;371(6494):264-267.

26. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A 1951;37(4):205-211.

27. Pauling L, Corey RB. The pleated sheet, a new layer configuration of polypeptide chains. Proc Natl Acad Sci U S A 1951;37(5):251-256.

28. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 1974;13(2):211-222.

29. Koehl P, Levitt M. Structure-based conformational preferences of amino acids. Proc Natl Acad Sci U S A 1999;96(22):12524-12529.

30. Blanco F, Ramirez-Alvarado M, Serrano L. Formation and stability of beta-hairpin structures in polypeptides. Curr Opin Struct Biol 1998;8(1):107-111.

31. Kortemme T, Ramirez-Alvarado M, Serrano L. Design of a 20-amino acid, three-stranded beta-sheet protein. Science 1998;281(5374):253-256.

32. Das C, Nayak V, Raghothama S, Balaram P. Synthetic protein design: construction of a four-stranded beta-sheet structure and evaluation of its integrity in methanol-water systems. J Pept Res 2000;56(5):307-317.

33. Richardson JS, Richardson DC. The de novo design of protein structures. Trends Biochem Sci 1989;14(7):304-309.

34. Yan Y, Erickson BW. Engineering of betabellin 14D: disulfide-induced folding of a beta-sheet protein. Protein Sci 1994;3(7):1069-1073.

35. Lim A, Makhov AM, Bond J, Inouye H, Connors LH, Griffith JD, Erickson BW, Kirschner DA, Costello CE. Betabellins 15D and 16D, de Novo designed beta-sandwich proteins that have amyloidogenic properties. J Struct Biol 2000;130(2-3):363-370.

36. Pessi A, Bianchi E, Crameri A, Venturini S, Tramontano A, Sollazzo M. A designed metal-binding protein with a novel fold. Nature 1993;362(6418):367-369.

37. Nanda V, Rosenblatt MM, Osyczka A, Kono H, Getahun Z, Dutton PL, Saven JG, Degrado WF. De novo design of a redox-active minimal rubredoxin mimic. J Am Chem Soc 2005;127(16):5804-5805.

38. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure 1997;5(8):1093-1108.

39. Mason JM, Kokkoni N, Stott K, Doig AJ. Design strategies for anti-amyloid agents. Curr Opin Struct Biol 2003;13(4):526-532.

40. Frid P, Anisimov SV, Popovic N. Congo red and protein aggregation in neurodegenerative diseases. Brain Res Rev 2007;53(1):135-160.

41. Kammerer RA, Kostrewa D, Zurdo J, Detken A, Garcia-Echeverria C, Green JD, Muller SA, Meier BH, Winkler FK, Dobson CM, Steinmetz MO. Exploring amyloid formation by a de novo design. Proc Natl Acad Sci U S A 2004;101(13):4435-4440.

42. Kim W, Kim Y, Min J, Kim DJ, Chang YT, Hecht MH. A high-throughput screen for compounds that inhibit aggregation of the Alzheimer's peptide. ACS Chem Biol 2006;1(7):461-469.

43. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature 2007;450(7167):259-264.

44. Hu X, Wang H, Ke H, Kuhlman B. High-resolution design of a protein loop. Proc Natl Acad Sci U S A 2007;104(45):17668-17673.

45. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins 2004;55(3):656-677.

46. Gray JJ, Moughon SE, Kortemme T, Schueler-Furman O, Misura KM, Morozov AV, Baker D. Protein-protein docking predictions for the CAPRI experiment. Proteins 2003;52(1):118-122.

47. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 2000;97(19):10383-10388.

48. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35(2):133-152.

49. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34(1):82-95.

50. Dunbrack RL, Jr., Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6(8):1661-1681.

51. Voigt CA, Gordon DB, Mayo SL. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. J Mol Biol 2000;299(3):789-803.

52. Dottorini T, Vaughan CK, Walsh MA, LoSurdo P, Sollazzo M. Crystal structure of a human VH: requirements for maintaining a monomeric fragment. Biochemistry 2004;43(3):622-628.

53. Koide A, Bailey CW, Huang X, Koide S. The fibronectin type III domain as a scaffold for novel binding proteins. J Mol Biol 1998;284(4):1141-1151.

54. Olson CA, Roberts RW. Design, expression, and stability of a diverse protein library based on the human fibronectin type III domain. Protein Sci 2007;16(3):476-484.

55. Huang J, Koide,A.,Nettle,K.,Greene,G.,Koide,S. Conformation-specific affinity purification of proteins using engineered binding proteins: Application to the estrogen receptor. Protein Expression and Purification 2006(47):348-354.

56. Leahy DJ, Hendrickson WA, Aukhil I, Erickson HP. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. Science 1992;258(5084):987-991.

57. Hamill SJ, Cota E, Chothia C, Clarke J. Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. J Mol Biol 2000;295(3):641-649.

58. Richardson JS, Richardson DC. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci U S A 2002;99(5):2754-2759.

59. Doig AJ. Thermodynamics of amino acid side-chain internal rotations. Biophys Chem 1996;61(2-3):131-141.

CHAPTER 2

PROTEIN DESIGN SIMULATIONS SUGGEST THAT SIDE CHAIN CONFORMATIONAL ENTROPY IS NOT A STRONG DETERMINANT OF AMINO ACID ENVIRONMENTAL PREFERENCES

Xiaozhen Hu and Brian Kuhlman*

Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC, 27599

Keywords: Computational Protein Design, Side Chain Entropy, Protein Stability

*corresponding author

This work was published in Proteins: Structure, Function and Bioinformatics(2006)Mar 15;62(3):739-48.

ABSTRACT

Loss of side chain conformational entropy is an important force opposing protein folding and the relative preferences of the amino acids for being buried or solvent exposed may be partially determined by which amino acids lose more side chain entropy when placed in the core of a protein. To investigate these preferences we have incorporated explicit modeling of side chain entropy into the protein design algorithm, Rosetta. In the standard version of the program the energy of a particular sequence for a fixed backbone depends only on the lowest energy side chain conformations that can be identified for that sequence. In the new model, the free energy of a single amino acid sequence is calculated by evaluating the average energy and entropy of an ensemble of structures generated by Monte Carlo sampling of amino acid side chain conformations. To evaluate the impact of including explicit side chain entropy, sequences were designed for 110 native protein backbones with and without the entropy model. In general, the differences between the two sets of sequences are modest, with the largest changes being observed for the longer amino acids: methionine and arginine. Overall, the identity between the designed sequences and the native sequences does not increase with the addition of entropy, unlike what is observed when other key terms are added to the model (hydrogen bonding, Lennard-Jones energies and solvation energies). These results suggest that side chain conformational entropy plays a relatively small role in determining the preferred amino acid at each residue position in a protein.

INTRODUCTION

Protein folding is a competition between the formation of favorable contacts, the loss of conformational entropy, and added strain. In addition to adopting a relatively fixed backbone structure, many of the side chains in a folded protein only sample a subset of the rotamers accessible to them in the unfolded state. A variety of independent methods have been used to estimate the average change in conformational side chain entropy upon folding [1-17]. The consensus from these studies is that approximately 0.5 kcal·mol^{-1} of side chain entropy is lost per dihedral angle fixed in the folded structure. It has also been proposed that the probability of an amino acid being placed in a buried position in a protein is proportional to how much entropy will be lost when the side chain is fixed in a single conformation. Amino acids with greater degrees of freedom (Lys, Arg and Met) are expected to be disfavored in buried positions[18,19]. Because the polar residues are on average intrinsically more flexible than the non-polar amino acids, it is not straightforward to determine the relative role of solvation and entropic effects in the environmental preferences of the amino acids. Protein design simulations provide one approach for deciphering the relative importance of these two effects.

Recently, there has been considerable success in the area of computational protein design as a variety of computer programs have been developed for identifying low energy sequences for target protein structures[20-25]. These models generally consist of two primary components: an energy function for evaluating the favorability of a particular sequence and a search protocol for scanning through sequence space. The models are frequently tested by redesigning naturally occurring proteins and comparing the redesigned sequences to the native sequences. Often the redesigned sequences are noticeably similar to the native sequences, and the usefulness of a specific term in the energy function can be determined by repeating the comparison without the energy term. In one study, Koehl and Levitt used protein design simulations to show that the intrinsic propensity of the amino acids for the

various types of secondary structure is a natural consequence of Lennard-Jones interactions and hydrophobic burial, thus indicating that a separate term did not need to be added to capture these preferences.[5,26-30]

Several approaches have been used to incorporate side chain entropy in protein design simulations. One common method is to assume that all residues in the protein are fixed in a single side chain conformation, and therefore the change in side chain entropy upon folding only depends on amino acid composition and the average side chain entropy of the various amino acids in the unfolded state[31-33]. This method is attractive because it is compatible with rotamer optimization protocols such as Dead End Elimination that require pair wise additive energy functions. However, this approach does not differentiate between buried and surface residues, and the practical effect is to only perturb the amino acid composition of the designed sequences. Several laboratories have developed a mean-field approach in which each residue is simultaneously populated by all possible rotamers at probabilities related to their energy with neighboring rotamers[4,34-36]. From this protocol it is straightforward to calculate side chain entropy at each position and include this in the free energy of the protein. One limitation of the mean-field method is that it is not entirely physical; it is not possible for a single residue to simultaneously occupy two conformations.

Farid and co-workers used a two layer Monte Carlo optimization protocol to incorporate explicit side chain entropy in protein design calculations[37]. The inner layer used Monte Carlo sampling of side chain conformational space to calculate the average energy and entropy of fixed sequences, while the outer layer was used to scan through sequence space. Here, we will use a similar approach in large-scale protein design simulations to determine if side chain entropy plays a significant role in determining the environmental preferences of the amino acids. A similar comparison has not been made previously, and these results will indicate if it is useful to include explicit side chain entropy calculations in protein design simulations.

METHODS

Rosetta. The Rosetta algorithm has been described previously [38,39]. It uses a Monte Carlo search procedure with simulated annealing to identify low energy amino sequences and side chain conformations for target protein structures. The side chains are modeled using Dunbrack's backbone dependent rotamer library [40,41]. Starting from a random sequence single amino acid substitutions or rotamer changes are accepted based on the Metropolis criterion. The energy function is a linear combination of the following terms: a 12-6 Lennard-Jones potential, the Lazaridis-Karplus implicit solvation model [42], an explicit orientational dependent hydrogen bonding term [43], the relative free energy of the various rotamers (as modeled by -lnP(rot|aa,phi,psi)) and a statistically based pair term that gives a weak bonus for putting unlike charges near each other [44].

In addition, each amino acid is assigned a reference energy that controls how often a particular amino acid is placed during a design simulation and represents to some degree the free energy of that amino acid in the unfolded state. The reference values and weights on each of the energy terms are parameterized to best reproduce native sequences. The weights used in this study are the same as those used previously to design a novel protein structure [39], with the exception that repulsive portion of the Lennard-Jones potential was dampened to account for the use of fixed backbones. It is important to note that the amino acid reference energies implicitly account for the various amounts of side chain entropy that each amino acid has in the unfolded state. This will control how often a particular amino acid is observed in the designed sequences, but it will not play a significant role in determining the environmental preferences of the amino acids.

It should also be noted that in addition to the reference energies, other terms in the Rosetta energy function incorporate some entropic effects, but none of these terms are expected to significantly

overlap with a side chain entropy term. The Lazaridis-Karplus solvation model is designed to model desolvation energies, and therefore implicitly accounts for the change in water entropy associated with the hydrophobic effect. The term representing side chain torsion energies (-lnP(rot|aa,phi,psi)) relates to the relative free energy of each rotamer and therefore may depend in part on vibrational entropy within each rotamer. It should not incorporate, however, the conformational entropy that we are modeling in this study that results from switching between rotamers. The pair term is based on the probability that two amino acids will be found near each other and is related to a free energy. It is difficult to determine if this term includes any effects from side chain entropy, if it does, there will be some double counting with our new explicit term for side chain entropy. To insure that this term is not skewing our results we have repeated the native sequence recovery tests without the pair term. The effect of including explicit side chain entropy in the Rosetta model is nearly identical with and without the use of the pair term (data not shown).

Incorporating explicit side chain entropy into Rosetta. **Figure 2.1** outlines our approach for incorporating side chain entropy into Rosetta. It is a two layer approach; the inner layer is used for calculating the free energy of a fixed sequence on a fixed backbone, while the outer layer is used to scan through sequence space. The inner layer uses Monte Carlo sampling to generate an ensemble of structures with a variety of side chain conformations. Each round of this procedure involves:

1) switching a single residue (chosen at random) to a new Dunbrack rotamer
2) evaluating the new energy of the protein
3) accepting the perturbation if it passes the Metropolis criterion.

At the completion of each round the energy of the protein is added to a running sum that is later divided by the number of rounds to determine the average energy of the ensemble (U in equation 1). After each round, it is also determined which rotamer is present at each sequence position and these are added to running sums that are later used to calculate the probability of observing a rotamer at

each sequence position (p in eq. 2)[2]. These probabilities are used to calculate the side chain conformational entropy (S) and Helmholtz free energy (A) of the system:

$$A = U - TS \qquad (1)$$

$$S = -R\sum_{i=1}^{nres}\sum_{r=1}^{nrot} p(r,i)\ln(p(r,i)) \qquad (2)$$

where *nres* is the number of residues in the protein, *nrot* is the number of possible rotamers at each sequence position, and T is the temperature. T was set to a physiologically relevant temperature (310 K) for these calculations. Average energies and rotamer probabilities were calculated after the system was equilibrated for

(3 * nrotamers) rounds, where nrotamers is equal to the number of rotamers being considered at each sequence position times the number of residues being redesigned. Similar results were obtained if the system was equilibrated for 5 * nrotamers rounds.

The outer layer uses Monte Carlo sampling to scan through sequence space(**Figure 2.1**). The procedure begins with a random sequence. Each round of optimization involves:

1) making a random single amino acid mutation
2) evaluating the free energy of the new sequence with a repacking simulation (equation 2)
3) accepting the mutation if it passes the Metropolis criterion.

For the outer layer the temperature is set high at the beginning of the simulation and gradually cooled to 0 K. 100 * number of residues sequence substitutions are used per simulation. Because a complete repacking simulation is performed after each sequence change, the double layer protocol is considerably slower than a standard sequence optimization simulation with Rosetta (60 times slower). To increase computational speed the protocol was modified so that only residues within 10Å of the mutated residue were repacked. We found that the average energies and entropies for

residues further away from the mutation site did not change, and therefore it was possible to use values saved from the previous repacking simulation. The modified protocol was 1.5 times faster, and gave the same results as when all residues are repacked following each mutation.

The quality of our results will depend in part on how accurately we can pack amino acid side chains (the inner layer of our protocol). The accuracy of side chain packing algorithms is often evaluated by removing the side chains from naturally occurring proteins and rebuilding the side chains from scratch. The procedure is evaluated by determining the fraction of the amino acids that are placed in the correct side chain conformation. To insure that Rosetta performs satisfactorily on this test we rebuilt the side chains on 57 high resolution crystal structures. With the standard Dunbrack rotamer library, 80% of the buried positions had both their chi 1 and chi 2 angles predicted within 40 degrees of the angles observed in the crystal structure. These results are similar to what has been achieved with other side chain placement algorithms [12,45-48].

Calculating side chain conformational entropies through complete enumeration. One assumption of equation 2 is that the rotamer probabilities at the various sequence positions are independent of each other, or in other words, that there is no covariant motion between side chains. If there is covariant motion equation 2 will overestimate the side chain entropy of the system. To test the validity of this assumption we used complete enumeration of rotamer configurations for 6 residue clusters to calculate the energies of all possible rotamer combinations. These energies were used to generate a partition function for the cluster and calculate the relative probabilities of each possible packing combination state (equation 3). These probabilities were then used to calculate the entropy of the system (equation 4), and compared with results obtained by Monte Carlo sampling as described above.

$$p_i = \frac{e^{-E_i / K_B T}}{\sum_{i=1}^{nstates} e^{-E_i / K_B T}} \qquad (3)$$

$$S = -R \sum_{i=1}^{nstates} p_i \ln(p_i) \qquad (4)$$

K_B is the Boltzmann constant and T is the temperature (310K). 816 clusters in 110 proteins were used for this comparison. Side chains outside of the cluster were held fixed during the complete enumeration protocol and the Monte Carlo sampling protocol.

Calculating side chain conformational entropy from a 80 ns molecular dynamics simulation of eglin C. To further check if covariant side chain motion reduces the total entropy of a protein we examined a 80 ns molecular dynamics simulation of eglin C from a previous study.[49] Eglin C remains folded throughout this simulation in a conformation similar to the crystal structure. Previously, Lee and co-workers used this trajectory to calculate order parameters for the side chains, and there was a good agreement between the calculated values and order parameters measured with NMR. To calculate side chain entropy from the simulation dihedral angles for each side chain were extracted from every 0.36 picosecond and binned into rotamers based on Dunbrack's rotamer definitions[41]. Rotamer frequencies were used to calculate side chain entropy using two separate approaches. The first approach was to treat each site independently, i.e., use the rotamer probabilities from single residues to calculate entropy (equation 2), and the second was to use probabilities of rotamer pairs to calculate entropy (equation 5):

$$S = -\frac{R \sum_{i=1}^{nres-1} \sum_{j=i+1}^{nres} p_{(i,j)} \ln(p_{(i,j)})}{nres - 1} \qquad (5)$$

where *nres* is the number of residues in the protein, $p_{(i,j)}$ is the probability of seeing a rotamer pair (i,j) during the whole simulation. As a control, we also calculated entropy by treating each chi angle in the protein independently (equation 6).

$$S = -R \sum_{chi\_bin=1}^{nchi\_bin} p_{(chi\_bin)} \ln(p_{(chi\_bin)}) \qquad (6)$$

Each chi angle in the protein is divided into bins (*chi_bin*) based on Dunbrack's rotamer definitions, and the probability that a given side chain is in a specific bin ($p_{(chi\_bin)}$) is determined by averaging over the molecular dynamics simulation.

Native sequence recovery tests. Rosetta was used to design sequences for 110 proteins ranging in size from 50 to 150 residues (see **supplementary Table 2.6** for pdb codes), and the designed sequences were compared to the wild type sequences. Cysteines were held fixed during these simulations because the Rosetta energy function for disulfide formation still needs to be refined. Residues were defined as buried if they had greater than 18 neighbors (C_α atoms within 10Å), and surface if they had less than 13 neighbors.

RESULTS AND DISCUSSION

Covariant changes in side chain position do not significantly reduce the side chain conformational entropy of a protein. Before determining the effects of side chain entropy on sequence design, we first tested if the total side chain entropy of a protein could be accurately calculated by assuming the rotamer probabilities at each sequence position were independent of each other (equation 2). Because proteins are coupled systems, the preferred side chain conformation at one position may depend on the conformations of neighboring residues. One consequence of covariant motion is that the total

number of populated states, and hence entropy, will be lower than if each residue moved independently of its neighbors.

To test the importance of covariant motion between amino acid side chains we used two alternative methods for measuring side chain conformational entropy. For the first method we enumerated through all possible rotamer combinations for 6 residue clusters from naturally occurring proteins and calculated the energy of each state. Residues outside of the cluster were held fixed. The energies were then used to derive the probability of each state and the total entropy of the system (equations 3 and 4). For the second method we used Monte Carlo sampling of amino acid rotamers to create an ensemble of structures. Rotamer probabilities for individual residues were calculated from the resulting ensembles and total side chain entropy was calculated assuming the rotamer probabilities at each sequence position were independent of each other (equations 1 and 2). The same 6 residue clusters were used for this approach as were used for the complete enumeration.

As expected, we measure higher entropies when we assume that the residues behave independently, but in general the differences are very small (**Table 2.1**). This suggests that there is not significant covariant motion between side chains, and that equation 2 is suitable for calculating side chain entropy during design simulations. A similar conclusion was reached by Leach et al. when using the A* algorithm to explore rotamer packing proteins[50]. To further test this result we also examined side chain motion from an 80 ns molecular dynamics simulation of eglin C. Side chain conformations were assigned to bins corresponding to the Dunbrack rotamers, and rotamer frequencies were used to calculate side chain conformational entropy. Entropy was calculated with two approaches, in the first case the rotamer probabilities for individual residues were used (equation 2) while in the second case probabilities for rotamer pairs were tabulated (equation 5). The two results were similar, suggesting again that there is not significant covariant motion between amino acid side chains. In contrast, if we treat each torsion angle in a side chain independently, then we calculate a higher value for side chain

entropy (**Table 2.1**). This result indicates, as expected, that the torsion angles within an amino acid side chain do not behave independently.

Side chain entropy and energy as a function of burial. Introducing the side chain entropy and free energy model into our energy function should have two competing effects. Flexible side chains will be rewarded for being able to sample multiple conformations, but at the same time they will be penalized if those states are not iso-energetic and the average energy of the ensemble is greater than the energy of the most favorable conformation. To examine the relative strength of these two effects we performed two sets of repacking simulations on a large set of naturally occurring protein structures (2832 pdb files). In the first case we used the standard Rosetta model to identify the lowest energy side chain configuration for each protein and recorded the energy of each type of amino acid as a function of burial. In the second case we performed repacking simulations at 310 K and recorded entropies (equation 2) and average energy for each amino acid as a function of burial.

The longer amino acids have the highest average values of conformational entropy and show the greatest difference between surface and buried positions (**Figure 2.2**). For instance, the average side chain entropy (TS) for arginine is 1.60 kcal / mol on the surface of a protein and 0.77 kcal / mol in the core of a protein, while for a valine the same values are 0.18 and 0.09 kcal / mol (**Table 2.2**). However, the longer amino acids also show the greatest differences between average energy and the energy of the most favorable rotamer. On the surface of proteins the average energy of arginines when free to sample multiple conformations is on average 0.62 kcal / mol greater then the energy of the most favorable rotamer. The net result is that the penalty for burying an arginine in the core of a protein is not as great as would be suggested from just examining the side chain entropy term. In general, including explicit side chain flexibility in the scoring function does not appear to dramatically perturb the relative energies of the amino acids at buried and exposed positions. The difference is only greater than 0.3 kcal / mol for 4 amino acids: Met, Arg, Gln and Glu. The

difference is smaller for lysine (0.26 kcal / mol) because lysine has a strong intrinsic preference to be in the extended conformation as evidenced by the Dunbrack rotamer library and high level quantum mechanics calculations [51].

In these simulations we have been modeling the amino acid side chains using Dunbrack's backbone dependent rotamer library. This library does not allow for small perturbations of chi angles within a rotamer. To test whether increasing the rotamer library would significantly perturb our results we increased the rotamer library by allowing for rotamers that had their chi 1 angles perturbed +/- one standard deviation from the most preferred chi 1 angle. These perturbations are generally around 10 degrees and are based on the standard deviations in the Dunbrack library. As other groups have observed previously [2,50,52], with the expanded rotamer library the absolute side entropy of the amino acids goes up significantly but the difference between surface exposed and buried positions is largely unchanged (**Table 2.3**). This suggests that vibration of torsional angles within rotamers will not be a key determinant of whether an amino acid prefers to buried or exposed.

Protein Design Simulations with Explicit Side Chain Entropy. To test if side chain entropy plays a large role in determining the environmental preferences of the amino acids Rosetta was modified to include explicit side chain entropy and free energy calculations (see methods) and sequences were designed for 110 naturally occurring protein sequences. Although Rosetta uses a stochastic search procedure to search for low free energy sequences, independent simulations for a single protein produce very similar sequences (> 70% identity between simulations), indicating that the protocol does not get trapped in false minima located far from the global minima. To control the overall frequency that each amino acid is used during design, Rosetta, assigns a unique reference value to each amino acid. Because the purpose of these simulations were to determine the role of side chain entropy in the environmental preferences of the amino acids, we parameterized a unique set of

reference values for each set of simulations so that the amino acids were designed at native-like frequencies.

Overall, including side chain entropy as a new energy term did not have a large effect on the recovery of native sequences (**Table 2.4**). 32% of the residues in the design set were kept as the native amino acid in the simulations with and without the entropy term. As anticipated, the largest changes are observed for the more flexible amino acids (**Table 2.5**). With the explicit side chain entropy model 49% of methionines are designed in the core while with standard Rosetta 55% of methionines are placed in the core. Arginine shows the largest changes, 20% are placed in the core without explicit side chain entropy while only 12% are placed in the core with explicit side chain entropy. For most amino acids, the environmental preferences were not significantly perturbed by adding the explicit side chain entropy. Most likely this is because gains in favorable side chain entropy are generally accompanied by an increase in average energy (**Figure 2.2**), and the sum of these two effects is considerably smaller than the magnitude of other terms in the energy function.

Results with and without side chain entropy contrast sharply with simulations performed with and without the Lazaridis-Karplus solvation model. Without the solvation model there are dramatic changes in the environmental preferences of the amino acids, and the identity between the designed sequences and the wild type sequences falls to 19% (**Table 2.4**). So although the longer polar residues may be partially disfavored from buried positions because of a loss in conformational entropy, it is clear that desolvation energies are a much stronger factor than side chain entropy in determining the environmental preferences of the amino acids.

It is important to point out that the relative importance of side chain entropy does depend on the energy function that is being used, and larger effects may be observed with other potential energy

functions. In these cases, native sequence recovery tests will provide an excellent approach for determining the usefulness of the side chain entropy term for protein design.

CONCLUSION

We draw two main conclusions from these studies. First of all, the agreement between side chain entropies calculated by complete enumeration (equation 4) and those calculated by Monte Carlo sampling (equation 2) suggest that most side chains rotate independently of each other. To determine if this was primarily an artifact of using a fixed backbone, we also examined side chain motion in an 80-ns trajectory of eglin C and observed little covariant motion between amino acid rotamers. Similar results have also been reported for molecular dynamics simulations with calmodulin and the fibronectin family of proteins [53,54], although correlated motion has been observed for a pair of residues in ubiquitin [55].

Secondly, we observe that the incorporation of explicit side chain entropy and free energy calculations into Rosetta does not substantially increase our ability to recapitulate native sequences in protein design simulations. However, we did observe a reduction in the number of methionines and arginines placed in buried positions that was consistent with the environmental preferences of these amino acids in naturally occurring proteins. In general, our results suggest that side chain entropy plays a relatively small role in determining the environmental preferences of the amino acids.

ACKNOWLEDGMENTS

We would like to thank Dr. Jan Hermans and Dr. Andrew Lee for allowing us to analyze their molecular dynamics simulation of eglin C.

FIGURES

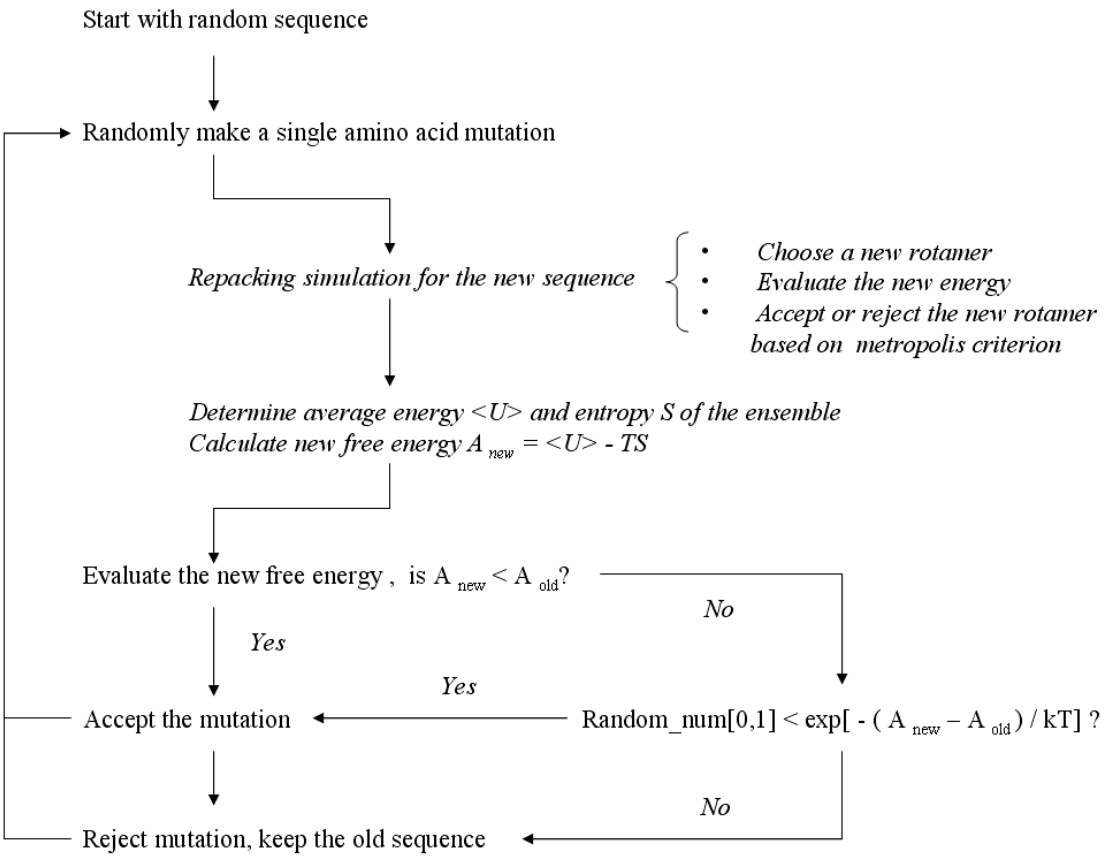


Figure 2.1 Algorithm for incorporating side chain entropy and free energy into Rosetta

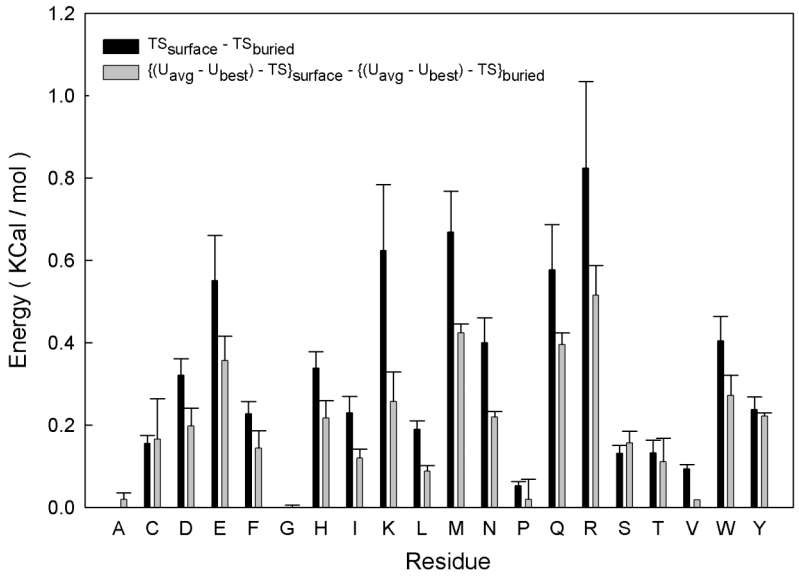


Figure 2.2 Changes in side-chain conformational entropy and free energy between surface and buried positions

The black bars show the change in entropy (TS) when a residue is buried while the grey bars compare average free energies ($U_{avg} - TS$) obtained with the explicit side chain entropy model to the energies that are calculated with the standard Rosetta model (U_{best}). The grey bars indicate the net effect that the explicit side chain entropy model has on the environmental preferences of the amino acids. Values are derived from repacking simulations with 2832 pdb files using the standard Rosetta model and the explicit side chain entropy model (equation 2). See Table 2.2 for a more complete breakdown of these results.

TABLES

Table 2.1 Average side chain entropy per residue as calculated with a variety of approaches These results suggest that covariant motion between side chain rotamers does not significantly reduce side chain conformational entropy.

		TS^a (kcal/mol)	TS^b (cluster) (kcal/mol)	TS^c(chi) (kcal/mol)
Clusters with Rosetta	Surface	0.62	0.60	0.68
	Boundary	0.44	0.43	0.48
	Buried	0.26	0.24	0.28
	Average	0.47	0.45	0.51
Eglin C MD simulation		0.45	0.45^*	0.53

TS^a : calculated from Monte Carlo sampling method by treating each residue independently (equation 2) . TS^b : calculated from complete enumeration of rotamer combination of 6-residue clusters (equations 3 & 4). TS^{b*}: calculated from rotamer pairs from MD simulation (equation 5). TS^c : calculated by treating each torsion angle independently(equation 6). T=310K.

Table 2.2 Energies and entropies as a function of environment
This table illustrates the compensation that takes place between entropy and energy when the explicit side chain entropy model is incorporated into Rosetta. On the surface residues have more conformational entropy, but in addition the average energy of the rotamer ensemble (U_{avg}) at each sequence position is significantly greater than the energy of the most favorable rotamer (U_{best}). The last pair of columns (($U_{avg} - U_{best}$) − TS) shows the net effect of using the explicit side chain entropy model.

Amino acid	**U_{avg} [a]**		**TS**		**U_{avg} [a] – U_{best} [b]**		**(U_{avg} [a] – U_{best} [b]) - TS**	
	Surface	Buried	Surface	Buried	Surface	Buried	Surface	Buried
ALA	-0.79	-1.82	0.00	0.00	-0.01	0.01	-0.01	0.01
CYS	1.60	0.80	0.38	0.22	0.11	0.12	-0.27	-0.10
ASP	-0.74	-1.50	0.79	0.47	0.37	0.24	-0.42	-0.23
GLU	-0.56	-1.32	1.34	0.79	0.49	0.29	-0.85	-0.50
PHE	-0.84	-3.10	0.29	0.07	0.14	0.06	-0.15	-0.01
GLY	-1.55	-1.36	0.00	0.00	0.00	-0.01	0.00	-0.01
HIS	0.10	-0.94	0.75	0.42	0.29	0.17	-0.46	-0.25
ILE	-0.84	-2.84	0.42	0.19	0.22	0.11	-0.20	-0.08
LYS	-0.22	-1.09	1.46	0.83	0.75	0.38	-0.71	-0.45
LEU	-1.02	-2.74	0.34	0.15	0.20	0.10	-0.14	-0.05
MET	-0.21	-1.84	1.17	0.50	0.43	0.19	-0.74	-0.31
ASN	-0.65	-1.00	1.01	0.61	0.45	0.27	-0.56	-0.34
PRO	-1.10	-2.03	0.32	0.27	0.13	0.10	-0.19	-0.17
GLN	-0.72	-1.25	1.26	0.69	0.47	0.30	-0.79	-0.39
ARG	-0.57	-1.30	1.60	0.77	0.62	0.31	-0.98	-0.46
SER	-0.48	-0.93	1.17	1.04	0.20	0.23	-0.97	-0.81
THR	-0.84	-1.57	0.87	0.74	0.17	0.14	-0.70	-0.60
VAL	-0.94	-2.69	0.18	0.09	0.12	0.05	-0.06	-0.04
TRP	-0.26	-2.08	0.49	0.08	0.20	0.06	-0.29	-0.02
TYR	-0.77	-2.43	0.70	0.47	0.14	0.13	-0.56	-0.34

The values in the table are derived from repacking simulations with 2832 pdb files using the standard Rosetta model and the new explicit side chain entropy model (equation 2).
U_{avg} [a] : average energy derived from repacking simulations at 310K
U_{best} [b] : the energy of the most favorable rotamer calculated from the standard Monte Carlo simulated annealing protocol

Table 2.3 Entropies as a function of rotamer library size

Including sub-rotamers (ex1) to more finely sample side chain conformational space increases the overall conformational entropy of buried and exposed positions, but does not dramatically perturb the change in entropy between surface and buried positions.

Amino Acid	default rotamer set[a]			ex1[b]		
	$TS_{Surface}$	TS_{Buried}	ΔTS	$TS_{Surface}$	TS_{Buried}	ΔTS
ALA	0.00	0.00	0.00	0.00	0.00	0.00
CYS	0.38	0.22	0.16	1.02	0.86	0.16
ASP	0.79	0.47	0.32	1.44	1.06	0.38
GLU	1.34	0.79	0.55	2.00	1.40	0.60
PHE	0.29	0.07	0.22	0.90	0.57	0.33
GLY	0.00	0.00	0.00	0.00	0.00	0.00
HIS	0.75	0.42	0.33	1.39	0.93	0.46
ILE	0.42	0.19	0.23	1.07	0.83	0.24
LYS	1.46	0.83	0.63	2.11	1.43	0.68
LEU	0.34	0.15	0.19	0.99	0.76	0.23
MET	1.17	0.50	0.67	1.83	1.12	0.71
ASN	1.01	0.61	0.40	1.66	1.18	0.48
PRO	0.32	0.27	0.05	0.98	0.93	0.05
GLN	1.26	0.69	0.57	1.92	1.28	0.64
ARG	1.60	0.77	0.83	2.26	1.35	0.91
SER	1.17	1.04	0.13	1.82	1.68	0.14
THR	0.87	0.74	0.13	1.53	1.38	0.15
VAL	0.18	0.09	0.09	0.84	0.73	0.11
TRP	0.49	0.08	0.41	1.04	0.46	0.58
TYR	0.70	0.47	0.23	1.31	0.90	0.41

default rotamer set[a] : no sub-rotamers are used for all residues

ex1[b] : use extra chi 1 sub-rotamers for all residues

The values in the table are derived from repacking simulations with the same pdb files and same Rosetta model as those used in Table 2.2.

Table 2.4 Comparison of native sequence recovery rates

Comparison of native sequence recovery rates for design simulations with and without the explicit side chain entropy model (110 protein structures were used). The general result is that the explicit side chain entropy model does not lead to large changes in sequence recovery, as opposed to when the solvation potential is removed from the model. In each case the reference values used by Rosetta were reparameterized to ensure that the amino acids were designed with native-like frequencies.

Amino acid	**Redesigned (Raw Counts)**				Fraction designed correctly		
	Native	No Entropy[a]	Explicit Entropy[b]	No Solvation[c]	No Entropy	Explicit Entropy	No Solvation
VAL	601	608	633	543	0.46	0.48	0.14
ILE	430	436	460	425	0.43	0.45	0.12
LEU	666	697	709	701	0.48	0.49	0.11
MET	147	152	147	141	0.16	0.12	0.05
PHE	292	363	360	281	0.52	0.47	0.14
GLY	595	518	449	411	0.73	0.67	0.62
ALA	651	560	586	637	0.35	0.35	0.18
PRO	309	363	358	262	0.58	0.61	0.44
TRP	94	127	142	148	0.40	0.34	0.24
TYR	253	290	290	297	0.25	0.21	0.26
SER	474	442	452	426	0.20	0.19	0.19
THR	450	433	432	480	0.19	0.20	0.18
ASN	334	328	337	354	0.18	0.18	0.19
GLN	315	380	333	360	0.10	0.08	0.06
ASP	518	468	482	578	0.21	0.22	0.19
GLU	631	692	663	655	0.18	0.20	0.07
ARG	351	325	343	363	0.10	0.12	0.07
LYS	658	607	609	688	0.15	0.16	0.14
HIS	135	115	119	154	0.10	0.12	0.05
CYS	189	189	189	189	1.00	1.00	1.00
Total	8093	8093	8093	8093	**0.32**	**0.32**	**0.19**

a : standard Rosetta energy function

b : explicit entropy and free energy is included in the energy function (equation 2)

c : standard Rosetta without the solvation model

Table 2.5 Environmental preferences of the amino acids

Environmental preferences of the amino acids in design simulations with and without the side chain entropy model (110 protein structures were used).

Amino acid	% amino acids buried			% amino acids surface		
	Native	No Entropy[a]	Explicit Entropy[b]	Native	No Entropy	Explicit Entropy
VAL	0.55	0.53	0.55	0.17	0.19	0.17
ILE	0.54	0.58	0.57	0.17	0.10	0.12
LEU	0.51	0.45	0.48	0.15	0.18	0.14
MET	0.46	0.55	0.49	0.20	0.13	0.17
PHE	0.61	0.63	0.59	0.11	0.10	0.10
GLY	0.19	0.19	0.20	0.48	0.48	0.47
ALA	0.37	0.50	0.55	0.33	0.19	0.15
PRO	0.15	0.15	0.17	0.60	0.59	0.56
TRP	0.55	0.48	0.46	0.15	0.17	0.20
TYR	0.51	0.30	0.30	0.18	0.27	0.30
SER	0.15	0.07	0.06	0.54	0.71	0.67
THR	0.17	0.08	0.09	0.41	0.63	0.57
ASN	0.13	0.04	0.05	0.54	0.81	0.77
GLN	0.13	0.06	0.04	0.46	0.67	0.74
ASP	0.11	0.11	0.10	0.62	0.65	0.60
GLU	0.10	0.10	0.06	0.54	0.47	0.59
ARG	0.12	0.20	0.11	0.49	0.27	0.45
LYS	0.09	0.21	0.18	0.52	0.34	0.35
HIS	0.24	0.23	0.20	0.36	0.41	0.38
CYS	0.70	0.70	0.70	0.09	0.09	0.10

a : standard Rosetta energy function

b : explicit entropy and free energy is included in the energy function (equation 2)

SUPPLEMENTAL MATERIAL

Table 2.6 PDB codes used in the design simulation

PDB ID	CHAIN ID	PDB ID	CHAIN ID
1A62	_	1IG5	A
1A8O	_	1IGD	_
1AAC	_	1IGQ	A
1ABA	_	1IIB	A
1AIL	_	1IQZ	A
1AWD	_	1J75	A
1B0N	A	1JHG	A
1B67	A	1JO8	A
1BBZ	A	1K61	A
1BF4	A	1KQ1	A
1BKB	_	1KTH	A
1BKF	_	1KU3	A
1BKR	A	1KW4	A
1BRF	A	1L9L	A
1BX7	_	1LDD	A
1C4Q	A	1LJO	A
1C5E	A	1LKK	A
1C75	A	1MGQ	A
1C9O	A	1MGT	A
1CC8	A	1MHN	A
1CKA	A	1NG2	A
1CTJ	_	1NKD	_
1CUK	_	1NME	B
1CZP	A	1O13	A
1D3B	A	1OAI	A
1D4T	A	1ON2	A
1D7Y	A	1OR7	A
1DD3	A	1OR7	C
1DJ7	B	1PLC	_
1DP7	P	1PSR	A
1E0B	A	1PTF	_
1EN2	A	1PWT	_
1ERV	_	1QTN	B
1EZG	A	1R69	_
1F9M	A	1RB9	_
1FJL	A	1RRO	_
1FK5	A	1SEM	A
1FPO	A	1TAF	A
1FR3	A	1TMY	_
1FS1	B	1UTG	_
1FSE	A	1VFY	A
1G2B	A	1VIE	_
1G2R	A	1WAP	A
1G3P	_	1YCC	_
1G6X	A	2HDD	A

1G8F	A	2IGD	_
1GCQ	C	2IHL	_
1GEF	A	2MCM	_
1GUT	A	2PHY	_
1HG7	A	2PVB	A
1I07	A	2TRX	A
1I0V	A	2UAG	A
1I27	A	4RXN	_
1I2T	A	7FD1	A
1I5Z	A	256B	A

_ : no chain ID

REFERENCES

1. Pickett SD, Sternberg MJ. Empirical scale of side-chain conformational entropy in protein folding. J Mol Biol 1993;231(3):825-839.

2. Doig AJ, Sternberg MJ. Side-chain conformational entropy in protein folding. Protein Sci 1995;4(11):2247-2251.

3. Abagyan R, Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 1994;235(3):983-1002.

4. Koehl P, Delarue M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. J Mol Biol 1994;239(2):249-275.

5. Blaber M, Zhang X, Lindstrom JD, Pepiot SD, Baase WA, Matthews BW. Determination of α-helix propensity within the context of a folded protein. Sites 44 and 131 in bacteriophage T4 lysozyme. J Mol Biol 1994;235:600-624.

6. Creamer TP, Rose GD. Alpha-helix-forming propensities in peptides and proteins. Proteins 1994;19(2):85-97.

7. Lee KH, Xie D, Freire E, Amzel LM. Estimation of changes in side chain configurational entropy in binding and folding: general methods and application to helix formation. Proteins 1994;20(1):68-84.

8. Creamer TP, Rose GD. Side-chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities. Proc Natl Acad Sci U S A 1992;89(13):5937-5941.

9. Tuffery P, Etchebest C, Hazout S. Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. Protein Eng 1997;10(4):361-372.

10. Street AG, Mayo SL. Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. Proc Natl Acad Sci U S A 1999;96(16):9074-9076.

11. Vasquez M. Modeling side-chain conformation. Curr Opin Struct Biol 1996;6(2):217-221.

12. Liang S, Grishin NV. Side-chain modeling with an optimized scoring function. Protein Sci 2002;11(2):322-331.

13. Gautier R, Tuffery P. Critical assessment of side-chain conformational space sampling procedures designed for quantifying the effect of side-chain environment. J Comput Chem 2003;24(15):1950-1961.

14. Schafer H, Smith LJ, Mark AE, van Gunsteren WF. Entropy calculations on the molten globule state of a protein: side-chain entropies of alpha-lactalbumin. Proteins 2002;46(2):215-224.

15. Kussell E, Shimada J, Shakhnovich EI. Excluded volume in protein side-chain packing. J Mol Biol 2001;311(1):183-193.

16. Creamer TP. Side-chain conformational entropy in protein unfolded states. Proteins 2000;40(3):443-450.

17. Cole C, Warwicker J. Side-chain conformational entropy at protein-protein interfaces. Protein Sci 2002;11(12):2860-2870.

18. Doig AJ, Gardner M, Searle MS, Williams DH. Thermodynamics of Side Chain Internal Rotations - Effects on Protein Structure and Stability; 1993. 557-566 p.

19. Doig AJ. Thermodynamics of amino acid side-chain internal rotations. Biophysical Chemistry 1996;61:131-141.

20. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82-87.

21. Desjarlais JR, Handel TM. De novo design of the hydrophobic cores of proteins. Protein Science 1995;4:2006-2018.

22. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J Mol Biol 1987;193:775-791.

23. Koehl P, Levitt M. De novo protein design. I. In search of stability and specificity. J Mol Biol 1999;293(5):1161-1181.

24. Havranek JJ, Harbury PB. Automated design of specificity in molecular recognition. Nat Struct Biol 2003;10(1):45-52.

25. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;282:1462-1467.

26. Richardson JM, Lopez MM, Makhatadze GI. Enthalpy of helix-coil transition: missing link in rationalizing the thermodynamics of helix-forming propensities of the amino acid residues. Proc Natl Acad Sci U S A 2005;102(5):1413-1418.

27. Penel S, Doig AJ. Rotamer strain energy in protein helices - quantification of a major force opposing protein folding. J Mol Biol 2001;305(4):961-968.

28. Koehl P, Levitt M. Structure-based conformational preferences of amino acids. Proc Natl Acad Sci U S A 1999;96(22):12524-12529.

29. Kinnear BS, Jarrold MF. Helix formation in unsolvated peptides: side chain entropy is not the determining factor. J Am Chem Soc 2001;123(32):7907-7908.

30. Avbelj F, Fele L. Role of main-chain electrostatics, hydrophobic effect and side-chain conformational entropy in determining the secondary structure of proteins. J Mol Biol 1998;279(3):665-684.

31. Filikov AV, Hayes RJ, Luo P, Stark DM, Chan C, Kundu A, Dahiyat BI. Computational stabilization of human growth hormone. Protein Sci 2002;11(6):1452-1461.

32. Looger LL, Hellinga HW. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. J Mol Biol 2001;307(1):429-445.

33. Angrand I, Serrano L, Lacroix E. Computer-assisted re-design of spectrin SH3 residue clusters. Biomol Eng 2001;18(3):125-134.

34. Lopez de la Paz M, Lacroix E, Ramirez-Alvarado M, Serrano L. Computer-aided design of beta-sheet peptides. J Mol Biol 2001;312(1):229-246.

35. Kono H, Saven JG. Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. J Mol Biol 2001;306(3):607-628.

36. Mendes J, Baptista AM, Carrondo MA, Soares CM. Implicit solvation in the self-consistent mean field theory method: sidechain modelling and prediction of folding free energies of protein mutants. J Comput Aided Mol Des 2001;15(8):721-740.

37. Shenkin PS, Farid H, Fetrow JS. Prediction and evaluation of side-chain conformations for protein backbone structures. Proteins 1996;26(3):323-352.

38. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 2000;97(19):10383-10388.

39. Kuhlman B, Dantas G, Ireton G, Varani G, Stoddard B, Baker D. Design of a novel globular protein fold with atomic level accuracy. Science 2003;302:1364-1368.

40. Dunbrack RL, Jr., Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol 1994;1(5):334-340.

41. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661-1681.

42. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins: Struct Func Genet 1999;35:132-152.

43. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 2003;326(4):1239-1259.

44. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved Recognition of Native-Like Protein Structures Using a Combination of Sequence-Dependent and Sequence-Independent Features of Proteins. 34 1999:82-95.

45. Bower MJ, Cohen FE, Dunbrack RL, Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267(5):1268-1282.

46. Peterson RW, Dutton PL, Wand AJ. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Sci 2004;13(3):735-751.

47. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 2001;311(2):421-430.

48. Dunbrack RL, Karplus M. Backbone-dependent rotamer library for proteins-An application to side-chain prediction. J Mol Biol 1993;230:543-574.

49. Hu H, Clarkson MW, Hermans J, Lee AL. Increased rigidity of eglin c at acidic pH: evidence from NMR spin relaxation and MD simulations. Biochemistry 2003;42(47):13856-13868.

50. Leach AR, Lemon AP. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. Proteins 1998;33(2):227-239.

51. Butterfoss GL, Hermans J. Boltzmann-type distribution of side-chain conformation in proteins. Protein Science 2003;12(12):2719-2731.

52. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. Curr Opin Struct Biol 2000;10(2):139-145.

53. Best RB, Clarke J, Karplus M. What Contributions to Protein Side-chain Dynamics are Probed by NMR Experiments? A Molecular Dynamics Simulation Analysis. J Mol Biol 2005;349(1):185-203.

54. Prabhu NV, Lee AL, Wand AJ, Sharp KA. Dynamics and entropy of a calmodulin-peptide complex studied by NMR and molecular dynamics. Biochemistry 2003;42(2):562-570.

55. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M. Simultaneous determination of protein structure and dynamics. Nature 2005;433(7022):128-132.

CHAPTER 3

COMPUTER-BASED REDESIGN OF A β-SANDWICH PROTEIN SUGGESTS THAT EXTENSIVE NEGATIVE DESIGN IS NOT REQUIRED FOR *DE NOVO* β-SHEET DESIGN

Xiaozhen Hu[1], Huanchen Wang[1], Hengming Ke[1] and Brian Kuhlman[1*]

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

[*]corresponding author.

This work is in press in Structure.

ABSTRACT

The *de novo* design of globular β-sheet proteins remains largely an unsolved problem. It is unclear if most designs are failing because the designed sequences do not have favorable energies in the target conformations or if more emphasis should be placed on negative design, i.e. explicitly identifying sequences that have poor energies when adopting undesired conformations. We tested if we could redesign the sequence of a naturally occurring β-sheet protein, tenascin, with a design algorithm that does not include explicit negative design. Denaturation experiments indicate that the designs are significantly more stable than the wild type protein and the crystal structure of one design closely matches the design model. These results suggest that extensive negative design is not required to create well-folded β-sandwich proteins. However, it is important to note that negative design elements may be encoded in the conformation of the protein backbone which was preserved from the wild type protein.

Keywords: Computational Protein Design, *De Novo* Protein Design, β-sheet Design, Negative Design

INTRODUCTION

Approximately one quarter of all protein domains are made entirely from β-strands and connecting loops [1]. β-sheets and β-barrels form relatively rigid structures that serve as excellent scaffolds for loops that can evolve new molecular recognition capabilities; antibodies are an excellent example of this. Despite the obvious importance of β-sheet proteins, we still do not understand them well enough to design them from first principles. Most *de novo* designed β-sheet proteins are prone to aggregation, and there are no *de novo* designs of an all β-sheet protein with more than three β-strands that have been validated with a NMR or crystal structure [2-6]. In contrast, several *de novo* designs of all helical or mixed α/β proteins have been validated with high resolution structures [7-10].

There may be several reasons why designed globular β-sheet proteins are prone to misfolding and aggregation. Many β-sheet proteins have greater sequence separation between contacting residues (high contact order) and therefore fold more slowly than helical and mixed α/β proteins [11]. Slower folding rates may allow more time for misfolding, domain swapping and aggregation. β-sheet proteins (designed and naturally occurring) are generally enriched in amino acids with a high intrinsic propensity to form β-strands [12-17]. While these amino acids are energetically favorable for the target β-sheet structure, they also have a high propensity to aggregate into fibrils or form undesired strand-strand interactions[18-20]. β-strands in two-layer β-sheet proteins often have an alternating repeat of hydrophobic and hydrophilic residues; this type of repeat is known to promote undesired strand-strand interactions [21]. β-sheet proteins that do not form barrels have exposed β-strands that may be well suited for forming edge-to-edge interactions. Indeed, it has been observed that naturally occurring β-sheet proteins contain negative design elements that protect them from unwanted edge-edge interactions [22]. These include placing charged residues on both sides of the edge strand, using

bulges and prolines to prevent optimal hydrogen bonding, and protecting the edge with other portions of the protein.

How many negative design elements are needed to create a well-folded globular β-sheet protein? Is it necessary to explicitly destabilize associations between non-native strand pairings or does the identification of a low free energy sequence for a target structure implicitly destabilize most competing states? In one study on *de novo* designed β-sheet proteins, the placement of a charged residue on the inward side of putative edge strands was shown to stabilize the monomer versus the aggregated state [23]. This result suggests that negative design elements may not need to be spread throughout the entire sequence. However, high resolution structures have not been solved for these designs, so it is not known if they are adopting the target structure. Other studies in *de novo* β-sheet design have also produced monomeric proteins, but in these cases it is also not certain if the proteins are adopting the target topology [24-26]. A recent design of a Rubredoxin mimic is most likely adopting the target fold, but in this case the energy gained from metal binding may preclude the need for extensive negative design [27].

In a previous study we used the design module of the molecular modeling program Rosetta to design a new amino acid sequence for the third FNIII domain of the protein tenascin [28]. This domain has 89 residues and forms a Greek Key fold with three β−strands in one sheet and four β-strands in the second sheet. Sheet 1 is formed by strands 1, 2 and 5. Sheet 2 is formed by strands 3, 4, 6 and 7. The side chains were removed from the protein and computational protein design was used to redesign the protein with no explicit knowledge of the wild type sequence. The only energy gap that was explicitly optimized was between the folded state and a reference energy that models the unfolded state and is based on amino acid composition. Rosetta's energy function is dominated by terms that model van der Waals forces, steric repulsion, desolvation energies, torsion energies and

hydrogen bonds [9,29]. Unfortunately, the designed protein, called TEN-D1, aggregated and we were not able to characterize it. This design may have failed because we did not identify a favorable sequence for the target state, or it may have failed because we did not sufficiently destabilize misfolded and aggregated states. Here, we further pursue this question by characterizing a new set of redesigns for the third FNIII domain of tenascin, but with an energy function that has been specifically parameterized for β-sheet design. As before, we do not include any explicit negative design in the protocol.

RESULTS

Reparameterizing the Rosetta Energy Function

The energy function used by Rosetta for protein design is a weighted sum of a damped 12-6 Lennard-Jones term, an implicit solvation model, an orientation dependent hydrogen bonding term, knowledge-based torsion energies and a set of reference values that control the relative favorability of the 20 amino acids [29]. The weights on these terms have been set to maximize the native sequence recovery during the complete redesign of whole proteins [30]. Our standard training set has a mixture of all helical, mixed α/β proteins, and all β proteins. For these studies we assembled a set of 121 high-resolution structures of all β-sheet proteins. The standard Rosetta energy function was used to design sequences for the proteins in the training set and the sequences were compared to the wild type sequences. Overall sequence identity was similar to what we have observed previously, but the fraction of hydrophobic residues in the redesigned sequences was higher than in the naturally occurring sequences (67% versus 53%, **Supplementary Table 3.3 and Table 3.4**). To create more native-like sequences, iterative rounds of perturbing the amino acid reference values and redesigning the proteins were used to arrive at a set of reference values that accurately reproduce the

hydrophobic/hydrophilic preferences of the naturally occurring β-sheet proteins (**Supplementary Table 3.3 and Table 3.4**). The goal of our fitting procedure is to improve our ability to perform positive design and find low energy sequences for target structures. However, by adjusting the amino acid reference values and therefore perturbing the overall amino acid composition of the protein we may be implicitly including negative design in our protocol. In this regard, our experiments are testing the importance of explicit negative design with the constraint that overall amino acid composition has been set to resemble naturally occurring β-sheet proteins.

Computational Redesign of Tenascin

Tenascin (pdbcode : 1ten) was used as the starting model for fixed backbone design. All the sidechains were removed from the protein except Tyrosine 869. Tyrosine 869 was not allowed to vary because it forms a sidechain backbone hydrogen bond that is important for the stability of the protein [31]. Rosetta prefers to put a phenylalanine at this position because the tyrosine rotamers used during the simulation do not allow for a low energy hydrogen bond. This residue was mutated to a phenylalanine in our previously published redesign of tenascin TEN-D1. 100 independent design trajectories were used to look for low energy sequences. The Rosetta full atom energies in the redesigned models varied between -220 and -215 kcal / mol. The lowest energy model, called TEN-D2, was chosen for experimental characterization.

A second round of design simulations were performed with an additional surface area- based packing score (SASAprob) included in the optimization procedure [32]. The SASAprob score examines the difference in solvent accessibility computed with a 0.5 Å probe and a 1.4 Å probe (the size of water). The difference in these two terms will be greater for underpacked proteins. The score is formulated as a probability based on average values measured for naturally occurring proteins. To optimize this score during a design simulation we have developed a rapid algorithm for computing solvent

accessible surface areas during protein design simulations. Our design picked from the first round of simulations, TEN-D2, has a SASAprob score of 0.46, indicating that it is more tightly packed than 46% of the proteins in the PDB. From the second round of simulations, we chose a design called TEN-D3, with a SASAprob score of 0.52 and a total score of -216 kcal / mol.

TEN-D2 has 53 mutations and TEN-D3 has 51 mutations when compared to the wild type sequence (**Figure 3.1, Table 3.1**). Our previously characterized sequence, TEN-D1, had 58 mutations. Highest sequence similarity is seen in the protein core; out of 20 buried residues, 9 were mutated in TEN-D2, and 8 were mutated in TEN-D3. The number of charged residues in the redesigns is significantly different than in the wild type protein. 20% of the wild type residues are negatively charged (Asp or Glu), while only 8% of the redesigns are negatively charged. The most highly conserved amino acids in the redesigns are proline, glycine and threonine. Four out of five prolines, three out of five glycines and ten out of 12 threonines are conserved.

Experimental Characterization

Both TEN-D2 and TEN-D3 were expressed in bacteria and experimentally characterized using a variety of biophysical methods. Size-exclusion chromatographies of the two redesigns suggest they are both monomeric (data not shown). There is good dispersion in the one-dimensional ^{1}H NMR spectra indicating that both redesigns are well-folded (**Figure 3.2**), and there are amide protons with chemical shifts above 8.5 ppm, indicative of β-sheet structure. Additionally, the circular dichroism (CD) spectra of the proteins are consistent with β-sheet structure. To probe the stability of the redesigns CD signal was monitored as a function of temperature and concentration of chemical denaturant at a single wavelength. Both TEN-D2 and TEN-D3 unfold at temperatures that are significantly higher than the wild type protein, the proteins unfold above 90 °C and 80 °C respectively (**Figures 3.3 and 3.4, Table 3.2**). The T_m for wild type tenascin is 58 °C. However,

unlike the wild type protein, the thermal unfolding curves for the redesigns are not reversible at pH 7. It has been shown that high net charges can help solubilize proteins in the unfolded state [33]. Consistent with this hypothesis, TEN-D2 refolds reversibly when the pH is dropped below the pK_a of the acidic side chains, increasing the net charge of the design (**Figure 3.3.D**).

Denaturation induced by guanidine hydrochloride was monitored with circular dichroism to measure the stability. Both redesigns fold reversibly in chemical denaturant and are significantly more stable than the wild type protein (**Figure 3.4**). The extrapolated free energies of folding are -11.9 and -8.7 kcal / mol respectively. The wild type protein has a free energy of folding of -5.1 kcal / mol. Interestingly, the m-values (slope of free energy versus [GuHCl]) are larger for the redesigns. This suggests that the redesigns bury more hydrophobic surface area upon folding than the wild type protein [34].

Structure Determination

The crystal structure of the TEN-D3 was determined at 2.4 Å resolution by X-ray crystallography (**Supplementary Table 3.5**) to verify that the structure matches the design model. Overall, there is a good match between the crystal structure and the design model, the root-mean-square deviation (RMSD) between the crystal structure and the design model is less than 0.8 Å for all heavy atoms of the protein (**Figure 3.5**). 82 percent of the sidechains have the same chi1 rotamer as designed and all the rotamers in the core have the same conformation (chi1 and chi2) as designed. Greater differences were seen on the surface; although several designed salt bridges on the protein surface were observed in the crystal structure. These include interacting pairs, Arg 74 and Asp 48, Asp 43 and Arg 37, and Glu 62 and Arg 37.

DISCUSSION

60% of the residues in the tenascin redesigns are not a direct reflection of natural protein evolution, but rather were chosen solely based on a calculated free energy difference between the target structure and a reference state that only depends on amino acid composition. Despite the simplicity of this design criterion, the proteins fold into the target structure. Similar findings have been reported for all helical, mixed α/β proteins, and small three stranded β-sheet proteins [6,28,35,36]. Our result suggests that the majority of amino acids in tenascin have not been explicitly selected to prevent misfolding, but rather selection for a low free energy target structure is sufficient to destabilize alternative folds. This result is not obvious *a priori*, given the fact that small stretches of sequence rich in β-sheet propensity are prone to association and the possible number of non-native strand pairings is much greater than native pairings.

Our results do not indicate that negative design is not important for *de novo* β−sheet design, but they do suggest that it may be sufficient to only focus on a limited number of negative design elements. For instance, the backbone conformation of tenascin appears to include negative design elements. Unwanted edge-to-edge β-strand interactions are most likely destabilized by a β-bulge in strand 1, the shortness of strand 5 and prolines in strand 7. All of these elements are preserved in our redesigns. Additionally, negative design elements may be encoded in the residues that are preserved from the wild type sequence. It is interesting that our designs do not include charged residues on the inward pointing face of the edge strands. Other design elements, such as the prolines in strand 7, must be preventing association between edge strands.

It is striking that the redesigned sequences are considerably more stable than the wild type sequences. Similar results have been observed when redesigning other protein folds with computational protein design software [28,37]. An increase in the m-values for chemical denaturation suggests that the designs

bury more hydrophobic surface area upon folding. This increase is consistent with the addition of extra hydrophobic residues in the redesigns and may explain the increase in protein stability.

Our results are encouraging in that they suggest that the *de novo* design of a β-sandwich protein may be possible without extensive consideration of strand mis-pairings. Despite this fact, *de novo* design is still a very challenging problem. To create a protein from scratch, it is necessary to identify a protein backbone that allows for tight packing of the side chains and allows for hydrogen bonding to buried polar groups. It is especially challenging to ensure that backbone polar groups in the connecting loops have hydrogen bond partners. Many of these polar groups are removed from solvent, and in naturally occurring proteins are engaged in sidechain-backbone hydrogen bonds. It will be exciting to see if new techniques in computational protein design that allow for backbone sampling and sequence design will allow these hurdles to be overcome.

EXPERIMENTAL PROCEDURES

Sequence Optimization Simulations

Fixed backbone design simulations were performed with svn version 9242 of Rosetta. The standard full atom energy function was used except for the following changes: the reference values were reparameterized to maximize the native sequence recovery test, the desolvation penalty for histidine was increased by varying the ddGfree parameter for histidine nitrogens from -4.0 to -9.0, and the Lennard-Jones potential was set to a linear slope at 0.85 of the van der Waals radius (instead of 0.6). Dunbrack's backbone dependent rotamer library was used with extra chi 1 torsion angles for all residues and extra chi 2 torsion angles for aromatic residues. The command line used for the simulations was: *Rosetta.gcc -s 1ten.pdb -design -fixbb -use_bw -ex1 -ex2aro_only -extrachi_cutoff 1 -resfile resfile -ndruns 100 (-use_sasa_pack_score)*

.

Protein Expression and Purification

Genes for the redesigned proteins were synthesized in-house with PCR extension of commercially purchased overlapping oligonucleotides from Operon [38]. The genes were inserted into *E. coli.* expression vector pET21b, with a linker "GSLE" followed by C terminal 6x His tag. The proteins were expressed in the *E. coli.* BL21 strain at 37 °C with 0.5 mM IPTG used for induction. The proteins were purified with a Ni^{++} affinity column followed by size-exclusion chromatography (Superdex-75).

NMR

The two redesigned proteins (~0.4 mM) were equilibrated in 20 mM sodium phosphate, 0.15 M NaCl, pH 7.2 buffer and one-dimensional ^{1}H NMR spectra were recorded at 25 °C on a Varian Inova 600 MHz spectrometer. NMR data were processed with NMRPipe [39].

Circular Dichroism

CD data were collected on a JASCO J-810/815 CD spectrometer using a 0.1 cm cuvette with 40 uM proteins. The CD signal was monitored at 215 nm as a function of temperature (4 – 96 °C). The fraction of unfolded protein was calculated assuming that the CD signal of the unfolded and folded protein varies linearly with temperature. GuHCl induced chemical denaturation experiments were recorded at 222 nm. The free energy calculations were obtained with a two-state assumption.

Crystallization, X-ray Diffraction and Structure Determination

The hanging-drop vapor diffusion method was used for crystallization trials. TEN-D3 with the concentration of 12 mg / mL in 100 mM NaCl, 20 mM Tris buffer at pH 7.4, was mixed with an equal volume of well buffer of 0.1 M sodium dihydrogen phosphate, 0.1 M potassium dihydrogen phosphate, 0.1 M MES, pH 6.5, 2.2M NaCl and100 mM urea. 20% glycerol was used as the

cyroprotectant. Diffraction data of TEN-D3 were collected at the Beamline x29A at Brookhaven National Laboratory.

The data were indexed and processed with the program HKL2000 [40]. The structure of TEN-D3 was solved by molecular replacement using the programs MolRep [41] and Phaser [42]. Wild type tenascin (PDB code 1TEN) was used as the initial search model. The model was then refined against the synchrotron data to 2.4 Å resolution. O [41] was used to build the model and CNS [43] was used to refine the structure. The geometry of the final model was assessed with the program PROCHECK [44].

Accession code: The coordinates and structural factors of TEN-D3 have been deposited into the RCSB Protein Data Bank with PDB ID code 3B83.

Acknowledgements

We thank beamline X29 at National Synchrotron Light Source for diffraction data collection. This research was supported by an award from the W.M. Keck foundation and the grant GM073960 from the National Institutes of Health. We declare no conflict of interest.

FIGURES

```
TEN-WT          LDAPSQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDENQYSIGN 60
TEN-D1          LPPPYNITVTNIGPTTAVLVYVRSESPSDGYNITFGTKNDDSDRVTVTLPSENTSYVITN 60
TEN-D2          LQPPFNITVTNITLTTAVVKWLPAQLPVEGYLVTYGRKNDPSDETTVDLTSSITSLTLTN 60
TEN-D3          LQPPFNIKVTNITLTTAVVTWQPPILPIEGILVTFGRKNDPSDETTVDLTSSITSLTLTN 60

TEN-WT          LKPDTEYEVSLISRRGDMSSNPAKETFTT 89
TEN-D1          LKPNTTFQITIRSQNGDKSSPPVSTYFTL 89
TEN-D2          LTPNTEYEVRIVARNGNLYSPPVSTTFKT 89
TEN-D3          LEPNTTYEIRIVARNGQQYSPPVSTTFTT 89
```

Figure 3.1 Sequences of the wild type and three redesigned proteins
TEN-WT: wild type; TEN-D1,TEN-D2,TEN-D3: redesigned sequences. The TEN-D1 sequence is from a previously published study [28].

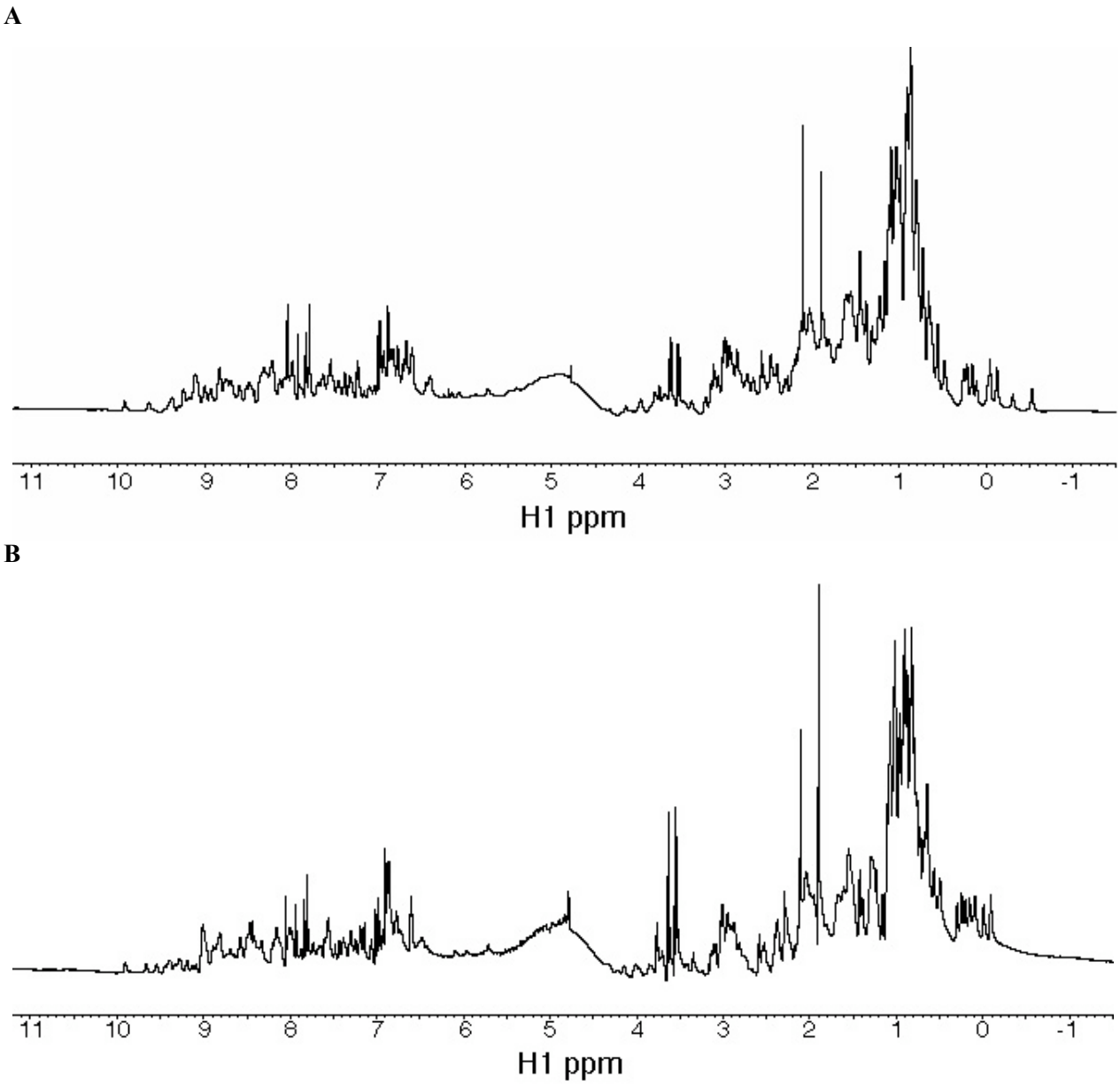


Figure 3.2 One-dimensional 1H spectra of the redesigned proteins
A: TEN-D2. B: TEN-D3.

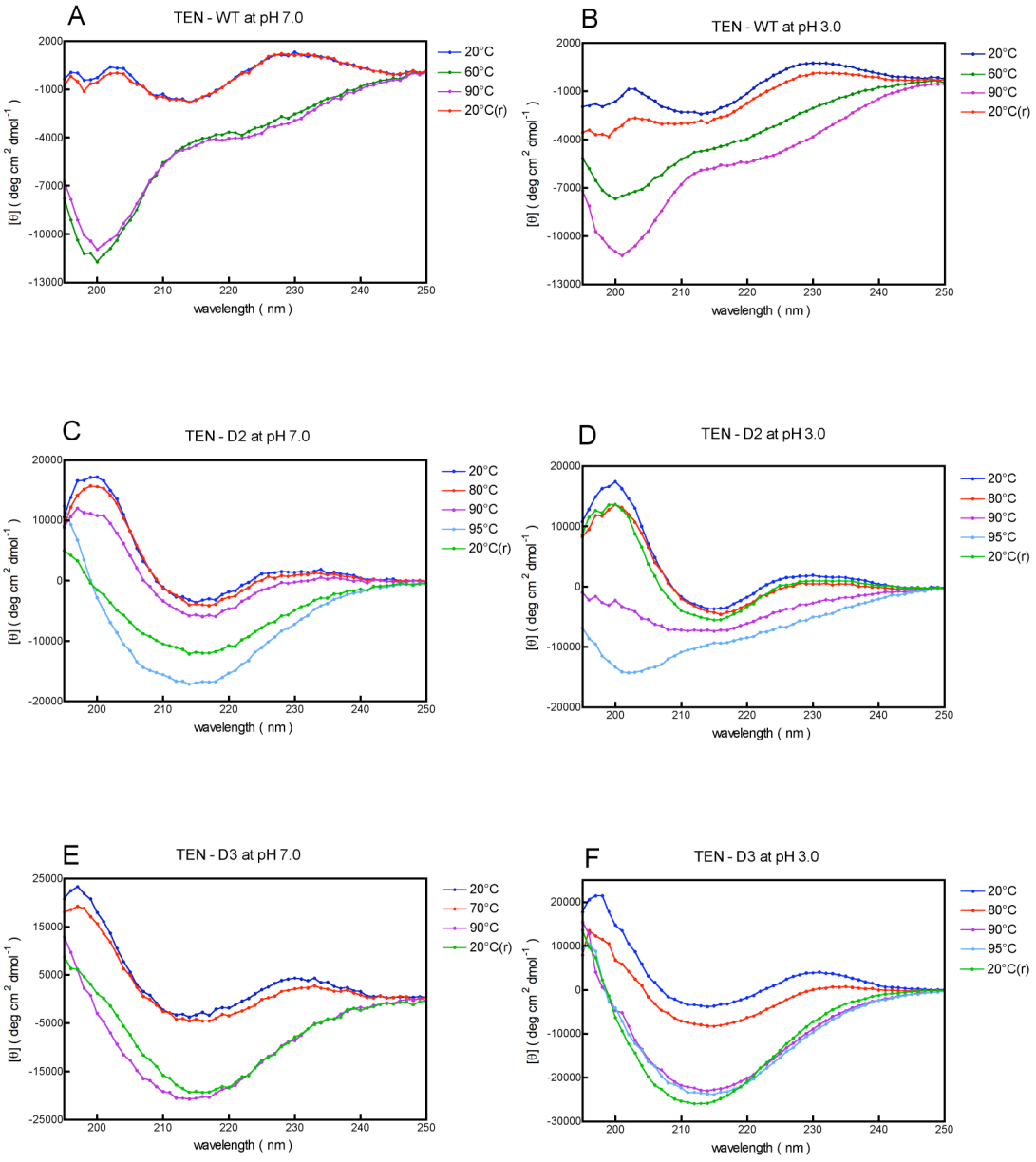


Figure 3.3 CD spectra of the wt and redesigns
Circular dichroism spectra of the wild type tenascin and the redesigned proteins at neutral and acidic pH with different temperatures. 20 °C(r) represents that the temperature was cooled back to 20 °C. A: TEN-WT at pH 7.0, B: TEN-WT at pH 3.0, C: TEN-D2 at pH 7.0, D: TEN-D2 at pH 3.0, E: TEN-D3 at pH 7.0, and F: TEN-D3 at pH 3.0.

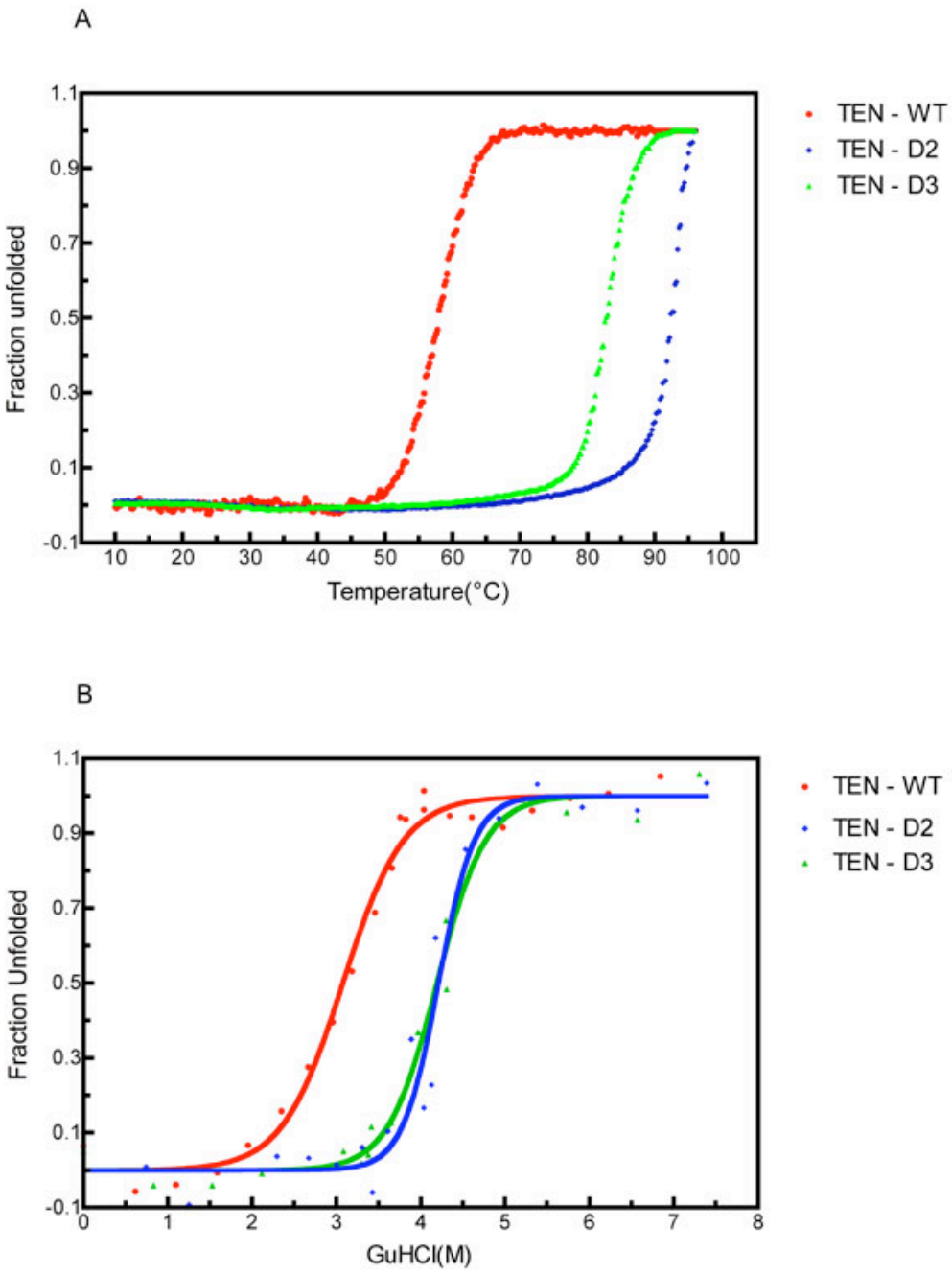


Figure 3.4 Temperature and chemical denaturation
Temperature and chemical denaturation as monitored by circular dichroism. A: Thermal unfolding of the wild type tenascin and the redesigned proteins. B: Chemical denaturation of the wild type tenascin, TEN-D2 and TEN-D3.

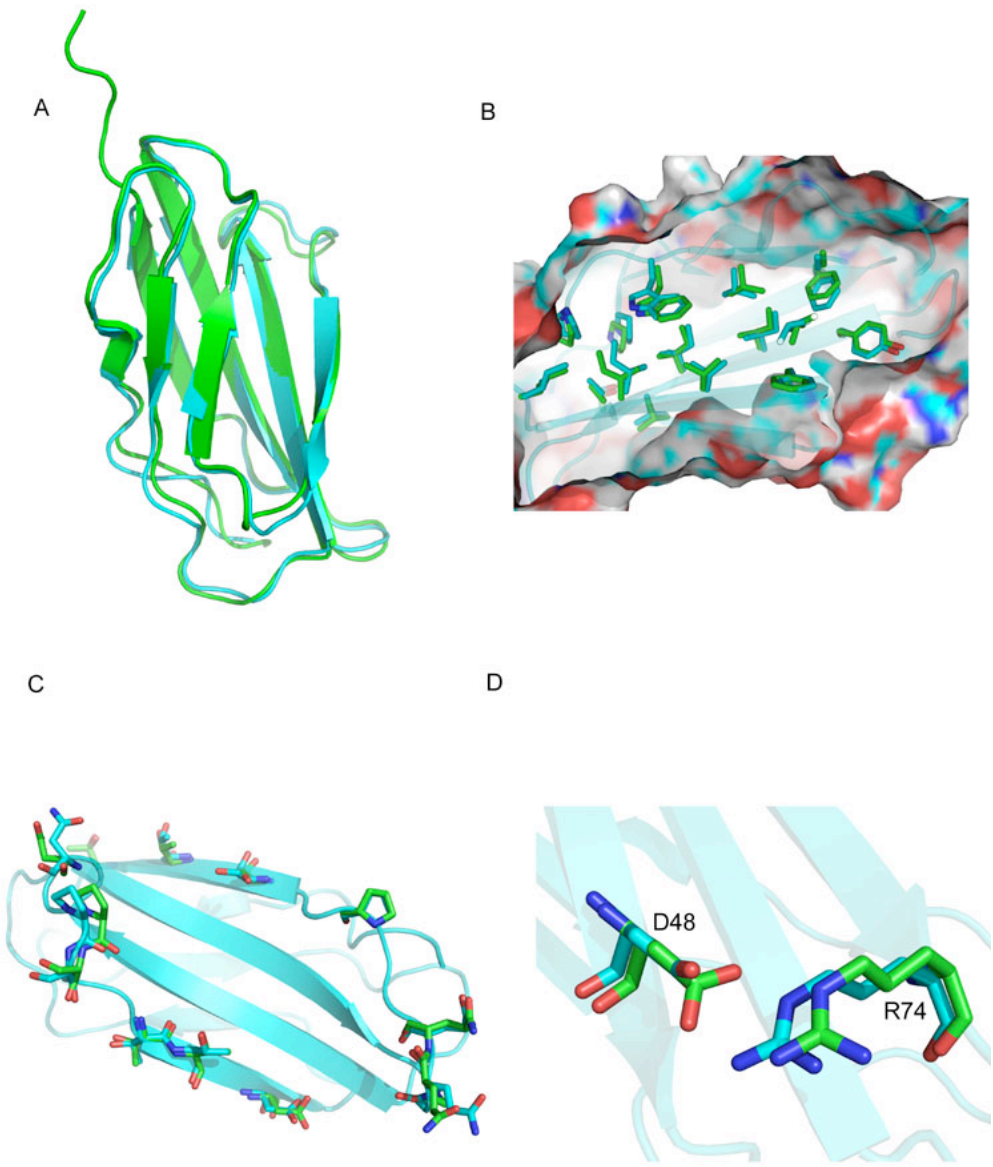


Figure 3.5 Structure alignment beteen the design model and the crystal structure
Structure alignment between the designed model (cyan) and the crystal structure of TEN-D3 (green). A: backbone only, B: buried residues, C: selected surface residues, D: a designed salt bridge between Asp 48 and Arg 74.

TABLES

Table 3.1 Sequence features of wild type and redesigned tenascin

Protein	TEN-WT	TEN-D1	TEN-D2	TEN-D3
MW (Da)	9895.9	9729.7	9800.0	9790.1
Theoretical PI	4.15	4.99	5.10	4.72
Fraction of positively charged residues	0.09	0.07	0.07	0.06
Fraction of negatively charged residues	0.20	0.08	0.08	0.08
Fraction of hydrophobic residues	0.38	0.37	0.42	0.42
Sequence identity to WT (overall)	/	31/89	36/89	38/89
Sequence identity to WT (buried*)	/	9/20	11/20	12/20
Sequence identity to TEN-D1 (overall)	31/89	/	45/89	48/89

*buried – Buried residues have more than 19 neighbors within 10Å.

Table 3.2 Thermodynamic parameters of wild type and redesigned tenascin

Protein	Tm(°C)	$\Delta G_U^{H_2O}$ (kcal mol^{-1})	m-GuHCl (kcal mol^{-1} M-^{1})
TEN–WT	58	5.1±1.2	1.7±0.3
TEN–D1	/	/	/
TEN–D2	>90	11.9±4.7	2.8±1.1
TEN–D3	>80	8.7±2.0	2.1±0.4

SUPPLEMENTARY MATERIAL

Table 3.3 Comparison of native sequence recovery rates for design simulations with the standard weight and modified beta sheet weight

Amino acid		**Redesigned (Raw Counts)**		**Fraction designed correctly**	
	Native	std[a]	bw[b]	std[a]	bw[b]
VAL	874	757	861	0.48	0.51
ILE	515	646	541	0.52	0.51
LEU	699	853	724	0.53	0.54
MET	157	156	159	0.17	0.17
PHE	297	602	321	0.54	0.45
GLY	793	825	765	0.81	0.79
ALA	635	545	638	0.34	0.41
PRO	393	767	425	0.82	0.62
TRP	130	330	151	0.41	0.39
TYR	301	604	339	0.3	0.3
SER	609	412	604	0.16	0.22
THR	640	510	625	0.23	0.26
ASN	385	381	376	0.23	0.25
GLN	291	260	272	0.07	0.06
ASP	522	390	516	0.19	0.24
GLU	620	317	641	0.11	0.2
ARG	411	323	423	0.12	0.17
LYS	601	390	610	0.1	0.13
HIS	166	90	167	0.07	0.15
CYS	119	0	0	0	0
Total	9158	9158	9158	0.35	0.37

a : standard energy function
b : modified beta sheet weight

Table 3.4 Environmental preferences of the amino acids in design simulations with the standard weight and modified beta sheet weight

Amino	% amino acids buried			% amino acids surface		
acid	Native	std[a]	bw[b]	Native	std[a]	bw[b]
VAL	0.62	0.64	0.58	0.15	0.15	0.2
ILE	0.64	0.54	0.62	0.14	0.15	0.11
LEU	0.63	0.4	0.49	0.13	0.31	0.19
MET	0.42	0.53	0.56	0.22	0.13	0.14
PHE	0.63	0.49	0.72	0.15	0.23	0.06
GLY	0.31	0.32	0.31	0.39	0.37	0.39
ALA	0.46	0.67	0.61	0.27	0.08	0.16
PRO	0.28	0.22	0.31	0.42	0.5	0.4
TRP	0.52	0.3	0.5	0.14	0.3	0.16
TYR	0.51	0.23	0.37	0.16	0.35	0.21
SER	0.22	0.44	0.28	0.45	0.28	0.45
THR	0.26	0.35	0.27	0.35	0.37	0.44
ASN	0.2	0.08	0.1	0.48	0.76	0.74
GLN	0.21	0.15	0.07	0.41	0.55	0.6
ASP	0.14	0.15	0.13	0.52	0.53	0.53
GLU	0.14	0.19	0.14	0.49	0.39	0.47
ARG	0.3	0.28	0.3	0.31	0.23	0.21
LYS	0.12	0.23	0.22	0.52	0.32	0.3
HIS	0.24	0.52	0.48	0.37	0.23	0.17

a : standard energy function
b : modified beta sheet weight

Table 3.5 X-Ray diffraction data collection and refinement statistics

Data collection	
Crystal	TEN-D3
Resolution range (Å)	30-2.4
Total reflections	3,217,388
Unique reflections	43,149
Completeness	99.5%
R_{merge}	0.076
Space group	$P42_12$
Unit cell dimensions	a = b = 126.329 Å c = 134.661 Å

Refinement	
Crystal	TEN-D3
Resolution range (Å)	30-2.4
R	0.24
Rfree	0.29
rmsd bond length (Å)	0.007
rmsd bond angle	1.505
No. of protein atoms	5956
Average B factor for all atoms ($Å^2$)	50.22

REFERENCES

1. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH--a hierarchic classification of protein domain structures. Structure 1997;5(8):1093-1108.

2. Kortemme T, Ramirez-Alvarado M, Serrano L. Design of a 20-amino acid, three-stranded beta-sheet protein. Science 1998;281(5374):253-256.

3. Hughes RM, Waters ML. Model systems for beta-hairpins and beta-sheets. Curr Opin Struct Biol 2006;16(4):514-524.

4. Searle MS, Ciani B. Design of beta-sheet systems for understanding the thermodynamics and kinetics of protein folding. Curr Opin Struct Biol 2004;14(4):458-464.

5. Ramirez-Alvarado M, Kortemme T, Blanco FJ, Serrano L. Beta-hairpin and beta-sheet formation in designed linear peptides. Bioorg Med Chem 1999;7(1):93-103.

6. Kraemer-Pecore CM, Lecomte JT, Desjarlais JR. A de novo redesign of the WW domain. Protein Sci 2003;12(10):2194-2205.

7. Walsh ST, Cheng H, Bryson JW, Roder H, DeGrado WF. Solution structure and dynamics of a de novo designed three-helix bundle protein. Proc Natl Acad Sci U S A 1999;96(10):5486-5491.

8. Wei Y, Kim S, Fela D, Baum J, Hecht MH. Solution structure of a de novo protein from a designed combinatorial library. Proc Natl Acad Sci U S A 2003;100(23):13270-13273.

9. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302(5649):1364-1368.

10. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;282(5393):1462-1467.

11. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277(4):985-994.

12. Smith CK, Withka JM, Regan L. A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. Biochemistry 1994;33(18):5510-5517.

13. Nagano K. Logical analysis of the mechanism of protein folding. I. Predictions of helices, loops and beta-structures from primary structure. J Mol Biol 1973;75(2):401-420.

14. Minor DL, Jr., Kim PS. Measurement of the beta-sheet-forming propensities of amino acids. Nature 1994;367(6464):660-663.

15. Minor DL, Jr., Kim PS. Context is a major determinant of beta-sheet propensity. Nature 1994;371(6494):264-267.

16. Koehl P, Levitt M. Structure-based conformational preferences of amino acids. Proc Natl Acad Sci U S A 1999;96(22):12524-12529.

17. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 1974;13(2):211-222.

18. Pawar AP, Dubay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM. Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. J Mol Biol 2005;350(2):379-392.

19. Garcia-Castellanos R, Bonet-Figueredo R, Pallares I, Ventura S, Aviles FX, Vendrell J, Gomis-Rutha FX. Detailed molecular comparison between the inhibition mode of A/B-type carboxypeptidases in the zymogen state and by the endogenous inhibitor latexin. Cell Mol Life Sci 2005;62(17):1996-2014.

20. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 2004;22(10):1302-1306.

21. Hecht MH. De novo design of beta-sheet proteins. Proc Natl Acad Sci U S A 1994;91(19):8729-8730.

22. Richardson JS, Richardson DC. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci U S A 2002;99(5):2754-2759.

23. Wang W, Hecht MH. Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins. Proc Natl Acad Sci U S A 2002;99(5):2760-2765.

24. Yan Y, Erickson BW. Engineering of betabellin 14D: disulfide-induced folding of a beta-sheet protein. Protein Sci 1994;3(7):1069-1073.

25. Quinn TP, Tweedy NB, Williams RW, Richardson JS, Richardson DC. Betadoublet: de novo design, synthesis, and characterization of a beta-sandwich protein. Proc Natl Acad Sci U S A 1994;91(19):8747-8751.

26. Lim A, Makhov AM, Bond J, Inouye H, Connors LH, Griffith JD, Erickson BW, Kirschner DA, Costello CE. Betabellins 15D and 16D, de Novo designed beta-sandwich proteins that have amyloidogenic properties. J Struct Biol 2000;130(2-3):363-370.

27. Nanda V, Rosenblatt MM, Osyczka A, Kono H, Getahun Z, Dutton PL, Saven JG, Degrado WF. De novo design of a redox-active minimal rubredoxin mimic. J Am Chem Soc 2005;127(16):5804-5805.

28. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J Mol Biol 2003;332(2):449-460.

29. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66-93.

30. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 2000;97(19):10383-10388.

31. Hamill SJ, Cota E, Chothia C, Clarke J. Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. J Mol Biol 2000;295(3):641-649.

32. Leaver-Fay A, Butterfoss GL, Snoeyink J, Kuhlman B. Maintaining solvent accessible surface area under rotamer substitution for protein design. J Comput Chem 2007;28(8):1336-1341.

33. Lawrence MS, Phillips KJ, Liu DR. Supercharging proteins can impart unusual resilience. J Am Chem Soc 2007;129(33):10110-10112.

34. Myers JK, Pace CN, Scholtz JM. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. Protein Sci 1995;4(10):2138-2148.

35. Scalley-Kim M, Baker D. Characterization of the folding energy landscapes of computer generated proteins suggests high folding free energy barriers and cooperativity may be consequences of natural selection. J Mol Biol 2004;338(3):573-583.

36. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278(5335):82-87.

37. Malakauskas SM, Mayo SL. Design, structure and stability of a hyperthermophilic protein variant. Nat Struct Biol 1998;5(6):470-475.

38. Stemmer WP, Crameri A, Ha KD, Brennan TM, Heyneker HL. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. Gene 1995;164(1):49-53.

39. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 1995;6(3):277-293.

40. Otwinowski zaM, W. Processing of X-ray Diffraction Data Collected in Oscillation Mode. Methods in Enzymology 1997;276(Macromolecular Crystallography, part A):307-326.

41. Vagin A, Teplyakov A. An approach to multi-copy search in molecular replacement. Acta Crystallogr D Biol Crystallogr 2000;56(Pt 12):1622-1624.

42. Storoni LC, McCoy AJ, Read RJ. Likelihood-enhanced fast rotation functions. Acta Crystallogr D Biol Crystallogr 2004;60(Pt 3):432-438.

43. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54(Pt 5):905-921.

44. Laskowski RA, Moss DS, Thornton JM. Main-chain bond lengths and bond angles in protein structures. J Mol Biol 1993;231(4):1049-1067.

CHAPTER 4

HIGH RESOLUTION DESIGN OF A PROTEIN LOOP

Xiaozhen Hu, Huanchen Wang, Hengming Ke and Brian Kuhlman*

Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC, 27599

*corresponding author

This work was published in Proc Natl Acad Sci USA.(2007) Nov 6;104(45)17668-73.

ABSTRACT

Despite having irregular structure, protein loops often adopt specific conformations that are critical to protein function. Most studies in *de novo* protein design have focused on creating proteins with regular elements of secondary structure connected by very short loops or turns. To design longer protein loops that adopt specific conformations we have developed a protocol within the Rosetta molecular modeling program that iterates between optimizing the sequence and conformation of a loop in search of low energy sequence–structure pairs. We have tested the procedure by designing 10-residue loops for the connection between the second and third strand in the β-sandwich protein tenascin. Three low energy designs from 7200 flexible backbone trajectories were selected for experimental characterization. All three designs, called LoopA, LoopB and LoopC, adopt stable folded structures. High resolution crystal structures of LoopA and LoopB have been solved. LoopB adopts a structure very similar to the design model (0.46 Å RMSD) and all but one of the side chains are modeled in the correct rotamers. LoopA crystallized at low pH in a structure that differs dramatically from our design model. It forms a strand swapped dimer mediated by hydrogen bonds to protonated glutamic and aspartic acids. Gel filtration indicates that the protein is not a dimer at neutral pH. These results suggest that the high resolution design of protein loops is possible; however, they also highlight how small changes in protein energetics can dramatically perturb the low free energy structure of a protein.

INTRODUCTION

Protein loops often adopt specific conformations that are critical to protein function. Many protein active sites contain residues that are located in loops, and protein-protein interactions are frequently mediated by loops. Despite the clear importance of loops, most studies in *de novo* protein design have not involved the creation of longer loops that adopt specific conformations, but rather have focused on proteins that consist almost entirely of α-helices or β-strands connected by short turns or loops [1-8].

There are several reasons why designing ordered loops may be especially challenging. Unlike α-helices and β-sheets, the backbone hydrogen bonding potential of a loop is not automatically satisfied. For a loop conformation to have low free energy it is important that each polar group in the loop is hydrogen bonding with another group in the protein or is accessible to water. However, if too much of the loop is exposed to water then it is less likely that it will adopt a unique conformation. Additionally, unlike α-helices and β-strands there are not well determined amino acid preferences for forming a well-ordered loop. Protein engineers have shown that helices can be built by favoring sequences rich in alanine, leucine, lysine and glutamate, while the β–branched amino acids, threonine, valine and isoleucine, prefer to form β-strands [9-11]. Loops can be designed by favoring sequences enriched in glycine and polar amino acids, but sequences of this type are unlikely to form a unique conformation [12]. Well-ordered loops typically have diverse sequences that form specific tight packing interactions within the loop and with the rest of the protein.

Designing a new loop requires specification of the new backbone coordinates as well as the amino acid sequence. This is a difficult problem because most arbitrarily chosen protein backbones are unlikely to be designable, i.e. there will be no amino acid sequences that pack on the structure with

energies that are comparable to what are observed for naturally occurring proteins [13]. Designable backbones can be created using heuristics derived from naturally occurring protein structures or structure prediction protocols can be used in tandem with sequence optimization protocols to search for low energy sequence-structure pairs [2,6,14-18]. Here, we examine if the second approach, combining structure prediction with automated sequence optimization, can be used to design loop sequences that adopt unique conformations. Schliebs and co-workers have used a similar strategy to introduce 4 mutations into a 8-residue loop in triosephosphate isomerase [19]. In their study, Monte Carlo sampling of loop conformations was interspersed with hand-picked mutations.

A variety of methods have been developed to predict the structures of protein loops [20]. In general, loop prediction protocols have two primary components, a procedure for searching through the various conformations a loop might adopt and an energy function for evaluating the relative favorability of these conformations. Often conformational sampling is aided by using loop conformations from other proteins as starting points for structure optimization. We use the molecular modeling program Rosetta to perform sequence design and loop modeling. Rosetta was first developed for *ab initio* structure prediction, but has since been expanded to contain protocols for high resolution structure refinement, loop modeling, molecular docking and protein design [6,21-24]. The Rosetta energy function used for high resolution refinement and design emphasizes short range interactions: steric repulsion, Van der Waals interactions within 5.5 Å, torsion energies, hydrogen bonding and a desolvation penalty for bringing atoms close to other polar atoms [25,26]. Sequence and conformational space are searched with a Monte Carlo optimization procedure. Single amino acid substitutions or backbone torsion angle perturbations are evaluated with the Metropolis criterion. In the last stage of refinement the Monte Carlo moves are followed by gradient-based minimization of torsion angles before comparing the energy of the structure to the most recently accepted structure. To design protein loops we have combined Rosetta's loop modeling protocols with sequence optimization. The protocol iterates between refining the structure of a loop and designing a sequence

for the loop. We generally perform thousands of independent trajectories as each individual simulation eventually gets trapped in a local energy minimum.

To experimentally test the protocol we have examined whether we can design new backbone conformations and sequences for the 10 residue loop that connects the second and third β-strand from the third fibronectin type III domain from tenascin-C [27]. In these studies, the wild type loop is removed from the protein, and the new loop is designed from scratch. The naturally occurring loop that connects these strands forms a well ordered structure with low B-values in the crystal structure, and ^{15}N nuclear spin relaxation experiments have shown that the loop is fairly rigid in solution [28]. Insertion of 4 glycines into the same loop from the homologous protein FNfn10 lowers the stability of FNfn10 by 1.7 kcal / mol [29].

RESULTS AND DISCUSSION

Computational design. Alternate backbone conformations and sequences were designed for residues 24-33 of the the third fibronectin type III domain from tenascin-C [27]. A three step process was used to create the new loops. First, starting backbone conformations for loop design were picked from fragments of naturally occurring proteins. Second, iterative rounds of backbone optimization and sequence design were used to search for low energy sequence-structure pairs. Third, a variety of calculated energies were used to pick designs for experimental validation.

Starting loop structures were built by searching the PDB database for 12 residue fragments of naturally occurring proteins with endpoints that superimpose with low RMSD on the backbone atoms of residues 23 and 34 of tenascin. 142 fragments were identified with endpoint RMSDs less than 3 Å. These fragments were then grafted onto tenascin using Monte Carlo optimization and gradient-based minimization of backbone torsion angles with a scoring function that favored loop closure, low

energy backbone torsion angles, low energy backbone-backbone hydrogen bonds and the absence of clashes with neighboring backbone atoms. From these simulations, 36 low scoring backbone structures were selected for use as starting points for high resolution design. These loops varied in RMSD to the wild type loop from 0.7 Å to 2.5 Å (**supplementary Figure 4.7).**

Each starting structure was used to seed 200 independent sequence design and backbone optimization trajectories (7200 trajectories in total). Each trajectory consisted of 9 rounds of sequence optimization followed by torsion-based backbone and side chain optimization (see methods). During these runs only small backbone perturbations were made and the average backbone perturbation over the course of a trajectory was 0.3 Å RMSD. However, these small backbone perturbations often changed the most preferred sequence, out of the 10 loop residues it was common to see 4 or 5 mutations when comparing to the initial designed sequence, and the Rosetta full atom energy of the designed structure typically dropped between 5 and 10 kcal / mol (**Figure 4.1)**. In some cases the simulation fell into a local minimum after the first round of design and backbone optimization, while in other cases the energy continued to drop throughout the trajectory. For the 7200 design models, the Rosetta scores varied from -135 to -156 kcal / mol. Some starting structures produced lower energy, high-resolution models on average than other starting structures. 10 starting structures were represented in the final top 200 scoring models.

Selecting designs for experimental validation. Designs were selected for experimental validation using several criteria: the Rosetta full atom score, the number of buried polar groups without a hydrogen bonding partner [30], and the quality of packing in the protein as measured by the amount of molecular surface accessible to a 0.5 Å probe but not a water molecule (SASApack score, see methods). A first round of selection was made by eliminating all models with more than 2 unsatisfied hydrogen bonds in the region of the designed loop, a SASApack score greater than 2.0 and a total score greater than -148 kcal / mol. For reference, the wild type protein has a Rosetta score of -148

kcal / mol after relaxation, 2 unsatisfied hydrogen bonds in the region of the designed loop, and a SASApack score of 0.83. The hydrogen bond filter was the most stringent and removed 6608 models from consideration. 60 models were left after applying all three filters. From these 60, three sequences were selected for further study: the lowest scoring structure (LoopC), and two structures with low packing scores (LoopA and LoopB) (**Figure 4.2**). Each sequence differs from the wild type protein and the starting structure in at least 7 out of the 10 sequence positions.

The wild type loop on tenascin is stabilized by a set of hydrophobic residues that form closely packed interactions in the space between the beginning and end of the loop. Similar types of interactions are present in all three designs selected for experimental study, but the identities of the amino acids vary (**Figure 4.2**). The pairwise backbone RMSDs between the designs range between 0.9 and 1.6 Å (**supplementary Figure 4.8**). To evaluate if the designed sequences were specific for their respective backbones, the sequences were threaded onto the other design models and scored after rotamer repacking and every sequence favored the backbone for which it was designed (**supplementary Figure 4.9**). For example, LoopA has a score of -152; when the LoopA sequence is threaded onto the WT, LoopB and LoopC backbones the scores are -137, -140 and -143 respectively.

Structure prediction with the designed loop sequences. During the loop design protocol small perturbations in backbone motion are used to look for local minima in structure space, but the protocol does not include any explicit tests that would probe if the design sequence prefers significantly different alternative conformations. To computationally test if our designed sequences prefer the designed target conformations we performed structure prediction with the loop sequences. In these simulations, the only input into Rosetta is the sequence of the loop and the structure of the scaffold. Structure prediction is performed using fragment based insertion with a round of low resolution scoring followed by high resolution scoring [31]. A cyclic coordinate descent algorithm is used to close loops following insertions [32]. As is the usual strategy with Rosetta, thousands of

independent structure prediction simulations were run with each sequence. For all three sequences, the lowest energy structure predictions resembled the design models (backbone RMSD < 0.8 Å) and there was an increase in energy as the RMSD values for the models increased above 1 Å (**Figure 4.3)**. Pairwise backbone RMSDs were calculated between each of the designed loops and the lowest energy structure prediction for each loop (**supplementary Figure 4.8**). There was a closer agreement between matched pairs, i.e. LoopA structure prediction compared to the LoopA design, than there was between unmatched pairs, i.e. LoopA structure prediction compared to the LoopB design model.

The lowest energy structure predicted for LoopB is similar to the design model (backbone RMSD = 0.77 Å), but there are noticeable differences in the backbone positions of residues 24-26. In the predicted structure the carbon alpha positions of these residues are shifted ~1 Å towards the N-terminal tail of tenascin, and proline 24 is better packed in the structure prediction model. The SASAprob score of proline 24 is 0.29 in the structure prediction model and 0.09 in the original design model. During structure prediction with the LoopB sequence there were models created that more closely matched the LoopB design model (backbone RMSD < 0.3 Å), but these models scored worse (**Figure 4.3)**. The fact that our design procedure did not find the lower energy conformation for the LoopB sequence indicates that in the future that it will be advantageous to perform more aggressive sampling of conformational space, including *de novo* structure prediction, when iterating between sequence design and structure optimization.

Experimental characterization. All three of the designed proteins were expressed in *E. coli.* 1-dimensional ^{1}H NMR spectra of each protein indicate that they are well folded and adopt □-sheet structures (**supplementary Figure 4.10**). Circular dichroism was used to probe the thermal stability of the proteins (**Figure 4.4).** Loop B and Loop C have midpoints of thermal unfolding (T_m) that are similar to the wild type protein, while the T_m for LoopA is 15 degrees lower than the wild type protein. The folding of all the three proteins is reversible and highly cooperative (data not shown).

Crystal trays were set up for all three designs. Crystals that diffract at high resolution were obtained for LoopA and LoopB. The crystal structure for LoopB was solved to 1.45 Å resolution with an R-factor of 17% and an Rfree of 19%. There is good agreement between the design model and the crystal structure. The design model was superimposed on the crystal structure by aligning the loop residues and residues that make contact with the loop (residues 4-8 , 20-31 , 48-55 , 72-74) (**Figure 4.5)**. Based upon this superposition, the RMSD between the backbone atoms in the designed loop and the crystal structure was 0.46 Å . The lowest energy *de novo* structure prediction for loopB had a slightly better match to the crystal structure, the RMSD was 0.42 Å . In the crystal structure the backbone atoms of residues 24-26 were located between the design model prediction and the *de novo* prediction. The B-values for the atoms in the redesigned loop are comparable with the other loops in the protein and are all below 30 $Å^2$.

All of the side chain rotamers in the design model were predicted correctly except for Gln26. In the crystal structure the side chain of Gln26 forms hydrogen bonds with the backbone nitrogen and side chain oxygen of Asp2 (**Figure 4.5**) and has low B-values (<25). We were curious if this side chain was not predicted correctly because of changes in the backbone coordinates of the residues or because of the Rosetta energy function. To test between these two options the backbone coordinates of the crystal structure were used as the template for a Rosetta side chain repacking simulation with the designed sequence. Like the design model, the repacked structure did not have hydrogen bonds between Asp2 and Gln26. The Rosetta energy function disfavors the crystal structure rotamer for several reasons. Firstly, the rotamer adopted in the crystal structure is especially rare in the Dunbrack rotamer library [33,34] and Rosetta assigns an internal energy of 3.0 kcal / mol to this rotamer. The Rosetta preferred rotamer has a rotamer score of 0.8 kcal / mol. Secondly, Rosetta only assigns weak scores to the two putative hydrogen bonds, both have scores weaker than 0.5 kcal / mol because the distances and angles between the groups are suboptimal. Favorable hydrogen bonds in Rosetta score

near 2 kcal / mol. Thirdly, there is a desolvation penalty (> 3 kcal / mol) for removing the glutamine side chain from water. It is difficult to determine which of these scores is most misrepresenting the true energetics of Gln26, and this result demonstrates the challenge of balancing nearly equal and opposite energy terms.

LoopA crystallized at low pH (3.0). The resolution of the structure was 2.1 Å with an R_{free} value of 30%. Under these conditions the structure of the protein does not look similar to the design model. In the crystal the protein adopts a 50/50 mixture of domain swapped dimers and monomers (**Figure 4.6**). In the domain swapped structure the designed loop opens up to allow strands 1 and 2 to insert into the partner molecule. The loop is stabilized by a set of hydrogen bonds involving acidic side chains that appear to be protonated. Unfortunately, there is not clear electron density for the loop residues in the monomeric chain. This suggests that at low pH the loop is not adopting a specific conformation in the monomer. To further characterize the monomer-dimer equilibrium, gel filtration experiments were performed at a variety of pHs. At neutral pH the protein has an apparent molecular weight that is close to the predicted weight for a monomer, while at low pH the apparent molecular weight is 18 kDa, halfway between a monomer and dimer. These results, combined with the decreased stability of LoopA at pH 7, suggest that the designed loop may not have a strong preference for adopting the target conformation.

None of the scores that were used to evaluate the design models suggested that LoopA would be more prone to forming a domain swap interaction than the other loop designs. Domain swapping has been observed in many proteins, including *de novo* designed proteins [35,36]. From these studies it is clear that subtle changes in environment or sequence can promote swapping. Unlike the sequences of LoopB, LoopC and the wild type protein, LoopA does not have any prolines in the redesigned loop. In the domain-swapped crystal structure of LoopA, residues 23, 27, 29 and 30 have phi angles that are incompatible with a proline. This suggests that the prolines in the other designed loops could play a

role in preventing domain-swapping by disallowing conformational changes required for opening up the loop. However, it should be noted that in a previous study we used a designed proline to favor domain swapping [37]. In summary, the LoopA results highlight the diversity of structures that a designed sequence can adopt, and provides an example of where explicit negative design would be useful if the competing states could be identified *a priori*.

CONCLUSION

Our results indicate that with the current Rosetta energy function and sampling techniques it is possible to design a 10-residue loop with high accuracy. LoopB is stabilized by tight packing interactions between hydrophobic side chains in the center of the loop. In the future, it will be interesting and important to test if novel protein loops can be designed by forming new hydrogen bonding interactions. In this study we used *de novo* structure prediction simulations to test if our designed sequences prefer the target conformations. In all three cases the lowest energy structure prediction resembled the design models, but in the case of LoopB there were a few residues that were shifted ~1 Å from the design model. In the crystal structure of LoopB, these residues adopted a position that was between the design model and the structure prediction. In future designs with more complicated target structures, it may be even more useful to evaluate designed sequences with *de novo* structure prediction simulations. These simulations can provide templates for creating even lower energy sequence-structure pairs and they can be used to determine if negative design will be needed to disfavor competing states.

MATERIALS AND METHODS

Iterative backbone and sequence optimization. Rosetta's standard full atom energy function was used for structure prediction and sequence design [6,21]. Rotamer-based sequence optimization was

performed as described previously [38]. Dunbrack's backbone dependent rotamer library was used with extra sub-rotamers created by varying all chi1 angles and the chi2 angles on aromatic residues plus or minus one standard deviation from the most preferred chi angles [33]. During sequence design all 10 residues in the designed loop were allowed to vary to any amino acid except for cysteine, and the neighboring residues were allowed to adopt alternative side chain conformations. 9 rounds of iterative sequence design and backbone optimization were used to search for low energy sequence-structure pairs.

Backbone flexibility was restricted to the loop. Backbone optimization was performed using Monte Carlo optimization. Only torsion angles are explicitly varied during the procedure. A single Monte Carlo move consisted of 1) a small change to the phi and psi angles of the loop residues (up to 5 residues are varied simultaneously), 2) a quick optimization of side chain rotamers and 3) gradient-based optimization of backbone and side chain torsion angles. After performing these 3 steps the energy of the new structure is compared to the energy of the protein before the move and the Metropolis criterion is used to decide if the move should be accepted. Two types of moves were used to create the initial perturbation to the backbone: small random changes (~1 degree perturbations) and shear moves. A shear move consists of a small change to a phi angle compensated by a change in the opposite direction to the psi angle. Fast rotamer optimization was performed by cycling over each side chain once (in random order) and choosing the lowest energy rotamer given the current environment. Gradient based minimization was performed with a conjugate gradient protocol that calculates the first derivative of the energy function for each torsion angle that is being varied [21,22]. A score that favors a low RMSD deviation between the first and last residue of the loop and the protein scaffold was used to keep the loop closed during backbone optimization.

De novo structure prediction of loop sequences. The structures of the designed sequences were predicted using a recently developed Monte Carlo-based loop modeling protocol in Rosetta [39](Chu

Wang, Phil Bradley and David Baker, personal communication). The sequence of the loop is used to pick overlapping 3 residue fragments from the PDB with similar sequences. These fragments are then randomly combined to create a starting structure for optimization. The first round of optimization is performed with a low resolution model of the protein that favors good backbone torsion angles and backbone hydrogen bonding. The second round is performed in high resolution full atom mode and combines small and shear moves (see above) with gradient based minimization and Dunbrack's cyclic coordinate descent algorithm for loop closure [32]. Rotamer repacking is also performed after every 20 backbone trials. Thousands of Monte Carlo moves are considered. For each sequence, 20,000 independent trajectories were performed.

SASApack and SASAprob scores. SASApack score was used to evaluate the quality of packing of the design models. The solvent accessible surface area (SASA) of each residue was calculated with two different probes (0.5 Å and 1.4 Å) and the difference was compared to the average difference seen in PDB for a particular residue in a similar environment [40]. The SASAprob score is the probability of observing an amino acid in a similar environment with a higher SASApack score. A SASAprob score of 0.95 indicates that a residue is better packed than 95% of similar residues in the PDB.

Protein expression and purification. Genes for the redesigned proteins were constructed with cassette mutagenesis in the pET21b expression vector. The proteins were expressed in the *E. coli.* BL21 strain at 37 °C with 0.5 mM IPTG used for induction. The proteins were purified with a Ni^{+} affinity column followed by size-exclusion chromatography (Superdex-75).

NMR. The redesigned proteins (about 1 mM concentration) were equilibrated in 20 mM sodium phosphate, 0.5 M NaCl, pH = 7.4 buffer and one-dimensional ^{1}H NMR spectra were recorded at 25 °C on a 700 MHz Varian spectrometer.

Circular dichroism. CD data were collected on a JASCO J-810/815 CD spectrometer using 0.1cm cuvetta containing 50 uM samples. The CD signal was monitored at 200 nm as a function of temperature (10 – 90 °C). The fraction of unfolded was calculated assuming that the CD signal of the unfolded and folded protein varies linearly with temperature.

Crystallization, X-ray diffraction and structure determination. The hanging-drop vapor diffusion method was used for crystallization trials of the three designed proteins at room temperature. LoopA (about 20 mg/mL in 100 mM NaCl, 20 mM Tris buffer at pH 7.4) was mixed with an equal volume of well buffer of 2.0 M Ammonium Sulfate, pH = 3.0, 10% additive 0.1 M cupric chloride (from Hampton research) was added to the drop. 20% ethylene glycol was used as the cryoprotectant. Crystals of LoopB (about 35 mg/mL in 100 mM NaCl, 20 mM Tris buffer at pH 7.4) were grown against a well buffer of 3.8M sodium formate, 5% glycerol, pH = 7.5. 20% glycerol was used as the cryoprotectant. Diffraction data of LoopA were collected at the Advanced Photon Source at Argonne National Laboratory, Beamline 22-ID (SER-CAT). Diffraction data of LoopB were collected at the Beamline x29A at Brookhaven National Laboratory.

The data were indexed and processed with the program HKL2000 [41]. Both the structures of LoopA and LoopB were solved by molecular replacement using the program AmoRe [42] and Phaser [43,44]. Wild type tenascin (PDB code 1TEN) was used as the initial search model. The models were then refined against the synchrotron data to 1.45 and 2.1 Å resolution respectively (**supplementary Figure 4.11)**. Alternating cycles of model building with the program O [45] and refinement with the programs CNS [46] and refmac [47] were used to determine the final structure. The geometry of the final model was assessed with PROCHECK [48].

Accession codes : The coordinates and structural factors of LoopA and LoopB have been deposited

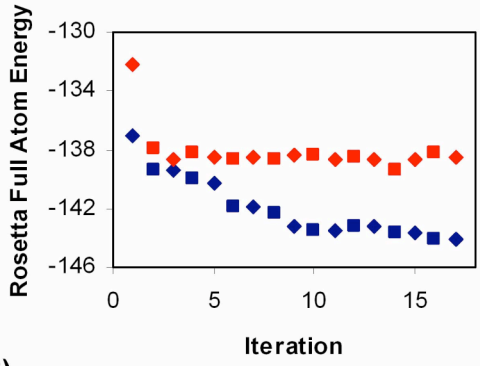
into the RCSB Protein Data Bank with PDB ID code 2RBL and 2RB8.

Acknowledgements.

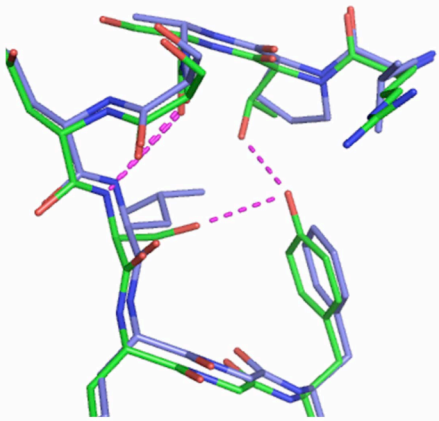
We thank Howard Robinson and Jillian Orans for collection of diffraction data from beamline X29 at NSLS and 22-ID (SER-CAT) at the Advanced Photon Source at Argonne National Laboratory. This research was supported by an award from the W.M. Keck foundation and the grant GM073960 from the National Institutes of Health.

FIGURES

A)



B)



C)

	Starting Sequence	Ending Sequence
Starting Structure	-133.1	-131.3
Ending Structure	-138.5	-139.4

Figure 4.1 Iterative optimization of a loop sequence and conformation
A) Two representative design trajectories are shown in red and blue. The diamonds indicate the energy of the protein after sequence design and the squares after optimization of backbone and side chain torsion angles. B) The starting (green) and ending (blue) models for the red trajectory (panel A) C) The starting and ending sequences scored on the starting and ending structures from the red trajectory (panel A).

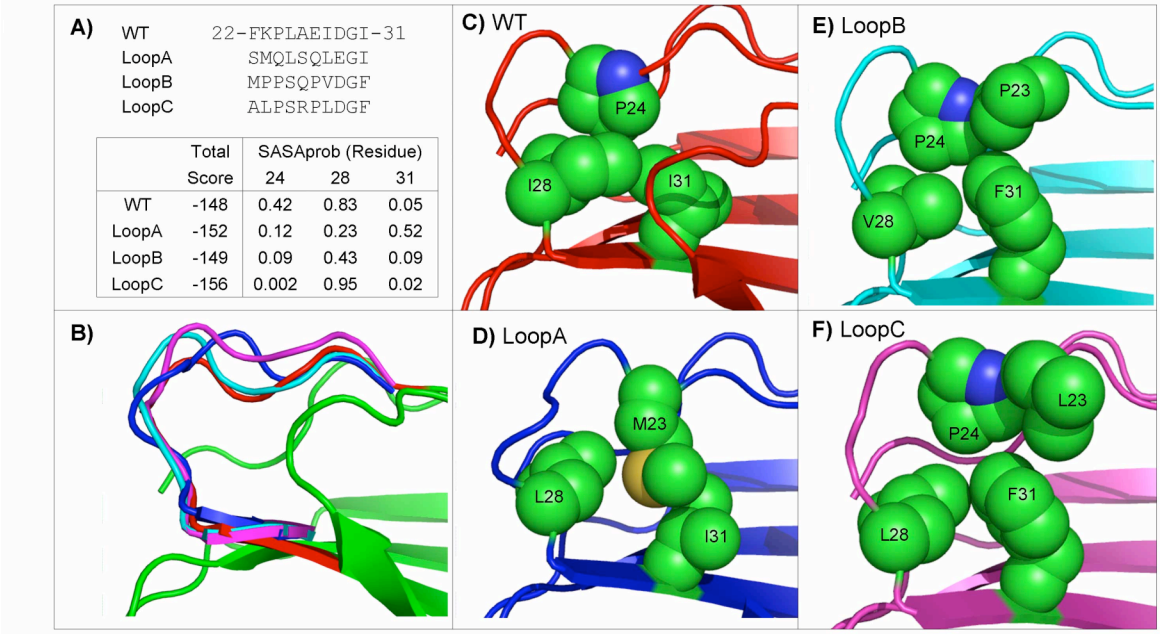


Figure 4.2 Models and sequences of the redesigned proteins
A) Sequences and scores of the redesigns. The SASAprob scores, which reflect the quality of packing for individual residues, are shown for residues buried in the center of the loop. A SASAprob score of 0.20 for a leucine indicates that 80% percent of the leucines in a similar environment in the PDB are better packed (see methods). B) Designed loops (blue – LoopA, cyan – LoopB, fuschia – LoopC) compared to the WT loop structure (red – WT). C-F) Models of the designed loops.

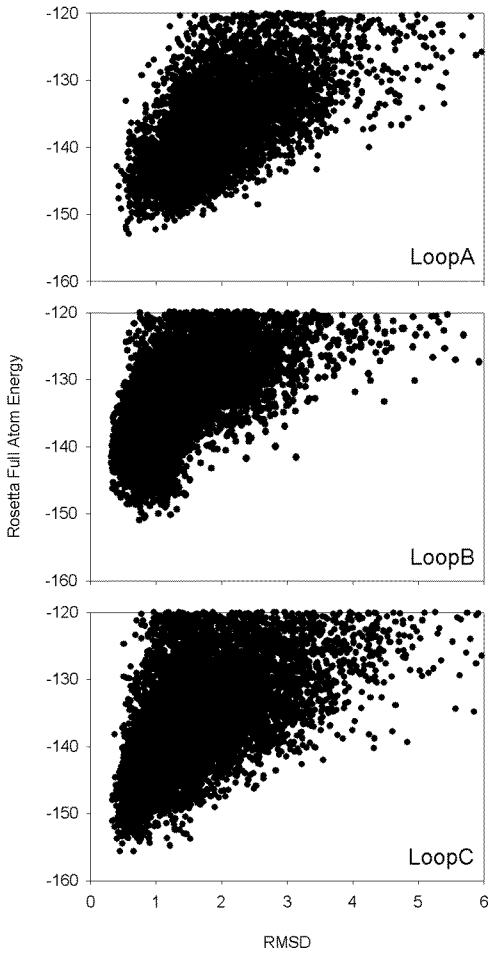


Figure 4.3 Structure prediction with the designed sequences
10,000 independent prediction trajectories were run for each design and the Rosetta energy of the models was plotted against the backbone RMSD of the predicted structures compared to the design models.

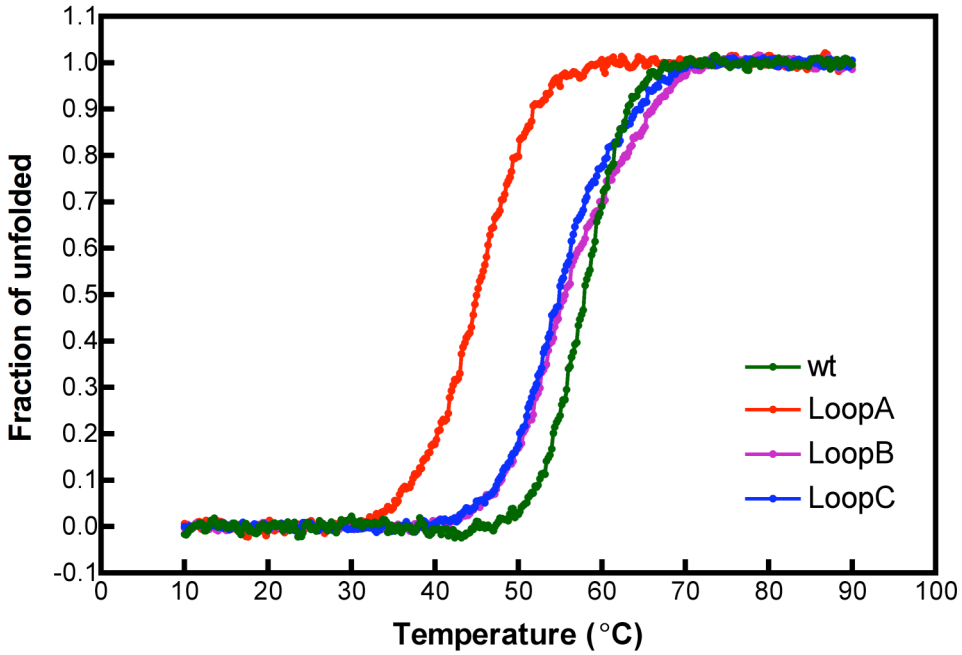


Figure 4.4 Thermal unfolding of the designed sequences as monitored with circular dichroism

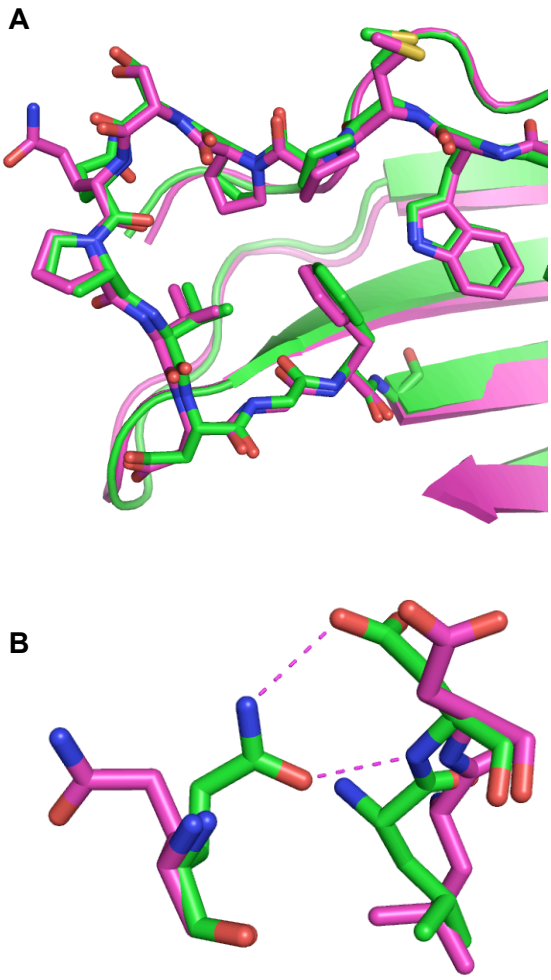


Figure 4.5 Alignment between the crystal structure and the design model
A) The crystal structure of LoopB (green) aligned with the design model of LoopB (mauve). The backbone atoms of residues 4-8, 20-31, 48-55, and 72-74 were used for the alignment. B) Close-up of glutamine 26.

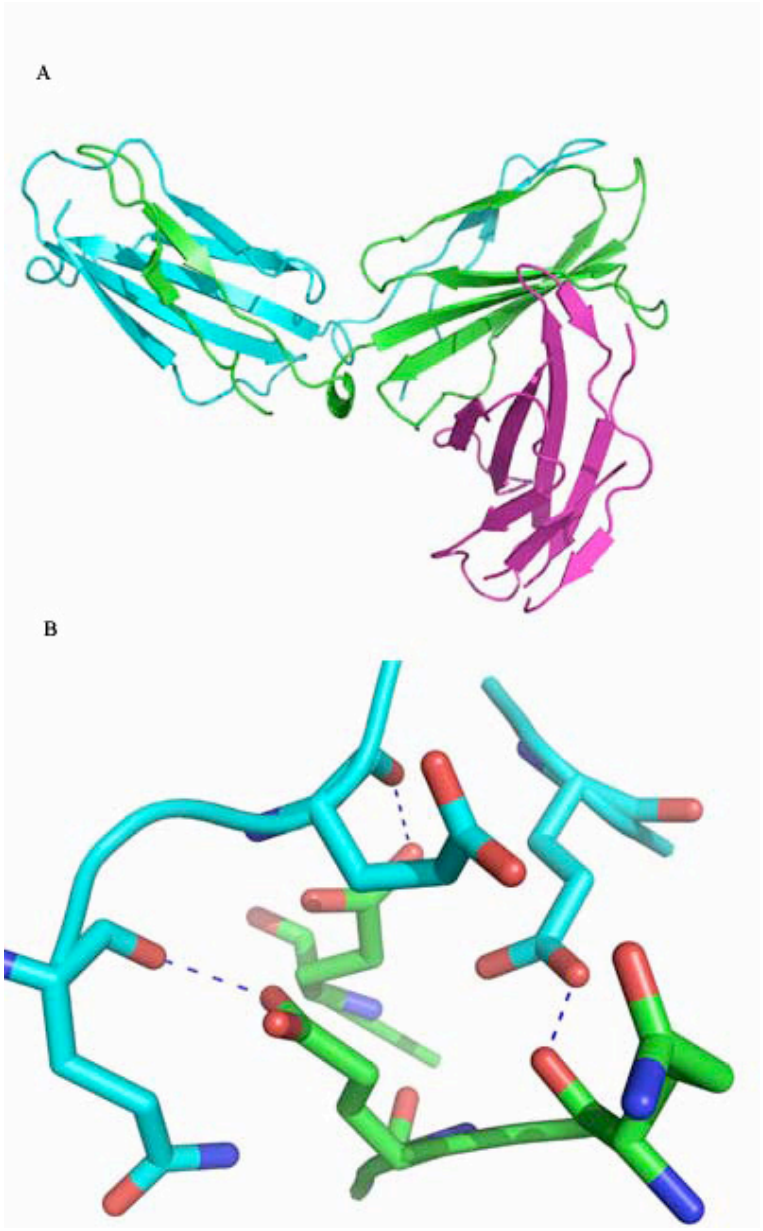


Figure 4.6 The crystal structure of LoopA
A) The crystal structure of LoopA at low pH. The repeating unit contains a domain-swap dimer (chain 1: cyan, chain 2: green) and a monomer (purple). Electron density is not present for the redesigned loop in the monomer. In the dimer the loop opens up and strands 1 and 2 insert into the partner molecule. B) The designed loop appears to be stabilized by protonated glutamic acid residues.

SUPPLEMENTARY MATERIALS

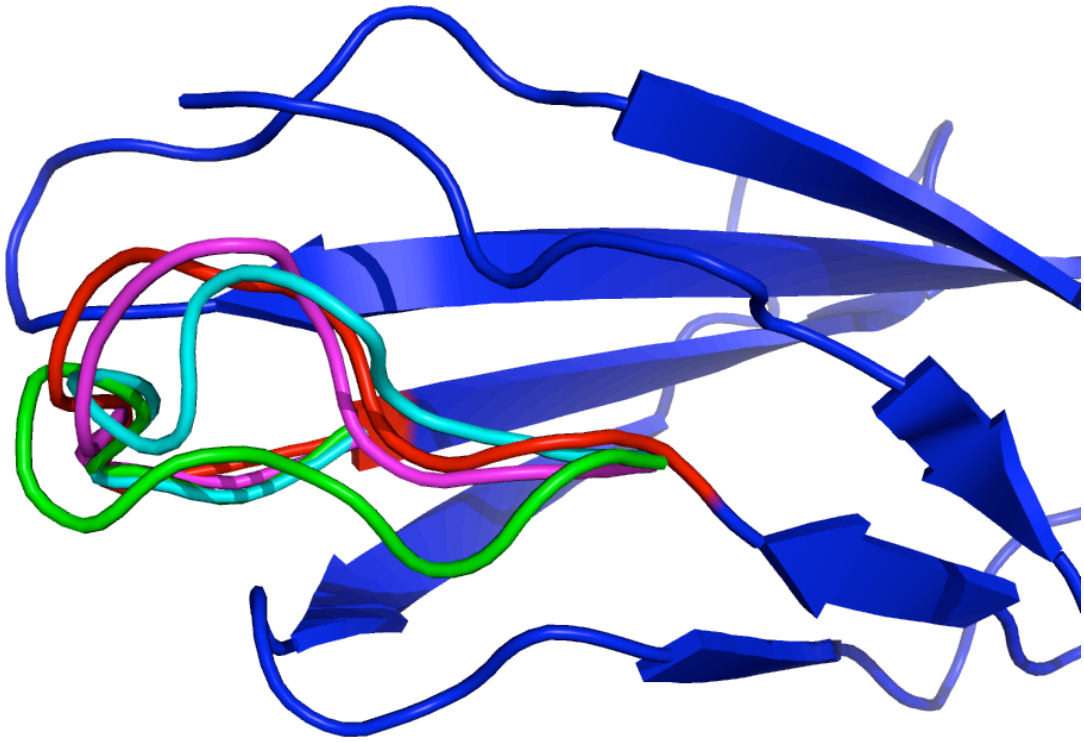


Figure 4.7 Representative set of starting structures used for loop design

	WT	LoopA	LoopB	LoopC	LoopA (sp)	LoopB (sp)	LoopC (sp)
WT	/	1.89	0.90	1.12	1.87	0.45	1.15
LoopA	1.89	/	1.49	1.58	0.58	1.82	1.50
LoopB	0.90	1.49	/	0.94	1.55	0.77	0.68
LoopC	1.12	1.58	0.94	/	1.46	1.14	0.47
LoopA(sp)	1.87	0.58	1.56	1.46	/	1.84	1.43
LoopB(sp)	0.45	1.82	0.77	1.14	1.85	/	1.10
LoopC(sp)	1.15	1.49	0.68	0.47	1.43	1.10	/

Figure 4.8 Alignment between the design models and structure predictions
Pairwise RMSDs between the design models and the lowest energy structure predictions (sp). The RMSDs are given for the backbone atoms in residues 22-33. The models were aligned using the fixed residues in the scaffold: residues 1-19 and 35-89.

	WT sequence	LoopA sequence	LoopB sequence	LoopC sequence
WT Structure	-147.8	-137.0	-144.4	-137.2
LoopA Structure	-147.1	-151.8	-136.0	-127.0
LoopB Structure	-143.6	-140.0	-148.8	-141.2
LoopC Structure	-144.2	-143.3	-147.8	-156.2

Figure 4.9 Energies of design models for different templates
Energies of design models created by threading the design sequences onto the various design templates. The side chain conformations were optimized with rotamer repacking before scoring.

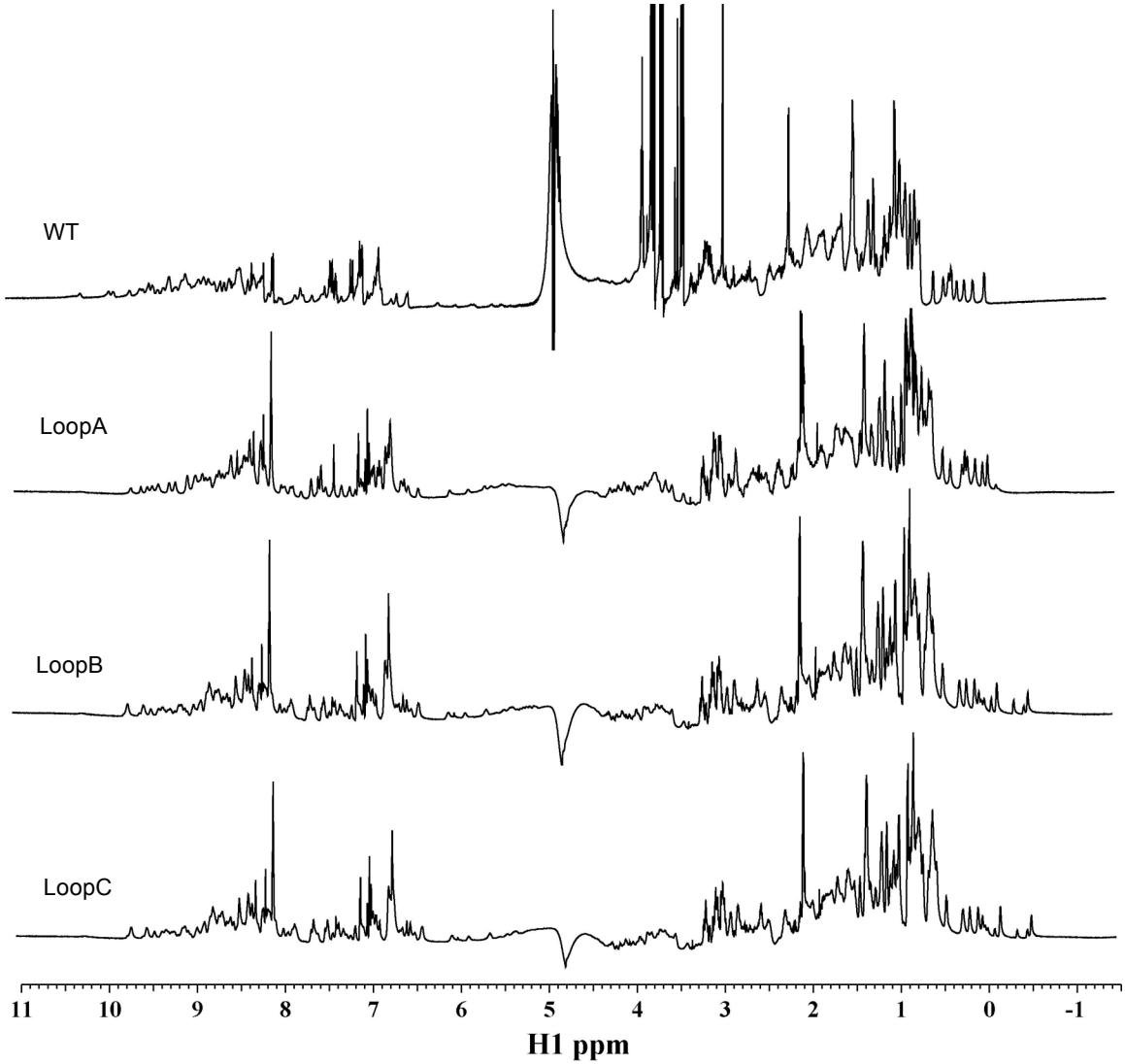


Figure 4.10 1-dimensional 1H spectra of the designed proteins

Data collection		
Crystal	LoopA	LoopB
Resolution range (Å)	49.004 – 2.101	27.185 – 1.450
Total reflections	1,366,022	573,177
Unique reflections	15,770	19,928
Completeness	86.0 % (60.3)	90% (33.9)
R_{merge}	0.054 (0.466*)	0.068 (0.191*)
Space group	H32	$P3_221$
Unit cell dimensions	a = b = 137.2 Å	a = b = 62.781 Å
	c = 86.682 Å	c = 53.793 Å

Refinement		
Crystal	LoopA	LoopB
Resolution range (Å)	70 – 2.1	30 – 1.45
R	0.25	0.17
Rfree	0.30	0.19
rmsd bond length (mc) (Å)	0.042	0.005
rmsd bond angle (mc)	3.330	1.379
No. of protein atoms	2014	724
Average B factor for all atoms ($Å^2$)	55.204	16.167

Figure 4.11 X-ray diffraction data
X-ray diffraction data collection and refinement (cryo solvent and wavelength , beam line, crystallization condition, shown in materials and methods)

REFERENCES

1. Hill RB, Raleigh DP, Lombardi A, DeGrado WF. De novo design of helical bundles as models for understanding protein folding and function. Acc Chem Res 2000;33(11):745-754.

2. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;282(5393):1462-1467.

3. Hecht MH, Richardson JS, Richardson DC, Ogden RC. De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. Science 1990;249(4971):884-891.

4. Hecht MH. De novo design of beta-sheet proteins. Proc Natl Acad Sci U S A 1994;91(19):8729-8730.

5. Kortemme T, Ramirez-Alvarado M, Serrano L. Design of a 20-amino acid, three-stranded beta-sheet protein. Science 1998;281(5374):253-256.

6. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302(5649):1364-1368.

7. Kuhlman B, O'Neill JW, Kim DE, Zhang KY, Baker D. Accurate computer-based design of a new backbone conformation in the second turn of protein L. J Mol Biol 2002;315(3):471-477.

8. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278(5335):82-87.

9. Regan L. Protein structure. Born to be beta. Curr Biol 1994;4(7):656-658.

10. Minor DL, Jr., Kim PS. Measurement of the beta-sheet-forming propensities of amino acids. Nature 1994;367(6464):660-663.

11. Munoz V, Serrano L. Helix design, prediction and stability. Curr Opin Biotechnol 1995;6(4):382-386.

12. Nagi AD, Regan L. An inverse correlation between loop length and stability in a four-helix-bundle protein. Fold Des 1997;2(1):67-75.

13. Butterfoss GL, Kuhlman B. Computer-based design of novel protein structures. Annu Rev Biophys Biomol Struct 2006;35:49-65.

14. Yin H, Slusky JS, Berger BW, Walters RS, Vilaire G, Litvinov RI, Lear JD, Caputo GA, Bennett JS, DeGrado WF. Computational design of peptides that target transmembrane helices. Science 2007;315(5820):1817-1822.

15. Looger LL, Dwyer MA, Smith JJ, Hellinga HW. Computational design of receptor and sensor proteins with novel functions. Nature 2003;423(6936):185-190.

16. Huang PS, Love JJ, Mayo SL. Adaptation of a fast Fourier transform-based docking algorithm for protein design. J Comput Chem 2005;26(12):1222-1232.

17. Desjarlais JR, Handel TM. Side-chain and backbone flexibility in protein core design. J Mol Biol 1999;290(1):305-318.

18. Fu X, Apgar JR, Keating AE. Modeling Backbone Flexibility to Achieve Sequence Diversity: The Design of Novel alpha-Helical Ligands for Bcl-x(L). J Mol Biol 2007;371(4):1099-1117.

19. Thanki N, Zeelen JP, Mathieu M, Jaenicke R, Abagyan RA, Wierenga RK, Schliebs W. Protein engineering with monomeric triosephosphate isomerase (monoTIM): the modelling and structure verification of a seven-residue loop. Protein Eng 1997;10(2):159-167.

20. Ginalski K. Comparative modeling for protein structure prediction. Curr Opin Struct Biol 2006;16(2):172-177.

21. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66-93.

22. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins 2004;55(3):656-677.

23. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 2003;331(1):281-299.

24. Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D. Progress in modeling of protein structures and interactions. Science 2005;310(5748):638-642.

25. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 2003;326(4):1239-1259.

26. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35(2):133-152.

27. Leahy DJ, Hendrickson WA, Aukhil I, Erickson HP. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. Science 1992;258(5084):987-991.

28. Carr PA, Erickson HP, Palmer AG, 3rd. Backbone dynamics of homologous fibronectin type III cell adhesion domains from fibronectin and tenascin. Structure 1997;5(7):949-959.

29. Batori V, Koide A, Koide S. Exploring the potential of the monobody scaffold: effects of loop elongation on the stability of a fibronectin type III domain. Protein Eng 2002;15(12):1015-1020.

30. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238(5):777-793.

31. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268(1):209-225.

32. Canutescu AA, Dunbrack RL, Jr. Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci 2003;12(5):963-972.

33. Dunbrack RL, Jr., Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6(8):1661-1681.

34. Bower MJ, Cohen FE, Dunbrack RL, Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267(5):1268-1282.

35. Liu Y, Eisenberg D. 3D domain swapping: as domains continue to swap. Protein Sci 2002;11(6):1285-1299.

36. Hom GK, Lassila JK, Thomas LM, Mayo SL. Dioxane contributes to the altered conformation and oligomerization state of a designed engrailed homeodomain variant. Protein Sci 2005;14(4):1115-1119.

37. Kuhlman B, O'Neill JW, Kim DE, Zhang KY, Baker D. Conversion of monomeric protein L to an obligate dimer by computational protein design. Proc Natl Acad Sci U S A 2001;98(19):10687-10691.

38. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 2000;97(19):10383-10388.

39. Wang C, Bradley P, Baker D. Protein-protein docking with backbone flexibility. J Mol Biol 2007;373(2):503-519.

40. Sood VD, Baker D. Recapitulation and design of protein binding peptide structures and sequences. J Mol Biol 2006;357(3):917-927.

41. Otwinowski Z, Minor W. Processing of X-ray Diffraction Data Collected in Oscillation Mode. Methods in Enzymology 1997;276:307-326.

42. Navaza J. AMoRe: an Automated Package for Molecular Replacement. Acta Crystallographica 1994;A50:157-163.

43. McCoy AJ, Grosse-Kunstleve RW, Storoni LC, Read RJ. Likelihood-enhanced fast translation functions. Acta Crystallogr D Biol Crystallogr 2005;61(Pt 4):458-464.

44. Storoni LC, McCoy AJ, Read RJ. Likelihood-enhanced fast rotation functions. Acta Crystallogr D Biol Crystallogr 2004;60(Pt 3):432-438.

45. Jones TA, Zou JY, Cowan SW, Kjeldgaard M. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Crystallogr A 1991;47 (Pt 2):110-119.

46. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 1998;54(Pt 5):905-921.

47. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 1997;53(Pt 3):240-255.

48. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 1993;26:283-291.

CHAPTER 5

DE NOVO DESIGN

ABSTRACT

De novo design is a rigorous test of our understanding of the protein folding process and is a very challenging problem. Despite many successes to date, *de novo* design of globular β-sheet proteins remains an unsolved problem. To improve our understanding of β-sheet proteins, we have used Rosetta to try to design β-sandwich proteins with Greek Key topology using only positive design. The designed proteins expressed at the anticipated size, and circular dichroism experiments indicated the designed proteins either aggregated with β-sheet content or are unfolded. Even after several generations of design, these proteins displayed low solubility. These results suggest that in the future, it will be worthwhile to incorporate negative design techniques to create well-folded β-sandwich proteins.

INTRODUCTION

De novo protein design has been a very attractive approach to test our understanding of the principles of the protein folding process[1,2]. The goal of *de novo* design is to predict a sequence that is not necessarily related to any naturally occurring protein sequence, and which will fold into a predefined three-dimensional structure. It is very challenging and computationally expensive because *de novo* design involves the construction and optimization of a new backbone template[3]. In the past few years there has been great progress in *de novo* design and a number of strategies have been developed for this approach[4-10]. Hecht *et al.* designed a four-helix bundle protein that seemed to be monomeric with helical secondary structure using a binary library approach[9]. The Degrado laboratory tried to design a four-helix bundle by incorporating some hydrophilic residues at buried positions, and they were able to convert the design from molten globule to a specific conformation. However, the high resolution structure shows that the protein does not adopt the target fold[11,12]. Harbury *et al.* created a four-helix coiled-coil with a novel super-helical twist[7]. Their approach was to search for a sequence that maximized the energy gap between a target structure and alternative structures. Another breakthrough result was the creation of a novel fold in the protein TOP7 by Kuhlman *et al.*[4] They used Rosetta to create a family of starting structures, which were optimized through iterations between sequence design and backbone minimization. This approach did not explicitly include negative design. The crystal structure of the design matches the design model very closely(RMSD 1.2 Å), and it is very stable (T_m > 100 °C).

Compared with α-helix design, β-sheet protein design has been less successful[2]. Since 1981, the Richardson group has been trying to design a series of β-sheet, bell-shaped proteins termed betabellins. For several generations, solubility was a big issue with these designs[13-15]. The most successful designs from their trials are betabellin 14D and betadoublet. They used an alternating hydrophobic/polar pattern for the primary sequence and a disulfide bond was used to connect two

identical subunits. Both proteins were water soluble at low pH; however NMR data suggested that they did not adopt a single unique conformation and are most likely molten globules. Kortemme *et al.* designed a 20-residue three-stranded β-sheet protein without disulfide bonds[16]. In this work they incorporated an aromatic cluster along one β-sheet which contributed to the stability. Sollazzo *et al.* used a portion of the immunoglobulin V_H domain to design a small protein that can bind metal upon folding[17]. Its solubility was poor (10 μM) which limited detailed structural analysis. Nanda *et al.* designed a mimic of the redox protein rubredoxin, which is one example of a functional *de novo* designed β-sheet protein[18].

These attempts at *de novo* design of β-sheet proteins are very encouraging; however, no design has yet been validated by a high resolution structure. The *de novo* design of modular β-sheet proteins is still very challenging. As we might expect, many of the *de novo* designed β-sheet proteins have very low stability and solubility, suggesting the need to incorporate negative design elements to design against unwanted misfolded and aggregated states. Negative design means trying to design a sequence that is unfavorable in alternative conformations.

To test if negative design is necessary for *de novo* design, we decided to design a β-sandwich protein from scratch using only positive design (search for a sequence that has low energy in the target conformation). Kuhlman *et al*. used this approach to design TOP7 successfully and we were curious to see whether this same approach could be used to design a β-sheet protein. To increase the chance of success, we chose a target fold found in nature, the fibronectin type III (FNIII) domain, an immunoglobulin-like (Ig-like) β-sandwich protein with a Greek Key topology. The Ig-like fold is one of the most common structures in the protein database. This domain is found in proteins of diverse functions with low sequence identity but with very similar three dimensional structures[19,20], suggesting that this fold is highly designable. It has three strands A, B and E in one sheet, and four

strands C, D, F and G in the other sheet, forming two interacting antiparallel β-sheets (**Figure 5.1**). Studies have shown that this FNIII domain is a very good structural scaffold for protein design because the residues contributing to stability are mostly in the core and those residues involved in function are mostly located in the loop regions[19-21]. The separation of function from folding allows the loops to be reengineered for different functional requirements. Because the loops are very tolerant to mutations, this fold has been proven to be a very good scaffold for monobody design[22]. Monobodies, small proteins that are considered antibody mimics, are easy to express and purify. This seems to be an ideal target fold for our *de novo* design and potentially could be useful for future molecular recognition applications.

MATERIALS AND METHODS

Computational design approach

The protocol for *de novo* design is shown in **Figure 5.2**, it mainly has four steps: 1) generating starting structures; 2) high resolution refinement of backbone conformation and sequence design; 3) rebuilding and refinement; 4) model selection for experimental validation.

Step 1: Generating starting structures

Our goal is to create a stable three-dimensional all β-sheet protein based on a specific design model. The first step for *de novo* design is to define the target structure. **Figure 5.3 A** shows the schematic representation of the target fold and the constraints used to specify the topology. A secondary structure was assigned to each sequence position and some short-range distance constraints were used to guide the assembly process; for instance, distance constraints between backbone nitrogens and carbonyl oxygens for stand pairing as shown in **Figure 5.3 A**. To create starting structures, a library

of three and nine residue fragments with the desired secondary structure was generated from the PDB[23]. A Monte Carlo optimization strategy was used to assemble these fragments to yield native-like models that maximize hydrophobic burial and hydrogen bonding (**Figure 5.3 B**)[4,23]. The advantage of Rosetta's proven fragment assembly strategy is that most of the local interactions are favorable because the backbone is built from small pieces of naturally occurring proteins. Structures satisfying the strand pairing constraints were minimized and passed to the next step: sequence design.

Step 2: High resolution refinement of backbone conformation and sequence design

The starting structures were generated without consideration of side chain packing, which means the starting structures might not have been designable (i.e. no low free energy sequence is available for the target structure). As we have seen in previous *de novo* design experiments using fixed-backbone sequence optimization to search for low energy sequences, no sequences for these backbone structures existed with energies comparable to energies of naturally occurring proteins[4]. To improve the models' energies, we used a protocol which couples sequence design and backbone movement by iterating between the sequence optimization for a fixed backbone and optimization of the backbone for a fixed sequence. The energies were improved dramatically with this iterative optimization procedure and were comparable to values we observed for fixed-backbone redesign of naturally occurring proteins.

The first round of sequence design uses a standard Metropolis Monte Carlo optimization procedure. The starting conformation is perturbed by a single rotamer substitution. If the substitution lowers the energy, it is accepted. If the energy is higher, the substitution is accepted with a probability.

The second round of backbone optimization protocol uses Monte Carlo optimization with gradient-based minimization. A variety of random changes to the backbone conformation are allowed[23]. One

type of change is a small perturbation of selected phi or psi angles. Another change involves the random perturbation of a selected phi angle and a compensating opposite rotation of the preceding psi angle. These two types of movements produce subtle local perturbation. A third random change is the insertion of a new fragment (using fragment libraries as before) along with variation of nearby backbone torsion angles to accommodate the change. After minimizing perturbation of a subset of phi and psi angles, the side chain torsion angles are rapidly optimized. The conformation produced by these changes is conjugate-gradient minimized to a local energy minimum and the perturbation is accepted or rejected based on the Metropolis criterion. Several thousand Monte Carlo moves followed by minimization are used for a round of backbone refinement.

Following backbone refinement, the sequence will be redesigned based on the new backbone conformation. Typically after 10 iterations of backbone optimization and sequence design, the sequence and energies do not change significantly. This protocol was used successfully to design a novel globular protein TOP7 [4]. In this work, the protein backbone did not move dramatically during this optimization procedure, but the subtle movement of the backbone usually lowered the energy of the structure significantly. For each starting structure, the protocol was used to obtain low energy structure-sequence pairs. To restrict our search to more soluble sequences, we allowed all amino acids except cysteine for all the loop positions and buried positions in the strands but restricted the surface positions in the strands to be polar amino acids. The reference energies were recalibrated to best reproduce the native sequences for a set of high resolution β-sheet proteins.

Step 3: Rebuilding and refinement

Step 2 allows us to design sequences with a flexible backbone, however, the backbone does not move dramatically during the optimization procedure. To introduce more backbone diversities so as to increase the conformational sampling, a newly developed protocol has been used to refine the

structures[24]. There are three steps in this protocol. First, identify the highly structurally variable regions of the protein, often mainly the loop regions, and rebuild them. Fresh coordinates for these regions were generated using the fragment insertion protocol. Cyclic coordinate descent[25] (CCD) was used to maintain the connectivity of the design model and proper closure of rebuilt sequences. This rebuilding protocol allows for the sampling a broad range of conformational spaces. The second step, full atom refinement, searches for local energy minima for the structures produced by the rebuilding process. The refinement has four steps: 1) random perturbation of backbone torsion angles; (2) rapid rotamer optimization; (3) gradient-based minimization to the local energy minimum; (4) evaluation of the Metropolis criterion. The refinement step ensures sampling of the lowest energy regions of the energy landscape.

Step 4: Model selection for experimental validation

Designed structures were selected for experimental validation with several filtering criteria: the number of unsatisfied hydrogen bonds, Rosetta energy and the quality of packing in the protein. Studies have shown that unsatisfied backbone polar groups are energetically quite expensive and unlikely[26], so the backbone polar groups should be involved in hydrogen bonds or solvent exposed. Unlike in fixed backbone design, structures from *de novo* design are more likely to have problems, particularly in hydrogen bonding and tight packing. The number of unsatisfied hydrogen bonds is the most stringent filter, so the first round of selection was made by filtering out those structures that have more unsatisfied hydrogen bonds than those seen in naturally occurring β-sheet proteins. The other two criteria were then applied to get the final designs. The final sequences were submitted to the NCI database with PSI-BLAST to confirm that the sequences were not related to any naturally occurring protein sequences[27].

First generation of design trials

We have tried to design β-sandwich proteins from scratch over four different generations, always using the fold topology as shown in **Figure 5.1**. For the first generation of designs, the short loops (AB, CD and FG) were all constructed with 2-residue hairpins that presumably would help to induce structure formation. The longer loops (BC, DE and EF) varied in size from 5 to 9 residues. Starting structures were assembled using small fragments from naturally occurring proteins, using distance constraints to force the desired strand pairing (see the Materials and Methods section for more details). From several thousand independent trajectories 76 starting backbone models were selected. Each starting structure was relaxed and followed by flexible sequence design. We did 10 independent trajectories for each relaxed structure and finally selected four design models (B001 ~ B004) for experimental characterization.

Second generation of design trials

Compared with the first generation designs, these designs were constructed with longer loops. We substituted two-residue hairpins (AB, CD) with longer loops. Earlier work demonstrated that conformational sampling is very important, so for this generation, we generated more decoys and selected 1303 starting backbone structures. More importantly, we also added one more step in the design simulation, which is to cut out the loop regions and rebuild them followed by high resolution refinement. Multiple iterations have been performed between sequence design and backbone optimization. This dramatically increased the conformational sampling and improved the energies. The Rosetta energies per residue for these designs became comparable to those of the naturally occurring β-sheet proteins.

Third generation of design trials

We constructed the models for the first and second generations based on some information from naturally occurring β-sheet proteins: the length of β-strands and the length of the loops. However, the models were still arbitrarily constructed and it is possible that the target structure is not designable(no low free energy sequence for that target fold). Studies have shown that loops and hairpins are very critical to stability and folding kinetics[28]. We noticed that naturally occurring β-sandwich proteins usually have long flexible loops instead of short hairpins, and decided to use a wild type β-sandwich protein, the tenth FNIII domain from human fibronectin[21](PDB code 1FNA), as a template to construct the fragments and constraint file. To avoid ending up with very native-like sequences, we excluded homologies in the fragment files. All the loops were quite long in this set of designs. **Figure 5.4** shows one example of a starting backbone model(long loops for AB, CD and FG). We generated 260 starting structures that closely resembled the final target fold; the RMSD to the template varied from 2 to 5 Å. Nine models (F1 ~ F9) were finally selected for experimental validation; their sequence identity to the template was about 20 percent.

Fourth generation of design trials

Rosetta uses an orientation-dependent hydrogen bonding term, derived from the distribution of three parameters (distance between the hydrogen and acceptor atoms δ, angle at the hydrogen atom θ and angle at the acceptor atom ψ, **Figure 5.4**) from the PDB database[29]. However, there is another important angle, X, which is the torsion angle given by rotation around the acceptor-acceptor base bond for sp^2 hybridized acceptors (**Figure 5.5**). This angle is missing from the current energy function. We calculated the X angle for naturally occurring β-sheet proteins and Rosetta generated decoys and found that the patterns are clearly very different (**Figure 5.6 A and Figure 5.6 B**). In nature, the X angle favors 0 degrees, however, Rosetta generated decoys do not. Our collaborators Jack Snoeyink and Ning Jin explicitly incorporated this X term into Rosetta energy function (**Figure**

5.6.C), and we used those starting structures from the second generation to do another round of sequence design with the modified energy function. Four models were selected for experimental validation (BN1 ~ BN4). The sequences for all the designs are listed in the **supplementary Table 5.2**.

EXPERIMENTAL PROCEDURES

Protein expression, purification and biophysical characterization

Genes for the designed proteins were either bought from a commercial gene synthesis corporation (GenScript) or synthesized in-house with PCR extension of many overlapping oligonucleotides bought from Operon[30]. Appropriate restriction enzymes were used to digest and subclone the sequences into the *E. coli* expression vector pET21b (Novagen) or pGex4T (Amersham) as fusions with either a 6x histidine tag or GST tag. Individual clones were screened for inserts and verified by DNA sequencing. The proteins were expressed in *E. coli* BL21 strain and were purified using affinity chromatography followed by ion exchange and gel filtration chromatography (Superdex-75). For those proteins that expressed in inclusion bodies, 6 M GuHCl was used to solubilize the pellet and then the soluble fraction was loaded onto an affinity column for further purification. GuHCl was then gradually removed by dialysis. The gel filtration was used to indicate if the protein was monomeric, oligomeric or aggregated. Analytical ultracentrifugation was used to confirm the oligomeric status if necessary. Circular dichroism (CD) was used to probe the secondary structure of the protein and measure whether it unfolds cooperatively with temperature or chemical denaturants. 1D-NMR experiments were used to confirm if the proteins were well folded. For those designs that appeared to be stable and well folded, we used X-ray or NMR to solve the structures.

RESULTS

Experimental Characterization

First generation

All four designed proteins (B001~B004) were expressed in *E. coli* and purified with nickel column affinity chromatography, concentrated and run on a size exclusion column. In all cases the proteins displayed apparent molecular weights in their gel filtration profiles, which were significantly higher than the expected weight of a monomeric protein. To determine the secondary structures of the designed proteins, CD spectra were recorded on Pistar-180 spectrometer. Spectra from 250 to 190 nm were collected using 0.1 cm cuvette containing 50 μM samples. As shown in **Figure 5.7 A**, the protein B002 has a single trough around 215 nm, which is characteristic of β-sheet proteins. Thermal denaturation was monitored at four different temperatures (25 °C, 50°C, 80°C and 95°C). The observed behavior suggested that the folding and unfolding under these circumstances were reversible. Although the CD spectra showed that the protein B002 were largely composed of β-sheet secondary structure, it was actually aggregated as evidenced by 1D NMR spectrum. 1D NMR data were collected on a 600 MHz Inova spectrometer and the spectrum showed very broad lines suggesting that the protein was poorly folded (**Figure 5.7 B**). From these data it was clear that the *de novo* designs were aggregating.

Second generation

Genes for five selected designs (B5~B9) were synthesized in-house and the genes were inserted into *E. coli* expression vector pGex4T1 with a GST tag. The proteins were purified with a GST affinity column, thrombin cleavage, then ion exchange and gel filtration columns. Gel filtration profiles of

the first three suggested that they were aggregating. B9 showed an apparent molecular weight of 24 kDa. To test secondary structure, CD experiments were performed on a Jasco J-810/815 CD spectrometer with 1mm path length cuvette. As shown in **Figure 5.8 A**, B9 showed a strong minimum at 190 nm suggesting the protein was unfolded. In the design model, there is a buried hydrogen bond between Tyr 17 and Ser 66 (**Figure 5.8 B**). We decided to mutate these two polar residues to hydrophobic residues and Rosetta picked Phe at position 17 and Ala at position 66 (**Figure 5.8 C**). We made these two point mutations, but the mutant showed very similar behavior to B9. Both the CD data and 1D NMR spectrum (**Figure 5.8 D**) showed the mutant was still unfolded.

Third generation

Nine designs (F1~F9) models were selected from this generation of designs. They were cloned into pet21b expression vector with a 6x histidine tag on the C-terminus. The proteins were expressed in *E. coli*, largely in the inclusion bodies. Gel filtration studies showed that all these designs aggregated. This aggregation is probably related to the sequence composition of these designs: as shown in **Table 5.1 C**, there are many more phenylalanines and prolines compared to the naturally occurring proteins.

Fourth generation

All the design models from this generation appeared to aggregate except BN1, which showed an apparent molecular weight of 20 kDa in the gel filtration profile. However, CD data showed that BN1 was unfolded (**Figure 5.9 A**). To help stabilize the protein, we introduced a disulfide bond between two strands from the two different sheets into the design model (**Figure 5.9 B**). The mutant, named ds2, showed behavior similar to BN1, and was also unfolded as evidenced by CD (data not shown).

DISCUSSION

In summary, we were able to express proteins of anticipated size with β-sheet tendencies, but they were either unfolded or aggregated. It is still unclear why the designed proteins are so prone to aggregation. Unfolded proteins may be associating with each other because some regions are intrinsically prone to aggregation. The protein may form a domain swap or it may not adopt a near native conformation. Nevertheless, from our design trials we have learned some important lessons that should deliver improvements in the future.

The Richardons' betabellins and betadoublet, as well as several of our design trials, all displayed limited solubility. An appropriate hydrophobic/hydrophilic residue composition is very important to a protein's stability. Several trends are observed by comparing the amino acid compositions of our designed models with naturally occurring proteins. If a designed protein has a relatively large number of hydrophobic residues, it is very likely to precipitate, as in our third generation of designs. All of the designed proteins display a relatively high percentage of hydrophobic residues, especially phenylalanine and proline, and none of these designs are soluble. However, if there are too few hydrophobic residues, it is hard to induce a self-associated folding process. Our design models B9 and BN1 both show a relatively low hydrophobic/polar ratio and both appear to be soluble but unfolded(**Table 5.1 E**). Thus, an optimal number of hydrophobic residues seems to be critical.

Hydrophobic interactions clearly increase stability, but to generate a well-folded protein, packing must be very compact. We noticed that overall the packing in our design models is not optimal. Rosetta has recently gained a tool named "packstat" which evaluates how well the protein is packed. The score increases from 0 to 1 as packing improves. We constructed a test set containing 112 naturally occurring β-sheet proteins from the PDB with high resolution and calculated their packstat scores. We found that the designed proteins have a lower packstat score than wild type proteins,

indicating poorer packing. **Figure 5.10** shows the histogram of the residue distribution of packstat scores in different environments for the design models (**Figure 5.10 A and Figure 5.10 C**) and naturally occurring β-sheet proteins (**Figure 5.10 B and Figure 5.10 D**). It is very clear that the naturally occurring proteins are much better packed than our design models, especially for the buried positions (**Figure 5.10 C and Figure 5.10 D**). Here we define buried residues as those with more than 19 neighbors (C_β distance within distance of 10 Å) and surface residues as those with less than 13. In the naturally occurring β-sheet proteins, the packstat scores are clustered near 0.9. The distribution of packstat scores for designed models is somewhat noisy, but we can still see the average packstat score is around 0.6, which is quite a bit lower than the wild type proteins. This result was quite shocking, and we were curious if this poor packing in the design models is due to the energy function or the sampling problem. Does this poor packing only happen in *de novo* design or is it a general problem in Rosetta? Given a designable backbone, will Rosetta do a better job in packing residues tightly? We took the set of native proteins and did fixed backbone design with different options and still the designed models did not pack as well as wild type proteins. This indicates that even if the backbone conformation is designable, Rosetta still has some problems in tightly packing the side chains. It remains to be seen if more conformational sampling is need or if the energy function needs to be changed to improve packing.

Another observation we have found is that in our design models, usually the number of positively and negatively charged residues is about equal, whereas in the naturally occurring proteins, the net charge is usually higher (**Table 5.1 E**). We believe that increasing the overall net charge of the protein can help to improve solubility because the charge repulsion may help to prevent precipitation and aggregation[31].

The successful de novo design of a β-sandwich protein still remains a grand challenge for the protein design field. Most β-sheet proteins designed from scratch aggregate. However, in nature, almost a quarter of all proteins are β-sheet and they are well-folded. Nature must have encoded some negative design elements to prevent misfolding and aggregation; for example, studies have shown that there is one proline that is highly conserved in the FNIII domains, but it is conserved for preventing aggregation, not for stability or function[32]. There are also some other structural features proposed by the Richardson laboratory, like short edge strands and β-bulges, that may contribute to preventing aggregation[32,33]. In our future design trials, we intend to include these negative design elements to our design simulation.

FIGURES

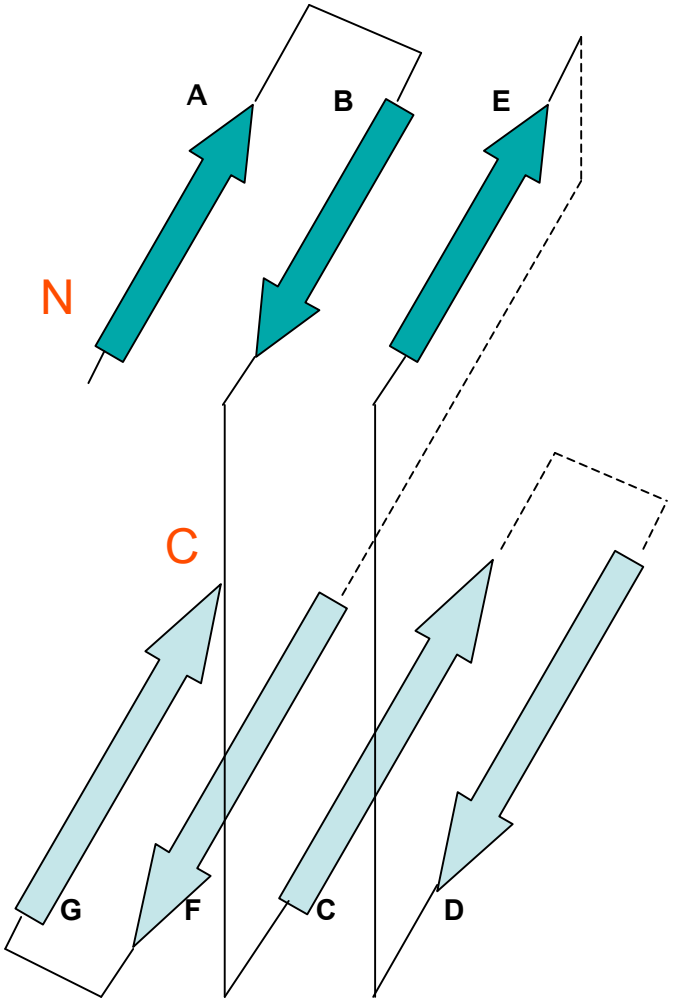


Figure 5.1 Target FNIII fold
Three strands A, B and E form the top sheet and strands C, D, F and G form the bottom sheet. Black lines connecting the strands represent the loop regions, the dotted lines represent the loops in the back of the bottom sheet.

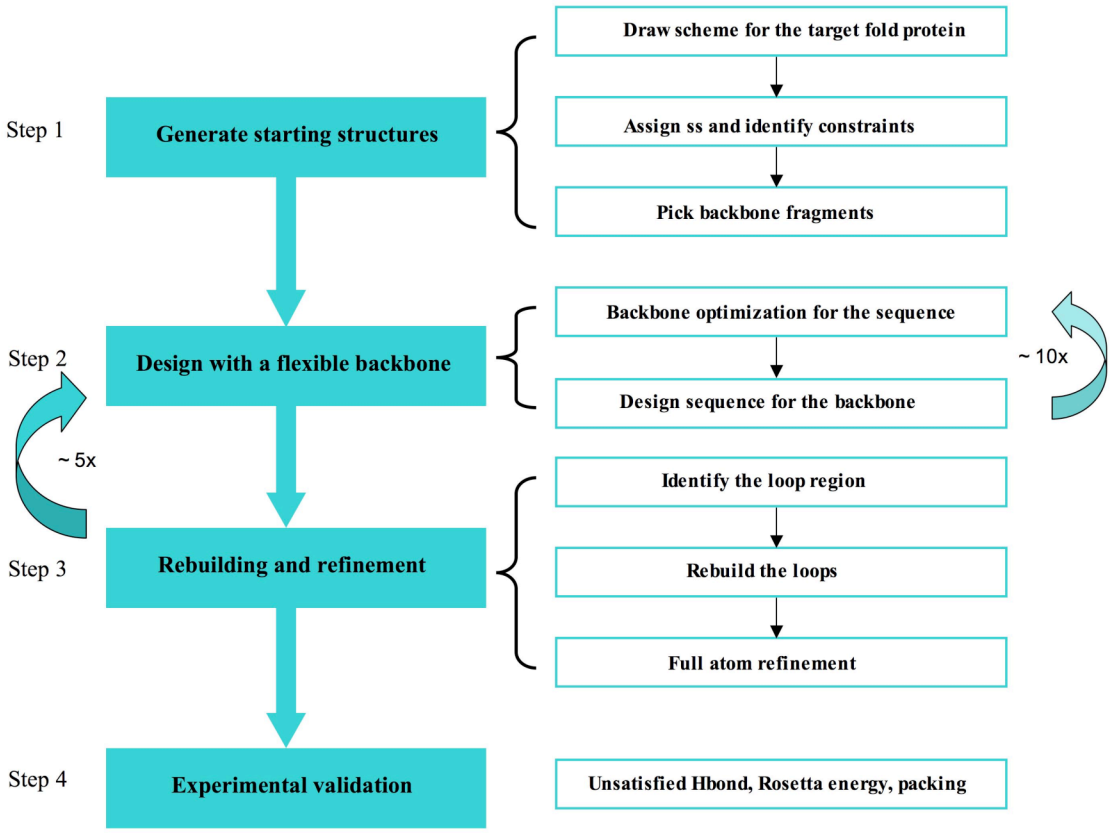
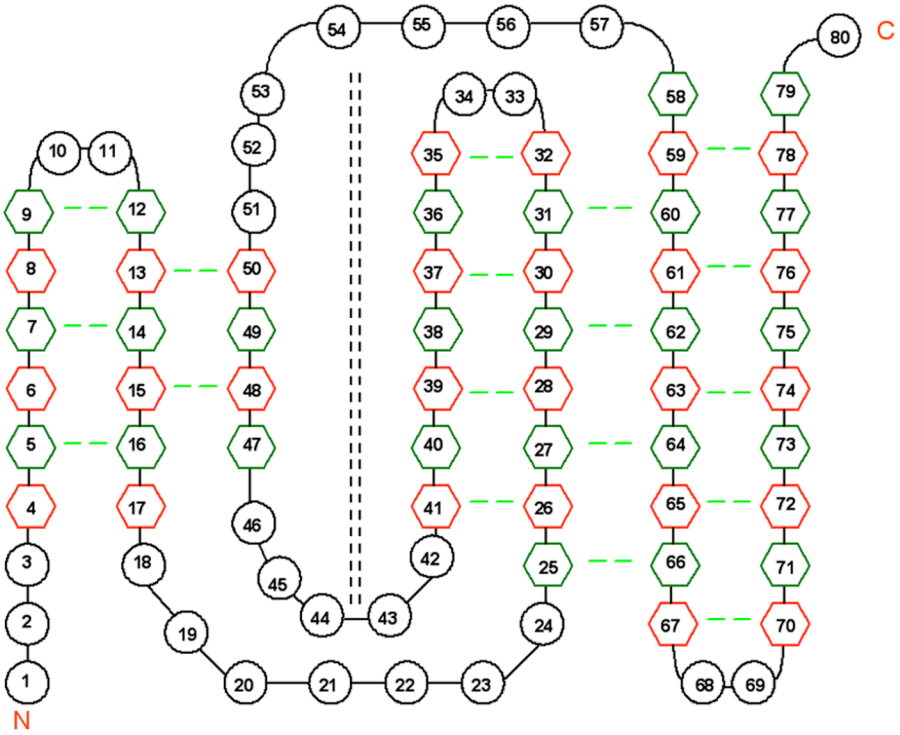


Figure 5.2 Protocol for *de novo* design

A



B

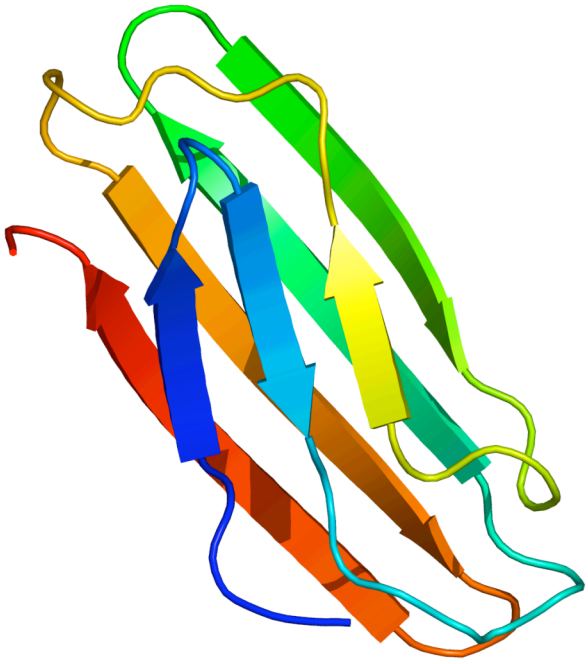


Figure 5.3 Schematic representation of the target fold
(A) Schematic of the target FNIII fold (top panel, hexagon represents β-strand, circular represents loop, green dashed line represents the distance constraint for hydrogen bonding); (B) An example of one of the starting structures built with Rosetta from the first generation

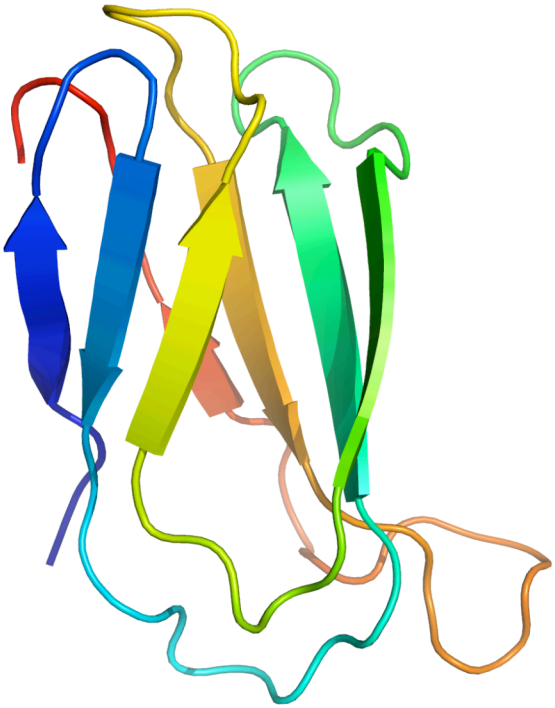


Figure 5.4 One example of the starting structures

One example of the starting structures built with Rosetta based on the wild type FN3 template (from the third generation).

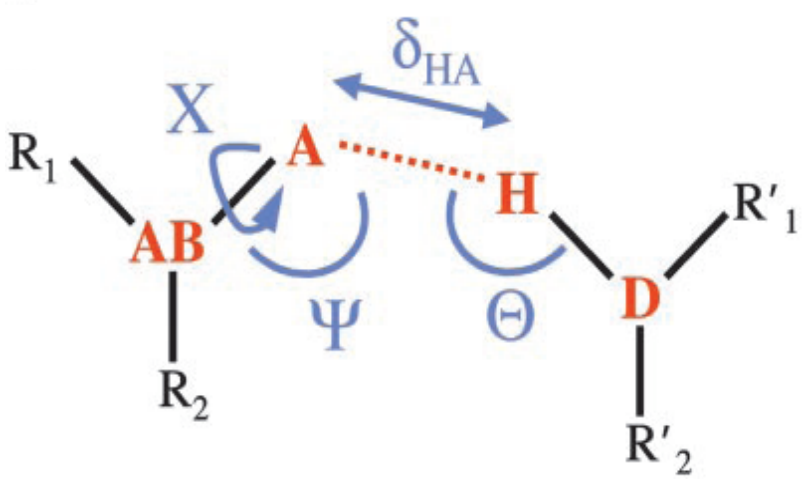
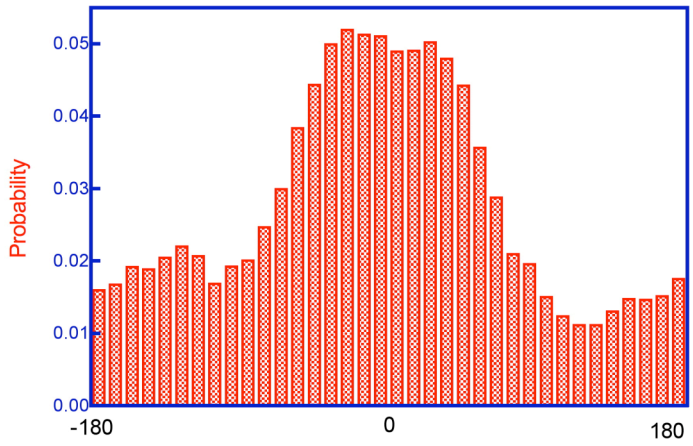
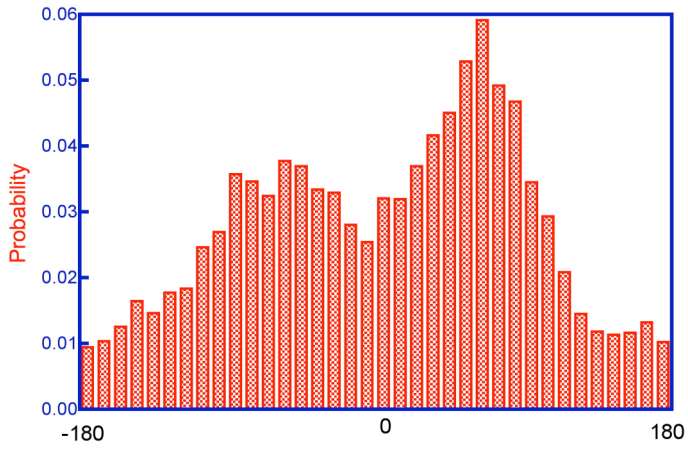


Figure 5.5 Schematic representation of the parameters used in Rosetta for hydrogen bonding potential
X is the dihedral angle between atoms R_1, AB, A and H. R_1: atom bound to the acceptor base; A: acceptor; D: donor; H: hydrogen [34]

A



B



C

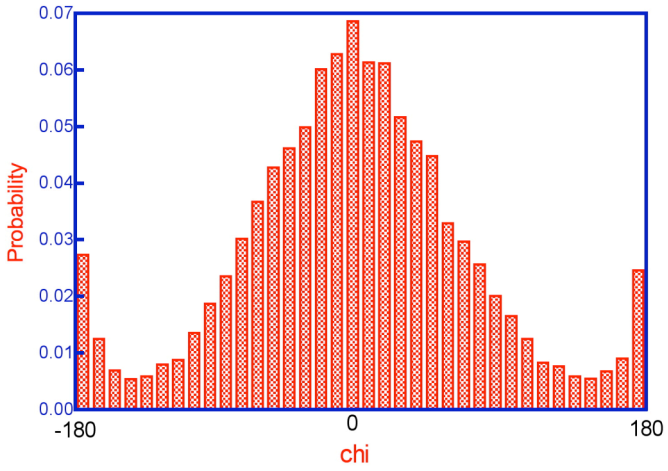
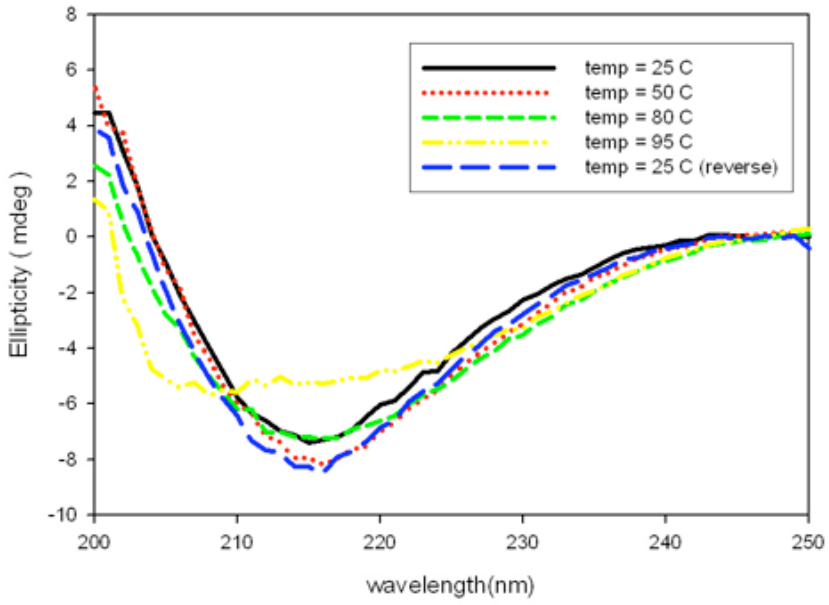


Figure 5.6 X angle distribution
(A) X angle distribution of naturally occurring β-sheet proteins. (B) X angle distribution of Rosetta generated decoys with the current version of Rosetta. (C) X angle distribution of Rosetta generated decoys after explicit incorporation of X term in the energy function

A



B

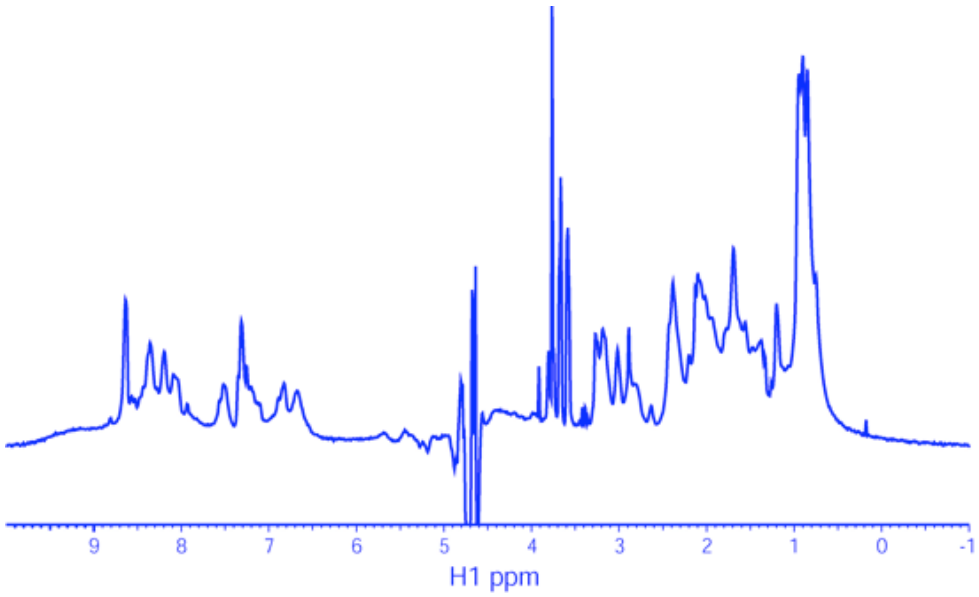


Figure 5.7 Spectra of B002
(A) Circular Dichroism spectra for protein B002 at different temperatures(Buffer condition: 50 mM NaPi, 150 mM NaCl, pH 3.5). A single minimum at 216 nm indicates that the protein was largely composed of β sheet secondary structure. The temperature was varied from 25 °C (black), 50 °C (red), 80 °C (green), 95 °C (yellow) and then cooled down back to 25 °C (blue). (B) 1D H^1 NMR spectrum of Protein B002 in pH 3.9, 150mM NaPi, 15mM Tris buffer at 37 °C on the 600MHz Varian spectrometer.

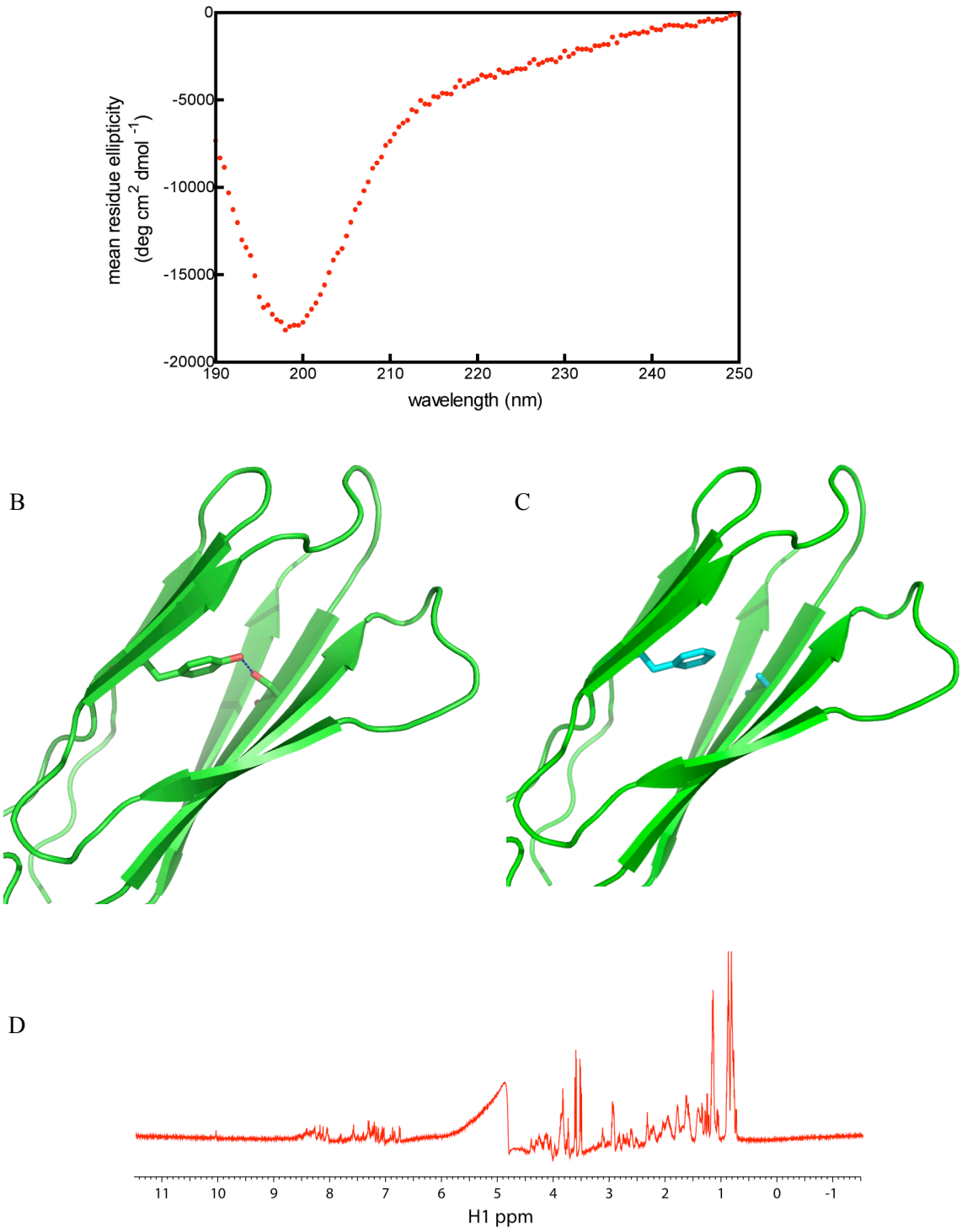
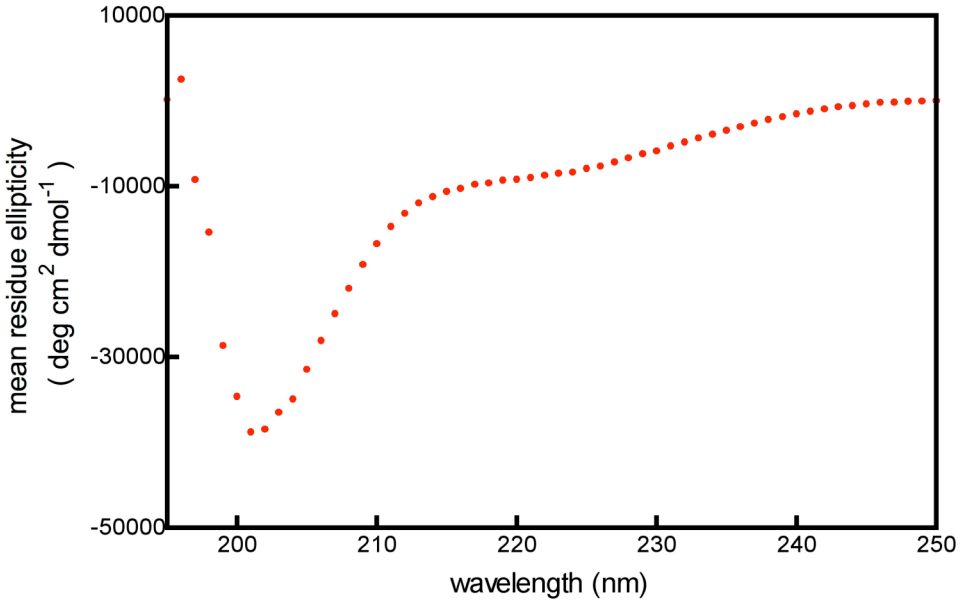


Figure 5.8 B9 and double mutant
(A) Circular Dichroism spectrum of B9 at 20 °C; (B) Design model of B9, Tyr 17 formed a hydrogen bond with Ser 66 in our design model; (C) Mutant of B9 (Y17F,S66A); (D) 1D NMR spectrum of the double mutant at 20 °C.

A



B

Figure 5.9 Spectra of BN1
(A) Circular Dichroism spectrum of BN1 at 20 °C; (B) Design model of ds2, a disulfide bond was introduced into the design model BN1(disulfide bond was shown as stick in brown).

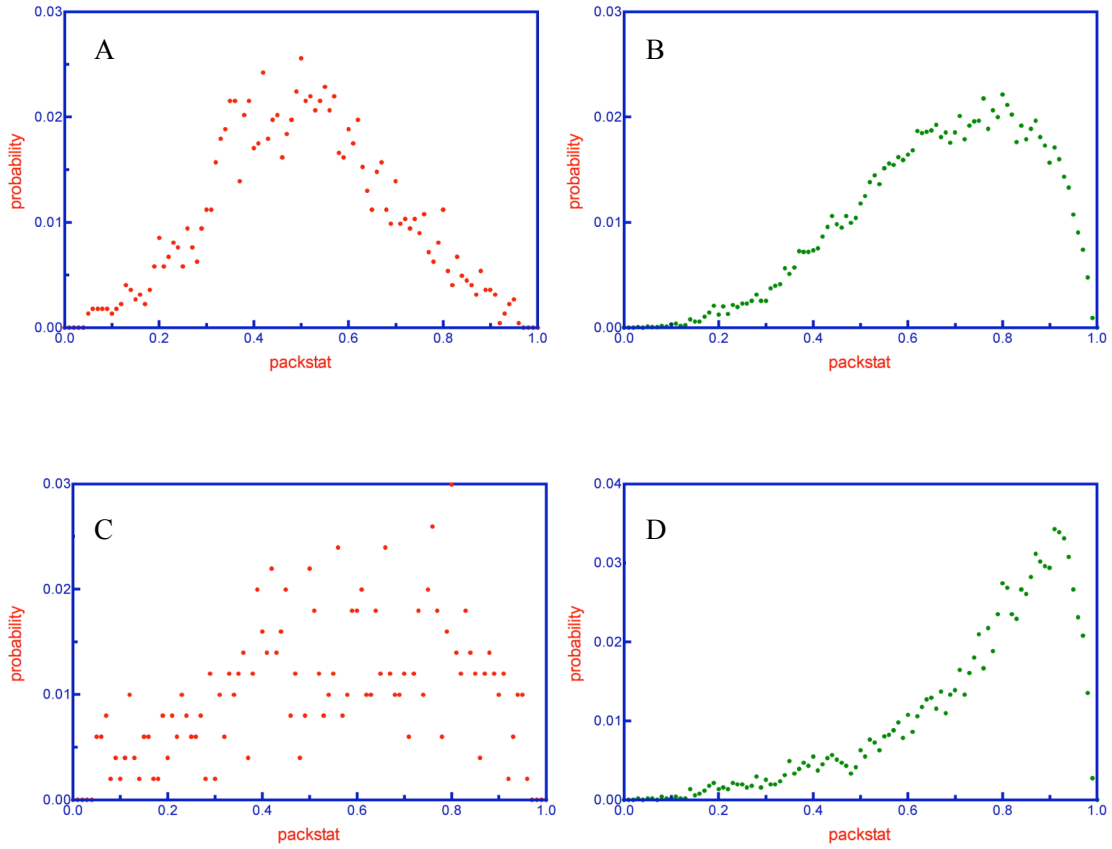


Figure 5.10 Comparison of packstat scores

Distribution of packstat scores for our *de novo* designed models (panel A and C) and naturally occurring β-sheet proteins (panel B and D). Panel A and B are for all the residues; C and D are for the buried residues (have more than 19 neighbors in 10 Å distance).

TABLES

Table 5.1 Sequence compositions for different generation of designs

(A) Sequence composition for the first generation of designs

Sequence	B001	B002	B003	B004
ALA	0.05	0.01	0.09	0.06
CYS	0.00	0.00	0.00	0.00
ASP	0.09	0.04	0.11	0.12
GLU	0.10	0.15	0.04	0.07
PHE	0.04	0.03	0.04	0.03
GLY	0.03	0.11	0.05	0.10
HIS	0.03	0.03	0.01	0.03
ILE	0.05	0.09	0.09	0.10
LYS	0.11	0.06	0.10	0.06
LEU	0.01	0.01	0.05	0.05
MET	0.01	0.00	0.01	0.01
ASN	0.09	0.06	0.10	0.05
PRO	0.05	0.03	0.03	0.03
GLN	0.01	0.06	0.03	0.01
ARG	0.04	0.07	0.03	0.05
SER	0.04	0.03	0.01	0.03
THR	0.10	0.07	0.10	0.11
VAL	0.10	0.14	0.11	0.05
TRP	0.00	0.00	0.01	0.00
TYR	0.06	0.01	0.00	0.04
Positive	0.17	0.16	0.14	0.14
Negative	0.19	0.19	0.15	0.20
Polar	0.24	0.23	0.24	0.20
Aromatic	0.10	0.04	0.05	0.06
Hydrophobic	0.30	0.39	0.42	0.40

(B) Sequence composition for the second generation of designs

Sequence	B5	B6	B7	B8	B9
ALA	0.06	0.05	0.01	0.05	0.01
CYS	0.00	0.00	0.00	0.00	0.00
ASP	0.11	0.11	0.02	0.11	0.07
GLU	0.05	0.05	0.11	0.05	0.07
PHE	0.04	0.07	0.05	0.04	0.02
GLY	0.00	0.05	0.02	0.01	0.07
HIS	0.02	0.01	0.05	0.02	0.02
ILE	0.10	0.10	0.11	0.07	0.08
LYS	0.07	0.08	0.08	0.10	0.07
LEU	0.05	0.02	0.01	0.04	0.06
MET	0.00	0.00	0.00	0.01	0.00
ASN	0.06	0.06	0.07	0.06	0.04
PRO	0.01	0.01	0.02	0.00	0.04
GLN	0.02	0.05	0.02	0.04	0.04
ARG	0.02	0.02	0.07	0.00	0.02
SER	0.08	0.07	0.06	0.06	0.13
THR	0.17	0.11	0.13	0.20	0.17
VAL	0.11	0.12	0.12	0.11	0.05
TRP	0.00	0.01	0.00	0.01	0.01
TYR	0.04	0.01	0.04	0.04	0.02
Positive	0.12	0.12	0.20	0.12	0.12
Negative	0.15	0.15	0.13	0.15	0.14
Polar	0.33	0.29	0.29	0.36	0.37
Aromatic	0.07	0.10	0.08	0.08	0.06
Hydrophobic	0.32	0.35	0.30	0.29	0.31

(C) Sequence composition for the third generation of designs

Sequence	F1	F2	F3	F4	F6	F7	F8	F9	F10	1fna	1ten
ALA	0.01	0.03	0.01	0.02	0.01	0.04	0.04	0.02	0.04	0.08	0.04
CYS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ASP	0.07	0.08	0.08	0.04	0.09	0.02	0.11	0.10	0.03	0.04	0.11
GLU	0.07	0.02	0.07	0.03	0.05	0.03	0.01	0.03	0.02	0.04	0.09
PHE	0.07	0.05	0.07	0.10	0.10	0.05	0.05	0.01	0.03	0.01	0.02
GLY	0.09	0.04	0.04	0.04	0.05	0.05	0.05	0.08	0.05	0.09	0.06
HIS	0.02	0.00	0.05	0.02	0.02	0.02	0.01	0.00	0.03	0.00	0.00
ILE	0.04	0.05	0.04	0.03	0.08	0.04	0.07	0.13	0.05	0.08	0.09
LYS	0.02	0.01	0.03	0.07	0.02	0.03	0.05	0.07	0.09	0.03	0.06
LEU	0.02	0.04	0.04	0.08	0.07	0.07	0.07	0.02	0.07	0.04	0.08
MET	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.03	0.02	0.00	0.01
ASN	0.02	0.05	0.07	0.07	0.05	0.04	0.02	0.01	0.05	0.02	0.03
PRO	0.15	0.12	0.13	0.13	0.12	0.11	0.14	0.11	0.16	0.08	0.06
GLN	0.05	0.04	0.08	0.04	0.02	0.03	0.04	0.02	0.01	0.01	0.02
ARG	0.11	0.05	0.08	0.08	0.07	0.04	0.04	0.01	0.01	0.05	0.03
SER	0.04	0.08	0.02	0.09	0.05	0.09	0.05	0.02	0.07	0.11	0.07
THR	0.04	0.14	0.02	0.04	0.07	0.16	0.11	0.20	0.11	0.13	0.13
VAL	0.08	0.10	0.04	0.04	0.04	0.10	0.02	0.08	0.08	0.10	0.04
TRP	0.01	0.00	0.02	0.01	0.01	0.00	0.02	0.00	0.02	0.01	0.01
TYR	0.07	0.05	0.10	0.05	0.05	0.03	0.07	0.05	0.03	0.07	0.03
Positive	0.15	0.07	0.16	0.16	0.11	0.10	0.11	0.08	0.13	0.09	0.09
Negative	0.13	0.10	0.14	0.08	0.14	0.05	0.12	0.13	0.05	0.09	0.20
Polar	0.16	0.32	0.19	0.24	0.20	0.33	0.23	0.25	0.24	0.27	0.26
Aromatic	0.14	0.11	0.19	0.16	0.16	0.09	0.14	0.07	0.09	0.09	0.07
Hydrophobic	0.41	0.41	0.32	0.35	0.38	0.43	0.40	0.47	0.48	0.46	0.38

(D) Sequence composition for the fourth generation of designs

Sequence	BN1	BN2	BN4	BN5
ALA	0.10	0.00	0.04	0.10
CYS	0.00	0.00	0.00	0.00
ASP	0.07	0.00	0.05	0.06
GLU	0.10	0.13	0.05	0.08
PHE	0.04	0.08	0.02	0.04
GLY	0.02	0.06	0.07	0.06
HIS	0.02	0.00	0.00	0.02
ILE	0.05	0.05	0.10	0.04
LYS	0.06	0.07	0.06	0.02
LEU	0.04	0.05	0.05	0.06
MET	0.01	0.01	0.01	0.00
ASN	0.10	0.10	0.10	0.05
PRO	0.00	0.04	0.01	0.02
GLN	0.08	0.05	0.06	0.02
ARG	0.07	0.07	0.04	0.11
SER	0.05	0.08	0.06	0.12
THR	0.10	0.08	0.14	0.14
VAL	0.05	0.07	0.11	0.04
TRP	0.01	0.00	0.00	0.01
TYR	0.05	0.06	0.05	0.01
Positive	0.15	0.14	0.10	0.15
Negative	0.17	0.13	0.10	0.14
Polar	0.32	0.31	0.36	0.33
Aromatic	0.10	0.14	0.07	0.06
Hydrophobic	0.26	0.27	0.38	0.31

(E) Overall comparison between the de novo design models and naturally occurring proteins

Generation	Proteins	Positive	Negative	Polar	Aromatic	Hydrophobic
1	B001	0.17	0.19	0.24	0.10	0.30
	B002	0.16	0.19	0.23	0.04	0.39
	B003	0.14	0.15	0.24	0.05	0.42
	B004	0.14	0.20	0.20	0.06	0.40
	Average(1)	0.15	0.18	0.23	0.06	0.38
2	B5	0.12	0.15	0.33	0.07	0.32
	B6	0.12	0.15	0.29	0.10	0.35
	B7	0.20	0.13	0.29	0.08	0.30
	B8	0.12	0.15	0.36	0.08	0.29
	B9	0.12	0.14	0.37	0.06	0.31
	Average(2)	0.14	0.14	0.33	0.08	0.31
3	F1	0.15	0.13	0.16	0.14	0.41
	F2	0.07	0.10	0.32	0.11	0.41
	F3	0.16	0.14	0.19	0.19	0.32
	F4	0.16	0.08	0.24	0.16	0.35
	F6	0.11	0.14	0.20	0.16	0.38
	F7	0.10	0.05	0.33	0.09	0.43
	F8	0.11	0.12	0.23	0.14	0.40
	F9	0.08	0.13	0.25	0.07	0.47
	F10	0.13	0.05	0.24	0.09	0.48
	Average(3)	0.12	0.10	0.24	0.13	0.41
4	BN1	0.15	0.17	0.32	0.10	0.26
	BN2	0.14	0.13	0.31	0.14	0.27
	BN4	0.10	0.10	0.36	0.07	0.38
	BN5	0.15	0.14	0.33	0.06	0.31
	Average(4)	0.14	0.14	0.33	0.09	0.31
All	Average(all)	0.13	0.13	0.27	0.10	0.36
	Wt*	0.09	0.20	0.26	0.07	0.38

Average(i:1~4) : average values for the ith generation, highlight in light turquoise
Average(all) : average values for all the design models (hightlight in turquoise)
Wt*: naturally occurring high resolution β-sandwich protein 1ten, highlight in green

SUPPLEMENTARY MATERIAL

Table 5.2 Sequences of all the design models

> B001.pdb
IPEPRVEYHNNKIEVHAPAGSSTARVKVEVKFNNKKYEDEFTQSDDTARFYNTPTGYDIDVK
VKVDTNNKDLEKEYTITM
> B002.pdb
QGQIEVHYENGKVRIHVPPQDDDGEVRGKARFNGKEIEIRVNRGEEEIEITGTQENSVIEVEV
KVTSNGQTVTKRFTVLG
> B003.pdb
PGNATVKTENGKLKIKVDVQISNALVKIEIDMNGTKIRWTFDVADAHLTVDNFPKTADIKVE
VRVDFNNIDANQTLDATG
> B004.pdb
YDNITITFENGKMKIDLGPGTTTADIEAEIDLDGSSIEGRATGGNADLDFDITVPGEKYRARVR
VKLNDYHHDKTIEVTQ

> B5.pdb
NASQDVTVKVQKTTIDVTYKLFNLDIVKIIVEFHPSDATTTDRKDFSASDDNATYTLISTNSSI
EVRVEIDLKNAHYETTITVT
> B6.pdb
VADATVDVQVQDNSITVEFRYNFTSKLDIVVEWKTNSNDDTQRLKVSGNQTSATFTGFPSG
KDVEIIIKIDGDFFHIEITVKAK
> B7.pdb
IQPPKVEITVHATKIEVKVETTQNSEFRIEVYVKRHNHNTIETRTISENRDSTRVLGFGNNHEY
EIIVRVDFSVTSYTFKTKIK
> B8.pdb
NASQDVTVKVQKTTIDVTYKLFNLDIVKIIVKFHGQDMTTEDEKDFSASDDNATWTTTSTNT
SYTVKVEIDLKNAHYETTITVT
> B9.pdb
LDFKQPRTSIQKDSIKYDVISGSTDNSTIIIEGHPESQNTTTTVKLSSSDNSLTLTGFPTGTKGVS
TIKVEGERAHLDYTETWE
> B10.pdb
LDFKQPRTSIQKDSIKFDVISGSTDNSTIIIEGHPESQNTTTTVKLSSSDNSLTLTGFPTGTKGVA
TIKVEGERAHLDYTETWE
> B11.pdb
LDCKQPRTSIQKDSIKFDVISGSTDNSTIIIEGHPESQNTTTTVKLSSSDNSLTLTGFPTGTKGVA
TIKVEGERAHCDYTETWE

> BN1.pdb
YNSKRAEIHFQDNTITARAVAGQYNDHRFEFRVDKENDNQQEKLRMTGSDNSATIKLTDVQ
KSAEAQARVQENNWEYETTYTIL
> BN2.pdb
LGKPEFRFTVRNNSLEVRVQPFNQGPERIKVEITEKNSNSETSFEVTGNQYTVTLSMSNSGTK
YEIRIKFEFNGYRYEQKYEFL
> BN4.pdb

SNSGQITVTLSNNSMTVKYLPGNKKIRTAVIVIKAQDENDIQTYTITGNIFTLTVTGVNGNLK
YIVEVRVEGDDERFSQTYTAQ
> BN5.pdb
NGPGTATITFHDGRVEFRAVASNQTNSEFRAEARPSSSSDGTERKASEQLDRITVEITDTTSSY
KARLELRGDNWTLRHTASLT
> ds1.pdb
YNSCRAEIHFQDNTITARAVAGQYNDHRFEFRVDKENDNQQEKLRMTGSDNSATIKLTDVQ
KSAEAQARVQENNWECETTYTIL
> ds2.pdb
YNSKRAEIHFQDNTITARCVAGQYNDHRFEFRVDKENDNQQEKLRMTGSDNSATIKLTDVQ
KSAEAQARCQENNWEYETTYTIL

> F1.pdb
VDMTWIRQGPYRVILHYPPPPTDVQYARFRVFKRDGGPPSYERERPPGSDHVDITGLEPGQE
YRVRIFYFSGDNNSPQTGPPQEFEFRVPK
> F2.pdb
NDFSFYPTSKTSVTVEAFPPQYDSRRVLVLVRDRTGDDVRATYTVPPNQTSTTITGLLPGYQY
EIIVFSDPPTNDMPNNAQPVSITITGPF
> F3.pdb
LDGSWHPIRQNDLLIDLKPYPSNYQFFRVRATHEEEDGWEREYTVPPQQDKIKFNNFQPGRH
YRVRIFPDPNQGFPPNYYDPYEVQYHHYP
> F4.pdb
LDGTFYPKRSNDLIVKLVPSDSPFQYFLAKAGHSRNEWQSRQFTRPSNLTSISFNNFLPGREY
KIYVFPFPNLGDPPQKRPPYEVRFKHTP
> F6.pdb
PPITIVEEDSDKFRITIPDFNLDVQFADHELWSETGTFPRLRFRIPGNLDSFEFTNLYPGYRYKG
RMFPISNNGDVPPQVFPYHLDISPYT
> F7.pdb
LDVRVTPTGLTTALGEFLPSSSTDTRHLVFVFKHENTAPQASYTQPPTSTTVTITGLVPGVEYI
AMFISRSGTNTSQPNNVPVKYKFRIPP
> F8.pdb
NYLQATFLSDYKAFFRWIPPDDDIQKFDVRTYDSSGSPPQYTYTAPPTLDKATITNLDPGQEY
RTKVIPRIGDGPGPPDTPPITLHLKSFW
> F9.pdb
IDITVIPEYKTEVRIKIDASTLDMQDITVIAYTEGGDTPTFTYTIPPTDKSMTVTGMPPGQKYKI
TVIIYPGNGTLPPITPDTTVTVDGDK
> F10.pdb
PHVHPKKIHLTKLIIKWQPPPVPLTRVIVIMKSSNGDVPTAKFTMPGNATSLEVNGLPPGAKY
KFDVFTWALPGTPDSNTPPYTSETSPNY

REFERENCES

1. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A. De novo design and structural characterization of proteins and metalloproteins. Annu Rev Biochem 1999;68:779-819.

2. Richardson JS, Richardson DC. The de novo design of protein structures. Trends Biochem Sci 1989;14(7):304-309.

3. Pokala N, Handel TM. Review: protein design--where we were, where we are, where we're going. J Struct Biol 2001;134(2-3):269-281.

4. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302(5649):1364-1368.

5. Butterfoss GL, Kuhlman B. Computer-based design of novel protein structures. Annu Rev Biophys Biomol Struct 2006;35:49-65.

6. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278(5335):82-87.

7. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;282(5393):1462-1467.

8. Hecht MH, Das A, Go A, Bradley LH, Wei Y. De novo proteins from designed combinatorial libraries. Protein Sci 2004;13(7):1711-1723.

9. Hecht MH, Richardson JS, Richardson DC, Ogden RC. De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. Science 1990;249(4971):884-891.

10. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, 3rd, Hilvert D, Houk KN, Stoddard BL, Baker D. De novo computational design of retro-aldol enzymes. Science 2008;319(5868):1387-1391.

11. Betz SF, Raleigh DP, DeGrado WF, Lovejoy B, Anderson D, Ogihara N, Eisenberg D. Crystallization of a designed peptide from a molten globule ensemble. Fold Des 1995;1(1):57-64.

12. Raleigh DP, Betz,SF and Degrado WF. A de Novo Designed Protein Mimics the Native State of Natural Proteins. J Am Chem Soc 1995;117(28):7558-7559.

13. Quinn TP, Tweedy NB, Williams RW, Richardson JS, Richardson DC. Betadoublet: de novo design, synthesis, and characterization of a beta-sandwich protein. Proc Natl Acad Sci U S A 1994;91(19):8747-8751.

14. Lim A, Makhov AM, Bond J, Inouye H, Connors LH, Griffith JD, Erickson BW, Kirschner DA, Costello CE. Betabellins 15D and 16D, de Novo designed beta-sandwich proteins that have amyloidogenic properties. J Struct Biol 2000;130(2-3):363-370.

15. Yan Y, Erickson BW. Engineering of betabellin 14D: disulfide-induced folding of a beta-sheet protein. Protein Sci 1994;3(7):1069-1073.

16. Kortemme T, Ramirez-Alvarado M, Serrano L. Design of a 20-amino acid, three-stranded beta-sheet protein. Science 1998;281(5374):253-256.

17. Pessi A, Bianchi E, Crameri A, Venturini S, Tramontano A, Sollazzo M. A designed metal-binding protein with a novel fold. Nature 1993;362(6418):367-369.

18. Nanda V, Rosenblatt MM, Osyczka A, Kono H, Getahun Z, Dutton PL, Saven JG, Degrado WF. De novo design of a redox-active minimal rubredoxin mimic. J Am Chem Soc 2005;127(16):5804-5805.

19. Hamill SJ, Steward A, Clarke J. The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. J Mol Biol 2000;297(1):165-178.

20. Clarke J, Cota E, Fowler SB, Hamill SJ. Folding studies of immunoglobulin-like beta-sandwich proteins suggest that they share a common folding pathway. Structure 1999;7(9):1145-1153.

21. Cota E, Steward A, Fowler SB, Clarke J. The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold. J Mol Biol 2001;305(5):1185-1194.

22. Olson CA, Roberts RW. Design, expression, and stability of a diverse protein library based on the human fibronectin type III domain. Protein Sci 2007;16(3):476-484.

23. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66-93.

24. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature 2007;450(7167):259-264.

25. Canutescu AA, Dunbrack RL, Jr. Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci 2003;12(5):963-972.

26. Fleming PJ, Rose GD. Do all backbone polar groups in proteins form hydrogen bonds? Protein Sci 2005;14(7):1911-1917.

27. Madde TL, Tatusov, R.L., Zhang, J. Applications of network BLAST server. Methods Enzymol 1996;266:131-141.

28. Wright CF, Christodoulou J, Dobson CM, Clarke J. The importance of loop length in the folding of an immunoglobulin domain. Protein Eng Des Sel 2004;17(5):443-453.

29. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 2003;326(4):1239-1259.

30. Stemmer WP, Crameri A, Ha KD, Brennan TM, Heyneker HL. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. Gene 1995;164(1):49-53.

31. Laurence M, Philips,KJ., Liu,DR. Supercharging proteins can impart unusual resilience. J Am Chem Soc 2007;129(33):10110-10112.

32. Steward A, Adhya S, Clarke J. Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. J Mol Biol 2002;318(4):935-940.

33. Richardson JS, Richardson DC. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci U S A 2002;99(5):2754-2759.

34. Morozov AV, Kortemme T, Tsemekhman K, Baker D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. Proc Natl Acad Sci U S A 2004;101(18):6946-6951.

CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

In our studies on the β-sheet protein design, we have the following conclusions. First of all, we developed a protocol to account for side chain entropy in Rosetta energy function, however, the incorporation of explicit side chain entropy and free energy calculations into Rosetta does not substantially increase our ability to recapitulate native sequences in protein design simulations. In general, our results suggest that side chain entropy plays a relatively small role in determining the environmental preferences of the amino acids.

Second of all, the fixed backbone design results are very encouraging, given a designable backbone, we were able to design a protein that is well-folded and much more stable with only positive design, suggesting that the *de novo* design of a β-sandwich protein may be possible without extensive negative design. However, one caveat is that the negative design elements may be encoded in the backbone conformation. It will be exciting to see if we can also design a protein that allow for backbone sampling.

Thirdly, we tried to design a part of the backbone for the wild type tenascin by removing a ten-residue loop from the wild type structure and rebuilding it with a specific conformation. This involved the optimization of both the conformational and sequence space. Two of the designs were crystallized and one of them matches the design model very well. This result indicates that with the current Rosetta energy function and sampling technique, it is possible to design a 10-residue loop with high accuracy. These results validate the design protocol we used for *de novo* design.

Last but not least, our *de novo* design trials failed for several generations as the design models mostly aggregate. The successful de novo design of a β-sandwich protein still remains an unanswered question. However, from our failure we learned some lessons that could potentially useful for future design.

By comparing our design models with the naturally occurring β-sheet proteins, we see several different patterns in the sequence composition. The hydrophobic residues are important for driving folding process so it is important to have a certain amount of hydrophobic residues (~40%). More importantly, these residues should be packed well enough to form a compact core that can nucleate the folding process of the whole protein. We found that our designed models generally are poorly packed compared to the naturally occurring β-sheet proteins and it still remains to be seen if this is a problem of conformational sampling or energy function. All our design models have low solubility, suggesting the need to incorporate negative design elements to design against unwanted misfolded or aggregated states. Another approach is to try to increase the net charges to increase solubility.

Nature have encoded some negative design elements to prevent misfolding and aggregation, we should borrow those information for our future design, for instance, incorporating some β−sheet element breakers, like edge prolines, β−bulges, polar residues in the middle of edge strands *etc.* that may contribute to prevent aggregation. In our future design trials, we should include these negative design element features to our design simulation.