

# COMPUTER-BASED DESIGN OF B-SHEET CONTAINING PROTEINS

Doo Nam Kim

A dissertation submitted to the faculty at 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

Chapel Hill  
2016

Approved by:

Brian Kuhlman

Qi Zhang

Andrew Lee

Harold Erickson

Jane Richardson

© 2016  
Doo Nam Kim  
ALL RIGHTS RESERVED

## ABSTRACT

Doo Nam Kim: Computer-based Design of  $\beta$ -sheet Containing Proteins  
(Under the direction of Brian Kuhlman)

Protein design is an excellent test of the minimal determinants of protein structure. Although 70% of naturally occurring proteins contain  $\beta$ -sheets, most previous design efforts have been limited to  $\alpha$ -helix bundle proteins or the redesign of naturally occurring proteins. Here, we test and develop computer-based methods for designing proteins rich in  $\beta$ -strands. The molecular modeling program Rosetta was used for three separate design tasks: (1) the design of  $\alpha/\beta$  and  $\alpha+\beta$  proteins with a new method called SEWING, which builds proteins from pieces of naturally occurring proteins, (2) the stabilization of  $\beta$ -sheet proteins via the redesign of surface-facing residues, and (3) the *de novo* design of  $\beta$ -sandwich proteins. This research showed that it is possible to extend the SEWING method to non- $\alpha$ -helix proteins, allowing the incorporation of structural features found in nature, and that it is possible to dramatically boost protein thermal stability ( $> 25^\circ\text{C}$ ) with the redesign  $\beta$ -sheet surfaces. However, we also found that the *de novo* design of  $\beta$ -sandwich proteins still remains an elusive goal.

*“Yeah, the template file had this section.”*



## ACKNOWLEDGEMENTS

I deeply appreciate Brian Kuhlman, Tim Jacobs, and Bryan der for their limitless patience and teaching for me. Without them, I wouldn't be here. I thank my family: Chanmi Choi, Philip Kim and Eugene Kim. I thank my committee members for their precious time: Jane Richardson, Harold Erickson, Andrew Lee, and Qi Zhang. I thank Nikolay Dokholyan and Qi Zhang for teaching me while I rotated. I thank all Kuhlman lab members including Alex Kenan, Alex, Alex, Amanda Loshbaugh, Andrew Leaverfay, Andrew Lerner, Ben Stranges, Deanne Sammond, Doug Renfrew, Frank Teets, Grant Murphy, Gurkan Guntas, Hayretin Yumer, Jack Macquire, Jeffrey Jones, Jenny (Xiaozhen) Hu, Joseph Harrison, Kevin Houlihan (who gratefully taught me how to identify disulfide bonds using Rosetta), Mahmud Hussain, Matt O'Meara, Minnie Langlois, Oana Lungu, Ryan Hallet, Sharon Guffy, Steven Lewis, Stephan Kudlacek, Seth Zimmerman, and Wesley Roten.

I thank Rosetta community members including Noboyasu Koga, Enrique Marcos, Gabriel Rocklin, Tom Linsky and Christopher Bahl in Baker lab. I thank my collaborators: Thomas Szyperski, Jasmin Federizon, and RamaKrishna Pulavarti. I thank my rotation students: Kisurb Choi, Devin Rodriguez. I thank Bo Zhao for this teaching. I thank UNC computing center experts: Sandeep and Jenny Williams. I thank Barry Lentz for his encouragement. I thank Ryan Lucey for his dedicated edition.

## TABLE OF CONTENTS

ABSTRACT .....	III
ACKNOWLEDGEMENTS.....	V
LIST OF TABLES .....	IX
LIST OF FIGURES .....	X
LIST OF ABBREVIATIONS AND SYMBOLS .....	XI
CHAPTER 1: COMPUTATIONAL PROTEIN DESIGN .....	13
Introduction .....	13
Protein as therapeutics .....	13
Computational protein design .....	14
Challenges in computational protein design.....	15
Necessity of <i>de novo</i> protein design .....	16
Previous <i>de novo</i> protein design methods .....	17
Supersecondary structure based <i>de novo</i> protein design method .....	20
Conclusion and following chapters.....	21
CHAPTER 2: COMPUTATIONAL <i>DE NOVO</i> DESIGN OF A/B AND A+B PROTEINS .....	22
Introduction .....	22
Computational Design and Experimental Results .....	25
Design with three secondary structure substructures .....	25
Design with five secondary structure substructures .....	27
Increase net charge in an effort to monomerize well-folded $\alpha+\beta$ and $\alpha/\beta$ proteins.....	32

Reversion design in an effort to monomerize and better fold en8 design .....	34
Redesign with refolding in an effort to monomerize and better fold en8 design .....	37
<b>Conclusion .....</b>	<b>38</b>
<b>Supporting Materials .....</b>	<b>39</b>
 <b>CHAPTER 3: BOOSTING STABILITY OF B-SHEET PROTEINS</b>	
<b>BY SURFACE REDESIGN .....</b>	<b>47</b>
<b>Overview .....</b>	<b>47</b>
<b>Introduction .....</b>	<b>48</b>
<b>Results .....</b>	<b>52</b>
<b>Discussion .....</b>	<b>59</b>
<b>Materials and Methods .....</b>	<b>61</b>
Computational Design and Analysis of proteins .....	61
Protein Expression and Purification .....	61
Circular Dichroism .....	61
Fluorescence .....	62
 <b>CHAPTER 4: <i>DE NOVO</i> DESIGN EFFORTS FOR B SANDWICH PROTEINS .....</b>	
<b>Introduction .....</b>	<b>64</b>
<b>Computational Design and Experimental Results .....</b>	<b>66</b>
Design of Backbone by SEWING .....	66
Design of Backbone by Assembling $\beta$ -Strands with Random Backbone Angles .....	68
Design of Backbone by ab initio Folding .....	68
Design with Folding Nucleus Conservation .....	69
Design with Repopulation of Existing Backbones .....	69
Redesign with Native $\beta$ -sandwich Backbones .....	70
<b>Conclusion .....</b>	<b>72</b>

Supporting Materials .....	72
<b>CHAPTER 5: FUTURE DIRECTIONS .....</b>	<b>75</b>
Towards More Efficient Protein Design .....	75
Necessity of Experimentally Determined Structural Information .....	76
Towards a More Accurate Score Function .....	77
Towards Successful <i>De novo</i> Design of $\beta$ -sandwich proteins .....	79
Potential Applications of Designed Proteins.....	81
Possible Reason of Increased Expression Yield of High Net Charged Proteins .....	82
<b>APPENDING CHAPTER: USED ROSETTA INPUT PROTOCOLS.....</b>	<b>83</b>
Rationale .....	83
RosettaScripts .....	84
Flags .....	87
Resfile .....	88
References .....	89

## LIST OF TABLES

TABLE 1.1 NUMBER OF PROTEIN FOLDS AND STRUCTURES ACCORDING TO CLASSES .....	16
TABLE 1.2 SOME EXAMPLES OF COMPUTATIONAL <i>DE NOVO</i> PROTEIN DESIGNS. ....	18
TABLE 2.1 EXPERIMENTAL RESULTS OF A+B PROTEINS USING SMALL SUBSTRUCTURES .....	26
TABLE 2.2 EXPERIMENTAL RESULTS OF A+B PROTEINS USING LARGE SUBSTRUCTURES.....	28
TABLE 2.3 EXPERIMENTAL RESULTS OF A/B PROTEINS USING LARGE SUBSTRUCTURES.....	28
TABLE 2.4 EXPERIMENTAL RESULTS OF ORIGINAL AND SUPERCHARGED A/B AND A+B PROTEINS. ....	33
TABLE 2.5 PREDICTED SECONDARY STRUCTURE OF EN8 BY NMR .....	35
TABLE 2.6 REVERTED AND ORIGINAL A/B AND A+B PROTEINS.....	36
TABLE 3.1 COMPUTED STABILITIES FOR PROTEINS.....	55
TABLE 3.2 MEASURED STABILITIES FOR PROTEINS. ....	56
TABLE 4.1 REDESIGN WITH NATIVE B-SANDWICH BACKBONES.....	71

## LIST OF FIGURES

Figure 1.1 Overview of repeat protein design protocol .....	19
Figure 1.2 Overview of the SEWING method.....	20
Figure 2.1 Selected design models that cooperatively unfolded .....	23
Figure 2.2 Application of SEWING method using small substructures .....	24
Figure 2.3 An example that two large substructures are merged .....	24
Figure 2.4 en8 structure and its closest structure in nature. ....	29
Figure 2.5 HSQC of en8 protein. ....	30
Figure 2.6 CD spectra of en8 protein. ....	31
Figure 2.7 Buried unsatisfied hydrogen bond analysis. ....	35
Figure 2.8 An example of C-terminal redesign.....	36
Figure 2.9 The original en8 structure and one of the refolded redesigns.....	38
Figure S2.1 Biophysical characterization of ab1 protein. ....	40
Figure S2.2 Biophysical characterization of ab2 protein. ....	42
Figure 3.1 Concept of the charge zipper scheme for the TNfn3 $\beta$ -sandwich fold.....	51
Figure 3.2 Surface exposed $\beta$ -sheet residues for the various designs.....	53
Figure 3.3 WT TNfn3 and redesigns with surface exposed residues .....	54
Figure 3.4 CD of the Rosetta designs and the charge zipper designs. ....	57
Figure 3.5 $T_m$ measurements for the PS protein as a function of NaCl concentration.....	58
Figure 4.1 $\beta$ -sandwiches in various immune systems.....	64
Figure 4.2 Overall Greek-key motif topology of fn3 $\beta$ -sandwich. ....	65
Figure 4.3 $\beta$ -sandwich backbone design by SEWING.....	67
Figure 4.4 Packing comparison of various WT $\beta$ -sandwich and design tried ones.....	70
Figure 5.1 An example of refining atomic models into cryo-EM maps. ....	76
Figure 5.2 Distribution of amino acids in WT $\beta$ -sandwich proteins. ....	78
Figure 5.3 Molecular imaging of a mouse implanted with tumors .....	81

## LIST OF ABBREVIATIONS AND SYMBOLS

A	alanine
ADMET	absorption, distribution, metabolism, excretion, toxicity
CD	circular dichroism
CDR	complementarity determining region
CO	contact order
CSI	chemical shift index
°C	degree Celsius
<i>E. coli</i>	<i>Escherichia coli</i>
EM	electron microscopy
GMEC	global minimum energy conformation
GPU	graphic processing unit
GuHCl	guanidine hydrochloride
HSQC	heteronuclear single quantum coherence
IgG	immunoglobulin G
kDa	kilo dalton
LB	luria broth
MALS	multiple angle light scattering
MCSG	microlytic MCSG commercially available crystallization reservoir buffers
MHC	major histocompatibility complex
mRNA	messenger ribonucleic acid
μM	micro mole
NaCl	sodium chloride
NMR	nuclear magnetic resonance
PDB	protein data bank
REU	Rosetta energy unit

RMSD	root mean squared deviation
S	serine
SEWING	structure extension with native substructure graphs
SUMO	small ubiquitin-like modifier
T	threonine
TALOS	torsion angle likeliness obtained from shift and sequence similarity
T <sub>m</sub>	melting temperature
TMAO	trimethylamine N-oxide
TNfn3	type III domain of the protein tenascin (pdb code: 1ten)
ULP1	ubiquitin-like-specific protease 1
V	valine
Y	tyrosine



## **CHAPTER 1: COMPUTATIONAL PROTEIN DESIGN**

### **Introduction**

Protein engineering is commonly used to improve the effectiveness and commercial value of therapeutics and research reagents. For example, the DNA sequencing industry has benefited from re-engineered DNA polymerases that allow for more efficient sequencing (1). Proteins can be reengineered to adopt a wide array of conformations. For example, compared to DNA origami assemblies that are limited Watson-Crick base pairs as a sole building block, protein assemblies allow more diverse directionality and conformations using electrostatic interactions, hydrophobic interactions, and hydrogen bonds between backbone and side-chain polar groups (2).

The pharmaceutical market is by far the biggest sector in biotechnology, much larger than the nucleic acid sequencing or biomarker markets. Therefore, the pharmaceutical industry is one that drives the United States economy. Each year, the total nominal spending on medicines in the U.S. is \$425 billion (3), and the industry supports around 3.4 million jobs. Pharmaceutical drugs are derived from chemical synthesis or are a result of biopharmaceuticals, which can include vaccines, recombinant proteins, stem cell therapies and others. Drugs can treat, or cure diseases of the human body.

### **Protein as therapeutics**

In the pharmaceutical market, protein-based drugs such as antibodies and peptides are the fastest growing players due to lower toxicity and tighter specificity than more traditional small molecule drugs. Protein drugs tend to have little side effects other than some unwanted immunogenicity which can be reduced by prediction (4). While, the toxicity issue of the small molecules seems inevitable due to proteins' inherent promiscuous interactions to small ligands (5). Proteins perform many critical functions

in our body. They protect the body against viruses and bacteria, transmits signals that coordinate biological processes between cells, tissues, and organs, and provide structure and support for cells. Therefore, it has been said that “Almost everything in biomedicine could be impacted by an ability to build better proteins”(6). Traditional small molecules can be good drug candidates due to their ability to bind to small pockets in proteins, oral administration, cheaper manufacturing cost, and better-studied ADMET properties. However, protein drugs are often more suited for controlling protein-protein interactions and need shorter and less expensive clinical trials compared to small molecules. Due to its very specific binding to a target with CDR, the antibody is also used as antibody conjugated small molecules (ADC: antibody drug conjugates). Even much expected mRNA based therapeutics by many companies including CureVac (7) and Moderna Therapeutics (8) rely on protein translation once the mRNA enters the body. Indeed, it has been mentioned that “Molecular recognition is a major part of what distinguishes biology from chemistry” (9).

### **Computational protein design**

To maximize the usefulness of proteins, scientists have been computationally designing them using a variety of methods (10) (11) (12) (13). One established approach for rotamer-based sequence design is dead-end elimination, which removes rotamers that are not part of the global minimum energy conformation (GMEC) (14) (15) (16). However, for many design goals it is necessary to sample alternative backbone conformations, and in these cases the traditional dead-end elimination no longer guarantees that flexible-backbone GMEC will not be pruned (17).

The Rosetta molecular modeling program (18) uses Monte Carlo simulations to search for low energy states. This simulation often efficiently converges energy minima due to its score function and rotamer library, which are best optimized to reproduce x-ray crystallography derived structures. The Rosetta community has been designing proteins for a variety of purposes including increasing the net charge of proteins (often called supercharging, this has been shown to improve the thermostability and

affinity of antibodies by up to 30-fold) (19) (20). Rosetta has been used to design a variety of clinically relevant proteins including: bispecific antibodies (21), protein inhibitors that bind to a conserved region of the hemagglutinin of an influenza virus (10), pH-sensitive IgG binding protein (12), protein binders that have high affinity and selectivity for a steroid molecule (22), and the removal of T-cell epitopes to reduce immunogenicity (4). Aside from design, the Rosetta program has been used for docking between hemagglutinin and monoclonal antibodies (23) and making predictions in order to improve catalytic activity (24). These computational protein designs are often need to be complemented or followed by experimental methods such as phage and yeast display (10) or single B-cell technology.

### **Challenges in computational protein design**

Although there have been some successes (25) (26), protein interface design remains one of the most challenging goals in computer-based design (27). To address this protein interface design challenge more effectively, optimizing the protein design score function for interface design only (rather than common monomer design) was tried with some success (28) (29). The Shifman group believed that buried polar atoms often occurred at protein-protein interfaces. In addition to this, the group showed that increasing the weight of the electrostatic term led to some protein-protein interface design successes as seen from an antibody affinity maturation design case as well (30). Nature also uses explicit electrostatic interactions such as the 10 femtomolar  $K_d$  of the barnase-barstar protein complex (31). However, their claim of observing some buried polar atoms at native protein-protein interfaces is under active scrutiny because our group believes that buried polar atoms are hardly be found in native protein-protein interfaces, and these buried polar atoms often lead to failures in interface design (27). Other challenging computational protein design goals include *de novo* design and all  $\beta$ -sheet protein design (32).

## Necessity of *de novo* protein design

*De novo* design builds both backbones and sidechains from scratch (33). *De novo* design serves two main purposes; first, it tests our understanding of protein structure and folding, like Richard Feynman said, “what I cannot make, I do not understand”. Second, it opens the possibility of using previously undiscovered protein structures. Although more than 112,000 protein structures have been deposited (as of August in 2016) (34), it is believed that a much larger portion is still unexplored, when one considers the possible combination of secondary structure elements. For example, only 1,400 out of an estimated 10,000 possible different protein folds have been identified (as of 2014) (35). Indeed, when I designed 51 different  $\beta$ -sandwich,  $\alpha/\beta$ , and  $\alpha+\beta$  proteins *de novo*, no designed structure was the same as any of the already determined protein structures according to the Dali database (36). Therefore, it is assumed that there are many “never born proteins” (35). For example, among the four major classes of protein structures (Table 1.1), all  $\beta$ -sheet proteins (32) and non-repeat  $\alpha/\beta$  proteins whose N and C-terminal (in primary sequence) regions that are located at the edge (in 3-D) have not been successfully *de novo* designed.

Class	SCOP (37)		CATH (38)
	Number of folds	Number of structures	Number of folds
Mainly $\alpha$	284	7,534	397
Mainly $\beta$	174	10,570	241
$\alpha/\beta$	147	11,853	626
$\alpha+\beta$	376	10,950	

**Table 1.1 Number of protein folds and structures according to classes** (May, 2015) (39).

## **Previous *de novo* protein design methods**

There have been numerous efforts to design proteins *de novo* since the design of ALPHA-1 (40) and Betabellin (41) (Table 1.2). Researches using Rosetta as their primary design tool have been designing proteins *de novo* mostly by “fragment insertion” (42), which replaces backbone torsion angles semi-randomly with those of either three or nine residue long extracted fragments (33) (43) (44). Although these fragment insertion methods have been successful, most of them are limited by “idealized” or “canonical” backbone geometries, which often fail to use nature’s distinct features (45). Furthermore, as stated by Jacobs et al. (46), backbone generation using repeating units is limited to native geometries of those repeating units (Figure 1.1) (47) (48).

Year	Class	Designer	N-term & C-term (topologically)	Design blocks for backbone	Comment
2003	$\alpha/\beta$	Kuhlman (33)	At least one terminal is located in the middle	3 or 9 residues long fragments	Very stable design ("top7")
2012	$\alpha/\beta$	Koga (43)	At least one terminal is located in the middle	3 or 9 residues long fragments	Design of five folds
2014	A	Joh & DeGrado (49)	Both N and C-terminals are located at the edges	Tertiary templates	Stochastic search of parameterized backbone geometries in the space of Crick parameters
2015	Includes $\alpha/\beta$	Parmegiani (47)	Some designs have N and C-terminals that are located at the edges	Naturally occurring repeating units	Guided by constraints derived from existing structures
2015	$\alpha/\beta$	Lin & Koga (50)	At least one terminal is located in the middle	3 or 9 residues long fragments	Design of seven folds
2015	$\alpha/\beta$	Park (48)	Both N and C-terminals are located at the edges	Naturally occurring repeating units	Guided by constraints derived from existing structures
2015	A	Brunette (46)	Some designs have N and C-terminals that are located at the edges	Repeating a $\alpha$ -loop- $\alpha$ -loop structural motif	$\alpha$ helix bundle design
2015	A	Murphy (51)	At least one terminal is located in the middle	3 or 9 residues long fragments	$\alpha$ helix bundle design
2015	A	Doyle & Bradley (44)	Closed architecture	3 or 9 residues long fragments	$\alpha$ helix bundle design
2015	$\alpha+\beta$	Marcos (52)	At least one terminal is located in the middle	Parametric geometric constraints	Used naturally conserved sequences
2016	$\alpha/\beta$	Huang (53)	Closed architecture	3 or 9 residues long fragments	Guided by geometric constraints derived from existing structures
2016	A	Jacobs (45)	At least one terminal is located in the middle (CA01)	Super secondary structures	$\alpha$ helix bundle design
2016	A	Boyken (54)	At least one terminal is located in the middle	Parametric geometric constraints	$\alpha$ helix bundle design
2016	$\alpha/\beta, \alpha+\beta$	Kim (this thesis)	Both N and C-terminals are located at the edges	Super secondary structures	Used bigger super secondary structures than motif (55)

**Table 1.2 Some Examples of Computational *de novo* Protein Designs.** I listed those designs whose 3-D structures were determined at least by secondary structure level experimentally. I excluded sidechain-only *de novo* designs with native backbones as redesigns (15) (16) (56) (57). There are more *de novo* designs especially that were done by the non-Rosetta community (35), however I did not introduce them here.

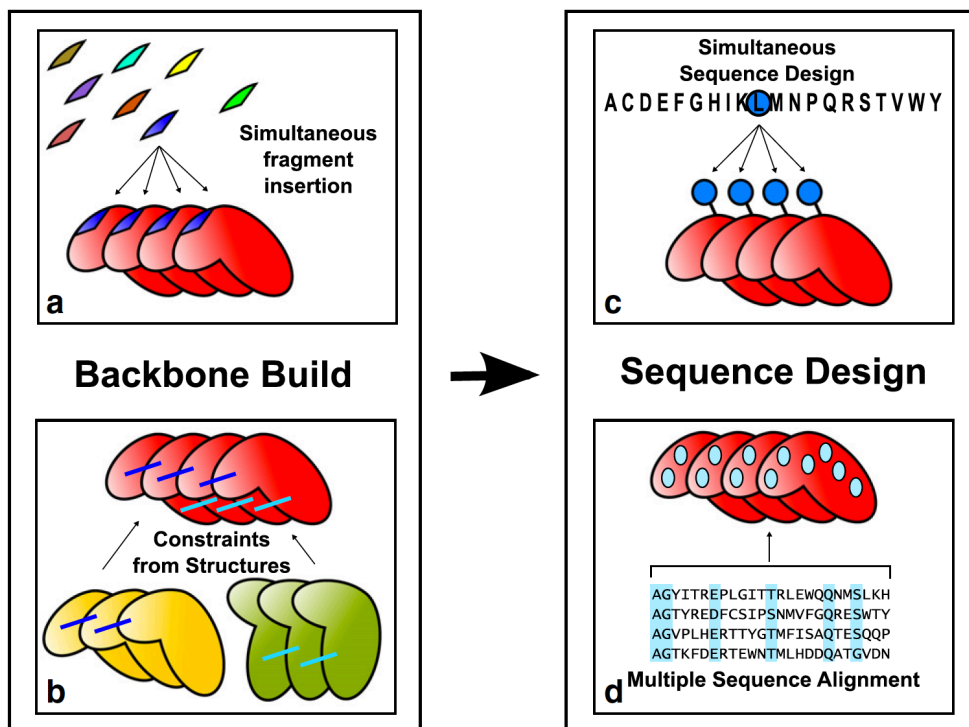
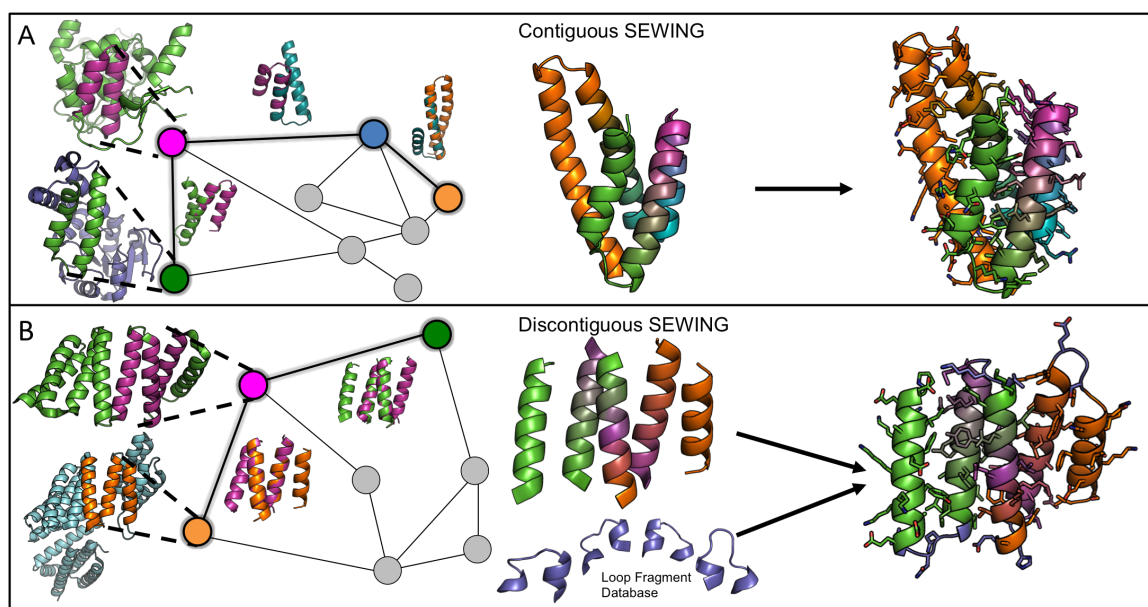


Figure 1.1 Overview of repeat protein design protocol (47)

## Supersecondary structure based *de novo* protein design method (SEWING)

To capture nature's unique geometric features, Jacobs et al. used protein supersecondary structures called "smotif" (45) (55) (58). Through this method, they were able to design helix backbone geometries that captured nature's unique features. They named this process "SEWING" (Structure Extension With Native substructure Graphs). They could extract not only native backbone geometries but also native sidechains that are often responsible for important capping, packing, hydrogen bonding and electrostatic interactions (Figure 1.2).



**Figure 1.2 Overview of the SEWING method.** (A) Contiguous SEWING workflow. (B) Discontiguous SEWING workflow. Each panel, from left to right: parental PDBs with extracted substructures; Graph schematic colored nodes indicate substructures contained in final design model, superimposed structures show structural similarity indicated by adjacent edges; Design model before sequence optimization and loop design; Final design models (58).



## Conclusion and following chapters

Protein engineering is important to better use protein's intrinsic biological properties. Computational protein design has been utilized because it reduces the time and cost of protein engineering. However, *de novo* design remains relatively challenging, especially when trying to preserve nature's unique geometric properties. To address this challenge, Jacobs et al. developed a new backbone design method, named SEWING, which has shown its  $\alpha$ -helix bundle design capability. Chapter 2 describes our efforts to apply this SEWING method to non- $\alpha$ -helix proteins, such as non-repeat  $\alpha/\beta$  and  $\alpha+\beta$  proteins. Chapters 3 and 4 focus on  $\beta$ -sandwich protein designs. Chapter 3 shows the charge zipping of  $\beta$ -sandwich proteins in an effort to induce the desired folding order of complex Greek key motif proteins. Chapter 4 presents various approaches for *de novo* design of  $\beta$ -sandwich proteins including the SEWING method.

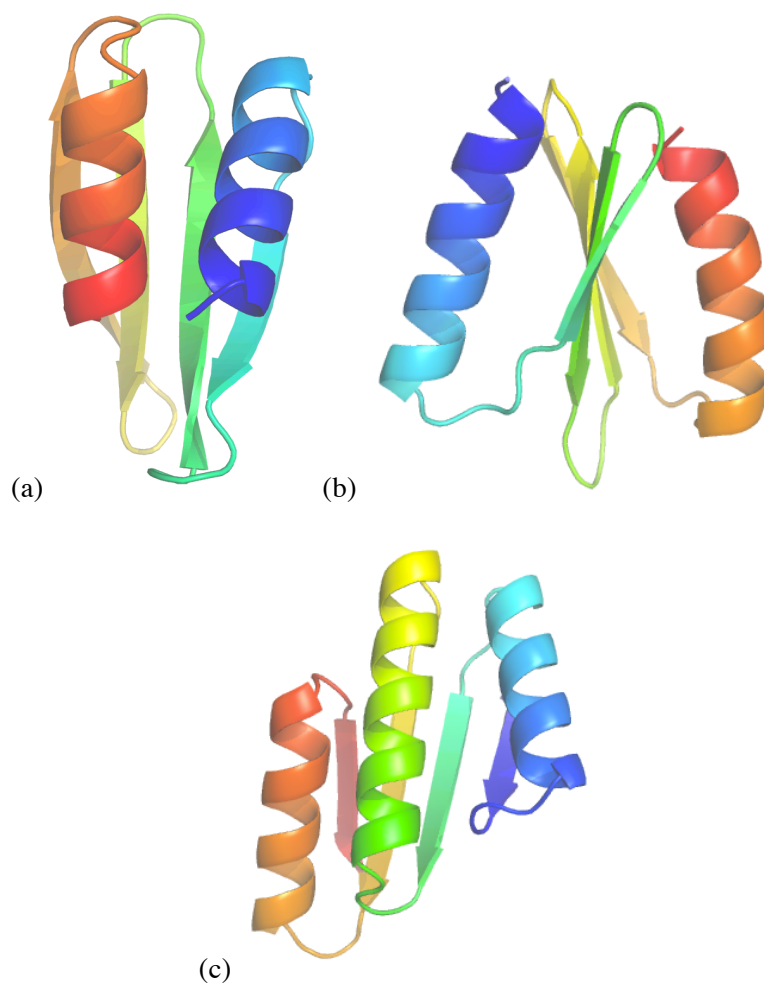
## CHAPTER 2: COMPUTATIONAL *DE NOVO* DESIGN OF A/B AND A+B PROTEINS

### Introduction

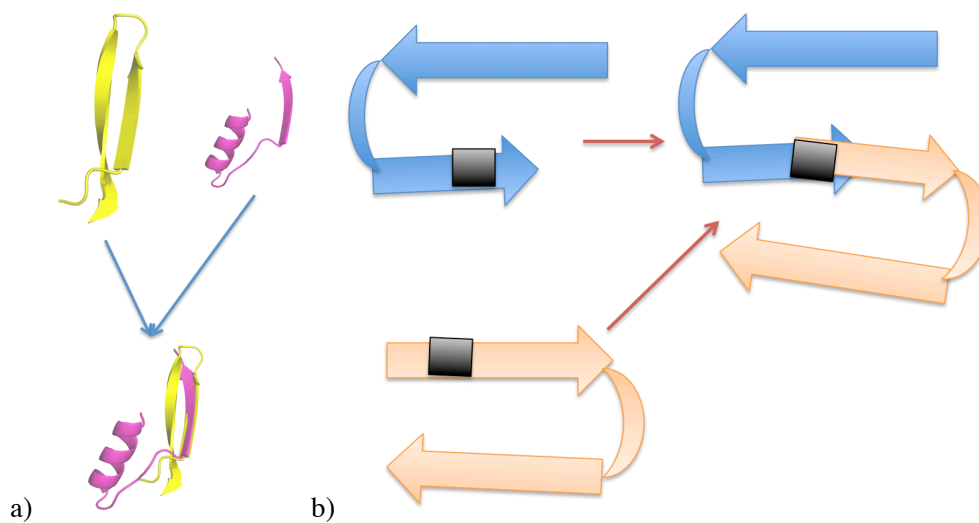
$\alpha/\beta$  class proteins have alternating secondary structures of  $\alpha$ -helix and  $\beta$ -strand. Examples of  $\alpha/\beta$  topology include  $\beta\alpha\beta\alpha\beta\alpha$  and  $\beta\alpha\beta\alpha\beta$ . Another topology are the  $\alpha+\beta$  class proteins which have  $\beta$ -meander structure(s) and are often found in single stranded RNA binding proteins (59). The common structure of this  $\alpha+\beta$  topology is  $\alpha\beta\beta\beta\alpha$ . Although these  $\alpha/\beta$  and  $\alpha+\beta$  class proteins constitute significant portions of whole protein structures (Table 2.1), there is only one successful *de novo* design for  $\alpha+\beta$  class so far (52). Previous *de novo* designs for  $\alpha/\beta$  class proteins are also fewer than those for  $\alpha$ -class proteins. Additionally all designs in this class, except repeat protein designs, have at least one terminal region (in primary sequence) that is located in the middle (in 3-D space) (Table 1.2).

Therefore, we tested whether we can design small globular non-repeat proteins that have N and C-terminal regions that are located at the edges. We were not sure whether this type of protein topology of a small size (less than 100 residues) could even stably fold. Because, when we searched our designed structures against all previously determined protein structures (36), we got only larger proteins containing this type of topology as the closest structures.

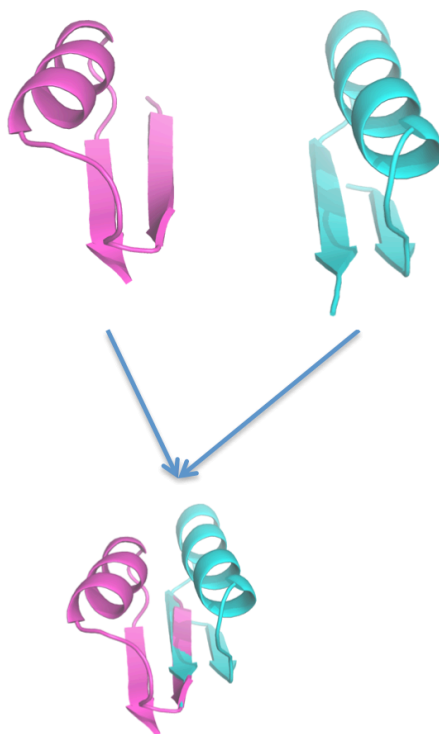
This ambitious design trial was done with a new backbone design method, SEWING (introduced in chapter 1). Firstly, we used three secondary structures as substructures for backbone assembly as Jacobs et al. used (45) to design  $\alpha+\beta$  proteins. Then we used five secondary structures as substructures for backbone assembly to design  $\alpha+\beta$  and  $\alpha/\beta$  proteins (Figures 2.1, 2.2 and 2.3).



**Figure 2.1 Selected design models that cooperatively unfolded.** a) “m4”, one of the  $\alpha+\beta$  designs that used three secondary structures as substructures for backbone assembly, b) “ab2”, one of the  $\alpha+\beta$  designs that used five secondary structures as substructures for backbone assembly, c) “en8”, one of the  $\alpha/\beta$  designs that used five secondary structures as substructures for backbone assembly



**Figure 2.2 Application of SEWING method using small substructures.** a) An example showing how two sets of three secondary structure based substructures are merged, b) An example of merged two  $\beta$ -hairpins. Black boxes represent regions that are used for superimposition.



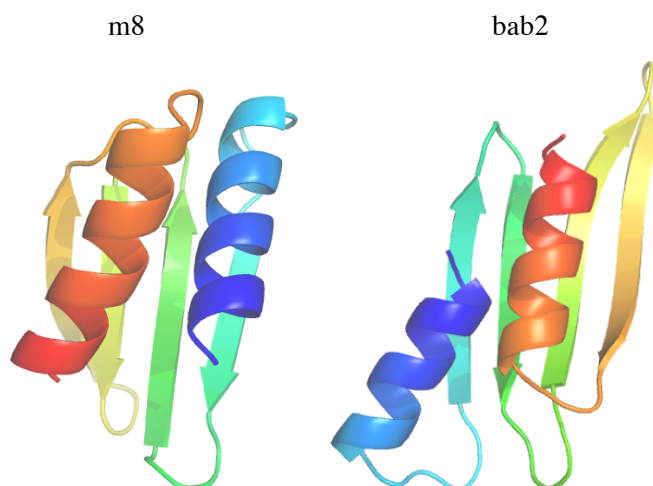
**Figure 2.3 An example that two large substructures are merged**

## Computational Design and Experimental Results

### *Design with three secondary structure substructures*

Using the previously described SEWING method (Chapter 1), I designed non-helix bundle proteins ( $\alpha+\beta$ ) using HLE, ELE, and ELH substructures (H: helix, L: loop, E: strand) (Figure 2.2). I assembled substructures only when they have at least 10 superimposable backbone atoms. However, none of the eight expressed designs (m1~8 stands for meander proteins) were stable enough for 3D structure determination (Table 2.1). Interestingly, even the disulfide bonded designs (including m4\_ss2, m8\_ss1) either aggregated before 1-D NMR or unfolded. However, at least these disulfide-bonded designs had improved melting temperatures than original non-disulfide bond designs as shown by later cooperative thermal unfolding. Therefore, these may indicate that although these designs folded into certain globular proteins, they might just lack further stability for 3-D structure determination. Therefore, I designed larger eight  $\alpha+\beta$  designs (bab 1~8 stands for bigger  $\alpha+\beta$  proteins), because larger proteins tend to be more stable (60) (61). However, most of these designs were not expressed as stable proteins either.

Design name	Residue length	Expression	Size exclusion	CD	Folding	Verdict
m1	70	Yes	Single peak	coil		Not folded
m2	68	Yes	Single peak	coil		Not folded
m3	68	Yes	Single peak	coil		Not folded
m4	68	Yes	Single peak	$\alpha+\beta$	47.6°C T <sub>m</sub>	Aggregated before 1-D NMR
m4_ss2	73	Yes	Single peak	$\alpha+\beta$	47.8°C T <sub>m</sub>	Aggregated before 1-D NMR
m5	65	Yes	Single peak			Not cleaved from SUMO by ULP1
m6	65	Yes	Single peak	coil		Not folded
m7	71	Yes	Single peak	coil		Not folded
m8	71	Yes	Single peak	$\alpha+\beta$	37.4°C T <sub>m</sub>	Aggregated before 1-D NMR
m8_ss1	76	Yes	Single peak	coil		
bab1	81	Yes (little)				Expression yield is too low
bab2	72	Yes (little)				Expression yield is too low
bab3	76	Yes	Single peak	coil	No cooperative unfolding	Not folded
bab4	74	No				No expression
bab5	78	Yes	Soluble aggregation			Aggregated
bab6	80	Yes				Not cleaved from SUMO by ULP1
bab7	78	Yes				Not cleaved from SUMO by ULP1
bab8	75	No				No expression



**Table 2.1 Experimental results of  $\alpha+\beta$  proteins using conventional small substructures (smotif).**

### *Design with five secondary structure substructures*

Nature often has important backbone geometrical features that aren't typically considered in *de novo* design. Therefore, to utilize native geometric backbone features more, we extracted larger substructures. Instead of conventional three secondary structure substructures, we extracted five secondary structure substructures (e.g. HLELE, ELELE, ELELH where H is helix, L is loop, and E is strand) and assembled them with *monte carlo* backbone assembly within the SEWING protocol (Figure 2.3). Unlike the previous approach, I assembled substructures only when they had at least 16 superimposable backbone atoms to favor capturing more designable backbones. Among the eight  $\alpha+\beta$  designs, two designs (ab1, ab2) seemed to be well folded and cooperatively unfolded according to CD and HSQC results (Figures S2.1, S2.2). However, they were multimeric (Table 2.2) and did not express enough for 3D structure determination even with increased level of glucose in LB media and auto-induction media. In addition to  $\alpha+\beta$  proteins, I also designed  $\alpha/\beta$  proteins using five secondary structure substructures (e.g. ELHLE). These eight  $\alpha/\beta$  designs were named en1~8 which stand for enumerated SEWING assembly of substructures. Backbones of these  $\alpha/\beta$  were generated after exhaustive enumerations of five secondary structure based substructures (62), because regular *monte carlo* backbone assembly in SEWING protocol did not generate diverse designable backbones due to limited five secondary structure based substructures in PDB (34). Among the eight  $\alpha/\beta$  designs, one design (en8) was expressed to the highest yields and folded quite stably according to HSQC (Table 2.3) (Figures 2.4, 2.5 and 2.6). The low expression yield of ab1, ab2, ab6, en4, en6 may be due to computational design failure or unexpected toxicity of these proteins in *E. coli* or unexpected ribosomal interaction or other complexities of the bacteriums' biology (63).

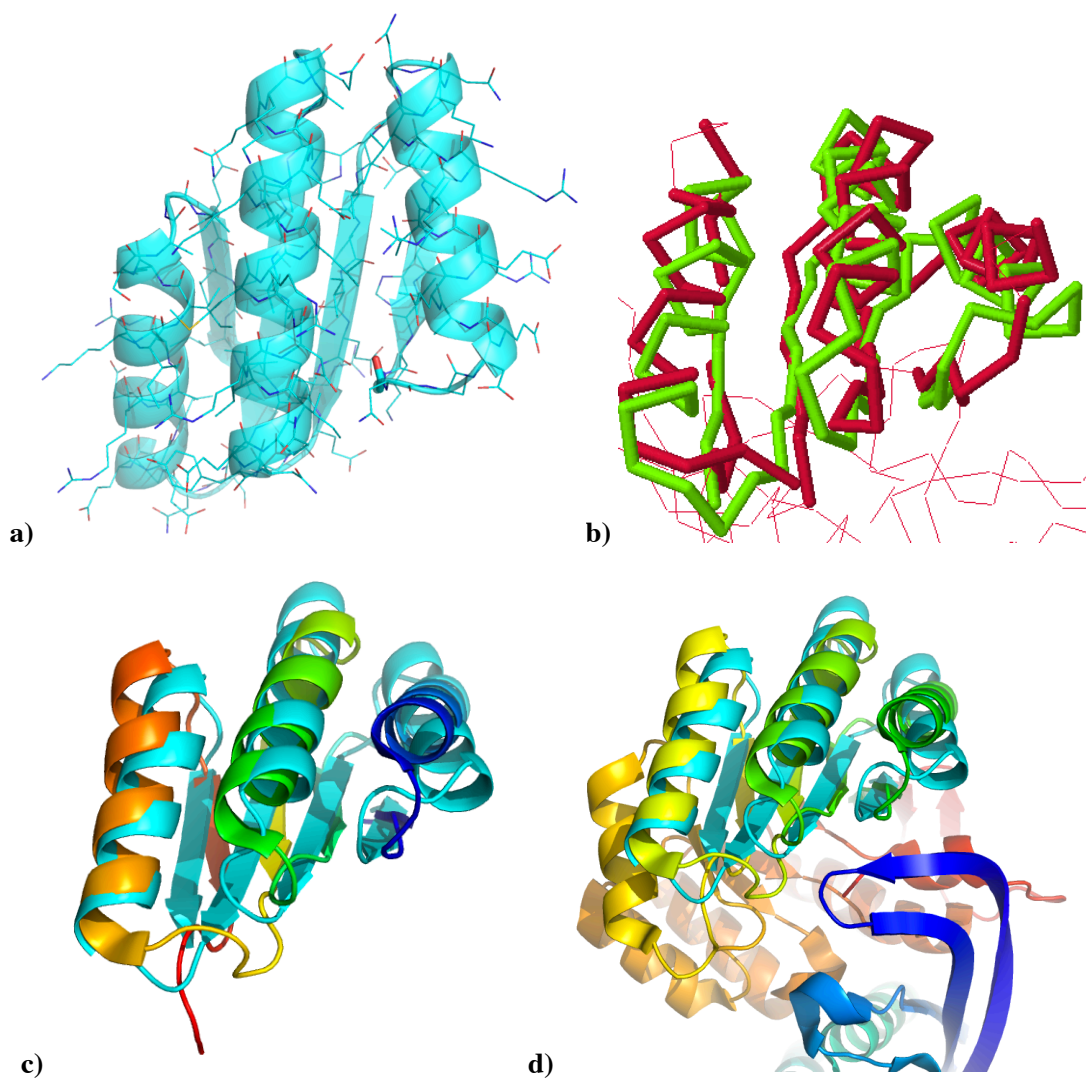
Design name	Protein length	Expression	Size exclusion chromatography	CD	Folding	Verdict
ab1	74	Yes	Single peak	$\alpha+\beta$	Cooperative unfolding	Well-folded (CD). However it tetramerized (MALS), and not well expressed for 3-D structure determination.
ab2	74	Yes	Single peak	$\alpha+\beta$	Cooperative unfolding	Well-folded (CD, HSQC). However it dimerized (MALS, NMR), and not well expressed for 3-D structure determination
ab3	74	Yes				Not cleaved from SUMO by ULP1
ab4	74	Yes				Not cleaved from SUMO by ULP1
ab5	74	Yes				Not cleaved from SUMO by ULP1
ab6	74	Yes				Expression yield is too low
ab7	74	Yes	Soluble aggregation			Aggregated
ab8	74	Yes	Single peak	Coil		Unfolded

**Table 2.2 Experimental results of  $\alpha+\beta$  proteins using largesubstructures.**

Design name	Protein length	Expression	Size exclusion chromatography	CD	Folding	Verdict
en1	75	Yes	Single peak	$\alpha/\beta$	No cooperative unfolding	Unfolded (1-D NMR)
en2	75	Yes	Single peak	$\alpha/\beta$	No cooperative unfolding	Unfolded (1-D NMR)
en3	75	Yes	Single peak	$\alpha/\beta$	No cooperative unfolding	Unfolded (1-D NMR)
en4	75	Yes	Single peak			Not that well expressed
en5	75	Yes	Single peak	$\alpha/\beta$	No cooperative unfolding	Unfolded (1-D NMR)
en6	82	Yes	Single peak			Not that well expressed
en7	83	Yes	Single peak	$\alpha/\beta$	Cooperative unfolding	Unfolded (1-D NMR)
en8	83	Yes	Single peak	$\alpha/\beta$	Cooperative unfolding	Well-folded except 1-20 residue region (CD, NMR). Dimer (MALS, NMR)

**Table 2.3 Experimental results of  $\alpha/\beta$  proteins using large substructures.**





**Figure 2.4 en8 structure and its closest structure in nature.**

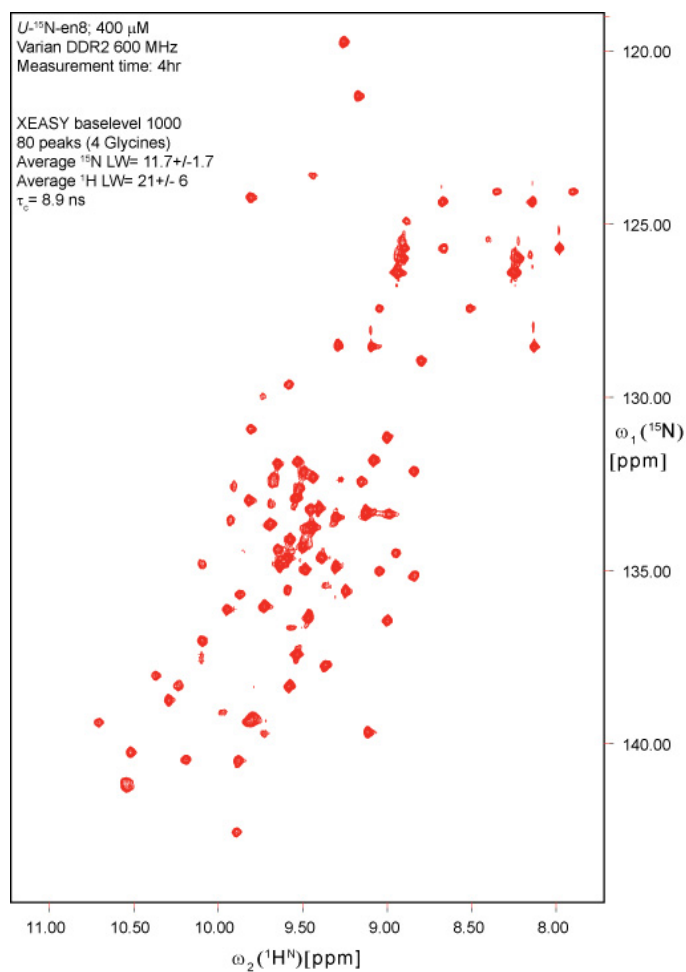
a) Designed structure of en8, red stick representation shows backbone oxygen atom of glycine at the 6<sup>th</sup> position, which could lack a satisfied hydrogen bond although it may be buried.

b) en8 (green) and its structurally closest structure (red) in nature (pdb code: 3der, chain A, 2.3 Å rmsd with en8 and 8.4 Z-score) according to Dali (36).

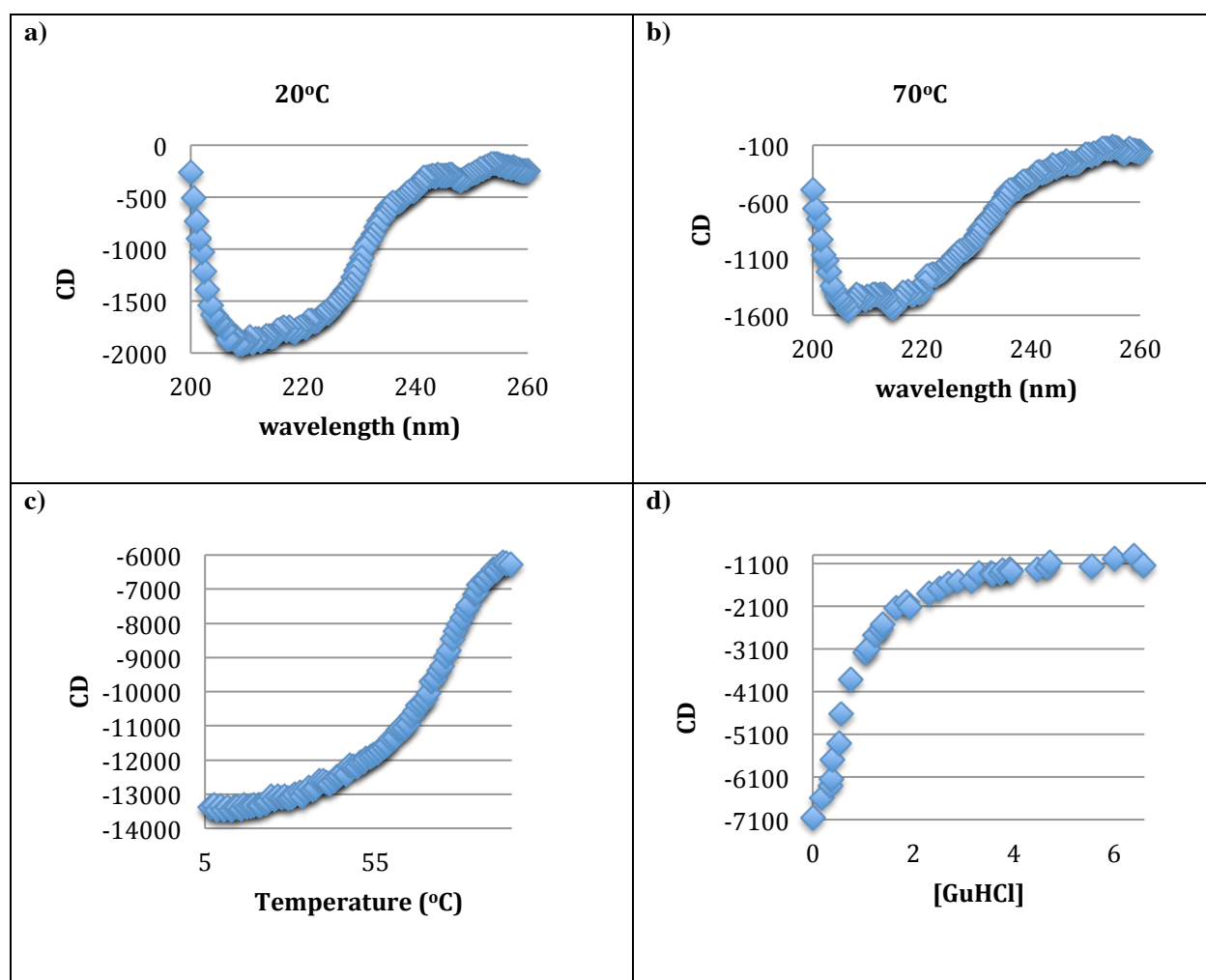
c) en8 (cyan) and its structurally closest structure (rainbow).

d) en8 (cyan) and its structurally closest structure (rainbow) with its original structure show that en8 structure seems feasible because there is no too close native backbones to a corresponding structure.

Except b), all illustrations were done by pymol (64).



**Figure 2.5 HSQC of en8 protein.** Well dispersed amide peaks show well-foldedness of the protein (400 $\mu\text{M}$  protein, in pH 7.0 20mM sodium phosphate + 150mM NaCl buffer, provided by RamaKrishna Pulavarti in Szyperski group)



**Figure 2.6 CD spectra of en8 protein.**

a) A full spectrum at 20°C (after extracting buffer only values, 30μM protein, 0.1mm width cuvette).

b) A full spectrum at 70°C. All other conditions are same as a).

c) Thermal melt (20μM protein, 1mm width cuvette, started after staying 10 minutes at 5°C, ramped 2°C per minute).

d) Denaturant unfolding (30μM protein, 1mm width cuvette, GuHCl is in mol)

All CD values are mean residue ellipticity ( $\text{deg} \cdot \text{cm}^2 / \text{dmol}$ ) and were taken at pH 7.0 20mM sodium phosphate + 150mM NaCl buffer.

*Increase net charge in an effort to monomerize well-folded  $\alpha+\beta$  and  $\alpha/\beta$  proteins*

Both MALS and NMR rotational correlation time from  $^{15}\text{N}$  average spin relaxation ( $\tau_c$  of en8 is 9 nano second, usually 0.5 ns/kDa) showed that the en8 is a homodimer in various protein concentrations and buffers (Table 2.4). MALS also showed that ab1 and ab2 are 4mer and homodimer, respectively. Multimerization of proteins is often necessary for distinct biochemical and biophysical properties, which additionally regulate functions at the post-translational level (65). However, we initially designed these proteins as a monomer. Additionally, the 3D structure determination of a monomer is easier than for a multimer.

Therefore, in an effort to monomerize en8, ab1, and ab2 designs, we supercharged them using the supercharge protocol in Rosetta (supercharge means increasing net charge into either the positive end or the negative end) (20). The rationale for increasing net charge is that net charges are expected to govern overall intermolecular interactions by affecting electrostatic interactions between surface residues, as a simulation study has shown, that net charges do influence long-range orientational distribution of the water surrounding the protein surfaces (66). Therefore, we expected that high net charged proteins tend to have more unfavorable interactions between monomers, preventing unwanted non-covalent or covalent formation between monomers.

However, all supercharged proteins were still multimeric according to MALS. For example, en8<sub>-6</sub>, which was supercharged to -6 net charge from -1 net charged en8, was dimerized still. Ab1<sub>-7</sub>, which was supercharged to -7 net charge from -3 net charged ab1, was folded as 4mer as well. Because the high net charge (-6) and the high salt concentration (1M NaCl) did not monomerize the dimerized en8, and because high TMAO (1.8M) made en8 even a 5mer, we hypothesized that en8 is a domain swapped obligate dimer rather than a transient electrostatic interaction derived dimer. This is a very preliminary idea, however the failure of crystal formation for en8<sub>-6</sub> with MCSG1~4 reservoir buffers (in total, 4 plates at 20°C with various protein concentrations did not produce protein crystals) may indicate that the en8<sub>-6</sub> is still partially unfolded. Because, it is widely believed that crystallography with flexible proteins is harder than stable proteins (67).

Design name	Net charge	Length	Size exclusion chromatography	MALS		CD
				Buffer and protein concentration	Oligomerization Determination	
ab1	-3	74	Single peak	20mM Na phosphate pH 7.0, 150mM NaCl, 500 $\mu$ 0	4mer	$\alpha$ + $\beta$
ab1_-7	-7	74	Single peak	20mM Na phosphate pH 7.0, 150mM NaCl, 760 $\mu$ 0		
ab2	-2	74	Single peak	20mM Na phosphate pH 7.0, 150mM NaCl, 113 $\mu$ M	Dimer	$\alpha$ + $\beta$
en8	-1	83	Single peak	20mM Na phosphate pH 7.0, 150mM NaCl, 840 $\mu$ 0	Dimer	$\alpha$ / $\beta$
				20mM Na phosphate pH 7.0, 1M NaCl, 621 $\mu$ 0		
				20mM HEPES-KOH pH 7.4, 150mM NaCl, 1mM DTT, 651 $\mu$ 0		
				20mM Na phosphate pH 7.0, 1.8M TMAO, 651 $\mu$ 0		
en8_-6	-6	83	Single peak	20mM Na phosphate pH 7.0, 150mM NaCl, 800 $\mu$ 0	Dimer	$\alpha$ / $\beta$
en8_A11V	-1	83	Single peak	20mM Na phosphate pH 7.0, 150mM NaCl, 500 $\mu$ 0	Dimer	
en8_S8T_A11V	-1	83	Single peak	20mM Na phosphate pH 7.0, 150mM NaCl, 500 $\mu$ 0	Dimer	

**Table 2.4 Experimental results of original and supercharged  $\alpha$ / $\beta$  and  $\alpha$ + $\beta$  proteins.**

### *Reversion design in an effort to monomerize and better fold en8 design*

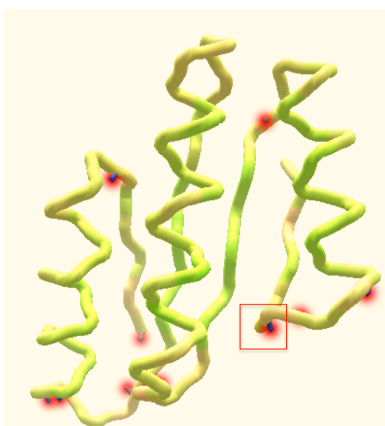
Predicted secondary structure by NMR showed that only residue numbers from 1 to 20 did not fold as we designed (Table 2.5). This partial unfolding of N-terminal region might have led to this possible domain swapping and failure of crystal formation for en8 with MCSG1~4 reservoir buffers (in total, 43 plates at 4~20°C with various protein concentrations did not produce crystals). There are possible reasons of this domain swapping. One is the possible existence of one buried unsatisfied backbone polar atom in the N-terminal region (Figure 2.1). However, it was not sure whether this backbone oxygen atom of glycine at the 6<sup>th</sup> position really lacks the needed hydrogen bond as both Kevin Houlihan's buried unsatisfied hydrogen bond detector and Foldit (68) did not identify it as a buried atom with an unsatisfied hydrogen bond (Figure 2.7). Consequently, we postulated that by introducing a few native interactions back in might prevent this domain swapping by offsetting the potential destabilizing effect of a buried unsatisfied polar atom in the N-terminal region.

Therefore, we came up with two reverted designs: en8\_A11V, and en8\_S8T\_A11V (Table 2.6). We reverted alanine to valine at the 11<sup>th</sup> position (en8\_A11V) because the presence of alanine in core or interface area is often not desired in many cases (53) (10). The reason that Rosetta chose the alanine at the 8<sup>th</sup> position in N-terminal helix for the original en8 design could be that alanine has the best helical propensity, while valine has the 7<sup>th</sup> worst helical propensity among all 20 canonical residues (69). Indeed, mutating less helical residues into alanine often improves stability of  $\alpha$ -helices (70). However, nature (the original pdb structure that SEWING used, UDP-GlcNAc 2-epimerase, pdb code: 4NES) may have used valine at the 8<sup>th</sup> position in helix to better fill the core void (Figure S2.3). The rationale for en8\_S8T\_A11V was that nature might have used threonine for the 8<sup>th</sup> position for a reason perhaps better capping the helix terminal region (Figure S2.4) in addition to a previously explained A11V. These redesigns of the N-terminal region seem plausible because the N-terminal and the C-terminal are designed as not interacting with each other. Protein sequence is known to govern protein folding, at least for a 127 residue-long protein (71). However, I suspected that if the protein's terminal regions are interacting with each other, redesign of one terminal region might make the protein fold totally differently. A partially

folded protein (“DA05”) of Jacobs (45) had N and C terminals, which are distant to each other as well, and C-terminal redesign made the DA05 to fold completely (Figure 2.8). However, both reversions (en8\_A11V and en8\_S8T\_A11V) dimerized as the original en8 did. Both reversions are currently undergoing crystal trials.

Secondary Structure	Residue # (designed)	Residue # (Taloz)	Residue # (CSI)
N-terminal loop	1	1-2	1-12
$\beta$ -strand (I)	2-5	3-4	None
$\alpha$ -helix (A)	9-20	11-16	13-15
$\beta$ -strand (II)	25-30	24-30	24-30
$\alpha$ -helix (B)	34-49	32-49	32-49
$\beta$ -strand (III)	54-59	54-59	53-59
$\alpha$ -helix (C)	63-76	63-75	64-75
$\beta$ -strand (IV)	79-82	79-82	78-82
C-terminal loop	83	83	83

**Table 2.5 Predicted Secondary Structure of en8 by NMR** (RamaKrishna Pulavarti in Szyperski lab confirmed that this en8 sample was fresh when he took NMR. Therefore en8 seems inherently partially folded).

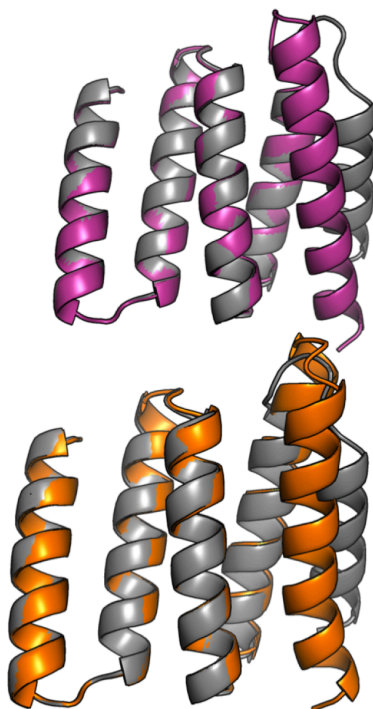


**Figure 2.7 Buried unsatisfied hydrogen bond analysis.** In this en8 designed structure, polar atoms without satisfying hydrogen bonds are represented as glowing red. One red glowing blue atom in red box represents backbone nitrogen of asparagine at the 7<sup>th</sup> position. Kevin Houlihan’s buried unsatisfied hydrogen bond detector identified residue 3, 23, and 24 have possible unsatisfied hydrogen bonds, not residue 6. The analysis was done by Foldit (68).

Design name	Normalized REU	Holes <sup>a)</sup>	Packstat <sup>a)</sup>	Folding score <sup>b)</sup>	Oligomerization <sup>c)</sup>
En8	-2.46 (-2.50 ~ -2.43)	0.69 (0.43 ~ 0.92)	0.68 (0.63 ~ 0.70)	0.72 (7,000 decoys)	Dimer (MALS)
En8_A11V	-2.48 (-2.49 ~ -2.47)	0.86 (0.76 ~ 1.02)	0.68 (0.65 ~ 0.71)	0.61 (5,000 decoys)	
En8_S8T_A11V	-2.47 (-2.49 ~ -2.45)	0.71 (0.37 ~ 1.01)	0.67 (0.62 ~ 0.71)	0.75 (5,000 decoys)	

**Table 2.6 Reverted and original  $\alpha/\beta$  and  $\alpha+\beta$  proteins.** Normalized REU is the total Rosetta score divided by total number of residues. All scores were calculated with `talaris2014_cart` score function after being relaxed. Each value of normalized REU, holes and packstat show mean (minimum ~ maximum).

- a) Both holes and packstat show how well a designed structure is packed. Lower value of holes or higher value of packstat means better packed.
- b) Folding score shows how forward-folding is funnel like (Equation S2.1). Essentially this represents how designed sequences are predicted to have designed secondary structures. Higher value of folding score means better funnel like (higher likelihood that designed protein will actually fold as designed).
- c) All three proteins were eluted as single peak from size-exclusion and went through MALS at 20mM Na phosphate pH 7.0 150mM NaCl as 500 micro M.

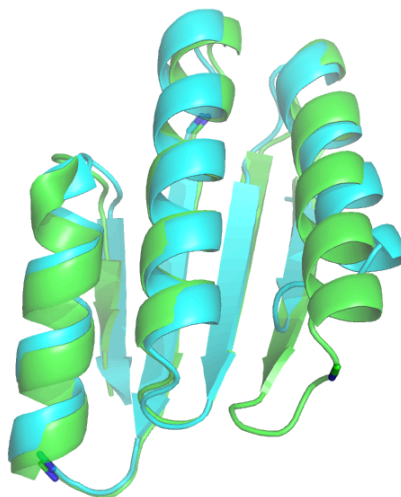


**Figure 2.8 An example of C-terminal redesign.** Design models of DA05R1 (purple) and DA05R2 (orange) superimposed on the original DA05 design model (gray). Because Jacobs et al. redesigned C-terminal region which is distant from N-terminal region, these redesigns may have been easier than the redesign of proteins whose C and N-terminal are near each other (58).



### *Redesign with refolding in an effort to monomerize and better fold en8 design*

Because the two reverted redesigns (en8\_A11V and en8\_S8T\_A11V) are still dimers, we completely redesigned the first 24 n-terminal residues by traditional *ab initio* refolding. Four redesigns (en8\_re01, en8\_re02, en8\_re03, en8\_re04) that are believed to satisfy three redesign goals are chosen for experimental validation (Figure 2.9). The first goal was to redesign the n-terminal  $\alpha$ -helix to be nearer to the next  $\alpha$ -helix because we believed that more closely packed secondary structures would be easier to design and more commonly found in nature. The second goal was to lengthen the n-terminal  $\beta$ -strand so that it has more hydrogen-bonds between it and the next  $\beta$ -strand. For example, the original en8 design had five backbone-backbone hydrogen bonds between the n-terminal  $\beta$ -strand and the next  $\beta$ -strand. On the other hand, refolded redesigns have seven to eight hydrogen bonds between these two  $\beta$ -strands. The third goal of this redesign was to remove possible polar atoms with unsatisfied hydrogen bond (Figure 2.7). Specifically, we tried to design backbone polar atoms to not face toward core of proteins as in the first successful *de novo* design of a globular protein (33). However, en8\_re01, en8\_re02, and en8\_re04 were dimerized again (en8\_re03 was not expressed enough). These are strong indications that an interface for this dimer is not likely the N-terminal region. Therefore, redesign of C-terminal region to have disulfide bond is underway. Having a disulfide bond in c-terminal such as between C-terminal beta-strand and C-terminal alpha-helix will likely lock up these two secondary structures to prevent possible domain swapping. Therefore, four redesigns of en8\_re04 to have a disulfide bond were done and will be experimentally characterized. There are other ideas of preventing domain swapping such as having glycine or proline in loop, or a shorter loop (72) (73) (74). However, en8 already has glycine in its loop that connects the C-terminal beta-strand and C-terminal alpha-helix. It is not certain whether the existence of proline in the loop helps to prevent or induce the domain swapping. The en8 has already short loops.

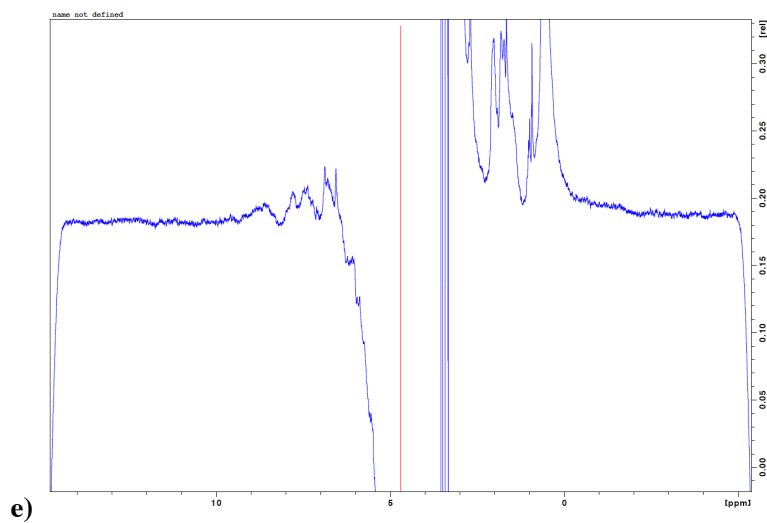
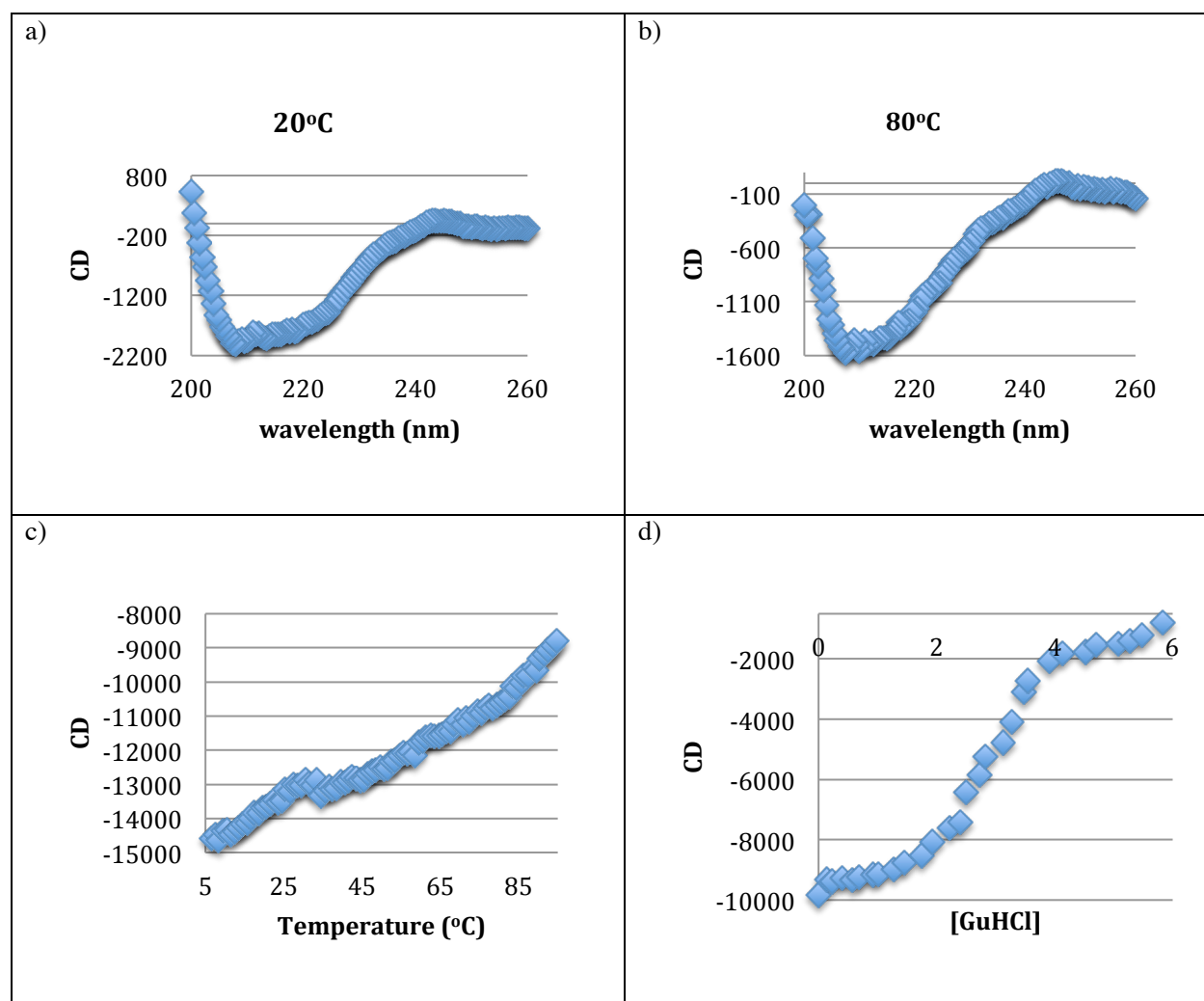


**Figure 2.9** The original en8 structure (in cyan) and one of the refolded redesigns (en8\_re01 in green).

## Conclusion

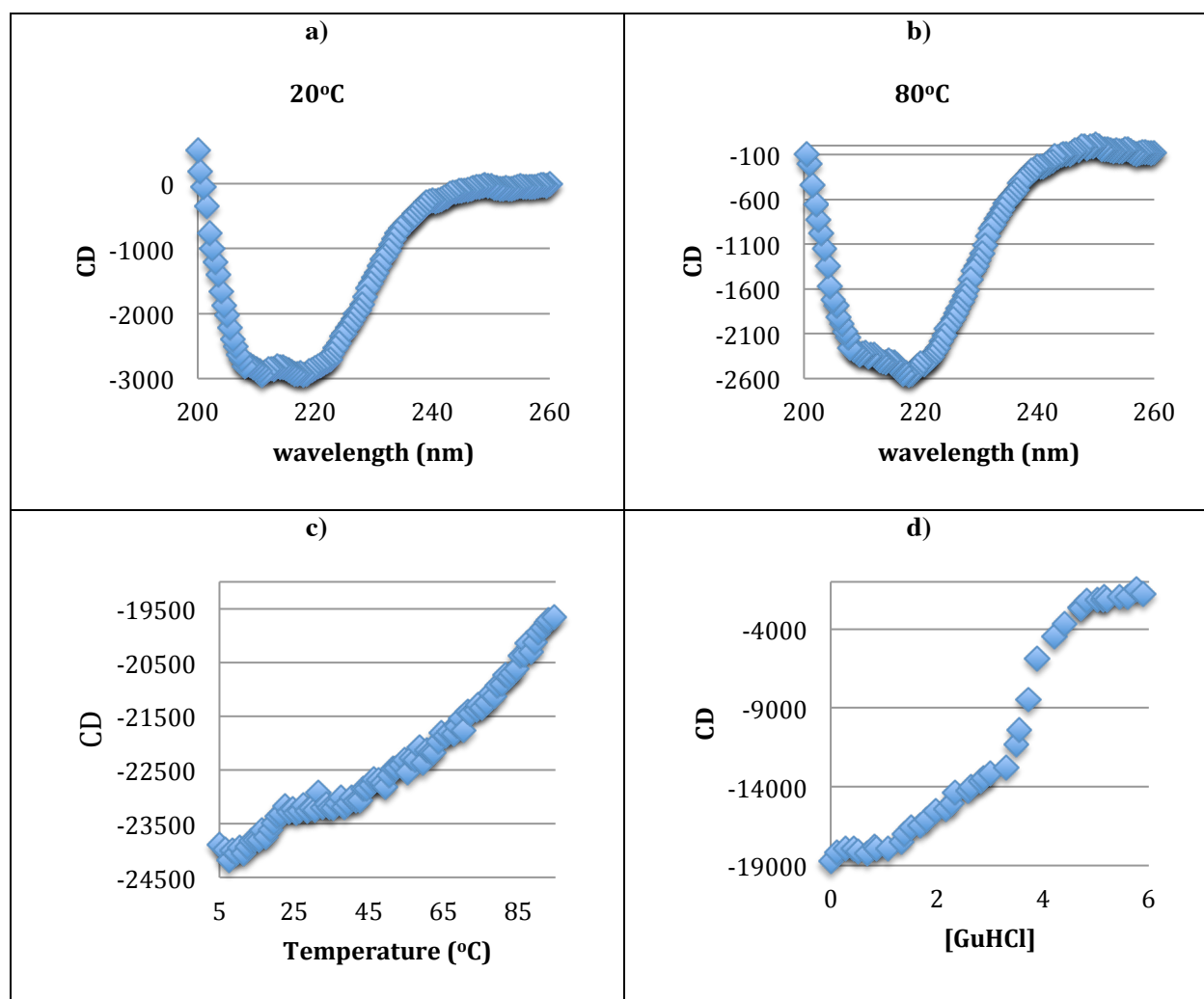
We have designed  $\alpha/\beta$  and  $\alpha+\beta$  proteins using supersecondary structures. One of the designs, en8, was expressed to high levels and mostly well folded. This is a significant improvement of computational protein design because it is the first proof that we can *de novo* design small globular non-repeat  $\alpha/\beta$  proteins whose N- and C-terminals are located at the edges (while, the sequence repetition of repeat proteins probably favor designed structures with internal repeats over alternative structures (63)). Additionally, this type of simple topology may serve as a nucleic acid binder (67). However, the original en8 design protein was a dimer and its N-terminal region (1~20 residues region) was partially unfolded. I have redesigned en8 into seven different variants in an effort to make it more well-folded. However, all of them are either dimerized or not expressed enough. Four new designs of en8\_re04 that have a disulfide bond will be experimentally characterized.

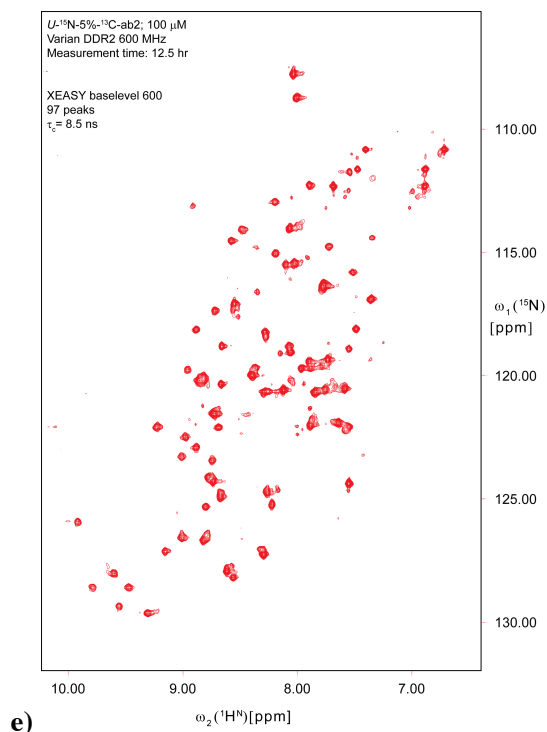
## Supporting Materials



**Figure S2.1 Biophysical characterization of ab1 protein.**

- a) A full spectrum at 20°C (after extracting buffer only values, 30μM protein, 0.1mm width cuvette).
- b) A full spectrum at 80°C. All other conditions are same as a).
- c) Thermal melt (15μM protein, 1mm width cuvette, started after staying 10 minutes at 5°C, ramped 2°C per minute).
- d) Denaturant unfolding (15μM protein, 1mm width cuvette, GuHCl is in mol), All CD values are mean residue ellipticity (deg\*cm<sup>2</sup>/dmol) and were taken at pH 7.0 20mM sodium phosphate + 150mM NaCl buffer.
- e) 1D-NMR with 31μM of ab1 protein with 64 scans at 25°C. The verdict whether the protein is well folded is inconclusive because of low sensitivity.



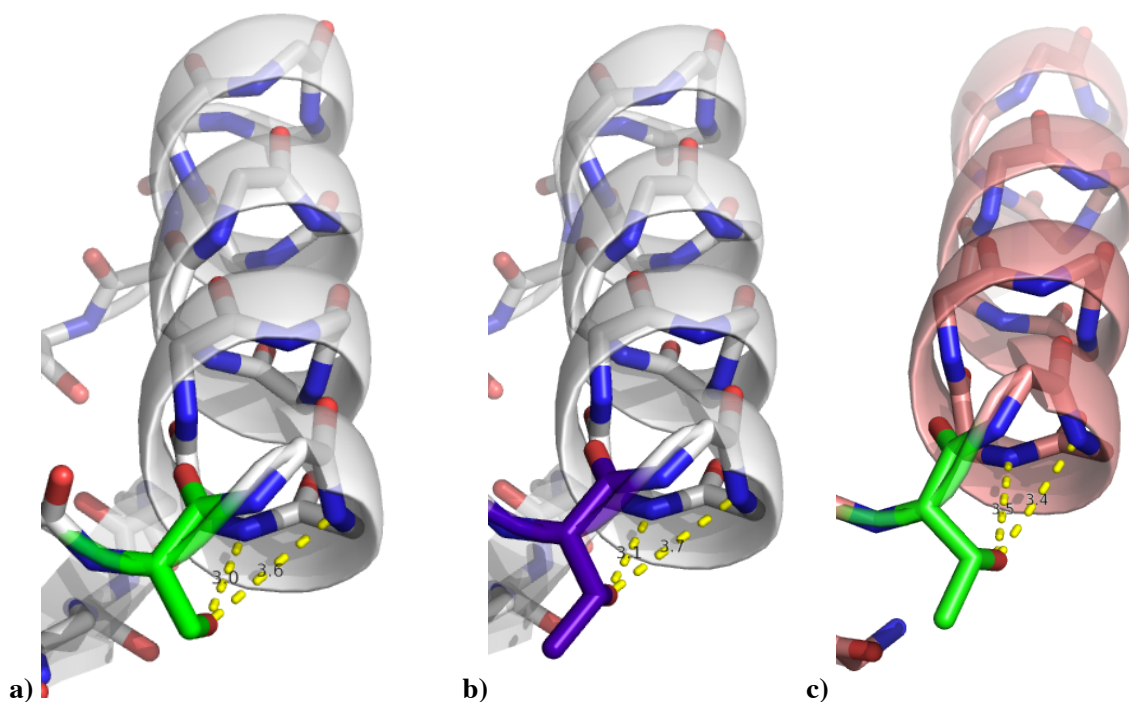


**Figure S2.2 Biophysical characterization of ab2 protein.**

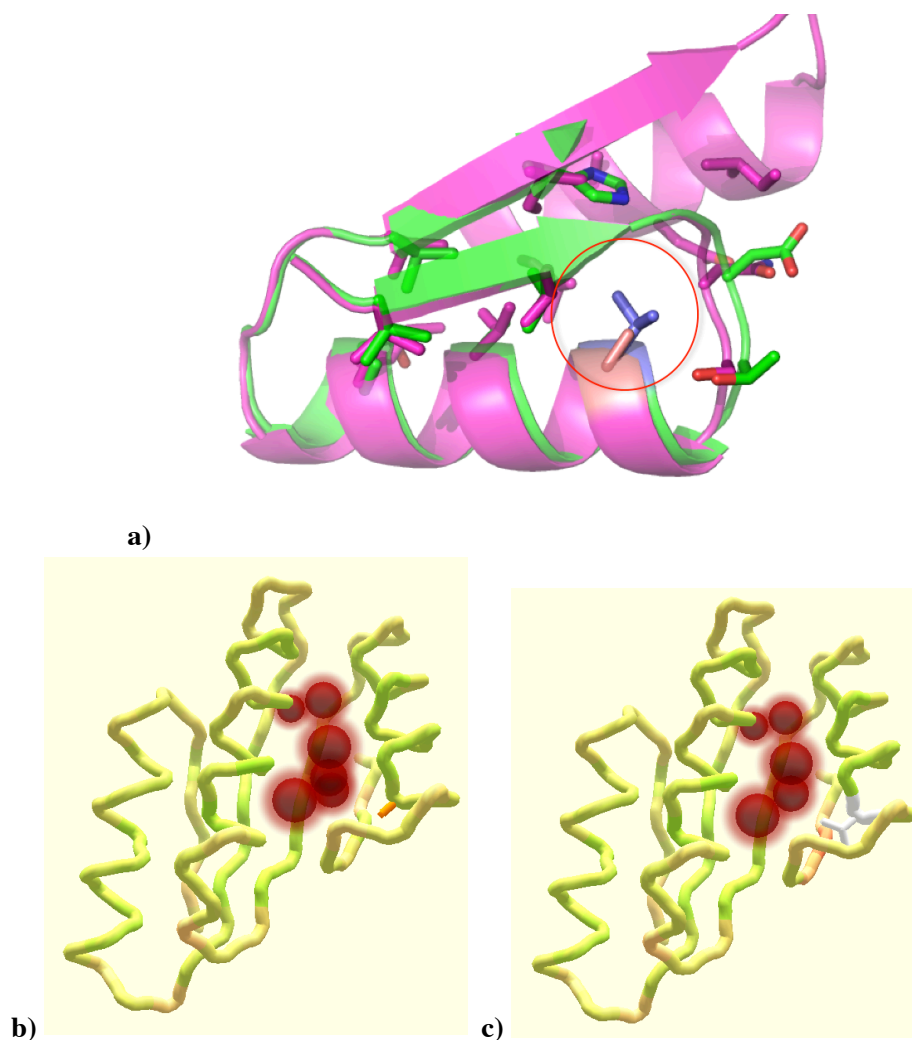
- a) A full spectrum at 20°C (after extracting buffer only values, 30μM protein, 0.1mm width cuvette).
- b) A full spectrum at 80°C. All other conditions are same as a).
- c) Thermal melt (15μM protein, 1mm width cuvette, started after staying 10 minutes at 5°C, ramped 2°C per minute).
- d) Denaturant unfolding (15μM protein, 1mm width cuvette, GuHCl is in mol), All CD values are mean residue ellipticity (deg\*cm<sup>2</sup>/dmol) and were taken at pH 7.0 20mM sodium phosphate + 150mM NaCl buffer.
- e) HSQC with 100μM of ab2 protein. The minor peaks seem to arise from minor contaminant (SUMO) rather than sample's heterogeneity (provided by RamaKrishna Pulavarti in Szyperski lab).

$$P(Near) = \frac{\sum_{i=1}^N \left( e^{\frac{-RMSD_i^2}{\lambda^2}} + e^{\frac{-E_i}{k_b T}} \right)}{\sum_{j=1}^N e^{\frac{-E_j}{k_b T}}}$$

**Equation S2.1 Folding score equation.** Value of  $\lambda$  is used as 2,  $k_b$  is Boltzmann constant,  $E_i$  is total Rosetta energy with  $i^{th}$  decoy (provided by Tom Linsky and Vikram Mulligan in Baker lab)



**Figure S2.3 Differences of helix capping between side-chain oxygen of the 8<sup>th</sup> position and backbone nitrogens of 10E and 11A.** a) original en8 design with serine at the 8<sup>th</sup> position, b) en8\_S8T design, c) native backbone that was used as substructure (node) during SEWING backbone assembly



**Figure S2.4 Analysis of void in core region.**

a) Native backbone for en8 design in green, 11V in native backbone in blue, designed backbone for en8 in magenta, 11A in designed backbone in salmon.

b) Void represented with red spheres in en8 design.

c) Void represented with red spheres in en8\_A11V design. A11V reversion in white shows reduced void volume in core region. Void was represented by foidit (75).



[protein sequences used for this de novo design for  $\alpha/\beta$  and  $\alpha+\beta$ ]

```
> m1
GDERKKKLEKKGFDVRKYEVRRNGEPKGYAVMAEKNKYWEIYVEENGQEKKETASTTEVAKRRVEKVMRL
> m2
TEEAkkAAELAKRAGDTGTEQQITLSQGREIRAWPNDDGSYVEIDTGKTTVRMPNAAEAAKQAAKAAN
> m3
TEEAKEAAKRAEEAGKKGTEMQITVSKGREFRVWPNSEGSTVQIDTGKTTYTASNAAEEAAKVAKKAVN
> m4
ESDFEEAVERAKRGEQVTYKNESNGTILEIRPTSQRFEFWRIENGEKRKKAIEVRGNNDDEMKAAREN
> m4_ss2
GCGESDFEEAVERAKRGEQVTYKNESNGTILEIRPTSQRFEFWRIENGEKRKKAIEVRGNNDDEMKAARENGC
> m5
LDEVLKRIEEIYKKGQKIAFRADVNGNELEVRDGDITIEFWLNGEKKSEVTGDDMDKVKEEMMRF
> m6
GERARKRFEEAGFETEKRGNEIYARYNGVEVKIIHDNGREETTAKMDPRDPEQQAKERAERAAQ
> m7
TEHQKEIEKVREKAIEKRGAPVETRGDTIRVEEPDGTAFYQIHGQTTRATTTTNPEEGEKKAREEAKKDSE
> m8
VRDELFEAMRAGKHPGQTFEYKDEDNNIIIRFTPEVRAFKNGEQKYTFKRDPREPEQAKAAAEDAIEKAL
> m8_ss1
CGGVRDELFEAMRAGKHPGQTFEYKDEDNNIIIRFTPEVRAFKNGEQKYTFKRDPREPEQAKAAAEDAIEKALGC
> bab1
TSEAEQEWKRKAIEKAAAREAKTKNKRVELYLTGPNNRVIRIEVWVDSNGNGQVNAAYADGQYFEVRTDPTKAIEEAFKRAMN
> bab2
VAADKAREIFEKTGSSHVTAQGTNLNGITFEVHYHPGAELRFEFRDGDVQRRRYQHSSLEEAERRAREAAK
> bab3
SDAEIEKKRLMKHPGQTIEVPVDSERRIYVESHDGRVEVRLYRNGQPERQTETSYPDGREHTWEAAKKAEYKRL
> bab4
TPKRHNEQAKEMHKNYQTKKTQQNGNTFFLFIEYRDGNETHMYIFVFRGTQREYHTKEPREFAIEKKARNL
> bab5
GDEEKERIREALEKGQEAHIKTTGGGQEILIVRHENGYRMEIRLNGEPQVERPNQSQEQLVKEAAKHAQELAERAK
> bab6
TEDDIKKAKELFKKLKEGKLQTIHARIQHDHGREIRIEARKKTDNEIEIRVWFYDGNKTEEMRFTPEAARRAEERAKS
> bab7
TEERIRDTKQEAHDKGFQTRTEKTERMNGQDYFLIEGGGIWFAFVKEDKDNNQETRFTATGSSPEEAERRARKKA
> bab8
STIDKIREKFKRNGGEEVQIRYSKGYWIFIIRKPGNVIVFVFLNGELKIHLVFDASKYDPEQARAEAEKEIEKQ
> ab1_mc_16_atoms_00210165_0189
TEAEKYATEIEKRLREKGIEARTYKKGNGIVIVAWDSTKIHVWIATEDTVKHEVTTASEEQLKELIRQFMEEAI
> ab1_-7
TEAEKYATEIEKRLREDGIEARTYKEGNGIVIVAWDSTKIHVWIATEDTVKHEVTTASEEQLKELIRQFMEEAI
> ab2_mc_16_atoms_00210165_0237
TEARKYATEIEKRLREKGIEARRIEKNGNGIVIVAFDSTKYHVYIATEDTVIHIEETTASEEQLKELLRQFMELAI
> ab2_-7
DEARKYATEIEKRLREDGIEARRIEEGNGIVIVAFDSTKYHVYIATEDTVIHIEETTASEEQLKELLRQFMELAI
> ab3_mc_16_atoms_00210165_0389
TEAEKYAEIEKRLREKKGIEARRIKKNGNGIVIVAYDSTKYHIYIATEDTVTHLETTASDEQLKELIRQAMETAI
> ab4_mc_16_atoms_00448602_0088
TYAEEMARKIMEELQKRGITATMFRSGNGIVIVTWDSTSWHFFVATSTRVEHYETTASEDQARKILEEFMKRAI
> ab5_mc_16_atoms_00448602_0252
TIAELARQIQEELDKRGITAEYRSGNGIVIVWWDSTSWHFFVATSTRVEHYETTASEDQARKLTEEYIRRAE
> ab6_mc_16_atoms_00448602_0276
TNAQEAAARKIEEELRKRGITATRYETGNGIVIVTWTSTSWHFFVATSTRVEHYETTASEDEARKLTKKYMKRAI
> ab7_mc_16_atoms_00448602_0371
TTAQEYARQIEEELRKRGITATVYETGNGIVIVHWDSTSWHFFVATETHVEHYETTASEDQARKLAKQYQREAI
> ab8_mc_16_atoms_00583666_0072
TEAQKVAEELRRRMDKNQGTGEIRVTDGEVEFRIRSGTEAHVRIENGQTTTVTKGSTKEEEKKKAIEKYREEV
> en1_144244_19_7_0114
TYEITGGDEEAAKKAEWWRGRHRTVKTMDSEFEEMRERYPEIPLKILHDDPEEARRLAEYQKKGLDVTWQP
```

> en2\_144244\_19\_7\_0117  
 TYEITGGDEEAAKKAEYWRGRHRTVKTCDMSEFEEMRERYPEIPLKVLHDDPEEARRIAEEYQKKGLDVTWQP  
 > en3\_144244\_19\_7\_0381  
 TYEITGGDEEAAKKAEYRRGRHRTVKTCDMSEFEKMRREYQITLKILHDDPEEARRLAEEYQKKGLDVTWSP  
 > en4\_144244\_19\_7\_0442  
 TYEITGGDEEAQKKAEEWWRGRHRTVKTTRMEFEAARERYQIPLKILHDDSEEARRLAEEYRKKGLDVTWST  
 > en5\_144244\_19\_7\_0455  
 EYEITGGDEEAAKKAEYWRGRHRTVKTCDMSEFEEMRERYPEIRLKILHDDPEEARRLAEEYQKKGLDVTWQP  
 > en6\_178065\_135\_428\_0354  
 TVRLRGRNLAEIVEKLARNGEKVITIEWKGTDESQRKIIIEAIIKRAAKHGGELEVEIKVTNEDEQRKMKKWASTADTQVRFKP  
 > en7\_178065\_327\_229\_0072  
 TVHVKGHSEEAERRIKEAIDRNLHVSIEIEGYNEERLRWAMKMAKEAQKKGAPVRVRIKNSDPEKLERARKIIESAGAEVEIT  
 > en8\_178065\_327\_229\_0254  
 TVHVKGNSDEAEERVRRRAIKNNQHVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRNSNPEELERARKIIESAGAEVEIT  
 > en8\_-6  
 TVHVKGNSDEAEERVRRRAIDDDQHVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRNDNPEELERARKIIESAGAEVEIT  
 > en8\_A11V  
 TVHVKGNSDEVEERVRRRAIKNNQHVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRNSNPEELERARKIIESAGAEVEIT  
 > en8\_S8T\_A11V  
 TVHVKGNTDEVEERVRRRAIKNNQHVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRNSNPEELERARKIIESAGAEVEIT  
 > en8\_re01\_S\_00000034\_wo\_angle\_cst\_0025  
 SRMTAKVTQNTPEEVKKAMDMLRKAANKNNMEVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRNSNPEELERARKIIESAGAEVEIT  
 > en8\_re02\_S\_00004588\_wo\_angle\_cst\_0019  
 KQKQVTVSDSQPPEISKEMAKFVQTAQKQKLSVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRNSNPEELERARKIIESAGAEVEIT  
 > en8\_re03\_S\_00017243\_0030  
 KEVKVTVSKDDPTEKVRKAFKAKRAASNKYMVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRNSNPEELERARKIIESAGAEVEIT  
 > en8\_re04\_S\_00018189\_0011  
 DRVTVTVSANTQPEHVKTAMDIAAEAAKNKLEVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRNSNPEELERARKIIESAGAEVEIT  
 > en8\_re04\_68C\_GGGC  
 DRVTVTVSANTQPEHVKTAMDIAAEAAKNKLEVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRCSNPEELERARKIIESAGAEVEITGGGC  
 > en8\_re04\_69C\_GGGC  
 DRVTVTVSANTQPEHVKTAMDIAAEAAKNKLEVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRNCNPEELERARKIIESAGAEVEITGGGC  
 > en8\_re04\_70C\_GGGC  
 DRVTVTVSANTQPEHVKTAMDIAAEAAKNKLEVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRNSCPEELERARKIIESAGAEVEITGGGC  
 > en8\_re04\_88C\_82C  
 DRVTVTVSANTQPEHVKTAMDIAAEAAKNKLEVKIEIEGYNEQILRDAMRLAKEAQKQGAPVRVEIRNSNPEELERARKIIESAGAEVEIT

## CHAPTER 3: BOOSTING STABILITY OF $\beta$ -SHEET PROTEINS BY SURFACE REDESIGN<sup>1</sup>

### Overview

$\beta$ -sheets often have one face packed against the core of the protein and the other facing solvent. Mutational studies have indicated that the solvent-facing residues can contribute significantly to protein stability, and that the preferred amino acid at each sequence position is dependent on the precise structure of the protein backbone and the identity of the neighboring amino acids. This suggests that the most advantageous methods for designing  $\beta$ -sheet surfaces will be approaches that take into account the multiple energetic factors at play including side chain rotamer preferences, van der Waals forces, electrostatics, and desolvation effects. Here, we show that the protein design software Rosetta, which models these energetic factors, can be used to dramatically increase protein stability by optimizing interactions on the surfaces of small  $\beta$ -sheet proteins. Two design variants of the  $\beta$ -sandwich protein from tenascin were made with 7 and 14 mutations respectively on its  $\beta$ -sheet surfaces. These changes raised the thermal midpoint for unfolding from 45°C to 64°C and 74°C. Additionally, we tested an empirical approach based on increasing the number of potential salt bridges on the surfaces of the  $\beta$ -sheets. This was not a robust strategy for increasing stability, as three of the four variants tested were unfolded.

---

<sup>1</sup> This chapter previously appeared as an article in the Protein Science. The original citation is as follows: D. N. Kim, T. M. Jacobs, B. Kuhlman, "Boosting protein stability with the computational design of  $\beta$ -sheet surfaces". Protein Sci. 25, 702–710 (2016) (76).

## Introduction

Approximately one quarter of all known protein structures are comprised almost exclusively of  $\beta$ -strands and connecting loops (77). These proteins often adopt  $\beta$ -sandwich or  $\beta$ -barrel folds in which it is common for one face of a  $\beta$ -sheet to point towards the hydrophobic core of the protein while the other face points towards solvent. As would be expected, the core facing residues play a critical role in determining protein stability as they form tight van der Waals and hydrogen bonding interactions with other residues in the protein. However, the solvent-facing residues can also play a strong role in dictating protein stability, as they frequently form specific interactions with residues from neighboring  $\beta$ -strands as well as nearby residues on the same  $\beta$ -strand. For this reason, there has been considerable effort aimed at understanding the sequence and structure features that contribute to  $\beta$ -sheet stability (19) (20) (78) (79).

Mutagenesis studies and statistical analyses of naturally occurring  $\beta$ -sheets have shown that some amino acids have a greater intrinsic propensity to adopt  $\beta$ -strands. The  $\beta$ -branched amino acids (Ile, Val, and Thr) and aromatic residues are overrepresented in  $\beta$ -strands, while the charged amino acids (Arg, Lys, Glu, and Asp) and turn residues (Gly and Pro) are underrepresented. Similar studies have also examined the preferences for various amino acids to be placed near each other on adjacent  $\beta$ -strands (80) (81). Two of most favored pairings are aromatic pairs and the formation of salt bridges using aspartate or glutamate paired with arginine or lysine. These preferences have been used widely to design and stabilize model  $\beta$ -hairpins and  $\beta$ -sheets constructed from synthetic peptides (82) (83), but there have been relatively few studies that have focused on using these principles for the large-scale redesign of  $\beta$ -sheets that are incorporated in folded proteins.

An important feature of  $\beta$ -sheets in well-folded proteins is that they are fairly rigid, and each residue in the sheet has a unique set of phi and psi angles as well as a unique set of neighbors, each with distinct geometries that dictate which direction side chains will be projected. An important consequence of this variability is that although there are general preferences for particular amino acids and amino acid pairs to stabilize  $\beta$ -sheets, the preferred amino acid at a specific residue position depends strongly on the precise structure surrounding that residue (81). This complexity and diversity suggests that the most

advantageous methods for designing  $\beta$ -sheets will be approaches that take into account the multiple factors that contribute to stability including: side chain rotamer preferences, van der Waals interactions, hydrogen bonding, desolvation effects, and electrostatics (84).

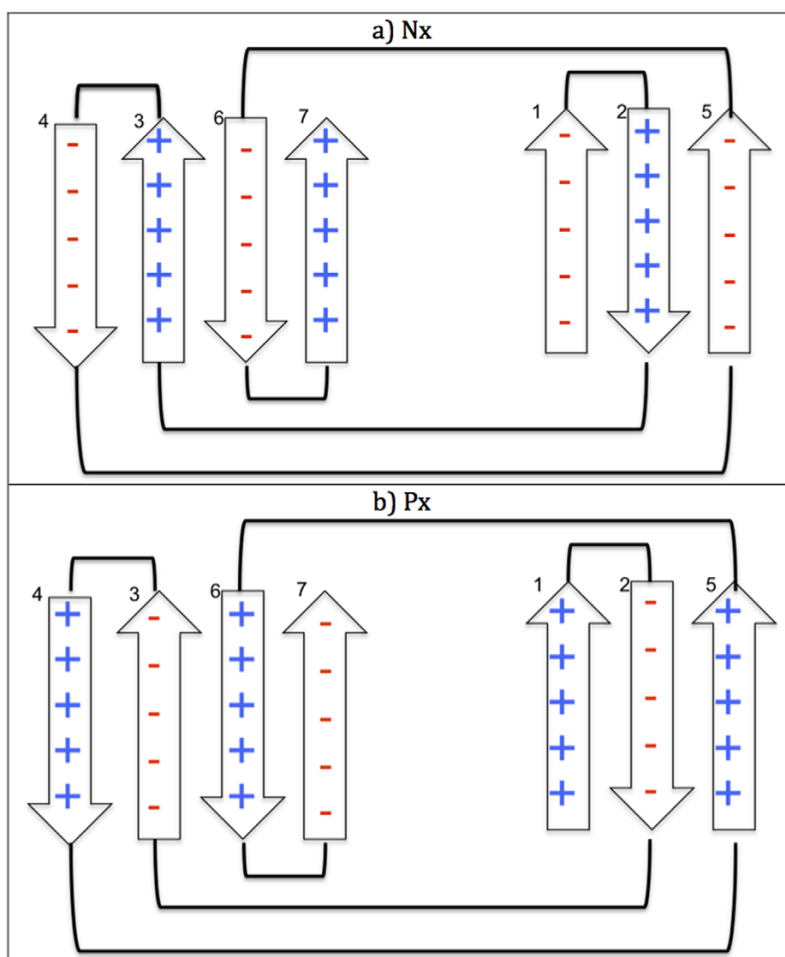
Over the last 20 years, methods for computational protein design have emerged as a powerful approach for optimizing sequences based on multicomponent energy functions. These protocols have been used to stabilize proteins, design new protein structures and interactions, and more recently create large macromolecular assemblies (85) (33) (86). In these studies,  $\beta$ -sheet surfaces have been designed in the context of larger goals, but there have been few studies that have specifically probed how effective these approaches are at designing  $\beta$ -sheet surfaces. For instance, is it possible to dramatically stabilize naturally occurring proteins by just redesigning their  $\beta$ -sheet surfaces? Mayo and coworkers optimized an energy function for the design of  $\beta$ -sheet surfaces and tested the protocol on the redesign of  $\beta$ -sheets from two proteins, in one case there was a modest decrease in protein stability and in the other case the melting temperature increased by 8°C (87). In this study, we used the molecular modeling program Rosetta to redesign  $\beta$ -sheet surfaces of the fibronectin type III domain of the protein tenascin (TNfn3, pdb code: 1ten (88)). This original TNfn3 structure starts with residue numbers 802R and 803L, while it starts with 1L in this chapter. Rosetta omits the first Arg residue due to its partial definition. Rosetta's default management of protein structure uses "pose" that rearranges the residue number to start with 1).

TNfn3 forms a Greek key fold with three  $\beta$ -strands in one sheet and four  $\beta$ -strands in a second sheet. It has been studied extensively as a model system for protein folding and stability (89) (90) and previous studies have demonstrated that its stability can be improved via mutation. In most cases, the stabilizing mutations have been located in the protein core, or the redesigns included a mixture of mutations from various regions of the protein (91) (92) (93).

Unlike the Mayo study, we did not employ an energy function and modeling protocol specifically created for  $\beta$ -sheet surfaces, but rather used the all atom energy function in Rosetta, which has been parameterized with a diverse set of sequence design and structure prediction tests (94) (95). The primary components of the energy function are a damped Lennard-Jones potential that models dispersion forces

and steric repulsion, an implicit solvation model that penalizes the burial of polar groups, an orientation-dependent hydrogen bonding term that has been parameterized to be used with damped Coulomb electrostatics, and knowledge-based terms that score dihedral preferences and the intrinsic preferences of the amino acids to be in alternative secondary structures. The Coulomb electrostatics term is a more recent addition to the Rosetta force field that has been benchmarked computationally (95), but few experimental tests have been performed with it.

In addition to designing  $\beta$ -sheet surfaces with Rosetta, we also tested an empirical approach based on increasing the number of salt bridges (glutamate or aspartate paired with lysine or arginine) between strands on the surface of the  $\beta$ -sheets. This approach was inspired by previous studies that demonstrated that arrays of salt bridges could be used to favor the formation of heterodimeric over homodimeric coiled-coils (96). Charge repulsion between like charged groups disfavored homodimers while charge attraction favored the heterodimers. A significant challenge in the design of  $\beta$ -sheet proteins is how to specify which  $\beta$ -strands will pair with each other. This is especially problematic for tertiary folds in which strands distant in primary sequence are paired in the final folded structure. Kinetically, it is more straightforward for strands close in primary sequence to pair, and many structure prediction algorithms suffer from predicting too many local contacts when performing *ab initio* structure prediction on  $\beta$ -sheet proteins (97). TNfn3 is an excellent example of a protein with a topology that is difficult for design and prediction and contains  $\beta$ -strand contacts distant in primary sequence; it includes strand pairing between the third and sixth  $\beta$ -strands as well as the second and fifth  $\beta$ -strands. Interestingly, we observed that through mutation it is possible to place charged residues on TNfn3 in such a way that every  $\beta$ -strand has the opposite charge of the  $\beta$ -strands that are paired with it, and that  $\beta$ -strands that are close in primary sequence, but are not paired in the final structure, end up with the same charge (Figure 3.1). We reasoned that this arrangement of charges should favor the folding and stability of the protein by creating favorable electrostatic interactions in the folded state, while simultaneously disfavoring kinetically accessible misfolded states.



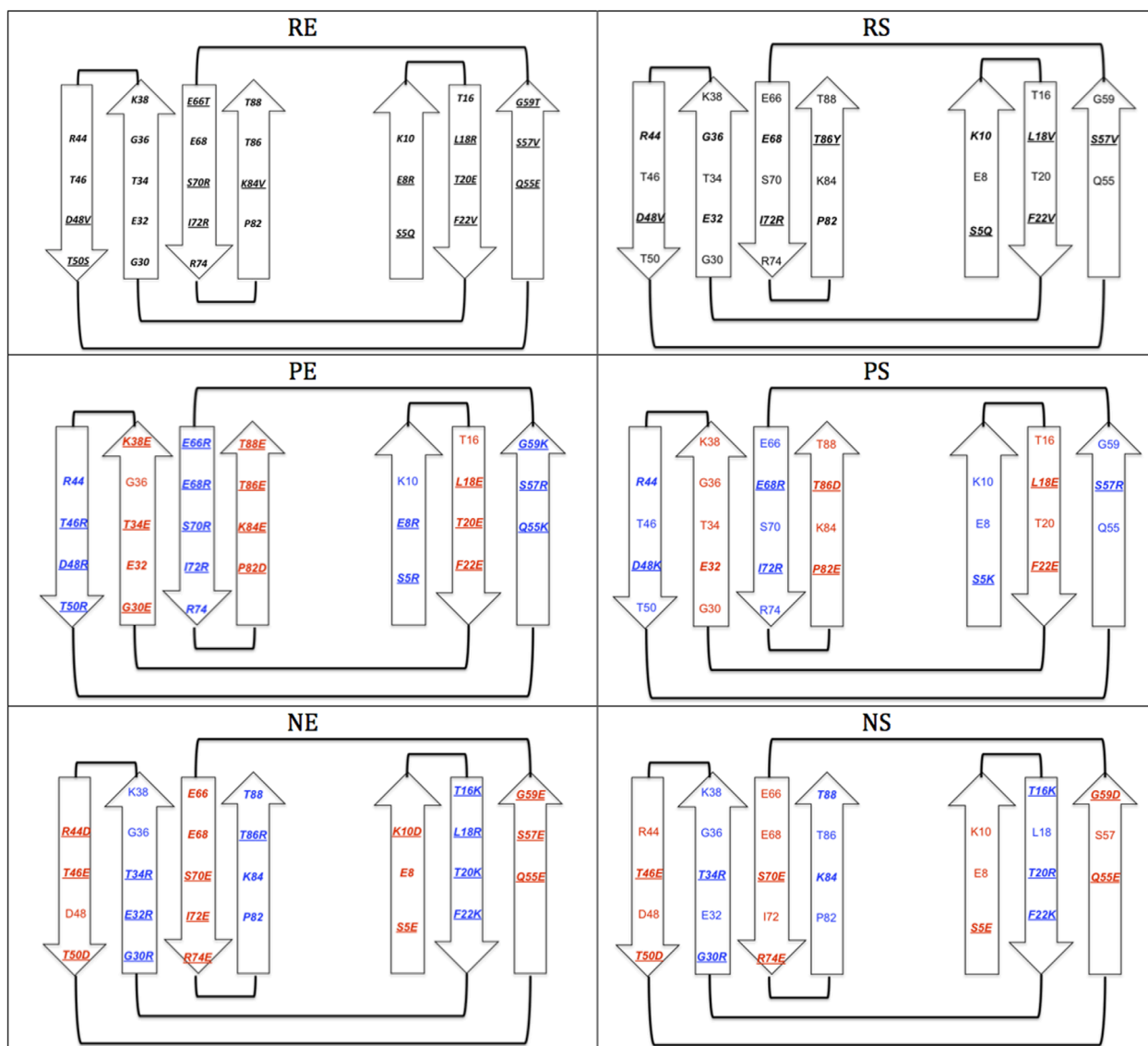
**Figure 3.1 Concept of the charge zipper scheme for the TNfn3  $\beta$ -sandwich fold.** By mutating residues on the surface exposed faces of the two  $\beta$ -sheets it is possible to create a scenario where every strand is paired with a strand of opposite sign in three-dimensional space, but strands that are close in primary sequence, but are not paired, have the same charge. (a) A charge zipper that starts with a negatively charged  $\beta$ -strand. (b) A charge zipper that starts with a positively charged  $\beta$ -strand.

Both mutational studies and statistical analyses of  $\beta$ -sheet sequences indicate that there is a strong energetic bonus for placing lysines and arginines across from glutamates or aspartates in  $\beta$ -sheets, while there is an energetic penalty for placing like charged amino acids near each other (80) (98). However, charged residues also have lower intrinsic preferences for adopting  $\beta$ -strands (81) (99) (100). This suggests that although charge patterning may stabilize the desired pair interactions, the new charged residues may also disfavor  $\beta$ -strand formation.

## Results

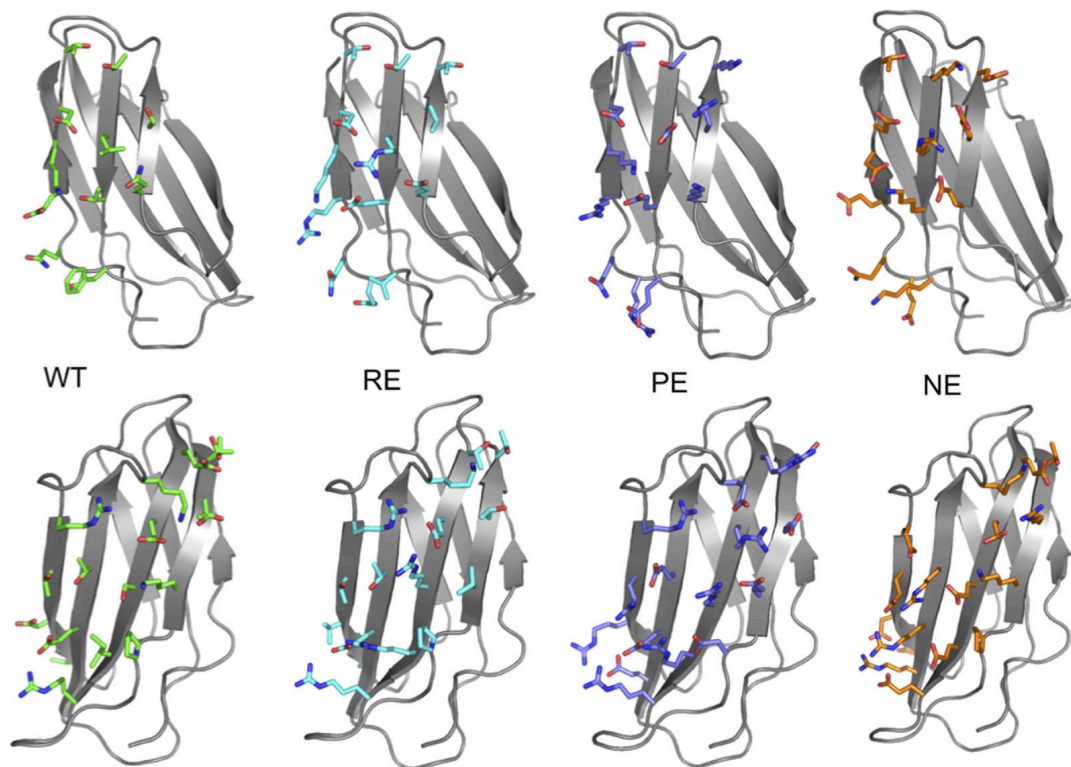
To test the Rosetta design protocol and energy function on  $\beta$ -sheet surfaces we designed and characterized two variants of TNfn3. In the exhaustive simulation, all surface positions on both  $\beta$ -sheets of TNfn3 were allowed to vary. This included 18 positions on the four-stranded sheet and 10 positions on the three-stranded sheet (Figure 3.2). All amino acids except for cysteine and proline were allowed at each position. Interestingly, Rosetta only mutated 5 residues on the four-stranded sheet and mutated 8 residues on the three-stranded sheet (Figure 3.3). All but one residue on strands 3 and 4, which are in the four-stranded sheet, were kept as the wild-type amino acid. We refer to this design as RE, for Rosetta exhaustive. The total calculated energy for RE is -195 REUs (Rosetta Energy Units, negative values are more favorable) relative to -180 REUs for the wild type protein. The hydrogen bond score is more favorable for RE compared to the WT protein (-14 vs. -10 REUs), as well the electrostatics term (-67 vs. -62; Table 3.1). New interactions predicted to occur in RE include hydrogen bonds between T66 and E68, E68 and R70, and D11 with R18.





**Figure 3.2 Surface exposed  $\beta$ -sheet residues for the various designs.** Mutated residues are underlined, and the residues which were allowed to vary in the simulations are shown in bold italic. RE (Rosetta designed exhaustively), RS (Rosetta designed sparsely), (PE) exhaustively designed charge zipper starting with positively charged  $\beta$ -strand, (PS) sparsely designed charge zipper starting with positively charged  $\beta$ -strand, (NE) exhaustively designed charge zipper starting with negatively charged  $\beta$ -strand, (NS) sparsely designed charge zipper starting with negatively charged  $\beta$ -strand.

We redesigned sidechains of surface residues with the native backbone of TNfn3, so that the redesigned proteins had alternating patterns of positive and negative charges (Figure 3.1) (76). In total, we redesigned two control designs (RE, RS), and four charge zipper designs (PE, PS, NE, NS) (Figures 3.2, 3.3, and Table 3.1). Because the 36<sup>th</sup> position at WT backbone seriously favors glycine, we kept glycine as is.



**Figure 3.3. Wild type TNfn3 (WT) and redesigns (RE, PE, NE) with surface exposed residues displayed in sticks.** The top row shows the 3-stranded  $\beta$ -sheet and the bottom row shows the 4-stranded  $\beta$ -sheet. The structures are oriented in the same fashion as the illustrations shown in Figure 3.2, in that on the 3-stranded sheet G59 is at the top right, and in the 4-stranded sheet residue 88 is at the top right. T86 and T88 in the WT protein are shown with two alternative conformations as observed in the crystal structure.

**Table 3.1 Computed Stabilities for Proteins.**

Type	Name	Salt bridges <sup>a</sup>	Total energy <sup>b</sup>	Coulomb term <sup>c</sup>	hbond_sc <sup>d</sup>	Solvation <sup>e</sup>	vdw <sup>f</sup>
Reference	WT <sup>g</sup>	2	-180	-62	-10	226	-359
Rosetta designed	RE	6	-195	-67	-14	232	-368
	RS	4	-190	-67	-11	229	-365
Charge Zipper	NE	14	-188	-77	-15	241	-372
	NS	10	-183	-69	-12	234	-370
	PE	18	-192	-89	-20	248	-373
	PS	6	-182	-71	-13	242	-370

<sup>a</sup> Number of salt bridges on the  $\beta$ -sheet surfaces.

<sup>b</sup> Total energy for the protein as computed with Rosetta (unit is REU) (101).

<sup>c</sup> Coulombic electrostatic potential with a distance-dependent dielectric (unit is REU) (95).

<sup>d</sup> Sidechain-sidechain hydrogen bond energy (unit is REU).

<sup>e</sup> Lazaridis-Karplus solvation energy (unit is REU).

<sup>f</sup> van der Waals (=“Lennard-Jones attractive between atoms in different residues”+“Lennard-Jones repulsive between atoms in different residues”; unit is REU).

<sup>g</sup> Fibronectin type III domain from tenascin (PDB code: 1ten).

We also performed a design run in which only 8 residues were allowed to vary on the 4-stranded sheet and 5 residues on the 3-stranded sheet (Figure 3.2). These residues were picked to emphasize the formation of new pair contacts between strands. Residues 44, 36, 68, and 86 were all varied and form a line across the 4-stranded  $\beta$ -sheet, similarly with residues 48, 32, 72, and 82. This design simulation produced a sequence with 3 mutations on the 4-stranded sheet and 4 mutations on the 3-stranded sheet. We refer to this design as RS, for Rosetta sparse. The total calculated energy for RS was -190 REU. As with the exhaustive design, there were improved hydrogen bonding and electrostatics energies compared to the wild type sequence with scores of -67 and -11 REUs respectively. New interactions included a hydrogen bond between E32 and R72, and a tight valine-valine interaction formed between V18 and V57.

To test the empirical approach of explicitly adding more salt bridges to  $\beta$ -sheet surfaces we constructed four variants of TNfn3. In two of the variants we varied most of the residues that were varied in the RE (exhaustive) simulation. In one of these cases, we started the charge patterning with the first  $\beta$ -strand forced to be negatively charged, while in the second case we started with the first  $\beta$ -strand positively charged. We refer to these designs as NE, for negative exhaustive, and PE, positive exhaustive. In PE, the first, fourth, fifth and sixth  $\beta$ -strands are positively charged, while the other strands are

negatively charged. The reverse is true for NE. To pick which charged residues were placed at each residue position, we performed a constrained design simulation with Rosetta where residues on the positive strands were constrained to lysine or arginine, and residues on the negative strands were constrained to be aspartate or glutamate. The final PE and NE designs have 22 and 19 mutations, respectively, and were predicted to include 18 and 14 surface salt bridges, respectively. Interestingly, the total score for the PE design is more favorable than the score for the NE design, -192 versus -188 REUs. One contribution to this difference is that the NE design results in a higher net charge for the protein (-14) compared with PE (-5) and wild type (-9; Table 3.2).

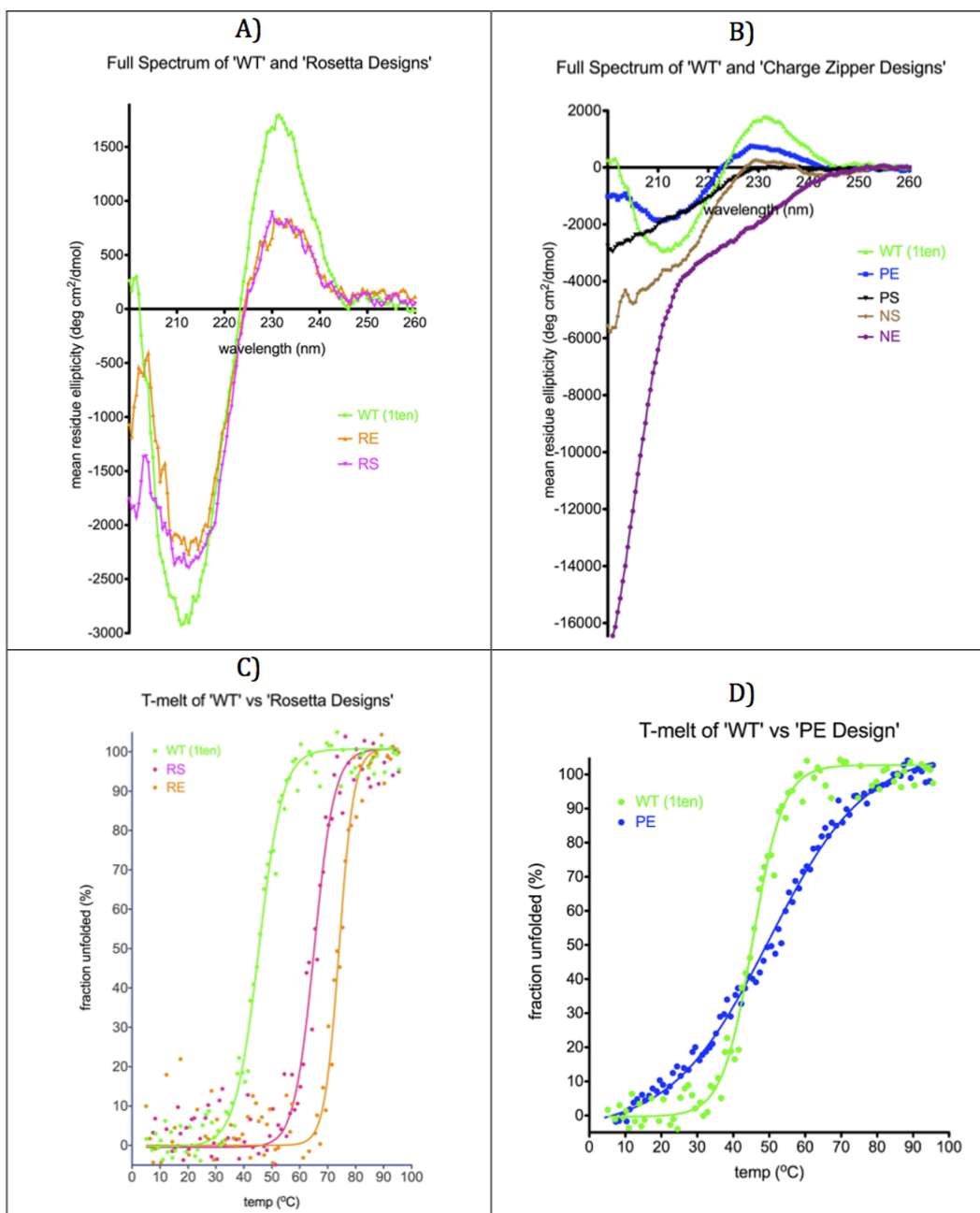
**Table 3.2 Measured Stabilities for Proteins.**

Type	Name	Net charge	Mutations	$T_m$ (0M NaCl)	$T_m$ (1M NaCl)	$m$ (Kcal/mol * M)	$\Delta G_u$ (0M GdnHCl) (Kcal/mol)
Reference	WT	-9	0	45.4°C	58.2°C	2.42	3.82
Rosetta designs	RE	-5	14	74.1°C	82.2°C	2.73	6.86
	RS	-7	7	64.1°C	77.7°C	2.56	5.27
Charge Zipper	NE	-14	19	Not folded	Not folded	Not folded	Not folded
	NS	-12	12	Not folded	Not folded	Not folded	Not folded
	PE	-5	22	49.9°C	47.6°C	N/A	N/A
	PS	-6	9	Not folded	48.3°C	N/A	N/A

In addition to the charge patterned exhaustive designs, we also created a PS (positive sparse) and a NS (negative sparse) design. These simulations used the same charge patterning rules that were used for the PE and NE designs. The PS design has 9 mutations relative to wild type and the NS design has 12 mutations.

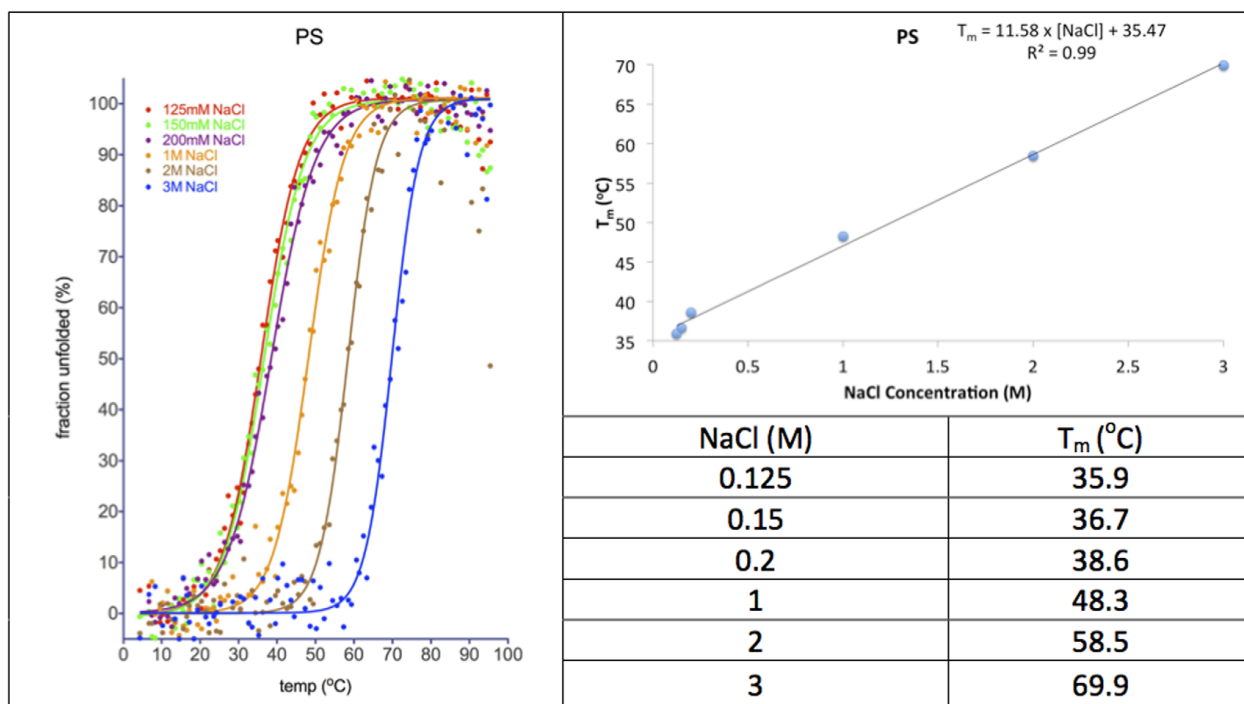
All six of the designs (RE, RS, PE, PS, NE, and NS) along with the wild type protein were expressed in *E. coli* and purified with metal affinity chromatography followed by gel filtration. Circular dichroism was used to determine if the proteins were folded. At low concentrations of salt, RE, RS, and PE all exhibited a CD spectrum consistent with a folded  $\beta$ -protein, while PS, NE, and NS have CD spectra indicative of random coil (Figure 4). The thermal stabilities of the folded proteins were measured by monitoring the CD signal at 220 nm as a function of temperature. Both of the Rosetta designed sequences were dramatically stabilized relative to the WT protein with thermal unfolding temperatures of

74.1°C (RE) and 64.1°C (RS) compared with 45.4°C for the wild type protein. Like the WT protein, the designs also refolded when returning to room temperature. These experiments were performed with 0M NaCl. At a concentration of 1 M sodium chloride, the designs were also more stable than the wild type protein, 58.2°C (WT), 82.2°C (RE), and 77.7°C (RS). Similar increases in stability were observed for RE and RS in chemical denaturation experiments with guanidine hydrochloride (Table 3.2).



**Figure 3.4** CD spectra and thermal denaturation experiments of the Rosetta designs (panels A and C) and the charge zipper designs (panels B and D).

Of the charge zipper designs, only PE is folded at low concentrations of salt and has a thermal unfolding temperature that is 5°C greater than the wild type protein. Interestingly however, PE is not stabilized by salt like the wild type protein, and at 1M NaCl has a thermal stability that is 11°C lower than the wild type protein. Intrigued by the dramatic changes in stability with changes in salt concentration, we examined NE, NS, and PS to determine if they could be induced to fold by adding salt. NS and NE did not fold, but PS was dramatically stabilized with the addition of NaCl. The thermal unfolding temperature of PS varied linearly with salt and the protein reached a thermal unfolding temperature of 48°C in 1M NaCl and 70°C at 3M NaCl (Figure 3.5).



**Figure 3.5**  $T_m$  measurements for the PS protein as a function of NaCl concentration.

## Discussion

Our results demonstrate that protein stability can be dramatically increased by redesigning only the solvent exposed face of small  $\beta$ -sheet proteins. Since the Rosetta design protocol aims to optimize several energetic features, including van der Waals contacts, intrinsic secondary structure preferences and electrostatic interactions, it is not straightforward to assign the increase in stability to any single feature. However, it is interesting that like WT TNfn3 both of the Rosetta designs, RE and RS, are stabilized by high salt concentrations. This suggests that the stability of these variants is not entirely dependent on the formation of salt-bridges between oppositely charged amino acids, as these interactions are predicted to become weaker at higher salt concentrations. Consistent with this conclusion, explicitly placing oppositely charged amino acids on the surface was not a simple recipe for boosting the stability of TNfn3. Three of the four charge zipper designs failed to fold at low salt concentrations. The charge zipper design that does fold, PE, is unlike the other TNfn3 variants, in that it is destabilized by high salt concentrations. This suggests that the redesign did have the intended effect of making protein stability more dependent on surface electrostatic interactions. In contrast to our results with  $\beta$ -sheets, surface salt bridges have been shown to have a more dominant role in stabilizing helical proteins (96) (102) (70). This is likely to be in part because the charged amino acids, Arg, Lys, Glu, and Asp have a higher intrinsic propensity to be in helices compared with  $\beta$ -strands (100).

One of our goals in testing charge patterning on TNfn3 was our hope that it would provide a way to dictate, which  $\beta$ -strands would pair with each other, and in particular destabilize pairing between strands that are close in primary sequence but are not intended to be paired. We thought that this would be a simple approach to incorporate in the de novo design of  $\beta$ -sandwich proteins, a problem that is still unsolved. The results suggest that charge patterning does not provide a simple solution, and indicate that the correct strand pairing will need to be specified by the many different structural features that go into determining  $\beta$ -sheet stability.

It is striking that in the design simulation where all residues on the surfaces of the  $\beta$ -sheets were allowed to vary, Rosetta only mutated 14 out of 28 residues. This is despite the fact that the design

simulation starts from a completely random sequence, and uses a stochastic sampling protocol to find a low energy sequence. This suggests that most native residues on the  $\beta$ -sheet surfaces of TNfn3 are already optimized for stability, and highlights the fact that every residue in a  $\beta$ -sheet is in a unique environment, where the most favorable residue depends on the precise positioning of neighboring backbone atoms (103).



## Materials and Methods

### *Computational Design and Analysis of proteins*

We redesigned the  $\beta$ -sheet surfaces on the WT fibronectin type III  $\beta$ -sandwich from tenascin (PDB code: 1ten) using the molecular modeling program Rosetta to perform rotamer-based sequence optimization in combination with backbone refinement. The protocol iterated five times between the PackRotamersMover (rotamer optimization) and the FastRelax protocol (backbone refinement) (18). The script used to perform these simulations is provided in the Supporting Information. Residues not allowed to change their amino acid identities were allowed to adopt different rotamers (“NATAA”); 1,000–10,000 independent simulations were performed for each set of design parameters (80–800 cpu hours spent, number of design trajectories did not affect greatly the final design selection), and the lowest energy sequence for each set was selected for experimental characterization.

### *Protein Expression and Purification*

All proteins were expressed using a 6-Histidine tagged PQE-80L vector in the BL21\* strain of E. coli. Isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG) was used at 0.4~0.8 OD<sub>600</sub> to induce and the proteins were expressed overnight at 18°C. Cell pellets were sonicated, and after additional centrifugation, supernatant was applied to a Ni-NTA column (GE healthcare). The purified solutions were further purified by size exclusion chromatography (GE healthcare HiLoad 16/60 Superdex 75 pg or HiLoad 16/600 Superdex 200 pg).

### *Circular Dichroism*

Secondary structure identification and melting temperature measurement were performed using circular dichroism with JASCO J-815 CD spectrometer. All measurements were done with 20  $\mu$ M protein concentration. All mean residue ellipticity values shown in this article are CD values of protein sample

after extracting CD values of buffer only. Data Integration Time (D.I.T) for ellipticity measurements was increased to 8 seconds from 4 seconds especially when high concentration of NaCl was used as buffers. When high concentration of NaCl was used as buffers, analysis of full spectrum of the ellipticity was not meaningful when wavelength is less than 205 nm. Nonlinear regression (sigmoidal dose-response) was used to fit all melting temperatures by Prism software ver. 5.0a (104). Similar thermal unfolding temperatures were obtained by fitting the data to the Gibbs Helmholtz equation with nonlinear regression by Mathematica 10 (105).

### *Fluorescence*

All chemical denaturations were evaluated by measuring fluorescence emission spectra (310–400 nm) with a Fluoromax 3 spectrofluorometer. Similar as in Gilbreth et al. (91), we plotted fluorescence intensity vs. [GdnHCl] at wavelength 365 nm after excited at 295 nm. All measurements were performed with 5  $\mu$ M protein concentration at 20 mM sodium phosphate pH 7.0 except PS where the measurement was done in 20 mM sodium phosphate pH 7.0 and 100 mM NaCl.

[Sequences of studied proteins]

> WT\_1ten

LDAPSQIEVKDVTDTTALITWFKPLAEIDGIELTYGIKDVPGDRTTIDLTEDENQYSIGNLKPDTTEYEVSLRSRRGDMSSNPAKETFTT

> RE

LDAPQQIRVKDVTDTTARIEWVKPLAEIDGIELTYGIKDVPGDRTTIVLSEDENQYVITNLKPDTTYEVRLRSRRGDMSSNPAVETFTT

> RS

LDAPQQIEVKDVTDTTAVITWVKPLAEIDGIELTYGIKDVPGDRTTIVLTEDENQYVIGNLKPDTTEYEVSLRSRRGDMSSNPAKEYFTT

>NE\_NATAA\_keep\_36G\_1ten\_chain\_A\_res-renum\_starting\_w\_2L\_res-renum\_0951 (lowest energy --> -1.78 REU)

LDAPQIEVDDVTDTKARIKWKKPLAEIDRIRLRYGIKDVPGDDTEIDLDEDENEYEIENLKPDTTEYEVELESERGDMSSNPAKERFTT

>NS\_NATAA\_keep\_36G\_1ten\_chain\_A\_res-renum\_starting\_w\_2L\_res-renum\_01847

LDAPQIEVKDVTDTKALIRWRKPLAEIDRIELRYGIKDVPGDRTEIDLDEDENEYSIENLKPDTTEYEVELISERGDMSSNPAKETFTT

>PE\_NATAA\_keep\_36G\_1ten\_chain\_A\_res-renum\_starting\_w\_2L\_res-renum\_0671

LDAPRQIRVKDVTDTTAEIEWEKPLAEIDEIELEYGIEDVPGDRTRIRLREDENKYRIKNLKPDTTRYRVRLRSRRGDMSSNDAAEEEFET

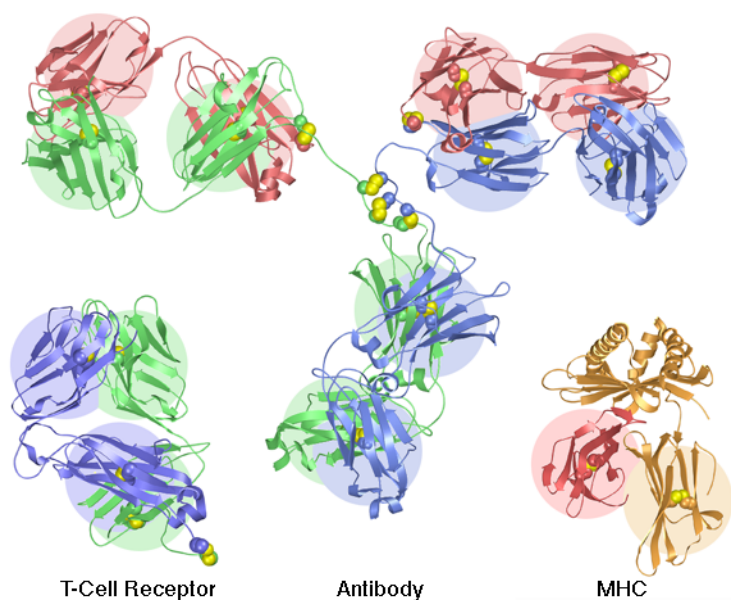
>PS\_NATAA\_keep\_36G\_1ten\_chain\_A\_res-renum\_starting\_w\_2L\_res-renum\_0646

LDAPKQIEVKDVTDTTAEITWEKPLAEIDGIELTYGIKDVPGDRTTIKLTEDENQYRIGNLKPDTTEYRVSLRSRRGDMSSNEAKEDFTT

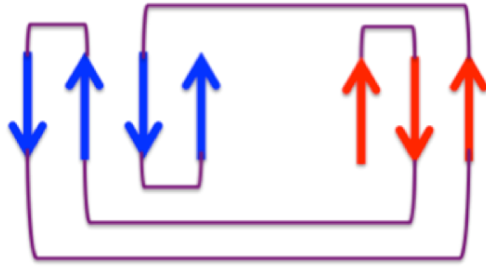
## CHAPTER 4: *DE NOVO* DESIGN EFFORTS FOR $\beta$ SANDWICH PROTEINS

### Introduction

Despite the significant prevalence of all- $\beta$  proteins (23% of all known structures (77)), the mechanism of folding is not well understood (106). This is especially true with regard to  $\beta$ -sandwich proteins, which constitute the highest percentage of domains among the mainly  $\beta$ -class (107), and are often found in various immune systems including antibodies (Figure 4.1). Among the  $\beta$ -sandwich proteins, we used the fibronectin type III domain of the protein tenascin (TNfn3), because, interestingly, it has a complex Greek-key motif (Figure 4.2). Unlike simple  $\beta$ -meanders and  $\alpha$ -helices, this Greek-key motif protein has high contact orders. This means that this motif has many residues that are distant to each other in primary sequence, but near each other in 3-D space. This high contact order makes this protein a more challenging target to study protein folding (108)(109)(110).



**Figure 4.1**  $\beta$ -sandwiches (identified in blobs) in various immune systems (111).



**Figure 4.2 Overall Greek-key motif topology of fn3  $\beta$ -sandwich**

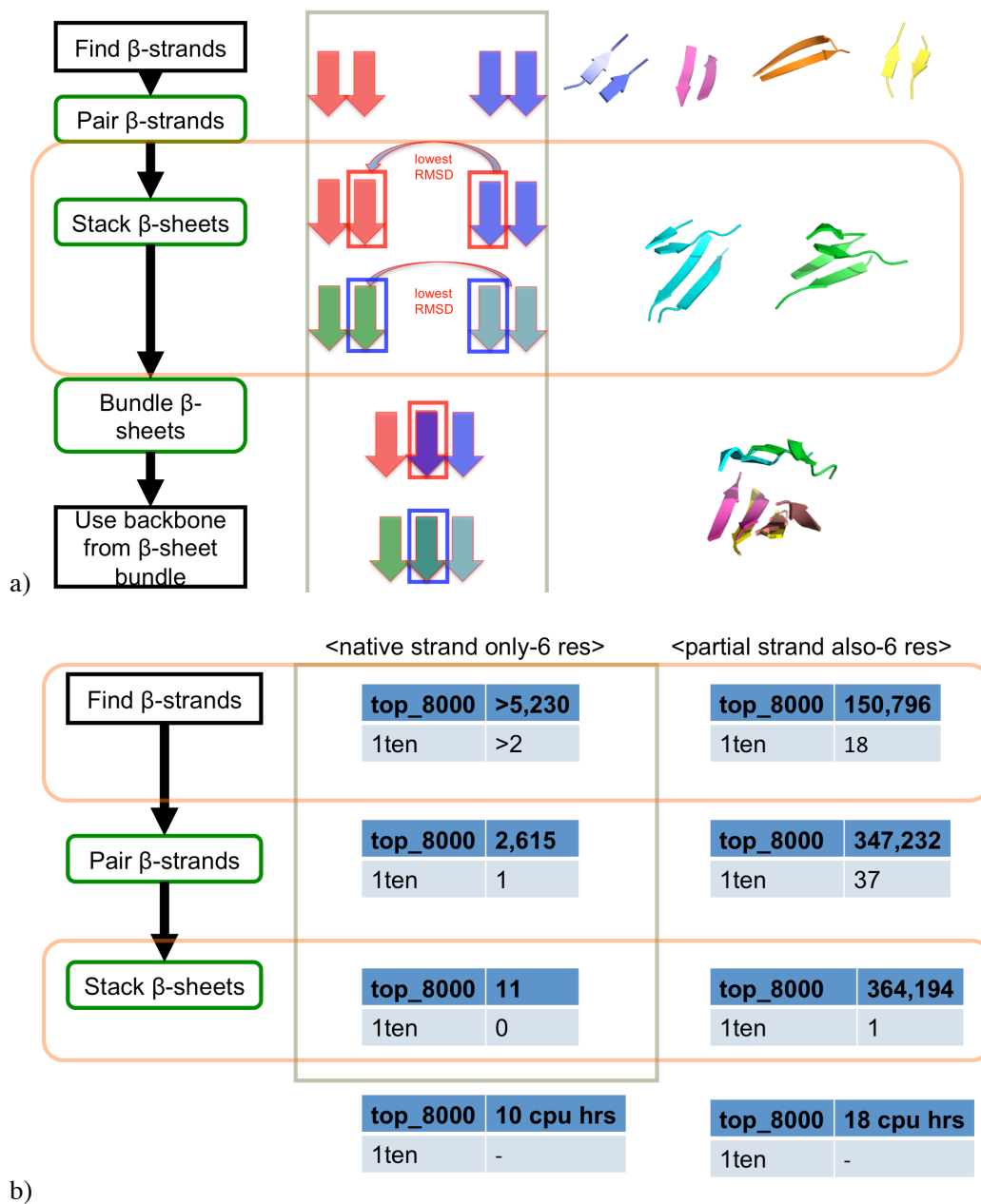
The fibronectin fold, as an extracellular matrix protein, has many biological implications also because it is crucial for cell adhesion, growth, and survival during embryogenesis, wound healing, and maintenance of normal tissue architecture (112) (113) (114). Because of these biological implications, there has been much research on folding and unfolding of this fibronectin fold (115) (116) (89) (117) (118) (119) (120). In order to better understand structural features and to engineer for various applications, attempts have been made to design this fold *de novo* for more than a decade (121).

## Computational Design and Experimental Results

*De novo* design of  $\beta$ -sandwich proteins has been elusive a goal. For example, 22 designs by Xiaozhen Hu did not show design success (121). We thought that one of the main reasons of these failures was backbone design problem, because we could successfully redesign sidechains with a native  $\beta$ -sandwich backbone (57). Therefore, we tried to design backbones with three alternative approaches to best sample ideal backbones: backbone design by SEWING, backbone design by assembling  $\beta$ -strands with randomly perturbed backbone torsion angles, and backbone design by *ab initio* folding.

### *Design of Backbone by SEWING*

We designed Greek-key motif  $\beta$ -sandwich backbones by SEWING (for all- $\beta$ -sheet proteins, we assembled bundles of four  $\beta$ -strands). What we had aspired to achieve was intricate “ideal” native distances and angles among native  $\beta$ -strands that are not easily captured (Figure 4.3) (122). However, unlike  $\alpha$ -helices (surveyed by Timothy Jacobs), there were not enough bundles of four  $\beta$ -strands for the assembly. This is largely because  $\beta$ -sheet structures are curved compared to  $\alpha$ -helices.



**Figure 4.3  $\beta$ -sandwich backbone design by SEWING.** a) overall concept, b) initial results. Top 8,000 is a collection of high quality protein crystal structures (123). 1ten is the pdb code. Native strand means each  $\beta$ -strand is used one time for assembly. Partial strand means each  $\beta$ -strand can be used multiple times because each strand is partially used. 6 res means 6 residues are used for assembly.

### *Design of Backbone by Assembling $\beta$ -Strands with Random Backbone Angles*

We generated five  $\beta$ -sandwich designs (S1~5 which stand for small) by assembling backbones of randomly perturbed backbone torsion angles using a PyRosetta utility (written by Brian Kuhlman). What we tried to achieve was better sampling of possible “designable” backbone conformations computationally. However, S3 was expressed as a soluble aggregate, and the other 4 designs were found in pellets (sequences of these designs are in supporting materials). Because many of them did not bind to Congo Red, these designs may not have formed patterned amyloid fibrils. These designs seemed completely unfolded because they did not bind to 1-anilino-8-naphthalene sulfonate (ANS). Therefore, known explicit negative designs may not have rescued these designs (124).

### *Design of Backbone by *ab initio* Folding*

Three  $\beta$ -sandwich designs (L1~3 stands for large) were generated by traditional *ab initio* folding (33) (42). This method folds a linear polypeptide chain with pre-extracted backbone torsion angles. When it folds, it replaces current backbone torsion angles with these pre-extracted ones for either 3 or 9 residues long. Therefore, it often is referred as fragment insertion. To achieve our desired  $\beta$ -sandwich topology, we applied various geometrical constraints during the folding simulation. However, L1 was expressed as a soluble aggregate, and the remaining two designs were found in pellets as well (sequences are in supporting materials). I expected that these L1~3 designs might have folded, because larger backbones tend to be more stable (60) (61). However, because these designs were larger than previous designs (S1~5), these designs would have been more difficult to computationally refold.

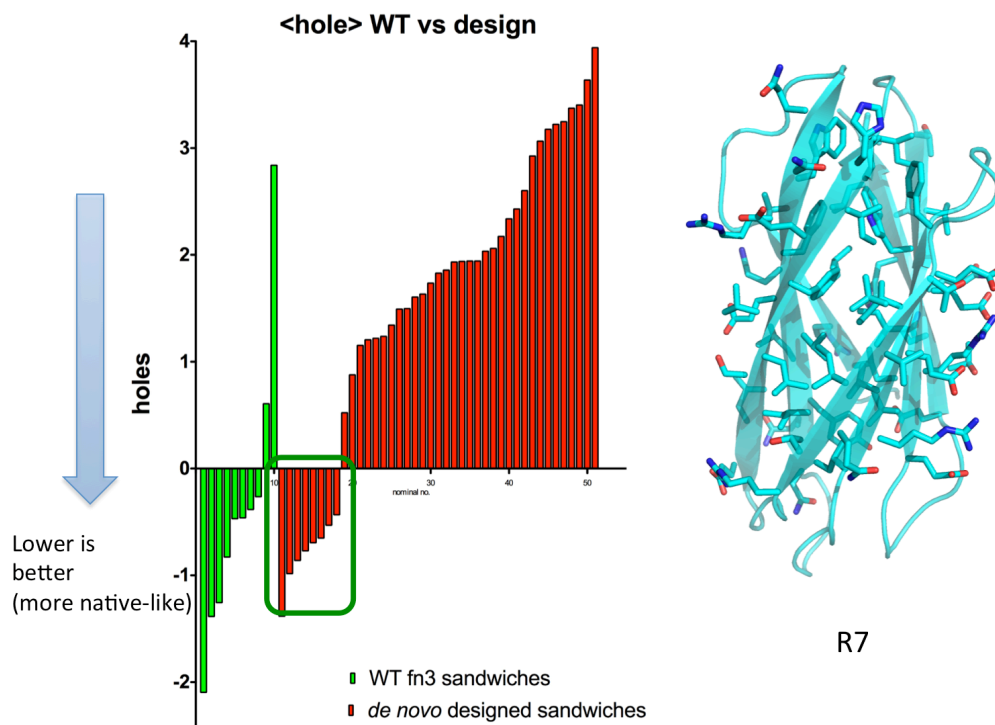


### *Design with Folding Nucleus Conservation*

Three  $\beta$ -sandwich designs (C1, C3, C4 which stand for conserved) were created by imitating a known folding nucleus. Although this is not a pure *de novo* approach, the known folding nucleus is highly conserved (125). Rosetta mutated known tyrosine corner into a phenylalanine. We used Rosetta chosen ones although we knew that phenylalanine is known to compromise some stability (126). Another highly conserved tryptophan was kept. However, all designed proteins were expressed as soluble aggregates (sequences are in supporting materials).

### *Design with Repopulation of Existing Backbones*

We generated eight  $\beta$ -sandwich designs (R1~8 which stand for repopulated) via repopulating fragments of a native  $\beta$ -sandwich protein (pdb code: 1L9N). Although this is not longer a pure *de novo* design approach, we hypothesized that it might sample backbone conformations more effectively. Furthermore, we improved sidechain packing in our designs through cartesian minimization (127) (Figure 4.4). However, all designed proteins were found in pellets (sequences are in supporting materials).



**Figure 4.4 Packing comparison of various WT  $\beta$ -sandwich and design tried ones.** The ones in green box shows R1~8. “WT”  $\beta$ -sandwich protein with 0.7 hole is redesigned  $\beta$ -sandwich protein (pdb code: 3b83). WT  $\beta$ -sandwich protein with 2.8 hole is the NMR determined structure (pdb code: 1k85). Rosetta hole calculation algorithm tends to give poor score to NMR derived structures.

#### *Redesign with Native $\beta$ -sandwich Backbones*

Although we had succeeded a redesign with native  $\beta$ -sandwich protein, our 19 *de novo* design of  $\beta$ -sandwich proteins fell short. Therefore, we wanted to determine whether a current general purpose Rosetta score function could reliably redesign native  $\beta$ -sandwich proteins. In total, we redesigned 14 designs (re1~7 and re9~15, which stand for redesign) and expressed them (Table 4.1) (sequences are in supporting materials). Except a fixed design of 1yq2 (re05), most designs were either aggregated or not expressed. Therefore, it seems obvious that we may need to use used  $\beta$ -sheet optimized score function as before (128). This conclusion was made with various supporting factors: all used genes were codon-optimized, most non-expressing designs were tried to be induced at both 18°C and 37°C and 50ml and 1.5 L LB media cultures, and this is a large-scale test (we used 7 native backbones, if we include one more

native backbone, it would be 8 backbones in total, the native backbone that we excluded came from pdb code 3hn3, this protein was expressed by *Mus musculus* before, when we tried to express extracted  $\beta$ -sandwich proteins after fixed backbone redesign and flexible backbone redesign, both redesigns were not expressed in *E. coli*, however we excluded for analysis because there was not guarantee that even WT protein is expressed by bacterial system).

Original pdb file	Backbone	Name	MW (kDa)	Conserved Y	RMSD (Å)	Sequence recovery (%)	Expression
1ten	WT	1ten		yes	0	100	WT
	fixed	re01	12.4	yes	0	37	No expression
	flexible	re09	12.2	yes	1.0	28	No expression
2r2c	WT	2r2c		yes	0	100	WT
	fixed	re02	14.6	yes	0	32	Aggregate
	flexible	re10	14.4	no (F)	1.2	29	No expression
1bgl	WT	1bgl		no	0	100	WT
	fixed	re03	14.5	no	0	25	Aggregate
	flexible	re11	14.2	no	1.4	20	Aggregate
1fnf	WT	1fnf		yes	0	100	WT
	fixed	re04	13.3	yes	0	38	Aggregate
	flexible	re12	13.0	yes	0.8	35	Aggregate
1yq2	WT	1yq2		no	0	100	WT
	fixed	re05	14.4	no	0	38	$\beta$ -sheet
	flexible	re13	14.8	no	0.9	39	No expression
3r8q	WT	3r8q		yes	0	100	WT
	fixed	re06	14.0	no (F)	0	37	Aggregate
	flexible	re15	13.0	no (H)	1.1	30	No expression
3tes	WT	3tes		yes	0	100	WT
	fixed	re07	12.6	yes	0	43	Retry needed
	flexible	re16	12.7	yes	0.7	30	No expression

**Table 4.1 Redesign with Native  $\beta$ -sandwich Backbones.**

## Conclusion

It has been very challenging to design all  $\beta$ -sheet proteins *de novo* (32). Our 22 designs by Xiaozhen Hu and 19 designs by Doo Nam Kim were all not successfully folded (121). These failures were somewhat unexpected because all these designs (including Xiaozhen Hu's 22 designs) had pretty good Rosetta total scores when calculated with the original "score 12" score function "score 12" (33) with explicit electrostatic terms. Evidently, the Rosetta total score alone is not enough to predict well designed  $\beta$ -sandwich proteins.

## Supporting Materials

> S1

MTAGGTHKNGKFFVDVHGKGNVDVDVEIRAGNVRRRYRVIGGPGERIEFQGD TAGEARVYVRDGTWEQTLYLK

> S2

MRVEIRMDNGWVRVDAEGDGPVKVDAQSRGGGYDYQVRTELVGRREITVQGAPNGEVRVRLRENGEEREFEQYK

> S3

MRFRGTRKGNVVEVEAKGDGTGNYEIKAKGGGQEYRYRAPTYGGSSWRGKIYPGGTFRVEVRIGNLNTEGEFK

> S4

MEYEIRVDNGEYVFRVGNSHRPEGEVRVFYNGIHIQKGTKYGDGQKIRVRGYLNGQVHFRFGGDNEDYEIVLG

> S5

MELRFELTG DYGRGFLRGYGYAELGFYVIFNGDETGFTTGGYGGQEFKFQGYPTGRYRLIARSGGEELRYEYE

> L1

MVEGRMEVRIHNGRAIGYGYVWSKTNPQKL RWRGVIVAGGVGAEFETGSDDDGTSLTVTVDTKGKTGVIRGRGEARGQKNGQEFRS  
EVENTPD

> L2

MSEARAEIGVHNGRLRVRVYGRSDSQGQENRMRGYMIVGGYRYETTGDGPPGATGFDIEVEHEGREGQA HFRGKVRGNINGDEQEYE  
TQYGSL

> L3

MAQLQTGYHWRNGHLGGLVRIIDNGAGGRGYVEIYVRGGNV DYKLQTEFPENGSHVEAEVHGEGNTKDFKV FARAVFYINGVEYRF  
EIRDGAG

> C1

PTTGRVSVQVQGN AVTVEWQGGDSVTRITLRF TDGDGNPTSLTVEADGN GDRRITVPVPSGSFTIEITVESGTGTTLTQTVDLAG

> C3

DSSSRVSATENG PAMTVTWTGTTRDISRITVTF TDGGGTGKVTLEVTGN GDTSVTVPVEGGRFKVTIRLEESNGTTDELEVDGSQ

> C4

PEPTQVTITRTGDNVTVTWTGEEEIREVTLTFTDARGQDSKRTVSAVGETVTTVTIPVPLDQYRVTVSLESKNGTTLTLTIDASG

> R1

PPIVRLELDQTHIPGLPILFTARIQNHQPTVLKNIRIEFLEQDPKITHEYVVGPLRPGEEVVIQFLHLPPTTGTLTHIWNILPGTTTIVVTMT  
LTI

> R2

TETLVVTVEPNPEPGKEFRVEIRVTNLTNLPWRNIRVEIRLPGHITVPIEIIFTTLPPGQTIVIETSHIPTRGTITFELRVETEPQTDIETSYNL  
TI

> R3

IPTLILTVEPNIEPGKEFTVTIVVTNLTNLPWRNIEVNILLPGHIDVPITIIFTHLPPGQTIVITTSIIPRTTGTLFQLEVNTPEQTDIRTSYTLTI

> R4

THLFTIDIPGVPEPGKPHLIRVTFINLSNTKHEGGTVNVEIPGLHTGRITVTFEPRPPGQTHTYPLILIPQQTGRITILFELHTSTQRHIEFRD  
TVL

> R5

TVELELILVNPHTEGETFILTLRFTNKTPVPVNPVQAEWWKTLTRTETQTYQIPTITPGETVNVKVEVTIPTRGVTTFEYRMTAELQPEEV  
RQTVMSD

> R6

TPTLDLQLITQSVPGTTFILELRITNLTNVPLRGLHVELLVPPHLTTPLSWQPKPSTPGEVVEWVQEFLKTTGTFTLIYNWRHETLPVLTH  
TRTMTH

> R7

PSNIRLQTLTTLVPGQNWRFEVLYQNSSTTVLVNIRVEIEFKNQPGRTVVIQTRPLTPGELLSLVFEIHIPPGTLVWEIRVEAKGFTLETQ  
QHTLVV

> R8

PDPLRLQLLVPIELGKTVIWELRFTNLQPTVLTNIRLEIRIENQTGLPREQVYPPPLPGQTMIFTIITYTITVPGTTRLELHVEAPGVPTLVVSQ  
TTIT

> re1\_1\_1\_TEN\_A\_0001\_VAL\_0967

LESAQNIQVIKITKTTAVVIFLPSTDPVQGYEMTYGYKEDPSDRVTVVLTSSIDHYVITNLKPNNAVYIVRIVARNGDLKSDSTARTYKT

> re2\_1\_1\_2R2C\_B\_0001\_VAL\_0848

TKVLVFAEARWIGKNYIEIRAIGQVLPGYTVSPAFMRVLLMLWEVQSFVNGDLGDYFVPGKVYVYTHEVNPPEGAPLDQNKYAVKV  
EIYSSQTGEVYAEIVVSIKPPG

> re3\_1\_2\_1BGL\_A\_0001\_VAL\_0334

PSLFDTYLDGQILIVRSNDSSESRYVLRGRQALNGVVLSTSEVRAHATARGEQYVVSTLTNTTDPGEVWLEYRFYQEPTDTSPPG  
AEMGNSKFKYAQVDKYYS

> re4\_1\_2\_1FNF\_A\_0001\_VAL\_0160

GDTESPTNIVFTNIGPDRVEVRWKPPPNRDLSGFDVRWHHRNHKEKAYRQEVSPPLYIILTNLKPNGEYVVGVAAREGTADSDEARG  
 TVRTGKAPG  
 > re5\_1\_2\_1YQ2\_A\_0001\_VAL\_0537  
 PASIELHLTYPPGGRVVLEVRNNSKSDASMFLLFYRVMINGVVATYGVKEARGSNALSAGEATKIELPPFPSSSTGKTVVEVVAKWK  
 LAGPTWPSGQPEGYTKLDMSASQG  
 > re7\_3R8Q\_sandwich\_1.clean\_0001\_VAL\_0651  
 TQEAPRNLTLEVGPHYLVIKWEPPPTQLGGFEVRVEPALKLGKAFNVWLPPNASSAVITGLLPNTEFEVRVFAVNGPEYSSPATAVVV  
 TKAPSG  
 > re8\_3TES\_A\_res-renum\_0001\_VAL\_0819  
 LEGPRNLRATNVTKKSIELRWEAAPNQFRAFVIEWEEARRRGTAAYRHEVHGSQRTYLLTGLLPNTTYEVAIQGRLNGQDSAPHSAFYTT  
 > re9\_1\_1\_1TEN\_A\_0001\_VAL\_0470  
 LKPAQNVHVTNTTSHALTALVLDASDDQVQGYLVMYGKADDSSDRVIGFASANDRYWLITNLEPGARYEVVVIANGNLHSDGNSTTF  
 VA  
 > re10\_1\_1\_2R2C\_B\_0001\_VAL\_0599  
 PVIEIQAWARWVGPKYIEVRVTVRTAPGYTVPEATVEVWLVTGGGVRAPVNGTDGQPAITGQTWVFYQVEPEPGYTLQSKFAINA  
 RVASAHTGQVWAEQVVPPIEPQG  
 > re11\_1\_2\_1BGL\_A\_0001\_VAL\_0967  
 RSYETRLDGRVILGISKADTDDSTNAVVEVHAHNGLTLYTGRYPANAPAGGAERVVLDQIKETTAAGEIWIEVYFRAKTAGDNYP  
 GYIKGYGKHKFSEVPDNLVSG  
 > re12\_1\_2\_1FNF\_A\_0001\_VAL\_0067  
 PPTLSPTDGRFTDVSARAVRVEWRPPDNLARGFIVVYFKKNDKSDMYVAVVPSSSTYYIATNLEPGAEEYVFRVAALLGTSISDSLEGT  
 QQTGLEPG  
 > re13\_1\_2\_1YQ2\_A\_0001\_VAL\_0226  
 PAEQLHLTFPPGGKVLLLVINNSDTSASLFLYRVMDNGDVKRYGVKEARGSNGLTAGETSVQVLEFPFPQPTGKTVVEVIARFK  
 VTTSRAPSGYPHGYTRLDTLEQK  
 > re14\_1\_2\_3HN3\_A\_0001\_VAL\_0203  
 GTFVAAVYVWTTQRGTSVVRFLIVVLGSQDYRIVVIIRDSNGQPQGGSSGSTGEAKVPNGKFAQPTGQHKYAHYHYMEVILYARTDN  
 GWEMYIYIQTVPSTG  
 > re15\_3R8Q\_sandwich\_1.clean\_0001\_VAL\_0170  
 TEIPPQNLTVLEIGHPHYTVIWFAPPTQLTGFKVIVTAKEDRGTSLVVWQHATSNYAVITGMLPNTTHEVRVIAVNGPEYSSPAKHVYTT  
 KEESR

## CHAPTER 5: FUTURE DIRECTIONS

### Towards More Efficient Protein Design

After about three months, I chose what I thought were the best eight designs for the first round of *de novo* designs of  $\beta$ -sandwich proteins. Because I heard that a more senior member of my lab (Ben Stranges), had tried a lot of experiments for his designs, I wanted to save valuable experiment time by only trying the best designs. Consequently, I lost some valuable opportunity to gain feedback from experiments. It turned out that lessons learned from experimental validation were more important than the somewhat uncertain computational design methodology improvements. I learned that as long as a computational design methodology improvement is not that significant, fast experimental validation is very efficient in the long run.

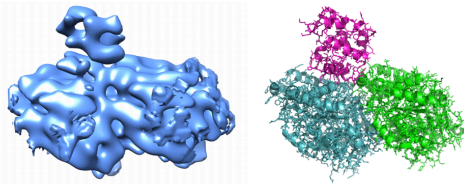
Also, I spent too much time trying to express and purify less promising designs such as S3, en4, and en6 proteins. Overall, I tried to express about 90 designed or WT proteins about 200 times, many proteins were tested multiple times with alternate expression and purification protocols. The rationale for these multiple tries was to confirm experimental validation or to find better ideal expression or purification conditions by repeating experiments in slightly different methods. However, the verdicts were eventually that these less-promising proteins were indeed either aggregated or were expressed not enough. Most of the time, when the expression yield or purification was not good (“not well behaving proteins”) the first time, additional trials were also fruitless. If I were to do Ph.D. training again, I would try to express or purify one or two times at the most. I hope that any protein designer who reads this article will design proteins much more efficiently based on my experiences. Also, sometimes using a concept or method from another field can be effective (129). For example, a statistical analysis borrowed from the stock market helped protein sequence analysis (130). Because the computational protein design field is

rapidly evolving especially among the Rosetta community, effective mutual learning and teaching from other protein designers is absolutely critical.

### Necessity of Experimentally Determined Structural Information

Most computational protein design methodology developments are iterations between computational design and experimental validation. It is critical to get experimental validation results, because they guide the next step during design methodology developments. In particular, experimental structural information is very important because we design protein structures. However, x-ray crystallography was challenging for us because crystals were not formed. Furthermore, 3D structure determination using NMR was difficult because the target protein (“en8”) was a homo-dimer. In the near future, the iteration between design and structure determination will be greatly facilitated because of cryo EM, which does not need protein crystallization and can elucidate atomic resolutions (Figure 5.1). Scientists have been pushing the lower size-limit and resolution of biological molecules by cryo EM; as of 2016, cryo EM can decipher structure of protein as small as 130 kDa and as accurate as 1.8 Å resolution (131) (132) (133).

#### • Re-refinement of 3J6P, resolution 8.2Å



METRIC	Original	<i>Phenix</i>
Map CC	0.596	0.743
RMSD (bonds/angles)	0.03/2.34	0.00/1.11
Clashscore	92.37	34.73
Rama. outl., %	2.03	0.54
Rotamer outl., %	26.21	0
C-beta deviations	2	0

949 residues, refinement time: 15 minutes

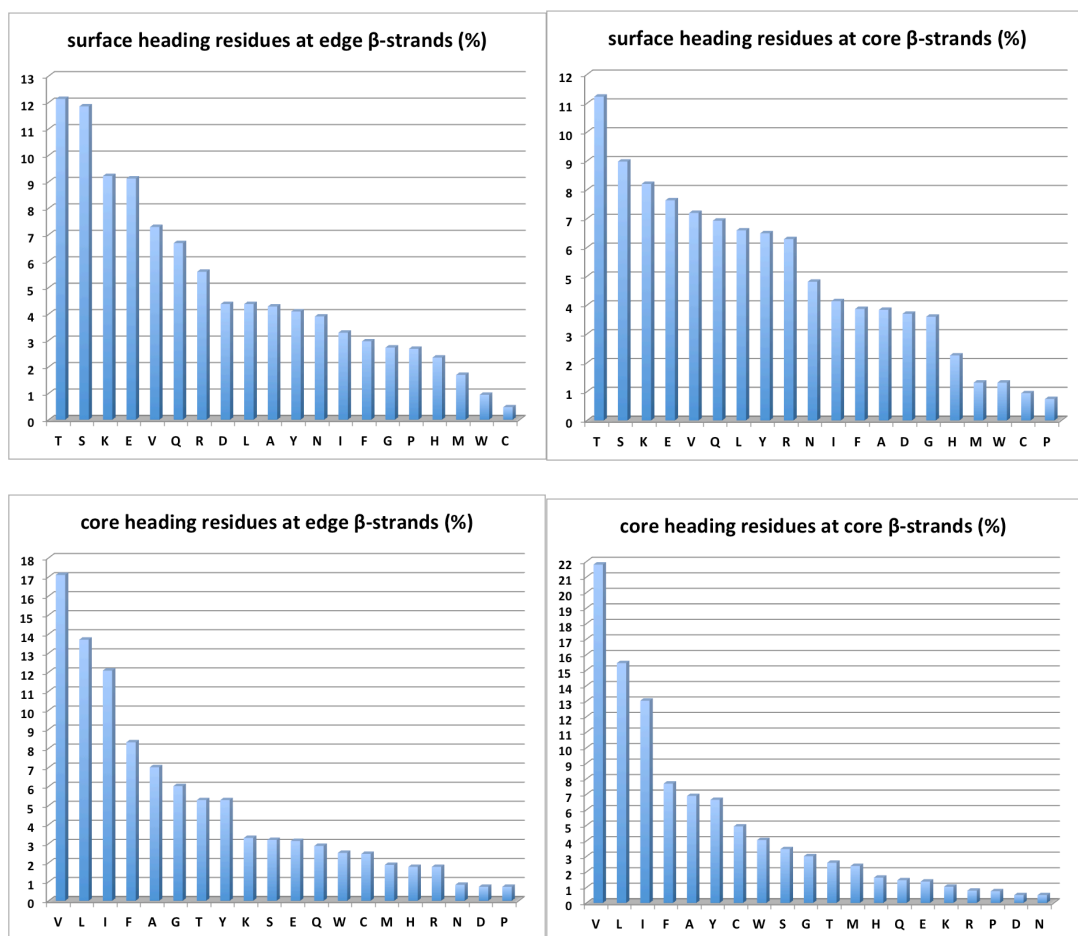
**Figure 5.1** An example of using Phenix real-space refinement to refine atomic models into cryo-EM maps (134).



## **Towards a More Accurate Score Function**

A score function is absolutely critical for higher success rates of computational protein design and folding (135). In fact, according to the random energy model, unfolded state energies are dependent only on amino acid composition rather than the specific arrangement of amino acids. Therefore, energy discrepancies between computational predictions and experimental results are due to force field flaws that account for folded state sequence energies (136).

For a more accurate score function, we need to pay more attention to reflect local properties instead of trying to achieve higher sequence and rotamer recovery rates with a global model. However, most of the recent Rosetta score function developments were done with monomer structures for a global model (94) (95). Therefore, even the latest score function (“talaris\_2014”) cannot exceed sequence recovery higher than 50%. Of course, I believe that no matter how much the Rosetta community strives to improve the sequence recovery rate with energy terms, we cannot reach 100% because nature seems to consider non-energetic factors such as aggregation prevention keepers and functional moieties. To truly reflect these complex native sequence properties, the score functions need to be applied differently according to relative location (surface, intermediate, and core). The score functions also need to be applied differently according to biological target (membrane proteins or exposed ones). These score functions also need to be applied differently according to secondary structure (helix, loop, and edge  $\beta$ -strand and core  $\beta$ -strand). For example, when I counted occurrences of amino acids according to relative orientation and locations of 203 WT  $\beta$ -sandwich proteins that have less than 120 residues, there were strong dependences on relative orientation (Figure 5.2).



**Figure 5.2 Distribution of amino acids in WT  $\beta$ -sandwich proteins.**

Indeed, electrostatic interactions are more often found at protein-protein interfaces, and explicit consideration of these interactions often brings successful protein designs (28) (29) (30) (31). I also observed minor dependence of amino acid distribution on relative location. This difference may have arisen from a negative design perspective to prevent unwanted aggregation (124). To better reflect local properties rather than relying on a global monomer model, one needs to use different training sets according to relative location (interface, surface, and core), biological target (membrane or exposed proteins), and secondary structure (helix, loop, and sheet) when fitting coefficients using optE (137). Of course, for real application of these score functions, Rosetta will apply different score functions depending on local environments.

Current Rosetta score functions have linear coefficients. Therefore, in certain cases, the interpretation of score functions regarding the energy term that specifically affects the observed outcome might seem straightforward. However, there are many score terms that are often redundant to each other due to their inherent properties of being knowledge base potential and molecular mechanics force fields (138). I need to confirm this claim by analyzing standard beta values using JMP statistical software. However, the fact that there are around 17 score terms (that some of them are dependent on each other) for a single score function, even after preliminary trials of reducing redundancies, is a strong indication that there are still significant redundancies among the score terms. Therefore, the interpretation of current score functions may not always be straightforward. Consequently, according to my machine learning experiences, the optimization of score terms' coefficients using neural network (to maximize sequence recovery rate) rather than linear regression appears to be a better choice.

### **Towards Successful *De novo* Design of $\beta$ -sandwich proteins**

*De novo* design of all  $\beta$ -sandwich proteins remains as one of the most difficult goals. Thorough analysis of native  $\beta$ -sandwich proteins by *SandwichFeatures* (139) shows that my eleven *de novo* designs (designed using: backbone assembly, *ab initio* folding, and conserving folding nucleus) have comparable geometric features with respect to: amino acid distribution, backbone dihedrals (ramachandran),  $\beta$ -sheet capping, buried unsatisfied H-bonds, chirality of sidechains, distance between facing  $\beta$ -sheets, electrostatic interactions, exposed hydrophobic areas, high contact order of backbone (folding order), interface area between facing  $\beta$ -sheets, length of components, native-like local structure (fragments), negative design to prevent aggregation, net charge, ratio of sheet/loop, right-handedness, and  $\beta$ -sheet shape. However, these eleven designs pack more poorly than native  $\beta$ -sandwich proteins. My newer eight proteins designed *de novo* (by repopulating existing backbone fragments and cartesian minimization) have all the native geometric features mentioned above in addition to packing comparatively to native proteins (R1~8, Figure 4.1). However, these designs still expressed in aggregated form. The factors that

we miss for successful beta-sandwich design may be understood from scientific analysis of conserved residues as in Poreski et al. (140). Many of these conserved residues seem to be critical for folding perhaps as a folding nucleus (117) that may not be easily captured by the rosetta score function.

As fourteen redesign tests (with seven native  $\beta$ -sandwich backbones in chapter 4) and our previous redesign of  $\beta$ -sandwich protein success (128) have shown, successful future design may need to use the  $\beta$ -sheet optimized score function (87). Indeed, when Xiaozhen et al. previously redesigned a WT  $\beta$ -sandwich protein with a  $\beta$ -sheet optimized score function (128), the sequence recovery was 42%. However, when I used a regular score function for redesigning WT  $\beta$ -sandwich proteins, the average sequence recovery was 36% (minimum 25% ~ maximum 43%) for fixed backbone design and 30% (minimum 20% ~ maximum 39%) for flexible backbone design (of course, the sequence recovery rate for the flexible backbone design is to some extent hypothetical).

In addition to this beta-sheet optimized score function, if we design better packed protein structures and improve poorly performing Rosetta samplings for high contact order proteins (110) and secondary structure prediction algorithms (109) for important forward folding trials (43), we may finally successfully design  $\beta$ -sandwich proteins *de novo*. Because forward folding for the full structure of beta-sheet structure is challenging, prediction of loop structure will only be a viable solution for now (74). I believe that charge zipping (chapter 3) of beta-sheet proteins will likely increase the success rate by inducing desired folding order (because there are many possible folding orders in Greek key motif proteins (141)). Furthermore, the ideal electrostatic interaction seems to improve thermal stability (140). Of course, our charge zipper designs (chapter 3) did not fold as we expected. These failures may be rescued by modifying reference values of charged residues as supercharging protocol did (19) (20) instead of forcing charged residues by resfile. As PE and PS folded while NE and NS did not fold, the 16th threonine may have to be kept as native. In addition to these computational methods, quick experimental validation of designed proteins' stabilities may be a viable alternative until computational based design method mature (142).

## Potential Applications of Designed Proteins

At times, basic research seems to lack application in the near future such as the knowledge based Rosetta score function (“score12”) that was used for *de novo* design of  $\alpha/\beta$  protein, *top7* (33). However, often times, long term practical applications are readily visible including designing a pH-sensitive IgG binding protein, an enzyme of a novel metabolic pathway, and influenza binding proteins (10) (12) (13). Likewise, newly discovered linear correlations between NaCl concentration and protein stability (chapter 3) may open many new protein-engineering goals such as the delivery of molecules according to salt concentrations, as the NaCl is often linked to many diseases (143). Once we attach a fluorophore to our fnIII  $\beta$ -sandwich protein, we may be able to easily visualize NaCl concentrations in our body in real time, similar to fluorophore-labeled antibodies targeting the antigen (Figure 5.3). If we engineer fnIII  $\beta$ -sandwich proteins’ loop regions to bind to certain targets, these  $\beta$ -sandwich proteins will bind to targets according to NaCl concentration and local temperature as well (144).



**Figure 5.3** Molecular imaging of a mouse implanted with tumors that bear a specific antigen (145).

*De novo* designed  $\alpha/\beta$  protein (en8) may be a useful scaffold because  $\alpha\beta$  class proteins tend to have longer half-life than  $\alpha$  and  $\beta$  class protein (146). Furthermore, the en8 protein has N and C-terminal residues that are located at the edges in 3-D spaces, which are expected to fold and unfold faster (147). Reasonably high melting temperature of the en8 (higher than 65°C) is encouraging because often proteins tend to be stable when they are large or have N and C-terminals that are located in the middle (63). The faster folding and unfolding rates may allow these proteins to serve as nucleic acid binding proteins where frequent binding and unbinding are desired rather than super tight binding. Indeed, nuclear complex proteins that act on DNA and RNA tend to be highly dynamic engaging in transient interactions with each other (67). In order to verify this hypothesis, folding and unfolding rate measurements may be necessary.

### **Possible Reason of Increased Expression Yield of High Net Charged Proteins**

Although some supercharged (high net charged) proteins ( $\beta$  barrel and  $\beta$  sandwich) tend to express lower yields (19) (20), my supercharged designs (chapter 2) expressed more than the original designs. This difference in expression yield may be explained by a difference in supercharging magnitude and protein fold. The reason that my supercharged  $\alpha/\beta$ ,  $\alpha+\beta$  proteins expressed more than their original designs could be explained by less favorable intermolecular interactions between monomers due to higher net charges. This explanation is very similar to glycosylation in nature. Glycosylation, the most prevalent post-translational modification of proteins, tends to make proteins more soluble. However, because water molecules interact more favorably with peptides rather than with glycans, promoted solubility by glycans seems to come more from steric inhibition of protein-protein contacts than from favorable solvent interaction with glycans (148).

## APPENDING CHAPTER: USED ROSETTA INPUT PROTOCOLS

### Rationale

Inspired by previous peptide design cases (149) (150), I've iterated "FastRelax" (for backbone conformation perturbation) and PackRotamers (for sidechain design) for both charge zipper designs (chapter 3) and for most of the *de novo* designs for  $\beta$ -sandwich proteins (chapter 4). However, instead of simple iteration between "FastRelax" and PackRotamers, "FastRelax" alone, which allows sequence design as repulsive weight is ramped up gradually, does the very similar job of flexible backbone design along with sequence design and seemed to better pack residues (101). This new method is essentially same with the "FastDesign" protocol in *Rosetta* software suite, which has more on-the-fly filters. I've used "FastRelax" with sequence design option for  $\alpha/\beta$  and  $\alpha+\beta$  *de novo* designs using SEWING (chapter 2) (151).

Here I present my last example of Rosetta scripts (152), a corresponding flag and resfile. Filters such as score\_norm and SecondaryStructureHasResidue are fairly useful in eliminating the need of later processing by perl or python.

## RosettaScripts

<ROSETTASCRIPTS>

<TASKOPERATIONS>

<RestrictToRepacking name=restrict/> Only allow residues to repack. No design.

<IncludeCurrent name=keep\_curr/>

Includes current rotamers (eg - from input pdb) in the rotamer set. These rotamers will be lost after a packing run, so they are only effective upon initial loading of a pdb!

<LayerDesign name=layerdesign make\_pymol\_script=1 layer=core\_boundary\_surface\_Nterm\_Cterm use\_sidechain\_neighbors=1>

<core>

<all append="AFGILMNPQVWYHKRST" />

<all exclude="CDE" />

</core>

<boundary>

<all append="AFGILMNPQVWYDEHKRST" />

<all exclude="C" />

</boundary>

<surface>

<all append="AGMNPQDEHKRST" />

<all exclude="CILVFWY" />

</surface>

</LayerDesign>

<ReadResfile name=resfile filename=min\_resfile.txt/>

</TASKOPERATIONS>

<SCOREFXNS>

<talaris2014\_cart weights=talaris2014\_cart>

Reweight scoretype=coordinate\_constraint weight=1/>

Reweight scoretype=res\_type\_constraint weight=1/>

</talaris2014\_cart>

<cen\_score weights=score3/> needed for env

</SCOREFXNS>



[illegible]

```

<SSShapeComplementarity name="ss_sc" verbose="1" loops="1" helices="1" />

<CavityVolume name="cav_vol" />

##### fragment assessment #####

<FragmentLookupFilter                name="faulty_fragments"                lookup_name="source_fragments_4_mer"
store_path="/nas02/home/k/i/kimdn/db/many/vall/VALL_clustered/backbone_profiler_database_06032014" lookup_mode="first" chain="1"
threshold="50" confidence="1" />

</FILTERS>

<MOVERS>

    <SwitchResidueTypeSetMover name=to_fa set=fa_standard/>

    <AssemblyConstraintsMover name=ACM native_rotamers_file=278711_5_1.rot native_bonus=1 native_pro_bonus=2/>

    <FastRelax      name=fastrelax      repeats=1      disable_design=false      scorefxn=talaris2014_cart      cartesian=1
task_operations=resfile,keep_curr,layerdesign delete_virtual_residues_after_FastRelax=1/>

    <Dssp name=dssp/>

</MOVERS>

<OUTPUT scorefxn=talaris2014_cart/>

<APPLY_TO_POSE>
</APPLY_TO_POSE>

<PROTOCOLS>

design

<Add mover_name="ACM"/>

<Add mover=fastrelax/>

to save time

<Add filter=must_have_core_res/>

real filters

<Add filter=bunsat/>

<Add filter=cav_vol/>

<Add filter=charge/>

<Add filter=env/>

<Add filter=exposed/>

<Add filter=faulty_fragments/>

```

```

    <Add filter=holes/>

    <Add filter=nres/>

    <Add filter=packstat/>

    <Add filter=percentage_of_ala/>

    <Add filter=rmsd/>

    <Add filter=score_norm/>

    <Add filter=ss_sc/>

    <Add filter=sspred/>

  </PROTOCOLS>
</ROSETTASCRIPTS>

```

## Flags

```

-s

/nas02/home/k/i/kimdn/lustre/side_chain_design/a_slash_b/w_max_4_loop_res_5_edge_res/led_to_E1_8/make_input_dirs/input_files/278711_5
_1.pdb

-holes::dalphaball /nas02/home/k/i/kimdn/db/DAlphaBall_sheffler/DAlphaBall_gccstatic

-jd2:delete_old_poses

-jd2:mpi_work_partition_job_distributor

-ignore_unrecognized_res

-nstruct 1

-packing:linmem_ig 10

-relax:constrain_relax_to_start_coords

-mpi_tracer_to_file mpi_tracer

-overwrite

-mute core

```

## **Resfile**

ALLAAxc

USE\_INPUT\_SC

start

1 A NOTAA CP

## REFERENCES

1. 2004. WO 2004/01897 A2, MODIFIED NUCLEOTIDES FOR POLYNUCLEOTIDE SEQUENCING. .
2. Bale, J.B., S. Gonen, Y. Liu, W. Sheffler, D. Ellis, C. Thomas, D. Cascio, T.O. Yeates, T. Gonen, N.P. King, and D. Baker. 2016. Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science*. 353: 389–394.
3. Pharmaceutical Products & Market. .
4. King, C., E.N. Garza, R. Mazor, J.L. Linehan, I. Pastan, M. Pepper, and D. Baker. 2014. Removing T-cell epitopes with computational protein design. *Proc. Natl. Acad. Sci. U. S. A.* 111: 8577–82.
5. Skolnick, J., and M. Gao. 2013. Interplay of physics and evolution in the likely origin of protein biochemical function. *Proc. Natl. Acad. Sci. U. S. A.* 110: 9344–9.
6. Service, R. 2016. This protein designer aims to revolutionize medicines and materials. *Science*. July 21st.
7. CureVac. .
8. Moderna. .
9. Strauss, C. 2015. The Perfect Fit. .
10. Fleishman, S.J., T.A. Whitehead, D.C. Ekiert, C. Dreyfus, J.E. Corn, E. Strauch, I.A. Wilson, and D. Baker. 2011. Computational Design of Proteins Targeting the Conserved Stem Region of Influenza Hemagglutinin. *Science*. 332: 816–821.
11. Song, Y., F. Dimaio, R.Y.R. Wang, D. Kim, C. Miles, T. Brunette, J. Thompson, and D. Baker. 2013. High-resolution comparative modeling with RosettaCM. *Structure*. 21: 1735–1742.
12. Strauch, E.-M., S.J. Fleishman, and D. Baker. 2014. Computational design of a pH-sensitive IgG binding protein. *Proc. Natl. Acad. Sci. U. S. A.* 111: 675–80.
13. Siegel, J.B., A.L. Smith, S. Poust, A.J. Wargacki, A. Bar-Even, C. Louw, B.W. Shen, C.B. Eiben, H.M. Tran, E. Noor, J.L. Gallaher, J. Bale, Y. Yoshikuni, M.H. Gelb, J.D. Keasling, B.L. Stoddard, M.E. Lidstrom, and D. Baker. 2015. Computational protein design enables a novel one-carbon assimilation pathway. *Proc Natl Acad Sci USA*. 112: 3704–9.
14. Desmet, J., M. De Maeyer, B. Hazes, and I. Lasters. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*. 356: 539–542.
15. Dahiyat, B.I., and S.L. Mayo. 1997. De Novo Protein Design: Fully Automated Sequence Selection. *Science*. 278: 82–87.
16. Dahiyat, B.I., C. Sarisky, and S.L. Mayo. 1997. De Novo Protein Design: Towards Fully Automated Sequence Selection. *J Mol. Biol.* 273: 789–796.

17. Georgiev, I., and B.R. Donald. 2007. Dead-End Elimination with backbone flexibility. *Bioinformatics*. 23: 185–194.
18. Leaver-fay, A., M. Tyka, S.M. Lewis, F. Lange, J. Thompson, R. Jacak, K. Kaufman, P.D. Renfrew, C.A. Smith, W. Sheffler, I.W. Davis, S. Cooper, A. Treuille, D.J. Mandell, F. Richter, Y.A. Ban, S.J. Fleishman, E. Corn, D.E. Kim, S. Lyskov, M. Berrondo, J.J. Havranek, S. Mentzer, Z. Popovic, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J.J. Gray, B. Kuhlman, D. Baker, and P. Bradley. 2011. ROSETTA 3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol*. 487: 545–574.
19. Miklos, A.E., C. Kluwe, B.S. Der, S. Pai, A. Sircar, R. a Hughes, M. Berrondo, J. Xu, V. Codrea, P.E. Buckley, A.M. Calm, H.S. Welsh, C.R. Warner, M. a Zacharko, J.P. Carney, J.J. Gray, G. Georgiou, B. Kuhlman, and A.D. Ellington. 2012. Structure-based design of supercharged, highly thermoresistant antibodies. *Chem. Biol*. 19: 449–55.
20. Der, B.S., C. Kluwe, A.E. Miklos, R. Jacak, S. Lyskov, J.J. Gray, G. Georgiou, A.D. Ellington, and B. Kuhlman. 2013. Alternative computational protocols for supercharging protein surfaces for reversible unfolding and retention of stability. *PLoS One*. 8: e64363.
21. Leaver-Fay, A., K.J. Froning, S. Atwell, H. Aldaz, A. Pustilnik, F. Lu, F. Huang, R. Yuan, S. Hassanali, A.K. Chamberlain, J.R. Fitchett, S.J. Demarest, and B. Kuhlman. 2016. Computationally Designed Bispecific Antibodies using Negative State Repertoires. *Structure*. 24: 641–651.
22. Tinberg, C.E., S.D. Khare, J. Dou, L. Doyle, J.W. Nelson, A. Schena, W. Jankowski, C.G. Kalodimos, K. Johnsson, B.L. Stoddard, and D. Baker. 2013. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*. 501: 212–6.
23. Thornburg, N.J., D.P. Nannemann, D.L. Blum, J.A. Belser, T.M. Tumpey, S. Deshpande, G.A. Fritz, G. Sapparapu, J.C. Krause, J.H. Lee, A.B. Ward, D.E. Lee, S. Li, K.L. Winarski, B.W. Spiller, J. Meiler, and J.E. Crowe. 2013. Human antibodies that neutralize respiratory droplet transmissible H5N1 influenza viruses. *J. Clin. Invest*. 123: 4405–4409.
24. Nannemann, D.P., K.W. Kaufmann, J. Meiler, and B.O. Bachmann. 2010. Design and directed evolution of a dideoxy purine nucleoside phosphorylase. *Protein Eng. Des. Sel*. 23: 607–616.
25. Stranges, P.B., M. Machius, M.J. Miley, A. Tripathy, and B. Kuhlman. 2011. Computational design of a symmetric homodimer using  $\beta$ -strand assembly. *Proc. Natl. Acad. Sci. U. S. A*. 108: 20562–7.
26. Der, B.S., M. Machius, M.J. Miley, J.L. Mills, T. Szyperski, and B. Kuhlman. 2012. Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. *J. Am. Chem. Soc*. 134: 375–85.
27. Benjamin Stranges, P., and B. Kuhlman. 2013. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Sci*. 22: 74–82.
28. Sharabi, O.Z., C. Yanover, A. Dekel, and J.M. Shifman. 2011. Optimizing Energy Functions for Protein – Protein Interface Design. *J Comp Chem*. 32: 23–32.

29. Shifman, J. 2016. Computational Protein Design – Design of Protein- Protein Interactions. .
30. Lippow, S.M., K.D. Wittrup, and B. Tidor. 2007. Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.* 25: 1171–1176.
31. Tidor, B., and L.-P. Lee. 2001. Barstar is electrostatically optimized for tight binding to barnase. *Nat. Struct. Biol.* 8: 73–76.
32. Hecht, M.H. 1994. De novo design of  $\beta$ -sheet proteins. *Proc Natl Acad Sci USA.* 91: 8729–8730.
33. Kuhlman, B., G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, and D. Baker. 2003. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science.* 302: 1364–1368.
34. Protein Data Bank. .
35. Woolfson, D.N., G.J. Bartlett, A.J. Burton, J.W. Heal, A. Niitsu, A.R. Thomson, and C.W. Wood. 2015. De novo protein design: How do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol.* 33: 16–26.
36. Dali protein structure database. .
37. Lo Conte, L., B. Ailey, T.J.P. Hubbard, S.E. Brenner, A.G. Murzin, and C. Chothia. 1997. SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res.* 25: 236–239.
38. Sillitoe, I., T.E. Lewis, A. Cuff, S. Das, P. Ashford, N.L. Dawson, N. Furnham, R.A. Laskowski, D. Lee, J.G. Lees, S. Lehtinen, R.A. Studer, J. Thornton, and C.A. Orengo. 2015. CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* 43: D376–D381.
39. SCOP browser. .
40. Eisenberg, D., W. Wilcox, S.M. Eshita, P.M. Pryciak, S.P. Ho, and W.F. DeGrado. 1986. The design, synthesis, and crystallization of an alpha-helical peptide. *Proteins Struct. Funct. Genet.* 1: 16–22.
41. B.W. Erickson, S.B. Daniels, P.A. Reddy, C.G. Unson, J.S. Richardson, and D.C.R. 1986. Betabellin: An Engineered Protein. *Comput. Graph. Mol. Model.* : 53–57.
42. Rohl, B.C.A., C.E.M. Strauss, K.M.S. Misura, and D. Baker. 2004. Protein Structure Prediction Using Rosetta. *Methods Enzymol.* 383: 66–93.
43. Koga, N., R. Tatsumi-koga, G. Liu, R. Xiao, T.B. Acton, and T. Gaetano. 2012. Principles for designing ideal protein structures. *Nature.* 491: 222–229.
44. Doyle, L., J. Hallinan, J. Bolduc, F. Parmeggiani, D. Baker, B.L. Stoddard, and P. Bradley. 2015. Rational design of  $\alpha$ -helical tandem repeat proteins with closed architectures. *Nature.* 528: 585–588.
45. Jacobs, T.M., B. Williams, T. Williams, X. Xu, A. Eletsky, J.F. Federizon, T. Szyperski, and B. Kuhlman. 2016. Design of structurally distinct proteins using strategies inspired by evolution. *Science.* 352: 687–90.

46. Brunette, T., F. Parmeggiani, P.-S. Huang, G. Bhabha, D.C. Ekiert, S.E. Tsutakawa, G.L. Hura, J.A. Tainer, and D. Baker. 2015. Exploring the repeat protein universe through computational protein design. *Nature*. 528: 580–584.
47. Parmeggiani, F., P.-S. Huang, S. Vorobiev, R. Xiao, K. Park, S. Caprari, M. Su, J. Seetharaman, L. Mao, H. Janjua, G. Montelione, J. Hunt, and D. Baker. 2015. A general computational approach for repeat protein design. *J. Mol. Biol.* 427: 563–575.
48. Park, K., B.W. Shen, F. Parmeggiani, P.-S. Huang, B.L. Stoddard, and D. Baker. 2015. Control of repeat-protein curvature by computational protein design. *Nat. Struct. Mol. Biol.* 22: 167–74.
49. Joh, N.H., T. Wang, M.P. Bhate, R. Acharya, Y. Wu, M. Grabe, M. Hong, G. Grigoryan, and W.F. DeGrado. 2014. De novo design of a transmembrane Zn<sup>2+</sup>-transporting four-helix bundle. *Science*. 346: 1520–4.
50. Lin, Y., N. Koga, R. Tatsumi-Koga, G. Liu, A.F. Clouser, G.T. Montelione, and D. Baker. 2015. Control over overall shape and size in de novo designed proteins. *Proc. Natl. Acad. Sci.* 112: E5478–E5485.
51. Murphy, G.S., B. Sathyamoorthy, B.S. Der, M.C. Machius, S. V Pulavarti, T. Szyperski, and B. Kuhlman. 2015. Computational de novo design of a four-helix bundle protein - DND\_4HB. *Protein Sci.* 24: 434–445.
52. Marcos, E., and D. Baker. Unpublished yet. .
53. Huang, P.-S., K. Feldmeier, F. Parmeggiani, D.A. Fernandez Velasco, B. Höcker, and D. Baker. 2016. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat. Chem. Biol.* 12: 29–34.
54. Boyken, S.E., Z. Chen, B. Groves, R.A. Langan, G. Oberdorfer, A. Ford, J.M. Gilmore, C. Xu, F. Dimaio, J.H. Pereira, B. Sankaran, G. Seelig, P.H. Zwart, and D. Baker. 2016. De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science*. 352: 680–687.
55. Fernandez-Fuentes, N., J.M. Dybas, and A. Fiser. 2010. Structural characteristics of novel protein folds. *PLoS Comput. Biol.* 6: e1000750.
56. Kortemme, T., M. Ramírez-Alvarado, and L. Serrano. 1998. Design of a 20-amino acid, three-stranded beta-sheet protein. *Science* (80-. ). 281: 253–6.
57. Hu, X., H. Wang, H. Ke, and B. Kuhlman. 2008. Computer-Based Redesign of a  $\beta$  Sandwich Protein Suggests that Extensive Negative Design Is Not Required for De Novo  $\beta$  Sheet Design. *Structure*. 16: 1799–1805.
58. Jaobs, T. 2015. De Novo Proteins Designed from Evolutionary Principles. Ph. D. Diss. Univ. North Carolina Chapel Hill. .
59. Personal communication with Qi Zhang. .
60. Main, Y.J. and E.R.G. 2009. Exploring the folding energy landscape of a series of designed consensus tetratricopeptide repeat proteins. *Proc Natl Acad Sci USA*. 106: 17383–17388.



61. Marcos, E. 2015. personal communication. .
62. Doo Nam Kim. EnumerateAssemblyMover. .
63. Huang, P.-S., S.E. Boyken, and D. Baker. 2016. The coming of age of de novo protein design. *Nature*. 537: 320–327.
64. The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. .
65. Sgourakis, N.G., O.F. Lange, F. Dimaio, I. André, N.C. Fitzkee, P. Rossi, G.T. Montelione, A. Bax, and D. Baker. 2011. Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. *J. Am. Chem. Soc.* 133: 6288–6298.
66. Chong, S.-H., and S. Ham. 2014. Interaction with the surrounding water plays a key role in determining the aggregation propensity of proteins. *Angew. Chem. Int. Ed. Engl.* 53: 3961–4.
67. Fernandez-Leiro, R., and S.H.W. Scheres. 2016. Unravelling biological macromolecules with cryo-electron microscopy. *Nature*. 537: 339–346.
68. Eiben, C.B., J.B. Siegel, J.B. Bale, S. Cooper, F. Khatib, B.W. Shen, F. Players, B.L. Stoddard, Z. Popovic, and D. Baker. 2012. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat. Biotechnol.* 30: 190–2.
69. Pace, C.N., and J.M. Scholtz. 1998. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* 75: 422–427.
70. Spek, E.J., a H. Bui, M. Lu, and N.R. Kallenbach. 1998. Surface salt bridges stabilize the GCN4 leucine zipper. *Protein Sci.* 7: 2431–2437.
71. Anfinsen, C.B. 1973. Principles that Govern the Folding of Protein Chains. *Science* (80-. ). 181: 223–230.
72. Luo, J., A. Teplyakov, G. Obmolova, T.J. Malia, W. Chan, S.A. Jacobs, K.T. O’Neil, and G.L. Gilliland. 2014. N-terminal beta-strand swapping in a consensus-derived alternative scaffold driven by stabilizing hydrophobic interactions. *Proteins Struct. Funct. Bioinforma.* 82: 1527–1533.
73. Teplyakov, A., G. Obmolova, T.J. Malia, J. Luo, S.A. Jacobs, W. Chan, D. Domingo, A. Baker, K.T. O’Neil, and G.L. Gilliland. 2014. C-terminal  $\beta$ -strand swapping in a consensus-derived fibronectin Type III scaffold. *Proteins Struct. Funct. Bioinforma.* 82: 1359–1369.
74. Hu, X., H. Wang, H. Ke, and B. Kuhlman. 2007. High-resolution design of a protein loop. *Proc. Natl. Acad. Sci. U. S. A.* 104: 17668–73.
75. Khatib, F., S. Cooper, M.D. Tyka, K. Xu, I. Makedon, Z. Popovic, D. Baker, and F. Players. 2011. Algorithm discovery by protein folding game players. *Proc. Natl. Acad. Sci. U. S. A.* 108: 18949–53.
76. Kim, D.N., T.M. Jacobs, and B. Kuhlman. 2016. Boosting protein stability with the computational design of  $\beta$ -sheet surfaces. *Protein Sci.* 25: 702–710.
77. Gerstein, M. 1998. How representative are the known structures of the proteins in a complete

- genome? A comprehensive structural census. *Fold. Des.* 3: 497–512.
78. Lassila, K., D. Datta, and S. Mayo. 2002. Evaluation of the energetic contribution of an ionic network to beta-sheet stability. *Protein Sci.* 11: 688–690.
  79. Lawrence, M.S., K.J. Phillips, and D.R. Liu. 2007. Supercharging proteins can impart unusual resilience. *J. Am. Chem. Soc.* 129: 10110–2.
  80. Smith, C.K., and L. Regan. 1995. Guidelines for protein design: the energetics of beta sheet side chain interactions. *Science.* 270: 980–982.
  81. Minor, D., and P. Kim. 1994. Context is a major determinant of  $\beta$ -sheet propensity. *Nature.* 371: 264–267.
  82. Riemen, A.J., and M.L. Waters. 2009. Design of Highly Stabilized beta-Hairpin Peptides through Cation- $\pi$  Interactions of Lysine and N-Methyllysine with an Aromatic Pocket. *Biochemistry.* 48: 1525–1531.
  83. Kiehna, S.E., and M.L. Waters. 2003. Sequence dependence of  $\beta$ -hairpin structure: Comparison of a salt bridge and an aromatic interaction. *Protein Sci.* 12: 2657–2667.
  84. Street, A.G., and S.L. Mayo. 1999. Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc. Natl. Acad. Sci. U. S. A.* 96: 9074–6.
  85. Hu, X., H. Wang, H. Ke, and B. Kuhlman. 2008. Computer-based redesign of a beta sandwich protein suggests that extensive negative design is not required for de novo beta sheet design. *Structure.* 16: 1799–1805.
  86. Shane Gonen, Frank DiMaio, Tamir Gonen, D.B. 2015. Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science.* 348: 1365–1368.
  87. Street, A.G., D. Datta, D.B. Gordon, and S.L. Mayo. 2000. Designing protein beta-sheet surfaces by Z-score optimization. *Phys. Rev. Lett.* 84: 5010–5013.
  88. Leahy, D.J., W. a Hendrickson, I. Aukhil, and H.P. Erickson. 1992. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science.* 258: 987–91.
  89. Lappalainen, I., M.G. Hurley, and J. Clarke. 2008. Plasticity Within the Obligatory Folding Nucleus of an Immunoglobulin-like Domain. *J. Mol. Biol.* 375: 547–559.
  90. Ng, S.P., K.S. Billings, L.G. Randles, and J. Clarke. Manipulating the stability of fibronectin type III domains by protein engineering. 384023.
  91. Gilbreth, R., and B. Chacko. 2014. Stabilization of the third fibronectin type III domain of human tenascin-C through minimal mutation and rational design. *Protein Eng. Des. Sel.* 27: 411–418.
  92. Jacobs, S.A., M.D. Diem, J. Luo, A. Teplyakov, G. Obmolova, T. Malia, G.L. Gilliland, and K.T.O. Neil. 2012. Design of novel FN3 domains with high stability by a consensus sequence approach. 25: 107–117.

93. Strickler, S.S., A. V Gribenko, A. V Gribenko, T.R. Keiffer, J. Tomlinson, T. Reihle, V. V Loladze, and G.I. Makhatadze. 2006. Protein stability and surface electrostatics: a charged relationship. *Biochemistry*. 45: 2761–6.
94. Leaver-Fay, A., M.J. O'Meara, M. Tyka, R. Jacak, Y. Song, E.H. Kellogg, J. Thompson, I.W. Davis, R. a. Pache, S. Lyskov, J.J. Gray, T. Kortemme, J.S. Richardson, J.J. Havranek, J. Snoeyink, D. Baker, and B. Kuhlman. 2013. Scientific benchmarks for guiding macromolecular energy function improvement. *Methods Enzymol*. 523: 109–143.
95. O'Meara, M.J., A. Leaver-Fay, M.D. Tyka, A. Stein, K. Houlihan, F. DiMaio, P. Bradley, T. Kortemme, D. Baker, J. Snoeyink, and B. Kuhlman. 2015. Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. *J. Chem. Theory Comput*. 11: 609–622.
96. O'Shea, E.K., K.J. Lumb, and P.S. Kim. 1993. Peptide “Velcro”: design of a heterodimeric coiled coil. *Curr. Biol*. 3: 658–667.
97. Ruczinski, I., C. Kooperberg, R. Bonneau, and D. Baker. 2002. Distributions of beta sheets in proteins with application to structure prediction. *Proteins Struct. Funct. Genet*. 48: 85–97.
98. Zhu, H., and W. Braun. 1999. Sequence specificity, statistical potentials, and three-dimensional structure prediction with self-correcting distance geometry calculations of beta-sheet formation in proteins. *Protein Sci*. 8: 326–342.
99. Minor, D.L., and P.S. Kim. 1994. Measurement of the beta-sheet-forming propensities of amino acids. *Nature*. 367: 660–663.
100. Fujiwara, K., H. Toda, and M. Ikeguchi. 2012. Dependence of alpha-helical and beta-sheet amino acid propensities on the overall protein fold type. *BMC Struct. Biol*. 12: 18.
101. Tyka, M.D., D.A. Keedy, I. André, F. Dimaio, Y. Song, D.C. Richardson, J.S. Richardson, and D. Baker. 2011. Alternate States of Proteins Revealed by Detailed Energy Landscape Mapping. *J. Mol. Biol*. 405: 607–618.
102. Walther, T.H., C. Gottselig, S.L. Grage, M. Wolf, A. V Vargiu, M.J. Klein, S. Vollmer, S. Prock, M. Hartmann, S. Afonin, E. Stockwald, H. Heinzmann, O. V Nolandt, W. Wenzel, P. Ruggerone, and A.S. Ulrich. 2013. Folding and Self-Assembly of the TatA Translocation Pore Based on a Charge Zipper Mechanism. *Cell*. 152: 316–326.
103. Kuhlman, B., and D. Baker. 2000. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. U. S. A*. 97: 10383–10388.
104. Prism, GraphPad Software, San Diego California USA ([www.graphpad.com](http://www.graphpad.com)). .
105. Mathematica, Wolfram Research, Inc. Champaign Illinois USA (<http://www.wolfram.com/mathematica>). .
106. Merkel, J.S., J.M. Sturtevant, and L. Regan. 1999. Sidechain interactions in parallel  $\beta$  sheets: the energetics of cross-strand pairings. *Structure*. 7: 1333–43.
107. Orengo, C. a, a D. Michie, S. Jones, D.T. Jones, M.B. Swindells, and J.M. Thornton. 1997. CATH

- a hierarchic classification of protein domain structures. *Structure*. 5: 1093–1108.
108. Kraemer-Pecore, C., J. Lecomte, and J. Desjarlais. 2003. A de novo redesign of the WW domain. *Protein Sci.* 12: 2194–2205.
  109. Kuhn, M., J. Meiler, and D. Baker. 2004. Strand-Loop-Strand Motifs : Prediction of Hairpins and Diverging Turns in Proteins. *Proteins Struct. Funct. Bioinforma.* 54: 282–288.
  110. Bonneau, R., I. Ruczinski, J. Tsai, and D. Baker. 2002. Contact order and ab initio protein structure prediction. *Protein Sci.* 11: 1937–1944.
  111. Cartoon representation of Ig domains in proteins of the immune system. .
  112. Wang, J.H.-C., B.P. Thampatty, J.-S. Lin, and H.-J. Im. 2007. Mechanoregulation of gene expression in fibroblasts. *Gene*. 391: 1–15.
  113. Ruoslahti, E., and J.C. Reed. 1994. Anchorage Dependence, Integrins and Apoptosis. *Cell*. 77: 477–478.
  114. Sottile, J., D.C. Hocking, and P.J. Swiatek. 1998. Fibronectin matrix assembly enhances adhesion-dependent cell growth. *J. Cell Sci.* 111: 2933–43.
  115. Cota, E., S.J. Hamill, S.B. Fowler, and J. Clarke. 2000. Two proteins with the same structure respond very differently to mutation: the role of plasticity in protein stability. *J. Mol. Biol.* 302: 713–25.
  116. Geierhaas, C.D., E. Paci, M. Vendruscolo, and J. Clarke. 2004. Comparison of the transition states for folding of two Ig-like proteins from different superfamilies. *J. Mol. Biol.* 343: 1111–1123.
  117. Cota, E., A. Steward, S.B. Fowler, and J. Clarke. 2001. The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold. *J. Mol. Biol.* 305: 1185–94.
  118. Gee, E.P.S., D.E. Ingber, and C.M. Stultz. 2008. Fibronectin unfolding revisited: modeling cell traction-mediated unfolding of the tenth type-III repeat. *PLoS One*. 3: e2373.
  119. Ohashi, T., and H.P. Erickson. 2011. Fibronectin aggregation and assembly: the unfolding of the second fibronectin type III domain. *J. Biol. Chem.* 286: 39188–99.
  120. Yu, T.-K., S.-A. Shin, E.-H. Kim, S. Kim, K.-S. Ryu, H. Cheong, H.-C. Ahn, S. Jon, and J.-Y. Suh. 2014. An Unusual Protein-Protein Interaction through Coupled Unfolding and Binding. *Angew. Chemie*. 53: 9784–9787.
  121. Hu, X. 2008. Computational design of  $\beta$ -sheet proteins. Ph. D. Diss. Univ. North Carolina Chapel Hill. .
  122. Kim, D.N. 2011. StrandBundleFeatures. .
  123. top 8000. .
  124. Richardson, J.S., and D.C. Richardson. 2002. Natural  $\beta$ -sheet proteins use negative design to avoid

- edge-to-edge aggregation. *Proc Natl Acad Sci USA*. 99: 2754–2759.
125. Hoxha, E., and S. Campion. 2014. Structure–Critical Distribution of Aromatic Residues in the Fibronectin Type III Protein Family. *Protein J*. 33: 165–173.
  126. Hamill, S.J., E. Cota, C. Chothia, and J. Clarke. 2000. Conservation of folding and stability within a protein family: the tyrosine corner as an evolutionary cul-de-sac. *J. Mol. Biol.* 295: 641–649.
  127. DiMaio, F., Y. Song, X. Li, M.J. Brunner, C. Xu, V. Conticello, E. Egelman, T.C. Marlovits, Y. Cheng, and D. Baker. 2015. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nat. Methods*. 12: 361–5.
  128. Hu, X., H. Wang, H. Ke, and B. Kuhlman. 2008. Computer-based redesign of a beta sandwich protein suggests that extensive negative design is not required for de novo beta sheet design. *Structure*. 16: 1799–805.
  129. Shaw, D.E. 2016. Interview before 2016 National Lecturer with David E. Shaw, PhD. .
  130. Ranganathan, R. 2014. A Model for Protein Design. .
  131. Gelfand, A. 2016. The Rise of Cryo-Electron Microscopy. *Biomed. Rev.* April 1st: 13–21.
  132. 2016. Potential for further improvements in single-particle electron cryomicroscopy. .
  133. Vinothkumar, K.R. 2015. Membrane protein structures without crystals, by single particle electron cryomicroscopy. *Curr. Opin. Struct. Biol.* 33: 103–114.
  134. 2015. Phenix Tools for Validated Refinement of Atomic Models into maps (low-resolution, Cryo-EM, X-ray or neutron). .
  135. Fleishman, S.J., and D. Baker. 2012. Role of the biomolecular energy gap in protein design, structure, and evolution. *Cell*. 149: 262–273.
  136. Alvizo, O., and S.L. Mayo. 2008. Evaluating and optimizing computational protein design force fields using fixed composition-based negative design. *Proc. Natl. Acad. Sci. U. S. A.* 105: 12242–12247.
  137. OptE score function optimization. .
  138. Conway, P., and F. DiMaio. 2016. Improving hybrid statistical and physical forcefields through local structure enumeration. *Protein Sci.* 25: 1525–1534.
  139. Doo Nam Kim. SandwichFeatures. .
  140. Porebski, B.T., A.A. Nickson, D.E. Hoke, M.R. Hunter, L. Zhu, S. McGowan, G.I. Webb, and A.M. Buckle. 2015. Structural and dynamic properties that govern the stability of an engineered fibronectin type III domain. *Protein Eng. Des. Sel.* 28: 67–78.
  141. Zhang, C., and S.H. Kim. 2000. A comprehensive analysis of the Greek key motifs in protein beta-barrels and beta-sandwiches. *Proteins*. 40: 409–19.
  142. Foit, L., G.J. Morgan, M.J. Kern, L.R. Steimer, A. a von Hacht, J. Titchmarsh, S.L. Warriner, S.E.

- Radford, and J.C. a Bardwell. 2009. Optimizing protein stability in vivo. *Mol. Cell.* 36: 861–71.
143. Cox, N., D. Pilling, and R.H. Gomer. 2012. NaCl potentiates human fibrocyte differentiation. *PLoS One.* 7: e45674.
  144. Hottest and coldest part of the Human Body. .
  145. Seth T Gammon, W.M.L., and M. Loechner. 2009. Carestream Molecular Imaging: imaging of cancer biology and relevant pathways in vivo. *Nat. Method.* 6: an4-an5.
  146. Broom, A., S.M. Ma, K. Xia, H. Rafalia, K. Trainor, W. Colón, S. Gosavi, and E.M. Meiering. 2015. Designed protein reveals structural determinants of extreme kinetic stability. *Proc. Natl. Acad. Sci. U. S. A.* 112: 14605–10.
  147. Broom, A., S. Gosavi, and E.M. Meiering. 2015. Protein unfolding rates correlate as strongly as folding rates with native structure. *Protein Sci.* 24: 580–587.
  148. Bagger, H.L., C.C. Fuglsang, and P. Westh. 2006. Hydration of a glycoprotein: Relative water affinity of peptide and glycan moieties. *Eur. Biophys. J.* 35: 367–371.
  149. Sood, V.D., and D. Baker. 2006. Recapitulation and design of protein binding peptide structures and sequences. *J. Mol. Biol.* 357: 917–927.
  150. Sammond, D.W., D.E. Bosch, G.L. Butterfoss, C. Purbeck, M. Machius, D.P. Siderovski, and B. Kuhlman. 2011. Computational Design of the Sequence and Structure of a Protein-Binding Peptide. *J. Am. Chem. Soc.* 133: 4190–4192.
  151. Lab, K. 2016. Refinement-of-SEWING-Assemblies. .
  152. Fleishman, S.J., A. Leaver-Fay, J.E. Corn, E.M. Strauch, S.D. Khare, N. Koga, J. Ashworth, P. Murphy, F. Richter, G. Lemmon, J. Meiler, and D. Baker. 2011. Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. *PLoS One.* 6: 1–10.