ADVANCES IN FORCE FIELD DEVELOPMENT AND SEQUENCE OPTIMIZATION METHODS FOR COMPUTATIONAL PROTEIN DESIGN

Thesis by

Premal S. Shah

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2005

(Defended March 30, 2005)

This thesis is dedicated to

D. J. Patel
and
R. C. Shah

Two men whose lives were defined by hard work, love for their families, and an unwavering desire to help everyone. They continue to inspire.

Acknowledgements

Only when I started thinking of writing this section of my thesis did I begin to truly appreciate the number of people who have influenced and contributed to my life at Caltech. So many people, both at Caltech and in the "real world" have made a difference in my life and my hope is that I do everyone justice here.

I would like to thank my advisor, Steve Mayo, for giving me the opportunity to work and learn in his lab. Steve's mentoring philosophy can be likened to that of Linus Pauling: fall on your own, get up on your own, and grow on your own. This tough love methodology is supplemented with Steve's amazing talent as a scientist, his business-savvy approach to today's science, and his desire to assure the comfort of his students. I have learned so much from him and hopefully, I have contributed to the overall goal of his lab.

The members of my thesis advisory committee, Profs Doug Rees, Frances Arnold, and Niles Pierce, have been a valuable resource; they've always been willing to answer my questions. It's been especially nice having Doug and Niles in the Broad Center. Knowing they're an arms length away is a luxury I haven't taken for granted.

I've always claimed that a research lab is the equivalent of a dysfunctional family. No place is this more true than at Caltech. A fusion of highly competent, opinionated, ambitious people can only make for some fond memories. I've had the good fortune of being surrounded by the world's most amazing (and interesting) people.

My two closest colleagues in lab, Possu Huang and Geoffrey Hom, have been with me from day one. We've shared trips to Vegas, Japan, and various other establishments. Possu is one of the best molecular biologists I know and Apple

Computers' most valuable public relations guy. Geoff is responsible for my love of poker and moving me towards the correct side of the moral spectrum. Both of them provide endless entertainment. Whether it's deciphering Possu's unique speaking ability or marveling at the many different poses Geoff strikes while sleeping under his desk, both of these guys have been an absolute joy to be around.

Rhonda Digiusto has been a friend above all, but also an irreplaceable resource in lab. She has always listened to my rants about the ups and downs in my life, answered early morning calls for assistance in lab, and made some of the best desserts I've ever had. It's safe to say we all dreaded the days Rhonda wasn't in lab. Not to be outdone in the dessert competition is Cynthia Carlson. Not only were her brownie treats the highlight of some weeks, Cynthia's organizational skills are in a league of their own; my friends around campus marvel at how smooth our existence in lab is because of her efforts. Along with Rhonda and Cynthia, Marie Ary completes the lab's trio of guardian angels. To say Marie helped me with my scientific writing would be a huge understatement. At some stretches, Marie probably spent more time with me than her own kids. If this thesis is a testament to anything, it's Marie's ability to teach the art of scientific writing.

The other person in lab who has been with me from the beginning is Scott Ross. Scott's scientific ability as an NMR spectroscopist is unmatched, but it is his infectious personality that makes him so pleasant to be around. Whether it's his stories of tasseling corn in Iowa or anecdotes from his own Caltech graduate days, Scott does everything with a smile and at 100%. He was a collaborator on several projects and is one of the best teachers I've ever been around. Scott always took the time to answer my questions regarding NMR spectroscopy and life in general and I thank him for that.

I've had the pleasure of growing close to several other members of Steve's lab. Jessica Mao was my neighbor in Braun labs and served as a personal psychologist. Jessica is one of the nicest people the world has to offer; never have I seen someone so upbeat and pleasant. Kyle Lassila and Eric Zollars have never let me get away with anything. They've kept me on my toes and served as sounding boards for late night rants following long and frustrating days. Kyle's bacteria spreader and ethanol stock have helped me on many occasions and Eric's upfront, in your face attitude can be refreshing in a world often plagued with false praise. They are also co-founders of the eppendorf-toss attention grabbing system which to this day drives me crazy.

I would not have been able to rise above the trauma of the 2004 Presidential election if it were not for Christina Vizcarra. Christina is one of the most informed people around and we've shared countless hours pondering the state of our country. Even though she's declined on several occasions, I'm confident Christina will run an effective campaign for me in my first congressional run.

John Love is the oldest kid I know. John's upside-down face is just one of many charming characteristics. He and I shared many good times together, and even after he left for Shangri-la, we've kept in touch and been good friends.

So many others in Steve's lab have influenced me. A complete list of all lab members who have contributed to this thesis appears at the end of this section.

One of the great things about Caltech is the intimate atmosphere of the campus. As a result I've grown close to many people outside the lab. Chris Otey was one of the first people I met when I arrived and we've been great friends right from the beginning. In addition to being a roommate for three years, he's been the defensive coordinator on

my football team, a reliable scientific colleague, and a compassionate friend. Chris is also responsible for my wasting countless days with incessant instant messaging. It's safe to say everyone needs a Chris in graduate school. I'm incredibly fortunate to have met him.

I met Angie Mah the same time I met Chris. Angie is truly unique in all aspects of life. An all business type of girl, Angie and I have cried together, laughed over the silliest things, and spent years worrying about our futures. A mention of Angie without commenting on her dog, Mackenzie, would be inappropriate. Mack is the only dog I've ever liked and I know Angie wouldn't be offended if I say I'm going to miss him more than her.

In the last couple of years I've become incredibly close to a couple of engineering guys who have educated me in the ways of wine, women, and their own "scientific method." Adam Olsen and Mohan Sankaran have made my last couple of years at Caltech an absolute joy ride. We've spent hours talking about the most ridiculous things in life, usually accompanied with several bottles of really cheap wine. What is so unique about our relationship is that each of us has a very different lifestyle and our motivations in life are completely dissimilar. It's this unique blend that makes us enjoy each other's company so much. I'm confident that even though we move forward in our own separate ways, we'll remain the best of friends.

Growing up in the Washington D.C. area, I was inundated with political talk shows. One of these was the "McLaughlin Group" and I always fantasized that one day I too would sit around a table discussing the political events of the week with other informed people. Well, even though we don't have our own TV show, Ryan Austin, Anders Olson, Terry Takahashi, Adam Frankel (all from Rich Roberts' lab) and I spent

lunch after lunch discussing the day's headlines and other discussion-worthy stories. Although the conversations often wavered towards the absurd, the bottom line was always the same: we had a great time. A mention of lunch without acknowledging Ernie would be wrong. Ernie is a Caltech establishment and has been the source of fuel for almost everyone at one point or another. Other members of Rich's lab that became good friends are Bill Ja, Shelley Starck, and Chris Balmaseda. Shelley's husband, Harry Green is also a friend.

Members of Niles Pierce's lab have been great floor-mates in the Broad Center. They're always there when you need advice on anything from science to the Tour de France. Robert Dirks and Justin Bois especially, have been great colleagues.

Although I've known her for only a short time, Ami Badani has been a welcome addition to my life. Her free-spirited and caring personality blends seamlessly with her ambition to be the best in all aspects of life. I thank her for the added perspective she has shared with me and her loving support; saying Ami is special is an enormous understatement.

Pradman Qasba was my first research mentor. His warm, caring attitude towards teaching helped me learn the fundamentals of research and propelled me to become excited about science.

Anyone who knows me can attest to how grateful I am for the support I receive from my family. My parents, Sunil and Rashmi Shah, have been nothing short of massive pillars of support throughout my life. They've taught me to look at the good in the world and always strive to be the best person I can be, regardless of the end result. They've let

me be my own person without letting me sever ties with my roots and culture. If I can be a tenth of the parents they are, I'll be a success.

My brother, Pratik Shah, is one of my best friends. Ironically, the farther I've moved away from him, the closer we've become. We've always been told we have completely different personalities, but what I've recently become increasingly aware of is just how similar we are. Among other things, Pratik and I share a love for sports that is downright freakish. Countless hours have been spent dreaming about another Redskins Super Bowl and an Orioles World Series. My return to the Washington DC area was motivated mostly by my desire to be close to my brother. Everyone should be as fortunate to have a brother who is such a good friend.

The West Coast part of the family has gone beyond the call of duty to make me feel comfortable and welcome. My aunts, Bharti Patel and Nayan Patel, and their families have welcomed me as one of their own. I am so grateful for their love and support and leaving them is one of the unfortunate consequences of moving.

I have too many cousins to exhaustively acknowledge here, but I want to thank Hemang and Vandana Patel for their support. Even from 3000 miles away, I felt I never really left them. I eagerly look forward to being around them again upon my return to their area. Also, here in Los Angeles, I've come to really know another cousin, Anup Patel. Growing up, we really didn't get to know each other, but my time here has afforded me the opportunity to become close to him. I hope that even after I leave, we continue to remain friends.

Finally, I want to mention my late grandmother, Champaben Patel. Never will I meet someone so devoted to assuring the well-being of her family. The only thing that

mattered to her was my happiness. Accolades and successes were never important to her; she loved all of her children unconditionally. I know she is looking down, proud to be a part of this moment.

Caltech has been an amazing place to do graduate work. Recruiting efforts here always emphasize the uniqueness of the institute that is "the size of a high school with the resources of Harvard." It is absolutely true. The relationships I've forged in my time here will stay with me for a lifetime. To anyone I've failed to mention, I apologize but thank you for your support.

Other members of the Mayo lab (in order of appearance):

Pavel Strop, Ben Gordon, Niles Pierce, Cathy Sarisky, Dan Bolon, Shannon Marshall, Chris Voigt, Julie Mayo, JJ Plecs, Dee Datta, Julia Shifman, Shira Jacobson-Rogers, Eun-Jung Choi, Oscar Alvizo, Josh Marcus, Peter Oelschlaeger, Ben Allen, Karin Crowhurst, Tom Treynor, Heidi Privett, Jennifer Keeffe, Mary Devlin, and Sarah Hamilton.

Abstract

The overall goals of computational protein design range from designing new protein folds and protein-protein interfaces to the *de novo* design of enzymes. All goals require that two equally challenging components of computational protein design be addressed. First, the physical model that describes a protein's intermolecular and intramolecular interactions must be accurate. Second, energetically optimal amino acid sequences must be identified from an enormous number of possibilities. This thesis describes work that makes progress in both these arenas. In addition, the effectiveness and applicability of computational protein design is demonstrated by tackling challenging design problems.

Improvements to the physical model have been made by developing a more accurate method for calculating rotamer (amino acid side-chain conformation) surface areas for use in our surface area-based hydrophobic solvation term. With this method, surface area errors were decreased dramatically and the experimental stabilities of proteins generated from computationally predicted sequences were improved. Also, our direct surface area calculation approach significantly reduced the compute time required for sequence optimization using dead-end elimination (DEE)-based algorithms.

Although DEE-based algorithms have been effectively used for many challenging design problems, the daunting task of sequence optimization can cause even the most efficient DEE-based methods to fail. We developed a sequence optimization technique called Vegas that combines elements of non-DEE-based as well as DEE-based algorithms. For design problems that were already tractable using DEE-based methods, Vegas delivered the GMEC in significantly less time. In cases where DEE-based

algorithms stalled and failed to deliver the GMEC, Vegas produced an answer that, at the time, was better than any other algorithm. This is illustrated by Vegas' solution to a challenging problem: the full sequence design of a 51-residue fragment of the *Drosophila* engrailed homeodomain (ENH). We generated a variant of ENH predicted by Vegas and compared its thermodynamic properties with a protein obtained using a Monte Carlo search. We found that the thermodynamic properties of the two molecules were identical. We also solved the solution structure of the Vegas-based molecule using nuclear magnetic resonance (NMR) spectroscopy and found that it folded accurately into the target fold.

Obtaining water soluble variants of membrane proteins might alleviate some of the problems encountered when working with them and facilitate our understanding of the different forces contributing to protein stabilities in membranes. We made progress in developing an automated design scheme that can generate water soluble variants of membrane proteins. We analyzed and compared the surfaces of membrane proteins and water soluble proteins, and developed a metric for altering membrane protein surfaces. Using this metric, we can design membrane protein surfaces using the ORBIT suite of protein design algorithms and convert them to those resembling water soluble protein surfaces. We tested this strategy on two proteins and although we have not been completely successful, we have established rules and guidelines that will aid future efforts towards achieving this goal.

Table of Contents

List of Tables and Figures

Tables

Figures

Abbreviations

GMEC	global minimum energy conformation
ORBIT	optimization of rotamers by iterative techniques
vdW	van der Waals
DEE	dead-end elimination
MC	Monte Carlo
SCMF	self-consistent mean field
GA	genetic algorithms
NMR	nuclear magnetic resonance
IMP	integral membrane protein
ENH	engrailed homeodomain
Gβ1	β1 domain of protein G
HERO	hybrid exact rotamer optimization
T2-solvation	type 2 solvation
T4-solvation	type 4 solvation
ΔG_{unfold}	free energy of unfolding
T_m	melting temperature
HEWL	hen egg white lysozyme
ANS	1-anilino-napthalene-8-sulfonate
NOE	nuclear overhauser effect
CD	circular dichroism
GPCR	G-protein coupled receptor
BR	bacteriorhodopsin

Chapter 1

Introduction

In 1954, Anfinsen and colleagues observed that a protein's amino acid sequence determines its folding pattern.[1] This discovery led to the creation of the field of protein structure prediction in which researchers attempt to identify a protein's three-dimensional structure from just its amino acid sequence. To achieve solutions to what is known as the protein folding problem, researchers strive to understand the molecular forces that drive proteins into their biological native states and the thermodynamic forces that govern a protein's intramolecular and intermolecular interactions. Great strides have been made in various aspects of protein structure prediction, yet 50 years after Anfinsen's studies, the field remains far from achieving its goal.

Whereas researchers tackling the protein folding problem attempt to predict the three-dimensional structure of a protein, the field of protein design seeks to identify optimal amino acid sequences that are compatible with an already known experimentally determined target fold. Although an incredibly challenging problem, protein design is slightly easier than the protein folding problem because of the large degeneracy associated with it (Figure 1-1); only one structure is associated with a given amino acid sequence, but a large number of amino acid sequences are compatible with a target fold.

Like protein structure prediction, the protein design paradigm is used to gain a better understanding of the physical forces that contribute to protein structure stabilization. In addition, protein design has been used to design novel protein folds,[2] introduce catalytic activity onto inert scaffolds,[3,4] and increase the thermodynamic stability of target folds.[5-8] Selective binding to ligands has also been performed using protein design[9] and more recently, the *de novo* design of protein-protein dimers has been

attempted.[10] Protein design techniques have also been utilized to increase the efficiency of directed evolution experiments.[11-14]

There are two main equally challenging components of protein design. First, the physical model describing the interactions of amino acid sidechains and the forces contributing to protein stability must be accurate. Second, sequences that provide the most energetically favorable structures must be identified from an enormous number of possible sequences.

The ORBIT (Optimization of Rotamers By Iterative Techniques) suite of algorithms was developed for the automated design of proteins.[15] ORBIT uses physically-derived terms inspired by the DREIDING[16] force field and includes terms for van der Waals (vdW) interactions, hydrogen bonding, electrostatics, and surface-area based hydrophobic solvation. The vdW term[17] contains a long range attractive component and a short range repulsive component and can be calculated using a Lennard-Jones 6-12 potential. The hydrogen bond scoring function[18] has a distance-dependent term and an angle-dependent term. The electrostatics term is Coulomb's law and approximates the interaction of two point charges in the presence of other charges or dipoles. The hydrophobic effect and the interaction of sidechains with solvent are considered in the surface area-based hydrophobic solvation term.[15,19] Nonpolar surface area burial is benefited while polar burial is penalized. A negative design term that penalizes nonpolar surface area exposure is also included.

Identifying the most favorable sequence in a protein design problem is a daunting computational challenge. Even a relatively small protein of 100 amino acids has 20^{100} ($\sim 10^{130}$) possible solutions. When different conformations of amino acids, or rotamers

(Figure 1-2),[20-22] are considered, the complexity is increased further. As a result, despite considerable progress in computing power, a computational exhaustive search of sequence space is nearly impossible. Sequence optimization in ORBIT is based primarily on the dead-end elimination (DEE)[23] theorem. Algorithms based on DEE[15,24-28] are useful because solutions are guaranteed to be the global minimum energy conformation (GMEC). This is desirable when making improvements to the physical model because it assures that the model is not compromised by identifying a non-optimal solution.[15] Also, if the physical model is accurate, the GMEC sequence provides the optimal stability for the target fold.

DEE-based algorithms have contributed to the success of many challenging design problems;[4,5,8,9,29] however, more ambitious designs can sometimes cause even the most effective DEE-based algorithms to stall. In such cases, alternative approaches may be employed. Algorithms based on Monte Carlo (MC) methods,[30,31] self-consistent mean field (SCMF) techniques,[32,33] and genetic algorithms (GA)[34,35] have all been effectively implemented. The major drawback when using these algorithms is that solutions are not guaranteed to be the GMEC.

The work detailed in this thesis contributes to both of the challenging aspects of protein design described above. Improvements to the physical model have been made by developing a more accurate model for the calculation of rotamer surface areas for the surface-area based hydrophobic solvation term. The previous model relied on a pairwise calculation of rotamer surface areas[15,19]—a restriction imposed by DEE-based algorithms.[23] While extremely effective, a pairwise decomposition of rotamer surfaces will over-count surface area burial while under-counting surface area exposure (Figure

1-3). Our model calculates rotamer surface areas directly from multi-body approximations generated from MC searches, thereby eliminating the need for pairwise approximations. We demonstrated both *in silico* and *in vitro* that our new model is more accurate and effective than the previous model; designed sequences had less surface area errors and led to proteins with improved stabilities. An added feature of using our new model is that compute times for DEE-based algorithms are significantly reduced.

The limitations of the sequence optimization algorithms mentioned above were also addressed. We developed a sequence optimization technique called Vegas that combines elements of non-DEE-based as well as DEE-based algorithms (Figure 1-4). For design problems that were already tractable using DEE-based methods, Vegas delivered the GMEC in significantly less time. In cases where DEE-based algorithms stalled and failed to deliver the GMEC, Vegas produced an answer that, at the time, was better than any other algorithm.

This is illustrated by Vegas' solution to a challenging problem: the full sequence design of a 51-residue fragment of the *Drosophila* engrailed homeodomain. Full sequence designs are difficult problems because of their enormous computational complexity; they also serve as robust tests for the accuracy of the physical model. When we commenced our work, only one full sequence design had been performed for which both thermodynamic and structural information had been obtained.[29] We generated a protein predicted by Vegas and compared its thermodynamic properties with a protein obtained using an MC search. We found that the thermodynamic properties of the two molecules were identical. We also solved the solution structure of the Vegas-based molecule using nuclear magnetic resonance (NMR) spectroscopy and found that it folded

accurately into the target fold. In the conclusions, we comment on the need to obtain the GMEC for challenging design problems.

An interesting challenge in structural biology is converting a membrane protein to a water soluble protein without altering the protein's structure. Membrane proteins serve many critical roles in cellular operations (Figure 1-5). These include acting as receptors for ligands, ion pumps, channels, and transport proteins. They also contribute to maintaining the structural integrity of the cell and mediate intercellular interactions. Membrane proteins also work as enzymes involved in metabolism and aid in cellular defense mechanisms.

The importance of this group of proteins might lead one to believe much structural work has been done. However, the problems associated with working with membrane proteins have prevented researchers from obtaining the plethora of structural data that has become so readily available for water soluble proteins. These problems include limited levels of expression, poor stabilities in detergent solutions, and greater difficulties in obtaining high-quality crystals that diffract well. Obtaining water soluble variants might alleviate some of these problems and facilitate our understanding of the different forces contributing to protein stabilities in membranes. For these reasons, it is desirable to develop an automated design scheme that can generate water soluble variants of membrane proteins.

We analyzed and compared the surfaces of membrane proteins and water soluble proteins, and developed a metric for altering membrane protein surfaces. Using this metric, we can design membrane protein surfaces using the ORBIT suite of protein design algorithms and convert them to those resembling water soluble protein surfaces.

We tested this strategy on two proteins and although we have not been completely successful, we feel our strategy effectively builds on previous work and makes progress towards the goal of designing water soluble variants of membrane proteins.

The last part of this thesis details work done in collaboration with Kevin Plaxco's laboratory. *De novo* designs performed with ORBIT are ideal for studying certain aspects of protein folding because our design algorithms lack the constraints or selective pressures for protein folding that are imposed in nature. Questions such as, "Are proteins selected based on folding rates or thermodynamic stability?" can be answered using *de novo* designed proteins. We found that our proteins, which were designed for improved thermodynamic stability, had significantly higher folding rates than the wild type, suggesting that nature's proteins select for stability rather than high folding rates. The way in which proteins confer their stability can also be studied using *de novo* designed proteins.

References

1. Anfinsen, C. B., Redfield, R. R., Choate, W. L., Page, J. & Carroll, W. R. (1954). Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *J Biol Chem* 207, 201-10.

2. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L. & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science* 302, 1364-8.

3. Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2004). Computational design of a biologically active enzyme. *Science* 304, 1967-71.

4. Bolon, D. N. & Mayo, S. L. (2001). From the Cover: Enzyme-like proteins by computational design. *Proc Natl Acad Sci U S A* 98, 14274-14279.

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

6. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* 305, 619-31.

7. Marshall, S. A., Morgan, C. S. & Mayo, S. L. (2002). Electrostatics Significantly Affect the Stability of Designed Homeodomain Variants. *J Mol Biol* 316, 189-199.

8. Morgan, C. S. (2000). Full sequence design of an alpha-helical protein and investigation of the importance of helix dipole and capping effects in helical protein design, California Institute of Technology.

9. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature* 423, 185-90.

10. Shukla, U. J., Marino, H., Huang, P. S., Mayo, S. L. & Love, J. J. (2004). A designed protein interface that blocks fibril formation. *J Am Chem Soc* 126, 13914-5.

11. Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z. G. (2001). Computational method to reduce the search space for directed protein evolution. *Proc Natl Acad Sci U S A* 98, 3778-83.

12. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nat Struct Biol* 9, 553-8.

13. Meyer, M. M., Silberg, J. J., Voigt, C. A., Endelman, J. B., Mayo, S. L., Wang, Z. G. & Arnold, F. H. (2003). Library analysis of SCHEMA-guided protein recombination. *Protein Sci* 12, 1686-93.

14. Otey, C. R., Silberg, J. J., Voigt, C. A., Endelman, J. B., Bandara, G. & Arnold, F. H. (2004). Functional evolution and structural conservation in chimeric

cytochromes p450: calibrating a structure-guided approach. *Chem Biol* 11, 309-18.

15. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci* 5, 895-903.

16. Mayo, S. L., Olafson, B. D. & Goddard III, W. A. (1990). DREIDING: A generic force field for molecular simulations. *The Journal of Physical Chemistry* 94, 8897-8909.

17. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* 94, 10172-7.

18. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci* 6, 1333-7.

19. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Fold Des* 3, 253-8.

20. Janin, J. & Wodak, S. (1978). Conformation of amino acid side-chains in proteins. *J Mol Biol* 125, 357-86.

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

22. Dunbrack, R. L., Jr. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* 230, 543-74.

23. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356, 539-542.

24. Gordon, D. B. & Mayo, S. L. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J Comput Chem* 19, 1505-1514.

25. Gordon, D. B. & Mayo, S. L. (1999). Branch-and-terminate: a combinatorial optimization algorithm for protein design. *Structure Fold Des* 7, 1089-98.

26. Pierce, N. A., Spriet, J. A. & Mayo, S. L. (2000). Conformational splitting: A more powerful criterion for dead-end elimination. *J Comput Chem* 21, 999-1009.

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

28. Gordon, D. B., Hom, G. K., Mayo, S. L. & Pierce, N. A. (2003). Exact rotamer optimization for protein design. *J Comput Chem* 24, 232-243.

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

30. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J Chem Phys* 21, 1087-1092.

31. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671-680.

32. Koehl, P. & Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol* 239, 249-75.

33. Koehl, P. & Delarue, M. (1996). Mean-field minimization methods for biological macromolecules. *Curr Opin Struct Biol* 6, 222-6.

34. Holland, J. H. (1992). *Adaptation in natural and artificial systems*, The MIT Press, Cambridge, Massachusetts.

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

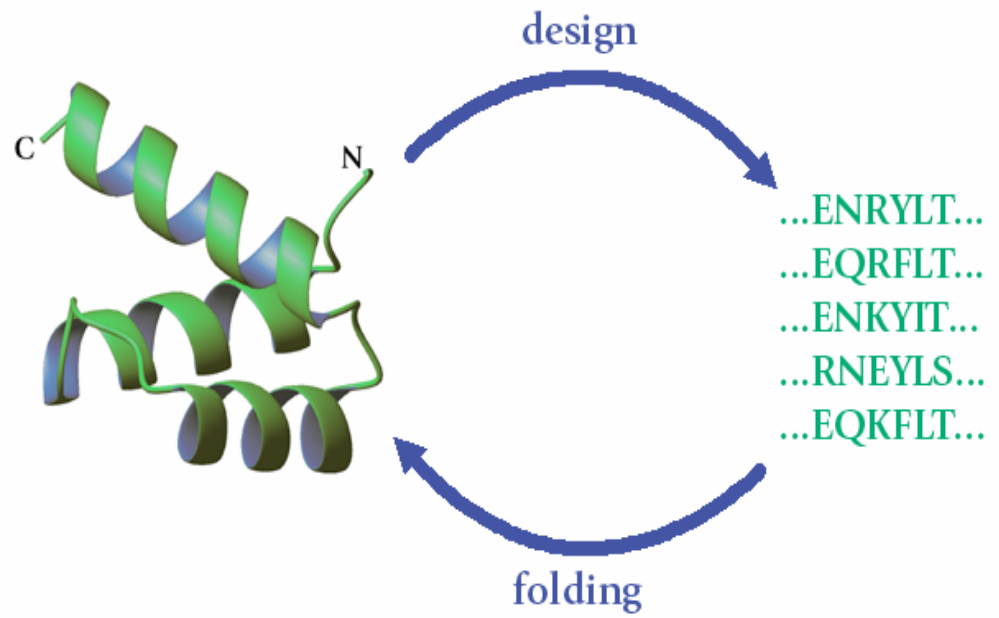


Figure 1-1: Protein design and the protein folding problem.
Protein design is the inverse of the protein folding problem. Only one structure is associated with one amino acid sequence, while many sequences are compatible with a given structure. Despite this advantage, protein design is a tremendous biophysical and computational challenge.

Figure from Shannon Marshall (California Institute of Technology)

A. B.

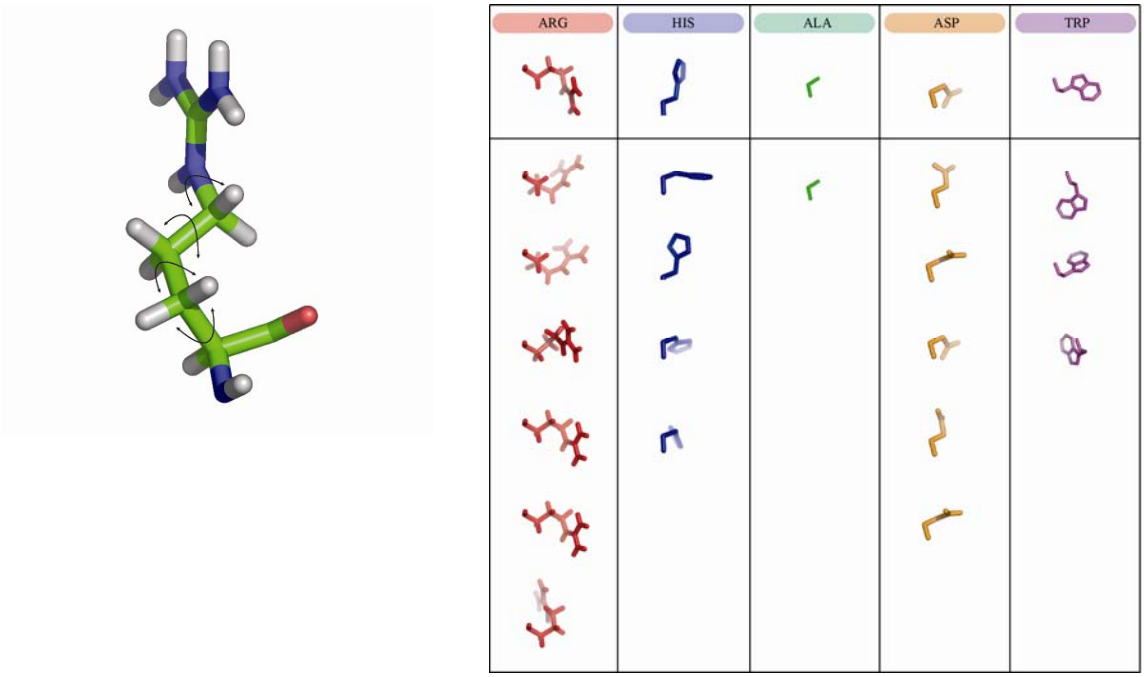


Figure 1-2: Different conformations of amino acid sidechains are represented by rotamers.
Amino acid sidechains can adopt many different conformations; each different conformation is called a rotamer. (A) Arginine sidechains possess several degrees of freedom and can be represented by many rotamers. (B) Different rotamers of some of the amino acids.

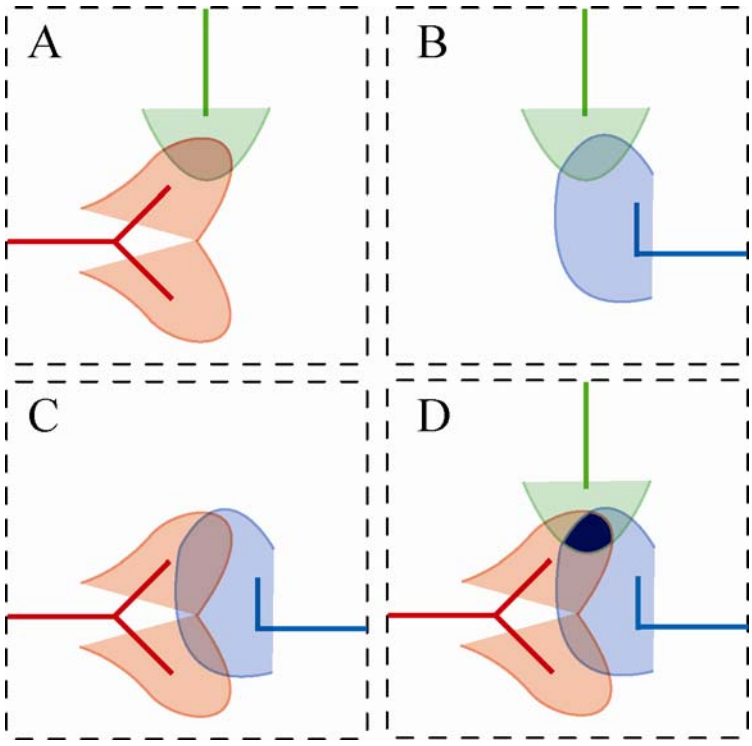


Figure 1-3: Pairwise decomposition of surface areas overestimates burial.
Because the GMEC is not known *a priori*, rotamer surface areas must be approximated. In our previously reported model, we use a pairwise decomposition of surface areas because use of DEE-based algorithms limits the interaction of rotamers to two bodies. Although pairwise surface areas can be calculated well (A-C), when three or more rotamers interact, there will be an overestimation of buried surface area (D, black area). Our new model eliminates the need for pairwise approximations because each rotamer is sampled in the context of a multi-body approximation of the GMEC obtained using an MC search. From this structure, the rotamer's exact surface area can be calculated.

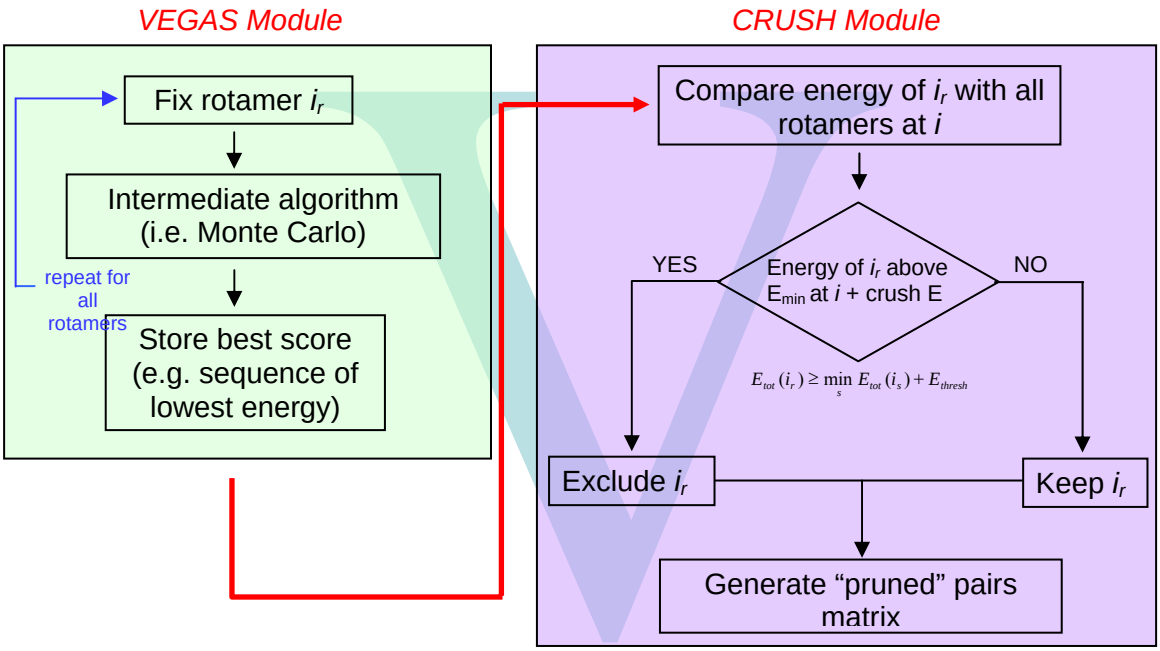


Figure 1-4: Schematic representation of the Vegas algorithm.
The Vegas algorithm prunes rotamer space by judiciously eliminating candidate rotamers by evaluating them in GMEC-like structures. High scoring rotamers, where the energy of the structure containing the candidate rotamer is the score, are eliminated. In this way, rotamers incompatible with the GMEC are eliminated. When a pruned rotamer space is handed to DEE-based algorithms, they proceed faster and more efficiently, and are able to provide solutions to design problems that are better than other optimization algorithms.

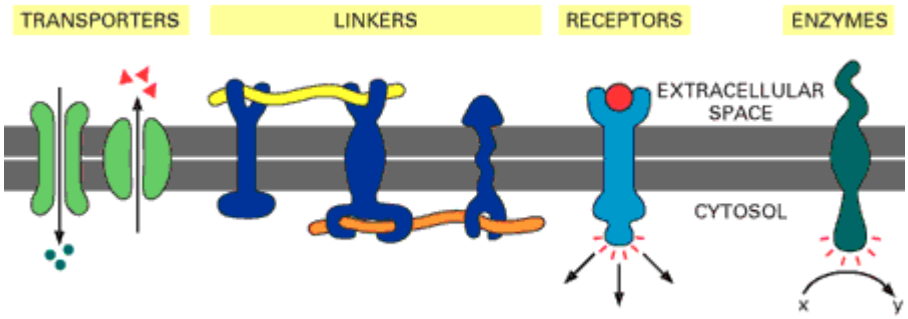


Figure 1-5: The various roles of membrane proteins.
Membrane proteins serve several roles in cells. They can act as transporters of everything from water to proteins, or be vital in maintaining cell structure by linking together various structural elements. Membrane proteins can take on the role of receptors that propagate signals throughout the cell and can exhibit enzymatic activity, catalyzing reactions.

Chapter 2

Direct Calculation of Rotamer Surface Areas in Protein Design

The text of this chapter has been adapted from a manuscript that was co-authored with Professor Stephen L. Mayo.

Premal S. Shah and Stephen L. Mayo, *Submitted* (2005)

Abstract

The incorporation of a surface area-based hydrophobic solvation term in the scoring functions used for protein design has been shown to improve the stabilities of designed variants. Use of a surface area-based solvation term requires the accurate calculation of the surface areas of the amino acid conformers (rotamers) used in side-chain selection. Current methods utilize pairwise approximations that overestimate surface area burial and underestimate surface area exposure. Although scaling factors have been used to compensate for these errors, the surface areas obtained are still inexact. We have developed a method that changes the nature of the approximation used in these calculations from a pairwise method to a full, multi-body method based on Monte Carlo (MC) simulations of the design space. Each rotamer is held fixed on the protein backbone and the MC search algorithm is used to estimate the optimal side-chain sequence and conformation for the rest of the molecule. The fixed rotamer's surface area is calculated directly from the MC-generated structure. With this method, surface area errors were decreased dramatically and the experimental stabilities of proteins generated from computationally predicted sequences were improved. Also, our direct surface area calculation approach significantly reduced the compute time required for sequence optimization using dead-end elimination-based algorithms.

Introduction

The hydrophobic effect is believed to play a dominant role in stabilizing the folded state of proteins.[1-3] The effect implies that hydrophobic (nonpolar) residues in proteins tend to be buried to minimize unfavorable interactions with polar solvent, resulting in energetically more stable molecules. The benefits of the hydrophobic effect have stimulated efforts to incorporate burial of nonpolar surface area as a design criterion in engineering proteins with improved stability. When a surface area-based hydrophobic solvation term was included in the scoring function used for protein design, the sequences produced were more stable;[4] their predicted energies also showed a better correlation with their experimental stabilities.[5] Appropriate use of such a term requires accurate calculation of the surface areas of the amino acid conformers (rotamers) used in side-chain selection. The sequence optimization algorithms commonly used in protein design are based on the dead-end elimination (DEE) theorem.[6] DEE-based algorithms[5,7,8] restrict calculation of sequence energies to include terms involving the interaction of, at most, two bodies. However, the surface area buried by three or more interacting rotamers can overlap, and with pairwise calculations, the overlapping area is counted more than once, resulting in values that are larger than the true buried surface area. Thus, buried surface area is overestimated and exposed surface area is underestimated. To compensate for this error, empirically validated scaling factors have been used, resulting in significant improvement.[5,9] But because of the pairwise restriction, the surface areas obtained are still inexact.

The exact surface area of each of the rotamers could be calculated directly, without using any pairwise approximations, if the location and identity of all the

surrounding atoms in the global minimum energy conformation (GMEC) were known. Unfortunately, the GMEC is not known *a prori*, but it can be accurately and efficiently approximated.

In this report, we use the Monte Carlo (MC) search algorithm[10,11] to approximate the GMEC. The structures generated using MC are then used to directly calculate exact rotamer surface areas. We call this new method type 4 solvation (T4-solvation), while referring to the previous pairwise approximation method as type 2 solvation (T2-solvation). Comparison with true surface areas revealed that T4-solvation calculates surface areas more accurately than T2-solvation. Experimental studies with two proteins showed that sequences predicted using T4-solvation also had stabilities that were comparable to or improved over those predicted using T2-solvation. In addition, T4-solvation significantly reduced the time required to perform the optimization calculations.

Direct Calculation of Rotamer Surface Areas

With T4-solvation, the approximation of surface areas is moved from a pairwise method to a full, multi-body method based on structures that approximate the GMEC. The rotamer is held fixed and MC searches are used to find the optimal sequence and conformation for the rest of the molecule. The rotamer's surface area is calculated in the context of the MC structure and this value is used in our surface area-based hydrophobic solvation term[5] to obtain the rotamer's solvation energy, which in turn is used to compute a new total energy for the rotamer. This is done for each rotamer, then the whole process is repeated using the updated energies. The updated energies allow MC searches to find

more accurate solutions with each iteration. To increase calculation efficiency, a threshold is specified that eliminates high-energy rotamers from subsequent iterations; in this case, the rotamer's surface area prior to elimination is used. We used five iterations for calculations in this report.

Results

T4-solvation surface areas are more accurate

To test the computational accuracy of T4-solvation, we performed side-chain placement calculations[6] on ten proteins spanning a wide range of sizes and compared the surface areas calculated using T4-solvation with the true surface areas calculated after the GMEC was known. Surface areas calculated from the GMEC using T2-solvation's pairwise approximation method were also compared with the true surface areas. Results are presented in Table 2-1. For both core and boundary regions, total buried surface areas calculated using T4-solvation compared well with the true total buried surface areas obtained from the GMEC (maximum of 0.1% error). Errors were also quite low, though slightly worse with T2-solvation.

The greatest difference is observed when comparing errors in exposed hydrophobic surface areas. For the core, the maximum error for T2-solvation was 141%, while T4-solvation produced a maximum error of only 7.7%. The T4 method also performed better than T2-solvation for the boundary.

T4-solvation produces stable variants

To further assess the validity of the T4-solvation method, we performed boundary designs on two proteins using T2- and T4-solvation and evaluated the physical properties of the proteins constructed from the predicted sequences. The two proteins used were the β1 domain of Streptococcal protein G (Gβ1) and the engrailed homeodomain (ENH). We used previously optimized variants of both proteins as starting sequences because they are thermodynamically more stable than the wild type and allow destabilizing variants to be produced without incurring problems with protein expression. We used a core-optimized variant of Gβ1, Gβ1-FII,[12,13] and a surface- and core-optimized variant of ENH, ENH-SC1.[14] The optimal sequences predicted from the design calculations are compared with the starting sequences in Figure 2-1. For Gβ1, of the 14 designed positions, use of T2-solvation predicted a sequence with 11 mutations (Gβ1-T2) and T4-solvation predicted a sequence with 13 (Gβ1-T4). For ENH, 11 positions were designed, and use of T2-solvation predicted 10 changes (ENH-T2), while T4-solvation predicted 8 (ENH-T4).

CD wavelength scans of both the Gβ1 variants are similar to Gβ1-FII (Figure 2-2). However, the minimum of the Gβ1-T4 spectrum is shifted slightly to a shorter wavelength compared to the other two proteins. This blue-shift often results from an increased number of Trp residues or from Trp residues in closer proximity to each other.[15] Gβ1-T4 contains an additional Trp residue at position 27 (Figure 2-1), which, in our predicted structure, is close to Trp-43. Mutation of Trp-27 to Ala or Phe eliminates this blue-shift (data not shown). One-dimensional proton nuclear magnetic resonance (1D ^{1}H NMR) spectra of the three Gβ1 proteins display sharp, narrow linewidths and high

dispersion—properties indicative of well-folded proteins (Figure 2-3). Temperature and chemical denaturation experiments show that Gβ1-T4 is more stable than Gβ1-T2 and as stable as Gβ1-FII (Table 2-2).

CD scans of the ENH variants produced spectra similar to ENH-SC1 and characteristic of α-helical proteins (Figure 2-4). 1D ^{1}H NMR spectra exhibit sharp, narrow linewidths and high dispersion (Figure 2-3). Temperature and chemical denaturation experiments indicate that both variants are significantly stabilized compared to ENH-SC1 (Table 2-2).

Computational optimization speed increases using T4-solvation

Use of T4-solvation improved the efficiency of the sequence optimization step in our protein design calculations. The Gβ1-T4 calculation converged to the GMEC almost five times faster than the Gβ1-T2 calculation (1.9 versus 8.8 processor hours). No differences in optimization times were seen with ENH designs. This was expected because we split the boundary into three regions and ran separate calculations on each; each of these calculations was very small and therefore converged extremely quickly. Separate calculations were run to remain consistent with previous ENH boundary designs.

Discussion

This study showed that the direct calculation of surface areas using T4-solvation is more accurate than previous methods that rely on pairwise approximations. These results suggest that an efficient search algorithm can be used to model the GMEC with sufficient accuracy to allow for the direct calculation of exact surface areas. We chose MC as our optimization algorithm because it has been used successfully in protein design[16-18] and has been shown to more accurately predict the GMEC than other commonly used approximate algorithms.[19] However, other optimization algorithms could be used.

The proteins predicted using T4-solvation had stabilities that were comparable to or improved over those predicted using T2-solvation. The improved stability of Gβ1-T4 may be due to the fact that it has fewer nonpolar residues than Gβ1-T2 (six versus eight). Expression levels were better for Gβ1-T4 than for Gβ1-T2. Also, Gβ1-T2 expresses poorly into the soluble fraction, suggesting aggregation. One explanation for these differences may be that because T2-solvation does not calculate surface areas accurately, exposed nonpolar surface area is under-penalized, which in turn leads to the prediction of sequences with more nonpolar residues. Gβ1-T2 buries more nonpolar surface area than Gβ1-T4 (2032.9 $Å^2$ versus 1758.9 $Å^2$) yet its higher exposure of nonpolar surface area (469.0 $Å^2$ compared to 395.3 $Å^2$) most likely is the dominating force that leads to lower stability. This is an effect observed previously in our laboratory; increasing the nonpolar content of boundary designs by even one position can lead to dramatic differences in protein stability and specificity.[20]

We noted that use of T4-solvation increased the computational speed of sequence optimization with our DEE-based algorithms. However, the total calculation time for the Gβ1-T2 calculation was only 43.3 processor hours compared to 144.96 processor hours for the Gβ1-T4 calculation; the calculation of rotamer surface areas is longer for T4-solvation. However, as mentioned above, we used five iterations for our T4-solvation calculation but found that reducing the number of iterations to three greatly reduced the compute time without compromising accuracy. In addition, varying the threshold can safely eliminate a larger subset of rotamers after each iteration, allowing subsequent iterations to proceed more rapidly. Given the increase in accuracy of surface areas and stability of proteins generated with T4-solvation, the slight increase in compute time is acceptable.

Materials and Methods

Computational methods

A description of force field potential functions and T2-solvation can be found in previous work.[5,9,21-23] An expanded version of the backbone dependent rotamer library described by Dunbrack and Karplus[24] was used. Surface areas were calculated using the Connolly algorithm[25] as described previously.[9] An automated algorithm was used that classified residue positions as core, boundary, or surface.[5] Side-chain placement calculations and sequence optimization for Gβ1 designs was done using HERO,[8] an extension of DEE. Sequence optimization for ENH designs was done as described previously.[20] Calculations were performed on either an SGI Origin 2000 supercomputer

with R10000 processors running at 195 MHz or on an IBM SP3 machine with Power3 processors running at 375 MHz.

For studies done with T4-solvation, our implementation of the MC algorithm can be found in previous work.[19] We used five annealing cycles of 10^6 steps per cycle. Five cycles were chosen to keep computation time to a minimum and 10^6 steps per cycle was selected to be consistent with previous studies in our laboratory. Increasing the number of cycles or steps per cycle did not result in greater surface area accuracy. Low and high annealing temperatures were 150 K and 4000 K, respectively. The threshold value for T4-solvation calculations was set conservatively at 100 kcal/mol.

For side-chain placement calculations, residue identities were kept fixed but their conformations were allowed to change. When one region was analyzed (i.e., core or boundary), the residues in the other two regions were kept fixed in identity and conformation.

Boundary positions for both Gβ1 and ENH designs are indicated in Figure 2-1. For Gβ1 designs, we allowed all 20 natural amino acids except Gly, Pro, Cys, and Met. For ENH designs, we used a fixed binary pattern of the B6 molecule in the Marshall and Mayo study[20] to partition the boundary region into core and surface. We allowed Ala, Val, Leu, Ile, Phe, Tyr, and Trp for positions classified as core; for positions classified as surface, we allowed Ala, Asp, Asn, Glu, Gln, His, Lys, Ser, Thr, and Arg.

Protein expression

The genes encoding the protein variants predicted by the calculations were constructed using recursive PCR[26] and cloned into a variant of the pET-11a (Novagen)

vector. DNA sequencing was used to confirm all sequences. Proteins were expressed in Bl21(DE3) *Escherichia coli* cells and isolated using a freeze-thaw method.[27] Purification was done with reverse-phase HPLC using a C8 prep column (Zorbax) with a linear acetonitrile-water gradient with 0.1% TFA. Protein masses were confirmed using electrospray or MALDI-TOF mass spectrometry.

Circular dichroism spectroscopy

CD analysis was performed on an Aviv 62A DS spectrophotometer. Gβ1 experiments were performed in 50 mM sodium phosphate buffer at pH 5.5. For wavelength scans and thermal denaturation experiments, 50 μM protein was used in a one mm pathlength cell. Thermal denaturations were performed from 1 ºC to 99 ºC with a step size of 1 ºC, equilibration time of 120 seconds, and a data averaging time of 30 seconds. Melting temperatures for Gβ1 variants and ENH-SC1 were determined by fitting a two-state transition.[28] Guandinium denaturation of Gβ1 proteins was done using an auto-titrator with 5 μM protein in a 10 mm pathlength cell. To maintain consistency during the experiment, stock solutions of guanidinium also contained 5 μM protein at the appropriate pH. A step size of 0.2 M, a mixing time of 10 minutes, and a data averaging time of 100 seconds were used. All ENH experiments were carried out at pH 4.5; all other parameters were the same as for the Gβ1 experiments. Guanidinium denaturation data were fit assuming a two-state transition and using the linear extrapolation model.[29] For thermal and chemical denaturations, a wavelength of 218 nm and 222 nm was used for Gβ1 and ENH proteins, respectively.

Nuclear magnetic resonance spectroscopy

A Varian 600 MHz spectrometer using a Varian triple resonance probe was used to obtain 1D ^{1}H NMR spectra. Samples were approximately 0.5 mM protein, 50 mM sodium phosphate in 10% $^{2}H_2O$. The pH was 5.5 and 4.5 for Gβ1 and ENH proteins, respectively.

References

1. Makhatadze, G. I. & Privalov, P. L. (1994). Hydration effects in protein unfolding. *Biophys Chem* 51, 291-304.
2. Pace, C. N., Shirley, B. A., McNutt, M. & Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *Faseb J* 10, 75-83.
3. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry* 29, 7133-55.
4. Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol* 5, 470-5.
5. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci* 5, 895-903.
6. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356, 539-542.
7. Looger, L. L. & Hellinga, H. W. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J Mol Biol* 307, 429-45.
8. Gordon, D. B., Hom, G. K., Mayo, S. L. & Pierce, N. A. (2003). Exact rotamer optimization for protein design. *J Comput Chem* 24, 232-243.
9. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Fold Des* 3, 253-8.
10. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J Chem Phys* 21, 1087-1092.
11. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671-680.
12. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* 94, 10172-7.
13. Su, A. & Mayo, S. L. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci* 6, 1701-7.
14. Morgan, C. S. (2000). Full sequence design of an alpha-helical protein and investigation of the importance of helix dipole and capping effects in helical protein design, California Institute of Technology.
15. Fasman, G. D., Ed. (1996). Circular dichroism and the conformational analysis of biomolecules. New York: Plenum Press.
16. Holm, L. & Sander, C. (1992). Fast and simple Monte Carlo algorithm for side chain optimization in proteins: application to model building by homology. *Proteins* 14, 213-23.

17. Hellinga, H. W. & Richards, F. M. (1994). Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci U S A* 91, 5803-7.

18. Godzik, A. (1995). In search of the ideal protein sequence. *Protein Eng* 8, 409-16.

19. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* 299, 789-803.

20. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* 305, 619-31.

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

22. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Structure Fold Des* 7, R105-9.

23. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr Opin Struct Biol* 9, 509-13.

24. Dunbrack, R. L., Jr. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* 230, 543-74.

25. Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221, 709-13.

26. Prodromou, C. & Pearl, L. H. (1992). Recursive PCR: a novel technique for total gene synthesis. *Protein Eng* 5, 827-9.

27. Johnson, B. H. & Hecht, M. H. (1994). Recombinant proteins can be isolated from E. coli cells by repeated cycles of freezing and thawing. *Biotechnology (N Y)* 12, 1357-60.

28. Minor, D. L., Jr. & Kim, P. S. (1994). Measurement of the beta-sheet-forming propensities of amino acids. *Nature* 367, 660-3.

29. Santoro, M. M. & Bolen, D. W. (1988). Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. *Biochemistry* 27, 8063-8.

Table 2-1: Error in calculation of exposed nonpolar surface area and total buried surface area for ten proteins of various sizes using the T2- and T4-solvation methods.

				Core						**Boundary**					
	Number of Residues			Surface Area		Surface Area Error				Surface Area		Surface Area Error			
Protein (PDB Code)	Total	Core	Boundary	Total Buried (Å^2)	Nonpolar Exposed (Å^2)	Total Buried (%)		Nonpolar Exposed (%)		Total Buried (Å^2)	Nonpolar Exposed (Å^2)	Total Buried (%)		Nonpolar Exposed (%)	
						T4	*T2*	*T4*	*T2*			*T4*	*T2*	*T4*	*T2*
1enh	54	9	21	2,344	34	0.0	1.3	0.0	78	3,840	474	0.0	0.9	0.0	4.2
1pga	56	10	15	2,232	12	0.0	0.4	0.0	67	2,680	565	0.0	1.9	0.0	7.1
1ubi	76	18	21	4,383	59	0.0	1.8	0.0	141	3,468	594	0.0	1.3	0.0	4.0
1mol	94	18	21	3,826	11	0.0	0.2	0.0	87	4,232	571	0.0	0.7	0.0	2.8
1kpt	105	26	31	5,246	83	0.0	0.2	0.0	13	4,936	553	0.0	0.6	0.0	2.7
4azu-A	128	41	28	9,028	156	0.1	1.1	1.0	51	4,801	867	0.0	0.1	0.0	0.6
1gpr	158	41	49	9,092	82	0.0	1.3	0.0	119	8,727	1,130	0.0	0.4	0.0	0.3
1gcs	174	53	36	11,658	160	0.1	0.3	7.7	32	6,803	677	0.0	1.2	0.0	4.5
1edt	265	96	73	20,820	209	0.0	0.4	1.3	41	13,535	1,196	0.0	0.0	0.3	3.7
1pbn	289	98	81	21,012	288	0.0	2.0	0.1	105	16,051	1,763	0.1	0.2	0.3	3.8

Table 2-2: Thermodynamic stability for starting sequence protein and designed variants of Gβ1 and ENH.

Protein	ΔG_{unfold}[a] (kcal mol^{-1})	m[b] (kcal mol^{-1} M^{-1})	T_m[c] (ºC)
Gβ1-FII	6.9	1.7	89.0
Gβ1-T2	3.5	1.8	64.7
Gβ1-T4	6.9	1.9	86.0
ENH-SC1	3.7	1.5	83.0
ENH-T2	5.0	1.0	> 99[d]
ENH-T4	5.0	1.1	> 99[d]

[a] Fee energy of unfolding at 25 ºC.
[b] Slope of ΔG_{unfold} *versus* denaturant concentration; measure of cooperativity.
[c] Melting temperature.
[d] Proteins were folded as monitored by CD at 99 ºC.

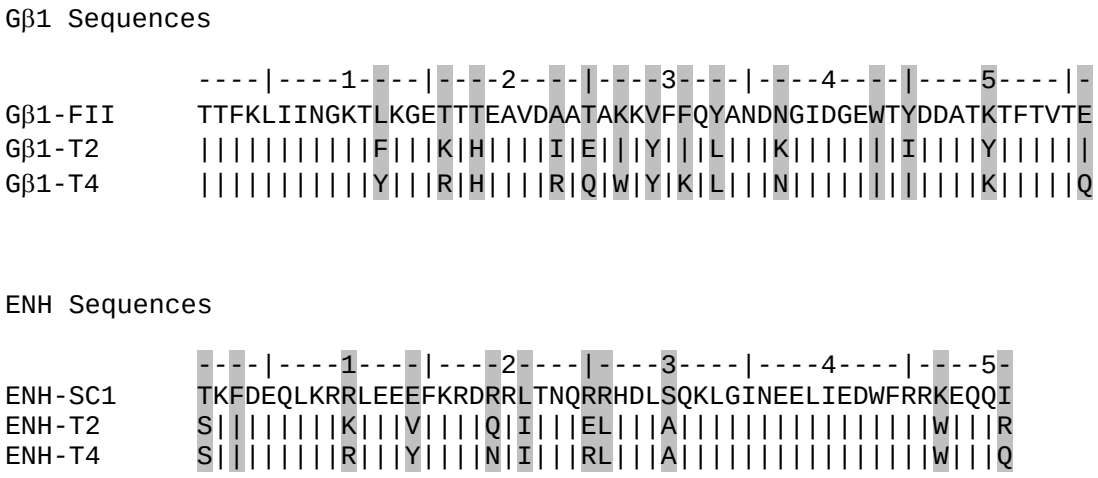


Figure 2-1: Comparison of designed sequences with wild type.
Optimal sequences predicted for the Gβ1 and ENH designs using T2- and T4-solvation. A bar indicates the same amino acid as the starting sequence (Gβ1-FII and ENH-SC1 for Gβ1 and ENH designs, respectively). Boundary positions are indicated in grey.

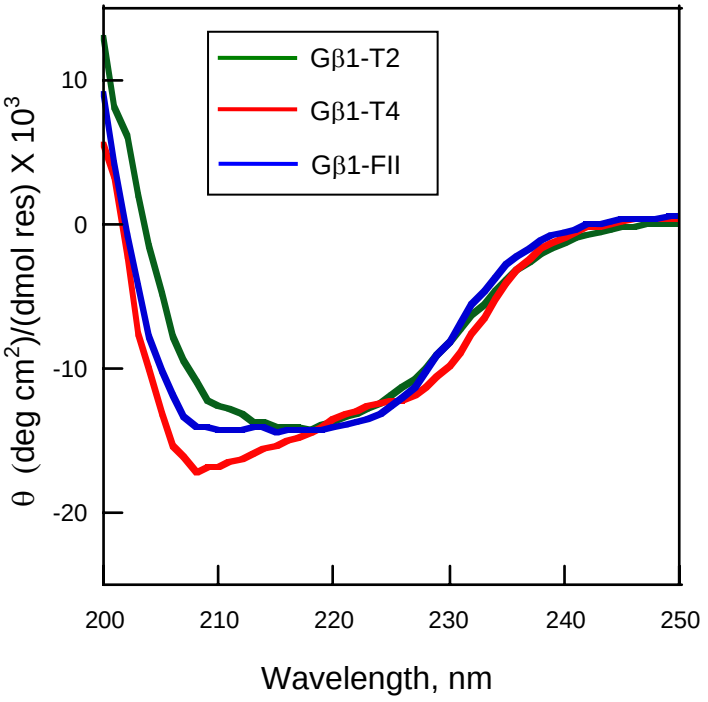
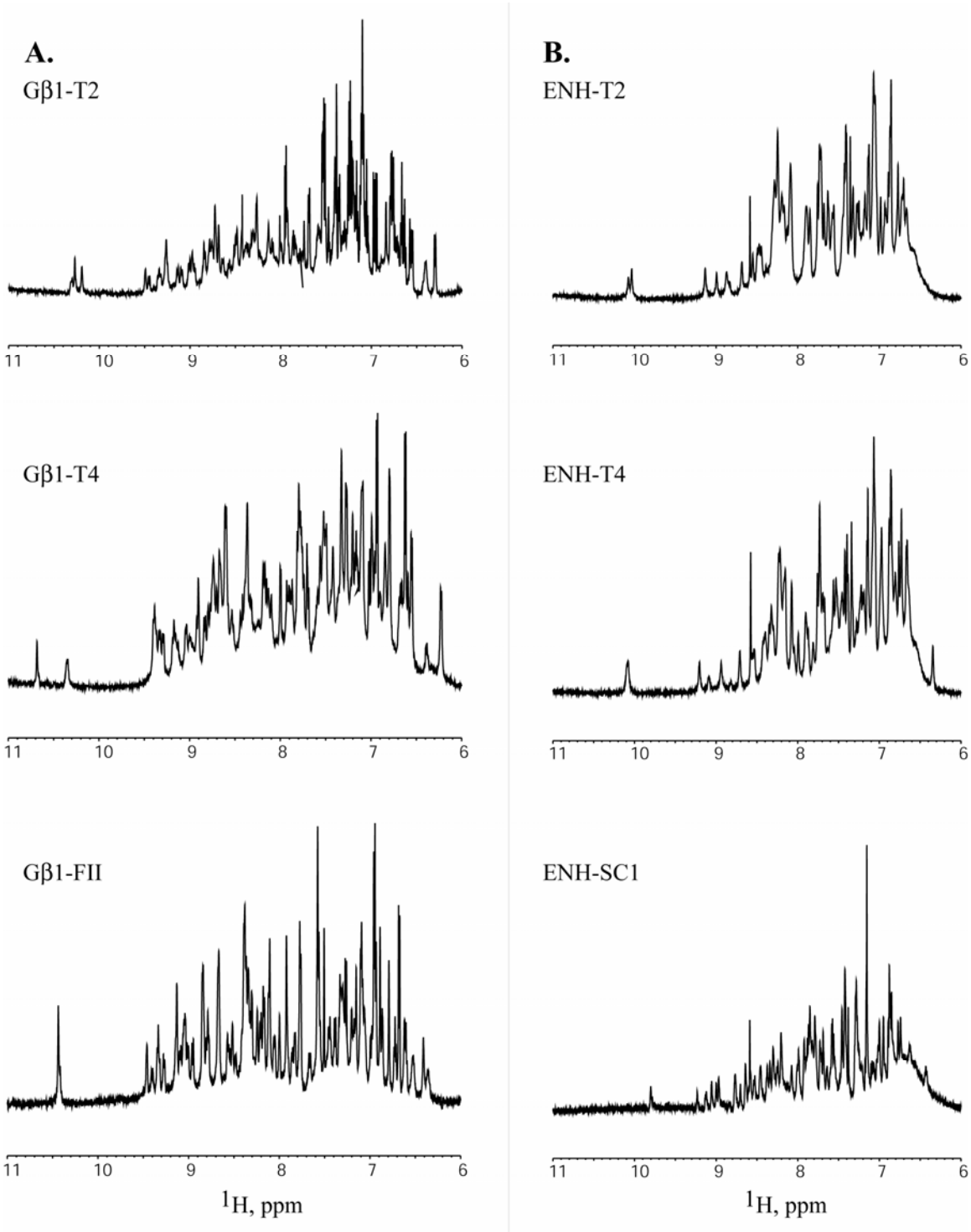


Figure 2-2: Far UV wavelength spectra of Gβ1 variants.
CD wavelength scans of starting sequence protein (Gβ1-FII) and designed variants (Gβ1-T2, and Gβ1-T4) measured at 25 ºC.

Figure on Following Page

Figure 2-3: NMR spectra of variants and wild type.
1D ^{1}H NMR spectra of (A) Gβ1 starting sequence protein (Gβ1-FII) and variants (Gβ1-T2, and Gβ1-T4) and (B) ENH starting sequence protein (ENH-SC1) and variants (ENH-T2 and ENH-T4). For clarity, only the aromatic and amide regions are shown. The sharp, narrow, well-dispersed lines are characteristic of well-folded proteins.


A.
Gβ1-T2
Gβ1-T4
Gβ1-FII
B.
ENH-T2
ENH-T4
ENH-SC1
11
10
9
8
7
6
^{1}H, ppm

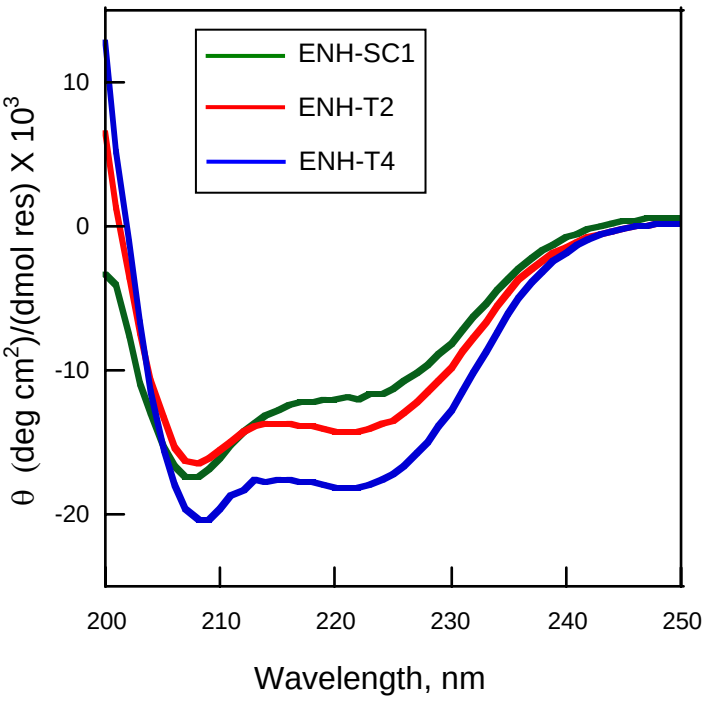


Figure 2-4: Far UV wavelength spectra of ENH variants.
CD wavelength scans of starting sequence protein (ENH-SC1) and variants (ENH-T2 and ENH-T4) measured at 25 ºC. Both designed variants are similar to ENH-SC1 and are typical of α-helical proteins.

Chapter 3

Preprocessing of Rotamers for Protein Design Calculations

The text of this chapter has been adapted from a published manuscript that was co-authored with Professor Stephen L. Mayo and Geoffrey K. Hom.

Premal S. Shah, Geoffrey K. Hom, and Stephen L. Mayo, *J. Comp. Chem.*, 25, 1797-1800 (2004).

Abstract

We have developed a process that significantly reduces the number of rotamers in computational protein design calculations. This process, which we call Vegas, results in dramatic computational performance increases when used with algorithms based on the dead-end elimination (DEE) theorem. Vegas estimates the energy of each rotamer at each position by fixing each rotamer in turn and utilizing various search algorithms to optimize the remaining positions. Algorithms used for this context specific optimization can include Monte Carlo, self-consistent mean field, and the evaluation of an expression that generates a lower bound energy for the fixed rotamer. Rotamers with energies above a user-defined cutoff value are eliminated. We found that using Vegas to preprocess rotamers significantly reduced the calculation time of subsequent DEE-based algorithms while retaining the global minimum energy conformation. For a full boundary design of a 51 amino acid fragment of engrailed homeodomain, the total calculation time was reduced by 12-fold.

Introduction

An important goal of computational protein design is to identify the amino acid sequence and side-chain orientations that correspond to the global minimum energy conformation (GMEC). However, searching for the GMEC is challenging due to the enormity of sequence space; even a small protein of 100 amino acids has 20^{100} ($\sim 10^{130}$) possible sequences. Accounting for side-chain flexibility by including different side-chain conformations called rotamers[1-3] further increases the combinatorial complexity. Consequently, exhaustive searches for the GMEC are almost always intractable.

Algorithms based on the dead-end elimination (DEE) theorem[4] have been developed to address combinatorial optimization problems in side-chain placement and protein design. If DEE-based algorithms converge, the solution is guaranteed to be the GMEC. As a result, not only are these algorithms useful when performing force field improvements or parameter optimization,[5,6] their use has proven to be successful for many challenging design problems.[7-11] While recent enhancements to DEE have allowed difficult designs to be performed,[12-15] more ambitious design problems can cause even the most effective DEE-based algorithms to stall. In addition, some calculations take an impractical amount of time to converge to the GMEC. In such cases, other algorithms may be employed. These include Monte Carlo (MC) methods,[16,17] genetic algorithms,[18,19] self-consistent mean field (SCMF) techniques,[20,21] and Branch-and-Bound methods.[22] Although these approaches can provide solutions when DEE-based algorithms stall, they typically have the drawback of not being able to guarantee that their solutions are the GMEC even when starting from a DEE-reduced rotamer space. As a result, there is still

ample motivation to develop techniques to improve or assist current DEE-based algorithms.

One approach is to reduce the number of rotamers in a calculation by eliminating a subset of rotamers prior to use of DEE-based algorithms. An example of this strategy can be found in the high-energy threshold reduction method.[23] In most cases, by eliminating rotamers possessing energies above a user-defined threshold, De Maeyer *et al.* were able to eliminate over one-third of rotamers without sacrificing the GMEC in side-chain placement calculations. Remaining rotamers were then evaluated with DEE. Here, we present a similar approach for protein design calculations; we prune rotamer space by judiciously eliminating rotamers, thus allowing DEE-based algorithms to proceed more efficiently. Our method, which we call Vegas, scores each rotamer at each position by fixing it in turn and using MC or SCMF to optimize the rest of the positions. The rotamer's score is the energy of the resulting solution. In addition, a rotamer's score can be calculated by evaluating an expression that generates a lower bound energy.[22] Rotamers remaining after the elimination step are passed on to a DEE-based algorithm. We can safely eliminate a large subset of rotamers without compromising the GMEC and we observe a significant reduction in total compute time.

Vegas

Vegas reduces the number of rotamers in protein design calculations by applying a rejection criterion after obtaining a score for each rotamer at each position. This is done by fixing the rotamer to be scored and using various optimization algorithms to generate a rotamer sequence for the rest of the molecule. The rotamer's score is the energy of the

resulting solution. In this report, two optimization algorithms were used: one based on Monte Carlo (MC) methods,[24] and another based on self-consistent mean field theory (SCMF).[24] In addition, a rotamer's score was also obtained by evaluating an expression that provided a lower bound energy[22] for the fixed rotamer. Rotamers with scores above the best score for that position plus a user-defined threshold value are eliminated. Remaining rotamers are then optimized with HERO,[15] an extension of DEE.

Results

We used two test cases to assess the effectiveness of Vegas. We started with the designs of different regions of a very small protein and increased the computational complexity with the second test case. Vegas' effectiveness was evaluated by its ability to retain the GMEC and increase computational efficiency. To check Vegas' performance in not eliminating GMEC rotamers, the GMEC was first obtained without Vegas in a reference calculation using HERO alone. The different versions of Vegas are referred to with an underscore between Vegas and the method used to obtain the rotamer score. For example, use of MC with Vegas is referred to as Vegas_MC.

Test case 1

We performed designs of the core, boundary, and surface regions of the β1 domain of protein G (Gβ1).[25] These small, relatively simple designs were done to demonstrate the ability of Vegas to safely apply a rejection criterion to eliminate rotamers without sacrificing the GMEC. Table 3-1 lists the number of rotamers eliminated as the threshold value is increased. All versions of Vegas performed equally well for core and boundary

designs; the most aggressive threshold value (5 kcal/mol) allowed about 90% of rotamers to be eliminated without losing the GMEC. Elimination was more difficult with surface residues. Compared to Vegas_MC, Vegas_SCMF and Vegas_Bound allowed for more aggressive threshold values to be applied without losing the GMEC.

Test case 2

A boundary design of a 51 amino acid fragment of the engrailed homeodomain (ENH)[26] was performed to determine Vegas' ability to increase computational efficiency without compromising accuracy (Figures 3-1 and 3-2). Vegas_MC and Vegas_SCMF retained the GMEC when threshold values of 10 kcal/mol and larger were used. At 10 kcal/mol, 72% and 64% of the 3571 total rotamers in the calculation were eliminated with Vegas_MC and Vegas_SCMF, respectively. Interestingly, a threshold of 5 kcal/mol for Vegas_MC produced the same amino acid sequence as the one in the GMEC; however, the conformations of some of the amino acids were different. We could not be as aggressive with Vegas_Bound; a minimum of 20 kcal/mol was required to obtain the GMEC. At this threshold, 41% of the rotamers were eliminated.

Although Vegas_MC and Vegas_SCMF allowed the use of more aggressive threshold values while retaining the GMEC, comparison of total calculation times shows Vegas_Bound to be more efficient (Figure 3-2). At a relatively conservative threshold value of 40 kcal/mol, Vegas_Bound obtained the GMEC almost four times faster than the reference calculation. At 20 kcal/mol, it produced the GMEC in only 8 processor hours—a 12-fold improvement over the reference calculation. In comparison, Vegas_MC was only

able to achieve a two-fold overall speed enhancement. Vegas_SCMF, on the other hand, actually caused the calculation to run two times slower than the reference calculation.

Discussion

Vegas is an efficient protein design tool that can reduce computational complexity without sacrificing the ability to obtain ground-state solutions. Its computational efficiency becomes more pronounced with increasing problem size. Vegas produced a 12-fold reduction in the time required to solve the boundary design of ENH, decreasing the total processing time from 92 to 8 hours. This increase in computational speed resulted from elimination of about 41% of the rotamers, without losing rotamers in the GMEC. The high efficiency of Vegas_Bound for this design compared to Vegas_MC and Vegas_SCMF (Figure 3-2) can be attributed to a dramatic difference in time for scoring the rotamers. The rotamer scoring times for Vegas_MC and Vegas_SCMF were 45 and 198 processor hours, respectively, while Vegas_Bound scored rotamers in less than one minute on a single processor.

The accuracy and increased efficiency provided by Vegas can extend the capabilities of protein design. For example, Vegas allows the use of larger rotamer libraries, which may provide lower energy solutions to design problems. Larger rotamer libraries have been shown to improve accuracy in side-chain placement calculations.[23] The use of Vegas can also allow more difficult designs to be performed and can facilitate the design of many features including functionally important properties.

A recent side-chain placement algorithm called FASTER[27] has shown promise when adapted to protein design (data not shown). Elements of FASTER could be

implemented as an additional rotamer-scoring method within Vegas. Vegas_FASTER, as well as Vegas with other optimization algorithms, is a viable option in the future. We used Vegas here as a preprocessor to HERO; however, Vegas is a general preprocessing method and can be combined with any relevant optimization algorithm.

Methods

Computational methods

A description of force field potential functions and their parameters can be found in previous work.[5,7,28-30] We used an expanded version of the backbone-dependent rotamer library described by Dunbrack and Karplus.[3] An automated algorithm was employed that classified residue positions as core, boundary, or surface.[5] For core positions, we allowed the selection of the amino acids A, V, L, I, F, Y, and W. For surface positions, we allowed A, S, T, D, N, H, E, Q, K, and R, and for boundary positions, we allowed all amino acids except G, P, C, and M. HERO and the bounding expression were implemented as described by Gordon et al.,[15] and MC and SCMF were implemented as described previously.[24] For MC, we used 5 annealing cycles of 10^6 steps per cycle. Low and high annealing temperatures were 150 K and 4000 K, respectively. For SCMF, we used initial and final temperatures of 20,000 K and 300 K, respectively, with the temperature lowered in 100 K increments. A convergence criterion of 0.001 and a pair-energy threshold of 100 kcal/mol were used.

Test case designs

In test case 1, we designed the core, boundary, and surface regions of Gβ1 (PDB code 1pga).[25] Core positions were 3, 5, 7, 9, 20, 26, 30, 34, 39, 41, 52, and 54. Boundary positions were 1, 12, 16, 18, 23, 25, 27, 29, 31, 33, 37, 43, 45, 50, and 56. Surface positions were 2, 4, 6, 8, 10, 11, 13, 14, 15, 17, 19, 21, 22, 24, 28, 32, 35, 36, 38, 40, 42, 44, 46, 47, 48, 49, 51, 53, and 55. Design of a region involved allowing all allowable amino acids for that region, while keeping the other two regions fixed in both identity and

conformation. Test case 2 was the boundary design of ENH (PDB code 1enh;[26] positions 1, 3, 10, 14, 19, 21, 25, 30, 47, and 51). Core and surface positions were kept fixed in identity but their conformations were allowed to change. All calculations were performed on an IBM SP3 running 375-MHz Power3 processors.

References

1. Janin, J. & Wodak, S. (1978). Conformation of amino acid side-chains in proteins. *J Mol Biol* 125, 357-86.
2. Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* 193, 775-91.
3. Dunbrack, R. L., Jr. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* 230, 543-74.
4. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356, 539-542.
5. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci* 5, 895-903.
6. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* 94, 10172-7.
7. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science* 278, 82-7.
8. Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol* 5, 470-5.
9. Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proc Natl Acad Sci U S A* 98, 14274-14279.
10. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* 305, 619-31.
11. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature* 423, 185-90.
12. Gordon, D. B. & Mayo, S. L. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J Comput Chem* 19, 1505-1514.
13. Pierce, N. A., Spriet, J. A. & Mayo, S. L. (2000). Conformational splitting: A more powerful criterion for dead-end elimination. *J Comput Chem* 21, 999-1009.
14. Looger, L. L. & Hellinga, H. W. (2001). Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J Mol Biol* 307, 429-45.
15. Gordon, D. B., Hom, G. K., Mayo, S. L. & Pierce, N. A. (2003). Exact rotamer optimization for protein design. *J Comput Chem* 24, 232-243.

16. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J Chem Phys* 21, 1087-1092.

17. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671-680.

18. Holland, J. H. (1992). *Adaptation in natural and artificial systems*, The MIT Press, Cambridge, Massachusetts.

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

20. Koehl, P. & Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol* 239, 249-75.

21. Koehl, P. & Delarue, M. (1996). Mean-field minimization methods for biological macromolecules. *Curr Opin Struct Biol* 6, 222-6.

22. Gordon, D. B. & Mayo, S. L. (1999). Branch-and-Terminate: a combinatorial optimization algorithm for protein design. *Structure Fold Des* 7, 1089-98.

23. De Maeyer, M., Desmet, J. & Lasters, I. (1997). All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold Des* 2, 53-66.

24. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* 299, 789-803.

25. Gallagher, T., Alexander, P., Bryan, P. & Gilliland, G. L. (1994). Two crystal structures of the β1 immunoglobulin-binding domain of Streptococcal protein G and comparison with NMR. *Biochemistry* 33, 4721-9.

26. Clarke, N. D., Kissinger, C. R., Desjarlais, J., Gilliland, G. L. & Pabo, C. O. (1994). Structural studies of the engrailed homeodomain. *Protein Sci* 3, 1779-87.

27. Desmet, J., Spriet, J. & Lasters, I. (2002). Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* 48, 31-43.

28. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr Opin Struct Biol* 9, 509-13.

29. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Structure Fold Des* 7, R105-9.

30. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Fold Des* 3, 253-8.

Table 3-1: Number of rotamers eliminated with varying threshold values for core, boundary, and surface designs of the β1 domain of protein G: Comparison using Vegas_MC, Vegas_SCMF, and Vegas_Bound.

	Core (413)[a]			Boundary (2663) [a]			Surface (4971) [a]		
Threshold (kcal/mol)	Vegas_ MC	Vegas_ SCMF	Vegas_ Bound	Vegas_ MC	Vegas_ SCMF	Vegas_ Bound	Vegas_ MC	Vegas_ SCMF	Vegas_ Bound
5	373	373	362	2254	2319	2357	4540[b]	3795[b]	3355[b]
10	332	337	323	1371	1495	1516	2995[b]	1901	1536
20	262	269	225	336	360	371	700	536	496
40	183	186	173	130	129	128	278	272	269
80	141	143	137	96	96	96	225	222	219
160	120	20	117	84	0	87	165	10	163

[a] Initial number of rotamers.
[b] Calculation failed to produce GMEC.

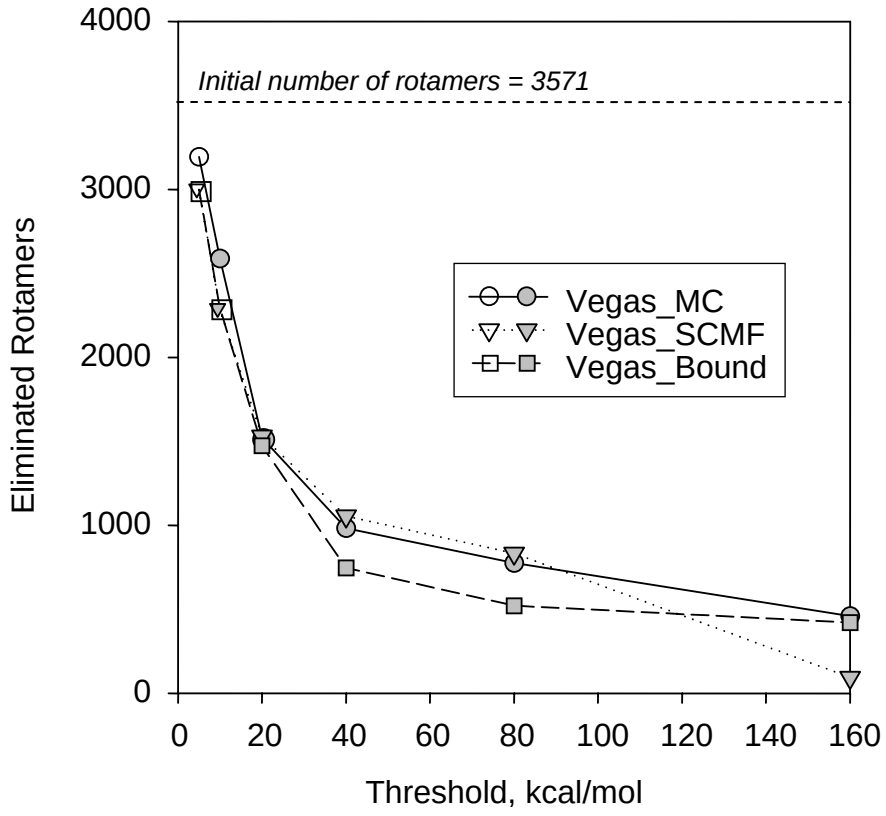


Figure 3-1: Elimination of rotamers with Vegas.
Number of rotamers eliminated with varying threshold values for the boundary design of engrailed homeodomain. The reference calculation (i.e., with HERO alone) contained 3571 rotamers. Threshold values that failed to produce the GMEC are shown with open symbols.

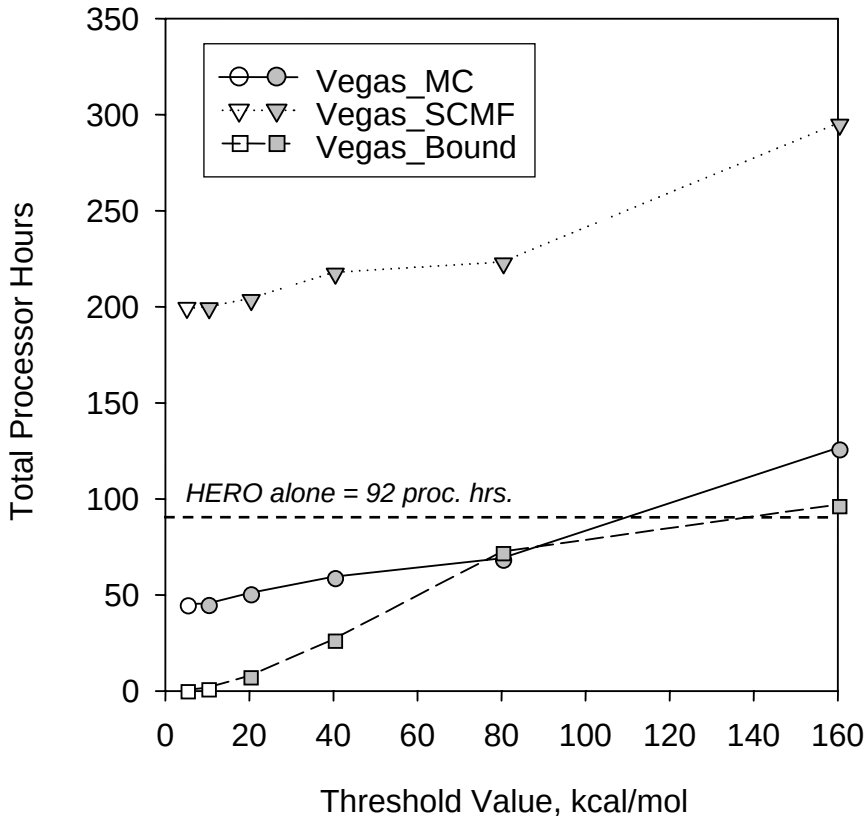


Figure 3-2: Total calculation times with Vegas.
Total calculation times for the boundary design of engrailed homeodomain. The reference calculation (i.e., with HERO alone) took 92 processor hours. Threshold values that failed to produce the GMEC are shown with open symbols.

Chapter 4

Thermodynamic and Structural Characterization of Full Sequence Designs

The text of this chapter has been adapted from a manuscript that was co-authored with Professor Stephen L. Mayo as well as Geoffrey K. Hom and Scott A. Ross.

Premal S. Shah, Geoffrey K. Hom, Scott A. Ross, and Stephen L. Mayo, *Submitted* (2005).

Abstract

Sequence optimization algorithms based on the dead-end elimination (DEE) theorem are preferred in computational protein design because if they converge, their solutions are guaranteed to be the ground-state solutions. However, the increasing size and complexities of designs can cause DEE-based algorithms to stall, failing to deliver a solution. We have used three alternate sequence optimization algorithms in concert with the ORBIT protein design software to simultaneously optimize every position of a 51 amino acid fragment of the *Drosophila* engrailed homeodomain. Two of the sequences obtained from the calculations were studied in detail. The optimized sequences share no statistical similarity to any known sequence and differ from the wild-type sequence by approximately 80%. Based on physical studies of the optimized variants, we conclude that the proteins are nearly identical to each other, displaying hallmarks of well-folded, all α-helical proteins. The thermodynamic stabilities of the designed variants were enhanced by approximately 2 kcal/mol over the wild-type protein at 25 °C. In addition, the designed variants have melting temperatures in excess of 100 °C compared to 43 °C for the wild type protein. We solved the solution structure of one of the designed variants and found that the protein folds accurately into the desired target fold. Knowledge that non-DEE-based sequence optimization algorithms can be used for large, challenging problems leading to variants with markedly improved stability and high specificity for the target fold allows for more ambitious protein design problems to be undertaken.

Introduction

Computational protein design seeks to find amino acid sequences compatible with a target fold. In general, the global minimum energy conformation (GMEC) is desired, since this sequence and conformation confers optimal stability for the fold, provided the physical forces governing protein structure and stability are accurately modeled. Obtaining the GMEC while simultaneously optimizing every position in a protein is a challenging combinatorial problem; for a relatively small 50-residue protein, the GMEC must be identified from 10^{65} possible amino acid sequences. When different conformers of amino acids (rotamers) are included, the complexity grows substantially, requiring the consideration of over 10^{100} rotamer sequences.

Many difficult designs[1-5] have been performed using algorithms based on the dead-end elimination[6] (DEE) theorem. DEE-based algorithms are ideal because if they converge, their solutions are guaranteed to be the GMEC. However, increasingly challenging design problems can prevent even the most effective DEE-based algorithms[7-10] from converging in any practical amount of time. Furthermore, in some cases, these algorithms stall and fail to converge entirely. As an alternative, non-DEE-based algorithms may be employed to obtain sequences compatible with a target fold. However, these algorithms also have their limitations: they do not necessarily provide the GMEC, and their performance has been shown to decay as the size of the design increases.[11]

Our goal was to determine whether the use of non-DEE-based algorithms on large, complex designs can provide solutions that are stable and assume the target fold. We undertook the full sequence design of a 51-amino acid fragment of the *Drosophila*

engrailed homeodomain (ENH). Non-DEE-based algorithms were required because DEE-based algorithms failed to converge. We used three algorithms: Monte Carlo[12,13] (MC), Vegas,[14] and FASTER.[15] MC is a commonly used stochastic search algorithm, Vegas is a rotamer pruning algorithm recently developed in our laboratory that is efficient for large designs, and FASTER is a fast and accurate side-chain placement method, which we adapted for protein design applications. The protein variants predicted with these algorithms were expressed, purified, and characterized thermodynamically. Furthermore, the solution structure of one of the variants was solved in order to assess whether the designed proteins adopt the desired target fold. This work adds to the small number of full sequence designs performed to date for which thermodynamic and structural studies have been perfomed.[16,17]

Results

Computational sequence optimization

We divided ENH[18] into core, boundary, and surface regions with an automated residue classification algorithm[19] and modeled the physical forces within each region with a potential energy function that includes van der Waals, electrostatic, solvation, and hydrogen bonding terms.[19-22] Only nonpolar amino acids were allowed in the core, while on the surface, only polar amino acids were considered. A fixed binary pattern was used that assigned boundary positions to either the core or the surface based on exposed surface area;[3] this fixed binary pattern has been shown to confer added stability to the ENH fold.[3] The amino acid identities of positions involved in helix capping and helix dipoles were further restricted as described previously.[4] To account for the torsional

flexibility of amino acids, a backbone-dependent rotamer library,[23] based on that of Dunbrack and Karplus,[24] was employed. The total initial search space for this calculation was 10^{111} rotamer sequences.

Our laboratory has successfully used DEE-based sequence optimization algorithms[7-10,16] to generate sequences for many design problems.[1,2,16] In this study, we initially attempted optimization with HERO,[10] an extension of DEE that performs more efficiently on large calculations. However, HERO stalled and failed to provide an answer. As a result, three non-DEE-based sequence optimization algorithms, MC, Vegas, and FASTER, were used to predict sequences compatible with the target ENH fold. The best rotamer sequences generated by Vegas and FASTER are identical and have simulation energies of -225.0 kcal/mol. This sequence (FSM1_VF) is a 39-fold mutant of the wild-type sequence (Figure 4-1). The best MC solution (FSM1_MC) has a slightly higher simulation energy (-223.4 kcal/mol) and is a 40-fold mutant of wild-type ENH and an 11-fold mutant of FSM1_VF. A BLAST[25] search indicated that the two optimized variants have no statistically significant similarity to any known sequence.

Physical characterization of ENH variants

Far ultraviolet (UV) circular dichroism (CD) spectroscopy of FSM1_VF and FSM1_MC revealed spectra characteristic of α-helical proteins (Figure 4-2). The spectra for the two variants are almost superimposable and are characteristic of α-helical proteins with minima at 208 and 222 nm. The spectra are also very similar to those for wild-type ENH as well as other well-folded ENH variants produced in our laboratory.[3,4,26] 1D ^{1}H nuclear magnetic resonance (NMR) spectroscopy performed on both proteins produced

spectra displaying the sharp, moderately-dispersed lines expected of a well-folded protein (Figure 4-3).

Thermal denaturations monitored by CD at 222 nm revealed that both proteins do not complete their unfolding transitions by 99 °C, indicating that they are still folded at this temperature (data not shown). In comparison, the wild type has a T_m of 43 °C (Table 4-1).[26] Chemical denaturations using guanidinium hydrochloride were performed to determine unfolding free energies (ΔG_{unfold}). The variants were over 2 kcal/mol more stable than the wild-type protein under similar conditions (Table 4-1).[27] This is a remarkable result considering that approximately 80% of the wild-type sequence was mutated to obtain our designed sequences.

ANS (1-anilino-napthalene-8-sulfonate) binding was used to further validate the structural integrity of the ENH variants. ANS selectively binds molten globule states of proteins.[28] Molten globules exhibit pronounced secondary structure and compactness but lack packed tertiary structure. Hen egg white lysozyme (HEWL) in 25% HFA (hexafluoroacetone hydrate) was used as a positive control; under this condition, HEWL binds ANS and exhibits molten globule characteristics.[29] Although the ENH variants showed some evidence of ANS binding, it was almost eight-fold lower than HEWL (data not shown). This slight ANS binding is most likely due to exposed hydrophobic patches rather than the result of binding to a molten globule state (see below).[28] Overall, the spectral and thermodynamic data indicate that the designed variants are very stable and are physically and structurally similar.

Solution structure of FSM1_VF

The solution structure of FSM1_VF was solved by NMR. Evidently due to the helical structure and relatively low sequence diversity of FSM1_VF (Figure 4-1), its NMR spectra display considerable chemical shift degeneracy. Thus, it was necessary to use both HNCACB/CBCA(CO)NH and HNCO/HN(CA)CO experiment pairs on uniformly ^{15}N, ^{13}C-labeled protein to sequentially assign backbone atom chemical shifts. Other standard double and triple resonance NMR experiments were then sufficient to achieve nearly complete assignment of side-chain atom chemical shifts. Over 1300 loose geometric constraints (interproton distances from NOEs, dihedral angles, and hydrogen bonds) on the structure were derived from NMR data (Table 4-2). The program ARIA[30] was used both to assign many of these constraints and to calculate an ensemble of structures consistent with them (Figure 4-4). The ensemble is of a precision typical for homeodomain NMR structures,[31] with 0.59 Å root mean square (r.m.s.) deviation to the mean for backbone heavy atoms of residues 3-45; the ensemble is also of good stereochemical quality, with 96.6% of residues in most-favored or allowed regions of ϕ,ψ space.

The calculated ensemble shows that FSM1_VF adopts the anticipated ENH fold. Helices 1 and 2 are well-defined, as is the tight turn between helices 2 and 3 and the first two turns of helix 3. The termini are poorly localized, as well as residues 18-20 in the loop between helices 1 and 2. Paucity of data makes the origin of this imprecision uncertain for the loop residues. However, intermediate $^{3}J_{HNHA}$ coupling constant values for residues 1-5, 46, and 48-51 suggest that the termini are disordered. Disorder in the backbone in these portions of the sequence is accompanied by side-chain disorder as

indicated by low χ_1 and χ_2 angular order parameters for nominal core residues W3, F43, F44, and F47.

We compared the FSM1_VF solution structure to the ENH crystal structure. The experimental structure closest to the mean of the ensemble in Figure 4-4 has a backbone r.m.s. deviation of 2.5 Å from the crystal structure for C_α atoms of residues 3-45 (Figure 4-5). The largest differences from the crystal structure were found at the termini and in the orientation of helix 3 with respect to helices 1 and 2. Indeed, solution structures of homoedomains uncomplexed to DNA frequently show disorder in both the N terminus and the C-terminal portion of helix 3.[31] In addition, the starting structure is a truncated version of the crystal structure due to lack of electron density at the C terminus. The crystal structure of ENH is thus quite possibly a nonphysical template for these regions of the molecule in solution. Furthermore, the different orientation of helix 3 could easily be an effect propagated from the disordered C terminus, and the disordered aromatic side chains in the termini could account for the modest ANS binding observed. For the remainder of the structure, FSM1_VF matches the template closely.

Discussion

Use of non-DEE-based algorithms

Non-DEE-based algorithms have been used to produce stable proteins;[17,32-35] however, most of these designs were restricted to the core and were less complex than the design performed here. A quantitative comparison showed that the performance of non-DEE-based algorithms decreases as the complexity of the problem increases.[11]

Performance was defined as the fraction of rotamers predicted incorrectly compared to the GMEC. The goal of the present study was to determine the effectiveness of non DEE-based algorithms on complex problems such as full sequence designs; that is, the ability to yield stable proteins that retain high structural specificity for the target fold. Baker and colleagues recently performed full sequence designs using MC with reasonable success;[17] however, the structures of the proteins have not yet been solved. In this study, we clearly demonstrate that three alternatives to DEE-based algorithms (MC, Vegas, and FASTER) can be used on complex problems to predict sequences with protein stabilities much higher than wild type. In addition, we verified that the designed variants have the same topology as the target fold, as shown by the solution structure of FSM1_VF.

These results suggest that many highly stable proteins can be obtained for complex design problems without identifying the GMEC. In fact, an MC search performed around the FSM1_VF sequence showed that there are at least 900 unique amino acid sequences with simulation energies between FSM1_VF (-225.0 kcal/mol) and our other stable variant, FSM1_MC (-223.4 kcal /mol). It is certainly plausible that all of these sequences would yield proteins that are equally stable and target fold-specific. Taken further, there are likely many sequences with simulation energies higher than that of FSM1_MC that would also adopt the target fold and possess stabilities higher than wild type.

The knowledge that very large, previously intractable designs can be successfully performed with non-DEE-based algorithms allows protein designers to tackle more ambitious problems. Catalytic activity can be designed onto larger scaffolds, improved stabilities can be obtained for larger proteins, and complex protein-protein interactions

can be studied. Larger rotamer libraries can also be used to enhance the accuracy of the solutions generated.

Methods

Computational modeling

Description of potential functions and parameters can be found in our previous work.[19-22,36,37] For ENH, we identified 11 core positions (7, 11, 15, 29, 33, 34, 35, 39, 40, 43, and 44), 11 boundary positions (1, 3, 10, 14, 19, 21, 25, 26, 30, 47, and 51), and 29 surface positions (2, 4, 5, 6, 8, 9, 12, 13, 16, 17, 18, 20, 22, 23, 24, 27, 28, 31, 32, 36, 37, 38, 41, 42, 45, 46, 48, 49, and 50). The fixed binary pattern of the B6 design in the Marshall and Mayo study[3] was applied to boundary residues. Residues 4, 22, and 36 were treated as helix N-capping positions; residues 5, 6, 23, 24, 37, and 38 as helix N-terminal dipole positions, and residues 16, 17, 31, 32, 49, and 50 as helix C-terminal dipole positions. The rules that govern these positions are described in previous work.[4]

Construction of mutants, protein expression, and purification

Genes encoding the ENH variants were made using recursive PCR techniques[38] and cloned into a modified pET11a (Novagen) vector. Recombinant protein was expressed by IPTG induction in BL21(DE3) hosts (Stratagene) and isolated using a freeze/thaw method.[39] Purification was accomplished using a linear acetonitrile/water gradient containing 0.1% TFA. Molecular weights were verified by mass spectrometry. The resultant protein was a 52-mer, with a methionine at the N terminus.

CD analysis

CD data were collected on an Aviv 62DS spectrometer equipped with a thermoelectric unit and an autotitrator. Wavelength scans and thermal denaturation experiments were performed in a 1 mm path length cell with 50 μM protein in 50 mM sodium phosphate at pH 5.5. Thermal melts were monitored at 222 nm. Data were collected every 1 °C with an equilibration time of 2 min and an averaging time of 30 sec. Guanidinium chloride denaturations were done in a 1 cm path length cell with 5 μM protein in 50 mM sodium phosphate at pH 5.5 and 25 °C. To keep the protein concentration constant, a saturated solution of guanidinium chloride was prepared with buffer that also included 5 μM protein. A 10 min mixing time and 100 sec averaging time were used. Data were fit and ΔG_{unfold} values were obtained using the linear extrapolation method.[40]

NMR spectroscopy

NMR experiments were performed at 20 °C on a Varian INOVA 600 spectrometer. Data was processed using nmrPipe[41] and analyzed using NMRview.[42] Backbone chemical shift assignments were obtained from 3D HNCACB, CBCA(CO)NH, HNCO, HN(CA)CO and HNHA spectra. 2D DQF-COSY and 3D C(CO)NH-TOCSY, ^{15}N-TOCSY-HSQC and HCCH-TOCSY spectra were used to assign aliphatic side-chain atom chemical shifts. Aromatic resonances were assigned from 2D DQF-COSY and TOCSY spectra and from 2D ^{13}C-CT-HSQC and (HB)CB(CGCD)HD and (HB)CB(CGCDCE)HE spectra. Exchange of backbone amide hydrogen atoms was

monitored by ^{15}N-HSQC spectra following suspension of protiated ^{15}N-labeled protein in deuterated buffer.

Structure determination

Distance restraints were derived from two 3D ^{13}C-NOESY-HSQC spectra (aliphatic and aromatic), a 3D ^{15}N-NOESY-HSQC spectrum and a 2D ^{1}H NOESY spectrum. All NOESY spectra were acquired with a 75 ms mixing time. $^{3}J_{HNHA}$ coupling constants were extracted from the HNHA spectrum. These were used, in combination with TALOS[43] analysis of chemical shifts, in the selection of dihedral angle restraints. Where TALOS and coupling constant analyses were consistent, both ϕ and ψ restraints were included. Where TALOS failed to make a prediction, a ϕ restraint was included if warranted by the coupling constant. Error bounds on dihedral restraints were set to ± 30°.

A set of 586 manually assigned NOE-derived distance restraints and 57 dihedral angle restraints were used as initial input for ARIA1.2.[30] ARIA identified 659 additional NOESY cross peaks, for a total of 953 unambiguous and 292 ambiguous distance restraints. At this stage, separate ARIA calculations were carried out fixing the methyl group stereochemistry of each V or L residue in the sequence in turn to obtain stereospecific assignments. In each case, one choice of assignments yielded an ensemble of structures with lower energies, lower χ_1 (and χ_2 for L residues) circular order parameters, and fewer NOE restraint violations than the alternate choice. Finally, the ensemble was examined for likely hydrogen bonds. Hydrogen bonds were judged to be present, and restraints included, if the amide proton had a hydrogen exchange protection

factor ≥ 1000 and if the residue was in a helix. Nineteen residues were thus restrained (1.3 Å < d_{NH-O} < 2.5 Å and 2.3 Å < d_{N-O} < 3.5 Å). Of 100 structures generated in a final ARIA calculation using all of these restraints, 43 had no NOE restraint violations > 0.5 Å and no dihedral angle restraint violations > 5°. This subset was analyzed with MOLMOL[44] and PROCHECK.[45]

References

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

2. Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proc Natl Acad Sci U S A* 98, 14274-14279.

3. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* 305, 619-31.

4. Marshall, S. A., Morgan, C. S. & Mayo, S. L. (2002). Electrostatics significantly affect the stability of designed homeodomain variants. *J Mol Biol* 316, 189-199.

5. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature* 423, 185-90.

6. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356, 539-542.

7. Gordon, D. B. & Mayo, S. L. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J Comput Chem* 19, 1505-1514.

8. Gordon, D. B. & Mayo, S. L. (1999). Branch-and-Terminate: a combinatorial optimization algorithm for protein design. *Structure Fold Des* 7, 1089-98.

9. Pierce, N. A., Spriet, J. A. & Mayo, S. L. (2000). Conformational splitting: A more powerful criterion for dead-end elimination. *J Comput Chem* 21, 999-1009.

10. Gordon, D. B., Hom, G. K., Mayo, S. L. & Pierce, N. A. (2003). Exact rotamer optimization for protein design. *J Comput Chem* 24, 232-243.

11. Voigt, C. A., Gordon, D. B. & Mayo, S. L. (2000). Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* 299, 789-803.

12. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J Chem Phys* 21, 1087-1092.

13. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220, 671-680.

14. Shah, P. S., Hom, G. K. & Mayo, S. L. (2003). Preprocessing of rotamers in protein design calculations. *Submitted*.

15. Desmet, J., Spriet, J. & Lasters, I. (2002). Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* 48, 31-43.

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

17. Dantas, G., Kuhlman, B., Callender, D., Wong, M. & Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J Mol Biol* 332, 449-60.

18. Clarke, N. D., Kissinger, C. R., Desjarlais, J., Gilliland, G. L. & Pabo, C. O. (1994). Structural studies of the engrailed homeodomain. *Protein Sci* 3, 1779-87.

19. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci* 5, 895-903.

20. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* 94, 10172-7.

21. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci* 6, 1333-7.

22. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Fold Des* 3, 253-8.

23. Dunbrack, R. (2002). Rotamer Libraries in the 21(st) Century. *Curr Opin Struct Biol* 12, 431.

24. Dunbrack, R. L., Jr. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* 230, 543-74.

25. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-402.

26. Morgan, C. S. (2000). Full sequence design of an alpha-helical protein and investigation of the importance of helix dipole and capping effects in helical protein design, California Institute of Technology.

27. Mayor, U., Johnson, C. M., Daggett, V. & Fersht, A. R. (2000). Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc Natl Acad Sci U S A* 97, 13518-22.

28. Semisotnov, G. V., Rodionova, N. A., Razgulyaev, O. I., Uversky, V. N., Gripas, A. F. & Gilmanshin, R. I. (1991). Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe. *Biopolymers* 31, 119-28.

29. Bhattacharjya, S. & Balaram, P. (1997). Hexafluoroacetone hydrate as a structure modifier in proteins: characterization of a molten globule state of hen egg-white lysozyme. *Protein Sci* 6, 1065-73.

30. Nilges, M., Macias, M. J., O'Donoghue, S. I. & Oschkinat, H. (1997). Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. *J Mol Biol* 269, 408-22.

31. Ledneva, R. K., Alekseevskii, A. V., Vasil'ev, S. A., Spirin, S. A. & Kariagina, A. S. (2001). [Structural aspects of homeodomain interactions with DNA]. *Mol Biol (Mosk)* 35, 764-77.

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

33. Lazar, G. A., Desjarlais, J. R. & Handel, T. M. (1997). De novo design of the hydrophobic core of ubiquitin. *Protein Sci* 6, 1167-78.

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

35. Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y. & Baker, D. (2002). Accurate computer-based design of a new backbone conformation in the second turn of protein L. *J Mol Biol* 315, 471-7.

36. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Structure Fold Des* 7, R105-9.

37. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr Opin Struct Biol* 9, 509-13.

38. Prodromou, C. & Pearl, L. H. (1992). Recursive PCR: a novel technique for total gene synthesis. *Protein Eng* 5, 827-9.

39. Johnson, B. H. & Hecht, M. H. (1994). Recombinant proteins can be isolated from E. coli cells by repeated cycles of freezing and thawing. *Biotechnology (N Y)* 12, 1357-60.

40. Santoro, M. M. & Bolen, D. W. (1988). Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants. *Biochemistry* 27, 8063-8.

41. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6, 277-93.

42. Johnson, B. A. & Blevins, R. A. (1994). NMR View - A computer program for the visualization and analysis of NMR data. *J Biomol NMR* 4, 603-614.

43. Cornilescu, G., Delaglio, F. & Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR* 13, 289-302.

44. Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* 14, 51-5, 29-32.

45. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* 8, 477-86.

Table 4-1: Thermodynamic data of variants and wild type.

Thermodynamic data[a]			
	Wild type	FSM1_VF	FSM1_MC
ΔG_{unfold} (kcal/mol)	1.9[b]	4.2	4.2
T_m (°C)	43[c]	>99	>99
m value[d] (kcal/mol M)	0.8[b]	1.3	1.2
C_m (M)[e]	1.5[b]	3.2	3.5

[a] All data were collected with protein in 50 mM phosphate, pH 5.5 unless noted. ΔG_{unfold} was calculated from experiments performed at 25 °C using guanidinium hydrochloride denaturation.
[b] Mayor *et al.*[27] (done at pH 5.8 at 25 °C using urea denaturation).
[c] Morgan[26] (done in 5 mM phosphate buffer, pH 4.5).
[d] Slope of ΔG_{unfold} *versus* denaturant concentration.
[e] Midpoint of unfolding transition.

Table 4-2: NMR structure statistics.

NMR structure statistics[a]	
Summary of restraints	
NOE distance restraints	1245
Unambiguous	953
Ambiguous	292
Hydrogen bonds[b]	19
Dihedral angle (ϕ,ψ) restraints[c]	57
R.m.s. deviation from restraints	
NOE restraints (Å)	0.024 ± 0.004
Dihedral restraints (°)	0.26 ± 0.12
R.m.s. deviation from idealized geometry	
Bonds (Å)	0.0037 ± 0.0002
Angles (°)	0.53 ± 0.03
Improper (°)	1.57 ± 0.14
Ensemble atomic r.m.s. deviations from mean structure[d] (Å)	
Backbone	0.59
All heavy	1.29
Ensemble Ramachandran statistics[e]	
Residues in most-favored region (%)	83.2
Additionally allowed region (%)	13.4
Generously allowed region (%)	2.3
Disallowed region (%)	1.1

[a] Statistics calculated for the ensemble of 43 structures (out of 100 calculated in ARIA[30]) which had no NOE restraint violations >0.5 Å and no dihedral restraint violations >5°.
[b] Each hydrogen bond yields two experimental restraints.
[c] Dihedral angle restraints were derived from HNHA analysis and chemical shift analysis with TALOS[43]. ψ restraints based on TALOS results were included if the HNHA and TALOS results were in agreement for the corresponding ϕ and if the residue was found to be in a helical conformation in structures calculated in the absence of angle restraints.
[d] Ensemble precision was calculated for residues 3-45.
[e] Ramachandran analysis was performed with Procheck.[45]

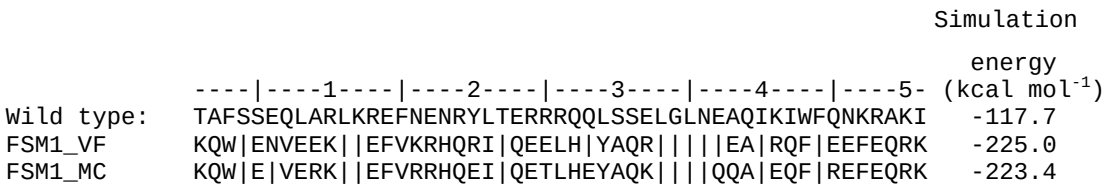


Figure 4-1: Comparison of designed sequences with wild type.
Sequence alignment and simulation energies of the wild-type sequence and the designed variants of ENH, FSM1_VF, and FSM1_MC. Positions that have the same identity as the wild type are indicated with a bar. FSM1_MC has 40 mutations and FSM1_VF has 39 mutations, differing from the wild-type sequence by 79% and 77%, respectively. FSM1_MC and FSM1_VF have all but 11 residues in common.

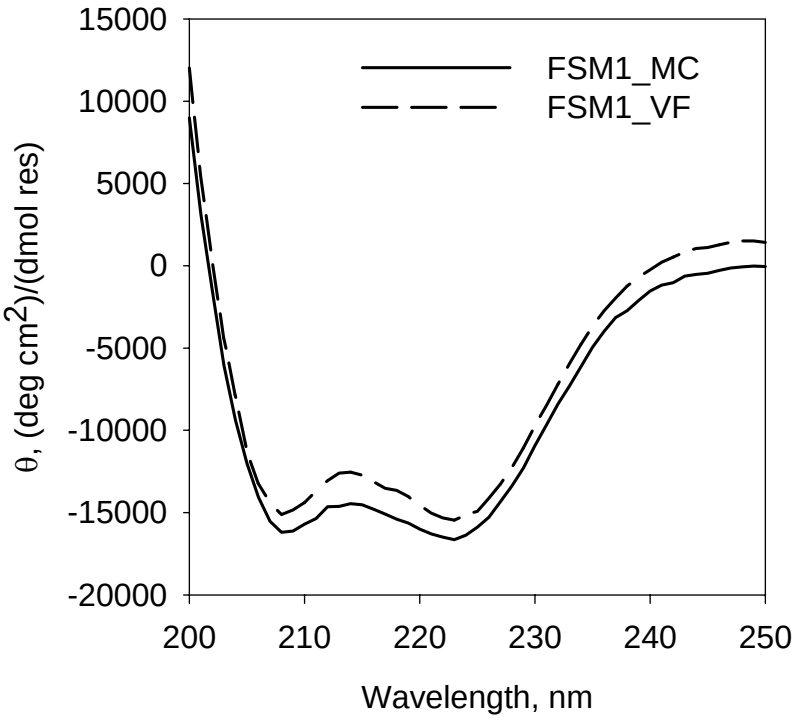


Figure 4-2: Far UV wavelength spectra of designed variants.
Circular dichroism wavelength scans of FSM1_VF and FSM1_MC. Spectra were obtained at 25 °C in 50 mM phosphate buffer at pH 5.5.

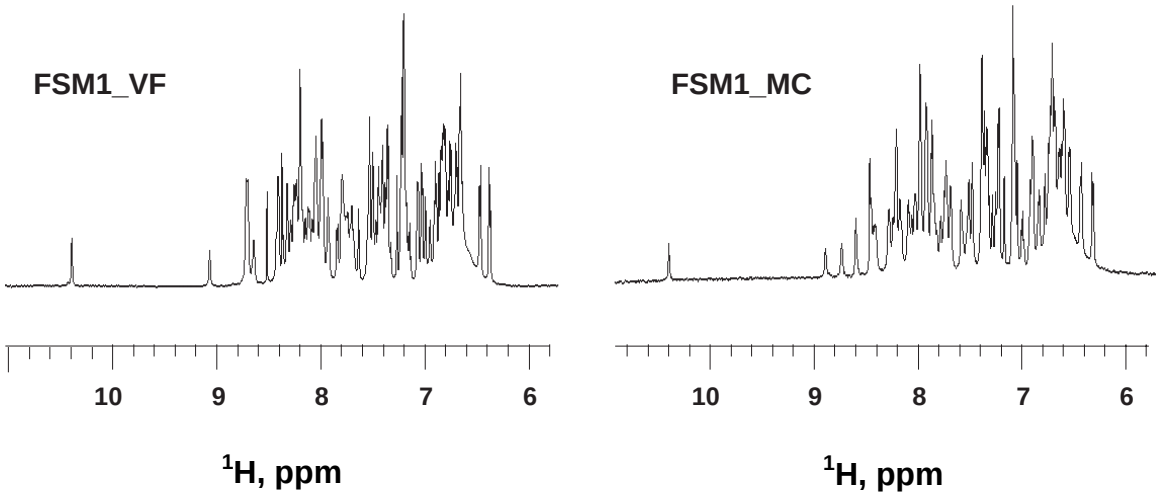


Figure 4-3: 1D ^{1}H NMR spectra of designed variants.
1D, ^{1}H nuclear magnetic resonance (NMR) spectra of FSM1_VF and FSM1_MC. For clarity, only the amide region is shown. The sharp, dispersed lines are characteristic of well-folded proteins.

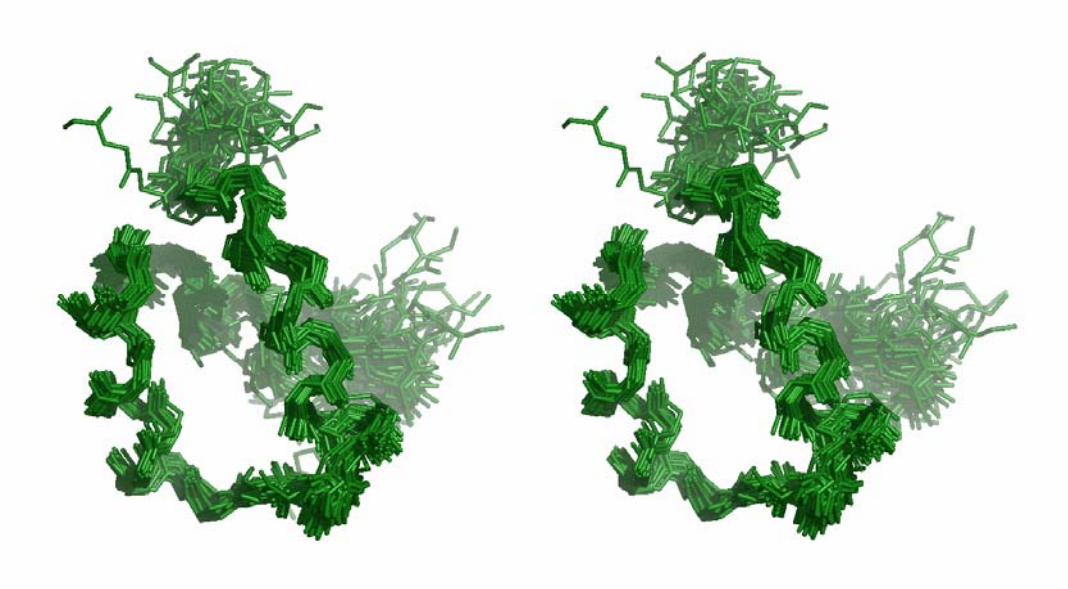


Figure 4-4: FSM1_VF ensemble.
Stereoview of the FSM1_VF structure ensemble. Best-fit superposition of 43 simulated annealing structures, showing the backbone. The N terminus is located at the top of the image.

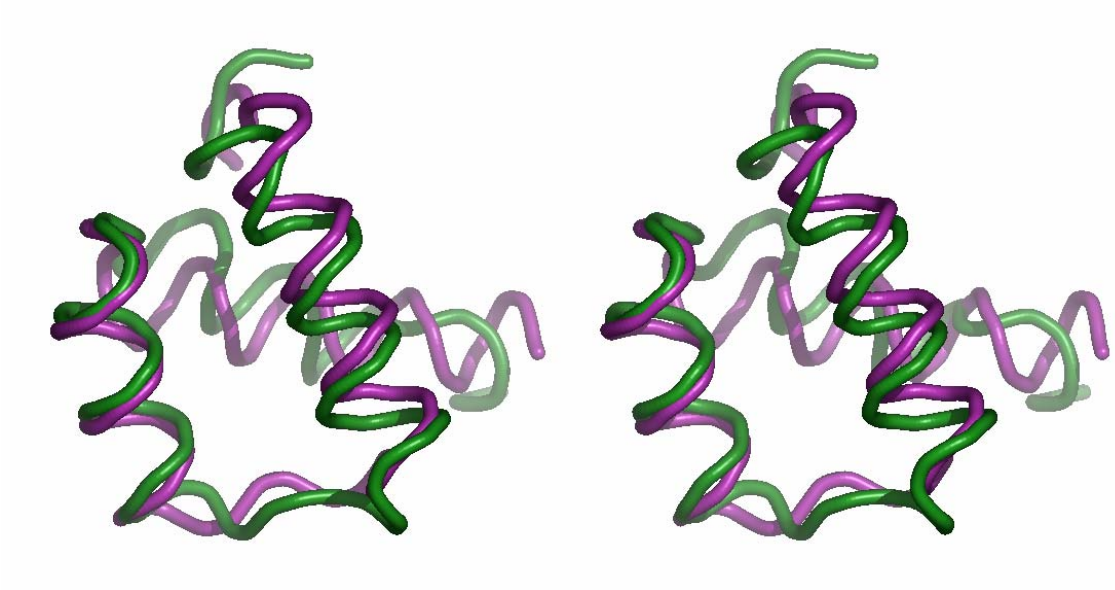


Figure 4-5: Superposition of FSM1_VF with crystal structure.
Stereoview of the backbones of FSM1_VF (green) and the crystal structure of ENH (purple). The r.m.s. deviation of C_α atoms of residues 3-45 is 2.5 Å

Chapter 5

Computational Design of a Water Soluble Variant of Bacteriorhodopsin

Abstract

The membrane protein bacteriorhodopsin acts as a protein pump in Halobacterium; isomerization of a covalently attached cofactor, retinal, drives the pump. Computational protein design techniques were used to design a water soluble variant of bacteriorhodopsin. Using exposed nonpolar surface area as a metric, we designed the surface of bacteriorhodopsin to resemble water soluble proteins; a database survey of water soluble proteins structures and membrane protein structures provided distributions centered at approximately 63% and 93%, respectively. The designed variant, which is a 58-fold mutant of the wild type, was expressed with high yields into inclusion bodies in *E. coli*, purified using a Ni^{2+} affinity column, and re-folded using rapid dilution. The protein is highly soluble and stable at concentrations up to 2.5 mg/ml in aqueous buffer, but was unable to incorporate retinal. As a result, the designed protein exists in equilibrium between monomer, dimer, and mostly high order aggregated states. Biophysical characterization showed that although the variant displays properties of helical proteins as measured by circular dichroism spectroscopy, it is most likely a molten globule—likely due to its inability to incorporate retinal.

Introduction

The number of protein structures available from the protein data bank (PDB)[1] has grown considerably over the past decade; almost 24,000 structures were available at the end of 2003 compared to approximately 1,700 structures at the end of 1993. However, due to the difficulty associated with determining structures of membrane proteins, only 83 have been deposited in the PDB as of this writing. The slow growth in the availability of high resolution x-ray structures of membrane proteins is primarily due to the difficulties typically encountered when working with them. These include low levels of protein expression, low stability in detergents, and the inability to generate high-quality crystals that diffract well. Obtaining water soluble variants of membrane proteins might alleviate some of these difficulties. In addition, water soluble variants that do not compromise structural integrity can provide insights into the different forces contributing to protein stabilization in membranes.

An important class of membrane proteins is the G-protein coupled receptors (GPCRs). Although sharing a conserved structure comprised of seven trans-membrane helices, the natural ligands for these receptors are extremely diverse. Ligand binding leads to conformational changes in GPCRs. These changes serve as a switch, transferring the signal to the trimeric guanine nucleotide binding regulatory proteins (G proteins), thus inhibiting or stimulating the production of intracellular secondary messengers such as cyclic adenosine monophosphate (cAMP) and Ca^{2+} ions.

From a drug discovery standpoint, GPCRs are prominent; 50% of all drugs launched in 2001 targeted GPCRs, producing worldwide sales exceeding $30 billion.[2] Design of novel drugs targeting GPCRs can be aided by obtaining water soluble variants

that maintain structural identity with the wild type; water soluble variants will afford researchers the opportunity to obtain more accurate binding constants in non-detergent environments, thus leading to better binding drugs.

A computational design approach[3-6] to generating water soluble membrane protein variants requires a high-resolution crystal structure as a starting point. To date, only one mammalian GPCR three-dimensional structure has been solved—bovine rhodopsin.[7] The paucity of GPCR structures and the low resolution of the single mammalian structure led us to use the structure of bacteriorhodopsin (BR) as a paradigm to probe our ability to convert GPCRs to water soluble proteins. BR is a light-driven proton pump from Halobacteria that possesses an all-*trans*-retinal whose isomerization drives the proton pump. Comparison of BR with bovine rhodopsin shows that their overall topologies are similar (Figure 5-1).[8] Superpositions using structural alignments alone reveal that the structures have a C_α root mean square deviation (RMSD) of only 2.13Å.[8] BR has been intensely investigated and the numerous structural studies have provided us with a high resolution structure (1.55Å, PDB code: 1c3w[9]).

An early hypothesis suggested that membrane proteins are "inside-out proteins" stabilized predominantly by polar interactions. This idea has been proven false. Structure analysis of known membrane proteins has revealed that their general structural features compare remarkably well with those of water soluble proteins.[10-12] The average hydrophobicity of the core is the same for both types of proteins, and the same interactions contribute to protein stability. However, the surfaces differ; membrane protein surfaces are predominantly nonpolar in composition, while water soluble protein surfaces are more polar. The common forces observed in both types of proteins have led

many researchers to suggest that water soluble variants of membrane proteins can be created by converting their surfaces to resemble those of water soluble proteins.

Computational protein design identifies optimal amino acid sequences that are compatible with a protein backbone. Our goal was to start with the high-resolution bacteriorhodopsin structure, redesign a more polar surface applying our previously determined rules for protein surface designs,[13,14] and experimentally validate the resulting variants by testing for solubility in water.

Results

Water soluble versus membrane protein surfaces

Effectively converting membrane protein surfaces to resemble those of water soluble proteins requires an understanding of how the surfaces differ. We calculated the nonpolar content of the surfaces of 16 membrane proteins and compared them with the nonpolar content of the surfaces of 49 water soluble proteins (Figure 5-2). For water soluble proteins, we observed a distribution centered at 64%. In contrast, membrane protein surfaces have a higher nonpolar content, with a distribution centered at 93%.

BR exists as a lipid-mediated trimer in the membranes of Halobacteria. Due to the nonpolar nature of lipid-mediated contacts, we were concerned that once the nonpolar surface of BR was converted to a more polar one, the trimer would not assemble. Therefore, we worked with only a monomer unit of BR; BR exists as a stable monomer in detergents. Wild-type BR (PDB code: 1c3w) has a nonpolar surface area of 93%. This surface area analysis served as a metric for our BR surface design; we used

computational design to convert BR so that its nonpolar surface area would fall within the water soluble protein regime.

Computational design of bacteriorhodopsin

We used an automated algorithm that divides a protein structure into core, boundary, and surface residues based on the residue's C_β distance from a solvent-accessible surface.[15,16] For the BR design, residues classified as core or boundary were kept fixed in both identity and conformation. The retinal cofactor is attached via a Schiff base mechanism to Lys 216 and was classified as a core residue. All nonpolar residues on the surface were designed using previously established rules.[13,14] Surface residues that were already polar were allowed to change in conformation but their identities were kept fixed. The four Gly residues on the surface were designed with the rules applied to nonpolar residues because their backbone phi/psi angles are in helical space. The ground-state sequence was identified from approximately 10^{160} possible sequences using an algorithm[3,17] based on the dead-end-elimination (DEE) theorem.[18] The predicted sequence (Figure 5-3), a 58-fold mutant of the wild type, has a nonpolar surface area of 63%—in the regime of our surveyed water soluble proteins.

Expression and purification of designed bacteriorhodopsin

The gene (Blue Heron Biotechnology) encoding the designed BR variant (WS-BR) was expressed in *E. coli* with a six-residue N-terminal His tag. Expression was carried out under various conditions to try to incorporate the retinal cofactor at this step (see Table 1). All *trans* retinal (Sigma) was added in various concentrations and

expression was carried out at temperatures ranging from 20 ºC to 42 ºC. We did not observe the incorporation of retinal as monitored by colorometric analysis and mass spectroscopy under any of the conditions tested. Purification was performed under denaturing conditions on a Ni^{2+} affinity column (Figure 5-4). Dialysis of the sample to remove denaturant resulted in a large amount of precipitate. As a result, the eluent was rapidly diluted 200-fold to reduce the denaturant concentration and then subjected to ultra concentration to reduce the volume. We found that rapid dilution with high concentrations of NaCl (1.5 M compared to 0.1 M) resulted in far less precipitation of WS-BR upon concentration. A 1.5 M NaCl concentration was therefore maintained in all subsequent protein analysis unless noted. The protein was stable at concentrations up to 2.5 mg/ml in aqueous buffer for up to a week. Electrospray mass spectrometry confirmed that the sample was the apo form of BR.

Attempts to incorporate retinal

The membranes of Halobacteria are referred to as “purple membranes” due to the deep purple color resulting from the retinal chromophore in BR. This purple color serves as a convenient spectroscopic assay to query for retinal incorporation. In addition, since the association of retinal to BR is covalent in nature (via a Schiff base mechanism), mass spectroscopy can be used to confirm the presence of retinal. Attempts to incorporate all-*trans* retinal were unsuccessful. Table 5-1 details the experiments performed at each step of the expression, purification, and refolding process.

Physical analysis of WS-BR

Far ultra-violet circular dichroism (CD) spectroscopy of WS-BR at 25 ºC (Figure 5) suggests a molecule that contains elements of secondary structure; unfolded proteins or random structure display a single minimum at 205 nm. The spectra of α-helical proteins display minima at 208 and 222 nm. WS-BR exhibits a broad minimum in the 208-220 nm range. Therefore, although WS-BR does not exhibit the hallmarks of unfolded proteins, it also does not resemble a typical α-helical protein.

Analytical gel filtration chromatography shows that the WS-BR sample was not mono-dispersed. We found a distribution of oligomeric states (Figure 5-6) with the majority of the sample existing as high-order oligomers and only small amounts appearing as monomers or dimers. Analytical ultracentrifugation results (data not shown) also suggest a distribution of states.

ANS (1-anilino-napthalene-8-sulfonate) binding was used to further validate the structural integrity of WS-BR. ANS selectively binds molten globule states of proteins. Molten globules exhibit pronounced secondary structure and compactness but lack packed tertiary structure. Hen egg-white lysozyme (HEWL) in 25% HFA (hexafluroacetone hydrate) was used as a positive control; under this condition, HEWL binds ANS and exhibits molten globule characteristics. WS-BR binds ANS in a concentration dependent manner (Figure 5-7). This result suggests that WS-BR is most likely a water soluble molten globule.

Temperature denaturation of WS-BR also confirmed that the protein is most likely a non-monodispersed molten globule (Figure 5-8). Under low (0.1 M) NaCl conditions (Figure 8A), WS-BR unfolds slowly and large amounts of precipitate are

observed following the experiment. At a much higher concentration of NaCl (1.5M, Figure 8B), WS-BR exhibits a more stable baseline and a sharper transition to the unfolded state; however, a precipitate is still observed following the experiment. These results suggest that although higher concentrations of NaCl help stabilize WS-BR, the salt merely helps by holding an aggregated molten globule-like species together longer.

Discussion

We used BR because of its structural similarity to the only mammalian GPCR structure that has been solved (Figure 5-1). In hindsight, the selection of BR as our first design target may have been too ambitious. However, the lessons learned from working with BR will help us identify other suitable targets and will facilitate attempts to make them water soluble.

Computational protein design seeks to find amino acid sequences compatible with a target fold without regard to the folding pathway traversed to obtain it. Disregarding the folding pathway can be a problem, however, when designing proteins that require specific folding pathways in order to incorporate ligands or cofactors. The folding pathway of BR has been well-studied[19-21] and it is hypothesized that the formation of an intermediate is required before retinal can be first loosely associated and then covalently attached (Figure 5-9). Once removed from the membrane, the folding pathway of BR is most likely disrupted or altered, preventing the formation of the requisite intermediate and therefore the incorporation of retinal.

The physical data obtained for WS-BR leads us to conclude that the molecule is an aggregate of molten globules that contains some helical character. These observations

are consistent with previous work. The binding of retinal provides stability to BR[22] but is not necessary for initiating helix formation.[19] Even after subjecting WS-BR to a myriad of conditions, we were unable to incorporate retinal into the protein. Without the incorporation of retinal, the computationally predicted structure of WS-BR contains a large void in its core that most likely leads to the observed molten globule states and subsequent aggregation.

A recent study by DeGrado and colleagues[23] demonstrated their ability to design water soluble analogues of the KcsA potassium channel. In light of their work and our results for WS-BR, we have identified certain criteria that should be met in our designs for the immediate future. First, an ideal design target should contain no cofactors that are incorporated during the folding of the molecule for the reasons mentioned above. Second, a protein whose functional unit is a homo-oligomer might be beneficial; the interfaces of the subunits can provide a driving force for folding, and the number of mutations required on each subunit will most likely be significantly less. However, it is important that the molecule not undergo any changes during the formation of quaternary structure.

While we were unable to obtain a properly folded WS-BR with retinal incorporated, our molecule was soluble in high concentrations in aqueous buffer for long periods of time. This suggests that our strategy for converting membrane proteins into water soluble variants is probably sound, but the choice of BR as an initial target may not have been prudent.

Materials & Methods

Nonpolar surface area calculations

We calculated the exposed nonpolar surface area of 48 water soluble proteins. The proteins were a subset of the Top-100 set from Richardson and colleagues (Table 5-2).[24] The exposed nonpolar surface areas were calculated using the coresurf_z program (J.J. Plecs, Caltech). Briefly, a 1.4 Å radius probe was rolled over the surface of the protein structure, generating a dot surface that was used to obtain surface areas. Nitrogen and oxygen atoms were considered polar while carbon atoms were considered nonpolar.

The surface area calculations for the membrane proteins we surveyed were performed in the same way as above. However, only the membrane-spanning region of the protein was analyzed. We therefore evaluated the most hydrophobic 30 Å stretch in the protein as determined by the number of carbon atoms (Figure 5-10). This follows the work of Spencer and Rees[25] and effectively identifies the membrane spanning region of membrane proteins.

Computational design of bacteriorhodopsin

The 1.55 Å structure of bacteriorhodopsin (PDB code: 1c3w)[9] was used as the template for our designs. All lipids and water molecules were eliminated from the structure file. The retinal was manually attached to Lys 216. Hydrogens were added using MOLPROBITY.[26] To relieve backbone strain and eliminate clashes, 50 steps of conjugate gradient minimization were performed on the molecule using SMIN from the ORBIT[3] suite of protein design programs.

Residues were classified as either core, boundary, or surface using an automated algorithm.[15,16] Residues that were classified as core or boundary were kept fixed in identity and conformation. Nonpolar residues (Trp, Tyr, Ala, Phe, Val, Ile, Leu, Met) that were classified as surface were designed to be either Asn, Glu, Gln, His, Lys, Arg, Asp, or Ala. Surface residues that were already polar were fixed in identity but allowed to change conformation. Furthermore, surface positions that were in helix N-capping positions or participated in helix dipoles were allowed to be designed according to previously established rules.[14] The four Gly residues had backbone conformations in helical phi/psi space and were therefore designed to be polar. A backbone-dependent rotamer library based on that of Dunbrack and Karplus[27,28] was used.

Expression and purification of WS-BR

The gene encoding the WS-BR amino acid sequence was purchased from Blue Heron Biotechnology and cloned into Novagen's pET15b vector downstream of a six-residue His tag. The protein was expressed in BL21(DE3) cells (Stratagene) under the control of an IPTG-inducible promoter. A final IPTG concentration of 1 mM was used to induce protein expression. Variations in the expression protocol were explored to induce incorporation of retinal (see Table 1). Cells were harvested following expression and lysed by sonication. Inclusion bodies were separated from the soluble fraction by centrifugation at 30,000 x *g* for 30 min. SDS gels showed that >99% of the protein expressed into the inclusion bodies. Inclusion bodies were dissolved in 6 M guanidinium HCl (GuHCl) and the protein was purified under denaturing conditions on a Ni^{2+} affinity column. The loading buffer was 6 M GuHCl, 0.1 M NaH_2PO_4, 0.1 M Tris-Cl, pH 8.0.

The column was washed with 6 M GuHCl, 0.1 M NaH_2PO_4, 0.1 M Tris-Cl, pH 6.3, and pH 5.9. The protein was eluted with 6 M GuHCl, 0.1 M NaH_2PO_4, 0.1 M Tris-Cl, pH 4.5.

The sample was refolded by rapidly diluting with 50 mM phosphate buffer, pH 7.2 so that a negligible concentration of GuHCl remained. Various concentrations of NaCl were included in the buffer (see Table 5-1) but no less than 0.1 M. In addition, many different conditions were used to try incorporating retinal (see above and Table 5-1). The sample was concentrated using Amicon's ultra-concentration apparatus. In 50 mM phosphate, 1.5 M NaCl, pH 7.2, the protein was stable at room temperature in concentrations of at least 3 mg/ml for up to a week.

Analysis of WS-BR

CD spectroscopy was performed on an Aviv 62DS equipped with a thermoelectric unit. The buffers for CD analysis varied depending on the state of the protein being analyzed. Temperature melts were done by incrementing the temperature in 1 ºC steps from 1 ºC to 99 ºC allowing the temperature to settle at each temperature for 2 min and using a 30 sec averaging time.

Analytical gel filtration chromatography was performed on Perkin Elmer's Biocad 710E using the S-75 column. Results were compared to molecular weight standards run under identical conditions.

References

1. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112, 535-42.
2. Klabunde, T. & Hessler, G. (2002). Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem* 3, 928-44.
3. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci* 5, 895-903.
4. Dahiyat, B. I. (1999). In silico design for protein stabilization. *Curr Opin Biotechnol* 10, 387-90.
5. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr Opin Struct Biol* 9, 509-13.
6. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Structure Fold Des* 7, R105-9.
7. Okada, T., Fujiyoshi, Y., Silow, M., Navarro, J., Landau, E. M. & Shichida, Y. (2002). Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography. *Proc Natl Acad Sci U S A* 99, 5982-7.
8. Teller, D. C., Okada, T., Behnke, C. A., Palczewski, K. & Stenkamp, R. E. (2001). Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of G-protein-coupled receptors (GPCRs). *Biochemistry* 40, 7761-72.
9. Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P. & Lanyi, J. K. (1999). Structure of bacteriorhodopsin at 1.55 A resolution. *J Mol Biol* 291, 899-911.
10. Langosch, D. & Heringa, J. (1998). Interaction of transmembrane helices by a knobs-into-holes packing characteristic of soluble coiled coils. *Proteins* 31, 150-9.
11. Bowie, J. U. (1997). Helix packing in membrane proteins. *J Mol Biol* 272, 780-9.
12. Rees, D. C., Komiya, H., Yeates, T. O., Allen, J. P. & Feher, G. (1989). The bacterial photosynthetic reaction center as a model for membrane proteins. *Annu Rev Biochem* 58, 607-33.
13. Morgan, C. S. (2000). Full sequence design of an alpha-helical protein and investigation of the importance of helix dipole and capping effects in helical protein design, California Institute of Technology.
14. Marshall, S. A., Morgan, C. S. & Mayo, S. L. (2002). Electrostatics significantly affect the stability of designed homeodomain variants. *J Mol Biol* 316, 189-199.
15. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science* 278, 82-7.

16. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* 305, 619-31.

17. Gordon, D. B. & Mayo, S. L. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J Comput Chem* 19, 1505-1514.

18. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356, 539-542.

19. Booth, P. J., Flitsch, S. L., Stern, L. J., Greenhalgh, D. A., Kim, P. S. & Khorana, H. G. (1995). Intermediates in the folding of the membrane protein bacteriorhodopsin. *Nat Struct Biol* 2, 139-43.

20. Booth, P. J. (2000). Unravelling the folding of bacteriorhodopsin. *Biochim Biophys Acta* 1460, 4-14.

21. Lu, H. & Booth, P. J. (2000). The final stages of folding of the membrane protein bacteriorhodopsin occur by kinetically indistinguishable parallel folding paths that are mediated by pH. *J Mol Biol* 299, 233-43.

22. Kahn, T. W., Sturtevant, J. M. & Engelman, D. M. (1992). Thermodynamic measurements of the contributions of helix-connecting loops and of retinal to the stability of bacteriorhodopsin. *Biochemistry* 31, 8829-39.

23. Slovic, A. M., Kono, H., Lear, J. D., Saven, J. G. & DeGrado, W. F. (2004). Computational design of water-soluble analogues of the potassium channel KcsA. *Proc Natl Acad Sci U S A* 101, 1828-33.

24. Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S. & Richardson, D. C. (1999). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol* 285, 1711-33.

25. Spencer, R. H. & Rees, D. C. (2002). The alpha-helix and the organization and gating of channels. *Annu Rev Biophys Biomol Struct* 31, 207-33.

26. Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). Structure validation by C_α geometry: ϕ,ψ and C_β deviation. *Proteins* 50, 437-50.

27. Dunbrack, R. L., Jr. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* 230, 543-74.

28. Dunbrack, R. (2002). Rotamer Libraries in the 21(st) Century. *Curr Opin Struct Biol* 12, 431.

Table 5-1: Different conditions explored in efforts to obtain a mono-dispersed WS-BR with the incorporation of retinal.[a]

Step	Condition	Result
Expression	*Temperature (20-42 ºC)*	Protein in inclusion bodies.
	Addition of all-trans-retinal (5-50 μM)	Protein in inclusion bodies; no indication of retinal incorporation pre/post purification.
	Induction time (1-5 h)	Increasing levels of protein until 3 h, but almost all in inclusion bodies.
Purification	*Ni^{2+} affinity chromatography under denaturing conditions*[b]	Pure protein, soluble in 6 M Gu-HCl.
Refolding	*Dialysis to remove denaturant (varying sample volume starting in 6 M Gu-HCl)* • Addition of all-*trans*-retinal (5-50 μM) • Addition of L-Arg (200 mM)	Large amounts of protein precipitation under all conditions.
	Rapid dilution to remove denaturant (varying final volume of Gu-HCl (20-100 μM)) • All dilutions carried out in 50 mM phosphate buffer • pH varied (5.0-9.0) • Presence of NaCl (100 mM, 500 mM, and 1.5 M) • Addition of TFE[c] (10-50%) • Addition of all-*trans*-retinal under most of the conditions listed (5-50 μM) • Temperature (4 ºC or room temp) • Addition of L-Arg (200 mM)	Little protein precipitation under all conditions but no evidence of incorporation of retinal. Protein remained stable in aqueous buffer at concentrations up to 2.5 mg/ml. A little less precipitation as NaCl concentration was increased. All conditions resulted in a non-mono-dispersed protein solution.

[a] This table is representative of variations performed at each step, but is not exhaustive.
[b] According to protocol from Qiagen Inc.
[c] 2,2,2-Trifluoroethanol.

Table 5-2: Water soluble proteins used to determine exposed nonpolar surface areas.

PDB Code	Length[a]	Nonpolar S.A. (%)[b]	PDB Code	Length[a]	Nonpolar S.A. (%)[b]
1aac	105	67	1ptx	64	67
1amn	174	59	1rcf	169	55
1aru	344	67	1rgeA	96	64
1benAB	51	68	1rroH	108	58
1bkf	107	64	1smd	496	64
1cem	395	60	1ttaA	127	65
1cnr	46	74	1whi	122	61
1cnv	283	61	1xyzA	319	62
1ctj	89	62	2cpl	164	64
1cus	213	64	2end	137	67
1fus	107	60	2erl	40	65
1igd	61	61	2hft	211	60
1iro	53	58	2mhr	118	68
1jbc	237	62	2msbA	111	65
1kap	481	63	2phy	125	59
1knb	185	65	2rhe	114	65
1lit	144	65	2rn2	155	63
1lkk	133	60	3b5c	85	63
1mla	307	66	3chy	127	63
1mrj	247	60	3ebx	62	64
1nif	332	65	3lzm	164	63
1phb	404	64	3pte	347	63
1plc	99	59	4fgf	123	62
1ptf	87	64	7rsa	124	62

[a] Length (number of residues) was determined from the structure file.
[b] Surface area (S.A.) calculated using the coresurf_z program (J. J. Plecs, California Institute of Technology).

Table 5-3: Membrane proteins used to determine exposed nonpolar surface areas.

PDB Code	Exposed Nonpolar S.A. (%)[a]	PDB Code	Exposed Nonpolar S.A. (%)[a]
1aij	94	1fx8	91
1bgy	89	1kzu	89
1bl8	92	1msl	90
1brx	93	1occ	91
1el2	92	1pcr	94
1eul	86	1qla	94
1f88	91	1qle	94
1fum	94	2prc	93

[a] Only the membrane spanning region was analyzed (see Methods). Surface are (S.A.) calculated using the coresurf_z program (J. J. Plecs, California Institute of Technology).

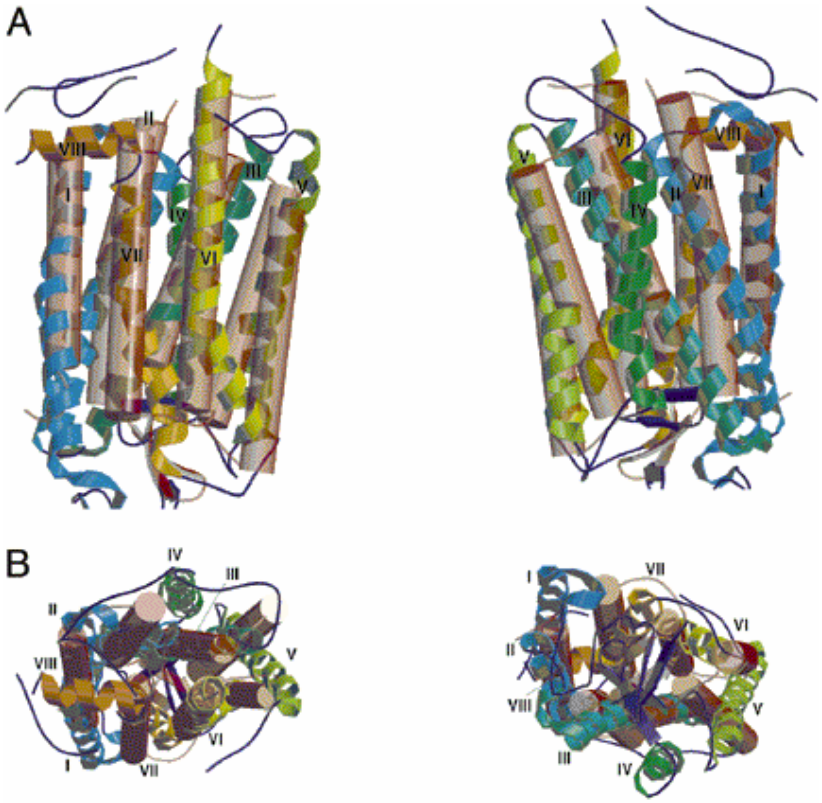


Figure 5-1: Superposition of bovine rhodopsin and bacteriorhodopsin.
Superposition of bacteriorhodopsin (pink transparent cylinders and connecting coil) on molecule A of bovine rhodopsin (colored helical ribbons and connecting coils). (A) On the right, the view is rotated 180º about the vertical axis. (B) On the left is a view of the top (cytoplasmic) surface of the molecules. Note the differences between helices IV and V in the two molecules. At the right is a bottom view of the molecules.

This figure was borrowed from Teller et al. (Reference 8).

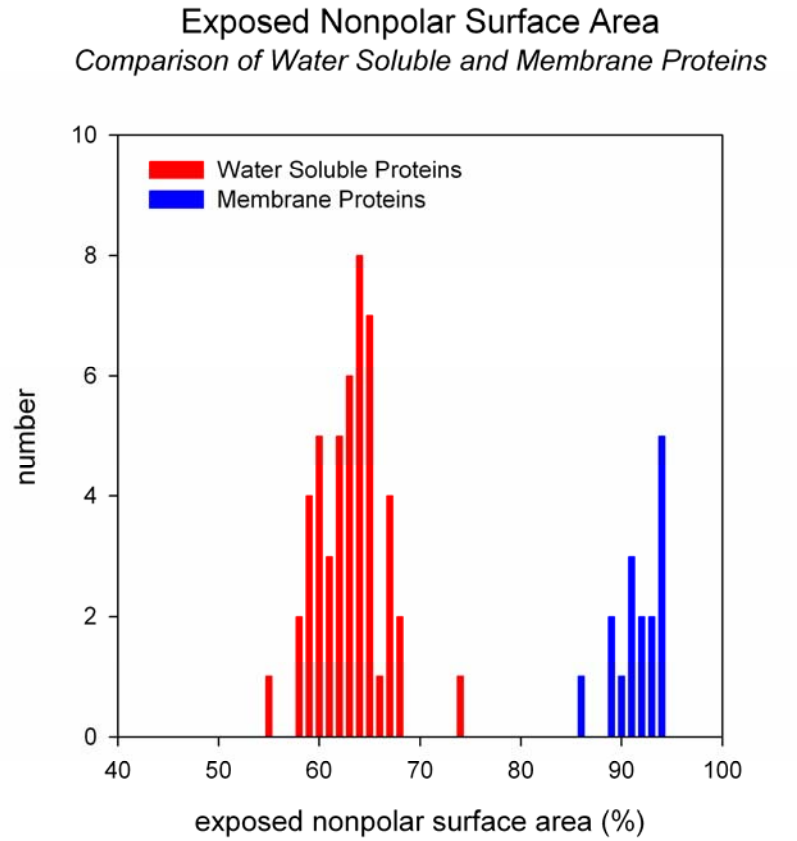


Figure 5-2: Exposed nonpolar surface areas of water soluble versus membrane proteins.
Database survey of exposed nonpolar surface areas of water soluble proteins compared to membrane proteins. The water soluble protein structures are a subset of Richardson and colleagues' Top100 set. The membrane protein structure files were obtained from Prof. D.C. Rees (California Institute of Technology). Only the membrane spanning regions of the membrane proteins were evaluated. Surfaces were analyzed using the coresurf_z program (J.J. Plecs, California Institute of Technology).

```
          *  **  *  **  ** **  *          *   *  *     *   * *   *          *   *  **     *
---------1---------2---------3---------4---------5---------6---------7---------8---------9---------100
MQAQITGRPEWIWLALGTALMGLGTLYFLVKGMGVSDPDAKKFYAITTLVPAIAFTMYLSMLLGYGLTMVPFGGEQNPIYWARYADWLFTTPLLLLDLAL
     TGRPEWEWLREGTDLMRDGTEEFRRKGEGVSDPDAKKFYHITTKVPEIAFTMYQSMLEGQGLTKVPFGGEQNPIYQARYQDWRETTPLLLEDLAL

         *  **  **  *   *  *  *  * **  **  *  **  *  **            *    *     **   *  *  **    * *
---------1---------2---------3---------4---------5---------6---------7---------8---------9---------200
LVDADQGTILALVGADGIMIGTGLVGALTKVYSYRFVWWAISTAAMLYILYVLFFGFTSKAESMRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEGAGIV
LVDADQGTIKALREADEEMIKTGLKGATTKEYSERERWWRQSTEAMKKILEVLREGF     SMRPEVDSTFKQLRNVTEKLWSKYPEVWQQGSEGQGNV

   *  *   *          *        *
---------1---------2---------3---------4---------
PLNIETLLFMVLDVSAKVGFGLILLRSRAIFGEAEAPEPSAGDGAAATS
PLNEETQLFMELDVSAKVGFGEILLRSRAIEG
```

Figure 5-3: Sequence alignment of wild-type bacteriorhodopsin and WS-BR.
The sequence of wild-type bacteriorhodopsin (blue) aligned with the designed water soluble variant, WS-BR (green). The 58 mutated positions are marked with an *. The gaps in the WS-BR sequence are due to lack of electron density in the solved X-ray structure; the wild-type sequence was used in these areas for purposes of gene construction.

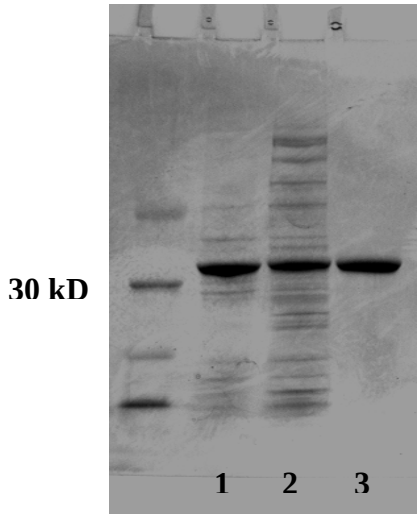


Figure 5-4: SDS-PAGE of WS-BR.
Gradient gel (4-20%) showing WS-BR at the appropriate molecular weight (~32 kD) following (1) cell lysis, (2) one column wash, and (3) column elution. The protein was estimated to be >99% pure.

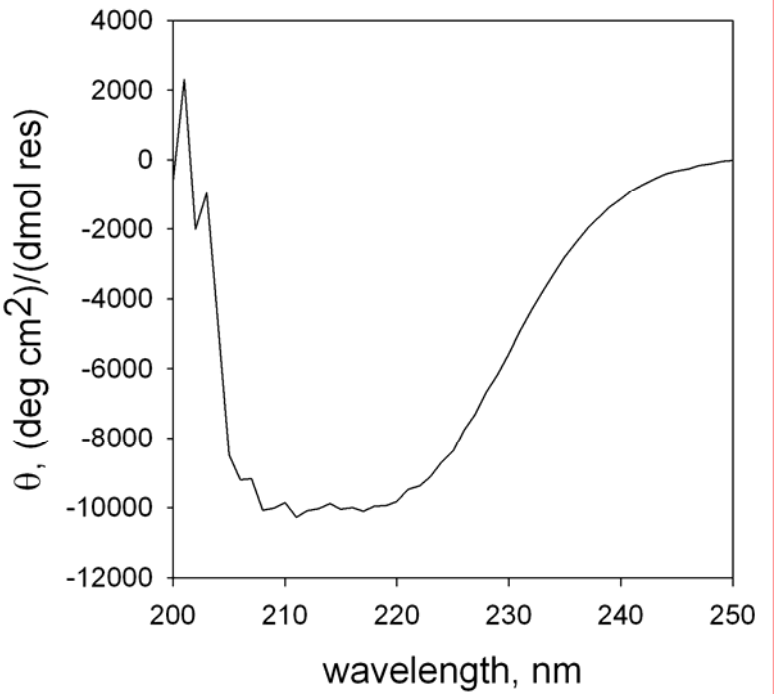


Figure 5-5: Far UV wavelength spectrum of WS-BR.
Circular dichroism wavelength scan of WS-BR. Spectrum was obtained at 25 ºC in 50 mM phosphate buffer, 1.5 M NaCl, pH 7.2. The average of three scans is shown.

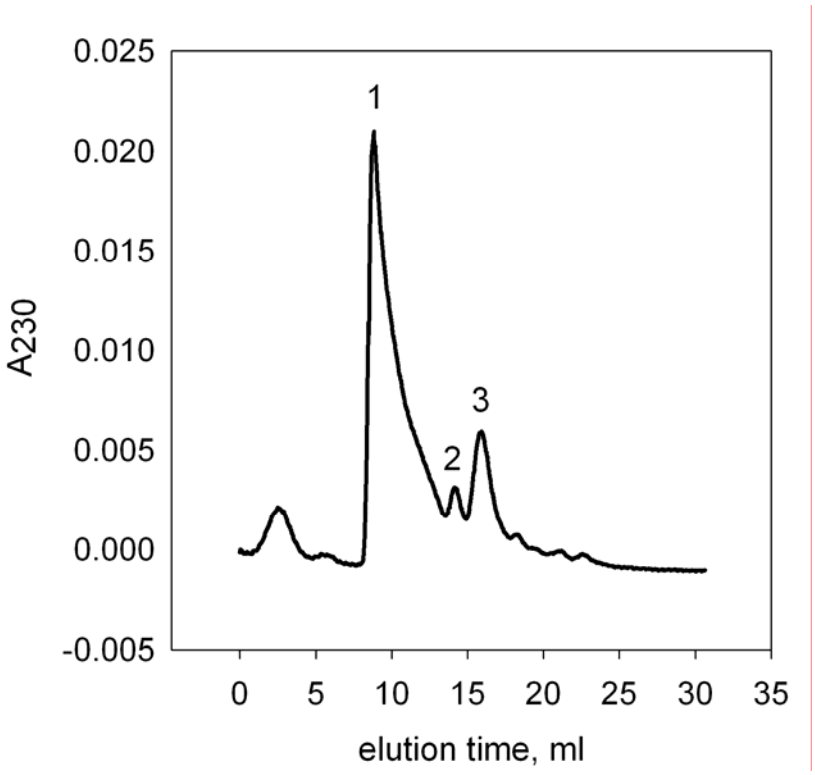


Figure 5-6: Analytical gel filtration chromatography of WS-BR.
The sample was not mono-dispersed as is evident by the various oligomeric states observed. Based on molecular weight standards (not shown), we estimated that the sample existed predominantly in high order oligomeric states (MW >200,000 D) (1). We also observed small amounts of the sample at the predicted molecular weights of a dimer (2), and monomer (3). The sample was in 50 mM phosphate buffer, 1.5 NaCl, pH 7.2.

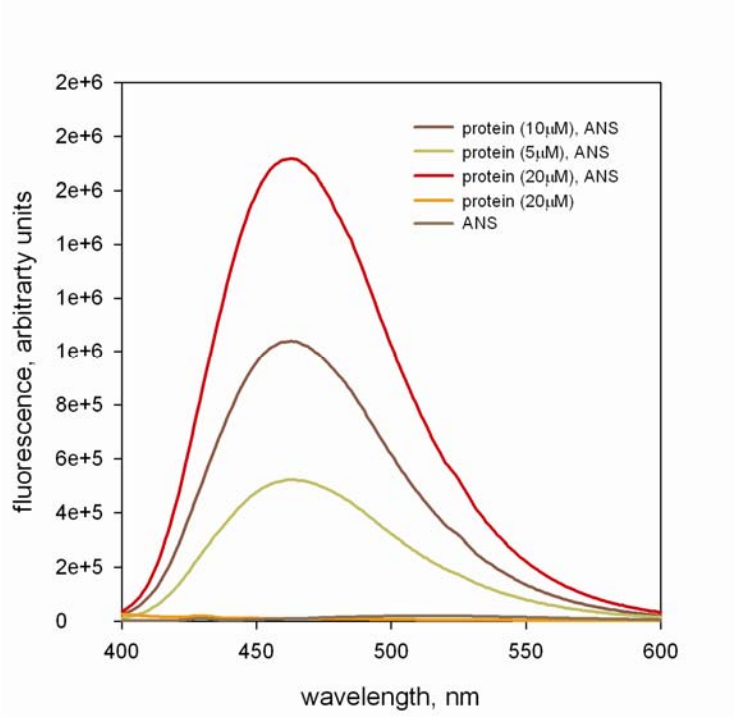


Figure 5-7: ANS binding experiment for WS-BR.
ANS (1-anilino-napthalene-8-sulfonate) was used to assess the physical state of WS-BR. ANS is a common marker for identification of protein molten globule states. Lysozyme in 25% HFA (hexafluroacetone hydrate) was used as a positive control (not shown). ANS binding increases as the protein concentration is increased.

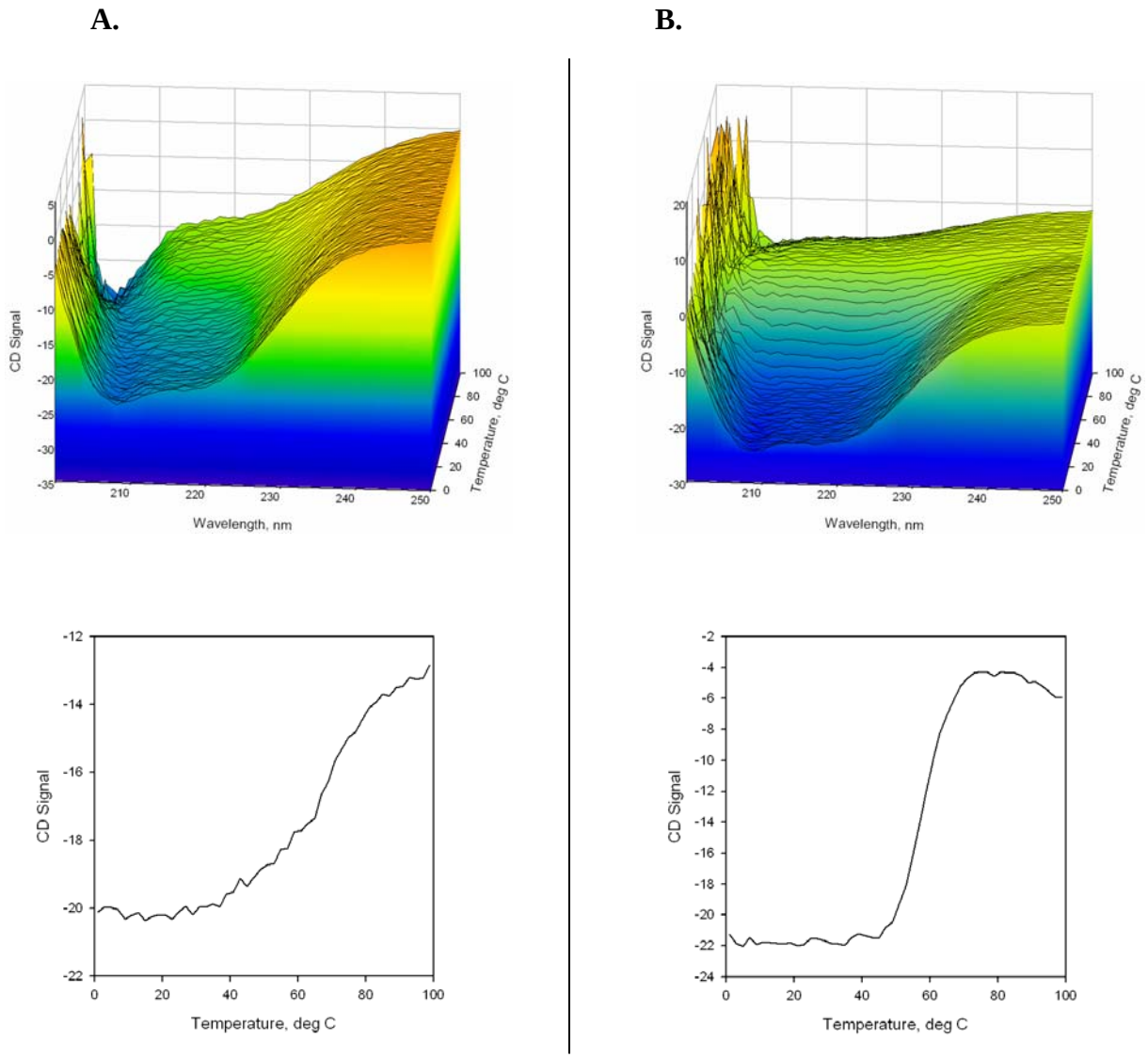


Figure 5-8: Temperature denaturations of WS-BR.
Temperature denaturations of WS-BR in 50 mM phosphate, pH 7.2 in (A) 0.1 M NaCl, and (B) 1.5 M NaCl. The top panels show far UV wavelength scans as a function of temperature. The bottom panels show one slice (222 nm) from the top panels.

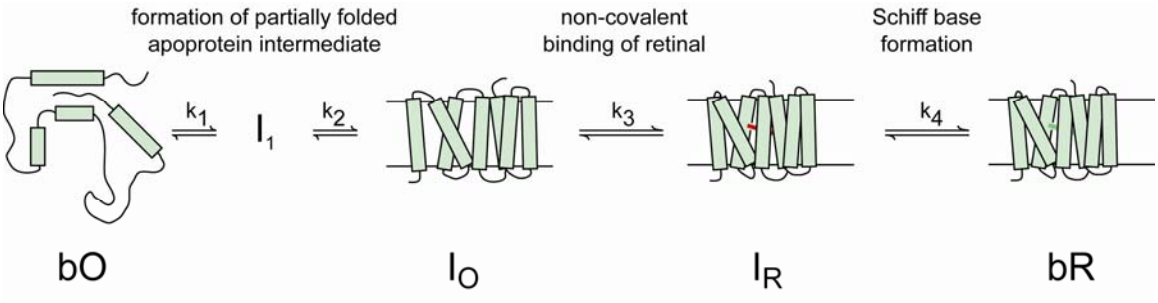


Figure 5-9: Hypothesized folding pathway of bacteriorhodopsin.
Experimental evidence suggests retinal is loosely attached following the appearance of the I_O intermediate state. Schiff base formation occurs after the I_R state to form functional bacteriorhodopsin.

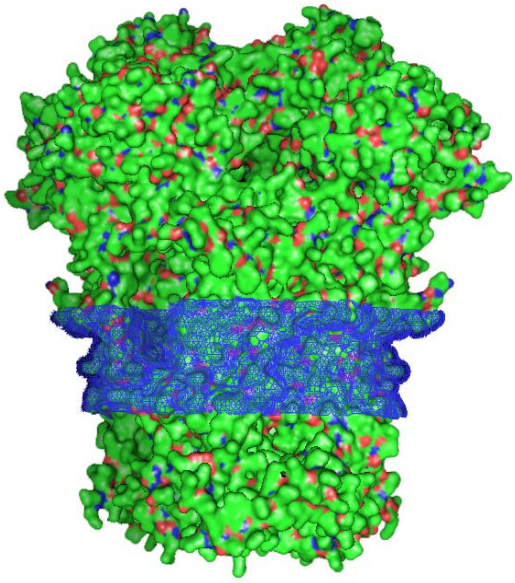


Figure 5-10: Membrane spanning region of a membrane protein.
Surface representation of the bovine cytochrome bc1 complex (PDB code: 1bgy). The blue mesh represents the membrane spanning region of the protein. Only this portion of membrane proteins was used to analyze exposed nonpolar surface areas. This region was determined by analyzing the most hydrophobic 30 Å stretch of the protein structure.

Chapter 6

NMR and Temperature Jump Measurements of *De Novo* Designed Proteins Demonstrate Rapid Folding in the Absence of Explicit Selection for Kinetics

The text of this chapter has been adapted from a published manuscript that was co-authored with Professors Stephen L. Mayo and Kevin W. Plaxco, as well as Blake Gillespie, Dung M. Vu, Shannon A. Marshall, and R. Brian Dyer.

Blake Gillespie, Dung M. Vu, Premal S. Shah, Shannon A. Marshall, R. Brian Dyer, Stephen L. Mayo, and Kevin W. Plaxco, *J. Mol. Bio*, 330, 813 (2003).

Abstract

We address the importance of natural selection in the origin and maintenance of the rapid folding of natural proteins by experimentally characterizing the folding kinetics of two *de novo* designed proteins, NC3-NCAP and ENH-FSM1. These 51-residue proteins, which adopt the helix-turn-helix homeodomain fold, share as few as 12 residues in common with their most closely related natural analog. Despite the replacement of up to 3/4 of their residues by a computer algorithm optimizing only thermodynamic properties, the designed proteins fold as fast or faster than the $35{,}000s^{-1}$ observed for this closest natural analog. Thus these *de novo* designed proteins, which were produced in the complete absence of selective pressures or design constraints explicitly aimed at ensuring rapid folding, are among the most rapidly folding proteins reported to date.

Introduction

Does natural selection play a direct role in defining and maintaining protein folding kinetics? Naturally occurring proteins fold far more rapidly than would be expected were the process a random search of conformational space,[1] suggesting at one extreme that rapid folding may be rare in the absence of explicit selective optimization. At the other extreme, it is possible that the selective pressures that ensure a stable native state inevitably produce biologically relevant folding rates.

Current theories of protein folding provide little indication of where proteins lie on the spectrum between these extremes. The nucleation-condensation model,[2] for example, suggests that selection for native state stability may be sufficient to ensure rapid folding, since the interactions that stabilize the native state also stabilize the transition state. In contrast, it is possible that stabilization of the native state may also lead to the stabilization of kinetic traps, slowing folding.[3,4] Unfortunately, simulation-based theories provide little quantitative indication of how frequently thermodynamically stable folds also exhibit the unfrustrated energy landscapes,[5] large energy gaps,[6] cooperative collapse[7] or (more generally) the lack of kinetic traps[8] that are associated with rapid kinetics. Some simulation studies have shown, however, that without explicit selective or design pressures aimed at ensuring a large energy gap, the majority of even thermodynamically stable heteropolymers fold extremely slowly (*e.g.,* ref. 6).

Experimental investigations of evolution's role in shaping folding kinetics similarly fail to resolve this issue. For example, the observed dependence of folding rates of simple proteins on native state topology[9] *may* reflect a limited role of evolution in defining folding kinetics, since kinetics are defined indirectly when selective pressures

define topology. But, in contrast, this relationship could represent a folding 'speed limit' beyond which explicit kinetic selection cannot further optimize rates. Thus the topology-dependence of folding rates in naturally-occurring proteins does not rule out the possibility that rapidly folding sequences are relatively rare among thermodynamically stable, but kinetically unoptimized, proteins. More generally, experimental studies couched in the context of naturally-occurring proteins, or close sequence analogs,[10-12] provide only limited insight into this issue since these proteins must fold rapidly in order to confer a selective advantage on the organism from which they were obtained.

The study of *de novo* designed proteins provides a means of circumventing this problem by eliminating folding kinetics as a design or selective constraint. With this motivation we describe here the refolding kinetics of a pair of *de novo* designed structural analogs of the engrailed homeodomain (En-HD) designated NC3-NCAP and ENH-FSM1. The proteins were designed using the ORBIT (optimization of rotamers by iterative techniques) suite of protein design algorithms to find compatible sequences for a target fold.[13] The ORBIT potential functions, which model the physical forces governing a protein's tertiary structure, were used in conjunction with optimization algorithms based on the dead-end elimination theorem (DEE)[14] to identify low-energy sequences. A detailed description of potential functions, parameters, and optimization algorithms is available from previous work (*e.g.,* refs. 15-17). Critical to the current study, ORBIT does not explicitly consider any aspect of folding kinetics.

Because the designed proteins are thermodynamically stable, it is reasonable to assume *a priori* that they fold rapidly relative to the Levinthal time. However, the fastest folding naturally-occurring protein folds at least a billion times more rapidly than the

slowest (*e.g.* refs 18,19), suggesting that a broad range of rates are consistent with thermodynamic stability. The question is thus, do *de novo* designed and naturally occurring proteins sharing a common topology fold with closely similar rates, or do their kinetics differ by many orders of magnitude? Addressing this question will provide insight into the relative importance of direct and indirect selective pressures in shaping folding kinetics.

Results

Moderate and distant relationships to naturally-occurring proteins

The designed proteins were built using residues 6-56 of the En-HD crystal structure as the template.[20] The design of NC3-NCAP has been described previously.[17] The resulting sequence shares 55% identity with the template molecule. ENH-FSM1 is a full sequence design of En-HD; details of the design will be published elsewhere (Shah *et al.*, manuscript in preparation). This molecule shares 25% sequence identity with the template and 37% identity with NC3-NCAP (Figure 6-1). A simple psi-BLAST search indicates that the sequence of the parent structure, En-HD from *D. melanogaster,* is the most closely related known sequence to NC3-NCAP. When sequence similarity rather than identity is considered, the homeodomain sequences of other organisms appear more closely related (67%). In contrast, a simple psi-BLAST search suggests that the fully redesigned ENH-FSM1 bears no statistically significant similarity to any known sequence when either identity or similarity are considered.

Well-folded, de novo designed proteins

Both NC3-NCAP and ENH-FSM1 fold to a stable, well-packed native state. Chemical denaturations indicate the proteins fold via a cooperative two-state process with stabilities of 2.9 and 3.4 ±0.2 kcal·mol^{-1}at 35°C, respectively (data not shown). While the structures of these proteins have not been determined, both exhibit the dispersed NMR and CD spectra characteristic of a folded protein (data not shown, see also ref. 17).

Folding kinetics of NC3-NCAP

NMR lineshape analysis indicates that NC3-NCAP folds extremely rapidly. The γ-methyl proton resonance of residue Leu11 undergoes a 550Hz chemical shift as the molecule unfolds. By monitoring the denaturant-dependent linebroadening of this resonance as a function of denaturant concentration (Figure 6-2), we have determined folding and unfolding rates across the unfolding transition (Figure 6-3). The measured folding rates decreased from 6,900s^{-1} and 2,200s^{-1} over the range 1.2 to 2.0M urea.

Laser temperature-jump (T-jump) relaxation studies confirm the rapid folding kinetics of NC3-NCAP. Unfolding was induced at urea concentrations corresponding to those in the NMR experiment by rapid T-jump from 15 to 35°C. The observed relaxation dynamics are well-fitted as a double exponential decay (see *e.g.,* Figure 6-4). The slower phase displays the significant denaturant dependence expected for an authentic unfolding phase. This phase was treated as a two-state, temperature-induced change in the equilibrium population of native and denatured states, and the refolding and unfolding rates extracted as $k_f = k_{obs}/(1+K_{UN})$ and $k_u = (k_{obs}\cdot K_{UN})/(1+K_{UN})$. K_{UN} was determined by CD-monitored equilibrium unfolding under identical conditions (data not shown). The

folding rates so derived are within experimental error of those predicted based on NMR lineshape analysis. Extrapolating the combined NMR/T-jump data to 0M urea, we estimate that NC3-NCAP folds and unfolds with rates of 29,000s^{-1} ($\ln k_f = 10.3 \pm 0.2$) and 230s^{-1} ($\ln k_u = 5.4 \pm 0.2$) in water, respectively (Figure 6-3).

The faster of the two phases observed in the T-jump experiment exhibits a rate of ~200,000s^{-1} (Figure 6-4) and is effectively independent of the denaturant concentration (data not shown). This rate is faster than any previously reported protein folding rate, but is similar to the rapid kinetics observed in T-jump studies of small helical peptides and some proteins (*e.g.,* refs. 21,22). Fersht and coworkers have reported that En-HD populates an equilibrium intermediate with native-like helical structure (A. Fersht, personal communication). The rapid phase we observe may reflect the population or thermal denaturation of a similar intermediate, or thermal relaxation of the denatured state. Because this phase is approximately an order of magnitude faster than the slower phase described above, it contributes little to the line broadening observed in the NMR unfolding experiment.

Folding kinetics of ENH-FSM1

ENH-FSM1 folds more rapidly than NC3-NCAP. As with NC3-NCAP, a highly shifted leucine methyl proton resonance in the folded state spectrum of ENH-FSM1 undergoes a 400Hz shift and line broadening as the molecule unfolds (Figure 6-2). Because lineshape-based rate determinations must be performed in the transition region, the kinetics of this molecule can only be determined at high denaturant concentrations. This necessitates long extrapolations to 0M urea. Nevertheless, analysis of the denaturant

dependence of this line broadening yields reasonably precise estimates of folding and unfolding rates in water of ~79,000s^{-1} ($\ln k_f$ = 11.3 ±0.5) and 920s^{-1} ($\ln k_u$ = 6.8 ±0.5) (Figure 6-5).

Both *de novo* designed proteins fold via poorly packed transition states. The relative solvent accessibility of the folding transition state of NC3-NCAP and ENH-FSM1, $m_f/(m_f - m_u)$, are 0.52 and 0.39, respectively. This is consistent with the observation that helical proteins generally exhibit relatively expanded transition states,[9] but inconsistent with the extremely compact transition state (0.85) reported for En-HD (A. Fersht, personal communication and ref. 23).

Comparison of kinetics and thermodynamics

An assumption of two-state folding is built into the fit of the linebroadening data. If the model is valid, however, the stabilities derived from T-jump and NMR-derived folding and unfolding rates will agree with stabilities determined from CD measurements. Kinetic and CD ΔG's for NC3-NCAP are 2.9 ±0.2 and 2.9 ±0.1 kcal-mol-1, respectively at. ΔG's for ENH-FSM1 are 2.7 ±0.4 and 3.4 ±0.1 respectively. This difference does not represent a statistically significant discrepancy at the 95% confidence interval.

Discussion

The folding rates of NC3-NCAP and ENH-FSM1 are within error of the 35,000s^{-1} reported for the analogous, naturally-occurring molecule,[23] placing them among the most rapidly folding proteins reported to date (*e.g.,* refs. 23-26). Similarly, Hill *et al.* have

reported microsecond folding kinetics for the *de novo* designed, 35-residue dimeric α_2D protein.[27] It thus appears that proteins produced in the absence of explicit design criteria aimed at ensuring rapid folding can fold as fast as the most rapidly folding natural proteins, and thus that rapid folding may be readily achieved even in the absence of direct selective pressure.

These observations are consistent with previous, albeit less direct, evidence that natural selection may play only a limited role in the determination of folding rates. For example, Baker and co-workers have used phage display selection techniques to generate 12 variant protein L sequences, 6 of which fold more rapidly than the parent sequence.[11] Similarly, two variant SH3 sequences in which approximately 50% of the residues were replaced via phage display fold as fast or slightly faster than the wild-type sequence.[10] Lastly, Serrano and co-workers have replaced up to 9 residues in the core of src-SH3 and found that the folding rates of 3 of 13 variants are accelerated up to 12-fold.[12] The ease with which folding kinetics are maintained or increased despite extensive mutations further supports the observation that folding kinetics are not the product of direct evolutionary optimization.

In contrast to these experimental observations, a large body of theoretical literature suggests that rapid folding is necessarily the product of direct selective pressure. For example, lattice polymer-based studies find that rapidly folding sequences will be produced only rarely in the absence of specific design or selective constraints. This is clearly illustrated by simulations of 27-mer lattice polymers, which indicate that only 3% to 15% of randomly selected sequences fold rapidly,[6,28,29] and the fraction of rapidly folding sequences may be yet smaller for 125-mer lattice systems.[30] This

predicted paucity of rapidly folding sequences is further supported by studies suggesting that folding kinetics are the product of extensive evolutionary optimization (*e.g.,* refs. 31,32) and that neither evolution[31,33] nor design[34-36] are likely to produce rapid folding unless kinetics are an explicit selection or design criterion. In contrast, our results suggest that if a large energy gap, low T_g or cooperative collapse are necessary to ensure rapid folding, then these properties are readily achieved even in the absence of direct selective pressure.

Why then do these proteins fold with biologically relevant rates? Several groups have studied the relationship between a protein's topology and the smoothness of its energy landscape. They have found that, while most structures are the unique ground state of only a few sequences, a small subset of structures are encoded by a large number of sequences.[37-40] Naturally-occurring protein folds are thought likely to represent such 'highly designable' structures.[37,40] Critically, these studies also suggest that highly designable structures almost invariably exhibit the smooth landscapes and unique ground states that theory associates with rapid folding.[37,39,40] The designed molecules characterized here, which fold to a naturally-occurring topology, apparently also exhibit the putatively linked properties of rapid folding and designability.

While the idea of designability predicts that proteins will fold on a biologically relevant timescale, it does not predict precisely where a protein's folding rate will fall on the billion-fold range of rates observed in nature. More specifically, it does not predict the strikingly similar folding rates of NC3-NCAP, ENH-FSM1 and En-HD. In contrast, a recent theory of folding kinetics termed the topomer search model predicts that if the folding energy landscape is smooth, folding rates will vary according to the native state

topology.[41] Our results are consistent with this suggestion, and show that the precise folding kinetics of designable proteins are determined indirectly by the selective pressures that define structure and thermodynamics.

Methods

Experimental Details

The designed proteins were expressed in *E. coli* and purified by HPLC as described.[17] NC3-NCAP was characterized in ^2H_2O, 50mM potassium phosphate buffered at pD 4.5. ENH-FSM1 was characterized in the same buffer at pD 5.5. In comparison, Mayor, *et al.* determined the folding kinetics of En-HD at pH 5.8.[23] Urea and guanidine HCl stocks were high purity grade (Pierce, USB), and were deuterated by dissolving in ^2H_2O and lyophilizing three times.

Equilibrium Thermodynamics

Unfolding thermodynamics were determined via circular dichroism (CD) at 222nm on an AVIV 202 Spectrometer (AVIV Instruments, Lakewood, NJ). Chemical denaturations with guanidine HCl were conducted using a Hamilton microlab 500 automatic titrator (Hamilton Company, Reno, NV) coupled to the spectrometer. Protein concentration was 5μM, and samples were equilibrated for 120s at each denaturant concentration. Data were fitted to a two-state model for unfolding, and thermodynamic parameters were determined as described elsewhere.[42]

^{1}H-NMR Data Collection and Analysis

1D-^{1}H NMR spectra were collected at 35°C and analyzed on a 500MHz Bruker. Lyophilized protein was dissolved at 1mM in $^{2}H_2O$ or ^{2}H urea solutions, buffered as described. 1,4-dioxane was employed as a temperature- and denaturant concentration-independent chemical shift standard. Individual spectra were 4096 points, and consisted of 1024 scans.

Both proteins display a significantly ring-current shifted resonance suitable for lineshape analysis of refolding kinetics. Specific labelling of NC3-NCAP demonstrated that this resonance is a γ-proton of Leu11; the resonance in ENH-FSM1 has not been conclusively determined, but model structures suggest it is also Leu11 (data not shown). These resonances were fit to a model of two-site chemical exchange to determine the molecules' denaturant-dependent refolding and unfolding rates (k_f and k_u) at each urea concentration.[24]

The determination of rates from lineshapes requires knowledge of the chemical shifts and linewidths of the folded and unfolded state resonances. Due to spectral overlap in the unfolded state, the random coil chemical shift and linewidth values for leucine γ-proton were adopted from the literature.[43] Given these four constants, the only fitted parameters in the analysis are the rates k_f and k_u.

T-jump measurements

The relaxation dynamics of NC3-NCAP were monitored by Trp fluorescence following a laser-induced T-jump. A pump pulse corresponding to the peak of a weak $^{2}H_2O$ near-IR absorption band ($\varepsilon = 10.1 cm^{-1}$ at 2μm) was used to maximize transmission

through the 100 μm path length cells, ensuring a nearly uniform temperature profile. Under these conditions, the diffusion of heat out of the interaction volume occurs in ~20ms. To avoid complications arising from sample cooling, data were analyzed over the range 1-200μs. All data were well-fitted as a double exponential relaxation process.

The T-jump spectrometer has been described previously.[44] The magnitude of the jump was calibrated (±1°C) by comparing the pump pulse-induced fluorescence change in a Trp sample with the equilibrium temperature-dependence of Trp fluorescence. 10,000 scans were collected and averaged for each sample. Samples were prepared as for the NMR experiments except that the protein concentrations were 80μM.

References

1. Levinthal, C. (1969). How to fold graciously. *Proceedings: Mossbauer Spectroscopy in Biological Systems, University of Illinois Bulletin* **67**, 22-24.
2. Fersht, A. R. (1995). Characterizing transition states in protein folding: an essential step in the puzzle. *Curr. Opin. Struct. Biol.* **5**, 79-84.
3. Kaya, H. & Chan, H. S. (2002). Towards a consistent Mmdeling of protein thermodynamic and kinetic cooperativity: how applicable is the transition state picture to folding and unfolding? *J. Mol. Biol.* **315**, 899-909.
4. Silow, M. & Oliveberg, M. (2003). High concentrations of viscogens decrease the protein folding rate constant by prematurely collapsing the coil. *J. Mol. Biol.* **326**, 263-271.
5. Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* **84**, 7524-8.
6. Sali, A., Shakhnovich, E. & Karplus, M. (1994). How does a protein fold? *Nature (London)* **369**, 248-251.
7. Klimov, D. K. & Thirumalai, D. (1996). Criterion the determines the foldability of proteins. *Phys. Rev. Lett.* **76**, 4070-4073.
8. Plotkin, S. S. & Wolynes, P. G. (2003). Buffed energy landscapes: Another solution to the kinetic paradoxes of protein folding. *Proc. Natl. Acad. Sci. USA* **100**, 4417-4422.
9. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985-94.
10. Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. & Baker, D. (1997). Functional rapidly folding proteins from simplified amino acid sequences. *Nat. Struct. Biol.* **4**, 805-809.
11. Kim, D. E., Gu, H. & Baker, D. (1998). The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl. Acad. Sci. USA* **95**, 4982-6.
12. Ventura, S., Vega, M. C., Lacroix, E., Angrand, I., Spagnolo, L. & Serrano, L. (2002). Conformation straion in the hydrophobic core and its implications for protein folding and design. *Nat. Struct. Biol.* **9**, 485-493.
13. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci.* **5**, 895-903.
14. Desmet, J., De Maeyer, M., Hazes, B. & Lasters, I. (1992). The Dead-end elimination theorem and its use in protein side-chain positioning. *Nature (London)* **356**, 539-542.

15. Dahiyat, B. I., Sarisky, C. A. & Mayo, S. L. (1997). *De novo* protein design: towards fully automated sequence selection. *J. Mol. Biol.* **273**, 789-96.

16. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.* **9**, 509-513.

17. Marshall, S. A., Morgan, C. S. & Mayo, S. L. (2002). Electrostatics significantly affect the stability of designed homeodomain variants. *J. Mol. Biol.* **316**, 189-199.

18. Wittung-Stafshede, P., Lee, J. C., Winkler, J. R. & Gray, H. B. (1999). Cytochrome b562 folding triggered by electron transfer: Approaching the speed limit for formation of a four-helix-bundle protein. *Proc. Natl. Acad. Sci. USA* **96**, 6587-6590.

19. Kern, G., Kern, D., Schmid, F. X. & Fischer, G. (1995). A kinetic analysis of the folding of human carbonic anhydrase II and its catalysis by cyclophilin. *J. Biol. Chem.* **270**, 740-745.

20. Clarke, N. D., Kissinger, C. R., Desjarlais, J., Gilliland, G. L. & Pabo, C. O. (1994). Structural studies of the engrailed homeodomain. *Protein Sci.* **3**, 1779-1787.

21. Williams, S., Causgrove, T. M., Gilmanshin, R., Fang, K. S., Callendar, R. H., Woodruff, W. H. & Dyer, R. B. (1996). Fast events in protein folding: helix melting and formation in a small peptide. *Biochemistry* **35**, 691-97.

22. Gilmanshin, R., Williams, S., Callender, R. H., Woodruff, W. H. & Dyer, R. B. (1997). Fast events in protein folding: relaxation dynamics of secondary and tertiary structure in native apomyoglobin. *Proc. Natl. Acad. Sci. USA* **94**, 3709-3713.

23. Mayor, U., Johnson, C. M., Daggett, V. & Fersht, A. R. (2000). Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl. Acad. Sci. USA* **97**, 13518-13522.

24. Burton, R. E., Huang, G. S., Daugherty, M. A., Fullbright, P. W. & Oas, T. G. (1996). Microsecond protein folding through a compact transition state. *J. Mol. Biol.* **263**, 311-22.

25. Spector, S. & Raleigh, D. P. (1999). Submillisecond folding of the peripheral subunit-binding domain. *J. Mol. Biol.* **293**, 763-768.

26. Myers, J. K., Pace, C. N. & Scholtz, J. M. (1995). Denaturant *m*-values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding [published erratum appears in *Protein Sci*. (1996) **5,** 981]. *Protein Sci.* **4**, 2138-48.

27. Hill, R. B., Bracken, C., DeGrado, W. F. & Palmer, A. G. I. (2000). Molecular motions and protein folding: characterization of the backbone dynamics and folding of α_2D using ^{13}C NMR Spin Relaxation. *J. Am. Chem. Soc.* **122**, 11610-11619.

28. Shakhnovich, E., Farztdinov, G., Gutin, A. M. & Karplus, M. (1991). Protein folding bottlenecks: a lattice monte carlo simulation. *Phys. Rev. Lett.* **67**, 1665-1667.

29. Sali, A., Shakhnovich, E. & Karplus, M. (1994). Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614-36.

30. Dinner, A. R., So, S.-S. & Karplus, M. (1998). Use of quantitative structure-property relationships to predict the folding ability of model proteins. *Proteins* **33**, 177-203.

31. Mirny, L. A., Abkevich, V. I. & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proc. Natl. Acad. Sci. USA* **95**, 4976-4981.

32. Li, L., Mirny, L. A. & Shakhnovich, E. I. (2000). Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. *Nat. Struct. Biol.* **7**, 336-342.

33. Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1998). A protein engineering analysis of the transition state for protein folding: simulation in the lattice model. *Fold. Des.* **3**, 183-194.

34. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1996). Improved design of stable and fast-folding model proteins. *Fold. Des.* **1**, 221-30.

35. Guo, Z. & Thirumalai, D. (1996). Kinetics and thermodynamics of folding of a *de novo* designed four-helix bundle protein. *J. Mol. Biol.* **263**, 323-43.

36. Betancourt, M. & Thirumalail, D. (2002). Protein sequence design by energy landscaping. *J. Phys. Chem. B* **106**, 599-609.

37. Govindarajan, S. & Goldstein, R. A. (1995). Searching for foldable protein structures using optimized energy functions. *Biopolymers* **36**, 43-51.

38. Li, H., Helling, R., Tang, C. & Wingreen, N. (1996). Emergence of preferred structures in a simple model of protein folding. *Science* **273**, 666-669.

39. Buchler, N. E. G. & Goldstein, R. A. (2000). Surveying determinants of protein structure designability across different energy models and amino-acid alphabets: a consensus. *J. Chem. Phys.* **112**, 2533-47.

40. Miller, J., Zeng, C., Wingreen, N. S. & Tang, C. (2002). Emergence of highly designable protein-backbone conformations in an off-lattice model. *Proteins* **47**, 506-512.

41. Makarov, D. E. & Plaxco, K. W. (2003). The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.* **12**, 17-26.

42. Pace, C. N. (1990). Conformational stability of globular proteins. *Trends Biochem. Sci.* **15**(1), 14-7.

43. Plaxco, K. W., Morton, C. J., Grimshaw, S. B., Jones, J. A., Pitkeathly, M., Campbell, I. D. & Dobson, C. M. (1997). The effects of guanidine hydrochloride

on the 'random coil' conformations and NMR chemical shifts of the peptide series GGXGG. *J. Biomol. NMR* **10**, 221-230.

44. Gulotta, M., Gilmanshin, R., Buscher, T. C., Callender, R. H. & Dyer, R. B. (2001). Core formation in apomyoglobin: probing the upper reaches of the folding energy landscape. *Biochemistry* **40**, 5137-5143.

Figure 6-1: Sequence alignment of designed variants and wild type.
Sequence alignment of the template, En-HD, and the *de novo* designed proteins NC3-NCAP and ENH-FSM1. Identity with the template is highlighted in gray. In the design of NC3-NCAP surface residues were varied, helix-capping and helix dipole propensities were optimized, and the identity of core residues was held fixed.[17] In contrast, all positions were allowed to vary in the design of ENH-FSM1. These sequences share 55 and 25% identity with En-HD, respectively, and 37% with each other. Only 10 residues are common among the three proteins.

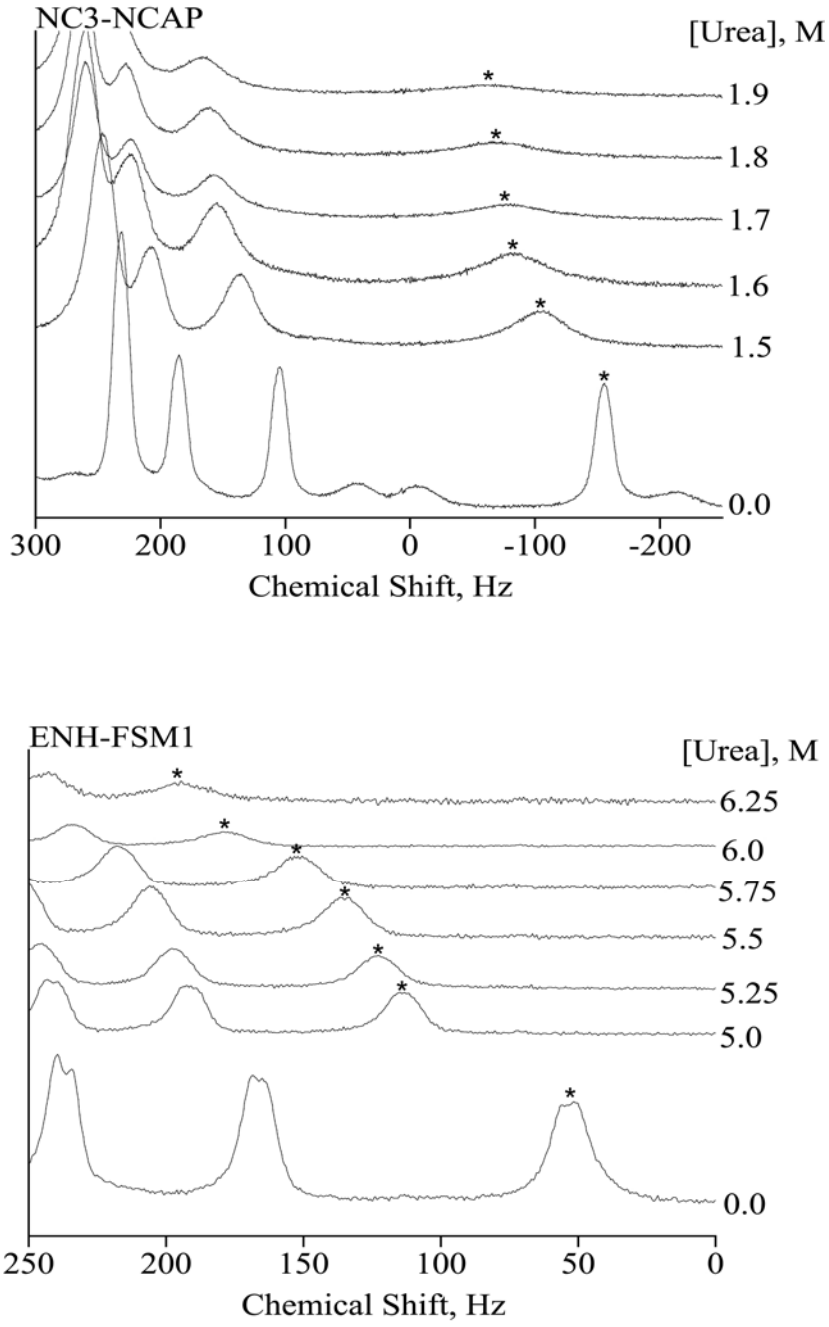


Figure 6-2: NMR data for designed variants.
Stacked plots of ^{1}H-NMR data for NC3-NCAP and ENH-FSM1 in the absence of denaturant and through their transition regions. As the molecules begin to unfold, ring shifted methyl resonances in both proteins shift and broaden until spectral overlap precludes further analysis.

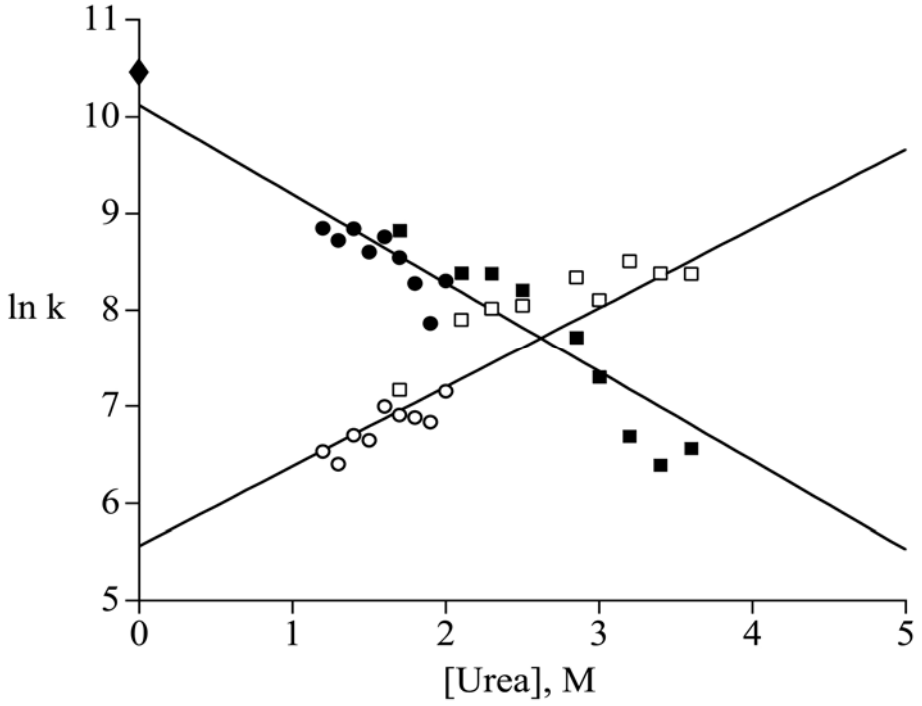


Figure 6-3: Folding and unfolding rates of NC3-NCAP.
Denaturant dependence of folding and unfolding rates for NC3-NCAP. Rates determined by both NMR lineshape (circles) and temperature-jump relaxation (squares), are effectively indistinguishable. The estimated folding and unfolding rates in water, determined by extrapolation to 0M urea, are 29,000s^{-1} and 230s^{-1}, respectively. The single diamond indicates the folding rate of the template molecule under similar conditions.[23]

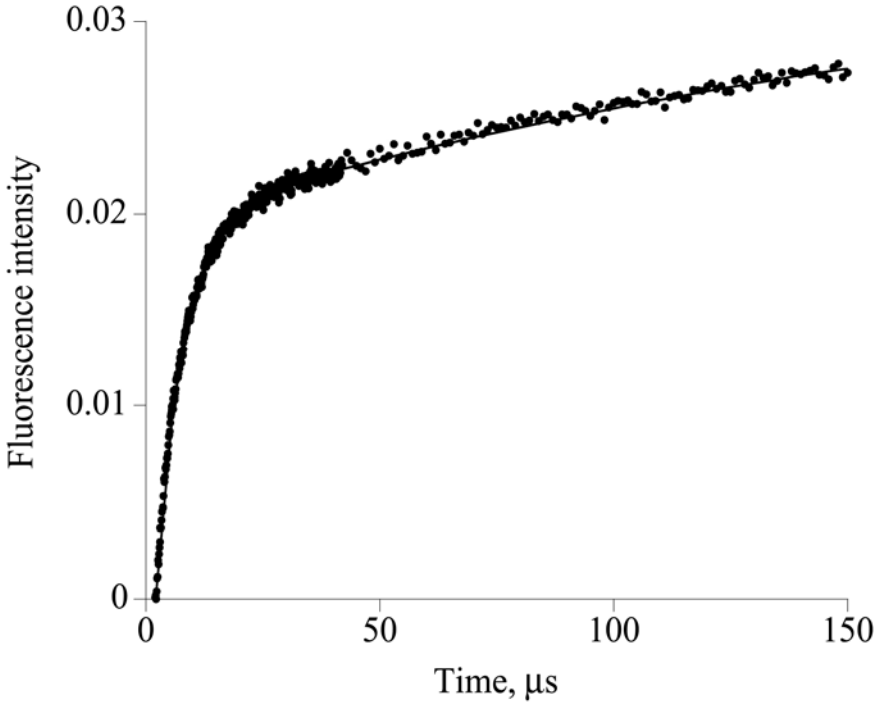


Figure 6-4: T-Jump experiments.
Temperature-jump relaxation data collected at 2.5M urea. The data are well-fitted to a double exponential decay. The faster relaxation (~200,000s^{-1}) exhibits little denaturant dependence and may represent either the thermal equilibration of the unfolded state after the T-jump or the population of a folding intermediate.[23] The folding and unfolding rates derived from the slower relaxation (4800s^{-1}) correspond well to the rates measured by NMR lineshape analysis (see Figure 2).

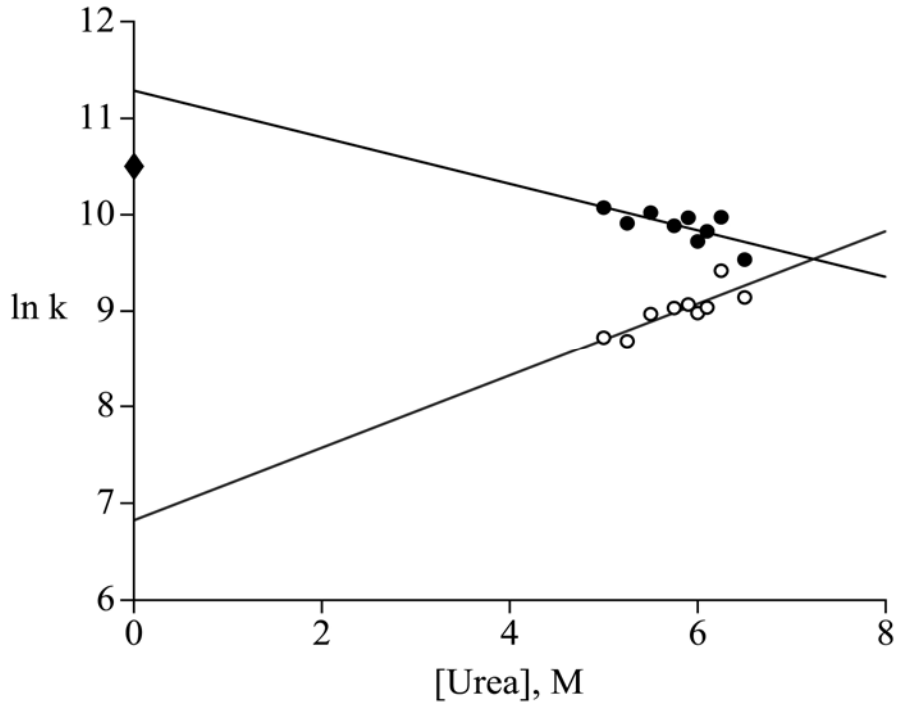


Figure 6-5: Folding and unfolding rates of ENH-FSM1.
Denaturant dependent folding kinetics of ENH-FSM1. NMR lineshape analysis predicts folding and unfolding rates in water of 79,000s^{-1} and 920s^{-1}, respectively. Because lineshape-based rate determinations must be performed in the transition region, the kinetics of this molecule can only be determined at high denaturant concentrations. This necessitates the long extrapolations to 0M urea. The single diamond indicates the folding rate of the template molecule under similar conditions.[23]

Appendix A

Baseline Correction Energies Provide More Natural Surface Amino Acid Compositions for ORBIT Designs

While performing surface designs of membrane proteins (Chapter 5), we observed that the distribution of allowable amino acids was often biased towards the longer amino acids that provide more favorable van der Waals contacts. Amino acids such as Lys, Arg, and Glu were selected at significantly higher frequencies than what was observed in nature's proteins. Indeed, surface designs of 30 structures that were performed using ORBIT revealed that amino acid compositions of designed surfaces were significantly different from surface compositions observed in nature (Figure A-1).

Typical surface designs using ORBIT limit the selection of amino acids to Asn, Asp, Gln, Glu, His, Lys, Ser, Thr, Ala, Arg. We therefore only analyzed this set of amino acids when performing the survey. Based on our observations (Figure A-1), we either penalized or benefited an amino acid (and all rotamers therein) to obtain designs producing a more wild-type-like composition. The energies were applied to a rotamer's rotamer-template energy and are listed in Table A-1.

After applying these energies, we performed designs on the same 30 structures above and compared the surface compositions to WT surface compositions (Figure A-2). We find a much higher correlation also (Figure 3, $r^2 = 0.85$) when we include baseline correction energies in our designs.

Without the use of baseline correction energies, longer amino acids are preferred because of the favorable van der Waals contacts that are provided with increased number of atoms. So in essence, baseline correction energies are simply correcting for this observed effect. Perhaps a more accurate way of addressing this issue would be to normalize van der Waals energies based on the number of atoms in an amino acid. Professor Mayo has implemented this into ORBIT; however it has not been fully tested.

Table A-1: Baseline correction energies

Amino Acid	Energy (kcal/mol)
Ala	-4.000
Asp	-1.687
Glu	-0.928
His	-0.120
Hsp[a]	-0.120
Lys	-1.460
Asn	-1.654
Gln	-0.909
Arg	0.500
Ser	-3.500
Thr	-3.000

[a] Protonated form of histidine

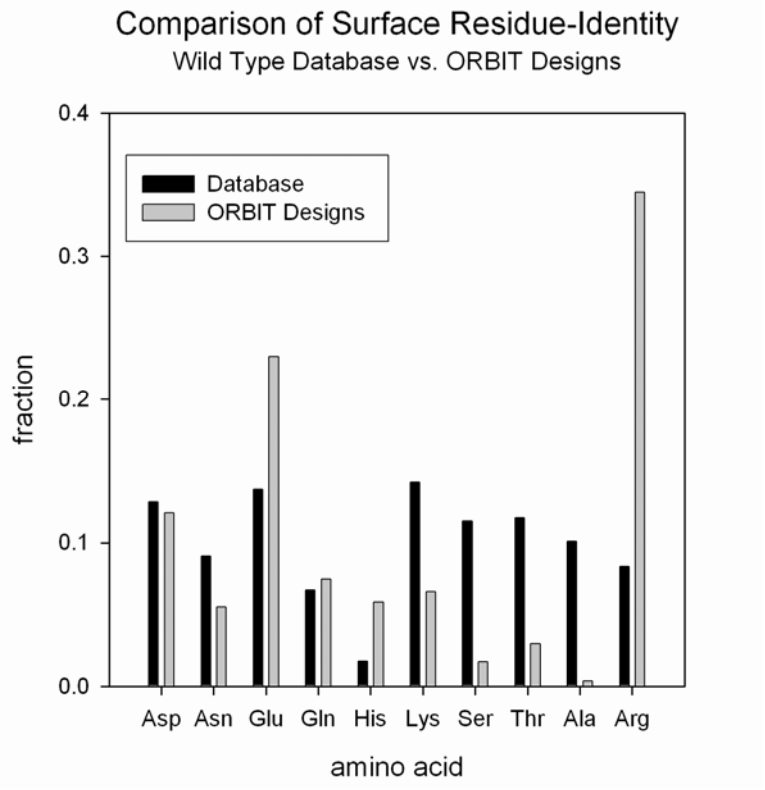


Figure A-1: Comparison of surface amino acid compositions.
Surface designs of 30 structures using standard ORBIT parameters revealed that there is a bias towards the selection of longer sidechains that provide favorable van der Waals contacts. This is significantly different from wild-type surfaces.

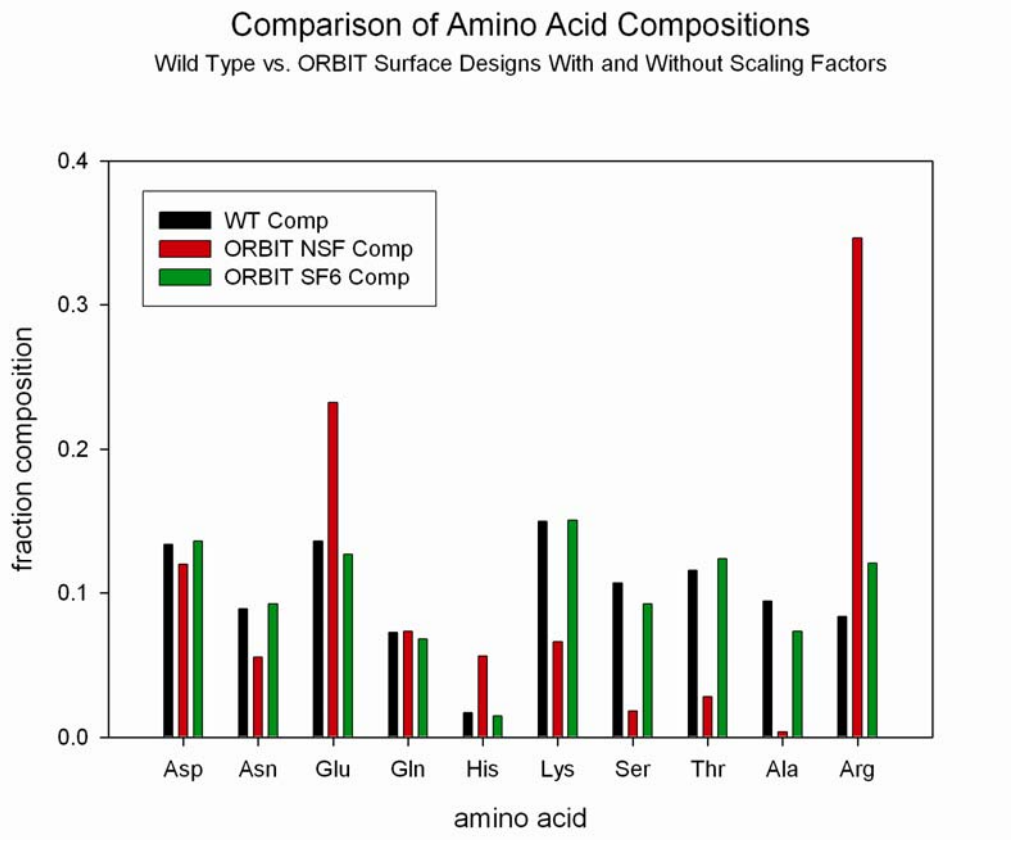


Figure A-2: Comparison of surface compositions with and without baseline correction energies.
Surface designs of 30 structures were performed with and without baseline correction energies and compared to wild-type compositions. Use of baseline correction energies provides compositions that resemble wild-type surfaces. Baseline correction energies were applied to a rotamer's rotamer-template energy.

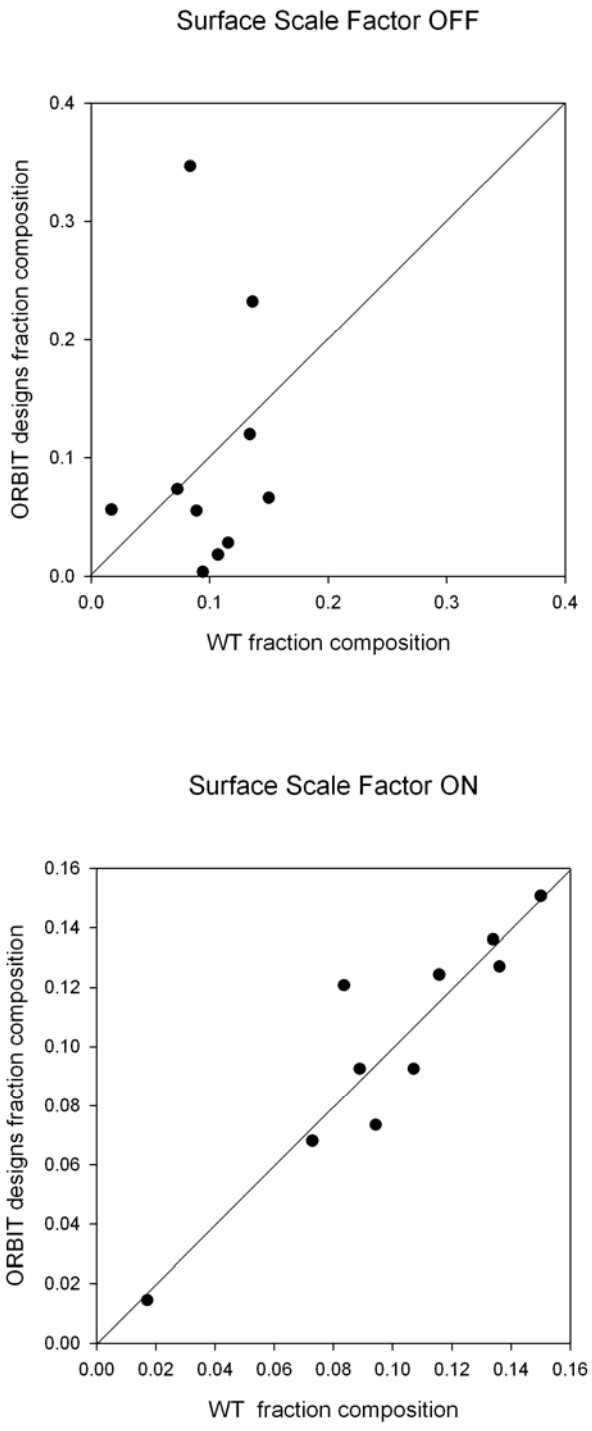


Figure A-3: Correlations between ORBIT surface designs and wild-type surfaces. Once baseline correction energies are applied to our set of 30 structures, a much higher correlation is observed between ORBIT designed surfaces and wild-type surfaces.

Appendix B

Chemical Shifts for FSM1_VF

Table B-1: Chemical shifts for FSM1_VF NMR structure sorted by residue.

Residue Number: Type	Atom Type	Chemical Shift *ppm*
1: MET	C_α	54.783
	H_α	4.070
	C_β	32.800
	H_β	2.090
	Cγ	30.730
	H_γ	2.520
	C_ε	16.720
	H_ε	2.046
	C	172.249
2: LYS	N	124.454
	HN	8.690
	C_α	56.162
	H_α	4.069
	C_β	32.845
	$H_\beta2$	1.330
	$H_\beta1$	1.480
	C_γ	24.624
	H_γ	1.210
	C_δ	29.070
	H_δ	1.550
	C_ε	41.964
	H_ε	2.887
	C	176.083
3: GLN	N	121.385
	HN	8.034
	C_α	55.763
	H_α	4.193
	C_β	29.189
	$H_\beta2$	1.750
	$H_\beta1$	2.078
	C_γ	33.909
	H_γ	2.212
	$N_\varepsilon2$	111.767
	$H_\varepsilon21$	7.409
	$H_\varepsilon22$	6.662
	C	174.705
4: TRP	N	118.965
	HN	7.499
	C_α	56.236
	H_α	4.774
	C_β	30.035
	$H_\beta2$	3.125
	$H_\beta1$	3.361
	$C_\delta1$	127.221
	$H_\delta1$	7.279
	$N_\varepsilon1$	130.794
	$H_\varepsilon1$	10.686
	$C_\zeta2$	114.643
	$H_\zeta2$	7.419
	$C_\eta2$	123.940
	$H_\eta2$	7.033
	$C_\zeta3$	121.894
	$H_\zeta3$	7.080
	$C_\varepsilon3$	120.447
	$H_\varepsilon3$	7.453
	C	175.314
5: SER	N	116.156
	HN	8.318
	C_α	57.731
	H_α	4.401
	C_β	64.614
	$H_\beta2$	3.918
	$H_\beta1$	4.157
	C	174.822
6: GLU	N	122.252
	HN	8.808
	Ca	58.685
	Ha	4.171
	Cb	29.381
	Hb	2.049
	Cg	36.250
	Hg	2.358
	C	177.898

Residue Number: Type	Atom Type	Chemical Shift *ppm*
7: ASN	N	117.705
	HN	8.446
	C_α	55.508
	H_α	4.506
	C_β	38.479
	$H_\beta2$	2.751
	$H_\beta1$	2.814
	$N_\delta2$	113.057
	$H_\delta21$	7.677
	$H_\delta22$	7.009
	C	177.119
8: VAL	N	120.856
	HN	7.725
	C_α	66.099
	H_α	3.518
	C_β	31.467
	H_β	2.095
	$C_\gamma2$	23.077
	$H_\gamma2$	0.948
	$C_\gamma1$	21.497
	$H_\gamma1$	0.638
	C	176.875
9: GLU	N	120.715
	HN	8.267
	C_α	60.133
	H_α	3.632
	C_β	29.202
	H_β	2.260
	C_γ	36.618
	H_γ	2.296
	C	178.193
10: GLU	N	117.601
	HN	8.112
	Ca	59.136
	Ha	3.998
	Cb	29.148
	Hb	2.121
	Cg	35.911
	Hg	2.390
	C	179.299

Residue Number: Type	Atom Type	Chemical Shift *ppm*
11: LYS	N	119.484
	HN	7.745
	C_α	59.084
	H_α	4.087
	C_β	32.279
	$H_\beta2$	1.910
	$H_\beta1$	1.940
	C_γ	25.476
	$H_\gamma2$	1.490
	$H_\gamma1$	1.609
	C_δ	29.204
	H_δ	1.550
	C_ε	42.189
	H_ε	2.945
	C	180.140
12: LEU	N	121.629
	HN	8.266
	C_α	57.927
	H_α	4.003
	C_β	41.146
	$H_\beta2$	1.016
	$H_\beta1$	1.383
	C_γ	26.400
	H_γ	1.270
	$C_\delta1$	23.407
	$H_\delta1$	0.253
	$C_\delta2$	24.445
	$H_\delta2$	-0.041
	C	178.395
13: LYS	N	118.621
	HN	8.287
	C_α	60.458
	H_α	3.873
	C_β	32.396
	$H_\beta2$	1.847
	$H_\beta1$	1.963
	C_γ	26.296
	$H_\gamma2$	1.370
	$H_\gamma1$	1.740
	C_δ	29.839
	H_δ	1.740
	C_ε	42.060
	H_ε	2.935
	C	179.394

Residue Number: Type	Atom Type	Chemical Shift *ppm*
14: GLU	N	119.195
	HN	8.026
	C_α	58.877
	H_α	4.028
	C_β	29.272
	H_β	2.110
	C_γ	35.859
	H_γ	2.337
	C	177.980
15: PHE	N	120.395
	HN	7.992
	C_α	61.821
	H_α	4.103
	C_β	39.522
	$H_\beta2$	3.075
	$H_\beta1$	3.106
	$C_\delta1$	131.295
	$H_\delta1$	6.644
	$C_\varepsilon1$	130.486
	$H_\varepsilon1$	6.681
	C_ζ	128.847
	H_ζ	6.632
	$C_\varepsilon2$	130.486
	$H_\varepsilon2$	6.681
	$C_\delta2$	131.295
	$H_\delta2$	6.644
	C	178.967
16: VAL	N	118.358
	HN	8.217
	C_α	65.813
	H_α	3.563
	C_β	31.660
	H_β	2.134
	$C_\gamma2$	22.886
	$H_\gamma2$	1.038
	$C_\gamma1$	21.554
	$H_\gamma1$	0.902
	C	177.758

Residue Number: Type	Atom Type	Chemical Shift *ppm*
17: LYS	N	119.329
	HN	7.727
	C_α	58.726
	H_α	3.995
	C_β	32.484
	H_β	1.894
	C_γ	25.282
	$H_\gamma2$	1.423
	$H_\gamma1$	1.554
	C_δ	29.513
	H_δ	1.670
	C_ε	42.137
	H_ε	2.940
	C	178.125
18: ARG	N	114.952
	HN	7.437
	C_α	56.205
	H_α	4.113
	C_β	30.574
	$H_\beta2$	1.510
	$H_\beta1$	1.610
	C_γ	27.467
	$H_\gamma2$	1.510
	$H_\gamma1$	1.610
	C_δ	43.193
	H_δ	2.040
	N_ε	84.988
	H_ε	7.461
	C	175.673
19: HIS	N	118.364
	HN	7.558
	C_α	55.012
	H_α	4.557
	C_β	28.431
	$H_\beta2$	2.530
	$H_\beta1$	2.945
	$C_\delta2$	119.809
	$H_\delta2$	6.872
	$C_\varepsilon1$	135.262
	$H_\varepsilon1$	8.169
	C	173.395

Residue Number: Type	Atom Type	Chemical Shift *ppm*
20: GLN	N	119.849
	HN	8.357
	C_α	57.603
	H_α	4.182
	C_β	29.202
	$H_\beta 2$	1.954
	$H_\beta 1$	2.065
	C_γ	34.219
	H_γ	2.352
	$N_\varepsilon 2$	112.964
	$H_\varepsilon 21$	7.559
	$H_\varepsilon 22$	6.849
	C	176.511
21: ARG	N	118.830
	HN	8.440
	C_α	55.332
	H_α	4.407
	C_β	29.700
	$H_\beta 2$	1.722
	$H_\beta 1$	1.879
	C_γ	27.222
	$H_\gamma 2$	1.530
	$H_\gamma 1$	1.560
	C_δ	42.249
	H_δ	3.178
	N_ε	84.645
	H_ε	7.236
	C	175.494
22: ILE	N	122.192
	HN	7.933
	Ca	60.398
	Ha	4.365
	Cb	39.709
	Hb	1.855
	Cg1	28.041
	Hg12	1.220
	Hg11	1.508
	Cd1	14.918
	Hd1	0.679
	Cg2	17.625
	Hg2	0.843
	C	175.243
23: THR	N	116.544
	HN	8.315
	C_α	60.311
	H_α	4.594
	C_β	71.371
	H_β	4.705
	$C_\gamma 2$	21.799
	Hγ21	1.304
	C	175.346
24: GLN	N	121.425
	HN	9.026
	C_α	59.327
	H_α	3.822
	C_β	27.928
	H_β	1.977
	C_γ	33.887
	H_γ	2.233
	$N_\varepsilon 2$	112.231
	$H_\varepsilon 21$	7.414
	$H_\varepsilon 22$	6.816
	C	178.403
25: GLU	N	119.131
	HN	8.779
	C_α	59.886
	H_α	4.071
	C_β	28.885
	$H_\beta 2$	1.929
	$H_\beta 1$	2.060
	C_γ	36.372
	$H_\gamma 2$	2.232
	$H_\gamma 1$	2.349
	C	179.068
26: GLU	N	120.177
	HN	7.757
	C_α	59.105
	H_α	4.011
	C_β	30.068
	$H_\beta 2$	1.927
	$H_\beta 1$	2.286
	C_γ	36.873
	H_γ	2.288
	C	179.745

Residue Number: Type	Atom Type	Chemical Shift *ppm*
27: LEU	N	121.600
	HN	8.218
	C_α	58.090
	H_α	4.002
	C_β	41.532
	$H_\beta2$	1.620
	$H_\beta1$	1.840
	C_γ	27.464
	H_γ	1.600
	$C_\delta1$	25.242
	$H_\delta11$	0.798
	$C_\delta2$	24.365
	$H_\delta21$	0.870
	C	177.931
28: HIS	N	117.889
	HN	8.702
	C_α	59.525
	H_α	4.081
	C_β	28.448
	H_β	3.400
	$C_\delta2$	119.747
	$H_\delta2$	7.171
	$C_\varepsilon1$	136.269
	$H_\varepsilon1$	8.502
	C	176.898
29: GLN	N	117.611
	HN	8.144
	C_α	58.731
	H_α	4.015
	C_β	28.341
	$H_\beta2$	2.096
	$H_\beta1$	2.177
	C_γ	34.082
	$H_\gamma2$	2.364
	$H_\gamma1$	2.546
	$N_\varepsilon2$	111.984
	$H_\varepsilon21$	7.599
	$H_\varepsilon22$	6.846
	C	178.391

Residue Number: Type	Atom Type	Chemical Shift *ppm*
30: TYR	N	120.981
	HN	7.936
	C_α	61.150
	H_α	3.992
	C_β	38.227
	$H_\beta2$	2.432
	$H_\beta1$	2.720
	$C_\delta1$	132.028
	$H_\delta1$	6.447
	$C_\varepsilon1$	117.657
	$H_\varepsilon1$	6.356
	$C_\varepsilon2$	117.657
	$H_\varepsilon2$	6.356
	$C_\delta2$	132.028
	$H_\delta2$	6.447
	C	176.952
31: ALA	N	119.500
	HN	8.291
	C_α	55.110
	H_α	3.521
	C_β	18.037
	H_β	1.268
	C	179.382
32: GLN	N	116.487
	HN	8.001
	C_α	58.744
	H_α	4.030
	C_β	28.502
	H_β	2.004
	C_γ	34.183
	$H_\gamma2$	2.201
	$H_\gamma1$	2.340
	$N_\varepsilon2$	111.803
	$H_\varepsilon21$	7.486
	$H_\varepsilon22$	6.789
	C	180.078

Residue Number: Type	Atom Type	Chemical Shift *ppm*
33: ARG	N	121.286
	HN	7.822
	C_α	58.733
	H_α	3.973
	C_β	29.583
	H_β	1.824
	C_γ	27.596
	$H_\gamma2$	1.564
	$H_\gamma1$	1.730
	C_δ	43.596
	H_δ	3.105
	N_ε	84.347
	H_ε	7.234
	C	178.198
34: LEU	N	117.225
	HN	7.402
	C_α	54.828
	H_α	4.076
	C_β	42.063
	$H_\beta2$	1.372
	$H_\beta1$	1.487
	C_γ	26.390
	H_γ	1.214
	$C_\delta1$	25.977
	$H_\delta1$	0.401
	$C_\delta2$	22.278
	$H_\delta2$	0.553
	C	177.110
35: GLY	N	107.041
	HN	7.688
	C_α	45.442
	$H_\alpha2$	3.702
	$H_\alpha1$	4.068
	C	174.465
36: LEU	N	120.082
	HN	7.362
	Ca	54.532
	Ha	4.408
	Cb	43.277
	Hb2	1.242
	Hb1	1.417
	Cg	27.850
	Hg	1.470
	Cd1	25.356
	Hd1	0.521
	Cd2	23.538
	Hd2	0.619
	C	176.067

Residue Number: Type	Atom Type	Chemical Shift *ppm*
37: ASN	N	119.435
	HN	8.323
	C_α	51.934
	H_α	4.633
	C_β	38.603
	$H_\beta2$	2.917
	$H_\beta1$	3.184
	$N_\delta2$	112.374
	$H_\delta21$	7.695
	$H_\delta22$	6.964
	C	175.291
38: GLU	N	120.477
	HN	8.741
	C_α	59.981
	H_α	3.800
	C_β	29.425
	H_β	2.056
	C_γ	36.238
	H_γ	2.362
	C	178.574
39: GLU	N	119.763
	HN	8.423
	C_α	59.225
	H_α	4.099
	C_β	28.744
	$H_\beta2$	1.992
	$H_\beta1$	2.086
	C_γ	36.222
	H_γ	2.318
	C	178.419
40: ALA	N	122.659
	HN	8.066
	C_α	54.785
	H_α	4.232
	C_β	18.820
	H_β	1.552
	C	180.221

Residue Number: Type	Atom Type	Chemical Shift *ppm*
41: ILE	N	118.192
	HN	8.082
	C_α	65.450
	H_α	3.457
	C_β	37.585
	H_β	1.800
	$C_\gamma1$	29.473
	$H_\gamma12$	1.730
	$H_\gamma11$	1.830
	$C_\delta1$	13.671
	$H_\delta1$	0.530
	$C_\gamma2$	17.947
	$H_\gamma2$	0.650
	C	177.050
42: ARG	N	119.922
	HN	8.066
	C_α	59.922
	H_α	4.012
	C_β	29.670
	H_β	1.981
	C_γ	27.437
	$H_\gamma2$	1.591
	$H_\gamma1$	1.778
	C_δ	43.290
	H_δ	3.240
	N_ε	83.973
	H_ε	7.685
	C	179.248
43: GLN	N	118.073
	HN	8.099
	C_α	58.959
	H_α	4.066
	C_β	28.186
	$H_\beta2$	2.230
	$H_\beta1$	2.270
	C_γ	33.893
	$H_\gamma2$	2.403
	$H_\gamma1$	2.553
	$N_\varepsilon2$	112.032
	$H_\varepsilon21$	7.695
	$H_\varepsilon22$	6.777
	C	178.204

Residue Number: Type	Atom Type	Chemical Shift *ppm*
44: PHE	N	121.120
	HN	8.364
	C_α	61.548
	H_α	4.054
	C_β	38.444
	$H_\beta2$	2.787
	$H_\beta1$	2.940
	$C_\delta1$	131.791
	$H_\delta1$	6.500
	$C_\varepsilon1$	130.731
	$H_\varepsilon1$	6.827
	C_ζ	129.925
	H_ζ	6.791
	$C_\varepsilon2$	130.731
	$H_\varepsilon2$	6.827
	$C_\delta2$	131.791
	$H_\delta2$	6.500
	C	177.659
45: PHE	N	117.646
	HN	8.491
	C_α	61.053
	H_α	4.203
	C_β	38.306
	$H_\beta2$	3.180
	$H_\beta1$	3.297
	$C_\delta1$	131.528
	$H_\delta1$	7.363
	$C_\varepsilon1$	130.916
	$H_\varepsilon1$	7.175
	C_ζ	129.222
	H_ζ	7.089
	$C_\varepsilon2$	130.916
	$H_\varepsilon2$	7.175
	$C_\delta2$	131.528
	$H_\delta2$	7.363
	C	177.906
46: GLU	N	119.461
	HN	8.172
	C_α	59.056
	H_α	4.080
	C_β	29.100
	$H_\beta2$	2.100
	$H_\beta1$	2.189
	C_γ	35.800
	$H_\gamma2$	2.314
	$H_\gamma1$	2.498
	C	178.498

Residue Number: Type	Atom Type	Chemical Shift *ppm*
47: GLU	N	118.056
	HN	7.805
	C_α	58.033
	H_α	4.064
	C_β	29.440
	H_β	1.945
	C_γ	35.988
	$H_\gamma2$	2.090
	$H_\gamma1$	2.333
	C	177.943
48: PHE	N	119.645
	HN	8.036
	C_α	59.447
	H_α	4.196
	C_β	39.415
	$H_\beta2$	2.620
	$H_\beta1$	2.885
	$C_\delta1$	131.690
	$H_\delta1$	7.026
	$C_\varepsilon1$	131.159
	$H_\varepsilon1$	7.209
	C_ζ	129.599
	H_ζ	7.213
	$C_\varepsilon2$	131.159
	$H_\varepsilon2$	7.209
	$C_\delta2$	131.690
	$H_\delta2$	7.026
	C	176.816
49: GLU	N	119.113
	HN	8.018
	C_α	56.652
	H_α	4.033
	C_β	29.668
	H_β	1.963
	C_γ	35.922
	H_γ	2.225
	C	176.626

Residue Number: Type	Atom Type	Chemical Shift *ppm*
50: GLN	N	118.828
	HN	7.836
	C_α	55.960
	H_α	4.201
	C_β	28.961
	$H_\beta2$	1.985
	$H_\beta1$	2.107
	C_γ	33.885
	$H_\gamma2$	2.350
	$H_\gamma1$	2.390
	$N_\varepsilon2$	112.590
	$H_\varepsilon21$	7.548
	$H_\varepsilon22$	6.844
	C	175.967
51: ARG	N	121.744
	HN	7.971
	C_α	56.094
	H_α	4.261
	C_β	30.390
	$H_\beta2$	1.698
	$H_\beta1$	1.808
	C_γ	27.096
	H_γ	1.588
	C_δ	43.332
	H_δ	3.088
	N_ε	84.878
	H_ε	7.232
	C	175.314
52: LYS	N	127.865
	HN	7.888
	C_α	57.578
	H_α	4.070
	C_β	33.470
	$H_\beta2$	1.650
	$H_\beta1$	1.746
	C_γ	24.800
	H_γ	1.333
	C_δ	29.110
	H_δ	1.620
	C_ε	42.300
	H_ε	2.929
	C	181.320