# **Computational Enzyme Design**

Thesis by

Daniel N. Bolon

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2002

(Defended January 16, 2002)

© 2002

Daniel N. Bolon

All Rights Reserved

## Acknowledgements

I am very appreciative of the good camaraderie and academic guidance that I have enjoyed the last few years. However, the initial transition into graduate school, including moving to a new coast, was a bit bumpy. In addition to missing friends and family, I remember feeling overwhelmed by the speed of research and the amount of knowledge that seemed required to design and carry out a thesis project. Having spent a few more years here, I have developed a deep appreciation for the Caltech community.

I would like to thank my research advisor, Stephen Mayo, for excellent ideas and scientific guidance. When I started in the lab, I proposed many unrealistic research projects to Steve. Instead of kicking me out of the lab, Steve suggested that I work on designing novel enzyme activity into an inert protein scaffold using a model of the reaction's transition state. It was a great idea, and I was fortunate enough to develop it into a workable foundation. In addition, I thank Steve for imparting his logical and farsighted approach to scientific research, his attention to detail, and his ability to present complicated ideas with great clarity.

I would like to thank Bill Goddard, Pamela Bjorkman, and Doug Rees for their guidance as my graduate committee. I have benefited greatly from their scientific expertise and career advice.

Thanks to everyone in the Mayo lab with whom I had a chance to interact. The combination of scientists with diverse backgrounds has made it a tremendous place to learn. In particular, I would like to thank Scott Ross for teaching me the theoretical and practical basics of protein nuclear magnetic resonance spectroscopy. Thanks to Fred Lee and Dirk Bökenkamp for their initial efforts in computational enzyme design. Thanks to

Sandy Malakauskus, Chantal Morgan, and Monica Breckow for teaching me the basics of circular dichroism, high-pressure liquid chromatography and how molecular biology works in the Mayo lab. Thanks to Ben Gordon and Arthur Street for explaining the details of the ORBIT protein design software and helping me to learn/debug FORTRAN. Thanks to Shannon Marshall for useful discussions and explaining electrostatic modeling. Thanks to Cynthia Carlson, Rhonda Digiusto, and Monica Breckow for making the lab run so smoothly. Thanks to Marie Ary for critical comments that have greatly improved my scientific writing. Thanks to everyone else in the Mayo lab and others involved in protein design at Caltech including: Niles Pierce, Caglar Tanrikulu, Robert Dirks, Bassil Dahiyat, Alyce Su, Cathy Sarisky, Julia Shifman, J.J. Plecs, Julie Archer, Possu Huang, Chris Voigt, Deepshikha Datta, Premal Shah, Geoffrey Hom, Shira Jacobson, Kirsten Lassila, Jessica Mao, Eric Zollars, and Eun-jung Choi.

Thanks to Pavel Strop for scientific discussions, going rock climbing, drinking beer, and blasting music in the lab after everyone had left. In this vein I also thank Lance Martin, Kirk Hansen, Brian King, Isaac Carrico, and Kent Kirshenbaum.

Financially I was supported by the Helen G. and Arthur McCallum Foundation, the Evelyn Sharp Graduate Fellowship, and a training grant from the National Institutes of Health. I also benefited from financial support for the Mayo lab from the Howard Hughes Medical Institute. Because of this funding, I was fortunate to always be able to carry out the experiments that I found worthwhile.

Thanks to all the people that make Caltech run so smoothly. Thanks to Liz Ayala, Gwen Murdock, and Alison Ross in Biology and Biochemistry. Thanks to Michael Werner, Angel Rooker and Chris Schuld in the Caltech HHMI office.

Thanks to my dad for teaching me math in grade school. Not being a teacher by profession, he took the time to prepare lessons every day. In addition to geometry, algebra and calculus, I learned a logical approach to problem solving that continues to help me today. While I didn't overly enjoy it at the time, today I am very appreciative. Thanks!

Thanks to the late David Bing for giving me my initial research opportunity. While I was an undergraduate he gave me the chance to work in his lab at the Center for Blood Research. His enthusiasm for science and life is something that I will always remember. Thanks to Janice Williamson and Susan Mitchell for teaching me practical molecular biology while in David's lab.

Thanks to my many inspirational teachers. In particular, thanks to Daniele Armaleo at Duke University for explaining how everything in a molecular biology lab works and making science fun. Thanks to Mrs. Gibbs at Brookline High School for her caring and clear instruction in chemistry.

Thanks to my mom and dad for being very encouraging and providing me with everything that I have needed since I was a wee baby.

I have really enjoyed my last few years here at Caltech thanks in large part to my family. I am very grateful for the time that we get to share. Thanks to Jennie, Cecelia and Anna.

### **Abstract**

The long-term objective of computational enzyme design is the ability to generate efficient protein catalysts for any chemical reaction. This thesis develops and experimentally validates a general computational approach for the design of enzymes with novel function.

In order to include catalytic mechanism in protein design, a high-energy state (HES) rotamer (side chain representation) was constructed. In this rotamer, substrate atoms are in a HES. In addition, at least one amino acid side chain is positioned to interact favorably with substrate atoms in their HES and facilitate the reaction. Including an amino acid side chain in the HES rotamer automatically positions substrate relative to a protein scaffold and allows protein design algorithms to search for sequences capable of interacting favorably with the substrate. Because chemical similarity exists between the transition state and the high-energy state, optimizing the protein sequence to interact favorably with the HES rotamer should lead to transition state stabilization. In addition, the HES rotamer model focuses the subsequent computational active site design on a relevant phase space where an amino acid is capable of interacting in a catalytically active geometry with substrate.

Using a HES rotamer model of the histidine mediated nucleophilic hydrolysis of *p*-nitrophenyl acetate, the catalytically inert 108 residue *E. coli* thioredoxin as a scaffold, and the ORBIT protein design software to compute sequences, an active site scan identified two promising active site designs. Experimentally, both candidate "protozymes" demonstrated catalytic activity significantly above background. In

addition, the rate enhancement of one of these "protozymes" was the same order of magnitude as the first catalytic antibodies.

Because polar groups are frequently buried at enzyme-substrate interfaces, improved modeling of buried polar interactions may benefit enzyme design. By studying native protein structures, rules have been developed within the scope of protein design that require core polar residues to largely satisfy their hydrogen bonding potential. Using this polar strategy to design the core of thioredoxin resulted in a protein that was thermodynamically stabilized relative to both the wt protein and a protein designed without core polar residues.

The enzyme design procedures presented here may serve as a platform to develop more detailed methods. It is hoped that the development and experimental testing of more detailed methods will continue to improve our understanding of enzyme mechanism and lead to the long-term goal of designing highly efficient enzymes.

# **Table of Contents**

Acknowledgements	
Abstract	V
Table of Contents	
List of Tables and Figures	X
Abbreviations	xiv
Chapter 1. Introduction	1
References	Ģ
Chapter 2. De Novo Design of Biocatalysts	11
Summary	12
Introduction	13
Rational Computational Design	13
Rational Homology Driven Design	16
Catalytic Antibodies	17
Directed Evolution	18
Conclusions	21
References	22
Chapter 3. Enzyme-like Proteins by Computational Design	28
Summary	29
Introduction	30
Materials and Methods	33
Computational	33
Experimental	36

Results and Discussion	37
Computational design	37
Experimental validation	40
Conclusions	47
References	49
Chapter 4. Polar Residues in the Protein Core of <i>Escherichia coli</i> Thioredoxin Are Important for Fold Specificity	54
Summary	55
Introduction	56
Materials and Methods	58
Analysis of polar residues at core positions in protein structures	58
Computational modeling	59
Mutagenesis and protein purification	59
CD analysis	60
Fluorescence analysis	60
DSC	60
NMR spectroscopy	61
Results	61
Discussion	68
References	74
Chapter 5. Prudent Modeling of Core Polar Residues in Computational Protein Design	80
Summary	81
Introduction	82
Materials and Methods	84

Database analysis and construction of hydrogen bond rules for protein core design	84
Computational modeling	85
Mutagenesis and protein purification	88
CD analysis	89
DSC	89
Results	90
Discussion	98
Conclusions	104
References	105
Appendix A. Residue Classification is a Good Predictor of Burial	110
Appendix B. Variation in Statistical Secondary Structure Propensity is Reduced at Core Positions Relative to at Surface Positions	
Appendix C. Summary of Thermodynamic Data	123

# **List of Tables and Figures**

Chapter 1		
Figure 1.1	Reaction coordinate illustrating transition state stabilization	3
Figure 1.2	Design of a high energy state rotamer	4
Figure 1.3	Overview of computational enzyme design	5
Chapter 2		
Figure 2.1	Comparison of computational enzyme design method with catalytic antibody method	14
Figure 2.2	Fitness landscape	16
Chapter 3		
Table 3.1	Partial charges on computed high energy state	34
Table 3.2	Top ten designs from active site scan	38
Figure 3.1	Structure of the high-energy state rotamer	32
Figure 3.2	Ribbon model of PZD2	40
Figure 3.3	Molecular surfaces comparing the designed active site of PZD2 with the wt scaffold	41
Figure 3.4	Kinetic model used to analyze the activity of PZD2	42
Figure 3.5	Velocity versus substrate concentration	42
Figure 3.6	Buffer corrected hydrolysis	44
Figure 3.7	Trapping of an acylated intermediate by mass spectrometry	45
Figure 3.8	Inhibitor analysis	46
Chapter 4		
Table 4.1	Experimental thermodynamic data	65

Figure 4.1	Number of polar residues at core positions as a function of protein size	62
Figure 4.2	Ribbon diagrams illustrating core polar residues in thioredoxin	63
Figure 4.3	CD measurements of thioredoxin variants	66
Figure 4.4	Proton NMR spectra of thioredoxin variants showing the amide region	67
Figure 4.5	Hydrogen-deuterium exchange of thioredoxin variants	68
Figure 4.6	Proton NMR spectra of thioredoxin variants showing the upfield methyl region	69
Chapter 5		
Table 5.1	Hydrogen bonding at core and surface positions	90
Table 5.2	Hydrogen bonding by different amino acid types at core positions	91
Table 5.3	Hydrogen bond requirements for polar residues at core positions	92
Table 5.4	Experimental thermodynamic data	97
Figure 5.1	Optimization progress as a function of processor time	93
Figure 5.2	Amino acid sequence comparison of thioredoxin core designs	94
Figure 5.3	Molecular models of wt thioredoxin and core designs	95
Figure 5.4	CD spectra of thioredoxin variants	96
Figure 5.5	Correlation of the number of core MCU atoms with protein size	99
Appendix A		
Figure A.1	Percent ASA burial of side chains classified as surface, boundary, or core	111
Figure A.2	Percent ASA burial of main chain carbonyl oxygen and amide hydrogen atoms classified as surface, boundary, or core	113

Appendix B		
Figure B.1	Normalized deviations of secondary structure propensity at surface and core positions	117
Figure B.2	$\alpha$ propensity	119
Figure B.3	$\beta$ propensity	120
Appendix C		
Table C.1	Experimental thermodynamic data	124

### **Abbreviations**

 $k_{\rm cat}$  catalytic rate

**K**<sub>m</sub> Michaelis constant

**TSA** transition state analogue

**HESR** high energy state rotamer

**ASS** active site scan

E. Coli Escherichia coli

**CD** circular dichroism

**DSC** differential scanning calorimetry

T<sub>m</sub> melting temperature

 $\Delta H$  enthalpy of unfolding

 $\Delta C_p$  difference in heat capacity between the unfolded and folded state

 $T_m$  melting temperature

**GdmCl** guanidinium chloride

 $\Delta G$  free energy of unfolding

m value slope of  $\Delta G$  versus denaturant concentration

C<sub>m</sub> midpoint of guanidinium chloride unfolding transition

**NMR** nuclear magnetic resonance

**HB** hydrogen bond

**SC** side chain

MC main chain

MCU main chain atoms that do not make MC-MC hydrogen bonds

# Chapter 1

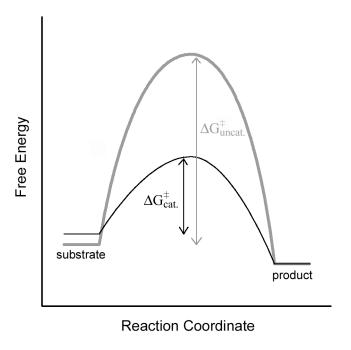
Introduction

The long-term objective of computational enzyme design is the generation of efficient protein catalysts for any chemical reaction. The ability to design enzymes for a given reaction has considerable practical application, particularly for the synthesis of pharmaceuticals.<sup>1</sup> However, the design of proteins with novel catalytic properties has met with extremely limited success.<sup>2</sup> This thesis develops and experimentally validates a general computational approach for the design of enzymes capable of catalyzing a novel target reaction.

Protein design methods with predictive capability as well as principles of enzymatic catalysis gained from classical biochemistry form the foundations for successful computational enzyme design. The use of rotamer (amino acid side chain identity and geometry) libraries combined with a powerful search algorithm capable of identifying the global minimum energy conformation (GMEC) was initially applied to side chain placement.<sup>3</sup> By allowing side chain identity to vary, protein sequences were designed to optimally interact with the folded backbone conformation.<sup>4</sup> A design cycle was then used to improve the predictive capability of a force field for protein design calculations.<sup>4,5</sup> The resulting ORBIT (Optimization of Rotamers By Iterative Techniques) protein design software successfully designed a 28 amino acid sequence that folded into the target conformation.<sup>6</sup> In combination with the basic principles of enzymatic catalysis (e.g., transition state stabilization),<sup>7</sup> the ORBIT protein design software was used to computationally design an active site into an inert protein scaffold.

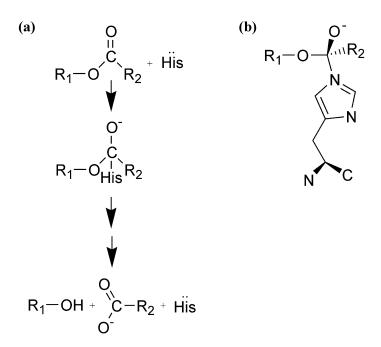
Natural enzymes accelerate reactions by precisely orienting catalytic side chains (e.g., in acid/base catalysis) and through transition state stabilization (Fig. 1.1). Thus, active site design should strive to position catalytic side chains to facilitate the reaction of

interest as well as arrange the active site to interact favorably with substrate atoms in a



**Figure 1.1** Lowering the free energy of activation  $\Delta G^{\ddagger}$ , by transition state stabilization and ground state destabilization results in faster reaction rates for the catalyzed reaction (black) compared to the un-catalyzed reaction (grey).

high-energy state. These principles were incorporated into ORBIT through the design of a high-energy state rotamer (HESR). The HESR includes substrate atoms in a high-energy state as well as at least one amino acid side chain geometrically positioned to facilitate the reaction (Fig. 1.2). Including an amino acid side chain in the HES rotamer automatically positions substrate relative to a protein scaffold and allows protein design algorithms to search for sequences capable of interacting favorably with the substrate. Because chemical similarity exists between the transition state and the high-energy state, optimizing the protein sequence to interact favorably with the HESR should lead to transition state stabilization. In addition, the HESR model focuses the subsequent computational active site design on a relevant phase space where an amino acid is capable of interacting in a catalytically active geometry with substrate.

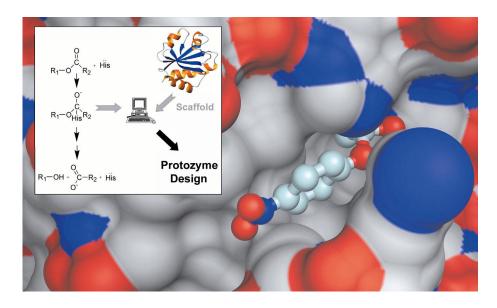


**Figure 1.2** Incorporation of catalytic residues and transition state stabilization in computational protein design through the design of a high-energy state rotamer. The reaction mechanism of histidine mediated nucleophilic ester hydrolysis (a) was used to design a high energy state rotamer for computational design (b).

As an initial experimental validation, we used a HES rotamer model of the histidine mediated nucleophilic hydrolysis of *p*-nitrophenyl acetate, the catalytically inert 108 residue *E. coli* thioredoxin as a scaffold, and the ORBIT protein design software to compute sequences (Chapter 3). By scanning this HESR at different positions within the target scaffold, two promising active site designs were identified within the target scaffold (Figure 1.3). Experimentally, both candidate "protozymes" demonstrated catalytic activity significantly above background. In addition, the rate enhancement of one of these "protozymes" was the same order of magnitude as the first catalytic antibodies.

High thermodynamic stability of the free protein is not required for efficient enzymatic catalysis. However, because enzyme activity is dependent upon the folded structure, the activity of enzymes decreases dramatically at temperatures sufficient to

denature the protein.<sup>8</sup> In early attempts at computational enzyme design in the scaffold of the 56 residue β1 domain of protein G, predicted sequences failed to fold, frustrating further enzymatic analysis (F. S. Lee, D. I. Bökenkamp, and S. L. Mayo; unpublished results). Therefore, for subsequent enzyme design efforts, the 108 residue *Escherichia coli* thioredoxin was selected as a scaffold because of its high thermostability<sup>9</sup> and moderate size.



**Figure 1.3** Computational design of an active site. A reaction mechanism and backbone scaffold are used to computationally predict protein sequences capable of binding substrate and catalyzing the reaction of interest. Molecular surface focusing on a designed active site with a catalytic histidine side chain positioned for nucleophilic attack on an activated ester (shown in ball and stick form).

Given that enzyme efficiency depends on interactions between the enzyme and substrate (e.g., transition state stabilization), the design of an effective active site should maximize beneficial interactions with the HESR provided that the protein sequence retains sufficient energetic stability to fold properly. As an initial approach to energetically favor interactions between the protein and the HESR, hydrophobic solvation was selectively applied to the HESR in the design of first generation "protozymes" (Chapter 3). Future calculations may be improved by the use of a more

detailed energy bias towards the HESR. For example, if the charge magnitude is increased at a particular substrate atom in the high-energy state relative to the ground state, then favorable electrostatic interactions with this atom in the HESR should lead to a reduction in the activation energy.

In parallel to the incremental development<sup>4,6,10,11</sup> of ORBIT for protein design, our initial computational enzyme design strategy forms a basic foundation upon which more detailed modeling methods may be developed. For example, the amino acid identity of non-HESR positions was initially optimized as either alanine or the wt identity (Chapter 3) in order to limit combinatorial complexity. The limited combinatorial complexity of these calculations made it feasible to optimize the entire protein in separate calculations with the HESR at different positions.

While the basic "protozyme" design strategy resulted in proteins with enzymatic activity, increased modeling detail may lead to the design of more efficient catalysts. For example, allowing a more diverse set of amino acid identities at non-HESR positions may allow for improved protein interaction with substrate atoms in the high-energy state. However, this strategy will increase the combinatorial complexity compared to restricting amino acid identity to wt or alanine, making the computational search more difficult. This increase in combinatorial complexity could potentially be offset by focusing the design on optimizing the amino acid identity at positions capable of directly interacting with the HESR.

The arrangement of side chains in active sites is frequently thermodynamically destabilizing.<sup>12</sup> However, the rest of the protein can provide enough thermodynamic stabilization to distinguish the native fold from alternative folds and help to position the

active site. In order to better understand the scaffold structure used in the "protozyme" designs, ORBIT was used to optimize the thioredoxin core for stability. In order to maximize hydrophobic burial, a driving force in stabilizing native protein conformations, 13 core residues were restricted to hydrophobic identity in this optimization. The restriction of residues to hydrophobic identity is common in core protein design. 4,6,11,14-19 The amino acid sequence that resulted from the ORBIT optimization of thioredoxin contained eight mutations, five of which were hydrophilic in the wild type (wt) protein. The designed sequence ("Core8") expressed poorly in E. coli relative to wt, and was dramatically destabilized relative to wt (Appendix, Table C.1). Sequences containing individual polar to hydrophobic mutations predicted by ORBIT were then experimentally characterized (Chapter 4). From this analysis, it was clear that the polar to hydrophobic mutations were largely responsible for the decreased thermodynamic stability of the core design relative to wt. This result prompted a database survey that revealed that the number of polar residues at core positions increases with protein size (Chapter 4). The strength of this correlation suggests that core polar residues play a structural role. Further experimental characterization in thioredoxin demonstrated that core polar residues could aid in fold specificity, the ability to energetically distinguish a single well-folded state from alternative structures (Chapter 4).

The accurate energetic modeling of buried polar interactions is made difficult by variations in local dielectric environments as well as variation in the charge distribution within molecules due to polarization. However, based on surveys of protein structures, it is possible to construct effective methods for modeling buried polar interactions within the scope of protein design. For example, there is clearly an energetic penalty for

burying un-satisfied hydrogen bond donors and acceptors.<sup>20</sup> Therefore, it is reasonable for protein design to attempt to fulfill the hydrogen bond potential of buried polar atoms. Based on this premise, hydrogen bonding by polar core side chains was analyzed in a set of protein structures in order to construct a prudent polar strategy for core protein design (Chapter 5). Using the prudent polar strategy to design the core of thioredoxin resulted in a molecule that was thermodynamically stabilized relative to both wt and a sequence designed without core polar residues. The basic concept of attempting to satisfy the hydrogen bond potential of buried atoms should extend to other design targets involving buried polar atoms, including protein-protein interactions and protein-small molecule interactions (e.g., enzyme design).

The initial success in the computational design of a "protozyme" for activated ester hydrolysis indicates the potential of ORBIT for functional protein design. However, the design of highly efficient enzymes remains a formidable challenge. Improvements in our understanding of the adaptive potential of different protein folds for function and continued improvements in the modeling of buried polar interactions should aid in this effort.

#### References

- Liese, A. & Filho, M. V. (1999). Production of fine chemicals using biocatalysis.
   Curr Opin Biotechnol 10(6), 595-603.
- 2. Corey, M. J. & Corey, E. (1996). On the failure of de novo-designed peptides as biocatalysts. *Proc Natl Acad Sci U S A* **93**(21), 11428-11434.
- 3. Desmet, J., Maeyer, M. D., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539-542.
- 4. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci* **5**(5), 895-903.
- 5. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr Opin Struct Biol* **9**(4), 509-513.
- 6. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science* **278**, 82-87.
- 7. Fersht, A. (1985). *Enzyme structure and mechanism*. 2nd edit, Freeman, New York.
- 8. Stenesh, J. (1993). Core topics in biochemistry, Cogno Press, Kalamazoo, MI.
- 9. Ladbury, J. E., Wynn, R., Hellinga, H. W. & Sturtevant, J. M. (1993). Stability of oxidized Escherichia coli thioredoxin and its dependence on protonation of the aspartic acid residue in the 26 position. *Biochemistry* **32**(29), 7526-7530.
- 10. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci* **6**(6), 1333-1337.
- 11. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* **94**(19), 10172-10177.

- Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc Natl Acad Sci U S A* 92(2), 452-456.
- 13. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155.
- 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(13), 5803-5807.
- 15. Desjarlais, J. R. & Handel, T. M. (1995). De novo design of the hydrophobic cores of proteins. *Protein Sci* **4**(10), 2006-2018.
- 16. Desjarlais, J. R. & Handel, T. M. (1999). Side-chain and backbone flexibility in protein core design. *J Mol Biol* **290**(1), 305-318.
- 17. Lazar, G. A., Desjarlais, J. R. & Handel, T. M. (1997). De novo design of the hydrophobic core of ubiquitin. *Protein Sci* **6**(6), 1167-1178.
- 18. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* **305**(3), 619-631.
- 19. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Structure Fold Des* **7**(5), R105-109.
- 20. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**(5), 777-793.

# Chapter 2

De Novo Design of Biocatalysts\*

<sup>\*</sup>Adapted from Bolon, D. N.; Voigt, C. A.; Mayo S. L. Curr. Opin. Chem. Biol. 2002, in press.

### Summary

The challenging field of *de novo* enzyme design is beginning to produce exciting results. The application of powerful computational methods to functional protein design has recently succeeded at engineering target activities. In addition, efforts in directed evolution continue to expand the transformations that can be accomplished by existing enzymes. The engineering of completely novel catalytic activity requires traversing inactive sequence space in a fitness landscape, a feat which is better suited to computational design. Optimizing activity, which can include subtle alterations in backbone conformation and protein motion is better suited to directed evolution, which is highly effective at scaling fitness landscapes towards maxima. Improved rational design efforts coupled with directed evolution should dramatically improve the scope of *de novo* enzyme design.

#### Introduction

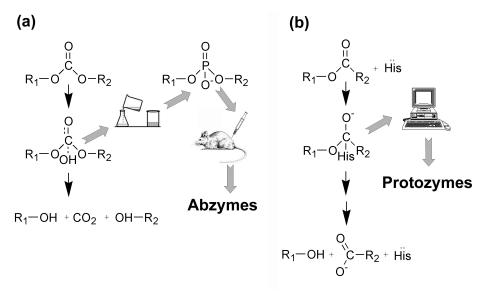
There are many properties of natural enzymes that make them appealing for chemical production. Natural enzymes are highly efficient catalysts with typical apparent second order rate constants ( $k_{cat}/K_m$ ) from  $10^6$  to  $10^8$  M<sup>-1</sup>sec<sup>-1</sup>. Naturally occurring enzymes are able to effectively catalyze difficult chemical transformations under mild conditions. Rate accelerations over background ( $k_{cat}/k_{uncat}$ ) can reach  $10^{17}$ . By binding substrates in well-defined active sites with catalytic side chains geometrically positioned, enzymes are highly selective catalysts. The enantioselectivity of enzymes is increasingly being used in the production of chiral pharmaceuticals. In addition, the general selectivity of enzymes results in few by-products, which, combined with their energy efficient operation under mild conditions, make enzymes environmentally friendly. Due to the high selectivity of enzymes, relevant natural enzymes do not exist for many industrially important transformations. For this reason, there is great practical interest in the design of enzymes with novel activities.

This review highlights the strengths of rational design, catalytic antibodies, and directed evolution for the design of novel enzymes. The strengths of different techniques are highly complementary, suggesting that combined approaches may have significant advantages over approaches that rely on a single method.<sup>5-7</sup>

#### **Rational Computational Design**

Computational protein design<sup>8</sup> starts with the coordinates of a protein main chain and uses a force field to identify sequences and geometries of amino acids that are optimal for stabilizing the backbone geometry. Even for small proteins, the number of

possible sequences far exceeds that which can be exhaustively searched. The development of powerful search algorithms to find optimal solutions has provided a major stimulus to the field. Studies correlating predicted and experimental stability have since been used to iteratively improve an empirical force field for protein design calculations. The combination of predictive force fields and fast search algorithms are now being applied to functional protein design.



**Figure 2.1** The similar procedures for computationally designed protozymes and traditional transition state anologue induced catalytic antibodies are illustrated. Computational enzyme design (a) models a reaction with the involvement of at least one catalytic amino acid side chain. A high-energy state of the reaction is then computationally modeled and used to identify active sites capable of binding substrate and orienting catalytic side chains. Catalytic antibody design (b) identifies a transition state in a reaction pathway. Transition state analogues are then synthesized and used to elicit antibodies with binding sites that are complementary to the transition state.

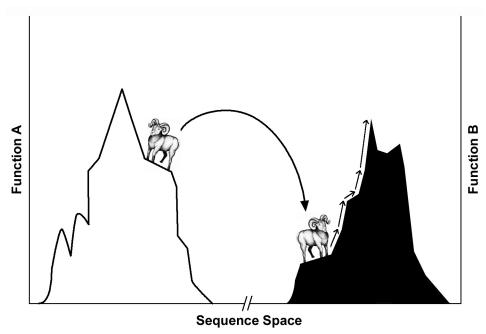
A novel active site for activated ester hydrolysis was computationally designed in the scaffold of the inert protein thioredoxin.<sup>12</sup> In a method analogous to catalytic antibodies, the high-energy state of a reaction was modeled with a side chain geometrically oriented for catalysis (Fig. 2.1A). Numerous low-energy geometries of this modeled to generate a library of high-energy state rotamers (HESR) for the design. To identify a novel active site, optimal sequences were computed with the HESR at different

positions in the protein. This active site scan (ASS) explicitly limited the search to a relevant phase space where the catalytic residue could be positioned to properly interact with the substrate. Ranking of different active sites from the scan based on substrate recognition successfully predicted "protozymes" that were experimentally shown to catalyze the reaction by the designed method of histidine mediated nucleophilic catalysis. While the catalytic activity of the designed protein was very modest, on the same order as the initial catalytic antibodies, the generalness of the approach promises interesting future results, including the ability to design proteins to catalyze reactions that are inaccessible by natural enzymes.

The diverse and powerful chemistry of metals makes design of metalloproteins a promising approach to enzyme design. Early pioneering work included the development of a computational method to identify sites in proteins capable of using side chains to complex metal atoms. By modeling iron with one primary coordination sphere ligated with di-oxygen, a number of metalloproteins were designed in the thioredoxin fold. These proteins were experimentally shown to bind iron and catalyzed a variety of oxygen chemistries.

The engineering of a completely new function into an inert protein scaffold is likely to require multiple coupled mutations where individual mutations may have no effect. In terms of a fitness landscape, engineering a novel function is likely to require traversing inactive sequence space (Fig. 2.2). Computational design benefits from the ability to search a larger space than possible by purely experimental methods and is well suited to designing novel function. Computational design is not currently an effective tool for increasing fitness, which may occur through subtle changes in backbone structure

and protein dynamics. However, current research is actively exploring these issues. The use of a flexible main chain was recently used to successfully design a left-handed coiled coil. The backbone geometry and sequence of a  $\beta$ -turn was recently designed in order to increase the folding rate and improve thermodynamic stability. Backbone flexibility has also been explored in the design of protein cores.  $^{18,19}$ 



**Figure 2.2** Fitness landscapes, illustrating a path to go from high fitness for function A to high fitness for a novel function B. The first jump to attain an initial level of Function B requires crossing an inactive region of sequence space. The extremely large sequence space that can be searched by computational design allows it to effectively jump over this inactive space. Directed evolution should then be an effective tool to iteratively improve function B to a high level of activity.

# Rational Homology Driven Design

The rational redesign of enzymes uses structural and/or sequence homology to graft desired properties of one enzyme onto another. The goals of homology driven experiments include engineering binding sites to fit different substrates as well as construction of new catalytic residues to alter mechanism and function.<sup>5</sup> Rational design

generally focuses on changing residues that contact substrate. The efficiency of most homology-designed enzymes has been poor relative to natural enzymes; however, some experiments have resulted in strikingly efficient enzymes.

The engineering of a catalytic Ser-His-Asp triad into a peptidyl-prolyl isomerase resulted in a remarkably efficient proline specific endopeptidase.<sup>20</sup> Based on structural alignment of 4-chlorobenzoyl-CoA dehalogenase and crotonase, engineering of two catalytically active acid/base residues as well as six positioning residues resulted in efficient crotonase activity in the dehalogenase.<sup>21</sup> Both of these designs started with significant catalytic machinery in place to polarize a carbonyl group within the substrate, and successfully included additional catalytic residues in order to support a related reaction.

Homology driven design has greatly improved our understanding of enzyme mechanisms and substrate specificity. This information together with more accurate and detailed physical models should greatly benefit future enzyme designs.

## Catalytic Antibodies

The use of transition state analogues (TSA) to elicit catalytic antibodies (abzymes) effectively uses the diversity of the immune system to identify proteins with substrate affinity and catalytic rate enhancement. The method includes synthesis of a TSA, immunization, isolation of monoclonal antibodies via hybridomas, and screening for catalytically effective antibodies (Fig. 2.1B). The technique of immunization with a TSA has been used to identify antibodies that successfully catalyze a number of different

reactions.<sup>22</sup> However, the catalytic efficiency of the resulting antibodies has been low relative to natural enzymes.<sup>23</sup>

In a significant advance, catalytic antibodies that utilize a nucleophilic mechanism were selected by reactive immunization and resulted in efficient catalysts.<sup>22</sup> Rather than a transition state analogue, a mechanism-based inhibitor was used to elicit the immune response. Antibodies that formed stable covalent attachments to the suicide inhibitor were effectively selected. This method was employed in the selection of an efficient abzyme with a nucleophilic lysine for aldol condensations.<sup>24</sup> The efficiency of this aldolase demonstrates the effectiveness of covalent catalysis. The ability to select for powerful catalytic groups and active sites with high transition state specificity could theoretically yield more efficient catalysts.

#### **Directed Evolution**

Directed evolution has emerged as a popular method for protein engineering. 4,25 Mimicking natural evolution, an initial parent gene is chosen and a diverse library of offspring genes is created through mutagenesis or recombination. A screen or selection is applied to the library and the mutants that exhibit the greatest improvement in the desired properties are chosen to become the parents to the next generation. This iterative search has generated large improvements in properties such as activity, specificity (including enantioselectivity<sup>26</sup>), stability, and has been used to evolve systems of enzymes, such as found in metabolic pathways. Directed evolution often discovers these improvements by making a few amino acid substitutions that collectively have an important functional

effect. For this reason, the method generally requires a starting protein with some activity towards the desired reaction.<sup>28</sup>

One of the critical steps in directed evolution is the creation of a screen or selection that is rapid enough to process millions of mutants while accurately measuring a desired property.<sup>29</sup> Screening by monitoring formation of the exact product of interest is desirable, but limits throughput. For this reason, the development of ingenious high throughput methods is an active area of research. The use of fluorogenic substrates for high throughput screening is a productive field that has been recently reviewed.<sup>30</sup> Other methods include the linking of substrate to phage particles, after which, product binding can be used to enrich for phage displaying active enzymes.<sup>31-34</sup> The *in vivo* three hybrid system is based on substrate competition.<sup>35</sup> In this system, the substrate of the reaction competes with a synthetic dimerizer and inactivates a transcription factor. Conversion of substrate to product shifts the system in favor of dimerization and can be followed by the expression of a reporter gene.

Most selection schemes indirectly screen for catalysis, often through binding events. One of the difficulties of this is that the mutants that are discovered tend to bind more tightly to the substrates, but are not necessarily better catalysts. To overcome this limitation, Iverson and co-workers coupled catalytic turnover with fluorescence, so that high throughput could be achieved using FACS (fluorescence-activated cell sorter). A fluorophore (F) and a quencher (Q) were tethered to a peptide substrate. Upon action of a protease, the substrate was cleaved, separating the F and Q moieties and disrupting the intramolecular quenching. Fluorescence is thus linked to a turnover event.

Because screening is often the limiting step, there have been several studies to optimize the mutation or recombination rate with respect to the number of mutants that can be screened.<sup>38</sup> Several groups have developed models that use mutagenesis data to expose the ruggedness of the fitness landscape, where non-additive effects between mutations are more prevalent as the ruggedness increases.<sup>39,40</sup> Using these models, it has been found that the optimal mutation rate decreases as the ruggedness increases<sup>38,41</sup> and that the ruggedness affects the optimal ratio of parent genes for recombination.<sup>42</sup>

Mutagenizing restricted regions of the gene can also optimize the search. A difficulty in doing this is deciding where the diversity should be targeted, as improvements can be propagated by subtle changes distant from the active site. 43-45 One approach has been to highly mutagenize the active site. The majority of the mutants created by this technique have drastically reduced fitnesses, thus requiring extremely large selection capabilities to discover improvements. Another successful approach has been to saturate residues where directed evolution previously discovered improvements. Overall, there is a need for algorithms that can aid in targeting regions of the protein structure that are likely to demonstrate improvement.

Several algorithms have been proposed that focus the diversity towards regions that are likely to preserve the structure, under the assumption that maintaining the stability of the structure is a prerequisite for discovering improvements. Towards this end, there is promise in merging computational and experimental combinatorial methods. Inverse folding algorithms can be used to predict the amino acid substitutions that are consistent with the parental structure. Mean-field techniques have been used to accelerate the calculation. 49,50 It is useful to condense the sequence information with an

entropy measure to identify those residues that are the most likely to retain the structure. The residues that least perturb the structure can then be mutagenized and screened for improvements in the desired properties.

#### Conclusions

The challenging field of *de novo* enzyme design has begun to yield proteins with impressive catalytic efficiency. However, the current methods are not sufficient to design efficient enzymes for many reactions. Most design efforts to date have focused on using a single method, a drawback given the potential of combining complementary methods. Computational design is well suited to optimizing direct interactions including catalytic machinery, but is not well suited for identifying distant mutations that may subtly alter the protein structure to enhance catalysis. Directed evolution effectively optimizes activity, but is not as well suited to introducing completely novel activity. A few recent efforts have taken advantage of the complementary traits of the different methods. Rational design was used to engineer a low level of phosphoribosyl-anthranilate isomerase (PRAI) activity into indole-3 glycerol phosphate synthase (IGPS) allowing selection in bacteria. Directed evolution was then used to generate a protein with highly efficient PRAI activity. Starting with an RNA binding antibody, structurally guided mutagenesis of a single catalytic histidine yielded moderate ribonuclease activity.<sup>51</sup> The combined use of computational design with directed evolution and antibody techniques should continue to result in improved *de novo* biocatalyst designs.

#### References

- 1. Radzicka, A. & Wolfenden, R. (1995). A proficient enzyme. Science 267, 90-93.
- 2. McCoy, M. (1999). Biocatalysis grows for drug synthesis. *Chem Eng News* **77**, 10-14.
- 3. Schmid, A., Dordick, J. S., Hauer, B., Kiener, A., Wubbolts, M. & Witholt, B. (2001). Industrial biocatalysis today and tomorrow. *Nature* **409**, 258-268.
- 4. Arnold, F. H. (2001). Combinatorial and computational challenges for biocatalyst design. *Nature* **409**, 253-257.
- 5. Cedrone, F., Menez, A. & Quemeneur, E. (2000). Tailoring new enzyme functions by rational redesign. *Curr Opin Struct Biol* **10**(4), 405-410.
- 6. Petsko, G. A. (2000). Enzyme evolution. Design by necessity. *Nature* **403**, 606-607.
- 7. Altamirano, M. M., Blackburn, J. M., Aguayo, C. & Fersht, A. R. (2000). Directed evolution of new catalytic activity using the alpha/beta- barrel scaffold. *Nature* **403**, 617-622.
- 8. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Structure Fold Des* **7**(5), R105-109.
- 9. Desjarlais, J. R. & Clarke, N. D. (1998). Computer search algorithms in protein modification and design. *Curr Opin Struct Biol* **8**(4), 471-475.
- Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr Opin Struct Biol* 9(4), 509-513.
- 11. Shimaoka, M., Shifman, J. M., Jing, H., Takagi, J., Mayo, S. L. & Springer, T. A. (2000). Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nat Struct Biol* **7**(8), 674-678.

- 12. Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational protein design. *Proc Natl Acad Sci U S A* **98**, 14274-14279.
- 13. Benson, D. E., Wisz, M. S. & Hellinga, H. W. (1998). The development of new biotechnologies using metalloprotein design. *Curr Opin Biotechnol* **9**(4), 370-376.
- Hellinga, H. W. & Richards, F. M. (1991). Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with predefined geometry. *J Mol Biol* 222(3), 763-785.
- 15. Benson, D. E., Wisz, M. S. & Hellinga, H. W. (2000). Rational design of nascent metalloenzymes. *Proc Natl Acad Sci U S A* **97**(12), 6292-6297.
- Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. (1998). Highresolution protein design with backbone freedom. *Science* 282, 1462-1467.
- 17. Nauli, S., Kuhlman, B. & Baker, D. (2001). Computer-based redesign of a protein folding pathway. *Nat Struct Biol* **8**(7), 602-605.
- 18. Su, A. & Mayo, S. L. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci* **6**(8), 1701-1707.
- 19. Desjarlais, J. R. & Handel, T. M. (1999). Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* **290**(1), 305-318.
- 20. Quemeneur, E., Moutiez, M., Charbonnier, J. B. & Menez, A. (1998). Engineering cyclophilin into a proline-specific endopeptidase. *Nature* **391**, 301-304.
- 21. Xiang, H., Luo, L., Taylor, K. L. & Dunaway-Mariano, D. (1999). Interchange of catalytic activity within the 2-enoyl-coenzyme A hydratase/isomerase superfamily based on a common active site template. *Biochemistry* **38**(24), 7638-7652.

- 22. Hilvert, D. (2000). Critical analysis of antibody catalysis. *Annu Rev Biochem* **69**, 751-793.
- 23. Stewart, J. D. & Benkovic, S. J. (1995). Transition-state stabilization as a measure of the efficiency of antibody catalysis. *Nature* **375**, 388-391.
- 24. Wagner, J., Lerner, R. A. & Barbas, C. F., 3rd. (1995). Efficient aldolase catalytic antibodies that use the enamine mechanism of natural enzymes. *Science* **270**, 1797-1800.
- 25. Petrounia, I. P. & Arnold, F. H. (2000). Designed evolution of enzymatic properties. *Curr Opin Biotechnol* **11**(4), 325-330.
- May, O., Nguyen, P. T. & Arnold, F. H. (2000). Inverting enantioselectivity by directed evolution of hydantoinase for improved production of L-methionine. *Nat Biotechnol* 18(3), 317-320.
- 27. Schmidt-Dannert, C., Umeno, D. & Arnold, F. H. (2000). Molecular breeding of carotenoid biosynthetic pathways. *Nat Biotechnol* **18**(7), 750-753.
- 28. Shao, Z. & Arnold, F. H. (1996). Engineering new functions and altering existing functions. *Curr Opin Struct Biol* **6**(4), 513-518.
- 29. Zhao, H. & Arnold, F. H. (1997). Combinatorial protein design: strategies for screening protein libraries. *Curr Opin Struct Biol* **7**(4), 480-485.
- 30. Wahler, D. & Reymond, J. L. (2001). Novel methods for biocatalyst screening. *Curr Opin Chem Biol* **5**(2), 152-158.
- Pedersen, H., Holder, S., Sutherlin, D. P., Schwitter, U., King, D. S. & Schultz, P.
   G. (1998). A method for directed evolution and functional cloning of enzymes.
   Proc Natl Acad Sci U S A 95(18), 10523-10528.

- 32. Demartis, S., Huber, A., Viti, F., Lozzi, L., Giovannoni, L., Neri, P., Winter, G. & Neri, D. (1999). A strategy for the isolation of catalytic activities from repertoires of enzymes displayed on phage. *J Mol Biol* **286**(2), 617-633.
- 33. Jestin, J.-L., Kristenson, P. & Winter, G. (1999). A method for the selection of catalytic activity using phage display and proximity coupling. *Angew. Chem. Int. Ed.* **38**(8), 1124-1127.
- 34. Atwell, S. & Wells, J. A. (1999). Selection for improved subtiligases by phage display. *Proc Natl Acad Sci U S A* **96**(17), 9497-9502.
- 35. Firestine, S. M., Salinas, F., Nixon, A. E., Baker, S. J. & Benkovic, S. J. (2000). Using an AraC-based three-hybrid system to detect biocatalysts in vivo. *Nat Biotechnol* **18**(5), 544-547.
- Olsen, M. J., Stephens, D., Griffiths, D., Daugherty, P., Georgiou, G. & Iverson, B.
   L. (2000). Function-based isolation of novel enzymes from a large library. *Nat Biotechnol* 18(10), 1071-1074.
- 37. Daugherty, P. S., Iverson, B. L. & Georgiou, G. (2000). Flow cytometric screening of cell-based libraries. *J Immunol Methods* **243**(1-2), 211-227.
- 38. Voigt, C. A., Kauffman, S. & Wang, Z. G. (2000). Rational evolutionary design: the theory of in vitro protein evolution. *Adv Protein Chem* **55**, 79-160.
- 39. Aita, T. & Husimi, Y. (1996). Fitness spectrum among random mutants on Mt. Fujitype fitness landscape. *J Theor Biol* **182**(4), 469-485.
- 40. Aita, T., Uchiyama, H., Inaoka, T., Nakajima, M., Kokubo, T. & Husimi, Y. (2000).

  Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type

- landscape: application to prolyl endopeptidase and thermolysin. *Biopolymers* **54**(1), 64-79.
- Matsuura, T., Yomo, T., Trakulnaleamsai, S., Ohashi, Y., Yamamoto, K. & Urabe,
   I. (1998). Nonadditivity of mutational effects on the properties of catalase I and its
   application to efficient directed evolution. *Protein Eng* 11(9), 789-795.
- 42. Aita, T. & Husimi, Y. (2000). Theory of evolutionary molecular engineering through simultaneous accumulation of advantageous mutations. *J Theor Biol* **207**(4), 543-556.
- 43. Spiller, B., Gershenson, A., Arnold, F. H. & Stevens, R. C. (1999). A structural view of evolutionary divergence. *Proc Natl Acad Sci U S A* **96**(22), 12305-12310.
- 44. Yano, T., Oue, S. & Kagamiyama, H. (1998). Directed evolution of an aspartate aminotransferase with new substrate specificities. *Proc Natl Acad Sci U S A* **95**(10), 5511-5515.
- 45. Oue, S., Okamoto, A., Yano, T. & Kagamiyama, H. (1999). Redesigning the substrate specificity of an enzyme by cumulative effects of the mutations of non-active site residues. *J Biol Chem* **274**(4), 2344-2349.
- 46. Skandalis, A. & Loeb, L. A. (2001). Enzymatic properties of rat DNA polymerase beta mutants obtained by randomized mutagenesis. *Nucleic Acids Res* **29**(11), 2418-2426.
- 47. Patel, P. H. & Loeb, L. A. (2000). Multiple amino acid substitutions allow DNA polymerases to synthesize RNA. *J Biol Chem* **275**(51), 40266-40272.

- 48. Miyazaki, K. & Arnold, F. H. (1999). Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function. *J Mol Evol* **49**(6), 716-720.
- 49. 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**(7), 3778-3783.
- 50. Kono, H. & Saven, J. G. (2001). Statistical theory for protein combinatorial libraries. Packing interactions, backbone flexibility, and the sequence variability of a main-chain structure. *J Mol Biol* **306**(3), 607-628.
- 51. Fletcher, M. C., Kuderova, A., Cygler, M. & Lee, J. S. (1998). Creation of a ribonuclease abzyme through site-directed mutagenesis. *Nat Biotechnol* **16**(11), 1065-1067.

# **Chapter 3**

**Enzyme-like Proteins by Computational Design**\*

<sup>\*</sup>Adapted from Bolon D. N.; Mayo S. L. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, 98, 14274-14279.

# **Summary**

We report the development and initial experimental validation of a computational design procedure aimed at generating enzyme-like protein catalysts called "protozymes." Our design approach utilizes a "compute and build" strategy that is based on the physical/chemical principles governing protein stability and catalytic mechanism. Using the catalytically inert 108 residue E. coli thioredoxin as a scaffold, the histidine mediated nucleophilic hydrolysis of p-nitrophenyl acetate as a model reaction, and the ORBIT protein design software to compute sequences, an active site scan identified two promising catalytic positions and surrounding active site mutations required for substrate binding. Experimentally, both candidate protozymes demonstrated catalytic activity significantly above background. One of the proteins, PZD2, displayed "burst" phase kinetics at high substrate concentrations, consistent with the formation of a stable enzyme The kinetic parameters of PZD2 are comparable to early catalytic intermediate. antibodies. But, unlike catalytic antibody design, our design procedure is independent of fold, suggesting a possible mechanism for examining the relationships between protein fold and the evolvability of protein function.

### Introduction

A prominent goal of protein design is the generation of proteins with novel functions, including the catalytic rate enhancement of chemical reactions at which natural enzymes are so efficient. The ability to design an enzyme to perform a given chemical reaction has considerable practical application for industry and medicine, particularly for the synthesis of pharmaceuticals. Significant progress has been made at enhancing the catalytic properties of existing enzymes through directed evolution. In contrast, the design of proteins with novel catalytic properties has met with relatively limited success. We present here a general computational approach for the design of enzyme-like proteins with novel catalytic activities.

The use of transition state analogs as haptens to elicit catalytic antibodies has been the most successful technique to date for generating novel protein catalysts.<sup>5</sup> Natural enzymes combine transition state stabilization with precisely oriented catalytic side chains. While a reactive hapten has been used in the generation of an antibody with a powerful nucleophile at the active site,<sup>6</sup> current catalytic antibody technology does not efficiently select for both catalytic side chains and tight non-covalent affinity in the same molecule. The relationship between the general backbone fold of an enzyme and its catalytic properties is not well understood. This is particularly relevant to catalytic antibodies that are currently constrained to the antibody fold and which have yet to show catalytic activity on par with natural enzymes.

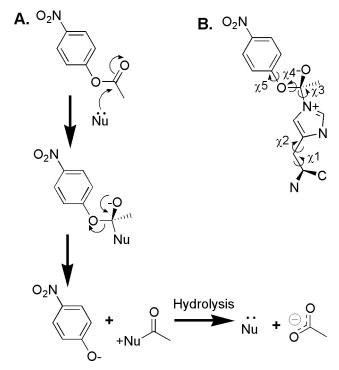
Rational, non-computational design efforts have recently succeeded at altering the catalytic reactions of two different enzymes.<sup>7,8</sup> Cyclophilin, a cis-trans isomerase of X-Pro peptide bonds was engineered into an endopeptidase by grafting a triad of catalytic

residues commonly found in serine proteases at the binding cleft.<sup>7</sup> In a more complicated design effort, Fersht and co-workers<sup>8</sup> were able to convert indoleglycerol-phosphate synthase into phosphoribosylanthranilate isomerase. The naturally occurring versions of these enzymes share a similar fold and catalyze consecutive reactions in the tryptophan biosynthetic pathway with the product of the isomerase being the substrate for the synthase. Based on a comparison of the crystal structures of both natural enzymes, loops were altered in the synthase to resemble the target isomerase, followed by directed evolution. These two successful design efforts both utilized starting proteins with substantial substrate affinity and altered existing active site residues in order to catalyze new reactions.

Compared to altering substrate specificity or catalytic mechanism, one of the most significant challenges of enzyme design utilizes a starting protein scaffold devoid of substrate binding and catalytic activities and introduces residues that convert the scaffold into an enzyme. Computational techniques have been used to design novel metal binding sites into proteins. By leaving one of the primary coordination spheres of the metal un-ligated by the protein, nascent metalloenzymes with a variety of oxygen redox chemistries have been generated. The general design of metalloenzymes to react specifically with more complicated organic molecules has not been demonstrated.

Numerous studies of natural proteins have revealed the basic principles of enzymatic catalysis, including proximity and orientation of substrate molecules, transition state stabilization, acid-base catalysis, and covalent catalysis.<sup>15</sup> The continued improvement in computer speed combined with the development and improvement of computational protein design algorithms<sup>16-18</sup> and forcefields<sup>19</sup> makes this an opportune

time to apply our understanding of enzyme mechanism to the rational, computational design of enzymes with novel functions. We model a high-energy state along a specified reaction pathway including the interaction of at least one catalytic residue with the high-energy state. The computational techniques described below, which have been incorporated into the ORBIT (Optimization of Rotamers By Iterative Techniques) protein design software, <sup>16</sup> identify catalytic sites within a protein structure as well as mutations necessary to accommodate substrate binding (that is, to build the active site).



**Figure 3.1** A) Nucleophile mediated catalysis of PNPA hydrolysis. B) High-energy state structure used in the computational active site scan. Labeled dihedral angles were varied as indicated in order to generate the set of high-energy state rotamers used in the design calculations.

The histidine mediated nucleophilic hydrolysis of p-nitrophenyl acetate (PNPA) into p-nitrophenol (PNP) and acetate (Fig. 3.1A) was selected as the target reaction. This reaction has a moderate activation barrier in pH neutral aqueous solution and is similar to the ester and carbonate hydrolysis reactions studied by the first catalytic antibodies.  $^{20,21}$ 

Nucleophile catalyzed hydrolysis is commonly achieved using a catalytic serine or cysteine in natural proteases. Histidine is often used as a general acid or base, but occasionally histidine acts as a nucleophile in phosphate transfer reactions. While histidine is not as powerful a nucleophile as activated serine or cysteine, it is sufficient to catalyze PNPA hydrolysis due to the moderate activation barrier of this reaction. The second order hydrolysis of esters by histidine side chains in a peptide based coiled coil system has been analyzed.<sup>22</sup> The peptides did not display substrate binding affinity, but the pKa and pH rate profile of reactivity towards PNPA could be modified by varying the local environment of the histidine<sup>23</sup> and in one variant, hydrogen bonding to the transition state appeared to provide second order rate enhancement.<sup>24</sup> For nucleophilic catalysis to be effective, the acylated enzyme intermediate must be unstable relative to substrate in buffer in order to increase the rate of product formation over the uncatalyzed buffer rate. An acyl-histidine intermediate would be less stable than either an acyl-serine or an acylcysteine intermediate. For these reasons, we modeled histidine as the nucleophile in our design.

#### **Materials and Methods**

#### Computational

For the histidine-PNPA high-energy state rotamers, a backbone independent rotamer library was generated that included nucleophilic attack by both the N $\delta$  and N $\epsilon$  atoms of histidine and attack on both enantiotopic faces of PNPA. The  $\chi 1$  and  $\chi 2$  dihedral angles were based on histidine dihedral angles in a survey of protein structures <sup>25</sup> and were expanded by  $\pm 1$  standard deviation from the reported values. Dihedral angles

for  $\chi 3$  (+/-30, +/-90, and +/-150 degrees),  $\chi 4$  (+/-60, and 180 degrees), and  $\chi 5$  (+/- 90 degrees) were based on canonical values for the relevant central atom hybridizations as described in the DREIDING force field.<sup>26</sup> Bond lengths and angles as well as additional dihedral angles were optimized using the DREIDING force field.<sup>26</sup> All other side chains were modeled using a backbone dependent rotamer library<sup>25</sup> that was expanded about  $\chi 1$  and  $\chi 2$  dihedral angles for aromatic residues, expanded about  $\chi 1$  dihedral angles for aliphatic residues, and unexpanded for polar residues as previously described.<sup>27</sup>

Table 3.1 Partial charges on computed high-energy state				
Nucleophilic attack by NE		Nucleophilic attack by Nδ		
Νδ1	-0.3	Ne2	-0.3	
Ηδ1	+0.5	Ηε2	+0.5	
Νε2	+0.6	Νδ1	+0.6	
Сζ	+0.4	Ce3	+0.4	
Οη1	-0.4	Οζ1	-0.4	
Οη2	-0.9	+Οζ2	-0.9	
Cı	+0.1	Сη	+0.1	
Сμ	+0.2	$C\lambda$	+0.2	
Nν	+0.6	Νμ	+0.6	
Oo1	-0.4	Ov1	-0.4	
Oo2	-0.4	Ov2	-0.4	

The calculations used an energy function based on the DREIDING force field.<sup>26</sup> A Lennard-Jones 12-6 potential with van der Waals radii scaled by 0.9 was applied to all side chain/backbone and side chain/side chain interactions as previously described.<sup>28</sup> An explicit hydrogen bond potential was applied to all groups that contained hydrogen bond donors and/or acceptors (including the high-energy state complex). The 12-10 hydrogen bond potential used a well depth of 8.0 kcal/mol, an equilibrium donor/acceptor distance of 2.8 Å, and a hybridization dependent angle term as previously described.<sup>29</sup> Electrostatic energies were computed using a distance dependent dielectric constant of

40r with partial atomic charges as previously described.<sup>29</sup> Partial atomic charges for the high-energy state rotamers were generated using the same bond polarization scheme used for the other amino acids and included a net +1 charge on the histidine moiety and a net -1 charge on the high-energy state substrate (Table 3.1). A surface area based atomic solvation potential was selectively applied to the high-energy state rotamers in order to favor interaction of the substrate with the protein (i.e., only surface area burial of atoms on the high-energy state rotamers were considered). This selective application of the solvation potential, which differs from our previous work, was required in order to reduce the tendency of generating predominately wild-type amino acid sequence solutions. Surface areas for the solvation potential were calculated using the Connolly algorithm<sup>30</sup> and the Lee and Richards definition of solvent accessible surface area<sup>31</sup> with unscaled atomic radii and an add-on radius of 1.4 Å. A hydrophobic burial benefit of 48 cal/mol/ $\mathring{A}^2$  and a hydrophobic exposure penalty of 76.8 cal/mol/ $\mathring{A}^2$  were used with the two-body decomposition method of Street and Mayo.<sup>32</sup> The solvation potential was not applied to polar atoms.

The hydrophobic solvent accessible surface area of substrate atoms was used to evaluate recognition. For this assessment, solvent accessible surface area was computed using the Connolly algorithm<sup>30</sup> and the Lee and Richards definition of solvent accessible surface area<sup>31</sup> with an add-on radius of 1.4 Å. The solvent accessible surface area of the substrate was determined using all atoms in the folded structure as well as using an unfolded reference that included only the high-energy state rotamer and the local backbone. Active site designs were ranked by the fraction accessible surface area defined as the ratio of hydrophobic accessible surface area in the folded structure and the

hydrophobic accessible surface area in the unfolded reference. Computed designs with zero total solvent accessible surface area for the substrates were considered inaccessible to substrate and were eliminated from further consideration.

## Experimental

The background mutation D26I was included in both PZD1 and PZD2. D26I was predicted by ORBIT in an independent calculation and results in increased thermodynamic stability similar to the previously reported D26A protein.<sup>33</sup> Position 26 is distal in space to the designed active sites in both PZD1 and PZD2.

The genes for PZD1 and PZD2 were constructed by site directed mutagenesis using the wild type thioredoxin gene (Invitrogen) cloned into PET-11A. Protein expression was induced with 0.5 mM IPTG from BL21(DE3) cells grown to mid log phase. Cells were lysed by sonication and pelleted twice at 20,000g for 30 minutes. The soluble fraction was brought to 60% acetonitrile, pelleted and rotory evaporated to half volume. For initial studies, purification was accomplished by reversed phase high performance liquid chromatography. For subsequent studies with PZD2, all protein samples were additionally purified by ion exchange and size exclusion chromatography to a purity of >99% as judged by silver stained PAGE.

PZD2 was dialyzed extensively against 10 mM sodium phosphate buffer at pH 6.95. Kinetic experiments at 22 °C were started by the addition of substrate dissolved in acetonitrile to buffer solution with and without PZD2 (final protein concentration of 4 μM). Protein concentration was determined by UV absorbance in 6 M guanidinium hydrochloride assuming an extinction coefficient of 12400 M<sup>-1</sup>cm<sup>-1</sup> at 280 nm. Product concentration was determined by the change in absorbance at 400.5 nm assuming an

extinction coefficient for deprotonated PNP of 19700  $M^{-1}cm^{-1}$ . Final acetonitrile concentration was 1% for all experiments. The steady state rate of hydrolysis by PZD2 was corrected for the buffer rate. Burst phase hydrolysis assays were performed at a substrate concentration of 1.6 mM. Protein concentration was 4  $\mu$ M, with the exception of wild type thioredoxin. Wild type thioredoxin was assayed at 50  $\mu$ M to improve signal to noise and extrapolated to a protein concentration of 4  $\mu$ M. The dead time for this experiment was approximately 30 seconds.

The +42 mass unit species detected in the trapping experiment is a result of replacement of a hydrogen (-1) with an acetyl group (+43). 100 µM PZD2 in 10 mM TRIS at pH 7.0 was reacted with 1.6 mM PNPA to steady state conditions and a mass spectrum acquired. The same protein solution without PNPA was used as a control. Burst phase kinetics in this buffer system yielded essentially identical results as in phosphate buffer.

## **Results and Discussion**

# Computational design

An important step in the design of a protein catalyst is locating the substrate relative to the protein scaffold. The location of binding sites for small molecule/protein docking is commonly evaluated by holding the protein fixed and performing a rotation/translation search of the small molecule using a grid based method.<sup>34</sup> The design of a protein catalyst requires sequence changes in the protein scaffold and a grid based approach would require optimization of the protein sequence for each substrate position making it computationally intractable. Instead we took advantage of the observation that

the vast majority of natural enzymes contain one or more side chains in highly specific geometric relation to substrate. In the case of PNPA hydrolysis catalyzed by histidine, the histidine side chain should be positioned with a nitrogen atom approaching the carbonyl carbon adjacent to the scissile bond. In a procedure reminiscent of transition state analog synthesis for catalytic antibody design, we modeled a high-energy state of histidine catalyzed PNPA hydrolysis as a series of side chain rotamers (Fig. 3.1B). Modeling a high-energy state and optimizing the surrounding protein sequence for binding to the high-energy state may lead to a reduction in activation energy for the reaction and catalytic rate enhancement.<sup>35</sup> Substrate affinity is explicitly included in the design because chemical similarity exists between the high-energy state and the substrate.

Table 3.2 Top ten designs from active site scan				
	Catalytic	Fraction		
	His	Hydrophobic		
Design	Position	Exposure	Active Site Mutations	
PZD1	12	0.11	F12H Y70A	
PZD2	17	0.15	F12A L17H Y70A	
PZD3	86	0.29	V86H I38A L42A L99A	
PZD4	72	0.34	I72H L79A	
PZD5	66	0.34	T66H F12A Y70A	
PZD6	6	0.36	None	
PZD7	39	0.37	A39H K57A	
PZD8	91	0.39	V91H T77A	
PZD9	49	0.39	Y49H K52A	
PZD10	77	0.43	T77H L79A T89A	

Ranked based on hydrophobic surface area burial of substrate atoms in the high-energy state complex. The top two designs, PZD1 and PZD2, were experimentally tested for catalytic activity.

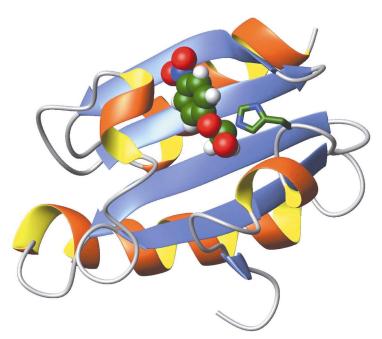
The well-studied protein  $E.\ coli$  thioredoxin<sup>36</sup> was selected as a scaffold due to its favorable expression properties, thermodynamic stability,<sup>37</sup> and successful history in computational design.<sup>14</sup> Naturally occurring thioredoxin can supply the reducing power for ribonucleotide reductase via its disulfide bond, but is essentially catalytically inert

with respect to PNPA binding and hydrolysis. The high thermodynamic stability of thioredoxin suggests that it can tolerate destabilizing mutations that are required to build an active site.<sup>38</sup> An active site scan was performed in order to identify favorable positions for the catalytic histidine as well as mutations necessary for substrate recognition and binding. In separate calculations using the ORBIT computational protein design programs, each position in the protein structure of thioredoxin was modeled using the set of high-energy state rotamers (Fig. 3.1B). This strategy systematically limits the search to the relevant phase space where histidine and the substrate are properly positioned to undergo chemistry. All other positions in the protein were allowed to chose (with proper consideration for rotamer flexibility) between their wild type identity and alanine in order to accommodate the substrate and to build the active site. After computing the optimal solution using algorithms based on the Dead-End Elimination theorem, <sup>39,40</sup> positions that changed to alanine can be subsequently allowed to change identity to other amino acids in order to form better interactions with the high-energy state rotamer. The results reported below are based on the first stage calculation that only allows mutation to alanine. The initial limited combinatorial complexity approach provides for tractable calculations and is expected to generate enzyme-like proteins that we call protozymes. As computer power and computational methods become more sophisticated, the need for a two-stage calculation will diminish. All 94 non-glycine, non-proline positions, a combinatorial complexity of approximately 10<sup>26</sup> amino acid sequences that corresponded to 10<sup>101</sup> rotamer sequences, were scanned in approximately two days on 14 195 MHz R10000 processors (Silicon Graphics) running in parallel. Designs (resulting from the limited combinatorial complexity calculations) with the

catalytic histidine at different positions in thioredoxin were ranked based on recognition of the high-energy state rotamer, using hydrophobic solvent accessible surface area of substrate atoms as a metric, as well as the simple requirement that the total solvent accessible surface area of substrate atoms be greater than zero in order to eliminate designs where the active site would not be accessible to solvent (Table 3.2).

# Experimental validation

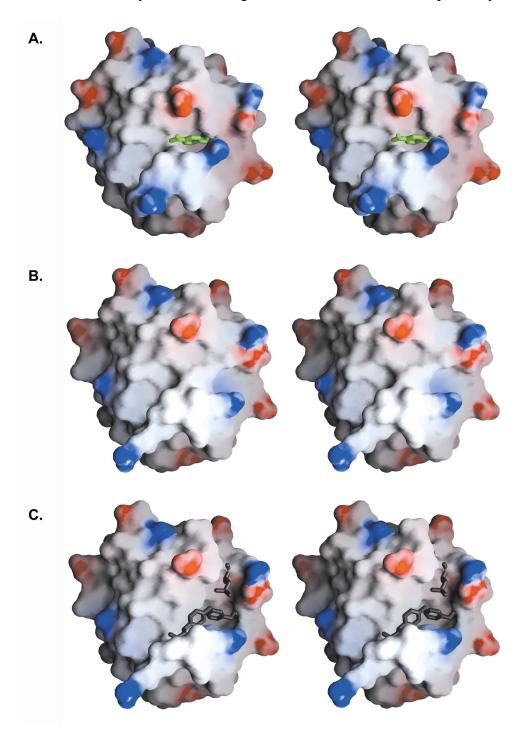
The top two sequences from the limited combinatorial complexity active site scan were selected for experimental analysis. Protozyme design 1 (PZD1) contains two mutations required to introduce the catalytic histidine and to build the active site (F12H and Y70A) while PZD2 contains three mutations (F12A, L17H, and Y70A).



**Figure 3.2** Computational design of PZD2. Ribbon diagram<sup>41</sup> illustrating the Nε of the catalytic histidine attacking the substrate molecule.

Experimentally, both proteins demonstrated catalytic hydrolysis of PNPA at a rate significantly above background. Based on preliminary kinetic experiments, PZD2

selected for further analysis. The designed structure of PZD2 computed by ORBIT



**Figure 3.3** Molecular surfaces <sup>42</sup> focusing on the active site of PZD2 with substrate atoms in green (A) and the corresponding region in the x-ray crystal structure <sup>43</sup> of the wild type scaffold (B & C). An active site cleft is present in the design of PZD2 that is largely filled in the wild type structure. Wild type residues that were mutated to create the active site are shown in panel C (F12, L17, and Y70). In the design of PZD2, all side chains were allowed to change rotamers, resulting in a slightly different surface compared to that of the wild type protein.

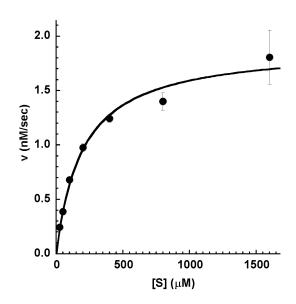
shows the substrate atoms in a cleft created above the  $\beta$  sheet and between two  $\alpha$  helices (Fig. 3.2). This cleft (Fig. 3.3A) is not present in the wild type thioredoxin x-ray structure<sup>43</sup> (Fig. 3.3B), indicating that the space creating mutations (F12A and Y70A) are necessary to create the putative substrate-binding site (Fig. 3.3C).

$$E + S \xrightarrow{K_S} ES \xrightarrow{k_2} P1 + EI \xrightarrow{k_3} E + P2$$

$$V = \frac{k_{cat}[E][S]}{K_m + [S]} ; k_{cat} = \frac{k_2 k_3}{k_2 + k_3} ; K_m = \frac{K_s k_3}{k_2 + k_3}$$

Figure 3.4 Kinetic model used to analyze the activity of PZD2.

The rate of PNPA hydrolysis by PZD2 at 22 °C in 10 mM sodium phosphate buffer at pH 6.95 was experimentally determined over a range of substrate concentrations using standard Briggs-Haldane steady state treatment (Fig. 3.4). The reaction velocity of



**Figure 3.5** Velocity versus substrate concentration for the hydrolysis of PNPA by PZD2.

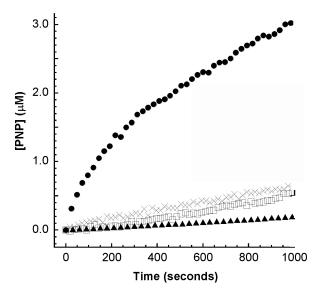
PZD2 demonstrated saturation kinetics with respect to increasing substrate concentration (Fig. 3.5) indicating that the molecule is acting as an enzyme-like protein with both

substrate affinity and catalytic rate enhancement. Hofstee analysis of the data gives a  $K_{\rm m}$  of  $170 \pm 20~\mu{\rm M}$ ,  $k_{\rm cat}$  of  $4.6 \pm 0.2 \times 10^{-4}~{\rm sec}^{-1}$ , and  $k_{\rm cat}/k_{\rm uncat}$  of 180 where  $k_{uncat}$  (2.5 × 10<sup>-6</sup> sec<sup>-1</sup>) is the rate of hydrolysis under the same conditions. The kinetics of PZD2 are comparable to those of the first catalytic antibodies:  $K_{\rm m}$  of 208  $\mu{\rm M}$  and  $k_{\rm cat}/k_{\rm uncat}$  of 770 for MOPC167,<sup>20</sup> and  $K_{\rm m}$  of 1.9  $\mu{\rm M}$  and  $k_{\rm cat}/k_{\rm uncat}$  of 960 for 6D4.<sup>21</sup>

De novo designed peptides have recently been designed which catalyze pnitrophenyl acetate hydrolysis by a second order mechanism (i.e., without substrate saturation). The second order rates for these peptides were determined under different conditions than those used here making direct comparisons difficult. At pH 5.1, rates of 0.29 M<sup>-1</sup>sec<sup>-1</sup> and 0.056 M<sup>-1</sup>sec<sup>-1</sup> have been reported.<sup>22,24</sup> Extrapolation based on the pH rate profiles<sup>22</sup> reported for the hydrolysis of similar activated esters suggests rates of approximately 0.7 and 0.2 M<sup>-1</sup>sec<sup>-1</sup> for p-nitrophenyl acetate hydrolysis at pH 7, respectively. The second order rate of hydrolysis by 4-methyl imidazole (determined under the conditions used in the analysis of PZD2) is 0.11 M<sup>-1</sup>sec<sup>-1</sup>. Computing a second order rate of hydrolysis for PZD2 as  $k_{cat}/K_m$  gives 2.7 M<sup>-1</sup>sec<sup>-1</sup>, which is a factor of 25 better than catalysis by 4-methyl imidazole and about an order of magnitude better than the peptide systems. It is extremely important to note, however, that the lack of substrate saturation in the second order systems complicates this analysis since substrate affinity was determined for PZD2. Substrate affinity is central to enzyme catalysis and clearly will be important in the design of protein-based catalysts for more complicated reactions.

At high substrate concentrations, kinetics with an initial "burst" phase are a common feature of natural enzymes and are a consequence of a kinetic bottleneck on the reaction pathway. For nucleophilic hydrolysis of PNPA, the rate of PNP formation

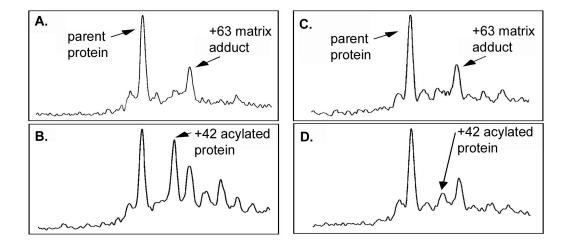
should decrease and plateau as the acyl-enzyme intermediate concentration approaches steady state. PZD2 displays burst phase kinetics at high substrate concentration (Fig. 3.6) consistent with the formation of an enzyme intermediate. The observed burst phase kinetics could also be a result of slow product release; however, this is unlikely given the kinetic parameters of PZD2. Based on the burst phase data, we estimate  $K_s = 1$  mM,  $k_2 = 3 \times 10^{-3}$  sec<sup>-1</sup>, and  $k_3 = 8 \times 10^{-4}$  sec<sup>-1</sup>. Analysis of the magnitude of the burst indicates that approximately 30% of PZD2 is in an active state. Using the burst magnitude as a metric of the enzyme concentration indicates a  $k_{cat}$  of  $1.5 \pm 0.1 \times 10^{-3}$  sec<sup>-1</sup>, approximately three times larger than the kinetic analysis using enzyme concentrations determined from protein UV absorbance.



**Figure 3.6** Buffer corrected hydrolysis of PNPA by PZD2 (•), PZD2 H17A (□), wild type thioredoxin (▲), and wild type L17H/D26I (×). Data are shown for high substrate concentration and equivalent low protein concentration.

Wild type thioredoxin is essentially inactive, but does show weak second order PNPA hydrolysis consistent with its single surface exposed histidine at position 6. Mutation of the designed catalytic histidine to alanine in PZD2 (H17A) results in a

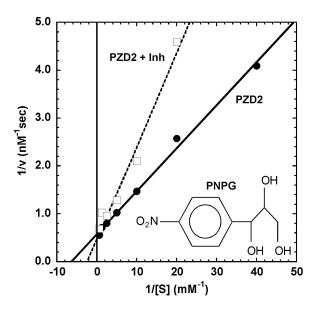
protein with catalytic activity similar to wild type thioredoxin indicating that the designed catalytic histidine at position 17 is necessary for the enzyme-like activity in PZD2 (Fig. 3.6). Mutation of the two other active site residues in PZD2 back to their wild type identities (A12F and A70Y) also results in a protein with activity similar to wild type. Additionally, at pH 5.7 the activity of PZD2 is almost entirely eliminated, consistent with protonation of the catalytic histidine (data not shown). The kinetic and mutational evidence strongly indicate that PZD2 is working as designed with H17 acting as a catalytic nucleophile and the space creating mutations (F12A and Y70A) forming a binding site for substrate.



**Figure 3.7** Trapping of an acylated intermediate by mass spectrometry: A) PZD2, B) PZD2 reacted with substrate, C) PZD2 H17A, D) PZD2 H17A reacted with substrate. A large increase in the population of a +42 species occurs upon reaction of PZD2 with substrate indicating the buildup of an acyl-enzyme intermediate. This +42 species is dramatically reduced for PZD2 H17A where the designed catalytic histidine was mutated to alanine. A small increase in the population of a +42 species is detected in PZD2 H17A upon reaction with substrate and is likely the result of acylation at the single surface exposed histidine at position 6. Consistent with this analysis, a small increase in the population of a double acetylated +84 product is detected upon reaction of PZD2 with substrate. A copper matrix adduct (+63) is present in all spectra.

According to its design, reaction of PZD2 with PNPA should yield an acylhistidine enzyme intermediate, increasing the population of a +42 mass unit species compared to free protein. Mass spectrometry of PZD2 clearly indicates increased

population of an acylated species upon reaction with PNPA (Fig. 3.7A&B). Acylation due to reaction with PNPA is essentially absent for the H17A mutant of PZD2 (Fig. 3.7C&D), indicating that acylation is dependent on H17. From the data it is not possible to distinguish between direct acylation of H17, or H17 mediated (acid/base or acylatransfer) acylation of a different residue.



**Figure 3.8** Lineweaver-Burk analysis of PZD2 catalyzed PNPA hydrolysis in the presence ( $\square$ ) and absence ( $\bullet$ ) of 10 mM PNPG.

Naturally occurring enzymes are frequently inhibited in a competitive manner by inert compounds resembling the substrate molecule. With this in mind, we assayed the inhibitory effects of p-nitrophenyl glycerol (PNPG), a highly soluble compound with structural similarity to PNPA. A double reciprocal analysis of PZD2 catalyzed PNPA hydrolysis in the presence and absence of inhibitor shows the hallmark features of competitive inhibition (Fig. 3.8) implying that PNPG is able to bind in the active site and block PNPA binding. A Dixon analysis, varying PNPG concentration at a fixed substrate concentration of 200  $\mu$ M yielded a  $K_i$  of 20 mM. The decreased binding affinity of PNPG relative to PNPA may be due to burial of polar hydroxyl groups against

hydrophobic regions in the active site or differences in the steric requirements for PNPG and PNPA. *p*-nitrophenyl phosphate was also tested, but it did not show detectable inhibition at concentrations up to 20 mM, suggesting that PZD2 binds preferentially to uncharged nitrophenyl molecules.

In order to further test the substrate specificity of PZD2 we assayed its kinetics with p-nitrophenyl propionate (PNPP), which has an additional methylene group compared to PNPA. Hofstee analysis indicates within error an identical  $k_{\rm cat}/k_{\rm uncat}$  and a  $K_{\rm m}$  of 110  $\pm$  20  $\mu$ M for PNPP, compared to 170  $\pm$  20  $\mu$ M for PNPA. Inspection of the designed structure of PZD2 with bound PNPA shows that space exists to accommodate the additional methylene of PNPP. Both favorable van der Waals interactions and hydrophobic burial afforded by the relatively open active site likely explain the slightly increased affinity of PZD2 for PNPP compared to PNPA.

# **Conclusions**

The computational design method presented here successfully identified enzyme-like active sites within a protein scaffold essentially devoid of relevent catalytic activity. While the reaction catalyzed (PNPA hydrolysis) and the level of activity of PZD2 are likely to be insufficient for practical applications, clear paths for improvement exist including further computational optimization and experimental directed evolution. Computational optimization of the identity of residues proximal in space to the high-energy state complex could result in improved shape and charge complementarity with the high-energy state, improving both substrate affinity and catalytic rate enhancement. Directed evolution, combining mutagenesis and recombination with screening and/or

selection, has been an effective method for increasing catalytic activity. The use of computational modeling to focus directed evolution experiments promises to increase the effectiveness of directed evolution methods. For the creation and optimization of novel protein catalysts, computational protein design and experimental directed evolution complement each other well. Use of the design methods presented here to generate proteins with novel catalytic activity and directed evolution to further enhance this activity is a promising path to the production of novel high efficiency protein catalysts.

The relationship between protein fold and function is a long standing question in enzymology. For example, from numerous structural studies, insight has been gained into the properties of the  $\alpha/\beta$  barrel fold that have facilitated its evolution to catalyze many different chemical reactions. With the methods presented here we may now address the evolvability of essentially any fold through the computational design of novel enzymatic function followed by directed evolution. The comparison of an antibody fold to an  $\alpha/\beta$  barrel fold, for example, may prove particularly insightful. Studies of this type may advance our understanding of the functional potential of protein folds, provide insight into natural evolution, and aid in the selection of scaffolds for future design efforts.

#### References

- 1. Liese, A. & Filho, M. V. (1999). Production of fine chemicals using biocatalysis. *Curr Opin Biotechnol* **10**(6), 595-603.
- 2. Arnold, F. H. & Volkov, A. A. (1999). Directed evolution of biocatalysts. *Curr Opin Chem Biol* **3**(1), 54-59.
- 3. Corey, M. J. & Corey, E. (1996). On the failure of de novo-designed peptides as biocatalysts. *Proc Natl Acad Sci U S A* **93**(21), 11428-11434.
- 4. Petsko, G. A. (2000). Enzyme evolution. Design by necessity. *Nature* **403**, 606-607.
- 5. Hilvert, D. (2000). Critical analysis of antibody catalysis. *Annu Rev Biochem* **69**, 751-793.
- 6. Wagner, J., Lerner, R. A. & Barbas, C. F., 3rd. (1995). Efficient aldolase catalytic antibodies that use the enamine mechanism of natural enzymes. *Science* **270**, 1797-1800.
- 7. Quemeneur, E., Moutiez, M., Charbonnier, J. B. & Menez, A. (1998). Engineering cyclophilin into a proline-specific endopeptidase. *Nature* **391**, 301-304.
- 8. Altamirano, M. M., Blackburn, J. M., Aguayo, C. & Fersht, A. R. (2000). Directed evolution of new catalytic activity using the alpha/beta- barrel scaffold. *Nature* **403**, 617-622.
- 9. Nixon, A. E., Firestine, S. M., Salinas, F. G. & Benkovic, S. J. (1999). Rational design of a scytalone dehydratase-like enzyme using a structurally homologous protein scaffold. *Proc Natl Acad Sci U S A* **96**(7), 3568-3571.

- Hellinga, H. W. & Richards, F. M. (1991). Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with predefined geometry. *J Mol Biol* 222(3), 763-785.
- Robertson, D. E., Farid, R. S., Moser, C. C., Urbauer, J. L., Mulholland, S. E.,
   Pidikiti, R., Lear, J. D., Wand, A. J., DeGrado, W. F. & Dutton, P. L. (1994).
   Design and synthesis of multi-haem proteins. *Nature* 368, 425-432.
- 12. Klemba, M., Gardner, K. H., Marino, S., Clarke, N. D. & Regan, L. (1995). Novel metal-binding proteins by design. *Nat Struct Biol* **2**(5), 368-373.
- Pinto, A. L., Hellinga, H. W. & Caradonna, J. P. (1997). Construction of a catalytically active iron superoxide dismutase by rational protein design. *Proc Natl Acad Sci U S A* 94(11), 5562-5567.
- Benson, D. E., Wisz, M. S. & Hellinga, H. W. (2000). Rational design of nascent metalloenzymes. *Proc Natl Acad Sci U S A* 97(12), 6292-6297.
- 15. Fersht, A. (1985). *Enzyme structure and mechanism*. 2nd edit, Freeman, New York.
- 16. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science* **278**, 82-87.
- 17. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science* **282**, 1462-1467.
- 18. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Structure Fold Des* **7**(5), R105-109.
- Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr Opin Struct Biol* 9(4), 509-513.

- 20. Pollack, S. J., Jacobs, J. W. & Schultz, P. G. (1986). Selective chemical catalysis by an antibody. *Science* **234**, 1570-1573.
- Tramontano, A., Janda, K. D. & Lerner, R. A. (1986). Catalytic antibodies. *Science* 234, 1566-1570.
- 22. Broo, K. S., Brive, L., Ahlberg, P. & Baltzer, L. (1997). Catalysis of hydrolysis and transesterification reactions of p-nitrophenyl esters by a designed helix-loop-helix dimer. *J Am Chem Soc* **119**, 11362-11372.
- 23. Broo, K. S., Brive, L., Sott, R. S. & Baltzer, L. (1998). Site-selective control of the reactivity of surface-exposed histidine residues in designed four-helix-bundle catalysts. *Fold Des* **3**(4), 303-312.
- 24. Nilsson, J. & Baltzer, L. (2000). Reactive-site design in folded-polypeptide catalysts--the leaving group pKa of reactive esters sets the stage for cooperativity in nucleophilic and general-acid catalysis. *Chemistry* **6**(12), 2214-2220.
- 25. Dunbrack, R. L., Jr. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side- chain prediction. *J Mol Biol* **230**(2), 543-574.
- 26. Mayo, S. L., Olafson, B. D. & Goddard, W. A., 3rd. (1990). DREIDING: a generic force field for molecular simulations. *J Phys Chem* **94**, 8897-8909.
- 27. Dahiyat, B. I., Sarisky, C. A. & Mayo, S. L. (1997). De novo protein design: towards fully automated sequence selection. *J Mol Biol* **273**(4), 789-796.
- 28. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* **94**(19), 10172-10177.
- 29. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci* **6**(6), 1333-1337.

- Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids.
   Science 221, 709-713.
- 31. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**(3), 379-400.
- 32. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Fold Des* **3**(4), 253-258.
- 33. Gleason, F. K. (1992). Mutation of conserved residues in Escherichia coli thioredoxin: effects on stability and function. *Protein Sci* **1**(5), 609-616.
- 34. Wang, J., Kollman, P. A. & Kuntz, I. D. (1999). Flexible ligand docking: a multistep strategy approach. *Proteins* **36**(1), 1-19.
- 35. Pauling, L. (1948). Chemical achievement and hope for the future. Am Sci 36, 51.
- 36. Holmgren, A. (1985). Thioredoxin. *Annu Rev Biochem* **54**, 237-271.
- 37. Ladbury, J. E., Wynn, R., Hellinga, H. W. & Sturtevant, J. M. (1993). Stability of oxidized Escherichia coli thioredoxin and its dependence on protonation of the aspartic acid residue in the 26 position. *Biochemistry* **32**(29), 7526-7530.
- 38. Hellinga, H. W., Wynn, R. & Richards, F. M. (1992). The hydrophobic core of Escherichia coli thioredoxin shows a high tolerance to nonconservative single amino acid substitutions. *Biochemistry* **31**(45), 11203-11209.
- 39. Desmet, J., Maeyer, M. D., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539-542.

- 40. Pierce, N. A., Spriet, J. A., Desmet, J. & Mayo, S. L. (2000). Conformational splitting: a more powerful criterion for dead-end elimination. *J Comput Chem* **21**, 999-1009.
- 41. Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* **14**(1), 51-55, 29-32.
- 42. Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**(4), 281-296.
- 43. Katti, S. K., LeMaster, D. M. & Eklund, H. (1990). Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution. *J Mol Biol* **212**(1), 167-184.
- 44. Voigt, C. A., Mayo, S. L., Arnold, F. H. & Wang, Z. (in press). *Proc Natl Acad Sci USA*.
- 45. Reardon, D. & Farber, G. K. (1995). The structure and evolution of alpha/beta barrel proteins. *Faseb J* **9**(7), 497-503.

# Chapter 4

# Polar Residues in the Protein Core of $E.\ Coli$ Thioredoxin Are Important for Fold Specificity

The data for Figure 4.1 was compiled by Arthur Street.

<sup>\*</sup>Adapted from Bolon, D. N.; Mayo S. L. *Biochemistry* **2001**, 40, 10047-10053.

### Summary

Most globular proteins contain a core of hydrophobic residues that are inaccessible to solvent in the folded state. In general, polar residues in the core are thermodynamically unfavorable except when they are able to form intra-molecular hydrogen bonds. Compared to hydrophobic interactions, polar interactions are more directional in character and may aid in fold specificity. In a survey of 263 globular protein structures, we found a strong positive correlation between the number of polar residues at core positions and protein size. In order to probe the importance of buried polar residues, we experimentally tested the effects of hydrophobic mutations at the five polar core residues in E. coli thioredoxin. Proteins with single hydrophobic mutations (D26I, C32A, C35A, T66L, T77V) all have cooperative unfolding transitions similar to wild type (wt) as determined by chemical denaturation. Relative to wt, D26I is more stable while the other point mutants are less stable. The combined five-fold mutant protein (IAALV) is less stable than wt and has an unfolding transition that is substantially less cooperative than wt. NMR spectra as well as amide deuterium exchange indicate that IAALV is likely sampling a number of low energy structures in the folded state, suggesting that polar residues in the core are important for specifying a well-folded native structure.

### Introduction

The clustering of hydrophobic residues in protein cores is a general feature of globular proteins in their native state.<sup>1,2</sup> In the denatured state, these hydrophobic residues become exposed to solvent. The interaction of hydrophobic amino acids with with hydrophobic-hydrophobic interactions results thermodynamically unfavorable ordering of solvent atoms in the denatured state. Thus, the burial of hydrophobic surface area upon protein folding is expected to thermodynamically stabilize the native state relative to the unfolded state.<sup>3,4</sup> Experimental evidence abounds for this stabilization. For example, increased hydrophobic burial in the design of the β1 domain of protein G resulted in a molecule with dramatically improved thermostability,<sup>5</sup> while cavity forming mutations that reduced hydrophobic burial were destabilizing in T4 lysozyme.<sup>6</sup> Simple lattice models suggest that the clustering of solvent-averse side chains in the protein interior is sufficient to encode a unique compact chain conformation.<sup>7</sup> Experimentally, the restriction of core residues to hydrophobic identity has been successfully employed in the design of a 28 residue  $\beta\beta\alpha$  motif.<sup>8</sup>

The transfer of polar amino acids from water to hydrophobic solvent results in a loss of hydrogen bonding and is thermodynamically unfavorable. In the unfolded state, polar residues are largely solvated and form hydrogen bonds to water. As expected, polar residues that are buried in the folded state and fail to make intra-molecular hydrogen bonds are destabilizing relative to hydrophobic residues. Buried polar residues can reduce the free energy of folding by forming intra-molecular hydrogen bonds. However, this decrease in free energy can be offset by gains in free energy due to

desolvation<sup>14</sup> and the entropy loss associated with fixing sidechains.<sup>15</sup> To address this energetic balance, a number of experimental studies have measured the thermodynamic effect of replacing a side chain that forms an intra-molecular hydrogen bond in the folded state with a residue that cannot form a hydrogen bond.<sup>16-19</sup> While disagreement exists, interpretation of the experimental data indicates that an intra-molecular hydrogen bond can stabilize the native state on the order of one kcal/mol.<sup>20</sup>

Structural specificity, the capacity to form a single unique folded structure, has been well studied in coiled coil systems where experimental techniques exist for rapidly identifying parallel or anti-parallel orientation as well as oligomeric state. Coiled coils contain a heptad repeat denoted abcdefg where positions a and d form the interface between the helices, positions e and g are at the interface boundary and positions b, c and f are oriented towards solvent. Kim and co-workers systematically changed residues in the GCN4 homodimeric coiled coil into the Fos-Jun heterodimeric coiled coil. It was found that eight polar residues at the e and g positions from Fos-Jun in the GCN4 background specified heterodimer formation over homodimer formation.<sup>21</sup> hydrogen bonds between an asparagine side chain at an a position from each monomer are found in the homodimeric coiled coil domain of GCN4.<sup>22</sup> When this asparagine is replaced by valine<sup>23</sup> or aminobutyric acid,<sup>24</sup> a mixture of dimers and trimers is formed. Crystal structures of the aminobutyric acid mutant demonstrate that the hydrophobic core is able to pack efficiently in both the dimeric and trimeric state.<sup>24</sup> A parallel heterodimeric coiled coil dubbed "peptide velcro" was successfully designed with an asparagine from each monomer forming buried hydrogen bonds as in GCN4 as well as charged residues at the e and g positions that are complemented in the parallel structure

but not the anti-parallel structure.<sup>25</sup> When the buried asparagine in this coiled coil was replaced with leucine, the resulting molecule formed a heterotetramer that lacked a unique orientation of the helices.<sup>26</sup> In coiled coil systems, buried polar residues clearly play a role in structural specificity.

The role of polar residues in the core of globular proteins has not been as extensively studied. We analyzed the number of polar residues at core positions in 263 globular proteins from a previously published structural data set.<sup>27</sup> We find a strong positive correlation between the number of polar residues at core positions and protein size. Active site residues that are often buried in proteins are not prevalent enough to fully account for the number of polar residues observed at core positions suggesting that core polar residues play a structural role. To evaluate this possible structural role, the effect of polar to hydrophobic mutations was experimentally determined in a model system, the 108 residue protein *E. coli* thioredoxin. The ORBIT (Optimization of Rotamers By Iterative Techniques) protein design software<sup>8</sup> was used to predict hydrophobic mutations compatible with the wild type protein structure.<sup>28</sup> Thermodynamic as well as NMR data indicate that core polar residues in thioredoxin aid in distinguishing a single well-folded structure from alternative structures.

#### **Materials and Methods**

Analysis of polar residues at core positions in protein structures

The orientation of the  $C\alpha$ - $C\beta$  vector relative to a solvent-accessible surface computed with only the template  $C\alpha$  atoms was used to determine core positions as previously described.<sup>29</sup> Briefly, a position was classified as core if the distance from its

 $C\alpha$ , along its  $C\alpha$ - $C\beta$  vector, to the solvent accessible surface was greater than 5.0 Å, and if the shortest distance from its  $C\beta$  to the surface was greater than 2.0 Å. The following ten amino acids were considered polar: Arg, Asn, Asp, Cys, Gln, Glu, His, Lys, Ser, Thr.

### Computational modeling

Simulations were performed using coordinates from the 1.7 Å x-ray structure of thioredoxin.<sup>28</sup> The five polar core positions in wt thioredoxin (D26, C32, C35, T66, T77) were restricted to the following seven amino acids: Ala, Val, Leu, Ile, Phe, Tyr, and Trp. Side chains at other core positions in the protein were allowed to change geometry while retaining their wt identity. The remainder of the structure was held fixed during the optimization. The resulting combinatorial complexity is 10<sup>28</sup> possible rotamer (side chain identity and geometry) sequences. Computational details and potential functions are described in our previous work.<sup>8</sup> Similar calculations were also performed that optimized the identity of each polar core residue independently.

# Mutagenesis and protein purification

The genes for all variants were constructed by inverse PCR<sup>30</sup> using the gene for wt thioredoxin (Invitrogen) cloned into pET-11a (Novagen). Mutations were confirmed by DNA sequencing. BL21(DE3) hosts (Invitrogen) were used for protein expression. Cells were grown to mid log phase and induced with 0.5 mM IPTG (ICN) for three hours at 37 °C. Cells were lysed by sonication and centrifuged twice at 20,000×g for 30 minutes. Acetonitrile (EM Science) was added to the soluble fraction to a final concentration of 60%. After centrifugation, the supernatant was evaporated to half volume and purified by reverse phase high performance liquid chromatography using an acetonitrile/water gradient containing 0.1% trifluoroacetic acid (Applied Biosystems).

Protein identities were confirmed by mass spectrometry. Protein concentration was determined by UV absorbance in 7 M GdmCl (ICN) based on an extinction coefficient of 13,700 M<sup>-1</sup>cm<sup>-1</sup> at 280 nm.<sup>31</sup>

### CD analysis

CD data were collected on an Aviv 62DS spectrometer equipped with a thermoelectric unit and using a 1 cm path length cell. Protein samples were 5  $\mu$ M in 50 mM sodium phosphate at pH 7.0. Guanidinium chloride denaturations were monitored at 219 nm at 25 °C for comparison to previously published data.  $^{32}$   $\Delta$ G values, m values, and error estimates were obtained by fitting the denaturation data to a two-state transition as described previously.  $^{33}$  GdmCl concentration was measured by refractometry.

### Fluorescence analysis

Fluorescent measurements were made on an SLM 8000 spectrofluorimeter. Chemical denaturation of wild type and IAALV (D26I/C32A/C35A/T66L/T77V) were determined by fluorescence. Protein samples were at 5  $\mu$ M in 50 mM sodium phosphate at pH 7.0. Samples were excited at 280 nm and chemical denaturation was followed by emission at a wavelength of 340 nm.

### **DSC**

Protein samples at ~0.7 mg/ml were dialyzed against 50 mM sodium phosphate buffer at pH 7.0. DSC data were collected on a N-DSC II instrument from Calorimetric Sciences Corp.<sup>34</sup> Samples were degassed under vacuum for 15 minutes prior to scanning. Scans were performed at a rate of 1.0 °C/min. Data were analyzed using the program cpcalc (Applied Thermodynamics).

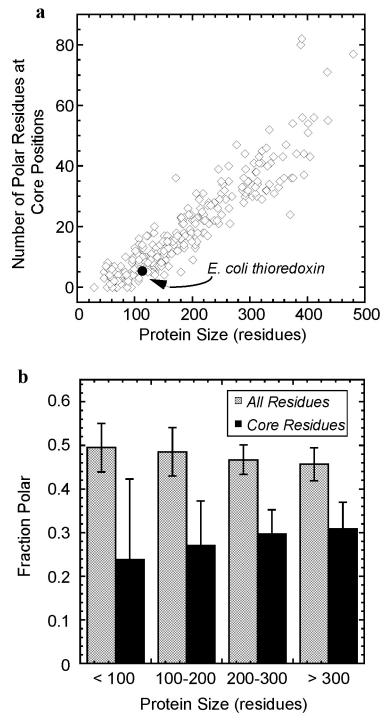
### NMR spectroscopy

Protein samples at 0.5 mM were prepared in 90:10 H<sub>2</sub>O:D<sub>2</sub>O buffered at pH\* 7.0 with 50 mM sodium phosphate. Spectra were acquired on a Varian Inova 600 MHz spectrometer at 25 °C. For hydrogen exchange studies, an NMR sample was prepared, the pH was adjusted to 7.0, and a spectrum was acquired to serve as an unexchanged reference. This sample was lyophilized, reconstituted in 99.9% D<sub>2</sub>O, and repetitive acquisition of spectra was begun immediately at a rate of 1 spectrum per 100 seconds. Data acquisition continued for ~20 hours, and then the sample was heated above its calorimetrically determined melting temperature for 15 minutes to fully exchange all labile protons. After cooling to 25 °C, a final spectrum was acquired to serve as the fully exchanged reference. The areas of all exchangeable amide peaks were normalized to that of a non-exchanging aliphatic proton peak.

#### Results

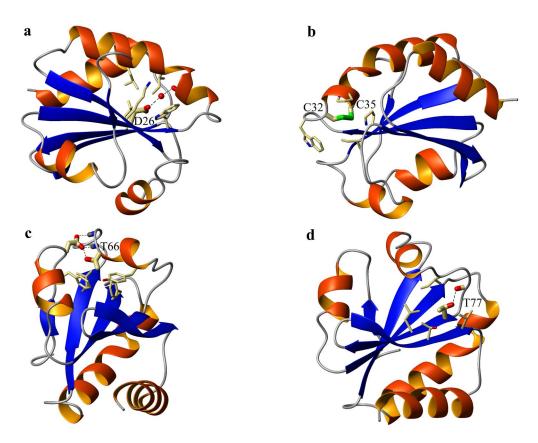
We find a strong positive correlation between the number of polar residues at core positions and protein size (Figure 4.1A). The largest protein in our analysis with no polar core residues was 94 amino acids. The number of polar residues at core positions as a function of the number of core residues was also analyzed (data not shown). According to a linear fit, a protein with eight or more core residues is expected to have at least one polar residue at a core position, and for proteins with a large number of core residues, approximately one-third of core residues are polar. The largest number of core residues in a protein with no core polar residues was 27. The average fraction of residues in the core that are polar increases slightly as protein size increases in the following ranges: less

than 100, 100-200, 200-300, and greater than 300 residues (Figure 4.1B). The average



**Figure 4.1** (a) Correlation of the number of polar residues at core positions with protein size. (b) Fraction of residues that are polar as a function of protein size. Error bars indicate one standard deviation from the mean.

within each size range is within a standard deviation of the other size ranges. Similarly, previous research has shown that bigger proteins tend to bury a larger percentage of the accessible surface area of their polar uncharged and charged atoms.<sup>35</sup> The average fraction of all residues that are polar decreases slightly with increased protein size within the same ranges and the average from each size range is again within a standard deviation of the other size ranges.



**Figure 4.2** Ribbon diagrams of the x-ray crystal structure of wild type *E. coli* thioredoxin<sup>28</sup> illustrating polar residues in the core and their surrounding contacts. (a) Asp26 makes a putative water mediated hydrogen bond to the carbonyl oxygen of Cys35. (b) Cys32 and Cys35 form an intra-molecular disulfide bond. (c) Thr66 makes a hydrogen bond to the side chain of Asp9. (d) Thr77 hydrogen bonds to the carbonyl oxygen of Gly74. Structure figures were generated using MOLMOL.<sup>36</sup>

Side chains of the five polar core residues (D26, C32, C35, T66, T77) in *E. coli* thioredoxin are greater than 95% buried, are distant in space from each other with the exception of the disulfide-bonded cystine, and make a range of hydrogen bonding

interactions in the 1.7 Å crystal structure.<sup>28</sup> D26, C32, C35, and T77 are conserved in human thioredoxin, which has a 0.79 Å rmsd from the E. coli structure and shares 29% sequence identity with E. coli thioredoxin.<sup>37</sup> The protein atoms surrounding D26 in E. *coli* thioredoxin are hydrophobic (Figure 4.2A). One of the O $\delta$  atoms of D26 is solvent inaccessible while the other is slightly solvent accessible via a narrow cleft in the protein. There is a water molecule in this cleft within hydrogen bond distance (2.7 Å) of D26. This water is also within hydrogen bond distance of the carbonyl oxygen of C35 (2.8 Å). A water molecule making similar contacts is present in both of the crystallographically independent molecules of the 1.7 Å structure 28 and the structure of a point mutant of E. coli thioredoxin, <sup>38</sup> and is found in five of six structures of human thioredoxin. <sup>37,39</sup> C32 and C35 form a redox active disulfide bond in wt thioredoxin. They are located close to the boundary between the protein core and surface. C32 and C35 are both slightly solvent exposed (Figure 4.2B); however, the sulfur atoms are solvent inaccessible. The sulfur atom of C32 is within weak hydrogen bond distance of the amide nitrogen of C35 (3.2 Å). T66 is solvent inaccessible (Figure 4.2C) and is within hydrogen bond distance of the side chain of D9 which makes putative hydrogen bonds with the amide protons of both T66 (3.0 Å) and G65 (2.9 Å). T77 is slightly solvent accessible (Figure 4.2D) and makes a putative hydrogen bond with the carbonyl oxygen of G74 (2.7 Å).

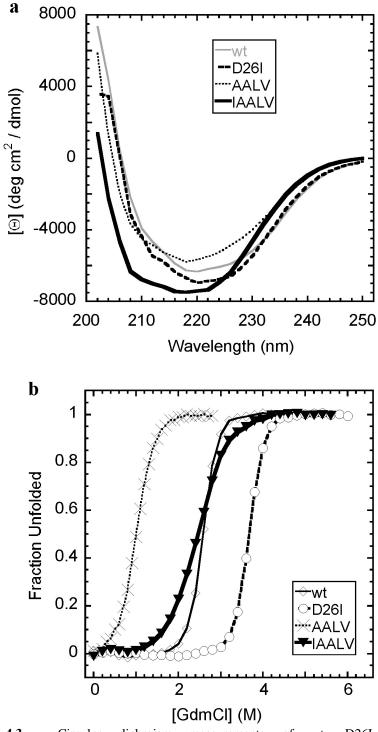
The ORBIT protein design calculations recommended the following five polar to hydrophobic mutations in thioredoxin: D26I, C32A, C35A, T66L, and T77V. Identical mutations were predicted when each position was optimized individually. Four of the point mutant proteins (C32A, C35A, T66L, T77V) were thermodynamically destabilized relative to wt while D26I was stabilized (Table 4.1). AALV, a mutant protein combining

the four destabilizing mutations (C32A/C35A/T66L/T77V), was thermodynamically destabilized relative to wt, but had a similar CD spectrum to wt (Figure 4.3A) and unfolding transition with similar cooperativity to wt as observed by chemical denaturation monitored by CD (Figure 4.3B). The five-fold mutant protein IAALV (D26I/C32A/C35A/T66L/T77V) is less stable than wt and has an unfolding transition that is substantially less cooperative than D26I, AALV, and wt as observed by chemical denaturation monitored by CD (Figure 4.3B). When monitored by fluorescence, chemical unfolding experiments also exhibit a substantially less cooperative unfolding transition for IAALV compared to wt (*m* values of 2.1 and 3.3 kcal mol<sup>-1</sup> M<sup>-1</sup> respectively). Comparison of the extent of chemical denaturation as monitored by CD and fluorescence does not indicate the presence of intermediate states.

Table 4.1 Experimental thermodynamic data							
Protein	T <sub>m</sub> a (°C)	ΔG <sup>b</sup> (kcal mol <sup>-1</sup> )	C <sub>m</sub> c (M)	m-valued (kcal mol <sup>-1</sup> M <sup>-1</sup> )			
wt	87	9.0	2.6	3.4			
wt reduced	76	6.3	1.8	3.6			
D26I	98	12.1	3.7	3.3			
C32A	74	5.3	1.6	3.4			
C35A	73	5.9	1.7	3.5			
T66L	85	7.6	2.4	3.1			
T77V	82	7.9	2.5	3.2			
AALV	55	2.8	1.0	2.8			
IAALV	66	3.9	2.4	1.6			

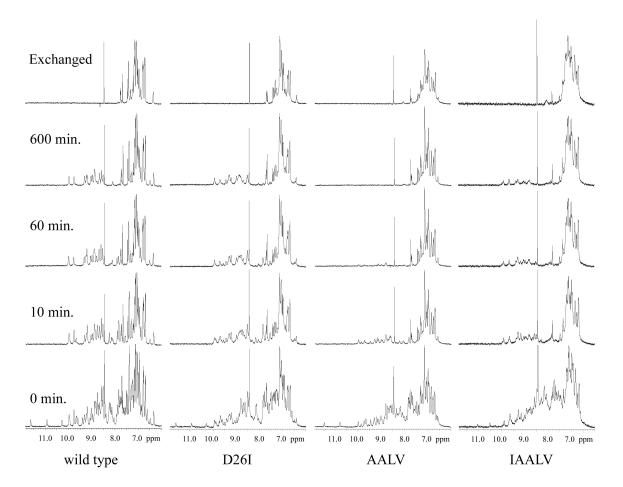
<sup>a</sup>Calorimetrically determined melting temperature. <sup>b</sup>Free energy of unfolding at 25 °C measured by guanidinium chloride denaturation (monitored by CD). <sup>c</sup>Midpoint of the guanidinium chloride unfolding transition. <sup>d</sup>Slope of  $\Delta G$  versus denaturant concentration plots.

The thermodynamic stability of D26I is 3.1 kcal/mol greater than wild type. Hydrogen exchange (Figures 4.4 & 4.5) shows that D26I protects slightly more amide

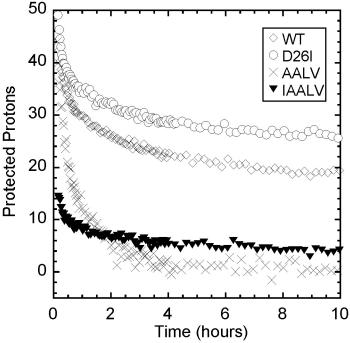


**Figure 4.3** Circular dichroism measurements of wt, D26I, AALV (C32A/C35A/T66L/T77V), and IAALV (D26I/C32A/C35A/T66L/T77V) thioredoxin. (a) Far-UV CD spectra. (b) Guanidinium chloride denaturation at 25 °C monitored by CD at 219 nm.

protons than wild type thioredoxin. The thermodynamic stability of IAALV is 1.1 kcal/mol greater than AALV. IAALV has very few protected amide protons relative to AALV; however, these protons are well protected (Figures 4.4 & 4.5). The proton NMR spectrum (Figures 4.4 & 4.6) of D26I shows slight line broadening relative to wild type thioredoxin. IAALV shows considerable line broadening and loss of chemical shift dispersion relative to AALV (Figures 4.4 & 4.6).



**Figure 4.4** Proton NMR spectra of wt, D26I, AALV (C32A/C35A/T66L/T77V), and IAALV (D26I/C32A/C35A/T66L/T77V) thioredoxin. The spectra at 0 min. were recorded in 90:10 H<sub>2</sub>O/D<sub>2</sub>O, and additional times refer to exchange time after resuspension of lyophilized sample in 99.9% D<sub>2</sub>O. The fully exchanged spectra were acquired after heating of the samples above their calorimetrically determined melting temperature. IAALV has broader lines and poorer dispersion than wt, suggesting that IAALV exhibits greater conformational flexibility. The persistence of amide peaks at 600 min. indicates well-protected protons in IAALV. Spectra at 0 min. were scaled to maximize the height of the viewable area. Within columns, spectra are scaled identically. The sharp peaks at 8.45 ppm are impurities.

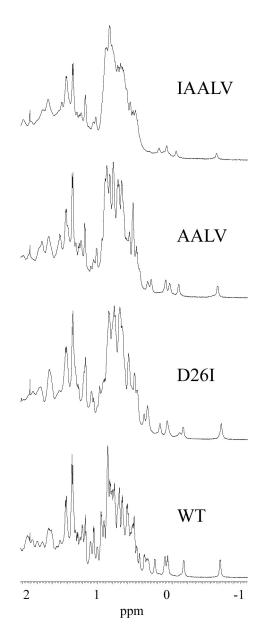


**Figure 4.5** Amide hydrogen-deuterium exchange of wt, D26I, AALV (C32A/C35A/T66L/T77V), and IAALV (D26I/C32A/C35A/T66L/T77V) thioredoxin. Total area of exchangeable peaks, expressed as number of protons, as a function of time at 25 °C and pH 7.0. Exchange rates calculated with the program SPHERE  $^{40,41}$  using the wt thioredoxin sequence predict a protection factor of  $10^4$  corresponds to an average half-life of 0.97 hours under these conditions.

### **Discussion**

The increased thermodynamic stability of D26I relative to wild type at neutral pH is similar to that reported for D26A.<sup>42</sup> The midpoint of guanidinium chloride denaturation increases by 1.1 M for D26I and 0.9 M for D26A<sup>42</sup> relative to wild type. In both cases, the increased thermodynamic stability relative to wt is likely due to the removal from the core of a charged atom whose hydrogen bond potential was unsatisfied.<sup>42</sup> The protected protons as well as the chemical denaturation data suggest that D26I is well folded. However, the line shape in the NMR spectrum of D26I is poorer than wt indicating some structural differences. The thermodynamic destabilization of T66L and T77V relative to wt is likely due to the loss of intra-

molecular hydrogen bonds without compensatory interactions with solvent for the previous hydrogen bond partners (O $\delta$  of Asp9 and O of Gly74).



**Figure 4.6** Proton NMR spectra of WT, D26I, AALV, and IAALV thioredoxin showing the upfield methyl region.

The loss in stability of the cysteine to alanine variants (C32A and C35A) relative to wt is largely accounted for by removal of the disulfide crosslink. Calorimetric studies by Sturtevant and co-workers<sup>43</sup> compared reduced wild type and C32S/C35S thioredoxin

and found an increased enthalpy of unfolding for C32S/C35S relative to the reduced wild type protein. From this study, it was concluded that stronger hydrogen bonding by the more polarizable hydroxyl of serine relative to the sulfhydryl of cysteine explained the difference in calorimetric unfolding. In addition, Raman spectroscopy indicates that the sulfhydryl group of each cysteine is a hydrogen bond donor in reduced thioredoxin. Hydrogen bonding by the sulfhydryl groups of the cysteine residues in reduced wild type thioredoxin, which would be partially eliminated in the C32A and C35A point mutant proteins, may explain the slight thermodynamic destabilization of C32A and C35A relative to the reduced wild type protein.

According to standard models, the m value reflects the change in hydrophobic surface area exposed upon conversion of the native state to the denatured state. The low m value of IAALV relative to wild type is consistent with greater hydrophobic surface area exposed in the native state of IAALV compared to wild type thioredoxin. Fluctuation between multiple low energy conformations in the native state may be responsible for some hydrophobic solvent exposure in IAALV. Decreased hydrophobic surface area exposed in the denatured state may also contribute to the low m value of IAALV relative to wild type. Alternatively, the presence of folding intermediates, while not observed in comparing the extent of chemical denaturation determined in the circular dichroism and fluorescence experiments, may be responsible for the difference in m values between IAALV and wt.

The slight increase in the number of protected protons for D26I relative to wild type thioredoxin indicates that the amide protons of D26I are less accessible to solvent.

This observation is consistent with the increased thermodynamic stability of D26I

relative to wild type. Both results suggest that D26I populates the unfolded state less than wild type thioredoxin.

IAALV is thermodynamically stabilized relative to AALV; however, this stabilization is greatly reduced compared to the D26I mutation in the wild type background (1.1 and 3.1 kcal/mol respectively). The extremely poor line shape and dispersion of IAALV relative to AALV, indicates a substantial increase in conformational heterogeneity. Increased conformational heterogeneity could result in poor packing and is consistent with the reduced thermodynamic stabilization of the D26I mutation in the AALV background compared to the wild type background. A small number of protons in IAALV are well protected from exchange with solvent, suggesting that in the family of folded structures putatively populated by IAALV, these protons are continually inaccessible to solvent exchange. In contrast, AALV appears well structured in the folded state but samples the denatured state and fully exchanges all of its protons over the time course of the hydrogen exchange experiments. Adding the D26I mutation to AALV likely results in a family of low energy structures in the folded state, all of which are stabilized relative to the denatured state.

Polar residues are prevalent at core positions in globular proteins. While some of these residues may be involved in active sites, the strong correlation with protein size suggests that most polar residues in the core are structurally important. McDonald and Thornton<sup>46</sup> found that the vast majority of buried polar atoms in proteins are satisfied by intra-molecular hydrogen bonds. They conclude that to offset the energetic cost of losing H-bonds with solvent upon folding, the majority of buried polar groups observed in folded protein structures form intra-molecular hydrogen bonds. Larger proteins are likely

to require more polar core residues in order to satisfy the hydrogen bonding potential of buried polar groups on the backbone, which may explain part of the correlation between protein size and number of polar residues at core positions.

Three slightly overlapping categories suffice to explain most core polar residues: active sites, thermodynamic stability, and structural specificity. Active sites use polar residues to perform chemistry and are often buried in clefts.<sup>47</sup> Polar residues can increase thermodynamic stability when they form intra-molecular hydrogen bonds to polar atoms that are solvent inaccessible (e.g., T66 and T77 in thioredoxin). The directional nature of polar interactions as well as their hydrophobic aversion allows core polar residues to determine a fold in a more specific manner than hydrophobic residues alone.

Previous studies have elucidated some of the factors that influence fold specificity. The relatively high conservation of core residues in sequence alignments clearly indicates the structural importance of core residues in terms of the hydrophobic effect and packing specificity. Based on this concept, a genetic library with the correct binary pattern to produce a 74 residue four helix bundle with a hydrophobic interior and hydrophilic exterior was shown to contain sequences that were helical, monomeric and had cooperative chemical denaturation. Consistent with these studies, we observe small proteins in our structural database survey with no polar residues at core positions. As the size of a protein increases, the probability increases of having a polar main chain atom that requires a core polar residue to satisfy its hydrogen bonding potential. Similarly, the number of folds that bury substantial hydrophobic surface area also increases with protein size. In larger proteins, polar residues at core positions likely aid in satisfying the

hydrogen bond potential of main chain polar groups as well as energetically distinguishing a single well-folded structure from alternative structures.

### References

- Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. J Mol Biol 105, 1-14.
- 2. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J Mol Biol* **196**, 641-656.
- 3. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155.
- 4. Honig, B. & Yang, A. S. (1995). Free energy balance in protein folding. *Adv Protein Chem* **46**, 27-58.
- 5. Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol* **5**(6), 470-475.
- Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E.
   P. & Matthews, B. W. (1992). Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science* 255, 178-183.
- 7. Dill, K. A., Bromberg, S., Yue, K., Fiebig, Klaus M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995). Principles of protein folding a perspective from simple exact models. *Protein Sci* **4**, 561-602.
- 8. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science* **278**, 82-87.
- 9. Nozaki, Y. & Tanford, C. (1971). The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J Biol Chem* **246**(7), 2211-2217.
- 10. Lesser, G. J. & Rose, G. D. (1990). Hydrophobicity of amino acid subgroups in proteins. *Proteins* **8**(1), 6-13.

- Myers, J. K. & Pace, C. N. (1996). Hydrogen bonding stabilizes globular proteins.
   Biophys J 71(4), 2033-2039.
- 12. Blaber, M., Lindstrom, J. D., Gassner, N., Xu, J., Heinz, D. W. & Matthews, B. W. (1993). Energetic cost and structural consequences of burying a hydroxyl group within the core of a protein determined from Ala → Ser and Val → Thr substitutions in T4 lysozyme. *Biochemistry* 32, 11363-11373.
- 13. Hendsch, Z. S., Jonsson, T., Sauer, R. T. & Tidor, B. (1996). Protein stabilization by removal of unsatisfied polar groups: computational approaches and experimental tests. *Biochemistry* **35**, 7621-7625.
- 14. Hendsch, Z. S. & Tidor, B. (1994). Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci* **3**, 211-226.
- 15. Rose, G. D. & Wolfenden, R. (1993). Hydrogen bonding, hydrophobicity, packing, and protein folding. *Annu Rev Biophys Biomol Struct* **22**, 381-415.
- 16. Byrne, M. P., Manuel, R. L., Lowe, L. G. & Stites, W. E. (1995). Energetic contribution of side chain hydrogen bonding to the stability of staphylococcal nuclease. *Biochemistry* **34**, 13949-13960.
- 17. Yamagata, Y., Kubota, M., Sumikawa, Y., Funahashi, J., Takano, K., Fujii, S. & Yutani, K. (1998). Contribution of hydrogen bonds to the conformational stability of human lysozyme: calorimetry and x-ray analysis of six tyrosine --> phenylalanine mutants. *Biochemistry* 37, 9355-9362.
- 18. Takano, K., Yamagata, Y., Kubota, M., Funahashi, J., Fujii, S. & Yutani, K. (1999).

  Contribution of hydrogen bonds to the conformational stability of human lysozyme:

- calorimetry and x-ray analysis of six Ser --> Ala mutants. *Biochemistry* **38**, 6623-6629.
- Shirley, B. A., Stanssens, P., Hahn, U. & Pace, C. N. (1992). Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry* 31, 725-732.
- 20. Pace, C. N., Shirley, B. A., McNutt, M. & Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *Faseb J* **10**(1), 75-83.
- 21. O'Shea, E. K., Rutkowski, R. & Kim, P. S. (1992). Mechanism of specificity in the Fos-Jun oncoprotein heterodimer. *Cell* **68**(4), 699-708.
- 22. O'Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science* **254**, 539-544.
- 23. Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. (1993). A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401-1407.
- 24. Gonzalez, L., Jr., Brown, R. A., Richardson, D. & Alber, T. (1996). Crystal structures of a single coiled-coil peptide in two oligomeric states reveal the basis for structural polymorphism. *Nat Struct Biol* **3**(12), 1002-1009.
- 25. O'Shea, E. K., Lumb, K. J. & Kim, P. S. (1993). Peptide 'Velcro': design of a heterodimeric coiled coil. *Curr Biol* **3**, 658-667.
- 26. Lumb, K. J. & Kim, P. S. (1995). A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry* **34**, 8642-8648.

- 27. Munoz, V. & Serrano, L. (1994). Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. *Proteins* **20**(4), 301-311.
- 28. Katti, S. K., LeMaster, D. M. & Eklund, H. (1990). Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution. *J Mol Biol* **212**(1), 167-184.
- 29. Dahiyat, B. I., Sarisky, C. A. & Mayo, S. L. (1997). De novo protein design: towards fully automated sequence selection. *J Mol Biol* **273**(4), 789-796.
- 30. Hemsley, A., Arnheim, N., Toney, M. D., Cortopassi, G. & Galas, D. J. (1989). A simple method for site-directed mutagenesis using the polymerase chain reaction. *Nucleic Acids Res* **17**(16), 6545-6551.
- 31. Holmgren, A. & Reichard, P. (1967). Thioredoxin 2: cleavage with cyanogen bromide. *Eur J Biochem* **2**(2), 187-196.
- 32. Kelley, R. F., Shalongo, W., Jagannadham, M. V. & Stellwagen, E. (1987). Equilibrium and kinetic measurements of the conformational transition of reduced thioredoxin. *Biochemistry* **26**(5), 1406-1411.
- 33. Santoro, M. M. & Bolen, D. W. (1988). Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alphachymotrypsin using different denaturants. *Biochemistry* **27**(21), 8063-8068.
- 34. Privalov, G., Kavina, V., Freire, E. & Privalov, P. L. (1995). Precise scanning calorimeter for studying thermal properties of biological macromolecules in dilute solution. *Anal Biochem* **232**(1), 79-85.

- Kajander, T., Kahn, P. C., Passila, S. H., Cohen, D. C., Lehtio, L., Adolfsen, W.,
   Warwicker, J., Schell, U. & Goldman, A. (2000). Buried charged surface in
   proteins. Structure Fold Des 8(11), 1203-1214.
- 36. Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* **14**(1), 51-55, 29-32.
- 37. Weichsel, A., Gasdaska, J. R., Powis, G. & Montfort, W. R. (1996). Crystal structures of reduced, oxidized, and mutated human thioredoxins: evidence for a regulatory homodimer. *Structure* **4**(6), 735-751.
- 38. Nikkola, M., Gleason, F. K., Fuchs, J. A. & Eklund, H. (1993). Crystal structure analysis of a mutant Escherichia coli thioredoxin in which lysine 36 is replaced by glutamic acid. *Biochemistry* **32**(19), 5093-5098.
- 39. Andersen, J. F., Sanders, D. A., Gasdaska, J. R., Weichsel, A., Powis, G. & Montfort, W. R. (1997). Human thioredoxin homodimers: regulation by pH, role of aspartate 60, and crystal structure of the aspartate 60 --> asparagine mutant. *Biochemistry* 36(46), 13979-13988.
- 40. Zhang, Y. Z. (1995). Protein and peptide structure and interactions studied by hdrogen exchange and NMR, University of Pennsylvannia.
- 41. Bai, Y., Milne, J. S., Mayne, L. & Englander, S. W. (1993). Primary structure effects on peptide group hydrogen exchange. *Proteins* **17**(1), 75-86.
- 42. Langsetmo, K., Fuchs, J. A. & Woodward, C. (1991). The conserved, buried aspartic acid in oxidized Escherichia coli thioredoxin has a pKa of 7.5. Its titration produces a related shift in global stability. *Biochemistry* **30**(30), 7603-7609.

- 43. Ladbury, J. E., Kishore, N., Hellinga, H. W., Wynn, R. & Sturtevant, J. M. (1994). Thermodynamic effects of reduction of the active-site disulfide of Escherichia coli thioredoxin explored by differential scanning calorimetry. *Biochemistry* **33**(12), 3688-3692.
- 44. Li, H., Hanson, C., Fuchs, J. A., Woodward, C. & Thomas, G. J., Jr. (1993). Determination of the pKa values of active-center cysteines, cysteines- 32 and -35, in Escherichia coli thioredoxin by Raman spectroscopy. *Biochemistry* 32(22), 5800-5808.
- 45. Schellman, J. A. (1978). Solvent denaturation. *Biopoly* **17**, 1305-1322.
- 46. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**(5), 777-793.
- 47. Stryer, L. (1995). Biochemistry. In *Biochemistry, Fourth Edition* Fourth Edition edit., pp. 191. W. H. Freeman and Company, New York, NY.
- 48. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* **247**, 1306-1310.
- Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993).
   Protein design by binary patterning of polar and nonpolar amino acids. *Science* 262, 1680-1685.

# Chapter 5

# **Prudent Modeling of Core Polar Residues**in Computational Protein Design

### Summary

Hydrogen bond interactions were surveyed in a set of protein structures. Compared to surface positions, polar side chains at core positions form a greater number of intra-molecular hydrogen bonds. Furthermore, the majority of polar side chains at core positions form at least one hydrogen bond to main chain atoms that are not involved in hydrogen bonds to other main chain atoms. Based on this structural survey, hydrogen bond rules were generated for each polar amino acid in core protein design. In the context of core protein design, these prudent polar rules were used to eliminate from consideration polar amino acid rotamers that do not form a minimum number of hydrogen bonds. As an initial test, the core of Escherichia coli thioredoxin was selected as a design target. For this target, the prudent polar strategy resulted in a minor increase in computational complexity compared to a strategy that did not allow polar residues. Dead end elimination was used to identify global minimum energy conformations for the prudent polar and no polar strategies. The prudent polar strategy identified a protein sequence that was thermodynamically stabilized by 2.5 kcal/mol relative to wild type thioredoxin and 2.2 kcal/mol relative to a thioredoxin variant whose core was designed without polar residues.

### Introduction

For design to succeed, the underlying model must accurately describe the physics involved. Under this principle, computational protein design has emerged as a powerful tool for studying the forces involved in stabilizing native protein structures. A design cycle, iterating between computational modeling and experimental analysis of designed sequences, has been used to improve the predictive capability of force fields for protein design.<sup>2</sup> Using the protein design cycle, it was found that including a benefit for hydrophobic burial improved the correlation between predicted and experimental stability. The energetic benefit of hydrophobic atoms interacting with a hydrophobic environment compared to an aqueous environment had previously been demonstrated by solvent transfer experiments.<sup>3</sup> The unfavorable interaction of solvent with hydrophobic molecules, compared to solvent-solvent interactions gained upon hydrophobic burial, is believed to explain this energetic benefit.<sup>4</sup> The burial of hydrophobic residues in the cores of folded proteins is thought to be a dominant force in stabilizing the native state over the denatured state.<sup>5</sup> Given the energetic benefit of burying hydrophobic atoms, protein design commonly restricts core positions to hydrophobic residues. 1,6-13 addition to maximizing hydrophobic burial, the restriction of core residues to hydrophobic identity results in a dramatic reduction in combinatorial complexity compared to the inclusion of polar residues.

Despite the energetic benefit of hydrophobic burial, analysis of folded protein structures indicates that many polar side chain atoms are buried in folded protein structures. We previously noted a strong positive correlation between protein size and the number of polar residues at core positions. This correlation suggests that polar

residues may play a structural role in energetically stabilizing the folded state and energetically distinguishing a single well-folded state from alternative conformations.<sup>16</sup> The mutation of polar residues involved in intra-molecular hydrogen bonding to hydrophobic residues frequently results in thermodynamically destabilized proteins.<sup>17-20</sup> In a survey of protein structures, McDonald and Thornton<sup>21</sup> found that the vast majority of buried polar atoms form intra-molecular hydrogen bonds.

These results suggest that core protein design could be improved by including polar residues provided that their hydrogen bond potential was largely satisfied. With the intent of constructing hydrogen bond rules for core protein design, we analyzed hydrogen bonding in a previously published set of unrelated high-resolution protein structures.<sup>22</sup> Hydrogen bond rules developed based on this survey together with a geometrically softened hydrogen bond potential form a prudent method for modeling polar residues in protein cores. This prudent polar strategy for core optimization was incorporated into the ORBIT (Optimization of Rotamers By Iterative Techniques) protein design software.<sup>9</sup> As an initial test, core residues of Escherichia coli (E. coli) thioredoxin were designed with prudent polar rules as well as with no polar residues. This core design differs from the perturbation study in Chapter 4 in that the disulfide bonded cystine residue was not optimized in this study. The cystine residue was retained here because the energetics of disulfide cross-links are not well modeled in protein design. The resulting sequences from the prudent polar and no polar designs were both thermodynamically stabilized relative to wild type (wt) thioredoxin. At 25 °C the prudent polar molecule was thermodynamically stabilized by 2.2 kcal/mol relative to the no polar molecule, indicating the energetic benefit of including polar residues in core protein design.

### **Materials and Methods**

Database analysis and construction of hydrogen bond rules for protein core design

We analyzed hydrogen bonding in a previously published dataset of 100 high-resolution protein structures.<sup>22</sup> Richardson and co-workers<sup>22</sup> previously modified the protein structures in this dataset by adding experimentally undetermined hydrogen atoms and positioning them to optimize hydrogen bonding and minimize van der Waals clashes (including amide flipping of asparagines and glutamines).

Side chains were classified as core, boundary, or surface based on the orientation of the  $C^{\alpha}$ - $C^{\beta}$  vector relative to a solvent-accessible surface computed with the template main chain heavy atoms (N, C,  $C^{\alpha}$ ) essentially as previously described.<sup>23</sup> A position was classified as core if the distance from its  $C^{\alpha}$ , along its  $C^{\alpha}$ - $C^{\beta}$  vector, to the solvent accessible surface was greater than 5.0 Å, and if the shortest distance from its  $C^{\beta}$  to the surface was greater than 2.0 Å. A position was classified as surface if the sum of the shortest distance from its  $C^{\beta}$  to the surface and the distance from its  $C^{\alpha}$ , along its  $C^{\alpha}$ - $C^{\beta}$  vector, to the solvent-accessible surface was less than 2.70 Å. A position was classified as boundary if it did not fit the criteria for both surface and core classifications. The same geometric criteria were used to classify main chain carbonyl and amide groups. For classification purposes, the bond distance of the N-HN and C-O bonds was adjusted to 1.54 Å to match with the  $C^{\alpha}$ - $C^{\beta}$  bond length for which the classification criteria were originally parameterized.

For our hydrogen bond analysis, we used the geometric hydrogen bond definition of McDonald and Thornton.<sup>21</sup> This hydrogen bond definition requires a donor-hydrogen-acceptor angle of greater than 90° and a hydrogen-acceptor distance of less than 2.5 Å, which corresponds to a heavy atom donor acceptor distance of approximately 3.5 Å for a linear donor-hydrogen-acceptor angle. Hydrogen bond donors and acceptors were those defined by McDonald and Thornton.<sup>21</sup> All hydrogen bonds meeting this criteria were identified within each protein structure in the dataset. In a subsequent step, side chainmain chain hydrogen bonds were searched to identify side chain hydrogen bonds to MCU atoms. Based on our survey, hydrogen bond requirements were developed for each of the polar amino acids in core protein design. For each amino acid, a minimum number of total hydrogen bonds as well as a minimum number of hydrogen bonds to desirable partners were determined. Desirable hydrogen bond partners are polar atoms that are likely to be buried from solvent and are defined here as MCU atoms as well as atoms of side chains that hydrogen bond to MCU atoms.

# Computational modeling

Simulations were performed using coordinates from the 1.7 Å x-ray structure of thioredoxin. All side chains identified as core were optimized in the calculations with the exception of Trp 28, Cys 32, and Cys 35. The side chain of Trp 28 adopts a non-canonical dihedral angle and is not well described in our rotamer library, while Cys 32 and Cys 35 form an intra-molecular disulfide bond. The remainder of the structure (including Trp 28, Cys 32 and Cys 35) was held fixed during the optimizations. All optimized side chains were modeled using a backbone dependent rotamer library that was expanded about  $\chi 1$  and  $\chi 2$  dihedral angles for aromatic residues, expanded about  $\chi 1$ 

dihedral angles for aliphatic residues, and unexpanded for polar residues as previously described.<sup>23</sup>

The calculations used an energy function based on the DREIDING force field.<sup>26</sup> A Lennard-Jones 12-6 potential with van der Waals radii scaled by 0.9 was applied to all described.8 interactions previously rotamer-template and rotamer-rotamer as Electrostatic energies were computed using a distance dependent dielectric constant of 40r (where r is the inter-atomic distance) with partial atomic charges as previously described.<sup>27</sup> A surface area-based atomic solvation potential was applied to hydrophobic atoms. Surface areas for the solvation potential were calculated using the Connolly algorithm<sup>28</sup> and the Lee and Richards definition of solvent accessible surface area<sup>29</sup> with unscaled atomic radii and an add-on radius of 1.4 Å. A hydrophobic burial benefit of 48 cal/mol/Å<sup>2</sup> and a hydrophobic exposure penalty of 76.8 cal/mol/Å<sup>2</sup> were used with the two-body decomposition method of Street and Mayo.<sup>30</sup> The solvation potential was not applied to polar atoms.

For the no polar and unrestricted polar calculations, an explicit hydrogen bond potential<sup>27</sup> was applied to all groups that contained hydrogen bond donors and/or acceptors. This hydrogen bond potential consisted of distance dependent and angle dependent terms:

$$E_{HB} = D_0 \left\{ 5 \left( \frac{R_0}{R} \right)^{12} - 6 \left( \frac{R_0}{R} \right)^{10} \right\} F(\theta, \phi, \varphi)$$

where  $R_0$  (2.8 Å) is the heavy atom donor-acceptor equilibrium distance,  $D_0$  (8 kcal/mol) is the well-depth, and R is the heavy atom donor-acceptor distance. The angle term is essentially a  $\cos^4$  term and is designed to penalize unfavorable geometries based on the

hybridization state of the donor and acceptor atoms as previously described.<sup>27</sup> For the no polar design, core positions were restricted to the following seven amino acids: Trp, Tyr, Phe, Leu, Ile, Val, Ala. For the unrestricted polar calculations, core positions were restricted to the following amino acids: Trp, Tyr, Phe, Leu, Ile, Val, Ala, Arg, Lys, Asp, Glu, Asn, Gln, Thr, Ser.

For the prudent polar calculation, core positions were initially restricted to the following amino acids: Trp, Tyr, Phe, Leu, Ile, Val, Ala, Arg, Lys, Asp, Glu, Asn, Gln, Thr, Ser. An explicit hydrogen bond potential was applied to all groups that contained hydrogen bond donors and/or acceptors. This potential differs from the default 12-10 potential described above in that it includes an angle independent energy ceiling:

$$E_{HB} = -D_1 + \left(D_0 - D_1\right) \left\{ 5 \left(\frac{R_0}{R}\right)^{12} - 6 \left(\frac{R_0}{R}\right)^{10} \right\} F(\theta, \phi, \varphi)$$

where  $D_1$  (6 kcal/mol) is the energy ceiling. In order to avoid benefiting grossly improper geometry, the hydrogen bond potential was only applied when 2.5 Å < R < 3.2 Å, and the donor-hydrogen-acceptor angle was greater than 90 degrees. This potential was used in part to account for the altered dielectric constant in protein cores as well as to account for backbone flexibility and the use of discrete rotamers. The low dielectric constant in the protein core should increase the electrostatic component of hydrogen bonds in the protein core.<sup>31</sup> The energy ceiling effectively reduces the angle dependence of the hydrogen bond potential, and is intended to energetically benefit residues that can form geometrically favorable hydrogen bonds given slightly greater side chain and backbone flexibility. The energy ceiling of 6 kcal/mol was chosen so that a core

hydrogen bond would be energetically similar in magnitude to the hydrophobic burial benefit of a methyl group in our forcefield ( $\approx 100 \text{ Å}^2 \times 0.048 \text{ kcal/mol/Å}^2 = 4.8 \text{ kcal/mol}$ ).

Desirable hydrogen bond partners were identified in the template atoms (portions of the molecule held fixed during optimization) using the same analysis program used for our database survey. In ORBIT, the pairwise energy matrix is calculated in two steps. In the first step, interaction energies between each rotamer and all template atoms are calculated and stored. In the second step, the energy between each pair of rotamers is calculated. During the rotamer-template energy calculation, polar residues that did not make the required number of hydrogen bonds were eliminated.

For all design strategies (unrestricted polar, no polar, and prudent polar), an algorithm based on the dead end elimination theorem<sup>32,33</sup> was used to search for the global minimum energy conformation. Calculations were performed on six 195 MHz R10000 processors (Silicon Graphics, Mountain View, CA) running in parallel.

# Mutagenesis and protein purification

The genes for all variants were constructed by inverse PCR<sup>34</sup> using the gene for wt thioredoxin (Invitrogen) cloned into pET-11a (Novagen). Mutations were confirmed by DNA sequencing. BL21(DE3) hosts (Invitrogen) were used for protein expression. Cells were grown to mid log phase and induced with 0.5 mM IPTG (ICN) for three hours at 37 °C. Cells were lysed by sonication and centrifuged twice at 20,000×g for 30 minutes. Acetonitrile (EM Science) was added to the soluble fraction to a final concentration of 60%. After centrifugation, the supernatant was evaporated to half volume and purified by reverse phase high performance liquid chromatography using an acetonitrile/water gradient containing 0.1% trifluoroacetic acid (Applied Biosystems).

Protein identities were confirmed by mass spectrometry. Protein concentration was determined by UV absorbance in 7 M GdmCl (ICN) based on an extinction coefficients at 280 nm of 13,700 M<sup>-1</sup>cm<sup>-1</sup> for the wt molecule, <sup>35</sup> 12,400 M<sup>-1</sup>cm<sup>-1</sup> for the prudent polar molecule, and 11,100 M<sup>-1</sup>cm<sup>-1</sup> for the no polar molecule.

### CD analysis

CD data were collected on an Aviv 62DS spectrometer equipped with a thermoelectric unit and using a 1 cm path length cell. Protein samples were 5  $\mu$ M in 50 mM sodium phosphate at pH 7.0. Guanidinium chloride denaturations were monitored at 219 nm at 25 °C for comparison to previously published data. <sup>16,36</sup>  $\Delta$ G values, m values, and error estimates were obtained by fitting the denaturation data to a two-state transition as described previously. <sup>37</sup> GdmCl concentration was determined by refractometry.

### **DSC**

Protein samples at ~0.7 mg/ml were dialyzed against 50 mM sodium phosphate buffer at pH 7.0. DSC data were collected on a N-DSC II instrument from Calorimetric Sciences Corp.<sup>38</sup> Samples were degassed under vacuum for 15 minutes prior to scanning. Scans were performed at a rate of 1.0 °C/min. Data were analyzed using the program cpcalc (Applied Thermodynamics). The free energy of unfolding as a function of temperature was calculated using the Gibbs-Helmholtz equation as previously described:<sup>39</sup>

$$\Delta G(T) = \Delta H_m - T\Delta S_m - \Delta C_p \left[ T_m - T \left( 1 - \ln \left[ \frac{T}{T_m} \right] \right) \right]$$

with  $T_m$ ,  $\Delta H_m$  and  $\Delta S_m$  values determined by DSC and a constant  $\Delta C_p$  of 1.66 kcal mol-1 K-1, which was previously determined<sup>40</sup> for wt thioredoxin.

### Results

In a previously published dataset of 100 unrelated, high-quality protein crystal structures at better than 1.7 A resolution,<sup>22</sup> we compared hydrogen bonding at surface and core positions for the following group of polar amino acids: Ser, Thr, Asn, Gln, Asp, Glu, Lys, and Arg (Table 5.1). We find that the majority of these polar side chains at core positions form intra-molecular hydrogen bonds. In contrast, the majority of these polar

<b>Table 5.1</b> Hydrogen bonding of polar side chains					
	Core	Surface			
Number of side chains	1835	3360			
Percent with no Intra-molecular HB	15.0	63.5			
Percent with HB to MC atom	62.7	16.3			
Percent with HB to MCU atom	53.6	12.7			
Percent with HB to crystallographic water	7.3	17.4			
<sup>a</sup> Includes Ser, Thr, Asn, and Arg.	Gln, Asp,	Glu, Lys,			

side chains at surface positions do not make intra-molecular hydrogen bonds. We find that most polar residues at core positions (62.7%) make hydrogen bonds to main chain atoms. Furthermore, the majority of polar residues at core positions (53.6%) make at least one hydrogen bond to a MCU atom (main chain atom not involved in hydrogen bonding to other main chain atoms). Very few polar core residues (7.3%) form hydrogen bonds to crystallographic water molecules, while a slightly greater proportion of surface polar residues (17.4%) form these interactions. At boundary positions, where the geometric classification criteria are insufficient to determine if a side chain will interact

with solvent or bury itself in the protein core, polar residues make an intermediate number of intra-molecular hydrogen bonds compared to surface and core positions (data not shown).

Table 5.2 Side chain hydrogen bonding at core positions							
Amino Acid	Total Number	Core Number	Percent Core	HB <sup>a</sup> per Core Residue	HBMCU <sup>b</sup> per Core Residue		
Ser	1204	360	30	1.5	0.6		
Thr	1131	365	32	1.4	0.5		
Asn	915	250	27	2.5	0.9		
Gln	633	131	21	2.3	0.9		

22

16

11

29

65

69

72

2.2

1.9

2.2

4.7

1.0

0.9

0.3

0.9

0.7

1.4

0.3

0.3

0.1

248

156

116

209

452

192

240

1134

999

1046

718 691

278

335

Asp

Glu

Lys

Arg

Tyr Trp

Cys

Hydrogen bonding by core side chains was analyzed for each polar amino acid type (Table 5.2). The partitioning of each amino acid between core and surface or boundary positions divides polar residues into two groups. The minority of the first group (Ser, Thr, Asn, Gln, Asp, Glu, Lys, and Arg) are located at core positions (11%-32%), while more than 65% of Trp, Tyr, and Cys residues and 45% of His residues are found at core positions. Amino acids in the first group (Ser, Thr, etc.) make more total hydrogen bonds and more hydrogen bonds to MCU atoms than amino acids in the second group (Trp, Tyr, Cys, and His). Within the first group, as the number of hydrogen bond donors and acceptors per amino acid type increases so does the total number of hydrogen bonds per residue. From this analysis hydrogen bonding requirements for polar core residues were constructed (Table 5.3). These rules require core polar residues to make a

His 376 171 46 0.2 0.3

<sup>a</sup>Total hydrogen bonds. <sup>b</sup>Hydrogen bonds to main chain atoms that are not involved in hydrogen bonds to other main chain atoms.

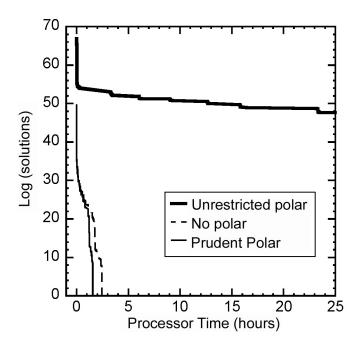
minimum number of total hydrogen bonds and a minimum number of hydrogen bonds to desirable partners. For example, in order for Asn to be considered at a core position, it must at a minimum form one hydrogen bond to a desirable partner and one additional hydrogen bond (two total H-bonds, one desirable H-bond). The number of total H-bonds required increases as the number of donors and acceptors per amino acid type increases while the number of desirable H-bonds required is never greater than one.

Table 5.3 HB requirements for core positions						
Amino Acid	Total H-Bonds <sup>a</sup>	Desirable H-Bonds <sup>b</sup>				
Ser	1	1				
Thr	1	1				
Asn	2	1				
Gln	2	1				
Asp	2	1				
Glu	2	1				
Lys	2	1				
Arg	4	1				
Tyr	0	0				
Trp	0	0				
Cys	*	*				
His	*	*				

<sup>\*</sup>The H-bonding trends seen in the database survey were deemed insufficient to generate prudent hydrogen bond requirements for these residues. <sup>a</sup>Identified using geometric H-bond definition described in Materials and Methods. <sup>b</sup>Desirable hydrogen bonds include H-bonds to MCU atoms as well as to atoms of side chains that hydrogen bond to MCU atoms

For the thioredoxin core design target, the amino acid identity of 32 positions was optimized. The wt sequence at these positions contains two Thr and one Asp, but is otherwise hydrophobic. For this design target, three different design strategies were employed: prudent polar, unrestricted polar, and no polar. For the prudent polar strategy, core polar rotamers were eliminated if they did not meet the hydrogen bond requirements

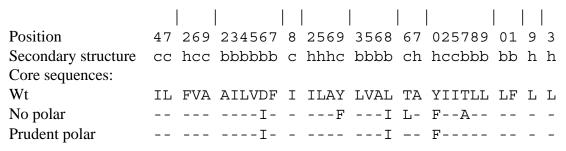
(Table 5.3). For the unrestricted polar strategy, polar rotamers were included without hydrogen bond requirements. For the no polar strategy, polar residues were excluded from the core as in previous design efforts.<sup>8,9,23</sup>



**Figure 5.1** DEE optimization progress as a function of processor time for the different design strategies. DEE successfully found the GMEC solution for the prudent polar (1.57 processor hours) and the no polar (2.44 processor hours) strategies, but stalled for the unrestricted polar strategy (10<sup>42</sup> solutions remained after 180 processor hours).

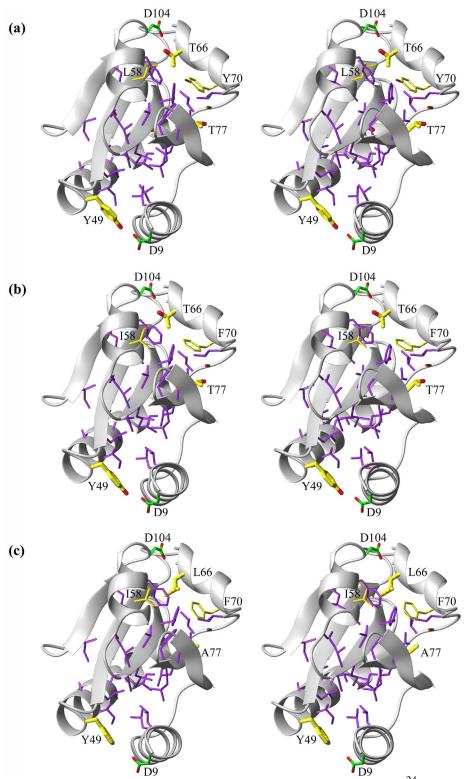
For the thioredoxin core design target, the prudent polar design strategy resulted in  $10^{50}$  possible solutions, the no polar strategy resulted in  $10^{49}$  possible solutions, and the unrestricted polar strategy resulted in  $10^{67}$  possible solutions. Computational optimization was performed using an algorithm based on the dead end elimination (DEE) theorem<sup>32,33</sup> and proceeded rapidly for the prudent polar and no polar strategies, but stalled for the unrestricted polar strategy (Fig. 5.1). The prudent polar global minimum energy conformation (GMEC) was found in 1.57 processor hours, while the no polar GMEC was found in 2.44 processor hours. After 180 processor hours, the unrestricted

polar optimization had stalled and was stopped with greater than  $10^{42}$  possible solutions remaining. The optimized sequence from the no polar strategy has six mutations (D26I, Y49F, L58I, T66L, Y70F, and T77A) corresponding to 81% identity relative to the wt sequence (Fig. 5.2). The optimized sequence from the prudent polar strategy has three mutations (D26I, L58I, T70F) corresponding to 91% sequence identity relative to the wt sequence (Fig. 5.2).



**Figure 5.2** Amino acid identity of positions included in the core optimization of thioredoxin. The first two lines indicate amino acid position. The third line indicates the secondary structure, as determined by MOLMOL, <sup>41</sup> for each position (h – helix; b - beta strand; c - coil, neither helix nor strand). The amino acids of the wt protein, as well as the two designed proteins, are shown in single letter code.

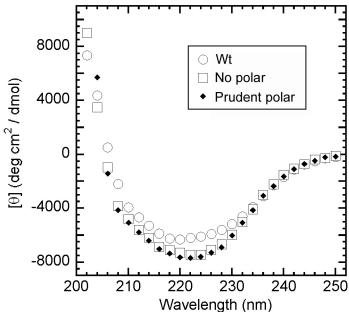
Models of the no polar and prudent polar designed structures indicate similar core packing compared to wt, but different H-bond interactions (Fig. 5.3). Of the six mutations (D26I, Y49F, L58I, T66L, Y70F, and T77A) in the no polar designed sequence, five remove polar groups relative to wt. In the wt structure, <sup>24</sup> three of these polar residues (Y49, T66, and T77) form direct intra-molecular H-bonds. These direct intra-molecular H-bonds are recovered in the prudent polar design. The hydroxyl group of Y49 is within hydrogen bond distance to a side chain carboxylate oxygen of D104. D104 is at a surface position and in contact with solvent. T66 is within hydrogen bond distance of the side chain of D9, which makes putative hydrogen bonds with the amide protons of both T66 and G65. T77 makes a putative hydrogen bond with the carbonyl



**Figure 5.3** Three-dimensional molecular representations of the wt thioredoxin crystal structure <sup>24</sup> (a) and models of the prudent polar (b) and no polar (c) designs. Side chains of the 26 core positions with identical amino acid identity are shown in purple. Side chains at the six variable positions (26 – located behind the b-sheet in these representations, 49, 58, 66, 70, and 77) are shown with carbon atoms colored yellow and oxygen atoms colored red. Side chains at two fixed positions (9, 104) involved in H-bonds to core side chains are shown with carbon atoms colored green and oxygen atoms colored red. Images were created with the program MOLMOL. <sup>41</sup>

oxygen of G74 in both the wt structure and the prudent polar design. The carbonyl oxygen of G74 is buried from solvent and is not involved in hydrogen bond interactions with other main chain atoms.

The prudent polar and no polar designed sequences have three common mutations relative to wt (D26I, L58I, and Y70F). In models of the prudent polar and no polar designs, I26, I58, and Y70 all make extensive hydrophobic contacts. In the structure of wt thioredoxin, D26 is hydrogen bonded through a bridging water molecule to the carbonyl oxygen of C35 (the carbonyl oxygen of C35 makes a putative hydrogen bond to the main chain amide of I38). In the wt structure, Hydrophobic atoms surround L58 and the hydroxyl group of Y70 contacts solvent but does not form intra-molecular H-bonds.



**Figure 5.4** Far-UV circular dichroism spectra of the wt, no polar, and prudent polar proteins. All three proteins appear to have similar secondary structure content and are well folded at 25 °C.

The wt, no polar, and prudent polar sequences were all expressed in *E. coli* with high yield (20-50 mg purified protein/L culture). Circular dichroism (CD) of the wt, no polar, and prudent polar proteins indicates that all three proteins have similar secondary

structure content and are well-folded at room temperature (Fig. 5.4). In addition, all three proteins denature cooperatively in the presence of increasing concentrations of guanidinium.

Table 5.4 Experimental thermodynamic data

Protein	$\begin{array}{c} \Delta G^a \\ (kcal\ mol^{\text{-}1}\ M^{\text{-}1}) \end{array}$	C <sub>m</sub> <sup>b</sup> (M)	m-value <sup>c</sup> (kcal mol <sup>-1</sup> )	$T_m^{d}$ (°C)	ΔH <sub>vh</sub> <sup>e</sup> (kcal mol <sup>-1</sup> )	$\begin{array}{c} \Delta H_{cal}^{f} \\ (kcal\ mol^{-1}) \end{array}$	$\beta^{\mathrm{g}}$
$Wt^h$	9.0	2.6	3.4	87	117	117	1.00
No polar	9.3	3.2	2.9	97	106	106	1.00
Prudent pola	ar 11.5	3.4	3.4	97	148	147	1.01

<sup>&</sup>lt;sup>a</sup>Free energy of unfolding at 25 °C. <sup>b</sup>Midpoint of guanidinium chloride unfolding transition. <sup>c</sup>Slope of  $\Delta G$  versus denaturant concentration plots. <sup>d</sup>Midpoint of calorimetric unfolding. <sup>e</sup> $\Delta H_{vh}$  Van Hoff's enthalpy. <sup>f</sup> $\Delta H_{cal}$  Calorimetric enthalpy. <sup>g</sup> $\Delta H_{cal}$  /  $\Delta H_{vh}$ . <sup>h</sup>from Bolon and Mayo. <sup>16</sup>

Chemical denaturation indicated that the no polar protein and the prudent polar protein were thermodynamically stabilized at 25 °C relative to the wt protein (Table 5.4). The no polar protein is 0.3 kcal/mol more stable and the prudent polar protein is 2.5 kcal/mol more stable than the wt protein. The m-value, which is a measure of cooperativity, is similar for all three proteins, though it is slightly reduced for the no polar protein relative to the other two proteins. The results of differential scanning calorimetry (DSC) indicate that  $T_m$ s for the no polar protein and the prudent polar protein are both increased by 10 °C relative to the wt protein. In addition, DSC indicates that the  $\Delta H$  of unfolding is 38% (41 kcal/mol) greater for the prudent polar protein than for the no polar protein (Table 5.4). In qualitative agreement with the free energies of unfolding determined from chemical denaturation, using the Gibbs-Helmholtz equation to calculate the free energies of unfolding from the calorimetric data indicates that the prudent polar

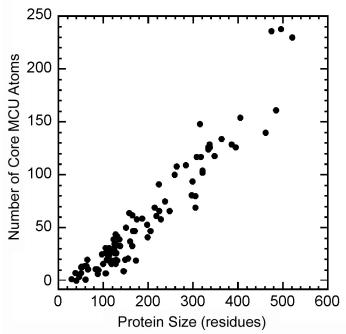
protein is thermodynamically stabilized over a temperature range from 0  $^{\circ}$ C to the  $T_{\rm m}$  (data not shown).

#### Discussion

The folding of a protein into a compact globular form requires the burial of both polar and non-polar atoms. Both statistical surveys<sup>21</sup> and mutational experiments<sup>17-20</sup> indicate the energetic benefit of forming intra-molecular hydrogen bonds between polar atoms that are buried from solvent. The energetic necessity of forming intra-molecular hydrogen bonds between buried polar atoms is also evident in the known structures of integral membrane proteins. In these proteins, the trans-membrane regions form either  $\beta$ -barrel structures of  $\alpha$ -helices in which solvent-shielded main chain atoms form intra-molecular hydrogen bonds.

As in membrane proteins, the majority of buried polar atoms in folded globular protein structures are main chain atoms.<sup>21</sup> The formation of secondary structure creates hydrogen bonds between many buried main chain atoms in globular proteins; however, some polar main chain atoms that orient towards the core are not involved in main chain-main chain hydrogen bonds. We note a strong positive correlation (Fig. 5.5) between protein size and the number of core MCU atoms (main chain atoms that aren't involved in hydrogen bonds to other main chain atoms). Polar side chains at core positions that are able to form intra-molecular hydrogen bonds to core MCU atoms may energetically favor the folded state, provided they also largely fulfill their own hydrogen bond potential. In our database survey, we note that the majority of polar side chains at core positions do form hydrogen bonds to MCU atoms (Table 5.1). As protein size increases and the

number of MCU atoms increases (Fig. 5.5) so should the demand for polar side chains as hydrogen bond partners. This may partly account for the strong positive correlation that has been noted between protein size and the number of polar side chains at core positions.<sup>16</sup>



**Figure 5.5** Correlation of the number of core MCU atoms with protein size. The number of main chain amide and carbonyl groups that are oriented towards the protein interior and are not involved in main chain hydrogen bonding increases as protein size increases. Satisfying the hydrogen bond potential of these atoms might in part explain the positive correlation between protein size and number of polar side chains at core positions. <sup>16</sup>

The group of amino acids with polar groups that tend to partition outside of the protein core (Ser, Thr, Asn, Gln, Asp, Glu, Lys, and Arg) make more hydrogen bonds to MCU atoms than the group that tends to partition inside the protein core (Tyr, Trp, Cys, and His). This can be rationalized based on the physical properties of the amino acid side chains. Compared to aliphatic side chains at core positions, members of the first group (Ser, Thr, etc.) energetically compete based primarily on their ability to form favorable polar interactions (H-bonds, electrostatics). In addition to polar properties, members of

the second group (Tyr, Trp, Cys, and His) have additional properties that may be important at core positions. Compared to aliphatic amino acids, Tyr and Trp have large hydrophobic surfaces that make them energetically beneficial for core packing. Cys can form covalent intra-molecular crosslinks that makes it energetically favorable. His contains two acid/base groups that titrate close to physiological pH, making it useful in active sites and controlling pH dependent protein dynamics.

Hydrogen bond rules for polar residues at core positions were constructed based on the database survey. These rules require that polar side chains at core positions hydrogen bond to MCU atoms as well as largely satisfy their own hydrogen bond potentials (Table 5.3). The relatively small number of hydrogen bonds to crystallographic water molecules by core polar residues (Table 5.1) may be due to the energetically unfavorable burial of the partner water molecules from bulk solvent. Due to the limited number of hydrogen bonds to crystallographic waters in our database survey and the likely energetic penalty of burying a water molecule, water mediated hydrogen bonds at core positions were not included in the prudent polar core design strategy. The prudent polar hydrogen bond requirements (Table 5.3) developed for core design will probably not transfer directly to boundary design because side chains at boundary positions are capable of interacting with solvent. However, the prudent polar strategy could be applied to boundary positions that are likely to be highly buried based on the generic classification scheme of Marshall and Mayo. <sup>13</sup>

Given the variable geometry seen for hydrogen bonds in protein structures,<sup>21</sup> as well as the energetic benefit<sup>17-20</sup> of forming hydrogen bonds to buried polar atoms, a geometrically softened hydrogen bond potential was used in combination with the

hydrogen bond requirements (Table 5.3) for core polar residues. For hydrogen bond geometry that meets minimum criteria, a hydrogen bond benefit of at least 6 kcal/mol was applied. In our force field, this minimum hydrogen bond benefit is slightly greater in magnitude than the energetic benefit of burying a methyl group. Thus, polar rotamers that meet the hydrogen bond rules (Table 5.3) will be able to energetically compete with hydrophobic residues in the design of protein cores.

For the thioredoxin core design target, the significant reduction in combinatorial complexity of the prudent polar strategy relative to the unrestricted polar strategy indicates that the hydrogen bond requirements (Table 5.3) effectively eliminated many polar side chain rotamers from consideration. This decrease in combinatorial complexity explains to the faster optimization trajectory for the prudent polar strategy compared to the unrestricted polar strategy (Fig. 5.1).

For the thioredoxin core, the initial combinatorial complexity of the prudent polar strategy was slightly greater than that of the no polar strategy. However, the computational search was moderately faster for the prudent polar strategy than the no polar strategy. The energetic contributions of the prudent polar strategy may simplify the search, allowing it to go faster. The Dead End Elimination (DEE) search algorithm<sup>42</sup> eliminates a rotamer r at position i if a rotamer t at positions i exists such that:

$$E(i_r) - E(i_t) + \sum_{j,j \neq i} \{ \min_{u} [E(i_r j_u) - E(i_t j_u)] \} > 0$$

where  $E(i_r)$  and  $E(i_t)$  are the rotamer-template energies of rotamer r and t respectively and  $E(i_rj_u)$  and  $E(i_tj_u)$  represent rotamer-rotamer energies. If all rotamer-rotamer interactions are zero, then the search simplifies to finding, at each position, the rotamer with the best rotamer-template interaction:

$$E(i_r) - E(i_t) > 0$$

Due to the rotamer-template hydrogen bond requirements and the softened hydrogen bond potential, the prudent polar strategy includes polar residues with highly favorable rotamer-template interactions, possibly shifting the energy balance in favor of rotamer-template interactions over rotamer-rotamer interactions, and simplifying the search.

Compared to the no polar sequence, the prudent polar sequence recovers three wt residues (Y49, T66, T77) involved in intra-molecular hydrogen bonding. These results suggest that the wt thioredoxin core is highly optimized to form hydrogen bonds between buried atoms. Our database survey as well as the survey of McDonald and Thornton<sup>21</sup> indicates that naturally occurring globular proteins tend to form hydrogen bonds to the majority of buried atoms. Given this observation, a reasonable design strategy would leave wt core polar residues fixed and optimize the identity of core positions whose wt identity is hydrophobic. This strategy should yield protein sequences that form well-folded structures. However, given the highly-coupled nature of protein cores, optimizing the identity of all core side chains using a reasonable strategy for modeling core polar residues may identify a GMEC solution with improved thermodynamic stability.

In addition, prudent polar modeling should be well suited to the core design of flexible 11,43 and *de novo* designed 44,45 backbones. Alternative main chain conformations generated by flexible backbone modeling may not allow buried hydrogen bond interactions that exist in the wt structure. For *de novo* designed backbones, there is no wt molecule to use as a basis for determining which core positions to fix as polar. The prudent polar strategy is independent of wt sequence, making it a good strategy for including polar rotamers in core design of flexible and *de novo* designed backbones.

The sequences predicted for the thioredoxin core by the no polar and prudent polar strategies both form folded protein structures similar to wt based on the CD spectra (Fig. 5.4). Differences in the number of Tyr residues in each protein may explain the slight variation seen between the CD spectra.

The increase in thermodynamic stability of the prudent polar and no polar designed proteins relative to the wt protein (Table 5.4) can largely be explained by the D26I mutation. The previously described D26A mutation results in a protein that is thermodynamically stabilized relative to wt. 46 This increase in thermodynamic stability is attributed to the removal of a charged polar group whose hydrogen bond potential is largely unsatisfied from the protein core. 46 This also explains the similar increase in thermodynamic stability of the D26I mutation. <sup>16</sup> The thermodynamic stability of D26I thioredoxin at 25 °C (12.1 kcal/mol)<sup>16</sup> is slightly greater than that of prudent polar thioredoxin (11.5 kcal/mol). This indicates that the L58I and Y70F mutations in prudent polar thioredoxin are slightly thermodynamically destabilizing relative to D26I thioredoxin. The physical basis for this slight destabilization is not readily apparent from analysis of the native thioredoxin structure, and may be due to interactions in the denatured state. While the prudent polar thioredoxin molecule is slightly less stable than a molecule with the D26I mutation alone, in terms of evaluating core protein design methods, the relevant comparison is between prudent polar thioredoxin and no polar thioredoxin.

The 2.2 kcal/mol increase in thermodynamic stability at 25 °C for prudent polar thioredoxin compared to no polar thioredoxin indicates the advantage of the prudent polar strategy over the previously used no polar strategy for core protein design. The energetic

difference between the prudent polar and no polar molecules can largely be accounted for by the interactions of T66 and T77 in prudent polar thioredoxin (L66 and A77 in no polar thioredoxin). Both T66 and T77 form hydrogen bonds to buried atoms in prudent polar and wt thioredoxin. Individual mutations at these positions to hydrophobic identity in the wt background each result in about a kcal/mol decrease in thermodynamic stability.<sup>16</sup>

### **Conclusions**

The prudent polar strategy improves core protein design relative to the no polar strategy by including polar side chain rotamers when they are likely to improve thermodynamic stability. The hydrogen bond requirements eliminate most core polar rotamers from consideration, resulting in a tractable increase in combinatorial complexity. The prudent polar strategy should improve core design for large protein applications, as larger proteins tend to have a greater number of core MCU atoms. The prudent polar strategy is well suited to core design with flexible and *de novo* designed backbones for which wt polar contacts are undetermined. In addition to core protein design, the prudent polar strategy may be useful in the modeling of other design targets that include buried polar groups such as protein-protein and protein-small molecule interfaces including computational enzyme design.

## References

- 1. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci* **5**(5), 895-903.
- 2. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Curr Opin Struct Biol* **9**(4), 509-513.
- 3. Nozaki, Y. & Tanford, C. (1971). The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J Biol Chem* **246**(7), 2211-2217.
- 4. Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation.

  \*Adv. Protein Chem. 14, 1-63.
- 5. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry* **29**, 7133-7155.
- 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(13), 5803-5807.
- 7. Desjarlais, J. R. & Handel, T. M. (1995). De novo design of the hydrophobic cores of proteins. *Protein Sci* **4**(10), 2006-2018.
- 8. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc Natl Acad Sci U S A* **94**(19), 10172-10177.
- 9. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science* **278**, 82-87.
- Lazar, G. A., Desjarlais, J. R. & Handel, T. M. (1997). De novo design of the hydrophobic core of ubiquitin. *Protein Sci* 6(6), 1167-1178.

- 11. Desjarlais, J. R. & Handel, T. M. (1999). Side-chain and backbone flexibility in protein core design. *J Mol Biol* **290**(1), 305-318.
- 12. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Structure Fold Des* **7**(5), R105-109.
- 13. Marshall, S. A. & Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* **305**(3), 619-631.
- 14. Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J Mol Biol* **105**, 1-14.
- 15. Lesser, G. J. & Rose, G. D. (1990). Hydrophobicity of amino acid subgroups in proteins. *Proteins* **8**(1), 6-13.
- Bolon, D. N. & Mayo, S. L. (2001). Polar residues in the protein core of Escherichia coli thioredoxin are important for fold specificity. *Biochemistry* 40(34), 10047-10053.
- Shirley, B. A., Stanssens, P., Hahn, U. & Pace, C. N. (1992). Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry* 31, 725-732.
- 18. Byrne, M. P., Manuel, R. L., Lowe, L. G. & Stites, W. E. (1995). Energetic contribution of side chain hydrogen bonding to the stability of staphylococcal nuclease. *Biochemistry* **34**, 13949-13960.
- 19. Yamagata, Y., Kubota, M., Sumikawa, Y., Funahashi, J., Takano, K., Fujii, S. & Yutani, K. (1998). Contribution of hydrogen bonds to the conformational stability of human lysozyme: calorimetry and x-ray analysis of six tyrosine --> phenylalanine mutants. *Biochemistry* 37, 9355-9362.

- 20. Takano, K., Yamagata, Y., Kubota, M., Funahashi, J., Fujii, S. & Yutani, K. (1999).
  Contribution of hydrogen bonds to the conformational stability of human lysozyme:
  calorimetry and x-ray analysis of six Ser --> Ala mutants. *Biochemistry* 38, 6623-6629.
- 21. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**(5), 777-793.
- 22. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of sidechain amide orientation. *J Mol Biol* **285**(4), 1735-1747.
- 23. Dahiyat, B. I., Sarisky, C. A. & Mayo, S. L. (1997). De novo protein design: towards fully automated sequence selection. *J Mol Biol* **273**(4), 789-796.
- 24. Katti, S. K., LeMaster, D. M. & Eklund, H. (1990). Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution. *J Mol Biol* **212**(1), 167-184.
- 25. Dunbrack, R. L., Jr. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side- chain prediction. *J Mol Biol* **230**(2), 543-574.
- 26. Mayo, S. L., Olafson, B. D. & Goddard, W. A., 3rd. (1990). DREIDING: a generic force field for molecular simulations. *J Phys Chem* **94**, 8897-8909.
- 27. Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci* **6**(6), 1333-1337.
- 28. Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709-713.
- 29. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**(3), 379-400.

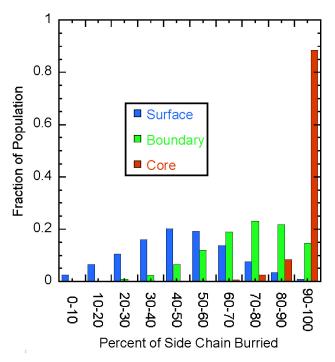
- 30. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Fold Des* **3**(4), 253-258.
- 31. Chong, L. T., Dempster, S. E., Hendsch, Z. S., Lee, L. P. & Tidor, B. (1998). Computation of electrostatic complements to proteins: a case of charge stabilized binding. *Protein Sci* **7**(1), 206-210.
- 32. Desmet, J., Maeyer, M. D., Hazes, B. & Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* **356**, 539-542.
- 33. Pierce, N. A., Spriet, J. A., Desmet, J. & Mayo, S. L. (2000). Conformational splitting: a more powerful criterion for dead-end elimination. *J Comput Chem* **21**, 999-1009.
- 34. Hemsley, A., Arnheim, N., Toney, M. D., Cortopassi, G. & Galas, D. J. (1989). A simple method for site-directed mutagenesis using the polymerase chain reaction. *Nucleic Acids Res* **17**(16), 6545-6551.
- 35. Holmgren, A. & Reichard, P. (1967). Thioredoxin 2: cleavage with cyanogen bromide. *Eur J Biochem* **2**(2), 187-196.
- 36. Kelley, R. F., Shalongo, W., Jagannadham, M. V. & Stellwagen, E. (1987). Equilibrium and kinetic measurements of the conformational transition of reduced thioredoxin. *Biochemistry* **26**(5), 1406-1411.
- 37. Santoro, M. M. & Bolen, D. W. (1988). Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl alphachymotrypsin using different denaturants. *Biochemistry* **27**(21), 8063-8068.

- 38. Privalov, G., Kavina, V., Freire, E. & Privalov, P. L. (1995). Precise scanning calorimeter for studying thermal properties of biological macromolecules in dilute solution. *Anal Biochem* **232**(1), 79-85.
- Becktel, W. J. & Schellman, J. A. (1987). Protein stability curves. *Biopolymers* 26(11), 1859-1877.
- 40. Santoro, M. M. & Bolen, D. W. (1992). A test of the linear extrapolation of unfolding free energy changes over an extended denaturant concentration range. *Biochemistry* 31(20), 4901-4907.
- 41. Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* **14**(1), 51-55, 29-32.
- 42. Goldstein, R. F. (1994). Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys J* **66**(5), 1335-1340.
- 43. Su, A. & Mayo, S. L. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci* **6**(8), 1701-1707.
- 44. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. (1998). High-resolution protein design with backbone freedom. *Science* **282**, 1462-1467.
- 45. Nauli, S., Kuhlman, B. & Baker, D. (2001). Computer-based redesign of a protein folding pathway. *Nat Struct Biol* **8**(7), 602-605.
- 46. Langsetmo, K., Fuchs, J. A. & Woodward, C. (1991). The conserved, buried aspartic acid in oxidized Escherichia coli thioredoxin has a pKa of 7.5. Its titration produces a related shift in global stability. *Biochemistry* **30**(30), 7603-7609.
- 47. Bolon, D. N. & Mayo, S. L. (2001). Enzyme-like proteins by computational design. *Proc Natl Acad Sci U S A* **98**(25), 14274-14279.

## Appendix A

Residue Classification is a Good Predictor of Burial

In the analyses of polar interactions (Chapters 4 & 5), residues at core positions are considered to be buried, residues at surface positions are considered to be in contact with solvent, and residues at boundary positions are considered to be able to be buried or in contact with solvent. In order to validate this argument, residue classification and solvent accessible surface area (ASA) have been determined for all side chains in a previously published set of high quality protein structures.<sup>1</sup>



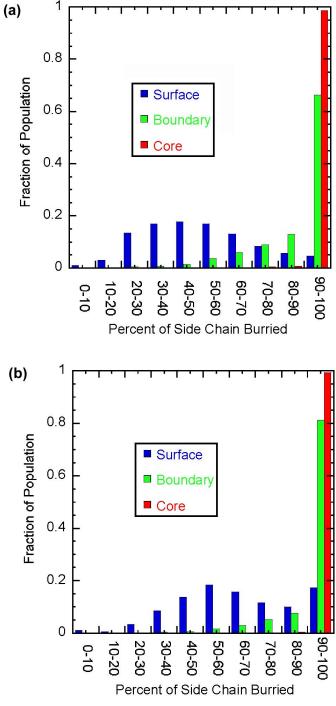
**Figure A.1** Percent ASA burial of side chains classified as surface, boundary, or core. All side chains with the exception of glycine are included.

Residue classification was performed as described in Chapter 5. ASA was calculated with the Connolly algorithm,<sup>2</sup> using the Lee and Richards definition<sup>3</sup> with a probe radius of zero, atomic radii from the DREIDING force field,<sup>4</sup> and an atomic add on radius of 1.4 Å. The fraction buried was the ratio of ASA in the folded state (surfacing all side chain atoms while excluding all other atoms) to ASA in the reference state

(surfacing all side chain atoms while excluding only the local backbone N, C,  $C^{\alpha}$ , and O atoms).

The vast majority (88%) of side chains at core positions bury greater than 90% of their ASA (Fig. A.1), indicating that core classification does signify that a side chain will likely be buried. The percent burial of surface side chains appears to be normally distributed with a mean of about 50% buried (Fig. A.1). Less than 1% of surface side chains bury greater than 90% of their ASA. Therefore, surface classification does indicate that a side chain is likely to interact with solvent. The fractional burial distribution of side chains at boundary positions is intermediate between that of side chains at surface and core positions (Fig. A.1).

The fractional burial distributions for main chain carbonyl oxygen and amide hydrogen atoms (Fig. A.2) also indicate that atoms at core positions are more likely to be buried than atoms at surface positions. Compared to side chains, carbonyl oxygen and amide hydrogen atoms bury a greater fraction of their ASA. This is likely due to their small total ASA and their close proximity to local side chains. Glycine side chain hydrogen atoms also bury a greater fraction of their ASA relative to larger side chains (data not shown).



**Figure A.2** Percent ASA burial of main chain carbonyl oxygen (a) and amide hydrogen (b) atoms classified as surface, boundary or core.

## References

- 1. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol* **238**(5), 777-793.
- 2. Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709-713.
- 3. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* **55**(3), 379-400.
- 4. Mayo, S. L., Olafson, B. D. & Goddard, W. A., 3rd. (1990). DREIDING: a generic force field for molecular simulations. *J Phys Chem* **94**, 8897-8909.

## Appendix B

Variation in Statistical Secondary Structure Propensity is Reduced at Core Positions Relative to at Surface Positions

Arthur Street re-compiled the data set of protein structures analyzed in this chapter and calculated main chain dihedral angles as well as residue classification.

Early efforts in protein fold prediction focused on identifying amino acid sequences likely to form either  $\alpha$  or  $\beta$  secondary structure. Statistical analyses of the distribution of the 20 natural amino acids in the different secondary structure elements, based on a set of known protein structures, have been used to calculate intrinsic secondary structure propensities. In addition, experimental propensities have been determined based upon thermodynamic analysis of a set of protein molecules, each with a different amino acid at a host site. Model systems of this nature have been developed for  $\alpha$  helical Analysis and  $\beta$  strand St

The total energetic effect of a mutation can be expressed as the sum of an intrinsic secondary structure propensity and a context dependent term:

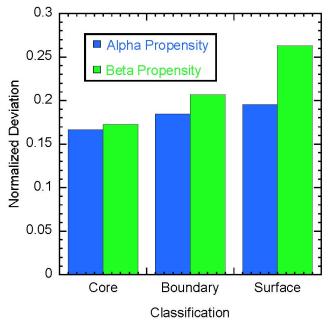
$$E_{total} = E_{intrinsic} + E_{context}$$

where  $E_{intrinsic}$  is independent and  $E_{context}$  is dependent on the rest of the protein sequence.  $E_{context}$  includes intra-molecular hydrogen bond, electrostatic, and van der Waal's interactions as well as solvation effects.  $E_{intrinsic}$  may include entropic contributions from the side chain as well as the local main chain.

The experimental systems developed to quantify secondary structure propensity can determine the total energetic effect of a mutation. However, it is not straightforward to measure the individual energetic contributions from intrinsic and context dependent terms. For this reason, it is desirable to control the environment local to the host site in order to minimize context dependent effects in the experimental systems.<sup>7,11</sup>

Assuming that protein sequences are highly optimized, the probability of an amino acid at a given position in a natural occurring protein should be related to the total functional effect of the amino acid. Under the further assumption that the primary

functional effect at most amino acid positions is to energetically favor the folded state, the probability of an amino acid at a given position should be related to the total energetic effect of the residue. Therefore, context dependent effects may influence secondary structure propensities derived from statistical analysis. In order to test this hypothesis, statistical secondary structure propensities were determined at surface and core positions. Core positions tend to have greater intra-molecular contact compared to surface positions (Appendix A) and their total energy should therefore have a larger context dependent contribution. This context dependent contribution should mask the intrinsic contribution and attenuate statistical propensity at core positions.

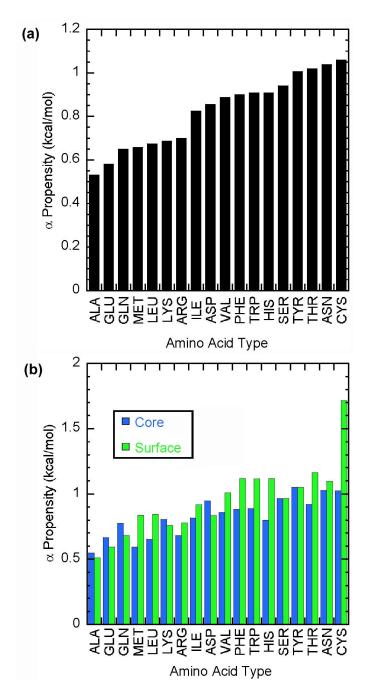


**Figure B.1** Normalized deviations of statistically derived secondary structure propensity at surface and core positions. In the core, both  $\alpha$  and  $\beta$  propensities appear attenuated relative to the surface.

A statistical analysis was performed on a previously published set of protein structures.<sup>3</sup> Positions were classified as surface, boundary or core as described in Chapter 4. Secondary structure propensity was determined as previously published.<sup>3</sup> The secondary structure space  $(\alpha, \beta, \text{ or other})$  of each amino acid was determined based on its

main chain dihedral angles. For each amino acid, the propensity to occupy each secondary structure space was calculated as the ratio of residues in the specific secondary structure space to the total number of residues. This statistical propensity was then converted to a free energy assuming a temperature of 300K ( $\Delta G = -RT \ln K$ ). In order to compare variation in propensity at surface and core positions, a normalized deviation was calculated as the ratio of the standard deviation to the mean (Fig. B.1).

This comparison indicates that the spread in  $\alpha$  and  $\beta$  propensity values is attenuated at core positions relative to surface positions. This result suggests that the relative energetic contribution of context dependent terms is greater in the core. This result is also consistent with a reduction in the absolute intrinsic energetic contribution. In  $\alpha$  helices, for example, the physical basis for propensity has been attributed to residual side chain entropy in the folded state. Residual side chain entropy at core positions may be reduced relative to residual side chain entropy at surface positions, resulting in an attenuation in the absolute intrinsic energetic contribution.



**Figure B.2**  $\alpha$  propensity calculated for all amino acids (a) and for amino acids at surface and core positions (b).

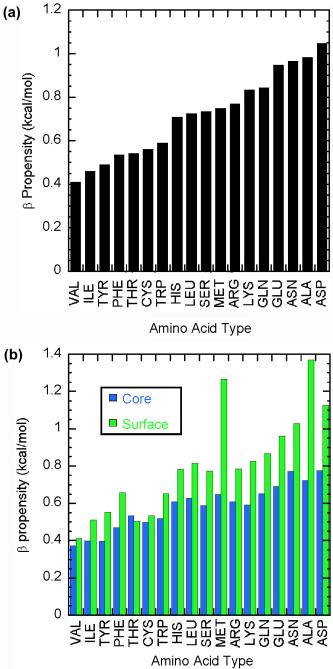


Figure B.3  $\,\beta$  propensity calculated for all amino acids (a) and for amino acids at surface and core positions (b).

### References

- 1. Chou, P. Y. & Fasman, G. D. (1978). Empirical predictions of protein conformation. *Annu Rev Biochem* **47**, 251-276.
- 2. Kim, P. S. & Baldwin, R. L. (1982). Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu Rev Biochem* **51**, 459-489.
- 3. Munoz, V. & Serrano, L. (1994). Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. *Proteins* **20**(4), 301-311.
- 4. O'Neil, K. T. & DeGrado, W. F. (1990). A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* **250**, 646-651.
- 5. Horovitz, A., Matthews, J. M. & Fersht, A. R. (1992). Alpha-helix stability in proteins. II. Factors that influence stability at an internal position. *J Mol Biol* **227**(2), 560-568.
- 6. Blaber, M., Zhang, X. J. & Matthews, B. W. (1993). Structural basis of amino acid alpha helix propensity. *Science* **260**, 1637-1640.
- 7. Chakrabartty, A., Kortemme, T. & Baldwin, R. L. (1994). Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci* **3**(5), 843-852.
- 8. Kim, C. A. & Berg, J. M. (1993). Thermodynamic beta-sheet propensities measured using a zinc-finger host peptide. *Nature* **362**, 267-270.

- 9. Minor, D. L., Jr. & Kim, P. S. (1994). Measurement of the beta-sheet-forming propensities of amino acids. *Nature* **367**, 660-663.
- 10. Minor, D. L., Jr. & Kim, P. S. (1994). Context is a major determinant of beta-sheet propensity. *Nature* **371**, 264-267.
- 11. Smith, C. K., Withka, J. M. & Regan, L. (1994). A thermodynamic scale for the beta-sheet forming tendencies of the amino acids. *Biochemistry* **33**(18), 5510-5517.
- 12. Creamer, T. P. & Rose, G. D. (1992). Side-chain entropy opposes alphahelix formation but rationalizes experimentally determined helix-forming propensities. *Proc Natl Acad Sci U S A* **89**(13), 5937-5941.

# Appendix C

**Summary of Thermodynamic Data** 

Table C.1 Experimental thermodynamic data

	Tm	$\Delta H_{vh}$	$\Delta H_{\text{cal}}$	β	$\Delta G$	$C_{\text{m}}$	m-value
Protein	(°C)	(kcal	mol <sup>-1</sup> )		(kcal mol <sup>-1</sup> )	(M)	(kcal mol <sup>-1</sup> M <sup>-1</sup> )
wt oxidized	87	117	117	1.00	8.8	2.58	3.4
wt reduced	76	97.0	103	1.06	6.2	1.77	
C32S/C35S	77	101	105	1.01	6.7	1.89	3.5
C32A/C35A	71	86.0	89.4	1.04	5.0	1.46	3.4
C32V/C35A	67	74.3	80.3	1.08	4.7	1.37	3.4
C32A/C35S	73	91.6	92.3	1.01	5.1	1.47	3.5
C32A reduced	74	94.2	95.2	1.01	5.2	1.56	3.4
C32S/C35A	75	104	104	1.00	6.1	1.77	3.5
C35A reduced	73	98.7	96.9	0.98	5.8	1.66	3.5
D26I	98	159	157	0.99	12.1	3.67	3.3
D26I reduced	84	152	149	0.98	10.2	2.97	3.4
D26A	96	139	139	1.00			
T66L	85	102	105	1.03	7.5	2.43	3.1
T77V	82	121	120	0.99	7.6	2.50	3.1
T66L/T77V	80	100	102	1.02	7.1	2.30	3.1
T66L/T77V reduced	64	84.7	85.9	1.01	4.4	1.44	3.0
AALV	55	48.7	49.0	1.01	2.8	1.01	2.8
Core5 <sup>a</sup>	66	65.7	60.2	0.92	3.8	2.82	1.3
Core6 <sup>b</sup>	89	111	112	1.01	7.7	3.37	2.3
Core8 <sup>c</sup>	66	64.0	63.6	0.99	4.7	3.18	1.5
Core8A <sup>d</sup>	93	104	103	0.99	6.6	3.54	1.9
2012011	,,,	101	105	0.77	0.0	5.5 1	1.,
Core_nopolar <sup>e</sup>	97	106	106	1.00	9.3	3.22	2.9
Core_prudent_polar <sup>f</sup>	97	148	147	0.99	11.5	3.41	3.4
g							
HisPatch <sup>g</sup>	81	118	118	1.00			
6Q	85	117	117	1.00			
K57F	84	119	121	1.02	8.7	2.68	3.3
D26I/K57F	04	117	141	1.02	12.1	3.90	3.1
D201/K3/1					12.1	3.70	J.1

 $^1\text{Midpoint}$  of calorimetric unfolding.  $^2\Delta H_{vh}$  Van Hoff's enthalpy.  $^3\Delta H_{cal}$  Calorimetric enthalpy.  $^4\beta$   $\Delta H_{cal}/\Delta H_{vh}$ .  $^5\text{Free}$  energy of unfolding at 25 °C.  $^6\text{Midpoint}$  of guanidinium chloride unfolding transition.  $^7\text{Slope}$  of  $\Delta G$  versus denaturant concentration plots.  $^a\text{D26I/C32A/C35A/T66L/T77V}.$   $^b\text{I4L}$ , D26I, Y49F, K57I, L58I, Y70F.  $^c\text{D26I}$ , C32A, C35A, Y49F, L58I, T66L, Y70F, T77V.  $^d\text{I4L}$ , D26I, Y49F, K57I, L58I, T66L, Y70F, T77A.  $^f\text{D26I}$ , Y49F, L58I, T66L, Y70F, T77A.  $^f\text{D26I}$ , L58I, Y70F.  $^g\text{E30H}$ , Q62H (JBC **271**:5059). All thermodynamic data were obtained in 50 mM NaPi, pH 7.0. Guanidinium denaturation experiments were performed at 25 °C with a protein concentration of 5 μM and where reduction is indicated the buffer was supplemented with 100 μM DTT. DSC experiments were performed with a protein concentration of around 1 mg/ml (≈100 μM) and where reduction is indicated the buffer was supplemented with 2 mM DTT.