Protein-Ligand Interactions: Docking, Design and Protein Conformational Change

Thesis by

Deepshikha Datta

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

California Institute of Technology

Pasadena, California, USA

2002

(Defended on December 16, 2002)

Abstract

Virtual ligand screening has proven to be a successful strategy in drug design. An in house-developed procedure (HierDock), a coarse grain docking method followed by a fine grain search procedure, was used to determine the binding site for sugars in the outer membrane protein A in *E.coli*, a key interaction in the pathogenesis of neonatal meningitis. These results are being further extended in suggesting possible peptide antagonists and drugs for therapeutic strategies.

Prediction of binding site of ligands in proteins, starting with the apo-protein is one of the challenges in the field of virtual ligand screening. HeirDock was modified for accurately predicting the ligand binding sites in apo-proteins that undergoes significant structural changes on binding to a ligand. The method was evaluated for finding the binding site for methionine in methionyl tRNA synthetase. We followed up on our understanding of binding mechanism in aminoacyl tRNA synthetases by attempting to design these enzymes to bind to non-natural amino acids. Using the computational protein design software (ORBIT), a phenylalanyl-tRNA synthetase variant that allows efficient *in vivo* incorporation of aryl ketone functionality into proteins was designed.

Ligand-induced conformation changes are commonly seen in proteins. We have developed a procedure by combining computational protein design with methods from mean-field theory to design protein sequences capable of switching between two completely different protein folds on chelating to metal. This method is potentially useful in characterizing protein sequence-structure relationships.

Acknowledgements

I have been inspired and encouraged by numerous people who have helped me in immeasurable ways to complete this thesis. First and foremost, I would like to thank my thesis advisor, Dr. Stephen Mayo, who has provided generous support throughout my stay at Caltech. I am very thankful to him for providing me the intellectual freedom to work on whatever research I found interesting. I consider myself very fortunate to have been a part of his research group.

I would also like to thank the members of my thesis committee, Dr. William Goddard, Dr. David Tirrell and Dr. Richard Roberts, for their active participation and feedback. I have had the unique opportunity to work closely with Dr. Goddard and Dr. Tirrell and have learned immensely through my interactions with them and with their group members.

Caltech's collaborative and interdisciplinary environment allowed me to collaborate with various group, both within Caltech and outside. Dr. Goddard provided me with the chance to work closely with the members of the Scripps Research Institute. In particular, I would like to thank Dr. Richard Lerner, Dr. Paul Wentworth and Dr. Anita Wentworth at Scripps for suggesting the antibody project and for participating in the various discussions that followed. I would like to thank Dr. Nemani Prasadrao at the Children's Hospital at USC and Dr. Kim Kwang at the John Hopkin's Medical Institute for their inputs on the neonatal mengingitis project. I am grateful to Pin Wang, Isaac Carrico and Dr. David Tirrell at Caltech for their efforts in making the tRNA synthetase project successful. Thanks to Blaine Mooers, Walter Basse and Professor Brian Matthews at the University of Oregon for putting in the time and effort in completing the

T-4 lysozyme project. Thanks to Dr. Nagarajan Vaidehi, David Zhang, Dr. Wely Floriano, and Dr. Xin Xu at the Material Simulation Center at Caltech for their contributions in numerous projects.

I am very appreciative of the good camaraderie that I have enjoyed in the Mayo lab. Thanks to both past and present members of the group with whom I have had a chance to interact. In particular, I would like to thank those with whom I directly worked on various projects: Kirsten Lassila, Arthur Street, Ben Gordon and Chris Voigt. Also, many thanks to Cynthia Carlson and Rhonda Digusto for their help on various occasions and to Marie Ary for her suggestions on scientific writing. I have really enjoyed the pleasant environment in the Material Simulation Center at Caltech. I am very grateful to Dr. Goddard and Dr. Nagarajan Vaidehi for always a being available to give me advice, guidance and encouragement.

Table of Contents

Appendices

List of Figures

List of Tables

Section I

Virtual Ligand Screening

Chapter 1

Introduction to Molecular Docking and Virtual Ligand Screening

Protein interactions with ligands, other proteins, or surfaces are controlled by a complex array of intermolecular interactions. Such interactions depend both on the specific interactions in the binding site as well as the non-specific forces outside the binding pocket. This interplay of specific and non-specific forces controls all protein interactions ranging from bimolecular collisions in solutions to adhesion between cells. The complexity of interactions between proteins and flexible target molecules, including other proteins, nucleic acids and small molecules, is often determined by the considerable flexibility of the protein binding sites and by the structural rearrangements that occur upon binding of the associated molecule.

A goal of many biophysical studies is to determine the molecular forces that control biological interactions and to use this information to rationally manipulate protein function by modifying the protein, the interacting ligand, or both. The forces that control protein behavior and their physical chemical origins are inferred from equilibrium binding kinetic measurements or are calculated with molecular models. Calculated energies are used to identify the role of the physical and chemical interactions in protein function and behavior. Although detailed calculations are feasible for small molecules, such calculations become prohibitive as the size and complexity of the biological macromolecules increase. Time-dependent forces between soft or mobile species add yet another degree of complexity, while static models of interactions do not describe the full range of parameters that influence biological behavior.

As we approach the post-sequencing phase of many genome projects, it is estimated that the number of potential drug targets will increase from about 500 to 5000-10000 in the next few years (Drews, 2000). The scope and the importance of structure-based drug design will also increase significantly. Virtual ligand screening using *in silico* methods can provide prospective leads and is a practical alternative to high-throughput screening of large compound libraries provided the binding modes and affinities of the distinct ligands can be predicted correctly. The docking and scoring problems countered in this endeavor are central to the theory of bimolecular interactions and are ultimately determined by the nature of the underlying binding energy landscape. However, the desired synergy of adequate conformational sampling combined with accurate evaluation of energetics has been difficult to achieve with any computational model.

Computational structure prediction of ligand-protein complexes using docking methods, like DOCK, FLEXx, and GOLD (Ewing *et al.*, 2001; Jones *et al.*, 1997; Kramer *et al.*, 1999), in combination with empirical scoring functions are used to predict ligand orientations in binding sites and binding affinities of ligands to proteins. While the binding geometries depend on the docking methods, binding energy estimates rely heavily on the potential functions used to calculate them. Knowledge based potentials follow rules based on statistical analysis of binding affinities and geometries of experimentally determined protein-ligand complexes. These rules are converted to "pseudo-potentials" which are then applied to score computer generated ligand orientations (Gohlke *et al.*, 2000; Muegge & Martin, 1999). A major concern with such methods is that they might only select for those

orientations that have been observed in the crystal structures used to derive the potential.

Regression-based scoring functions estimate binding affinities by adding up interaction terms derived from the weighted structural parameters of the complexes. The weights are assigned by regression methods by fitting predicted and experimentally determined affinities to a given set of training complexes (Bohm and Stahl, 1999). A concern with these methods is the dependence on the size, composition, and generality of the training set used to derive the weights (Tame, 1999). Moreover, such methods can only interpolate and thus, are unable to identify new molecular scaffolds that are not present in the training set. Nevertheless, improvements in regression-based methods have contributed to some encouraging examples demonstrating the potential of such techniques (Rognan *et al.*, 1999).

First-principle-based approaches also approximate binding free energy by adding up individual contributions of different interactions. However, the individual energy terms are derived from physico-chemical theory and are not determined by fitting to experimental affinities. In most cases, gas phase molecular mechanical contributions are combined with solvation free energies. Evaluation of solvation energy is a challenge both in terms of computational demands and accuracy (Massova and Kollman, 2000). The methods used to calculte solvation include implicit solvent methods like Possion-Boltzmann and surface generalized Born methods (Shoichet *et al.*, 1999). The gas phase energy calculations depend on the type of the force field, for example, AMBER, CHARMM, DRIEDING (Brooks

et al., 1983; Mayo *et al.*, 1990). First-principle based approaches often calculate the correct order of the binding free energies but the numbers generally exceed the experimental values significantly.

Protein flexibility and the dynamics of inter molecular interfaces can regulate binding affinity and specificity in molecular recognition. It has been suggested that structural stability and flexibility during molecular recognition are associated with the ruggedness of the underlying binding energy landscape and can be related to various functions, such as specificity or permissiveness in recognition. However, predicting the correct substructures of the protein-ligand complexes is extremely difficult, especially in cases where the binding site of a flexible protein is unknown. Hierarchical approaches incorporating both ligand and protein flexibilities have contributed to recent progress in ligand-protein docking. Such procedures include multistage docking approaches and a hierarchy of energy functions that aim to capture the subtleties of protein flexibility on ligand binding.

Our method of virtual ligand screening, called HierDock (Figure I-1), uses flexible ligand docking using DOCK 4.0 as a coarse grain search followed by fine grain dynamics based on first principles. An advantage of using DOCK 4.0 is that its fragment-based ligand reconstruction scheme can effectively generate a large number of ligand conformations in a very short time using a fast, although crude scoring function based on van der Waal's and columbic interactions. The ligand ensemble generated from DOCK is then passed on to the next level: a finegrain molecular dynamics simulation where the ligand is allowed to optimize in the binding pocket. The fine grain optimization step involves simulations with implicit

solvent and also allows protein flexibility. Our results are verified with experimental observations aimed at improving the fine grain optimizations scoring techniques. Besides trying to improve the HierDOCK procedure, we have also applied the technology in trying to answer interesting biological questions and have found our current technology can be leveraged to understand pathogenesis of *E. coli* meningitis (Chapter 2). By combining docking with quantum chemistry, we have delineated an interesting biological mechanism by which antibodies oxidize water molecules to produce hydrogen peroxide (Chapter 3). In searching for binding site for small ligands in large, flexible globular proteins, we have incorporated modifications in both the coarse and fine grain levels and have been successful in finding the binding region of methionine in methionine tRNA synthetase (Chapter 4).

References

Bohm, H. J. & Stahl, M. (1999). Rapid empirical scoring functions in virtual screening applications. *Med Chem Res* 9, 445-462.

Brooks Br, B. R., Olafson Bd, States Dj, Swaminathan S, Karplus M. (1983). Charmm - A Program For Macromolecular Energy, Minimization, And Dynamics Calculations. *Journal Of Computational Chemistry* 4(2), 187-217.

Drews, J. (2000). Drug discovery: a historical perspective. *Science* 287(5460), 1960-4.

Ewing, T. J., Makino, S., Skillman, A. G. & Kuntz, I. D. (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* 15(5), 411-28.

Gohlke, H., Hendlich, M. & Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* 295(2), 337-56.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 267(3), 727-48.

Kramer, B., Rarey, M. & Lengauer, T. (1999). Evaluation of the FLEXX incremental construction algorithm for protein- ligand docking. *Proteins* 37(2), 228-41.

Massova, I. & Kollman, P. A. (2000). Combined molecular mechanical and continuum solven approach (MM-PBSA/GBSA) to predict ligand binding. *Persp Drug Discov Des*(18), 113-135.

Mayo S.L., O. D. D., Goddard W.A. III. (1990). Dreiding : A generic force field for molecular simulations. *J Phys Chem* 94, 8897-8909.

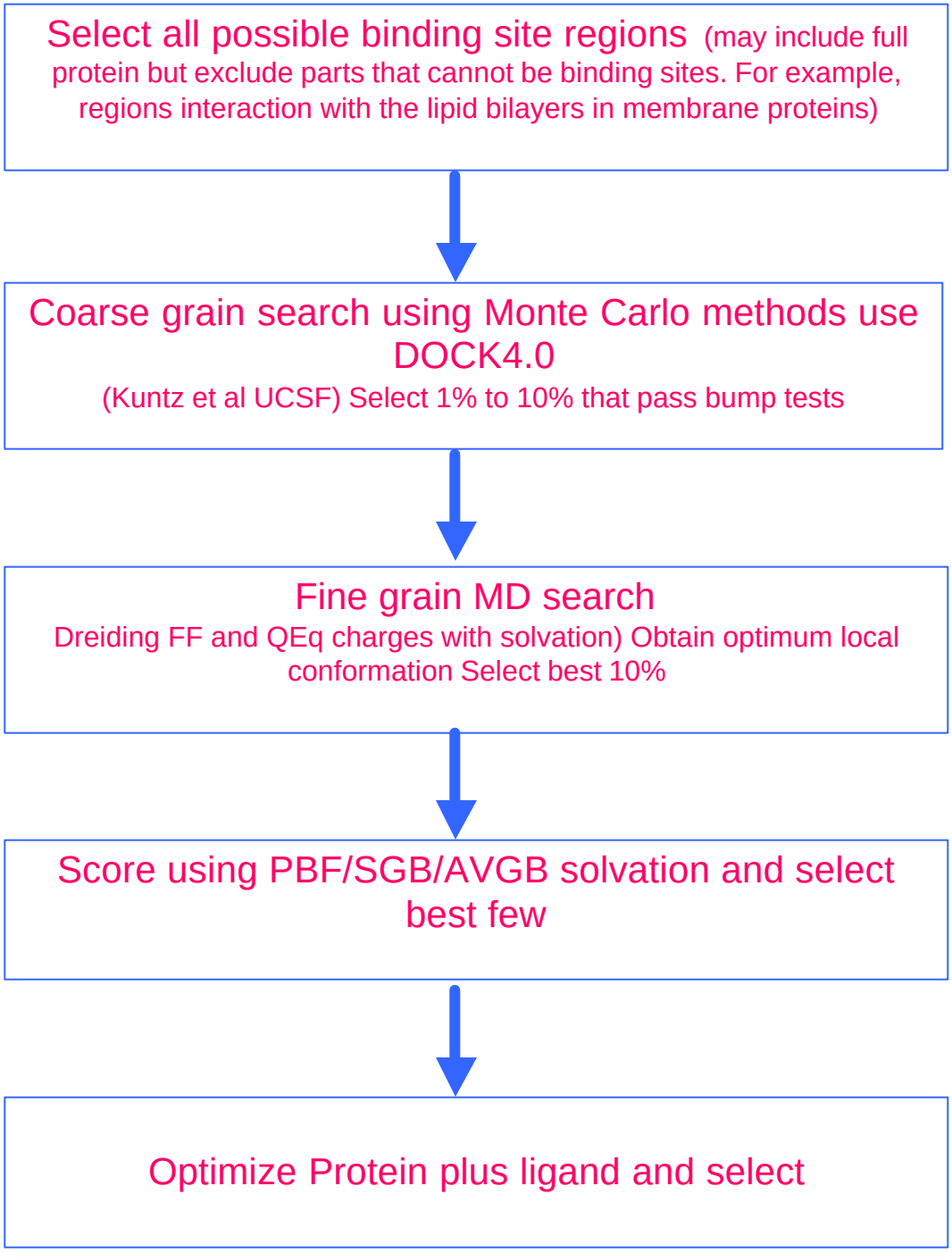
Muegge, I. & Martin, Y. C. (1999). A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem* 42(5), 791-804.

Rognan, D., Lauemoller, S. L., Holm, A., Buus, S. & Tschinke, V. (1999). Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem* 42(22), 4650-8.

Shoichet, B. K., Leach, A. R. & Kuntz, I. D. (1999). Ligand solvation in molecular docking. *Proteins* 34(1), 4-16.

Tame, J. R. (1999). Scoring functions: a view from the bench. *J Comput Aided Mol Des* 13(2), 99-108.

Figure I-1: A flowchart representing the steps of HierDock procedure.


Select all possible binding site regions (may include full protein but exclude parts that cannot be binding sites. For example, regions interaction with the lipid bilayers in membrane proteins)
Coarse grain search using Monte Carlo methods use DOCK4.0
(Kuntz et al UCSF) Select 1% to 10% that pass bump tests
Fine grain MD search
Dreiding FF and QEq charges with solvation) Obtain optimum local conformation Select best 10%
Score using PBF/SGB/AVGB solvation and select best few
Optimize Protein plus ligand and select

Chapter 2

Interaction of *E. coli* Outer Membrane Protein A with Sugars on the Receptors of the Brain Microvascular Endothelial Cells

Adapted from Datta D., Vaidehi N., Floriano W. B., Kim K. S., Prasadarao N. V. and Goddard III W. A, (2002); *Protein : Structure. Functions and Genetics (in press)*

ABSTRACT

E. coli, the most common gram-negative bacteria, can penetrate the brain microvascular endothelial cells (BMEC) during the neonatal period to cause meningitis. Experimental studies have shown that outer membrane protein A (OmpA) of *E. coli* plays a key role in the initial steps of the invasion process by binding to specific sugar moieties present on the glycoproteins of BMEC. These experiments also show that polymers of chitobiose (GlcNAcβ1,4-GlcNAc) block the invasion, while epitopes substituted with the L-fucosyl group do not.

We have used a hierarchy of coarse grain docking method with molecular dynamics (MD) to predict the binding sites and energies for interactions of GlcNAcβ1,4GlcNAc and other sugars with OmpA. The results suggest two important binding sites for the interaction of carbohydrate epitopes of BMEC glycoproteins to OmpA. We have identified one site as the binding pocket for chitobiose (GlcNAc1, 4GlcNAc) in OmpA, while the second region (including loop 1 and 2) may be important for recognition of specific sugars. We find that the site involving loops 1 and 2 has relative energies that correlate well with experimental observations. This theoretical study elucidates the interaction sites of chitobiose with OmpA and the binding site predictions are testable either by mutation studies or by invasion assays. These results can be further extended in suggesting possible peptide antagonists and drug design for therapeutic strategies.

Introduction

E. coli, a common gram-negative bacterium, causes meningitis during the neonatal period (1,2). The morbidity and mortality associated with this disease has remained significant with case fatality rates ranging from 15% to 40% of the infected neonates while ~50% of the survivors sustain neurological sequelae (1,3). Incomplete understanding of the pathogenesis and pathophysiology of *E. coli* meningitis has hampered the development of new therapeutic avenues, thus contributing to the high morbidity and mortality. For example, *E. coli* meningitis develops as a result of hematogenous spread. However, it is not clear how the circulating *E. coli* traverses across the blood-brain barrier, which contains a single cell lining of the brain microvascular endothelial cells (BMEC). Several cell surface structures (such as S-fimbriae) present in *E. coli* are involved in the pathogenesis of *E. coli* meningitis. However, one of the more important steps involves the interaction of OmpA with the BMEC. Prasadarao *et al* (4) showed that the expression of outer membrane protein A (OmpA) enhances the *E. coli* invasion of BMEC. Thus $OmpA^+$ *E. coli* strains invade BMEC with 50- to 100-fold higher frequency than $OmpA^-$ *E. coli* strains (4). Moreover, OmpA interacts with a 95 kDa BMEC glycoprotein for *E. coli* invasion, which is specifically expressed in endothelial cells of brain origin but not of systemic origin (N. V. Prasadarao private communication). To examine the specificity of interaction of the sugar moieties on Ecgp, the glycoprotein of BMEC, the BMEC was treated with wheat germ agglutinin (WGA), which blocked *E.coli* invasion (5). Since WGA is specific to binding of GlcNAcβ1-4GlcNAc epitopes, it was concluded that WGA binds to GlcNAc-β1-4GlcNAc epitopes

on the glycoprotein on BMEC, thus preventing invasion. However treatment of BMEC with WGA does *not* block the interaction of S-fimbriae (specific for NeuAc2,3-Galactose epitopes) binding to BMEC. This suggests that the inhibition is specific to chitobiose and not a mere steric hindrance by lectin. Other lectins such as ConA (specific to mannose) and AAL (specific to NeuAc2,3-Galactose) did not show such blocking activity (6). In addition, during invasion $OmpA^+$ *E. coli* induces actin filament rearrangement in BMEC, but this rearrangement is blocked significantly by both chitobiose and wheat germ agglutinin (7). On the other hand, $OmpA^-$ *E. coli* did not exhibit any effect on actin rearrangement (7). Moreover, chitobiose-sepharose chromatography shows binding of OmpA to chitobiose, suggesting that the observed inhibition is due to the direct interaction of OmpA with GlcNAcβ1-4GlcNAc epitopes.

Prasadarao and coworkers have further shown that masking of OmpA binding sites for GlcNAcβ1-4GlcNAc epitopes with these chito-oligomers significantly reduces the incidence of meningitis in newborn rat model of hematogenous meningitis, suggesting the biological relevance of this interaction (6). They tested simple disaccharides such as chitobiose, lactose, cellobiose and fucosyl substituted chitobiose for invasion assays and showed clearly that chitobiose blocks the invasion of BMEC by *E. coli* while lactose, cellobiose and fucosyl substituted chitobiose do not.

Although the GlcNAcβ1-4GlcNAc epitopes are universally present on many N-linked glycoproteins, studies by Prasadarao *et al*., indicate clearly that OmpA interaction with GlcNAcβ1-4GlcNAc epitopes on Ecgp is crucial for the establishment of the disease. It has been demonstrated (8-10) that the same sugar or amino acid present on different proteins can contribute to specificity of the pathogen (5,6). For example,

enteropathogenic *E. coli* binding to epithelial cells depends on type 1 fimbriae, which is specific to the mannose residue and can be blocked by mannose sugar (6). However, the same strain does not bind to BMEC very efficiently even though BMEC contain mannose residues. This suggests that either the protein sequence and conformation around the GlcNacβ1-4GlcNAc epitopes in Ecgp or the density of the carbohydrate-bearing proteins on BMEC differ from that of epithelial cells and might depend on the microenvironment of the cells. Thus these extensive biochemical studies provide evidence that OmpA interacts with chitobiose and that this interaction could be important in the initial steps of the *E.coli* pathogenesis.

Experiments have also shown that two short synthetic peptides (the hexamer, Asn27-Gly32 and the pentamer, Gly65-Asn69) generated from the N-terminal amino acid sequence of OmpA exhibit significant inhibition of OmpA-contributed *E. coli* invasion of BMEC (5). These results indicate that the interaction of amino acid residues in loops 1 and 2 of the OmpA interact with GlcNAcβ1-4GlcNAc epitopes for *E. coli* invasion of BMEC.

On the other hand, OmpA mediated *E. coli* invasion was not observed with systemic endothelial cells such as HUVEC (5). Further studies on systemic endothelial cells revealed that the majority of GlcNAcβ1-4GlcNAc epitopes are substituted with L-fucose, thus possibly blocking the interaction with OmpA for *E. coli* invasion.

These findings strongly implicate OmpA as the necessary microbial structure for the neurotropic nature of *E. coli* to invade BMEC, an important event in the pathogenesis of *E. coli* meningitis. Blocking of OmpA binding site by small molecules is a potential

strategy to prevent the penetration of *E. coli* into central nervous system. Thus, to lay the groundwork for rational drug design, we initiated studies of the molecular level interactions of OmpA with chitobiose and other disaccharides already tested experimentally.

OmpA is a highly conserved outer membrane protein of *E. coli* with a molecular weight of ~35 kDa (325 amino acids) containing eight transmembrane domains and four extracellular loops. The crystal structure of the N-terminal 1-171 residues in the transmembrane domain of OmpA has been solved (11) to a resolution of 2.5Å and further refined to 1.65Å (12). The crystal structure consists of a regular eight-stranded β-barrel with large water-filled cavities, but does not form a pore. The barrel is exceptionally long with an average length of 13 residues. The barrel interior is polar with salt bridge networks that form a barrier for passage of water or ions. Nevertheless, it contains water filled cavities that could serve as interaction sites with other proteins. The four extracellular loops are mobile and not well defined in the crystal. The high-resolution structure of OmpA (12) is better resolved in the transmembrane regions, but there are several missing residues in other regions. Hence our study uses the 2.5Å structure (pdb code: 1bxw) since this was complete (11). The hexamer, Asn-27-Gly-32 and pentamer, Gly65-Asn69, tested for inhibition of *E. coli* invasion are present in the extracellular loops L1 and L2.

We use the HierDock first principles simulation procedure (13) to predict the binding site of chitobiose in OmpA. The HierDock procedure finds two important interaction sites on OmpA favorable to chitobiose. Substitution of fucosyl group on

chitobiose makes the epitope very unfavorable for binding to OmpA, providing an explanation for why *E.coli* does not invade the systemic endothelial cells.

Methods and Validation

To obtain an accurate description of the binding of small molecules to a protein requires an accurate description of the forces between them plus an accurate description of the changes in solvation that accompany binding. Also required is a comprehensive conformation search over the potential binding sites while considering all plausible conformation changes in the ligand and the protein. The best computational methods are believed to provide reasonably accurate predictions of relative binding energies (14-16); however, for cases where the binding site is not known it is not practical to use these most accurate methods at every possible site. Thus we have developed the HierDock hierarchical strategy that starts with a coarse grain search (fast but not too accurate) over the full protein to identify the best sites for finer grain studies, which are followed by a succession of increasingly accurate but increasingly costly studies that ultimately include an accurate description of solvation and fully flexible protein and ligand. HierDock has previously been applied successfully to predict the binding of non-natural amino acids to phenylalanyl t-RNA synthetase (17) and to predict the affinity of odorants to the mammalian olfactory receptor (13), a membrane-bound protein.

HierDock uses a coarse grain Monte Carlo procedure [currently, DOCK4.0 (18)] to select an ensemble of conformations over which to do a hierarchy of more accurate (fine grain) Molecular Dynamics (MD) annealing (19) to optimize the ligand in the

various potential binding sites. The MD calculations use an all-atom forcefield (FF) with continuum solvation calculations of the energies and forces arising from solvation for both the ligand and the protein receptor (20,21). The Poisson-Boltzmann (PB) description (20) leads to accurate energies and forces to describe the solvent effects on the energies and structure of small molecules. However, PB is too slow for MD. Consequently, we use the Surface Generalized Born (SGB) method (21), which leads to a reasonably accurate description of the solvent effects at considerably less cost. We have found that the SGB method leads to results as accurate as the PB method in describing the electrostatic response of the solvent (22). These solvation methods have been used in our calculations of scoring functions.

FF Validation Studies: Critical elements of the HierDock protocol are the hierarchical sequence of conformational searching, the accuracy of the FF, and the inclusion of solvation in the calculating structures and binding energy.

We use the DREIDING FF (23) with CHARMM (24) charges for the protein and charge equilibration (25) charges for the carbohydrate/sugar ligands. This approach has been used previously to study the binding of chitin (a polysaccharide) to family 18 and 19 chitinases (26,27). Simulated annealing dynamics were reported (26) for hexaNAG substrate binding to family 18 chitinase. The RMS in coordinates for all atoms in the binding site was 2.13Å, which is within the crystal resolution.

In another study (28) the structure of sugars bound to proteins were predicted and compared to co-crystal structures available for these complexes. Calculations for the binding of L-arabinose and D-fucose to the L-arabinose binding protein (pdb code: 1abe) were performed. Starting with the crystal structure (1abe) the ligands were removed from

the binding pocket and using the HierDock protocol (DREIDING FF) and we docked L-arabinose and D-fucose to 1abe structure. The predicted structures of the protein/ligand complexes are in good agreement with the crystal structures (coordinate RMS error of 0.4Å compared to the crystal structure of L-arabinose in L-arabinose binding protein, 1abe). The coordinate RMS error for D-fucose binding to L-arabinose binding protein is 1.4Å compared to the corresponding crystal structure (1abf). Using a single target protein as a starting point we predicted the co-crystal structure of both L-arabinose and D-fucose binding to L-arabinose binding protein. This is in fairly good agreement with co-crystal structures, of resolution 1.7 and 1.9Å for 1abe and 1abf respectively. Other groups have also used the DREIDING FF for molecular dynamics for glycoproteins (29).

Summarizing, a number of studies validate that our FF and charges should lead to reasonable binding sites and energies for sugar protein complexes such as in the current study. Indeed we report here that the calculated binding energies for various sugars to OmpA correlate well with experimental invasion assays.

Application of HierDock protocol to sugars binding to OmpA.

A. Coarse grain docking ensemble

Starting with the crystal structure of OmpA, we removed the waters so that the volume of the receptor site will be explored more completely. The HierDock procedure was as follows.

1. *Sampling volume.* The docking site was not known for OmpA and hence the negative image of the entire receptor's molecular surface was mapped as shown in Fig. II-1, and filled with a set of overlapping spheres. A probe of 1.4 Å radius was used to

generate a molecular surface with 5 dots/Å. Sphere clusters were generated for the whole binding site using the program Sphgen.

2. *Defining regions for docking*: The sphere filled volume from step1 representing the possible binding sites, was partitioned into 12 regions. These regions included the 4 loops, the space in between the loops, and the space inside the barrel till half way within the transmembrane region. Since the barrel is very narrow and the ligands are reasonable big, it seemed improbable that the ligand would go deep into the barrel without any steric clashes. Moreover, the sugars are attached to Ecgp, the glycoprotein on the BMEC surface that will further prevent it from going far into the barrel cavity. Also, the barrel has internal polar networks that form a prominent barrier in the barrel interior with ordered water molecules inside the β barrel.

3. *Ensemble selection (scoring):* To generate an ensemble of docked receptor-ligand complexes, we used the program DOCK (version 4.0) to sample various orientations and conformations of the ligands in the receptor site. We used flexible ligand docking option in DOCK4.0, with torsion minimization of ligands. A non-distance dependent dielectric constant of 1 and a distance cutoff of 10 Å were applied for calculating protein-ligand coulombic interaction energy. The energy score in DOCK4.0 uses coulombic energy and van der Waals energy as described in reference 18. The conformations were ranked using energy scoring in DOCK4.0. The top 10-30 conformations were selected by energy score for each ligand in each of the 12 potential binding regions. These selected conformations were further used for fine grain molecular dynamics calculations.

B. Fine grain MD search

The next step in HierDock is to perform a fine grain MD with an all atom forcefield to optimize the ligand conformation inside the binding pocket.

1. *Annealing Molecular Dynamics*: For each of the 12 regions, the 20 best scoring docked conformations from DOCK were subjected to five cycles of annealing dynamics (each from 50K to 600K and back to 50K with 25 picoseconds of MD at each temperature). This allows for the optimization of the ligand conformation in the protein. One lowest energy conformation is stored from each cycle of annealing MD. The energies of the best-annealed structures were calculated using all atom DRIEDING forcefield (23) and SGB continuum solvation model. We find that this step of ligand optimization is critical in getting energetically favorable conformations for the complex (Protein plus ligand).

2. *Binding Energy Calculation*: The binding energies of the 20x5 annealed structures for each ligand in each of the 12 docking regions were calculated. Binding energies were calculated using DREIDING forcefield, and charge equilibration (24) charges using MPSim simulation code (19). Solvation effects for ligand binding were calculated using the SGB description of the continuum solvent model. The binding energy is calculated as the difference between the total energies of the complex (protein +ligand) and the sum of the protein and ligand energies. The best conformation from this pool was selected by the binding energy for each of the 12 docking regions.

Summarizing, the HierDock protocol uses a hierarchical strategy for conformation search and a corresponding scoring function to select a subset of structures for the next level. The coarse grain level (DOCK4.0) uses a crude scoring function including just

coulombic and van der Waals interactions of the ligand with the protein. Using this scoring function a subset of conformations generated by DOCK4.0 is selected for the annealing step. At the fine grain level (annealing MD level) the scoring function uses an all tom forcefield and continuum solvation method to calculate the binding energies and select the best-bound structure of the ligand in the protein.

C. Selection of the best two regions through application to Chitobiose

A complete scanning of all possible docking regions for OmpA was done with chitobiose. The structure of the ligands used is shown in Fig. 2.2. The final 1200 structures in 12 regions (100 conformations from each region) were scored using SGB solvation and DREIDING forcefield. Comparison of the binding energies in various regions showed that regions 1 and 2 shown in Figure II-3 have the best binding energies. Hence, these two regions were ranked are possible binding sites. The binding energy of conformations of chitobiose in regions 1 and 2 were about –15 kcal/mol and –20 kcal/mol, the difference being 5 kcal/mol, was small enough to treat these two regions on an equal footing. Hence these two regions were kept as possible binding regions for chitobiose. The binding energy of chitobiose in other regions was less than –5 kcal/mol and hence considered insignificant compared to regions 1 and 2.

D. Application to all four ligands

Having identified the probable binding sites for chitobiose, we then carried out the complete sequence of HierDock calculations over regions 1 and 2 for the three ligands using structure of the ligands shown in Figure II-2.

Fucosylated chitobiose

Cellobiose

Lactose

Thus, these ligands were docked and the structures were further annealed using MD and all the resulting complexes (OmpA+ligand) were optimized (energy minimized) using conjugate gradient method. Finally, the binding energies of these ligands were calculated and ranked in these two binding sites.

E. Optimization with the flexible binding site on OmpA

Protein flexibility is critical to determining the binding conformation and the critical ligand protein interactions energies. Hence, we performed five cycles of 50ps annealing MD heating from 50K to 600K and cooling from 600K to 50K for the best conformation of chitobiose bound to region 2. These simulations were performed allowing the ligand and all the amino acid residues that are in the top half of the barrel in Figure II-3 to be flexible but keeping the rest of the protein fixed. These simulations optimized the hydrogen bonds and van der Waals contacts made by the ligand to particular residues in the binding cavity. The final conformation from these simulations was used for all analysis.

Results and Discussion

Using the procedure described above we predicted the binding site and binding energies of sugars in OmpA. Regions 1 and 2 shown in Figure II-3 were ranked as the most favorable regions for binding of the best ligand (chitobiose) in OmpA. The binding energy of the best conformation of chitobiose is –15.42 kcal/mol in regions 1 and–20.50 kcal/mol in region 2. In contrast the calculated binding energy of chitobiose in all other

regions was worse than -5.0 kcal/mol. Hence we consider region 1 and region 2 as likely binding sites for chitobiose in OmpA. Indeed peptides from region 1 [between external loop structures loop1 (L1) and loop2 (L2)] have been tested experimentally for inhibition of *E.coli* invasion.

The binding energies of chitobiose, fucosylated chitobiose, lactose, and cellobiose in regions 1 and 2 are shown in Figure II-4. The result is that chitobiose has the best binding energy in both regions 1 and 2. This indicates that binding of chitobiose in these two regions of OmpA is specific. Experimentally, chitobiose dramatically reduced invasion (>95%) when BMEC are infected with *E. coli* K1 after pretreating with ligands (5), while none of the other ligands show reduced invasion. Thus we find a good correlation between the binding energies of the ligands to OmpA and the experimental invasion assay measurements. This suggests that binding to OmpA may be a necessary step in the pathogenesis of bacterial meningitis.

Figure II-5 shows the contribution of the valence and non-bond energies to the binding energy. Interestingly, lactose and cellobiose (both similar in size to chitobiose) do *not* show good binding energies. We attribute the lack of good binding energy for lactose compared to chitobiose to the absence of $NHCOCH_3$ group. The hydrogen bonds made by the $NHCOCH_3$ group in chitobiose to His152, His20, and Arg157 are absent in lactose and cellobiose. On the other hand fucosyl substituted chitobiose in region 1 has non-bond interactions similar to chitobiose but the binding is weakened by the steric clash with the bulky fucosyl substitution. In region 2, fucosyl chitobiose does not have favorable non-bond interactions since it loses hydrogen bonds to His152 and Arg 157.

Fucosyl chitobiose is a bulky ligand that does not bind to OmpA due to steric hindrance from the protein. We observed that during step B.1 of the HierDock protocol (i.e., the ligand annealing MD for fucosyl chitobiose with OmpA fixed in region 2) the sugar ring was forced into a boat conformation due to steric clashes with the fixed protein. We found that the ligand gets trapped in this conformation during the cooling phase of annealing MD, resulting in a strained state for fucosyl chitobiose. Such a transformation to the boat form does not occur if annealing MD is performed with *all* protein atoms are allowed to move along with the ligand. However, even with full relaxation (to the chair form) the binding energy of fucosyl chitobiose is still worse than chitobiose by 14.2 kcal/mol. We consider that annealing MD with all atoms movable is too expensive computationally to do for all docked structures. Thus we refined the HierDock procedure to identify the large strain energies (caused in this case by the chair to boat transformation) in the internal energy of the ligand during protein fixed annealing MD. Then for structures leading to high internal strain energy, the HierDock procedure now carries out annealing MD with all protein atoms movable when calculating the binding energy.

Chitobiose makes a number of critical contacts with residues on the loops in region 1 and region 2. In region 1, it makes significant contacts with residues Val 68, Ser 67, and Glu 69 on loop L2 and with Asn 27, Asn 28, and Ile 25 on loop L1 (Table 1). The NHCOCH3 group on chitobiose makes a critical hydrogen bond with Asn 27 as shown in Figure II-6. There is experimental support of a role for residues in region 1. A number of peptides have been tested experimentally for inhibition of *E. coli* invasion, but

only peptides from L1 and L2 were able to block invasion when compared against random peptides. This correlates very well with our predicted site in region 1.

Region 2 is in the water filled pocket between the extracellular loops. Chitobiose makes a number of contacts in this region including a number of van der Waals contacts with loops L1, L3, and L4, and with some residues in the β-barrel. The residues that are within 5Å of chitobiose are listed in Table 1. Chitobiose is partially buried in the cavity between the loops (Table 1) and present in the boundary between the β-barrel and the loops. Chitobiose makes electrostatic contacts with Trp 103, Arg 157 and His 152 (shown in Figure II-7). On annealing MD with the protein cavity movable, the loops come closer to complex more strongly with the ligand. Loops 1 and 4 show maximum displacements from the original crystal structure (Figure II-8). Annealing dynamics suggest that chitobiose stays at the mouth of the cavity rather than going far into the cavity. OmpA has a very narrow pore, which hinders chitobiose from inserting further into the cavity. We suggest experimental studies on invasion assays using the peptides from this list of residues in Table II-1.

In this study we correlated the binding energies to the percentage invasion results from invasion assay experiments. It should be noted that many other steps might be involved in the pathogenesis leading to invasion of BMEC by *E.coli*. The correlation between the calculated binding energies and invasion assay results suggests that the binding of OmpA is a necessary step in the pathogenesis. Of course there may be many other important factors. To further validate this model, we suggest several experimental tests of the predictions made by the model:.

1. Synthesize peptides from binding region 2 on OmpA and test the efficacy of these peptides for blocking of *E. coli* invasion. For example we recommend testing the following peptides for blocking invasion: QYHDTGLIH, QYHDT, HDTGL, TGLIH, DTGLI, GMVWRADTWS, GMVWR, ADTWS, VWRAD, KNHDT, NHDT, TNNIGDAHTIGTRPDNG, TNNIG, DAHTI, GTRPDNG, NIGDAHTIGTRPD, TRPDNG, NNIGDAHT, AHTIGTRPDN

2. Carry out point mutation studies for OmpA, targeting the polar and charged residues within 3Å of chitobiose in region 2. Our predicted binding site suggests that mutations of residues Asp117, Asn147, His152, Arg157, and Asp159 will lead to disruption of OmpA interaction with chitobiose.

3. Design small molecule drugs that would bind more strongly than chitobiose to regions 1 and 2 to inhibit the OmpA invasion of BMEC. We are currently using the data from our predicted sites to search for new compounds for experimental tests.

The glycoprotein (Ecgp) that interacts with OmpA has been sequenced and its two glycosylation sites have been recently identified by Prasadarao *et al.* (unpublished results). Based on these modeling studies we propose two possible mechanisms of OmpA binding to the sugars in Ecgp.

- One possibility is a two-step mechanism: Here the chitobiose is first recognized by the loop1 in region 1 and then transferred to region 2 for stronger binding. In this case it seems plausible that region 1 is a recognition region and region 2 is the binding region. This recognition mode involving a sequence of two interaction regions

has been shown for such other membrane proteins as CCR5 (30), a cofactor of CD4 in HIV invasion.

- A second plausible mechanism is that two sugar moieties from two glycosylated sites from Ecgp interact with both sites of OmpA at the same time. This might give stronger binding constants and additional selectivity.

Conclusions

We find two regions on OmpA that are potential interaction sites for GlcNAc1 and 4GlcNAc epitopes on glycoproteins of BMEC. For both the regions, we showed that OmpA of *E. coli*, binds most favorably to chitobiose as compared to cellobiose, fucosyl substituted chitobiose and lactose. The difference in the binding energies of OmpA with chitobiose and fucosylated chitobiose is of significance in understanding the pathogenesis of *E. coli* meningitis. Since most of the systemic cells of the body have fucosylated glycoproteins, this explains the specificity of *E. coli* invasion of BMEC. A good binding between OmpA and chitobiose arises from specific interactions of the NHCOCH3 groups on chitobiose with the OmpA residues. Lactose and cellobiose, although they have the same β1-4 linkage as chitobiose, they lack the NHCOCH3 group that is important in making favorable specific interactions. Fucosylated chitobiose, on the other hand, makes good contacts with the receptor and has a high non-bonded energy but it has an unfavorable valence energy.

The binding site predictions made are testable either by point mutation studies or by invasion assays. It is important to note that we have studied the interaction of only

sugars with OmpA, because the experimental invasion assays were tested with the same sugars.

Acknowledgements

This work was supported by NIH/NICHD (WAG) and NIH R29AI40567 (NVP). The facilities of the Materials and Process Simulation Center used in this project are supported also by DOE (ASCI ASAP), NSF (CHE and MRI), NIH, ARO-MURI, Chevron Corp., MMM, Beckman Institute, Seiko-Epson, Avery-Dennison Corp., Asahi Chemical, General Motors, and Kellogg's.

References

1. Gladstone, I.M., Ehrenkranz, R.A., Edbergt, S.C., and Baltimore, R.S., (1990), *Pedriatr Infect. Dis. J.*, **9**, pp 819-825.

2. Klein, J.O., Feigin, R.D., McCracken Jr., G.H., (1986), *Pediatrics*, 78 (suppl), pp 959-982.

3. Unhanand M., Mustafa M.M.,. McCracken G.H and Nelson J.D. (1993), *J.Pediatr.*, 122, pp 15-21.

4. Prasadarao N. V., Wass, C.A., Weiser J.N.,. Stins, Huang M.F, S. H., and. Kim K.S (1996) *Infect. Immun*., 64, pp 146-153.

5. Prasadarao N. V., Wass C. A, and . Kim K.S., (1996), *Infect. Immun* 64, pp 154-160.

6. Prasadarao N.V., Wass, C.A and Kim, K.S.,. (1997), *Infect. Immun*., 65, pp 2852-2860.

7. Prasadarao, N.V., Wass, C.A. Stins, M.F., Shimada, H., and Kim, K.S., (1999), *Infect. Immun* 67, pp. 5775-5783.

8. Isberg, R.R., Science, (1991), 252:934-938.

9. Kindberg G.M, Magnusson S, Berg T, Smedsrod B, (1990), *Biochem.J*., 270:197-203.

10. Lecuit M, Dramsi S, Gottardi C, Fedor-Chaiken M, Gumbiner B, and Cossart P. (1999), *EMBO J*. 18 3956-3963.

11. Pautsch, A. and Schulz, G.E. (1998), *Nat. Struct. Biol*. **5**, 1013-1017.

12. Pautsch, A. and Schulz, G.E. (2000), *J. Mol. Biol.* 298, 273-282.

13. Floriano W.B, Vaidehi N., Singer M.S., Shepherd G. M., Goddard III, W.A., (2000)., *Proc. Natl. Acad. Sci., USA*, **97**, 10712 -10716.

14. Orozco, R., and Luque F.J., (2000). *Chem Rev*, 100, 4187-4225

15. Kuntz I.D., Blaney J.M., Oatley S.J., Langridge R., Ferrin T.E. (1982). *J. Mol. Biol.* 161, 269-288.

16. Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J. (1998), *J. Comput. Chem*., 19: 1639-1662.

17. Wang, P., Vaidehi, N. Tirrell, D.A., and Goddard III, W.A., *J. Am Chem. Soc*. Submitted.

18. Ewing, T.A. & Kuntz, I.D. (1997). *J Comput. Chem.* 18, 1175-1189.

19. Lim K.T., Brunett S., Iotov M., B. McClurg, N. Vaidehi, S. Dasgupta, S. Taylor, and W.A. Goddard III, (1997), *J. Comp. Chem.*, 18, 501-521.

20. Tannor, D.J., Marten, B., Murphy, R., Friesner, R.A., Sitkoff, D., Nicholls, A., Ringnalda, M.N., Goddard III, W.A. and Honig, B., (1994), *J. Am. Chem. Soc*., 116, 11875-11882.

21. Ghosh , A., Rapp, C.S. & Friesner, R.A. (1998). *J. Phys. Chem. B* 102, 10983-10990.

22. Tang, Y., Ghiralando, G., Vaidehi, N., Kua, J., Mainz, D.A., Degrado, W.A. Goddard III, W.A., and. Tirrell, D.A., *Biochemistry*, (2001), 40, 2790-2796.

23. Mayo, S. L., Olafson, B.D. & Goddard III, W.A. (1990). *J. Phys. Chem.* 94, 8897-8909.

24. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T., Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe, M., Wiorkiewicz-Kuczera J, Yin D, Karplus M, (1998), *J. Phys. Chem.*B , **102:** 3586-3616 .

25. Rappé, A.K. and Goddard III, W.A. (1991). *J. Phys. Chem.* 95, 3358.

26. Brameld, K.A., and Goddard III, W.A., *J. Am. Chem. Soc.* (1998), 120, 3571-3580.

27. Brameld, K.A., and Goddard III, W.A., *Proc. Natl. Acad. Sci. USA*, (1998), 95, 4276-4281.

28. Floriano, W.B., Vaidehi, N. and Goddard III, W.A. (in preparation).

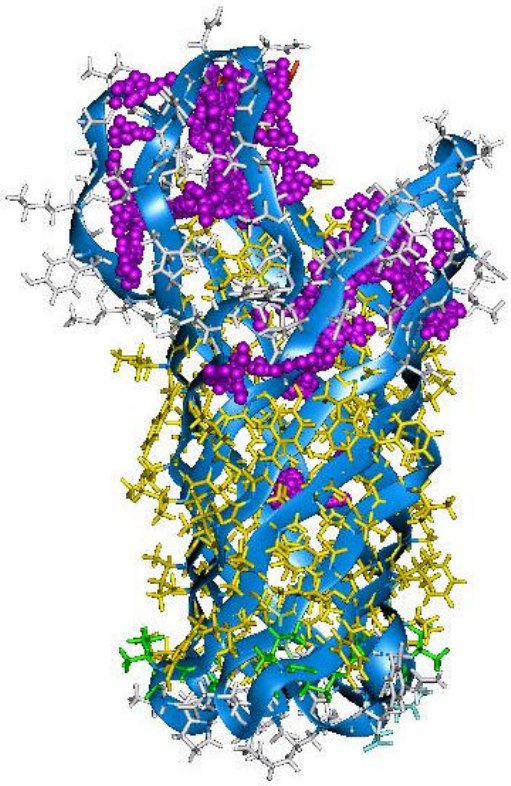
29. Gabriel J.L., and Mitchell W.M., *Proc. Natl. Acad. Sci. USA*, (1993), 90, 4186-4190.

30. Yang, J., Zhang, Y.W., Huang, J.F., Zhang, Y.P., and Liu, C.Q., Theo *Chem*, (2000), 505 199-210.

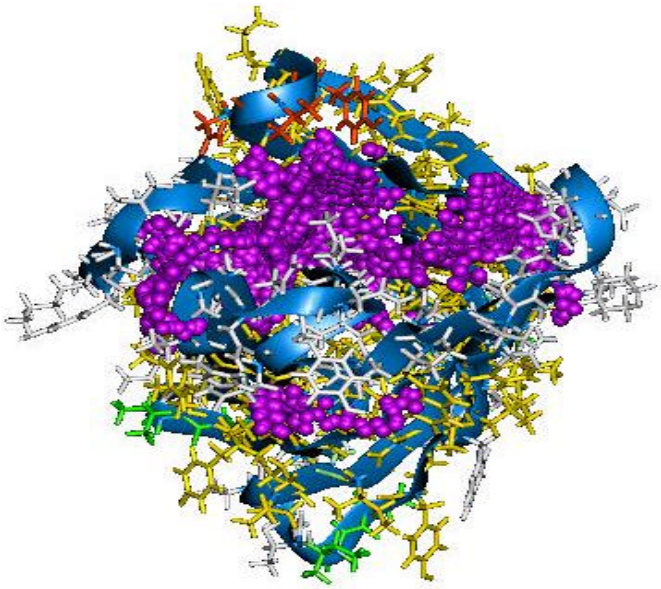
Table 1 : List of residues that are in the binding regions 1 and 2 of chitobiose. Residues listed here are within 5 Å of chitobiose in regions 1 and 2. Residues in bold are within 3 Å of chitobiose and make critical contacts with chitobiose.

	Loop1	**Loop2**	**Loop3**	**Loop4**	**Barrel**
Region1	Ile25 Asn26	Ser67 Val68			
	Asn27 Asn28	Glu69			
	Gly29 Pro30				
Region2	Gln18 Tyr19		Gly100	Thr145 Asn146	Asp106 Thr107
	His29 **Asp21**		Met101 Val102	**Asn147** Ile148	Tyr108 Ser109
	Thr22 Gly23		Trp103 Arg104	Gly149 Asp150	Lys114
	Leu24 Ile25		**Ala105** Asn115	Ala151 **His152**	
	His 20		His116 **Asp117**	Thr153 Ile154	
			Thr118	Gly155 Thr156	
				Arg157 Pro158	
				Asp159 Asn160	
				Gly161	

Figure II-1 Sphere filled volume representing the possible binding sites on OmpA. (a) shows the side view of the protein and (b) represents the top view.

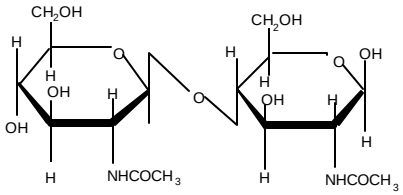


(a)

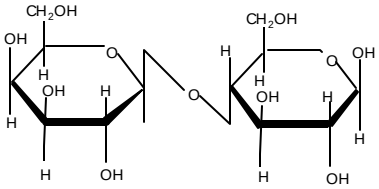


(b)

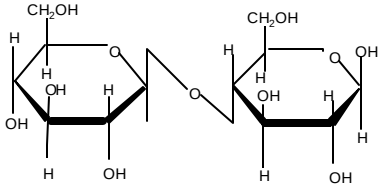
Figure II-2 Sugars used for Docking in the twelve regions of OmpA.



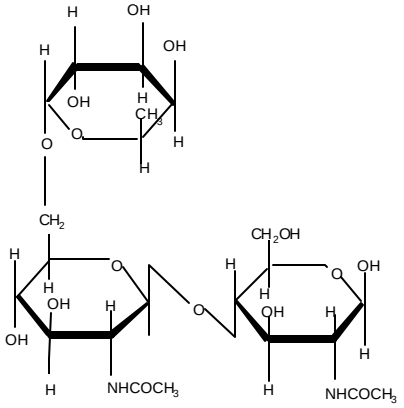
Chitobiose



Lactose



Cellobiose



Fucosylated chitobiose

Figure II-3 OmpA region 1 and region 2. Region 1 (weaker binding) is in between loops L1 and L2. Region 2 (stronger binding) is in the water filled cavity between the 4 loops.

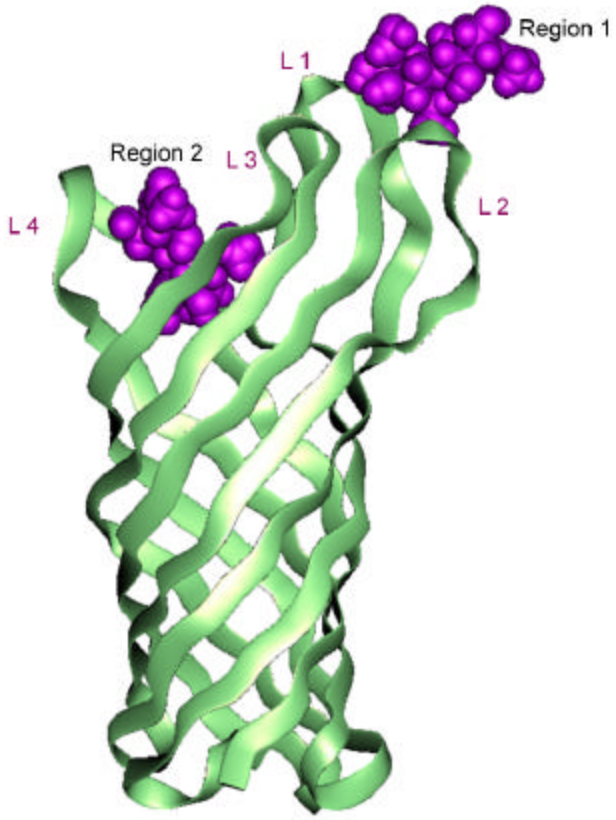

Region 1
L 1
Region 2
L 3
L 2
L 4

Figure II-4 Binding energies of the four ligands in regions 1 and 2. Chitobiose has the best binding energy in both regions.

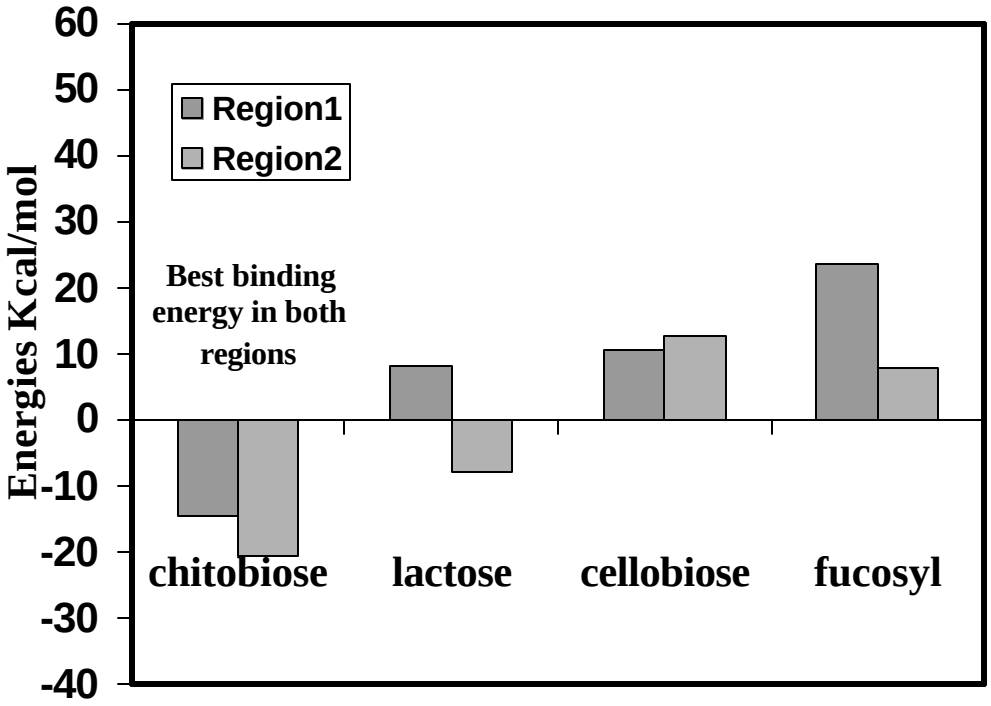
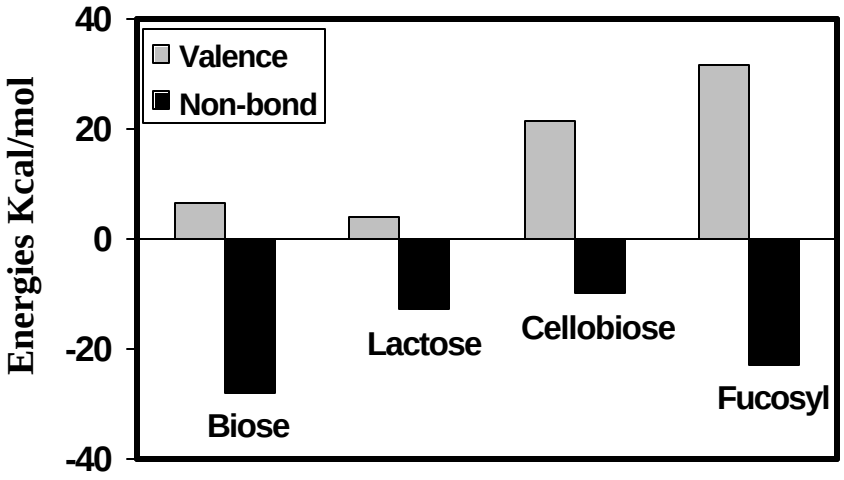
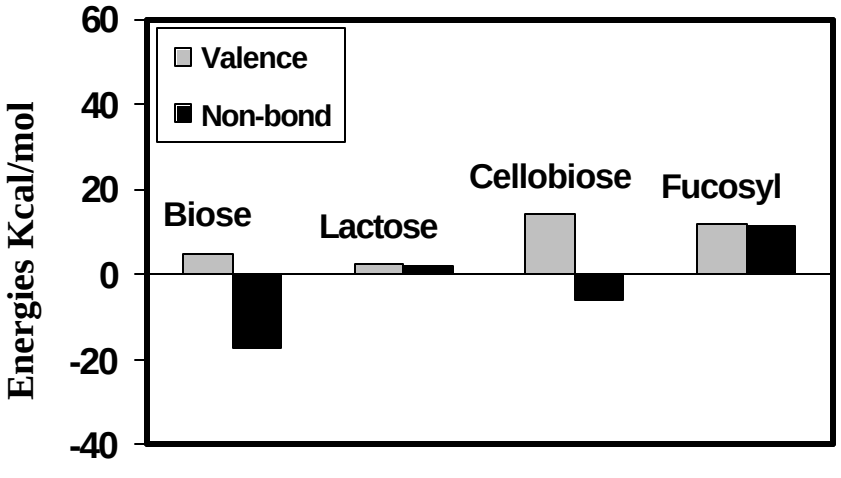

60
50
40
30
20
10
0
-10
-20
-30
-40
Energies Kcal/mol
Region1
Region2
Best binding energy in both regions
chitobiose
lactose
cellobiose
fucosyl

Figure II-5 Valence and non-bond energy contributions to the binding energy of lactose, chitobiose, cellobiose and fucosylated chitobiose in regions 1 and 2. **(a)** Valence energy components in region1. **(b)** Non-bond energy components in region1. **(c)** Valence energy components in region 2. **(d)** Non-bond energy components in region.


40
20
0
-20
-40
Energies Kcal/mol
Valence
Non-bond
Biose
Lactose
Cellobiose
Fucosyl

(a)


60
40
20
0
-20
-40
Energies Kcal/mol
Valence
Non-bond
Biose
Lactose
Cellobiose
Fucosyl

(b)

Figure II-6. Important residues within 5 Å of chitobiose in region 1 (between loops, L1 and L2).

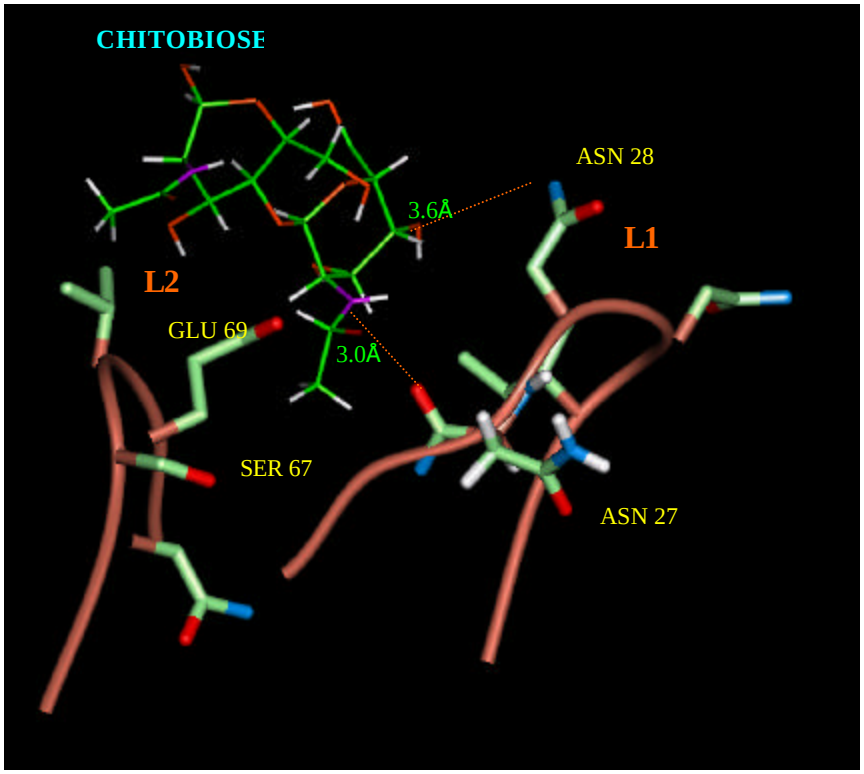

CHITOBIOSE
ASN 28
3.6Å
L1
L2
GLU 69
3.0Å
SER 67
ASN 27

Figure II-7. Residues within 3Å of chitobiose in region 2. Residues on loop4 (which moves significantly) make important contacts with chitobiose. Almost all residues within 3Å of chitobiose are either polar or charged

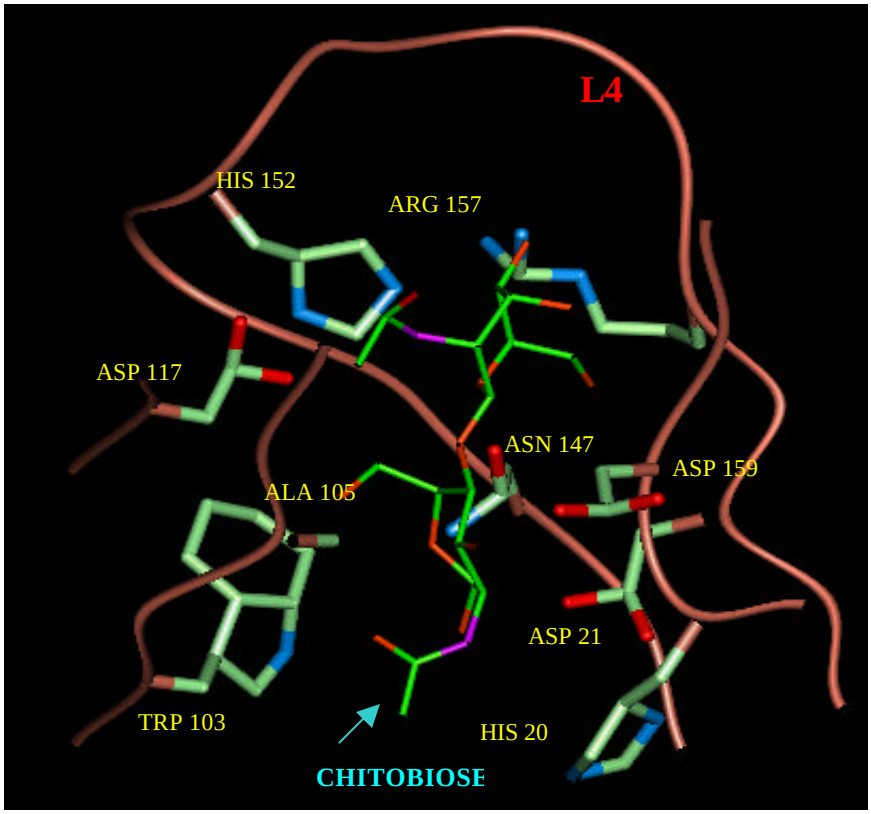
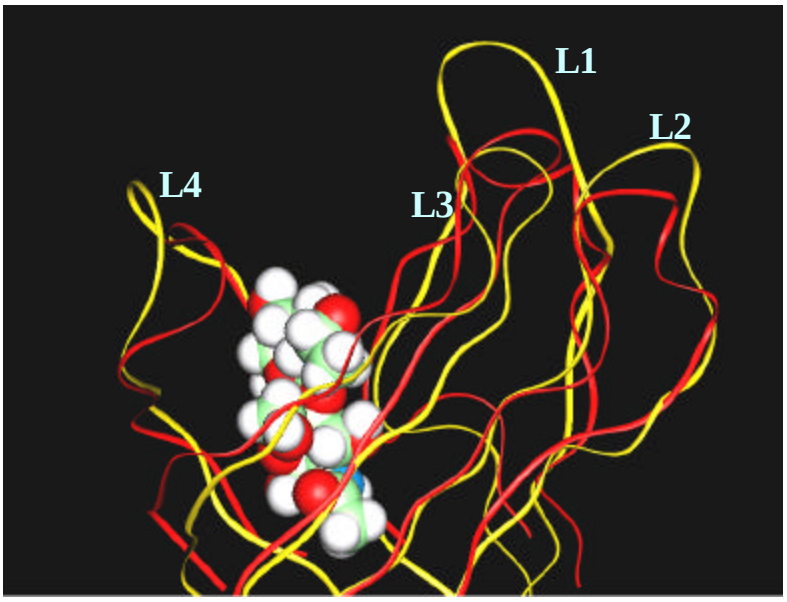

L4
HIS 152
ARG 157
ASP 117
ASN 147
ASP 159
ALA 105
ASP 21
TRP 103
HIS 20
CHITOBIOSE

Figure II-8 Four loops move significantly on binding to chitobiose. Loop4 and Loop1 show maximum change in structure.


L1
L2
L3
L4

Chapter 3

Mechanism for Antibody Catalysis of the Oxidation of Water by Singlet Dioxygen

Adapted from Datta. D., Vaidehi N., Xu X., and Goddard III W. A., Proc Natl Acad Sci U S A. 2002 ; 99: 2636-41.

Abstract

Wentworth and coworkers (Wentworth *et al.* (2001*), Science*, 293, 1806-1811**)** recently reported the surprising result that antibodies and T-cell receptors efficiently catalyze the conversion of molecular singlet oxygen ($^1O_2$) plus water to hydrogen peroxide (HOOH). Recently quantum mechanical (QM) calculations were used to delineate a plausible mechanism, involving reaction of $^1O_2$ with two waters to form HOOOH (plus H_2O), followed by formation of HOOOH dimer, which rearranges to form HOO-HOOO + H_2O, which rearranges to form two HOOH plus $^1O_2$ or $^3O_2$. For a system with ^{18}O H_2O, this mechanism leads to a 2.2:1 ratio of ^{16}O:^{18}O in the product HOOH, in good agreement with the ratio 2.2:1 observed in isotope experiments by Wentworth *et al.*

In this paper we use docking and molecular dynamics techniques (HierDock) to search various protein structures for sites that stabilize these products and intermediates predicted from QM calculations. We find that the reaction intermediates for production of HOOH from $^1O_2$ are stabilized at the interface of light and heavy chains of antibodies and T-cell receptors. This Inter Greek Key Domain Interface (IGKD) structure is unique to antibodies and T-cell receptors.

IGKD interface is *not* present in β_2-microglobulin, which does *not* show any stabilization in our docking studies. This is consistent with the experimentally observed lack of HOOH production in this system. Our results provide a plausible mechanism for the reactions and provide an explanation of the specific structural character of antibodies responsible for this unexpected chemistry.

1. Introduction

Recently, Wentworth *et al.* [1] reported surprising results that antibodies can convert molecular oxygen to hydrogen peroxide and they also showed that the antibodies *catalyze* the oxidation of H_2O to H_2O_2 by singlet oxygen molecules, $^1O_2$ [2]. This suggests that in addition to the well-known antigen recognition function of antibodies, they may also promote destruction of the molecules to which they bind. This could have implications in the function (and malfunction) of the immune system and in the evolution of this system.

Investigations of the long-term photo-production of H_2O_2 by antibodies and non-immunoglobulin proteins reveal a remarkable difference [2]. It was demonstrated that the sustained high concentrations of H_2O_2 produced recursively could not have been by the oxidation of the amino acids in the antibodies. Thus, production of H_2O_2 by antibodies remains linear for a much longer period than for all non-immunoglobulin proteins tested (up to > 50 mol equivalents of H_2O_2). Furthermore, if the H_2O_2 generated during the assay is removed, antibodies are able to resume H_2O_2 production at the same initial rate as at the start of the experiment while other proteins that produce H_2O_2 do so by the photo-oxidation of the amino acids (e.g. tyrosine, tryptophan) and are not able to resume the same initial rate of H_2O_2 production. These experiments strongly suggest that the antibodies play a catalytic role in converting $^1O_2$ plus water to H_2O_2.

Through isotopic labeling experiments Wentworth *et al.* [1] concluded that water was oxidized by the $^1O_2$ generated. However, the experiments have not provided a mechanism to understand *how* the antibodies and T-cell receptors (TCR) carry out this remarkable and unexpected chemistry. They observed that only antibodies and TCR

catalyze this reaction, which implies that these molecules probably have unique structural features not present in other proteins. One unique feature of these systems is the interfaces created by the Greek key motifs. However, β-microglobulin also has a Greek key motif but does not convert $^1O_2$ to H_2O_2.

The goal of this paper is to determine which sites in the antibodies (and TCR) play a role in the process by which $^1O_2$ interacts with H_2O to produce H_2O_2. A companion paper [3] presents quantum mechanical (QM) calculations that delineate plausible chemical reaction mechanisms for this chemistry, which are summarized in section 3. Briefly, this mechanism involves formation of HOOOH from the reaction of $^1O_2$ with H_2O dimer, followed by complexing with another HOOOH to form a dimer that rearranges to form two HOOH plus O_2. For a system with ^{18}O H_2O, this mechanism leads to a 2.2:1 ratio of ^{16}O:^{18}O in the product HOOH, in good agreement with the ratio 2.2:1 observed in isotope experiments by Wentworth *et al [1]*.

In this paper, we use docking and molecular dynamics (MD) techniques to search various protein structures for sites that stabilize these products and intermediates predicted from QM calculations. We used the HierDock docking and MD protocol [4] to find antibody sites that might stabilize the reaction intermediates. These HierDock studies considered high-resolution (<2.0Å) crystal structures known to catalyze this chemistry (Several Fab fragments of antibodies with varying sequence homology and TCR) and other structures (β2-microglobulin) known not to. We find that all antibodies and TCR have unique sites that stabilize the QM intermediates and products, whereas no such sites are found for the β2-microglobulin or other proteins. The deduced catalytic sites are at the interface of light and heavy chains of the antibody and TCR.

These results suggest a specific structural characteristic of antibodies that is responsible for this unexpected chemistry. Armed with such specific predictions it should be possible to design experimental tests that would help verify or discard some of the plausible mechanisms. The predictions about specific important sites in the antibody could be used to design mutation studies in the antibodies and TCR to provide detailed tests on the role of the antibody. Since the proposed mechanism does not require an energy or electron source (other than $^1O_2$) one might be able to use these insights to design nanoscale biomimetics to carry out this remarkable chemistry in very different environments.

Section 2 presents the methods used in the HierDock protocol, section 3 summarizes the QM results, section 4 describes the sites in antibody found to stabilize the catalytic intermediates, and section 5 discusses the results.

2.0 METHODS

To identify plausible catalytic sites in the antibodies, we used the HierDock [4] protocol to search the entire antibody structure for sites that would bind to the reaction intermediates in Eqn. 1 using the structures obtained from QM [3]. HierDock uses a hierarchy of coarse grain docking and fine grain MD methods (including continuum solvation forces) to sample possible binding sites for ligands in the protein to determine binding sites and energies. HierDock has been applied successfully to such membrane-bound proteins as the olfactory receptors [4] and outer membrane protein A of *E. coli* [5] and to phenylalanyl t-RNA synthetase [6].

In this paper we first used HierDock to search the entire Fab structure for low energy binding sites. Here we partitioned the entire Fab antibody structure into four

docking regions that could be searched in parallel. First, we carried out a coarse grain search in each region to generate a set of conformations for ligand binding. This procedure used DOCK 4.0 [7] to generate 20,000 configurations, of which 100 were ranked using the DOCK scoring function. Docking the intermediates and products of this reaction was done using rigid ligand option in DOCK4.0

We then selected the 20 best conformations from DOCK in each region and subjected each ligand to annealing MD to further optimize the conformation in the local binding pocket while allowing both the ligand and binding cavity (residues with an atom within 5Å of the binding ligand) to move. In this step, the ligand and the binding cavity in the protein were heated and cooled from 50K to 600K in steps of 10K (0.05ps at each temperature) for 1 cycle. This allows the protein cavity to readjust for the interaction with the ligand. This fine grain optimization was performed using MPSim [8] and a full atom forcefield (FF) (DREIDING)[9].

In addition we used the Surface Generalized Born (SGB) continuum solvent method [10] to obtain forces and energies resulting from the polarization of the solvent by the charges of the ligand and protein. This allows us to calculate the change in the overall binding conformation resulting from differential solvation to obtain accurate binding energies. The charges on the various ligands were obtained from quantum mechanics (Mulliken population densities at the atom centers), while the charges for the protein were from CHARMM22 [11]. A dielectric constant of 80.37 was used for the solvent field in the SGB calculation and 2.0 for the inside of the protein.

From the 20 trajectories of annealing calculations in each docking region, we selected the 20 best conformations. The relative binding energies of the 20 best

structures in each region were compared (DREIDING FF with solvation) to decide which of the four docking regions leads to good binding energies for the ligands. In addition to the binding energies we also examined the population density of good binding structures in each region. The most populated regions of good structures (structures with good binding energies) were chosen for analysis

3.0 Summary of results from QM calculations and plausible mechanisms

The QM studies [3] lead to plausible mechanisms for formation and decomposition of HOOOH and related compounds. The most plausible mechanism involves several steps:

a) Reaction of $^1O_2$ with two waters to form HOOOH plus H_2O (reaction 1 in Figure III-1)

b) Formation of HOOOH dimer

c) Unimolecular rearrangement of HOOOH dimer to form [HOO-HOOO + H_2O] (reaction 2 in Figure III-1)

d) Unimolecular rearrangement of this complex to form HOOH-OOO + H_2O

e) Unimolecular rearrangement of this complex to form HOOH product + HOOOOH (reaction 3 in Figure III-1)

f) Fission of the HOOOOH to 2 HOO and association to form cyclic HOO dimer (singlet or triplet)

g) Rearrangement of cyclic HOO dimer to form HOOH product plus $^1O_2$ or $^3O_2$.

For a system with ^{18}O H_2O, this mechanism leads to a 2.2:1 ratio of ^{16}O:^{18}O in the product HOOH, in good agreement with the ratio 2.2:1 observed in isotope experiments

by Wentworth *et al* [1]. Depending on the products from steps f and g, this QM based mechanism leads to a net reaction of

$$2\ {}^1O_2 + 2\ H_2O \rightarrow 2\ HOOH + {}^1O_2 \qquad (A)$$

$$2\ {}^1O_2 + 2\ H_2O \rightarrow 2\ HOOH + {}^3O_2 \qquad (B)$$

The net reaction in (A) has a molecularity of 2 HOOH formed from each $^1O_2$ is in agreement with the experimental results from Wentworth *et al.* This excellent agreement with the experiments gives some credence to the QM based mechanism. To determine sites in antibodies and TCR that might play a role in enhancing these catalytic processes, we searched for sites in the antibody that bind

The HOOOH product of reaction 1 (step a) (part of P1 and of R2)

(1a)

The HOOOH dimer (R2) of reaction 2 (step c)

(1b)

The HOOH product of steps e and g (part of P3 in reaction 3)

(1c)

Given the clusters of binding sites favorable for these stable intermediates or products, we also examined if they would stabilize the reaction intermediates

TS1: the H_2O-H_2O-$^1O_2$ transition state of reaction 1 (step a)

(2a)

TS2: the transition state for reaction 2 (step c)

(2b)

TS3: the transition state for reaction 3 (step e). (2c)

The gas-phase structures for important intermediates and complexes are summarized in Figure III-1. Some additional comments are:

- Xu et al [3] finds that the barrier for the direct reaction of $^1O_2$ with H_2O to form HOOOH is over 60 kcal/mol, whereas the reaction of $^1O_2$ with H_2O dimer (R1 of Figure III-1) has a barrier (TS1) of ~ 30 kcal/mol.
- Xu et al [3] find two stable structures for the monomer: trans (P1 of Figure III-1) and cis (shown in R2 of Figure III-1). The cis structure is 2.4 kcal/mol higher in energy than the trans structures. We docked both conformations.
- Xu et al [3] find 12 stable but distinct structures for the dimer, $(HOOOH)_2$. The most relevant for the formation of HOOH is R2 in Figure III-1. This structure is 4.9 kcal/mol more stable than the cis-monomer. We docked this dimer conformation.

4.0 The Catalytic Site in Antibodies for Catalytic Transformation of $^1O_2$ and H_2O to HOOH

4.1 Binding sites in the Fab antibody fragment [crystal structure (4c6.pdb)]

To seek plausible reaction sites for various steps in the QM mechanism we used the 1.2Å Fab crystal structure 4c6.pdb[12], which is the highest resolution Fab crystal structure available. The crystal structure was supplemented by adding hydrogens at standard geometries (as given by DREIDING) and hydrated counterions Na^+ and Cl^- were also added to charged side chain residues to maintain neutrality [13].

The crystal structure was optimized using the FF, charges, and continuum solvation methods described in section 2. This minimized structure has a coordinate RMS error of 0.71Å to all atoms of the crystal structure. (The experimental resolution of

the crystal structure is 1.2Å.) This indicates that the FF, charges, and solvation methods are sufficient to describe the system. We used this optimized 4c6 Fab structure in the HierDock protocols to search for sites in the 4c6 Fab structure that strongly bind HOOOH, HOOOH dimer, and H_2O_2 (see Eqn. 1). In addition, we examined the stabilization of the transition states in Eqn. 2 at the predicted binding sites for Eqn. 1.

4.1.1 Binding sites for HOOOH monomer and dimer

We find three sites (denoted I1, I2, I3) that strongly bind HOOOH monomer and dimer. Two of these sites (I1, I2) are at the interface of V_H and V_L and one site (I3) is between C_{H1} and CL, as shown in Figure III-2a. To help a reader to locate sites I1, I2, I3 in the 3D structure, table 1 lists the residues at each site within 5Å of the bound HOOOH dimer. It is interesting that near I1 is trp109 on the heavy chain that is conserved across all antibodies and could be a potential sensitizing residue for the singlet oxygen.

All three sites are at the interface of *two* Greek key domains and hence we call this interface region as "**Inter Greek Key Domain Interface**" or IGKD. The two xenon-binding sites reported by Wentworth *et al*. [1] in the 4c6 structure lie in the IGKD, very close to the sites I1 and I2. Thus Xe1 is 18.4Å from Site I1 while Xe2 is 11.8Å from site I1 and 13.0Å from site I2.

Since the QM predicted mechanisms require two H_2O for the reaction with $^1O_2$, we would expect that the reaction site should have ordered water clusters at this site. Indeed Figure III-4a and 4b show that the crystal has higher ordered water clusters at site I1, I2, and I3, with several water dimers and trimers. Although the QM calculations use a second H_2O to catalyze the reaction of $^1O_2$ with H_2O, it is possible that these IGKD sites

that stabilize the water clusters, might also be able to replace the catalytic role of the H_2O (that is protons from the amino acids surrounding these sties might play similar roles).

4.1.2 Binding sites for product H_2O_2

The same HierDock procedure was used to search for sites in the 4c6 antibody structure that would stabilize the product H_2O_2. Here we find the two clusters (P1 and P2 shown in Figure III-2b) containing most of the highest binding structures. P1 is at the base of the antigen-binding site, completely overlapping the Xe2 site reported by Wentworth *et al.* [1]. P2 is between the C_L and C_{H1} domains and overlaps region I3. Both P1 and P2 are in the hydrophobic region between the barrel like interface of the variable and constant domains. In contrast to I1, I2, and I3, sites P1 and P2, do *not* exhibit bound water in the crystal structure, indicating that they are buried hydrophobic pockets.

The results derived from our docking studies of the intermediates and the product suggest that this catalytic reaction takes place in the interface regions of the variable and constant domains. This is supported by experimental evidence that shows strongly bound water dimers and trimers in these regions and the Xe binding studies suggesting that these regions are hydrophobic. Both predicted regions seem to be ideal for the reactions because of their ability to stabilize the key intermediates of the reaction cascade.

We also verified that the sites I1, I2, and I3 also stabilize the transition states for the reaction by performing a HierDock calculation for TS1, TS2, and TS3 [defined in Figure III-1 and Eqn. (2)] in the I1, I2, and I3 regions of the 4c6 structure. The transition state structures were kept rigid in all these docking studies. We found that the transition states cluster favorably in these regions.

4.2 Binding sites for HOOOH dimers, monomers, and H_2O_2 in other Immunoglobulin Fab fragments.

The formation of H_2O_2 has been observed for a large number of antibodies (over 200), all of which have been observed (1,2) to catalyze the conversion of $^1O_2$ to HOOH. This suggests that the reaction center is highly conserved across all antibodies. This may seem surprising since these antibodies include a reasonable diversity in sequences. However, the sites I1-I3 and P1-P2 we find to be important are associated with a unique structural motif of the fold in antibodies (and TCR) which might be rather insensitive to sequence. To test whether these sites would stabilize the intermediates for a range of antibodies, we selected three high-resolution (< 2Å) Fab structures (pdb codes: 2fb4, 1c5c, 1e60) that have maximally diverse sequences. This selection of structures was accomplished using the ClustalW sequence alignment program[14]. The three Fab structures selected have sequence identities of 47% to 68% with each other and with 4c6.pdb.

HierDock was performed across the entire antibodies to prevent a bias towards any particular sequence in docking protocol. In each case we find three clusters corresponding to I1-I3 and two corresponding to P1-P2 at the same positions as for 4c6. Thus the bound HOOOH dimer and monomer cluster along the V_H and the V_L interface of the IGKD for all three additional structures. This study confirms that IGKD fold is important in the catalysis of this reaction and the commonality of the binding sites for different sequences supports the IGKD region as the catalytic site.

4.3 Predicted binding sites of intermediates in T-cell receptor (TCR)

Experimentally it is known that TCR produces HOOH from $^1O_2$, just as for antibodies. To determine whether our procedure would explain this observation, we examined TCR (pdb code: 1tcr), which has the Greek key motif and the IGKD just as in antibodies. Again we used the HierDock protocol to perform an unbiased search for binding site across all regions of the TCR.

We found that the HOOOH monomers and dimers cluster at the heavy and light chain interface (sites I1-I3) of the TCR. This is consistent with the experimental observation that TCR does produce H_2O_2. Since, the sequence similarity between 4c6 and TCR is only 25%, this suggests that the essential feature is structural not sequence specific. These results support the conclusion that it is the IGKD interface created by the arrangement of immunoglobulin domains that is required for the stabilization of the intermediates.

4.4 Predicted binding sites of intermediates in b2-microglobulin:

β2-microglobulin has the characteristic Greek key motif present in antibodies, but it is monomeric and hence does *not* have the barrel-like interfacial structure of the TCR and the Fab region of antibodies. Consequently, we use HierDock to perform an unbiased search for binding sites across all regions of ß$_2$-microglobulin (pdb code: 1duz) to find favorable binding regions for HOOOH monomer, its dimer, and the transition states. However, we found no common consensus-binding region for the monomer and dimer in β2 microglobulin. The bound structures did not have a high population of docked conformations in any one region.

This indicates that the immunoglobulin fold by itself is not sufficient to catalyze the reaction. Rather, we require an interface created by the arrangement of

immunoglobulin domains, IGKD, to create the environment required for the stabilization of the intermediates. This is consistent with the results of Wentworth *et al.*, who showed experimentally that β2 microglobulin does *not* produce H_2O_2 from $^1O_2$. We attribute the lack of H_2O_2 production in β2 microglobulin to the absence of a hydrophobic interface lined with organized water molecules. This suggests that the unique feature responsible for the catalysis is the IGKD (only present for antibodies and TCR), not the Greek key fold (which is present in all immunoglobins, including β2-microglobulin and other proteins).

5.0 DISCUSSION

5.1 Nature of binding site for HOOOH monomer and dimer:

The two catalytic sites predicted here are at the interface of light and heavy chains of the antibody, *a structure unique to antibodies and TCR.* This IGKD interface of two Greek key domains is shown in Figs 3a and 3b. The two binding sites are each located on the sides of the barrel-like structural motif [15] at the interface of V_H and V_L, as shown in the inset of Figure III-3b. This structure has the beta sheets of V_H and V_L separated by ~ 5 Å, favoring the binding of the water sheet observed experimentally. The residues lining these sites shown in bold face in table 1 are strictly conserved and those in *italics* are conservative replacements. These results were obtained by performing a clustalW sequence alignment of the 37 Fab sequences having structure resolved to within 2.0Å. Trp109 in the I1 binding site is conserved across all antibodies and could be a potential sensitizing residue for the singlet oxygen.

There are a number of well-ordered crystallographic waters on the sides of this interfacial barrel-like structure between the V_H and V_L, as shown in Figure III-4a and 4b.

These waters are ordered in dimers (O--O distances approximately 2.6Å), trimers (O--O distances varying from 2.6 to 3.3Å), and a pentamer cluster (see Figure III-4a) with distances of 2.54 to 2.79Å. This pentamer ring of H_2O is in region I2 (it is formed by the crystallographic waters: Wat 12, 54, 60, 249, and 339). The water dimers shown in Figure III-4b are Wat5 and Wat404. Such well-ordered water clusters can be observed only in high-resolution crystal structures, such as 4c6 structure with 1.2Å resolution.

We consider that these water clusters are the H_2O structures that react with $^1O_2$, to form HOOOH, which subsequently reacts with a second HOOOH or $^1O_2$ to form H_2O_2 and the other reactive intermediates discussed above. Thus the first step of our QM mechanism involves two waters in a dimer-like structure, just as in Figure III-4a or 4b, with one of the waters acting as a catalyst in this step.

The I1, I2 sites determined using HierDock seem quite appropriate for the reaction to generate HOOOH from $^1O_2$ plus two H_2O. This product HOOOH is also favorable in this same site or in I3. It is plausible that a second HOOOH (formed from an additional $^1O_2$ and another H_2O dimer) could remain in these regions to combine with the first to form HOOOH dimer at either I1-I2 or at I3. This could then form H_2O_2 as in the QM mechanism. This H_2O_2 might then migrate to the sites P1-P2 that we find most favorable for H_2O_2.

A closer look at the interface of light and heavy chains of all antibodies shows that the bottom of the channel or barrel is capped by polar amino acids. For most antibodies these are glutamines forming a hydrogen bond network, as shown in Figure III-3b. We suggest that these residues could serve two functions.

- One, they could gate the reactants and various intermediates from entering the hydrophobic channel. Instead, these intermediates would go to the side of the barrel at the interface of light and heavy chain as shown in Figure3a.
- Second, they could prevent the H_2O_2 formed from escaping from the bottom of the barrel. This might direct them to be released towards the antigen-binding site.

To determine if these glutamines play a role in capping the products from the $^1O_2$ chemistry, it would be interesting to examine systems where the glutamines are mutated to hydrophobic residues.

For Fab our studies of binding HOOOH and its dimer and of H_2O_2 suggest the model that the IGKD motif is essential for H_2O_2 production from singlet oxygen. Since the F_c structure of antibodies have one such IGKD interface compared to two in the Fab structure, this suggests that the efficiency of HOOH production in Fc should be half that of Fab. Indeed Wentworth and Lerner[16] have shown that F_c structures have half the efficiency of Fab structures.

5.2 Geometric Pathway for the conversion of $^1O_2$ to HOOH.

A schematic geometric roadmap based on our proposed mechanism is given in Figure III-5 (for the 4c6 Fab structure).

a) We assume that $^1O_2$ may enter the antibody from near the Xe1 (and Xe2) xenon-binding site to migrate through the hydrophobic environment of V_H and V_L to the IGKD interface region (sites I1 and I2).

b) Here it $^1O_2$ can convert the clustered waters at this site to HOOOH.

c) This HOOOH might react with a second $^1O_2$ or it might migrate to the I3 site where it could react with a second HOOOH. In either case this reaction produces two HOOH. The HOOH products of this reaction might migrate to sites P1 and P2.

d) Subsequently these HOOH might migrate towards the interior of the barrel where H_2O_2 (or other intermediate such as HOOOH or the $(HOO)_2$ dimer) could react with the antigen. This might mark it for destruction.

Such a destructive role of consistent with the observation that $^1O_2$ is produced in processes involved with the macrophage engulfing the antigen bound antibody.

6.0 Conclusions:

Based on the experiments by Wentworth *et al.* showing that antibodies can catalyze $^1O_2$ to oxidize water to form H_2O_2 and based on the QM computational studies of Xu et al showing that the chemical mechanism involves production of HOOOH and subsequent reactions to form a series of products culminating in H_2O_2, we searched various proteins for special sties compatible with this chemistry.

Our HierDock studies lead to the conclusion that the interfacial motif IGKD, *between* two Greek keys (present only in antibodies and TCR and not present in β2 microglobulin) is critical to catalysis of $^1O_2$ to oxidize water to form HOOOH and H_2O_2. For both antibodies and TCR, we found sites (I1-I3) in the region favorable for binding the HOOOH reaction intermediates and sites (P1-P2) favorable for the H_2O_2 product. Based on these docking results and on the QM calculations, we propose a sequence of steps by which antibodies can produce HOOOH and H_2O_2 from $^1O_2$. These results suggest that such reactive intermediates as HOOOH and $(HOO)_2$ and the product HOOH are favorably formed in the IGKD paired Greek key barrel region close to the antigen.

We speculate that the conversion of $^1O_2$ to HOOOH and/or HOOH might provide for a protective function against singlet oxygen (which can attack dienes and other molecules in cells).

Alternatively these reactive intermediates might react with the antigen to help make the protein recognized by the antibody more susceptible to attack by other enzymes in the macrophage. This might provide a defense mechanism against the proteins having antigens to these antibodies. Here the HOOOH and/or HOOH might react selectively against just the antigen recognized. Based on the detailed prediction of binding sites involved in various steps, one can imagine a variety of biological experiments that might test our QM and HierDock results. Thus elective mutations could be made to enhance or inhibit various steps. These results suggest a number of experimental tests and provide a guideline for how to build biomimetic nanoscale systems to producing HOOH (or HOOOH).

These computational studies provide mechanistic insight to the experimental observations by Wentworth *et al.* that antibodies and TCR can catalyze the conversion of $^1O_2$ plus water to H_2O_2. The results gives very close agreement with observed isotope ratio of 2.2:1. In particular the results explain the observed molecularity of 2.0 for the number of HOOH produced per $^1O_2$. This supports strong support for the QM mechanism.

6. Acknowledgments

We thank Richard Lerner for suggesting this problem and Albert Eschenmoser, Paul Wentworth, Anita Wentworth, Lyn Jones, and Kim Janda for helpful discussions.

We also thank Xueyong Zhu, Nicholas Larsen, and Ian Wilson for access to their 1.2Å structure for the chimeric Fab antibody prior to publication.

This research was funded by NIH (HD 36385-02). The facilities of the MSC used in these studies were funded by NSF-MRI, DURIP (ARO and ONR), and the Beckman Institute. In addition the MSC is funded by grants from DOE-ASCI-ASAP, ARO-MURI, NIH, NSF, Avery-Dennison, Asahi Chemical, Chevron, 3M, Dow Chemical, Nippon Steel, Seiko-Epson, and Kellogg's.

References

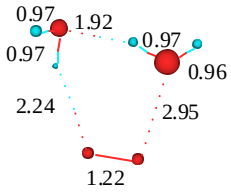
1. Wentworth, P., Jones, L.H., Wentworth A.D., Zhu, X.Y., Larsen, N.A., Wilson, I.A., Xu, X. Goddard, W.A., Janda, K.D., Eschenmoser, A., and Lerner, R.A., (2001*), Science*, **293,** 1806-1811.
2. Wentworth, A.D., Jones, L.H., Wentworth, P., Janda, K.D., and Lerner, R.A.,(2000), *Proc Natl Acad Sci* U S A,. **97**, 10930-10935
3. Xu, X., Muller, R.P. and W. A. Goddard, III., *Proc. Natl. Acad. Sci* USA (companion paper).
4. Floriano, W.B., et al., (2000), *Proc Natl Acad Sci* U S A, **97**(20) 10712-10716.
5. Datta, D., Vaidehi, N, Floriano, W.B.,. Kim, K.S., Prasadarao, N.V., and Goddard III, W.A., (2001), *Proteins: Structure, Function and Genetics* (submitted).
6. Wang, P., Vaidehi, N., Tirrell, D.A., and Goddard III, W.A., (2001), *J. Am. Chem. Soc.* (submitted).
7. Ewing, J.A., Kuntz, I.D., (1997), *J Comp Chem*, **18**, 1175-1189.
8. Lim, K-T. Brunett, S., Iotov, M., .McClurg, R.B., Vaidehi, N., Dasgupta, S., Taylor, S., and Goddard III, W.A.,(1997), *J Comp. Chem*.**18**, 501-521.
9. Mayo, S.L., Olafson, B.D., and Goddard III, W.A., (1990), *J. Phys. Chem*. **94**: 8897-8909.
10. Ghosh, A., Rapp CS, Friesner RA, (1998), *J. Phys. Chem. B*, **102** 10983-10990.
11. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T., Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe, M., Wiorkiewicz-Kuczera J, Yin D, Karplus M, (1998), *J. Phys. Chem.*B , **102:** 3586-3616 .
12. Zhu, X.Y., Larsen, N.A. and Wilson, I.A., private communication. We thank Dr. Wilson for providing the 1.2Å 4c6.pdb , Fab structure prior to publication.
13. Vaidehi , N., and Goddard III, W.A., (1997), *Proc. Natl. Acad. Sci. USA*, **94**, 2466-2471.
14. Eddy, S.R., *Multiple alignment using hidden Markov models.* Proc Int Conf Intell Syst Mol Biol, 1995. **3**: p. 114-20.

15. Branden, C., and J. Tooze, *Introduction to protein Structure*. 1998, New York: Garland Publishing Co. 306-309.
16. Wentworth, P. and Lerner, R.A. private communication

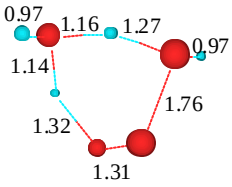
Table III-1 : List of residues in the 4c6 Fab structure in the three predicted binding sites I1, I2 and I3 of the HOOOH dimers. The bold face residues are strictly conserved across 37 aligned sequences of Fab. The residues in italics are conservative replacements.

I1		I2		I3	
V_L	V_H	V_L	V_H	C_L	C_{H1}
Ser 48	*Gln 3*	Asp 1	Lys 45	Val164	**Leu 147**
Lys 50	Leu 4	Pro 100	Glu 47	Leu 165	*Lys 149*
Arg 51	Gly 107	Tyr101	Trp 48	Asn 166	**Phe 172**
	Ser 108	Thr 102	Asn 61	*Ser 167*	*Ala 174*
	Trp 109		Pro 62	Ser 181	**Pro 173**
	Gly 110		Ser 63	Ser 182	Val 175
				Thr 183	**Tyr 181**
					Thr 182
					Leu 183
					Ser 184

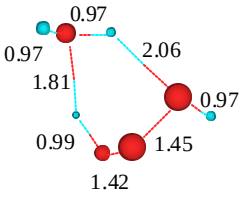
Figure III-1. Gas phase structures (optimized using quantum mechanics, see reference [3]) for various clusters and transition states. These structures were used in the docking studies.



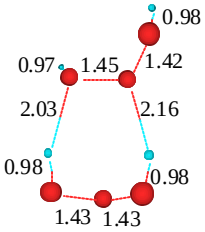
R1



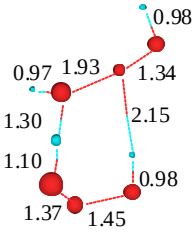
TS1



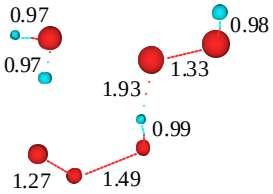
P1



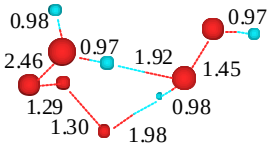
R2



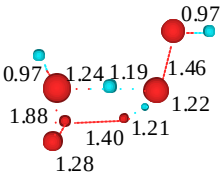
TS2



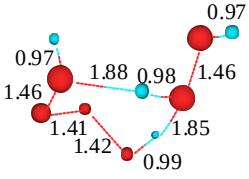
P2



R3

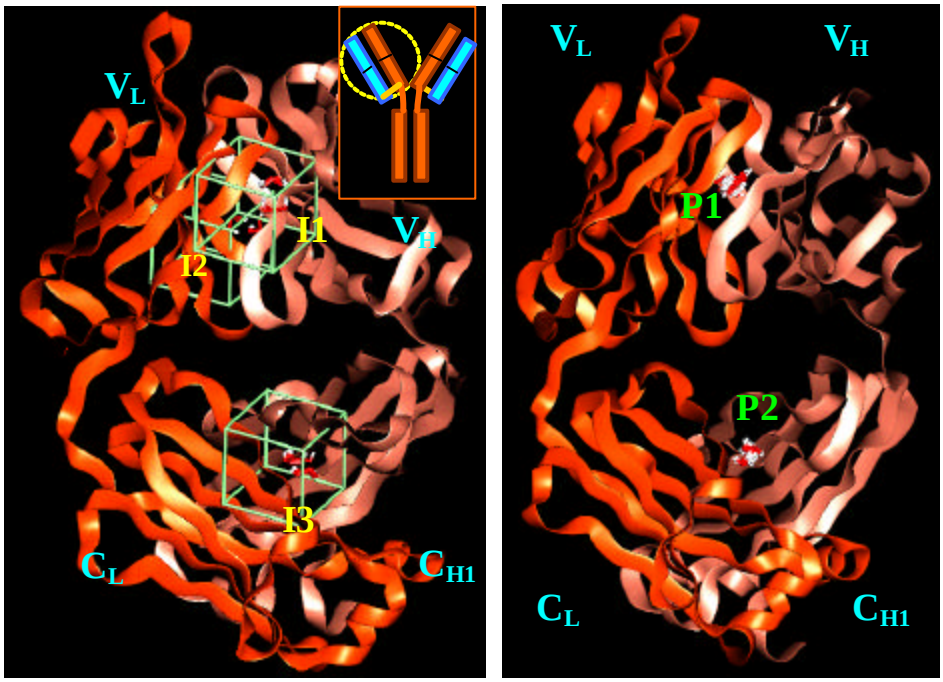


TS3



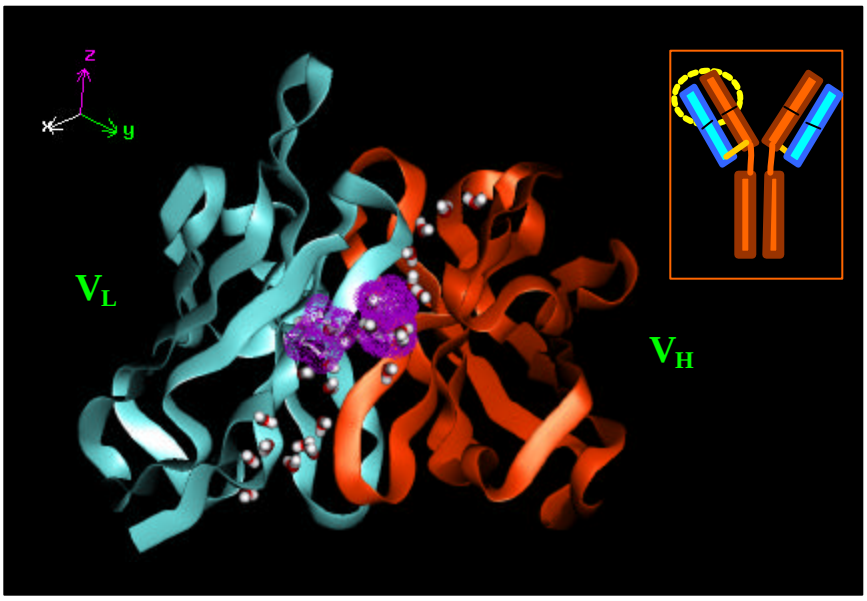
P3

Figure III-2. (a) Clustering sites for docking of HOOOH dimers. All sites are located between the V_L and V_H interface. This shows Regions I1 and I3 in front. Region I2 is opposite region I1 in the back. The inset shows where this region is relative to the overall immunoglobulin. Regions I1-I3 are in the Inter Greek Key Domain Interface (IGKD) unique to antibodies and TCR. (b) Clustering sites for docking of H_2O_2. Region P1 is situated within the Beta-barrel created by the V_H and V_L interface. P2 is located between the C_H and C_L interface. Regions P1-P2 are in the Inter Greek Key Domain Interface (IGKD) unique to antibodies and TCR.

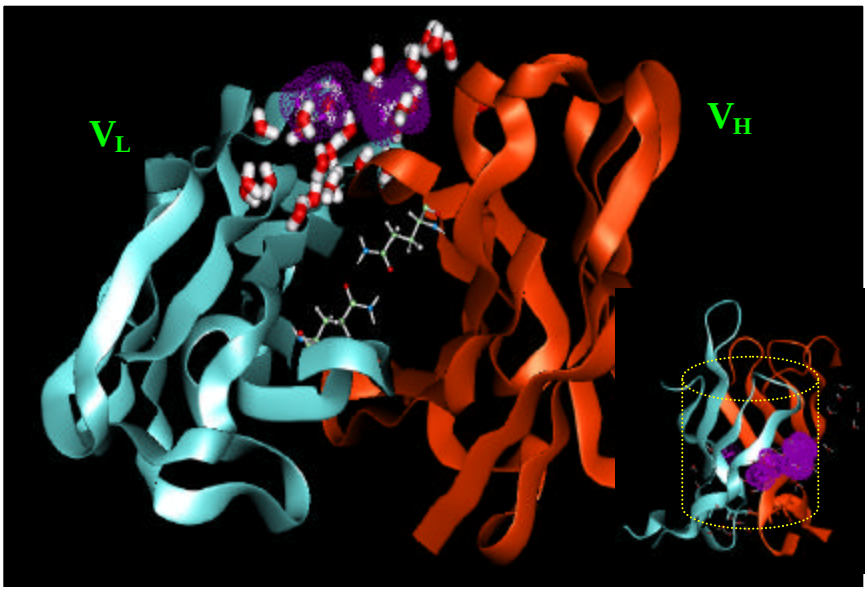


(a) (b)

Figure III-3. (a) The purple dots indicate two regions (A and B) of the Fab antibody fragment that bind strongly to the HOOOH dimer and which we conclude are plausible regions for the catalysis of $^1O_2$ plus H_2O dimer to form HOOOH. Here A is on the left and B is on the right. These sites are at the interface of the V_H and V_L in a region containing well-ordered crystallographic waters (shown with half bonds). Regions A and B are in the Inter Greek Key Domain Interface (IGKD) unique to antibodies and TCR. The inset shows a schematic antibody structure with a yellow circle to indicate the region magnified. (b) The structure in (a) is rotated 90º about the horizontal axis to show the hydrophobic channel bounded by Gln38 from V_L and Gln39 from V_H. This forms a hydrogen bond network at the mouth of the barrel. (Region A is again at the left) The inset shows the barrel like structure (containing two Greek keys) unique to antibodies that we suggest is critical to the catalysis.

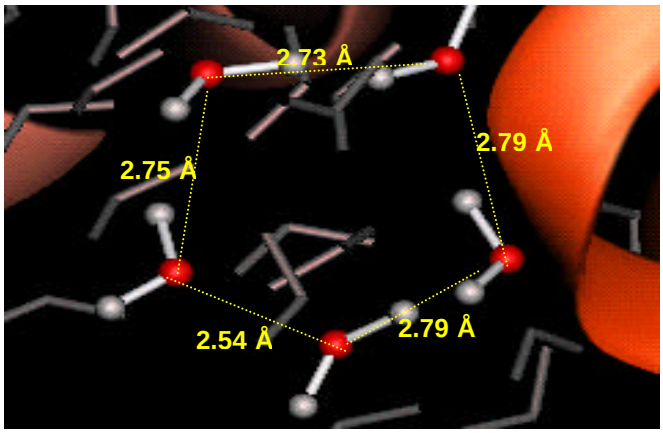


(a)

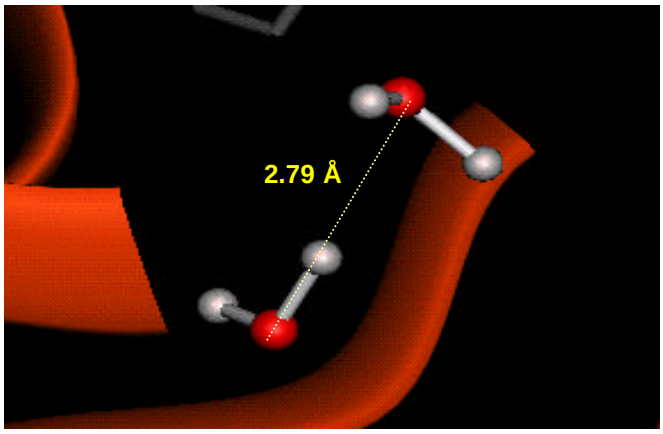


(b)

Figure III-4: Ordered water molecules found in the crystal structure at the IGKD interface of the Fab antibody fragment. (a) A pentamer ring of H_2O molecules with each hydrogen bonded to two others. (b) An example water dimer where hydrogen from one water molecule is pointing towards the oxygen of the other molecule

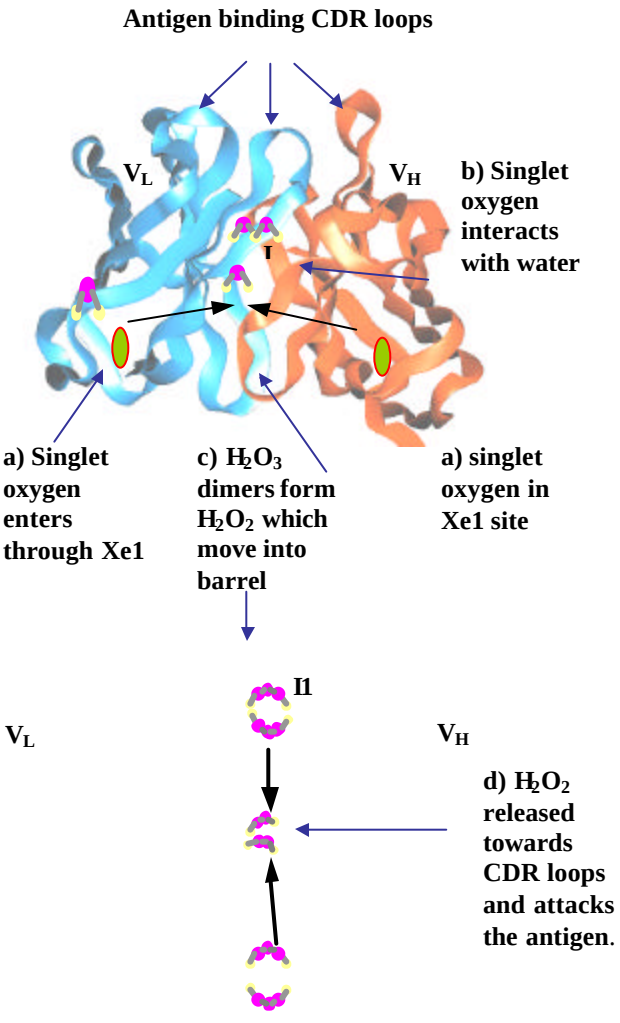


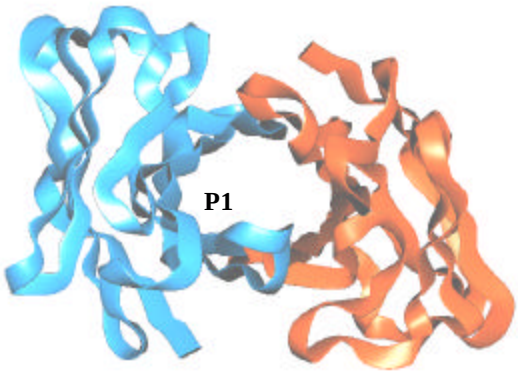
(a)



(b)

Figure III-5. The geometric pathway for the sequence of reactions converting $^1O_2$ and water to HOOOH and then to HOOH. Here we assume that $^1O_2$ enters the hydrophobic region near Xe1 (and Xe2). At I1 (or I2) it can react with a water dimer (or trimer) to form HOOOH. The HOOOH may stay at I1 (or I2) but it may go to I3, which does not have crystallographic waters. This HOOOH may react directly with a second $^1O_2$ or with the HOOOH from a previous reaction to form the HOOOH dimer. This may occur at I3. The HOOOH dimer can rearrange through a series of steps to form HOOH, which may go to sites P1 or P2 (there are no crystallographic waters at these points). Here the HOOH is positioned close to the region at which antigen may be bound (HOOOH may also go to this region). From here the HOOH (or HOOOH) might react directly with the part of a protein whose antigen is recognized by the antibody.


Antigen binding CDR loops
V_L
V_H
I
b) Singlet oxygen interacts with water
a) Singlet oxygen enters through Xe1
c) H_2O_3 dimers form H_2O_2 which move into barrel
a) singlet oxygen in Xe1 site
I1
V_L
V_H
d) H_2O_2 released towards CDR loops and attacks the antigen.


P1
I2

Chapter 4

Selectivity And Specificity Of Substrate Binding In Methionyl-tRNA Synthetase

Adapted from an unpublished manuscript coauthored with Nagarajan Vaidehi, David Zhang, Professor David A. Tirrell and Professor William A. Goddard III

Abstract

In vivo incorporation of amino acids in protein biosynthesis is a precisely controlled mechanism. The accuracy of this process is controlled to a significant extent by a class of enzymes called the aminoacyl tRNA synthetases. Aminoacyl tRNA synthetases achieve this control by a multi-step identification process that includes "physical" binding and "chemical" proofreading steps. However, the degree to which each synthetase uses these specificity-enhancing steps to distinguish their cognate amino acid from the non-cognate ones vary considerably. We have used Hier-Dock computational protocol to elucidate this binding mechanism in methionyl tRNA synthetase (MetRS) by first predicting the recognition site of methionine in the apo form of methionyl tRNA synthetase (apo-MetRS). We have developed this generalized procedure, which can be used to search for ligand binding region in globular proteins with no prior information about the binding site. We have further investigated the specificity of MetRS towards the binding of 19 other natural amino acids to both apo-MetRS and to the co-crystal structure of MetRS with methionine bound to it (co-MetRS). We have established through our computed binding energies that the discrimination towards the non-cognate substrate increases in the second step of the physical binding process that is associated with a conformation change in the protein.

Introduction

Specific recognition of amino acids by their corresponding tRNAs and aminoacyl-tRNA synthetase (aaRS) is critical for the faithful translation of the genetic code into protein sequence information. The aaRSs catalyze a two-step reaction in which amino acids are esterified to the 3? end of their cognate tRNA substrates [1]. In the first step, the amino acid and ATP are activated by the aaRS to form an enzyme-bound aminoacyl-adenylate complex. In the second step, the activated amino acid is transferred to the 3?-ribose of the conserved CCA-3? end of the cognate tRNA. The fidelity of protein synthesis depends, in part, on the accuracy of this aminoacylation reaction. aaRSs bind their cognate amino acid through a multi-step recognition process and correction mechanisms that include physical binding and a chemical proof reading [2]. The four major steps involved in the transfer of aminoacyl group to the t-RNA are:

1. Binding of amino acid and ATP
2. Conformational change in the aaRS induced by binding and formation of the aminoacyl-adenylate complex.
3. Proof reading of misactivated non-cognate aminoacyl adenylate complex
4. Transfer of aminoacyl to the tRNA and proof reading

The physical binding of the amino acid and ATP to aaRS is achieved in steps 1 and 2, which is accompanied by a conformation change in the aaRS. However, this binding event is necessary but not sufficient for the incorporation of the analog or the cognate amino acid during protein biosynthesis. Binding is followed by chemical

proof reading steps 3 and 4 which are termed as the pre-transfer and the post-transfer proof reading steps, respectively. With every step, the aaRS recognizes its cognate amino acid with increased specificity, while discriminating more efficiently against the non-cognate amino acids. However, the degree to which each aaRS uses the specificity enhancing steps varies considerably with regard to the twenty naturally occurring amino acids and the type of aaRS. For example, tyrosyl t-RNA synthetase has the highest specificity in the first binding step whereas isoleucyl tRNA synthetase, achieves maximum discrimination in the pre-transfer proofreading step [3-6].

Many research groups have focused on the use of *in vivo* methods for incorporating the non-natural amino acid analogs into proteins. It has been demonstrated that the wild-type translational apparatus can use non-natural amino acids with fluorinated, electroactive, unsaturated and other side chain functions [7-13]. However, the number of amino acids shown conclusively to exhibit translational activity *in vivo* is small, and the chemical functionality that has been accessed by this method remains modest. Only those analogs that are able to successfully circumvent the multi-step filter mechanisms of the natural synthetases eventually get incorporated.

With an increase in efforts of incorporating artificial amino acids *in vivo,* it has become vital to enhance our understanding of the molecular level mechanism at different steps that aaRSs utilize to ensure high fidelity in translation. A better understanding of this mechanism will also be very useful in allowing us to design mutants of aaRS for incorporation of specific analogs and also in suggesting analogs that are more efficiently

incorporated [14-16]. In this study, we have implemented a computational procedure to gain insight into the binding mechanism of methionyl-tRNA synthetase (MetRS). Computational methods are becoming increasingly important to understand the molecular level mechanisms that are not feasible with experiments and also for faster virtual screening of analogs prior to synthesis.

MetRS is a class I aaRS, and undergoes a large conformational change on upon binding to methionine. It is a dimeric protein and the crystal structures of *E. coli* MetRS in its apo form and as a co-crystal with its native ligand, methionine, have been solved to 1.85Å and 2.03Å resolution, respectively. [We refer to the apo form of MetRS protein structure as apo-MetRS(crystal) and the co-crystal structure of *E. coli* MetRS with methionine as Met/MetRS(crystal). Note that the protein conformations in both these crystal structures are different especially in the binding site. The symbol MetRS always denotes the *E. coli* MetRS unless otherwise specified.] Both *in vivo* incorporation of methionine analogs into proteins and their *in vitro* measurements of the rate of incorporation have been studied extensively and it has been demonstrated that MetRS is one of the more permissive aaRS for the incorporation of a large number of analogs [9]. We are interested in computationally determining the specificity of MetRS for the natural non-cognate amino acids and methionine analogs in the steps of amino acid recognition and binding. A better understanding of its binding mechanism would be useful to streamline a virtual screening approach for the incorporation of non-natural amino acids.

We have used the HierDock computational protocol to first predict the binding site of methionine in MetRS in the apo-MetRS (crystal) [17]. We scanned

through the entire protein (except the anticodon recognition region) for predicting the preferential binding site for methionine using no knowledge from the crystal structure of Met/MetRS (crystal). We refined the HierDock protocol and derived what we call as the "recognition site" which includes all the residues in the binding pocket of methionine as seen in Met/MetRS (crystal) however, methionine is oriented in this pocket with its side chain exposed to solvent. Our results indicate that the first step to amino acid binding is the recognition of the zwitterion part of the ligand which is referred to as the "recognition mode". We find that apo-MetRS is able to distinguish methionine from the non-cognate natural amino acids but has cysteine and serine as competitors. We also find that MetRS in the Met/MetRS (FF) protein structure has better discrimination for the twenty amino acids and once again methionine has the best binding energy in this structure with Gln as a close competitor.

The calculated binding energies of methionine analogs are correlated with either the *in vivo* incorporation results or the *in vitro* measurements of rate of the aminoacyl adenylate formation. We find that in Met/MetRS (FF) protein, the analog with high incorporation rates bind better than those that do not get incorporated. In an attempt to incorporate novel methionine analogs Kiick *et al.* reported that Homopropargylglycine (myag) replaces methionine most efficiently utilizing the natural translation apparatus of *E. coli* while cis-crotglycine (ccg) shows almost no incorporation [18]. Our calculated binding energies correlate well with the *in vivo* incorporation trends exhibited by these analogs and with the binding energies calculated by *in vitro* methods.

Methods

Preparation and Optimization of Structures.

Ligand Structures: Both the neutral and the zwitterion forms were used for all the twenty natural amino acids as well as the five methionine analogs. The ligand conformations were optimized in the extended conformation at the Hartree-Fock level of theory with a 6-31G** basis set, including solvation according to the Poisson-Boltzman functional using the Jaguar computational suite [19] (Schrödinger, Portland, OR). The Mulliken charges ascertained from this calculation were retained for the subsequent molecular mechanics simulations. The conformations of the 5 methionine analogs are shown in Figure IV-4a.

Preparation and Optimization of Protein Structures: The 2.03Å *E. coli* apo-MetRS structure was obtained from PDB database (pdb code: 1QQT) that included the fully active monomer α chain of a homodimer, crystal waters, and a zinc (II) ion [20]. CHARMM22 charges with the nonpolar hydrogen charges summed onto the heavy atoms were assigned to the α chain according to the parameters set forth in the DREIDING force field [21]. The protein was neutralized by adding counterions (Na^+ and Cl^-) to the charged residues (Asp, Arg, Glu and Lys) and subjected to a minimization of the potential energy by the conjugate gradient method using Surface Generalized Born continuum solvation method [22]. The RMS in coordinates (CRMS) of all atoms after minimization is 0.68Å and this structure is referred to as apo-MetRS(FF). Using the same procedure the co-crystal structure of *E. coli* MetRS (pdb code: 1F4L; resolution 1.85Å) was minimized and the CRMS for all atoms of

the minimized structure compared to the crystal is 0.57Å [23]. We refer to this structure as met/MetRS(FF). The CRMS values for both the structures are well within experimental error that demonstrates the proficiency of the FF used in present studies. The crystal waters and other bound molecules were removed for docking to maximize the searchable surface of the protein. We have used continuum solvation method for all structure optimizations and energy scoring in this study with an internal protein dielectric constant of 2.5 was employed for all calculations.

HierDock Protocol

We use the HierDock procedure, which has been applied successfully to study the binding of odorants to membrane-bound olfactory receptors [17, 24] for outer membrane protein A binding to sugars [15] and for Phenylalanine and its analogs binding to PheRS [25], [26]. The HierDock ligand screening protocol follows a hierarchical strategy for examining conformations, binding sites and binding energies. Such a hierarchical method has been shown to be necessary for docking algorithms [27]. The steps in HierDock involve using coarse grain docking methods to generate several conformations of protein/ligand complexes followed by molecular dynamics (MD) simulations including continuum solvation methods performed on a subset of good conformations generated from the coarse grain docking. Methods combining docking and MD simulations have been tested [28] but the main drawback of these tests were that only a single protein/ligand complex structure was kept from the coarse grain docking methods for MD simulations. This is risky considering that the coarse grain methods do not have reliable scoring functions that include solvation.

Free energy perturbation methods are generally regarded to lead to accurate free energies of binding but are computationally intensive and not readily applicable to a wide variety of ligands [29]. Our goal is to derive a fast hierarchical computational protocol that uses hierarchical conformation search methods along with different levels of scoring functions, which would allow screening of amino acid analogs for aaRSs. The three major steps in HierDock procedure in this paper are:

- First, a coarse grain docking procedure to generate a set of conformations for ligand binding. In this paper we used DOCK 4.0 [30, 31] to generate and score 20000 configurations, of which 10% were ranked using the DOCK scoring function.

- We then select the 20 best conformations for each ligand from DOCK and subject them to annealing molecular dynamics (MD) to further optimize the conformation in the local binding pocket, allowing the atoms of the ligand to move in the field of the protein. In this step the system was heated and cooled from 50K to 600K in steps of 10K (0.05 ps at each temperature) for 1 cycle. At the end of annealing MD cycle, the best energy structure is retained. Annealing MD allows the ligand to readjust in the binding pocket to optimize its interaction with the protein. This fine grain optimization was performed using MPSim [32] with DREIDING forcefield [21] and continuum solvation methods. We use the SGB continuum solvent method to obtain forces and energies resulting from the polarization of the solvent by the charges of the ligand and protein. This allows us to calculate the change in the ligand structure due to the solvent field and hence, obtain more realistic binding

energies that take into account the solvation effects on the ligand/protein structure. The annealing MD procedure generated 20 protein/ligand complexes for each ligand.

- For the 20 structures generated by annealing MD simulations for each ligand, we minimized the potential energy (conjugate gradients) of the full ligand/protein complex in aqueous solution using SGB. This step of protein/ligand-complex optimization is critical to obtaining energetically good conformations for the complex (cavity + ligand). Then we calculated binding energies as the difference between the total energy of the ligand-protein complex in solvent (ΔG(protein+ligand)) and the sum of the total energies of the protein (ΔG (protein)) and the ligand separately in solvent (ΔG(ligand)). The energies of the protein and the ligand in solvent were calculated after independent energy minimization of the protein and the ligand separately in water. Solvation energies were calculated using Poisson-Boltzmann continuum solvation method available in the software Delphi [33]. The non-bond interaction energies were calculated exactly using all pair interactions. Thus the binding energy is given by

$$\Delta\Delta G_{calc} = \Delta G(protein + ligand) - \Delta G(protein) + \Delta G(ligand) \qquad (1)$$

Since the structure optimizations included solvation forces using the SGB continuum solvent approximation with the experimental dielectric constant, we consider that the calculated energies are free energies [34]. This multi-step scanning procedure is based on docking *via* DOCK 4.0 coupled with fine-grain MM techniques. The coarse grain docked complex structures generated are

scored with FF and differential solvation, which effectively filters the docked complexes to isolate the top contenders. As demonstrated in Kekenes-Huskey 2002**,** Dock 4.0 structures vary erratically with rank, whereas filtering with MPSim optimization brings the best structures to the top of the rank list.

Scanning the entire apo-MetRS(FF) for predicting binding site of met: For the case of apo-MetRS(FF) we wanted to test the HierDock procedure for scanning the entire protein for the favorable binding site of met. However, it has not been tested for a case where the protein undergoes a large conformational change in the binding site after the ligand binding starting from apo-protein structure. The steps involved in the scanning procedure are:

1. *Mapping of possible binding regions*. A probe of 1.4 Å radius was used to trace a 4 dots/Å negative image of the protein molecular surface, according to Connolly's method [35]. The resulting data was used to generate clusters of overlapping spheres with the SPHGEN program. These spheres serve as the basis for the docking method.
2. *Definition of docking region*. The pockets of empty space of the receptor (apo-MetRS(FF)) surface represented by spheres were divided into 14 possible 10 Å wide overlapping cubes, which covered the entire protein surface. Each region was scanned to determine its suitability as a binding site. The site that contains the greatest number of lowest energy docked conformations is designated as the putative binding region.
3. *Prediction of binding site*: Steps 1 to 3 of HierDock procedure was performed with methionine as the ligand in all the 14 possible binding

regions in the entire apo-MetRS(FF). The orientations of the ligand in the receptor were generated by DOCK 4.0, using flexible docking with torsional minimization of the ligand, a continuum dielectric of 1.0 and a distance cutoff of 10 Å for the evaluation of energy.

4. *Selection of the most probable binding site and best configurations*: The best conformation from each region was determined using the buried surface area cutoff criteria for the ligand along with the binding energy. Such a buried surface area cutoff is required for filtering at the coarse grain level. An average of the most buried and the least buried conformer was calculated and all conformers whose buried surface area was lower than the average were eliminated from further analysis [36]. The conformations that passed the buried surface area filter were sorted by binding energies calculated using equation (1) and the conformation with the best binding energy in every region were compared between regions. All the complex energies were calculated. The region with the lowest energy binding energy calculated using equation (1) was selected as the preferential binding region.

Docking of ligand pool into the binding region and calculating relative binding energies*:* Steps 1 to 3 of HierDock procedure was performed for all the ligands in the ligand pool in the putative binding region and the relative binding energies for the best ligand conformations were calculated using equation (1). The ligands (20 natural amino acids and analogs of met) were ranked according to binding affinities to determine which ligands have the

highest affinity for the binding site. The best energy conformation of methionine in optimized apo-MetRS(FF) structure is the predicted structure of methionine in apo-MetRS(FF). We denote this predicted structure as met/apo-MetRS(FF).

Prediction of binding site for met/MetRS(FF) co-crystal structure: For the case of met/MetRS(FF) structure, the receptor was prepared by removing the methionine from the met/MetRS(FF) structure. The protein surface was mapped with spheres, as described above, and the binding regions was covered by a 12 Å $\times$ 12 Å $\times$ 12 Å box centered in the center of mass of methionine ligand. Only this region was used in subsequent docking. Steps 1 to 3 of HierDock procedure were performed using the same set of control parameters but only in the known binding region. The conformation with the best energy binding energy in this region calculated using equation (1), starting from the protein structure in met/MetRS (FF) is the predicted co-crystal structure of met/MetRS. We denote this predicted structure of methionine in MetRS co-crystal structure as met/MetRS (HierDock).

Docking of ligand pool into the binding site and calculating relative binding energies in met/MetRS (HierDock) We performed steps 1 to 3 of HierDock procedure for all 20 natural amino acids and the methionine analogs in the 12Å x 12Å x12Å binding region and the relative binding energies for the best ligand conformation for each ligand was calculated using equation (1). The ligands (20 natural amino acids and analogs of met) can then be ranked

according to binding affinities to determine which ligands have the highest affinity for the binding site.

Binding energy calculation of the 20 natural amino acids and methionine analogs in the conformation that activates the protein: HierDock protocol predicts the best energy conformation for each ligand (20 natural amino acids and methionine analogs) in the defined 12Å x 12Å x 12Å binding region in Met/MetRS(FF) structure. These predictions give rise to different preferred binding conformation for each ligand. However, the orientation that methionine adopts in the makes the necessary contacts required for the enzymatic activity is referred to as the "activation mode". To assess the relative binding energies of the twenty natural amino acids and their analogs in the activation mode perturbation calculations for all the ligands were performed as follows:

- An amino acid rotamer library [37] was used to generate all the conformations of each amino acid in the binding site, and a similar library was generated for the five methionine analogs.
- The best rotamer was chosen by matching each rotamer k in the binding site and evaluated with the following equation using the Dreiding force field:

$$E_k = \sum_{i,j}\left(\frac{q_i q_j}{4\pi\varepsilon r_{ij}} + D_e\left(\left(\frac{r_m}{r_{ij}}\right)^{12} - 2\left(\frac{r_m}{r_{ij}}\right)^{6}\right) + D_{HB}\left(5\left(\frac{r_{HB}}{r_{ij}}\right)^{12} - 6\left(\frac{r_{HB}}{r_{ij}}\right)^{10}\right)\cos^4\theta\right) \quad (2)$$

where i and j sum over all atoms in the ligand and protein residue residues in the binding site, q_i and q_j are partial charges on atoms i and j, respectively. r_{ij} is the distance between atoms i and j, and r_m and D_e are van der Waals distance and well

depth of atoms *i* and *j*, r_{HB} and D_{HB} are hydrogen bond distance and well depth, respectively. θ is the hydrogen bond angle between atoms *i, j* and their bridging hydrogen atom. The hydrogen bond term is only evaluated for hydrogen bond donor and acceptor atoms. To avoid over penalizing clash, the van der Waals radii were reduced to 90% of the standard values in the Dreiding force field.

- After the best rotamer was chosen for each ligand, the total energy was minimized in the presence of protein, and the binding energy was then calculated using equation (1) for each of the twenty natural amino acids in the "activation mode" and compared.

Results and Discussion

Prediction of Binding site of methionine in apo -MetRS(FF) and Met/MetRS(FF)

Figure IV-1 shows the location of region14 box in apo-MetRS(FF) which was determined to be the binding region by sifting through the 14 regions in apo-MetRS. The best conformation of methionine in this region shows methionine to be making electrostatic interactions with His301 and Asp52 (Figure IV-3c), the two amino acids that have been shown to play a significant role in methionine binding [38, 39]. His 301 to alanine mutation results in loss of the affinity for methionine and D52A mutation reduces the *Kcat* of the adenylation reaction by 4 folds indicating that it has a major role on the catalytic step in the formation of methionyl adenylate. Tyr15, another key amino acid determined by mutation analysis and has been structurally observed in the co-crystal structure to form the binding pocket for methionine [40], [23] is located within 5Å of the docked methionine. The main component of the binding energy in our predicted binding orientation comes from the electrostatic

interactions that methionine makes with Asp52 and His 301 followed by its Van der Waals interactions in this binding region.

The docked orientation in the apo enzyme occupies the identical position in the binding pocket as seen in the co-crystal structure (Figure IV-3a). However the orientation of methionine and the residues lining the binding pocket including parts of the protein backbone are very different in the two conformations of MetRS. Although methionine seems to be making electrostatic contact with Asp52 the anchoring residue, the side chain of methionine is not buried in the 7Å pocket. The reason for this is that we have used the unbounded structure of the synthetase, which, on binding to the amino acids undergoes significant conformation change. The co-MetRS structure suggests that the large solvent exposed cavity become reduced in volume as it gets partially filled with methionine and Y15, W253, F300, W229, F304 and Y251. These residues are significantly displaced from their apo-enzyme orientation as they reorient to form a hydrophobic pocket for methionine. In our predicted binding mode of methionine in the apo-enzyme, all these residues are within 5Å of methionine ligand. We expect this to be the initial binding orientation of methionine.

Another interesting observation that substantiates that the predicted orientation of methionine could be the initial binding mode is that methionine has one of the best binding energies of all 20 natural amino acids in this region. The specificity of this site further confirms that we have been able to find the correct binding region. Methionine has serine and cysteine as close competitors but they get eliminated as the protein undergoes conformation change. In an attempt to force the side chain of

methionine to be buried in the pocket we did annealing dynamics of the entire complex with solvation and reduced VDW radii of the ligand atoms. However, the orientation of methionine did not change.

Also, a number of residues within 5Å of methionine in this region are conserved among a large number of organisms. In a sequence alignment among 59 prokaryotes we find all the amino acids within 4Å of methionine in the predicted binding region are either strictly conserved or are conserved replacements. Of the 12 residues within 4Å of methionine, 7(Y15, D52, V252, W253, A256, Y260, H301) are strictly conserved and 5 (A12, L13, P14, P257, F300) are conserved replacements (Figure IV-3c). This is interesting considering that there are only 21 positions in the entire alignment that are strictly conserved and we find a third of them in our predicted binding region without any prior knowledge of the binding site. A binding search protocol for unliganded proteins followed by a sequence alignment analysis for the predicted binding region could provide more evidence on the accuracy of the predicted binding site and help in recognizing key amino acids lining binding pocket. Generally, one would expect to see conserved residues or conservative replacements in substrate binding sites in proteins across various species.

We also docked methionine in the binding region of the co-MetRS(FF). This test was performed to check if we were able to predict the crystallographic binding orientation of methionine in the binding pocket. This test was important to validate the accuracy of our docking protocol and the force field. Our predicted structure had a CRMS deviation of *0.*55Å from the crystal structure (Figure IV-3a).

IV b. Specificity for methionine in 1QQT and 1FTM

We docked all 20 amino acids and calculated their binding energies in the predicted binding region in apo-MetRS(FF**)** and the crystallographic binding site in Met/MetRS(FF). We also did perturbation studies of the natural amino acids in these two structures. The perturbation studies were done to analyze the binding energies of the non-cognate amino acids if they oriented in a similar conformation in the binding site as methionine.

Perturbation analysis: In the case of apo-MetRS(FF**)** closest competitors for methionine are serine and cysteine. However, as the enzyme undergoes conformation change, its ability to discriminate against these non-cognate residues increases significantly. It has been noticed that for most synthetases there is no absolute specificity for the cognate substrate in the sense of a "lock and key" model. For example, Yeast IleRS is not able to distinguish between Trp and Ile in the first step of binding because of the higher hydrophobic interactions gained by the non-cognate substrate. However, as the initial binding process is completed, the enzyme is able to discriminate against the non-cognate amino acids more easily [2, 5].

In Met/MetRS(FF), methionine has the best binding energy, and it has an energy difference of more than 20 kcal/mol with its closest competitors, Asn and Arg. The closest competitors from the first binding step (Leu, Glu and Gln) are discriminated against with a very high efficiency as the structure of the protein changes.

Docking analysis: The docking study was done predominantly to recognize possible competitors of methionine. It may be possible that a non-cognate amino acid binds at

the methionine binding pocket but does not make the critical interactions that methione makes in this binding pocket. In such cases, the amino acid may not be able to react with ATP and charge the tRNA. In apo-MetRS(FF), methionine has the best binding energy of –26.38 kcal/mol with leucine glutamine and glutamate as the closest competitors. In Met/MetRS(FF) methionine again has the best energy with glutamine and serine as the closest competitors. Gln, in its preferred binding site in Met/MetRS(FF) has its zwitterions part and the $?_1$ torsional angle in the same orientation as methionine at this site. Yet, its $?_2$ and $?_3$ angles are significantly different from that of methionine. The S^d of methionine makes two hydrogen bonds – one with the terminal oxygen of Tyr260 and the other with the backbone amide of Leu13. However, because of the difference in its binding mode, Gln is unable to make a hydrogen bond with Tyr260 and makes only a weak hydrogen bond with the backbone amide of Leu13 (O—H-N distance of 3.9Å).

One more observation is that the order of binding of the amino acids is identical in the docking analysis in apo-MetRS(FF) and the perturbation study in Met/MetRS(FF). It indicates that when the enzyme undergoes structural change, if all the amino acids were to bind in the binding mode of methionine in the co crystal structure, their order of binding would remain the same as indicated by the apo enzyme. However, the magnitudes of binding energies, which indicate the level of discrimination, would be very different.

IV c. Binding energies of analogs

To test the sensitivity of our simulation procedure, we wanted to test if we could get good correlation between the computed binding energies for the methionine analogs with experimental binding energies. We tested five methionine analogs of which four get incorporated into proteins with reasonable efficiency and for which the experimental binding energies are available. Ccg, which is a *cis*-form of tcg (Figure IV-4a), has the lowest incorporation efficiency and hence, it was used as a negative control for which we hoped to get the worst binding energy for this analog. Binding energy calculations of the methionine analogs were carried out in the conformation that activates the protein, i. e. , by perturbation analysis.

In the case of 1QQT, the binding energies of the analogs are all in the top 50% but are interspersed with the non-cognate natural amino acids This indicates that in this conformation, MetRS lack the capability to discriminate efficiently (Figure IV-4d). However, in the co-crystal structure, there is a clear preference for binding the analogs. The analogs and methionine have a binding energy range of -63.4 to -79.1 kcal /mol (Figure IV-4b). The closest competitor from the non-cognate set of natural amino acids has a binding energy of -35.0 kcal/mol. In this conformation, we also find a good correlation between experimentally observed binding energies and computed binding energies (Figure IV-4c). As we had expected, ccg has the worst binding energy and gets incorporated with the lowest efficiency whereas myag has the best computed binding energy and has been tested to be the most best methionine analog. This information could be useful for initial computational scanning of the analogs before experimental testing. The binding energy of ccg in

Met/MetRS(FF)could be used as a cutoff for designing new analogs and the ones that rank above the cut off could be experimentally tested for binding.

We analyzed the binding modes of ccg and tcg to understand what in particular about the cis-form of the ligand renders it to be an unfavorable ligand. We analyzed the non-bond energies of these ligands with all the residues lining the binding pocket and have tabulated our findings as pairwise interactions in Table 1. Ccg has a VDW clash with Ala12, the terminal hydroxyl group of Tyr260 and His301. At the same time, the cis orientation of terminal methyl group does not make the same favorable interactions with Ala 256 and Pro 267 as tcg (Figure IV-5). Since Tyr 260 and His 301 have an important role in the binding process as indicated by experiments, mutating them to smaller residues may be deleterious. On the other hand, it would be interesting to explore the effect of Ala to Gly mutation at position 12 on the incorporation of *cis* forms of various analogs.

MetRS has been observed to be extremely promiscuous and is able to incorporate substrates that are up to 340000 folds poorer than methionine. This could be attributed to the conformational flexibility of the active site of metRS that has not been modeled in our simulation. The active site conformation could be different for different analogs. However, we have performed our perturbation studies only on the co-metRS bound to the natural substrate. The active site flexibility may be important in enabling MetRS to activate methionine analogs with varying side chain functionalities. One more consideration is that we are comparing our simulated binding energies to experimentally derived binding energies that are further derived from ATP-PP_i exchange studies. ATP binding could have other structural effects on

the enzyme that were not modeled in our simulations. However, it is interesting to note that we are able to get reasonably good correlation even with the limitations in the simulations. One can expect to gain more insights into the mechanism of this system with advancements in the simulation procedures.

Conclusions

We have studied the specificity of MetRS for methionine in the first two steps of the binding process. We have demonstrated that its specificity increases in the second binding step where the enzyme undergoes a significant conformational change. We speculate that methionine first anchors to residues Asp52 and His301 with its side chain and as the protein undergoes conformation change due to substrate binding (either the amino acid, ATP, or both), the cavity opens up and methionine flips into the cavity. Multi-step binding mechanisms where the ligand-protein complexes display "induced-fit" have been illustrated in other protein. This has been attributed to the presence of energy gradients, or funnels, near the binding sites - the binding process initiates from a higher energy conformer and terminates in lower energy conformation [41].

When the structure to be docked is taken from the crystallized co-complex, predicting the fitted association is relatively straightforward as indicated by the docking study using met/metRS. Our study with the apo-MetRS illustrates that although determining the final bound conformation starting with the "free", "unbound" state of the enzyme is extremely difficult, a refined search method can be applied to predict the correct binding region for the ligand. The predictions can be

used to indicate the important residues in the binding regions that can be further tested by mutations studies. Therefore, for those enzyme crystal structures that are not co-crystallized with their substrates a powerful docking protocol, like HierDOCK can prove to be very useful in recognizing the binding region, even in cases where the protein is very flexible. If the molecules are relatively rigid and have smooth binding funnels with single or few minima, there is a higher likelihood that the docked conformation of the ligand in the "free", "unbound" state is the correct bound conformation since the conformational diversity of the protein is limited [42]. But in the case of proteins that undergo significant conformation changes on associating with the ligand, it is unlikely that the predicted ligand plus protein complex would be the correct structure. In the case of a flexible protein, like MetRS, that has a larger conformational diversity, achieving a correct prediction bound conformation is complicated since the bound conformation could be very different from the free, unbound structure. However, the complex predicted with the apo enzyme should be regarded as an important "recognition mode" for the system, a key step in its multi-step binding process, since even at this stage of binding it could show some level of discrimination. In apo-MetRS, both docking and perturbation analysis indicate that in this conformation the enzyme is able to eliminate more than 60% of the natural amino acids. One could imagine that if the final bound complex after the change in conformation was the only filtering mechanism for an enzyme, each amino acid would first have to bind at this site, followed by the structural change in the enzyme and then get eliminated. Such a process would be both time consuming and energetically expensive for the enzyme. A first level of filter at the apo-enzyme

conformation certainly seems to be more efficient screening mechanism adopted by flexible enzymes. It would be interesting to see how the procedure for binding site search performs in other apo-enzyme systems. We have already tested it for the predicting the binding site of phe in *thermus thermophillus* PheRS by scanning the entire apo-crystal structure of PheRS and have been able to find the correct binding site (unpublished results).

Binding site dynamics in enzyme brings in the question of enzyme specificity. An interesting observation about protein plasticity is that proteins displaying higher selectivity are also more rigid while those that more flexible can bind to a large number of substrates. Considering the conformational flexibility in the MetRS, as indicated by the substantial structural change in the co-crystal, it is not surprising that it is one of the more permissive aminoacyl tRNA synthetases.

References

1. Ibba, M. and D. Soll, Aminoacyl-tRNA synthesis. Annu Rev Biochem, 2000. 69: p. 617-50.

2. Freist, W. Accuracy of protein biosynthesis: quasi-species nature of proteins and possibility of error catastrophes. J Theor Biol, 1998. 193(1): p. 19-38.

3. Freist, W., H. Sternbach, and F. Cramer, Isoleucyl-tRNA synthetase from baker's yeast and from Escherichia coli MRE 600. Discrimination of 20 amino acids in aminoacylation of tRNA(Ile)-C-C-A(3'NH2). Eur J Biochem, 1987. 169(1): p. 33-9.

4. Freist, W., H. Sternbach, and F. Cramer, Isoleucyl-tRNA synthetase from baker's yeast and from Escherichia coli MRE 600. Discrimination of 20 amino acids in aminoacylation of tRNA(Ile)-C-C-A. Eur J Biochem, 1988. 173(1): p. 27-34.

5. Freist, W., H. Sternbach, and F. Cramer, Isoleucyl-tRNA synthetase from Escherichia coli MRE 600. Different pathways of the aminoacylation reaction depending on presence of pyrophosphatase, order of substrate addition in the pyrophosphate exchange, and substrate specificity with regard to ATP analogs. Eur J Biochem, 1982. 128(2-3): p. 315-29.

6. Freist, W. and H. Sternbach, Tyrosyl-tRNA synthetase from baker's yeast. Order of substrate addition, discrimination of 20 amino acids in aminoacylation of tRNATyr- C-C-A and tRNATyr-C-C-A(3'NH2). Eur J Biochem, 1988. 177(2): p. 425-33.

7. Deming, T.J., Fournier, M. J., Mason, T. L., & Tirrell, D. A., Biosynthetic Incorporation and Chemical Modification of Alkene Functionality in Genetically Engineered Polymers. J. Macromol. Sci. Pure Appl. Chem, 1997. A34: p. 2143-2150.

8. van Hest, J.C. and D.A. Tirrell, Efficient introduction of alkene functionality into proteins in vivo. FEBS Lett, 1998. 428(1-2): p. 68-70.

9. van Hest, J.C., Kiick, K. L., Tirrell D. A., Efficient Incorporation of Unsaturated Methionine Analogues into Proteins in Vivo. J Am Chem Soc, 2000. 122: p. 1282-1288.

10. Kothakota S, M.T., Tirrell DA, Fournier MJ, Biosynthesis Of A Periodic Protein Containing 3-Thienylalanine - A Step Toward Genetically-Engineered Conducting Polymers. Journal Of The American Chemical Society, 1995. 117: p. 536-537.

11. Dougherty MJ, K.S., Mason TL, Tirrell DA, Fournier MJ, Synthesis Of A Genetically Engineered Repetitive Polypeptide Containing Periodic Selenomethionine Residues. Macromolecules, 1993. 26: p. 1779-1781.

12. Budisa, N., Steipe, B., Demange, P. Eckerskorn, C., Kellermann, J., & Huber, R., High-Level Biosynthetic Substitution of Methionine in Proteins by its Analogs 2-Aminohexanoic Acid, Selenomethionine, Telluromethionine, and Ethionine in E. coli. Eur. J. Biochem, 1995. 230: p. 788-796.

13. Duewel, H., Daub, E., Robinson, V., & Honek, J. F., Incorporation of Trifluoromethionine into a Phage Lysozyme: Implications and a New Marker for Use in Protein 19F NMR. Biochemistry, 1997. 36(3404-3416).

14. Zhang, D., et al., Structure-based design of mutant Methanococcus jannaschii tyrosyl-tRNA synthetase for incorporation of O-methyl-L-tyrosine. Proc Natl Acad Sci U S A, 2002. 99(10): p. 6579-84.

15. Datta, D., et al., A Designed Phenylalanyl-tRNA Synthetase Variant Allows Efficient in Vivo Incorporation of Aryl Ketone Functionality into Proteins. J Am Chem Soc, 2002. 124(20): p. 5652-3.

16. Liu, D.R., et al., Engineering a tRNA and aminoacyl-tRNA synthetase for the site-specific incorporation of unnatural amino acids into proteins in vivo. Proc Natl Acad Sci U S A, 1997. 94(19): p. 10092-7.

17. Floriano, W.B., Vaidehi, N., Goddard, W. A., Singer, M. S., Shepherd, G. M., Molecular mechanisms underlying differential odor responses of a mouse olfactory receptor. Proc Natl Acad Sci U S A, 2000. 97(20): p. 10712-6.

18. Kiick, K.L., R. Weberskirch, and D.A. Tirrell, Identification of an expanded set of translationally active methionine analogues in Escherichia coli. FEBS Lett, 2001. 502(1-2): p. 25-30.

19. Tannor, D.J., Marten B, Murphy R, Friesner RA, Sitkoff D, Nicholls A, Ringnalda M, Goddard WA, Honig B, Accurate First Principles Calculation Of Molecular Charge-Distributions And Solvation Energies From Ab-Initio Quantum-Mechanics And Continuum Dielectric Theory. Journal Of The American Chemical Society, 1994. 116(26): p. 11875-11882.

20. Mechulam, Y., et al., Crystal structure of Escherichia coli methionyl-tRNA synthetase highlights species-specific features. J Mol Biol, 1999. 294(5): p. 1287-97.

21. Mayo SL, O.B., Goddard WA, Dreiding - A Generic Force-Field For Molecular Simulations. Journal Of Physical Chemistry, 1990. 94(26): p. 8897-8909.

22. Ghosh, A., Rapp, C. S., Friesner R. A., Generalized born model based on a surface integral formulation. Journal Of Physical Chemistry **B**, 1998. 102(52): p. 10983-10990.

23. Serre, L., et al., How Methionyl-tRNA Synthetase Creates its Amino Acid Recognition Pocket upon l-Methionine Binding. J Mol Biol, 2001. 306(4): p. 863-876.

24. Vaidehi, N., et al., Prediction of structure and function of G protein-coupled receptors. Proc Natl Acad Sci U S A, 2002. 99(20): p. 12622-7.

25. Wang, P., Vaidehi, N., Tirrell, D. A,, and Goddard III, W. A., Virtual Screening for Binding of Phenylalanine Analogs to Phenylalanyl-tRNA Synthetase. J Am Chem Soc, 2002. (in press).

26. Kekenes-Huskey, P., Vaidehi, N., Floriano, W. B., Goddard III, W. A., Submitted. 2002.

27. Halperin, I., Buyong, M., Wolfson, H., Nussinov, R., Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins : Structure, Function and Genetics, 2002. 47: p. 409-443.

28. Wang, J.M., Morin, P., Wang, W., Kollman, P. A, Use of MM-PBSA in Reproducing the Binding Free Energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. J. Am. Chem. Soc., 2001. 123: p. 5221-5230.

29. Kollman, P.A., Free energy calculations: Applications to chemical and biochemical phenomena. Chem. Rev., 1993. 93: p. 2395-2417.

30. Ewing, J.A., Kuntz, I.D., Critical evaluation of search algorithms for automated molecular docking and database screening. J Comp Chem, 1997. 18: p. 1175-1189.

31. Ewing, T.J., Makino, S., Skillman, A.G., Kuntz, I.D., DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput Aided Mol Des, 2001. 15: p. 411-28.

32. Lim, K.T., Brunett, S., Iotov, M., McClurg, R. B., Vaidehi, N., Dasgupta, S., Taylor, S., Goddard III, W. A., Molecular dynamics for very large systems on massively parallel computers: The MPSim program. Journal of Computational chemistry, 1997. 18: p. 501-521.

33. Rocchia, W., Alexov, E., Honig, B., Extending the Applicability of the Nonlinear Poisson-Boltzmann Equation: Multiple Dielectric Constants and Multivalent Ions. J. Phys. Chem. B, 2001. 105: p. 6507-6514.

34. Hendsch, Z.S., Tidor, B., Electrostatic interactions in the GCN4 leucine zipper: substantial contributions arise from intramolecular interactions enhanced on binding. Protein Science, 1999. 8: p. 1381-1392.

35. Connolly, M.L., Analytical Molecular-Surface Calculation. Journal Of Applied Crystallography, 1983. 16: p. 548-558.

36. Stahl, M. and H.J. Bohm, Development of filter functions for protein-ligand docking. J Mol Graph Model, 1998. 16(3): p. 121-32.

37. Bower, M.J., F.E. Cohen, and R.L. Dunbrack, Jr., Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology-modeling tool. J Mol Biol, 1997. 267(5): p. 1268-82.

38. Fourmy, D., et al., Identification of residues involved in the binding of methionine by Escherichia coli methionyl-tRNA synthetase. FEBS Lett, 1991. 292(1-2): p. 259-63.

39. Ghosh, G., et al., Activation of methionine by *Escherichia co*li methionyl-tRNA synthetase. Biochemistry, 1991. 30(40): p. 9569-75.

40. Kim, H.Y., et al., The relationship between synthetic and editing functions of the active site of an aminoacyl-tRNA synthetase. Proc Natl Acad Sci U S A, 1993. 90(24): p. 11553-7.

41. Zhang, C., Chen, J., DeLisi, C., Protein-protein recognition: Exploring the energy funnel near the binding sites. **Proteins:** Structure, Function and Genetics, 1999. 34: p. 255-267.

42. Tsai, C.J., Kumar, S., Ma, B., Nussinov, R., Folding funnels, binding funnels, and protein function. Protein Sci, 1999. 8: p. 1181-1190.

Table IV-1.Energy analysis for CCG and TCG analogs

	CCG			TCG		
Residue	VDW	Coulomb	H-bond	VDW	Coulomb	H-bond
ASP 52	0.438	-21.47	-10.246	0.364	-21.427	-9.839
LEU 13	-1.255	-6.045	-9.898	-1.762	-6.142	-10.19
TYR 15	-2.173	-7.773	-0.677	-3.282	-6.162	-0.122
TRP 253	-3.779	-1.934	0.000	-3.779	-1.879	0.000
ILE 297	-2.097	-0.454	0.000	-0.967	-1.585	0.000
PRO 14	-0.973	-1.467	0.000	-1.931	-0.526	0.000
HIS 301	-0.102	-1.189	0.000	-1.216	-1.074	0.000
PRO 257	-0.670	-0.122	0.000	-1.486	-0.080	0.000
ILE 293	-0.273	-0.145	0.000	-1.110	-0.232	0.000
TYR 260	-0.227	-0.116	0.000	-1.780	0.706	0.000
ALA 256	-0.941	0.616	0.000	-1.414	0.601	0.000
VAL 252	-0.233	-0.025	0.000	-0.227	-0.047	0.000
ALA 12	-0.084	0.673	0.000	-0.144	0.081	0.000

Figure IV-1: Sphere filled volume of MetRS representing the possible binding sites in the enzyme.The search volume was divided into 14 regions as indicated by the cubic boxes.The binding site was found in the box colored in red.

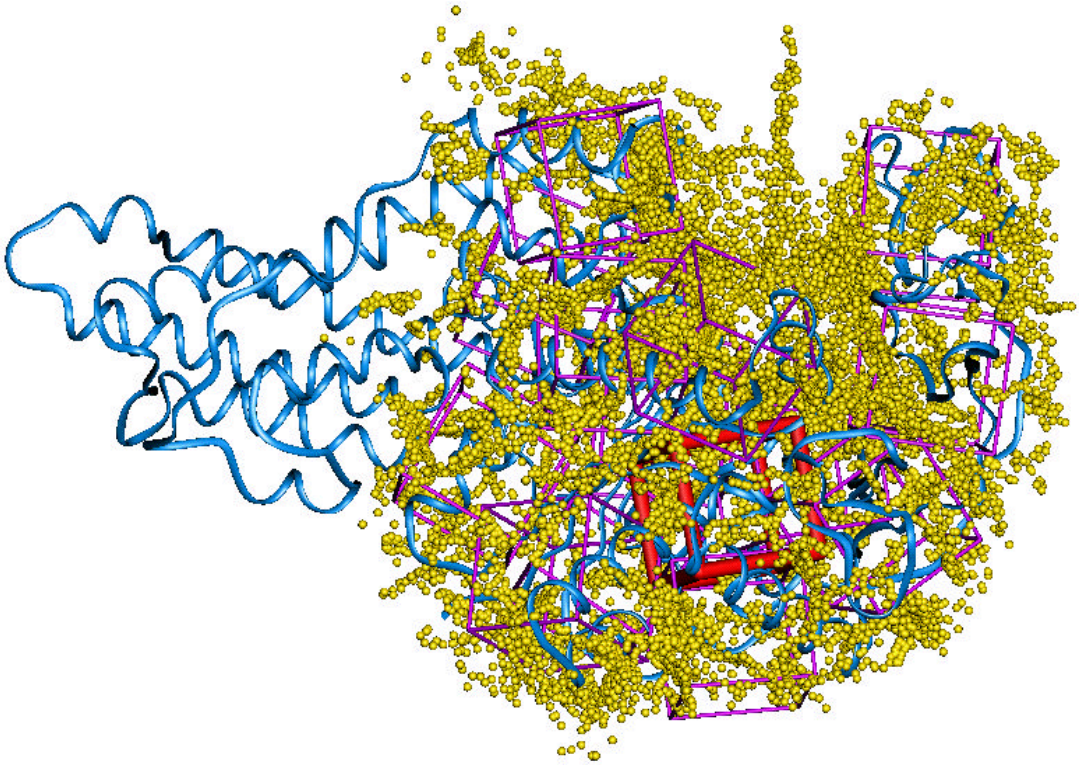


Figure IV-2: Binding energies of all 20 amino acids in the methionine binding site in Met/MetRS(FF) and apo-MetRS(FF).(a) shows binding energies of the 20 aminoacids when docked in the predicted methionine binding site in apo-MetRS(FF) and (b) shows the binding energies generated from perturbation analysis at the same site.(c) reports the binding energies generated from docking all 20 amino acids in the crystallographic methionine binding site in Met/MetRS(FF) and (d) indicates binding energies calculated from perturbation analysis at the same site.

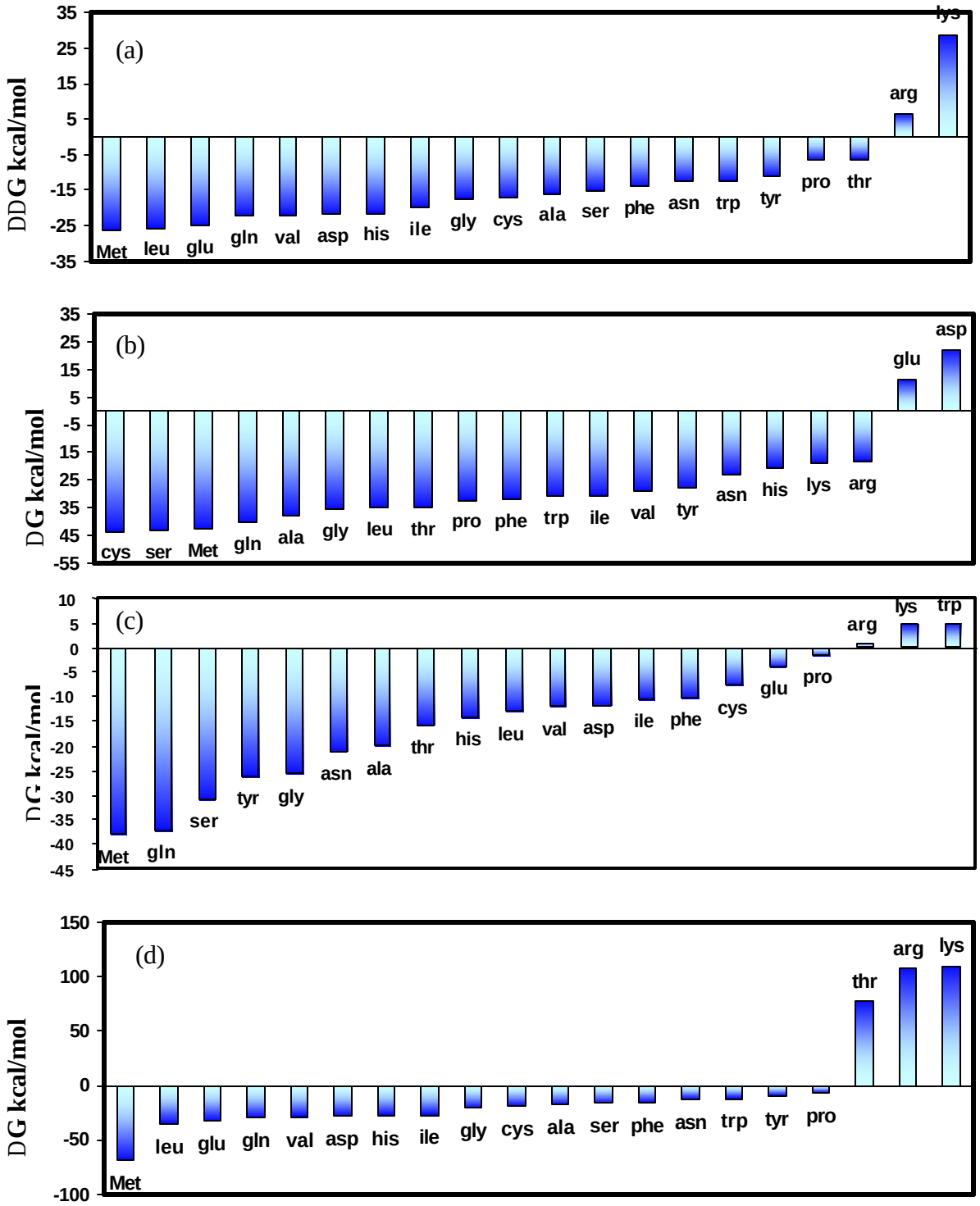

(a)
DDG kcal/mol
35
25
15
5
-5
-15
-25
-35
Met leu glu gln val asp his ile gly cys ala ser phe asn trp tyr pro thr arg lys
(b)
DG kcal/mol
35
25
15
5
-5
15
25
35
45
-55
cys ser Met gln ala gly leu thr pro phe trp ile val tyr asn his lys arg glu asp
(c)
DG kcal/mol
10
5
0
-5
-10
-15
-20
-25
-30
-35
-40
-45
Met gln ser tyr gly asn ala thr his leu val asp ile phe cys glu pro arg lys trp
(d)
DG kcal/mol
150
100
50
0
-50
-100
Met leu glu gln val asp his ile gly cys ala ser phe asn trp tyr pro thr arg lys

Figure IV-3(a) Binding site of methionine in apo-MetRS(FF) and Met/MetRS(FF).Amino acids lining the binding pocket are shown in purple for apo-MetRS(FF) and in green for Met/MetRS(FF).Methionine orientation from perturbation analysis in Met/MetRS(FF) in shown in red and its conformation from docking in apo-MetRS(FF) is colored blue.Residues closest to methionine that undergo the largest conformation changes are labeled.

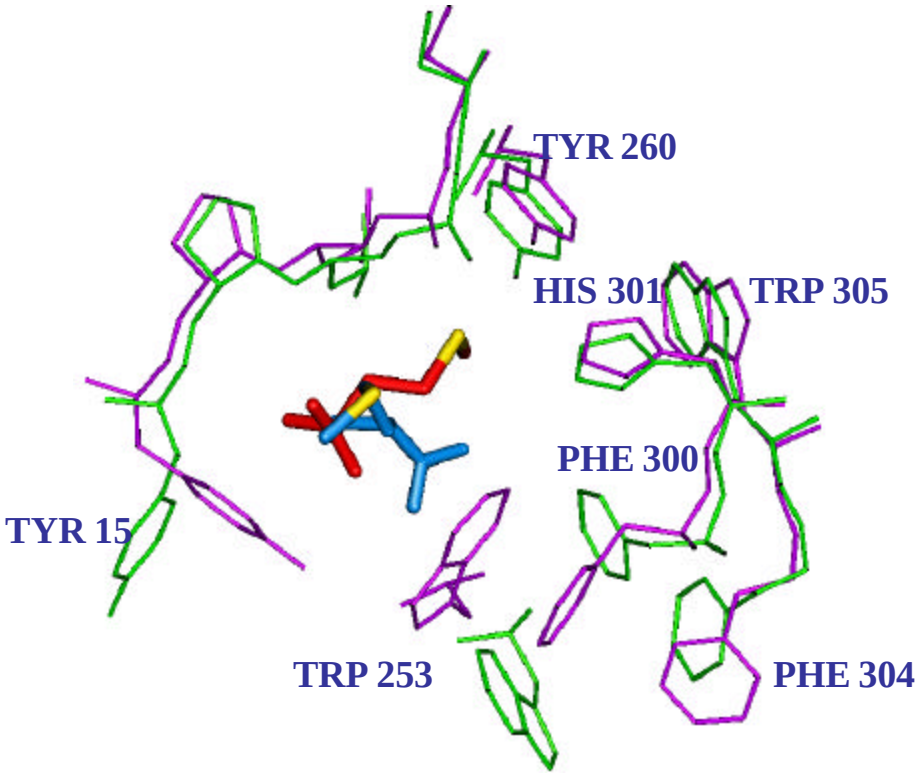

TYR 260
HIS 301
TRP 305
PHE 300
TYR 15
TRP 253
PHE 304

Figure IV-3(b) The S^{δ} of methionine makes two hydrogen bonds – one with the terminal oxygen of Tyr260 and the other with the backbone amide of Leu13 in the docked conformation in Met/MetRS(FF).The crystal structure orientation on methionine is shown in blue.The CRMS between the two conformations is 0.55 Å.

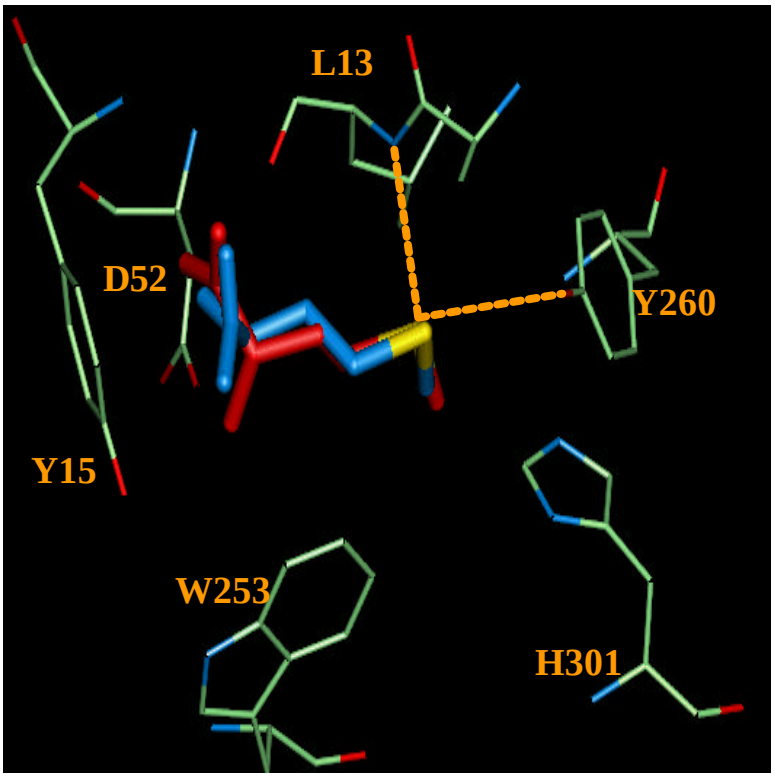

L13
D52
Y260
Y15
W253
H301

Figure IV-3(c) The predicted binding site for methionine in apo-MetRS(FF).The conserved residues within 4 Å are labeled in gold and the conserved replacements are labeled in aqua.

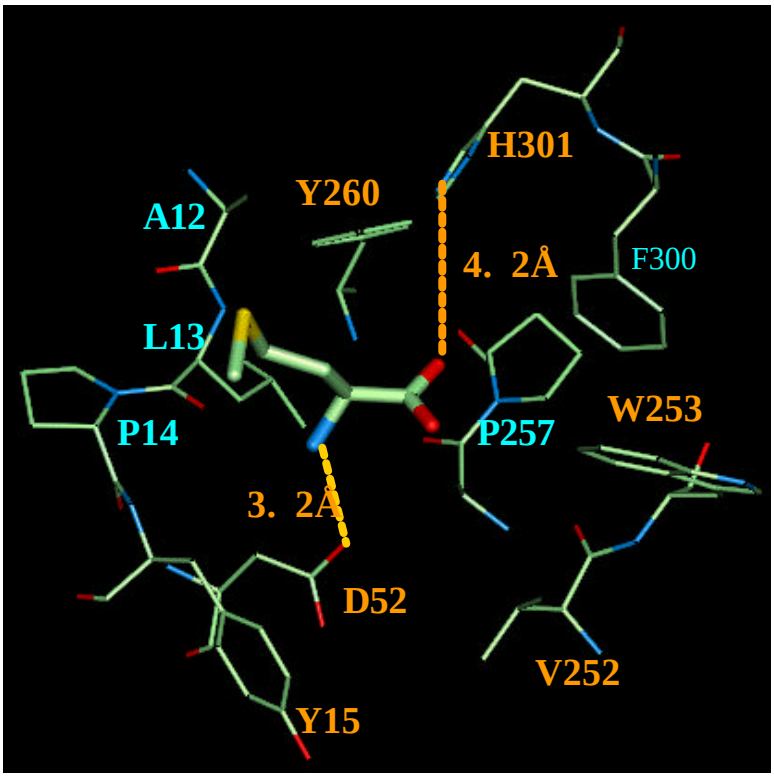
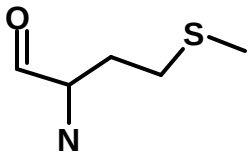
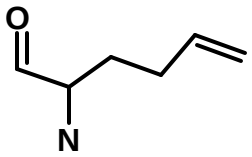

H301
Y260
A12
4. 2Å
F300
L13
W253
P14
P257
3. 2Å
D52
V252
Y15

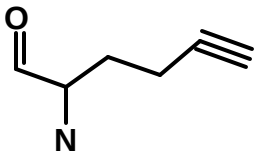
Figure IV-4: (a) Structures of methionine and its analogs used in this study.L-methionine (met), homoallylglycine (mhag), homopropargylglycine (myag), norleucine (nleu), *trans*-crotglycine (tcg) and *cis*-crotglycine (ccg).



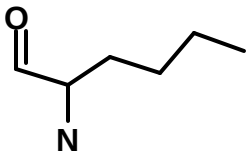
Met



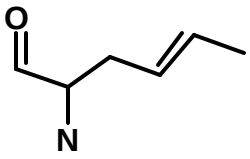
mhag



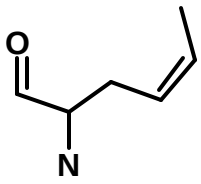
myag



nleu



tcg



ccg

Figure IV-4 (b) Binding energies of the analogs in the binding site of Met/MetRS(FF) calculated using perturbation method.

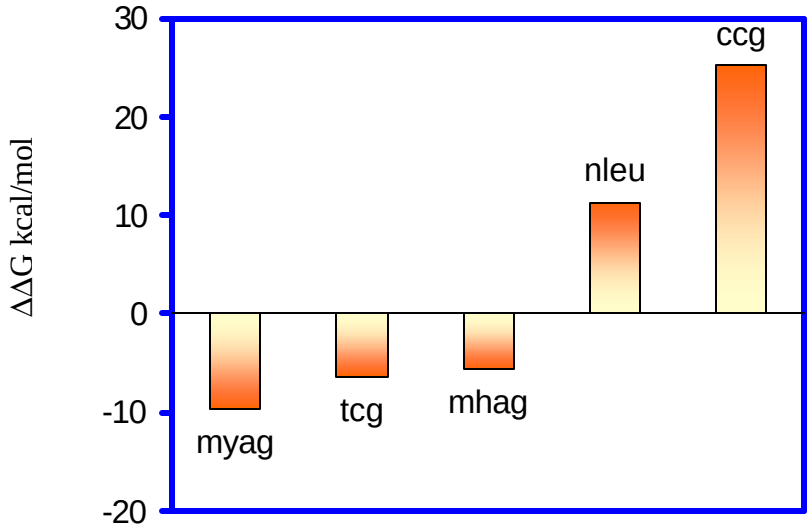

30
20
10
0
-10
-20
ΔΔG kcal/mol
myag
tcg
mhag
nleu
ccg

Figure IV-4(c) shows the correlation between the calculated binding energies and the experimentally observed ??G with respect to methionine.

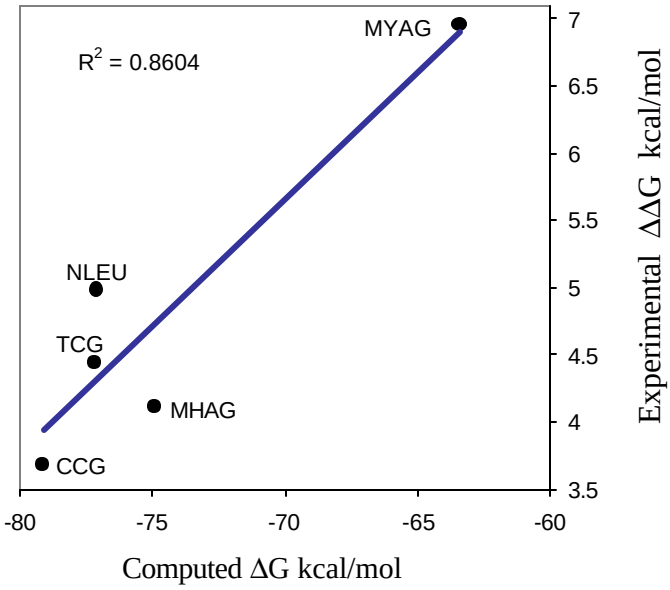

MYAG
R^2 = 0.8604
NLEU
TCG
MHAG
CCG
7
6.5
6
5.5
5
4.5
4
3.5
-80
-75
-70
-65
-60
Experimental ΔΔG kcal/mol
Computed ΔG kcal/mol

Figure IV-4(d) Binding energies of analogs along with the natural amino acids in the binding site of apo-MetRS (FF).Analogs are represented in shades of pink.

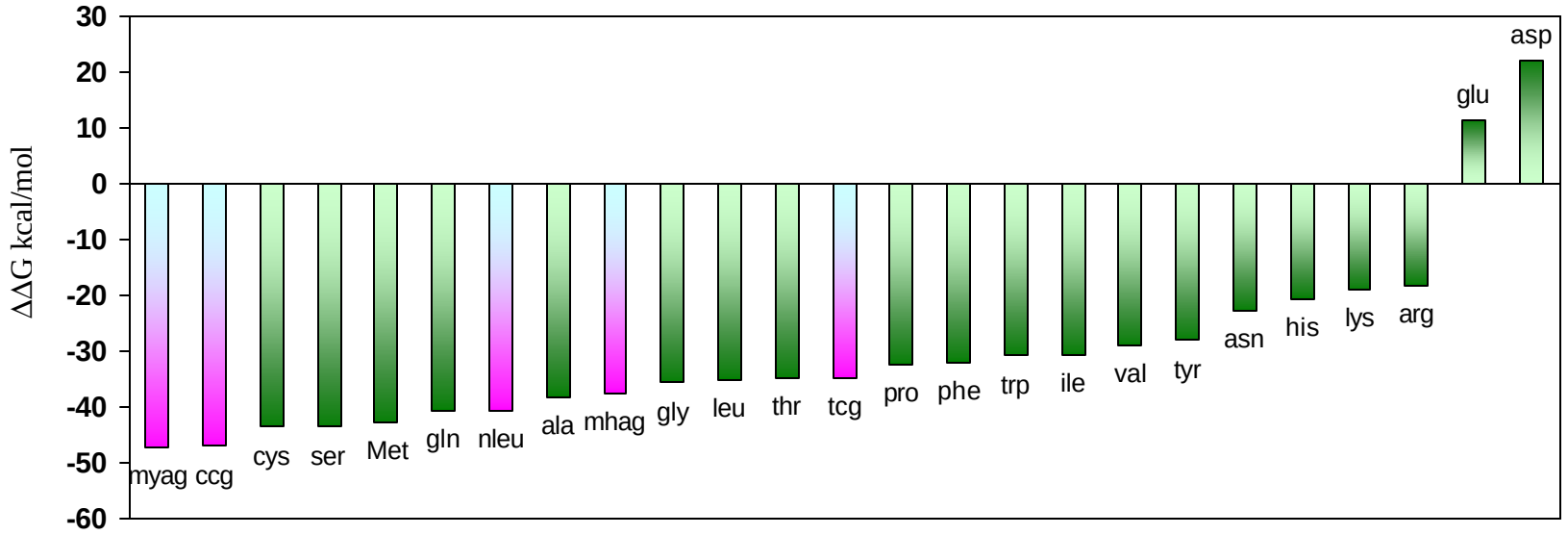
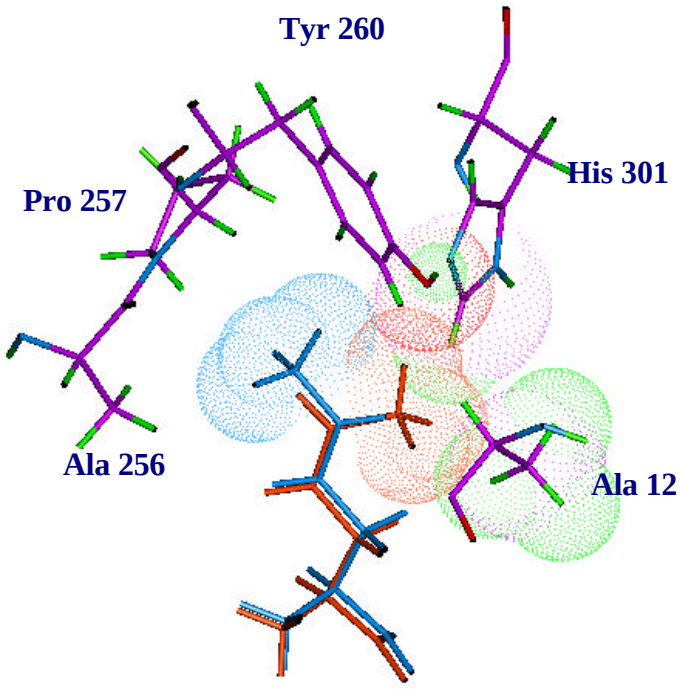

ΔΔG kcal/mol
30
20
10
0
-10
-20
-30
-40
-50
-60
myag
ccg
cys
ser
Met
gln
nleu
ala
mhag
gly
leu
thr
tcg
pro
phe
trp
ile
val
tyr
asn
his
lys
arg
glu
asp

Figure IV-5: The binding modes of tcg and ccg, shown in the binding pocket of Met/MetRS(FF), were predicted by perturbation analysis at this site.Ccg (orange) has VDW clashes with Ala12, His301 and Tyr 260 and at the same time, *cis* orientation of the terminal methyl group created a void near Ala 256 and Pro 257.Tcg (Pink) is shown to fill that void and also avoid the unfavorable VDW interactions.


Tyr 260
His 301
Pro 257
Ala 256
Ala 12

Section II

Protein Design

Chapter 5

Introduction to Protein design

One of the goals in computational protein design is to develop algorithms for predicting amino acid sequences that would adopt a specified three-dimensional structure. It is an extremely progressive area of research and has been successfully applied to engineer proteins with improved stabilities and activities [refer to [1]for a recent review]. Developments in this field are beginning to have an impact on biotechnology and further advancements in improving design strategies are expected to expand the range of applicability of computational protein design to larger and more complex biological systems. Besides providing as a suitable tool for designing proteins for medicinal and industrial purposes, the development of protein design tools should also confer a deeper insight into the principles that underlie protein sequence-structure relationship.

There are three primary aspects of computational protein design that, although quite distinct from one another, are significantly interdependent. The progress achieved by several groups in this field shows that improvements in design techniques have been made possible through refinements in these areas. The first aspect concerns the energy expression used to assess and score the relative fitness of different amino acid sequences with respect to the desired protein fold. The second area deals with the way in which the protein design problem, the model, is represented. The model provides a framework for describing the target fold and its flexibility, the amino acids allowed for design positions, and the rotamer library used to represent the possible side chain conformations. The third area of enhancement is the search strategy used to scan the enormous combinatorial complexity of possible sequences and selecting those that are optimal for a given fold.

Since the inception of computational protein design, elements of a suitable energy

expression or force field to rank the desirability of an amino acid sequence for a particular backbone structure have been suggested and evaluated. Alterations and additions to the energy terms to improve the correlation between computed and experimentally observed properties are usually achieved by iterating between theory and experiment [2]. The force-field terms describing the non-bond interactions are usually explored for improvement, while the bonded energies are taken from commonly used molecular mechanics force fields [3]. Since the rotamers derived from protein databases generally have good internal energies, and for most design schemes, rigid backbone structures are used, the usefulness of improving "bonded" energies has not been rigorously demonstrated in protein design.

For protein cores, a force field that models packing specificity is usually sufficient to design a well-folded protein [2, 4]. For designing protein surfaces however, energy terms that properly balance non-bonded polar interactions also need to be included in the potential function. Chapter 7 describes an approach used to derive a potential function for designing ß-sheet surfaces. This method was used to create a plastocyanin variant with enhanced thermostability. The derived potential placed more importance on electrostatic interactions that led to the selection of charged residues on protein surfaces. An important next step in understanding β-sheet stability was to define the role of side-chain ionic interactions. Chapter 8 outlines a study that evaluates the interaction energy of a three-residue ionic network constructed on the ß-sheet surface of protein G.

Except for a few notable exceptions, the models used to represent protein design problems in most cases, do not allow for backbone flexibility; side-chain flexibility is incorporated by selecting amino acid rotamers from a library of discrete conformations

[5-9]. If necessary, one may use rotamer libraries with different levels of resolution. However, because the size of the design problem grows exponentially with the increase in the number of rotamers, using a finely descretized rotamer library is often unfeasible. The rigidity of such a framework ignores the possibility of backbone shifts to accommodate mutations and an incomprehensive rotamer library could lead to the selection of incorrect side chain conformations. These limitations of the design procedure are highlighted in Appendix 1. This section reports a study on the redesign of the core residues in T-4 lysozyme where a significant shift in the protein backbone is observed in the crystal structure of a designed variant.

The third area of thrust in computational protein design deals with refinements in the search strategy. Searches for the optimal sequence for a target protein fold are achieved using various deterministic and stochastic combinatorial optimization algorithms. However, as structural targets get larger, it has become necessary to find more powerful methods to address the increased combinatorial complexities. Efficient algorithms that take into account the limitations of computing power and computational time are being developed and applied to numerous design problems.

The ultimate goal of automated protein design is not only to be able to generate amino acid sequences that are compatible with the given backbones, but also to ensure that the selected sequences are able to perform specific functions. Intermolecular interactions lie at the core of protein function in a wide range of fundamental biochemical processes. Proteins function through their interactions with ligands, other proteins, or surfaces and these interactions are controlled by a complex array of intermolecular forces. In many instances, binding to ligands induces structural changes that allow, for

example, signal transduction across large distances. Designing ligand binding sites and engineering ligand-induced conformation changes in proteins are very important applications of computational protein design.

Chapters 9 and 10 delve into the application of computational techniques in the area of ligand-protein interactions. Chapter 9 describes the applicability of protein design to alter the specificity of a known binding site to enable it to bind to alternate ligands. An aminoacyl tRNA synthetase with altered ligand specificity was designed and was subsequently shown to be capable of incorporating an artificial amino acid *in vivo.* Chapter 10 explores the possibility of using computational methods to manipulate ligand-induced conformational change. The methodology in this study combines computational protein design with techniques from mean-field theory to generate sequences that undergo substantial conformational changes upon ligand binding. The design approach and the results in this study will provide important insights and information that will aid future design efforts in this direction.

References

1. Kraemer-Pecore, C.M., Wollacott, A. M., Desjarlais, J. R., Computational protein design. Curr Opin Chem Biol., 2001. **5**: p. 690-5.
2. Dahiyat, B.I. and S.L. Mayo, Protein design automation. Protein Sci, 1996. **5**(5): p. 895-903.
3. Gordon, D.B., S.A. Marshall, and S.L. Mayo, Energy functions for protein design. Curr Opin Struct Biol, 1999. **9**(4): p. 509-13.
4. Dahiyat, B.I. and S.L. Mayo, Probing the role of packing specificity in protein design. Proc Natl Acad Sci U S A, 1997. **94**(19): p. 10172-7.
5. Harbury, P.B., Tidor, B., Kim P. S., Repacking protein cores with backbone freedom: structure prediction for coiled coils. Proc Natl Acad Sci U S A., 1995. **371**: p. 8408-12.
6. Harbury, P.B., Plecs, J. J., Tidor, B., Alber, T., Kim P. S., High-resolution protein design with backbone freedom. science, 1998. **92**: p. 8408-12.
7. Su, A. and S.L. Mayo, Coupling backbone flexibility and amino acid sequence selection in protein design. Protein Sci, 1997. **6**(8): p. 1701-7.
8. Dunbrack, R.L. and M. Karplus, Backbone-dependent rotamer library for proteins. Application to side- chain prediction. J Mol Biol, 1993. **230**(2): p. 543-74.
9. Dunbrack, R.L. and M. Karplus, Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol, 1994. **1**(5): p. 334-40.

Chapter 6

A Designed Apoplastocyanin Variant that Shows Reversible Folding

Adapted from Datta D., Mayo S. L. ,Biochem Biophys Res Commun. 2002 Aug 30;296(4):988

Abstract

Plastocyanin, like many other metalloproteins, does not undergo reversible folding, which is thought to be due to an irreversible conformational change in the copper-binding site. Moreover, apoplastocyanin's ability to adopt native tertiary structure is highly salt-dependent, and even in high salt it has an irreversible thermal denaturation. Here we report a designed apoplastocyanin variant, PCV, that is well-folded and has reversible folding in both high and low salt conditions. This variant provides a tractable model for understanding and designing protein ß-sheets.

Introduction

Plastocyanins are small (97-104 amino acids) copper-binding proteins that function in photosynthesis to catalyze electron transfer from cytochrome-*f* of the membrane-associated cytochrome-*b6f* complex to P700+ in photosystem-I. Like all plastocyanins, the poplar plastocyanin used in this study is a small Greek key ß-sandwich protein (10,500 Da) with eight ß-strands, a single turn of a-helix, and a copper site coordinated by two histidines, a cysteine, and a methionine in a distorted tetrahedral geometry (1, 2) (Fig 7.1).

Plastocyanin has been used to assess the role of ß-turns in dictating protein structure (3) and to elucidate folding mechanisms in ß-sheet proteins (4). It has also been a useful model for studying electron transport in Type 1 copper proteins and in plants (5). However, plastocyanin has a major drawback as a model for understanding protein thermodynamics. The thermal denaturation of plastocyanin occurs at 61 °C and is irreversible (6). It has been suggested that the overall protein denaturation occurs after the disruption of the copper-binding site and that during the thermal transition the geometry of this site changes from tetrahedral, in the native form, to square planar, in the denatured state (7).

The apo form of plastocyanin in low salt conditions has an altered circular dichroism (CD) spectrum compared to the holo protein, suggesting that the apo protein has a significantly reduced ß-sheet content and is unstable in solution (8). Apoplastocyanin (apoPC) adopts the folded conformation only in the presence of high salt concentrations (1 M NaCl or 0.5 M Na_2SO_4), and under these conditions, its CD spectrum is almost identical to that of holo PC (4). In order to address both the

irreversibility of folding and the salt dependency, we used the ORBIT (Optimization of Rotamers By Iterative Techniques) protein design software to redesign the copper-binding site of poplar apoPC. The design resulted in a variant in which two metal coordinating residues, Cys84 and His37, were replaced with an alanine and a valine, respectively. The designed mutant, PCV, adopts the folded conformation in both high and low salt concentrations and has a completely reversible denaturation under high salt conditions.

Results and Discussion

In the process of redesigning the copper-binding site, calculations were run using different van der Waals scale factors. A scale factor of 0.9 has typically been used for protein core design (9). In order to replace the metal binding site in plastocyanin, van der Waals scale factors smaller than 0.9 were explored (10). PCV was generated using a scale factor of 0.85 where His37 and Cys84 are mutated to a valine and an alanine, respectively.

PCV and apoPC were compared for their thermal stability and reversibility of folding in high and low salt conditions. PCV appears folded in both high and low salt concentrations while apoPC lacks the native conformation in low salt conditions. ApoPC can only adopt its folded conformation in high salt concentrations or when reconstituted with copper to yield holoplastocyanin. PCV has a completely reversible two-state thermal transition in high salt and is almost completely reversible in low salt conditions (Fig VI-2a). Also, thermal denaturation results show that PCV is marginally stabilized compared to apoPC (Fig VI-2b).

PCV is a salt-independent, metal-free variant of plastocyanin that expresses well and is easy to purify. It should therefore be an excellent model system for protein studies. We have already found it to be useful in understanding and designing ß-sheet surfaces (11).

Experimental Methods

Computational design: Simulations were performed using coordinates from the X-ray structure of apoplastocyanin (PDB code: 2pcy). The metal binding site residues (Fig VI-1) were classified as core and boundary using the RESCLASS residue classification program as described previously (9). Cys84, His37 and Met92 were classified as core residues and His87 as a boundary residue. Eight hydrophobic amino acid types (Ala, Val, Leu, Ile, Phe, Tyr, Met, and Trp) were considered at the three core positions. All other residues as well as the backbone were held fixed. To maximize core packing, the radius scale factor for van der Waal's interactions was varied from 0.9 to 0.7 in steps of 0.5. Computational details, potential functions and parameters for van der Waals interactions, solvation and hydrogen bonding are described in our previous work (12).

Gene synthesis, mutagenesis and protein purification: The wild type (PC) plastocyanin gene (13) was constructed using recursive PCR technique (14) and was cloned into the pET-11a expression vector (Novagen). PCV was obtained by site directed mutagenesis (15) on the constructed plastocyanin gene using inverse PCR. Protein expression was carried out in *E. coli* strain BL21(DE3). The cells were grown in 2? YT at 37 °C. Recombinant protein expression was induced by adding IPTG to cells at an OD_{600} of 0.75. Cells were then incubated at 30 °C for an additional five

hours. To isolate apoPC, cells were sonicated in the presence of 10 mM DTT (dithiothrietol) to prevent intermolecular disulfide bond formation. Purification was accomplished by reverse phase high performance liquid chromatography. Two peaks corresponding to a 99-residue form (lacking an N-terminal methionine) and a 100-residue form (including the N-terminal methionine) were observed. The 100-residue form was used for all analysis.

CD analysis: Circular Dichroism (CD) data were collected on an Aviv 62 DS spectrometer equipped with a thermoelectric cell holder using a 1 mm path length cell. Protein samples were at a concentration of 70 µM in 50 mM potassium phosphate buffer at pH 7.0 (low salt) and potassium phosphate buffer containing 0.5 M sodium sulfate (high salt). Wavelength scans were carried out from 200 nm to 260 nm using increments of 2 nm. Thermal melts were monitored at 210 nm and data were collected every 2 °C with an equilibration time of 2 min and an averaging time of 40 seconds. The melting temperatures were determined by evaluating the maximum of a d?/dT versus T plot.

References

1. Guss, J. M. and Freeman, H. C. (1983) Structure of oxidized poplar plastocyanin at 1.6 A resolution. *J Mol Biol* **169**, 521-63.

2. Guss J. M., Harrowell P. R., Murata M., Norris V. A., Freeman H. C. (1986) Crystal structure analyses of reduced (CuI) poplar plastocyanin at six pH values. *J Mol Biol* **192**, 361-87

3. Ybe, J. A. and Hecht, M. H. 1996. Sequence replacements in the central beta-turn of plastocyanin. *Protein Sci* **5**, 814-24.

4. Koide, S., Dyson, H. J., Wright, P. E. 1993. Characterization of a folding intermediate of apoplastocyanin trapped by proline isomerization. *Biochemistry* **32**, 12299-310.

5. Hope, A. B. (2000) Electron transfers amongst cytochrome f, plastocyanin and photosystem I: kinetics and mechanisms. *Biochim Biophys Acta* **1456**, 5-26.

6. Freire, E. and Biltonen, R. L. (1978) Statistical thermodynamics of thermal transitions in macromolecules I. Theory and application to homogenous systems. *Biopolymers* **17**, 463-479.

7. Milardi, M., Carmelo, R. L., Grasso, D. (1994) Extended theoretical analysis of irreversible protein thermal unfolding. *Biophys Chem* **52**, 183-189.

8. Li, H. H. and Merchant, S. (1995) Degradation of plastocyanin in copper-deficient Chlamydomonas reinhardtii. Evidence for a protease-susceptible conformation of the apoprotein and regulated proteolysis. *J Biol Chem* **270**, 23504-10.

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

10. Strop P. and Mayo, S. L. (1999) Rubredoxin Variant folds without Iron. *J Am. Chem. Soc*. **121**, 2341-2345.

11. Street, A. G., Datta, D., Mayo, S.L. (2000) Designing protein beta-sheet surfaces by Z-score optimization. *Phys Rev Lett* **84**, 5010-3.

12. Gordon, D. B., Marshall, S. A., Mayo, S. L. (1999) Energy functions for protein design . *Current Opinion in Structural Biology* **9**, 509-513

13. Ybe, J. A. and Hecht, M. H. (1994) Periplasmic fractionation of *Escherichia coli* yields recombinant plastocyanin despite the absence of a signal sequence." *Protein Expr Purif* **5**, 317-23

14. Casimiro, D. R., Toy-Palmer, A., Blake, R. C., Dyson, H. J. (1995) Gene synthesis, high-level expression, and mutagenesis of Thiobacillus ferrooxidans rusticyanin: His 85 is a ligand to the blue copper center. *Biochemistry* **34**, 6640-8.

15. 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,** 6545-51.

16. Kraulis, P. J. (1991) MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallography* **24**, 946-950

Figure VI-1: Ribbon diagram of the X-ray crystal structure of wild type poplar apoplastocyanin illustrating residues in the copper-binding site. This figure was created using MOLSCRIPT (16).

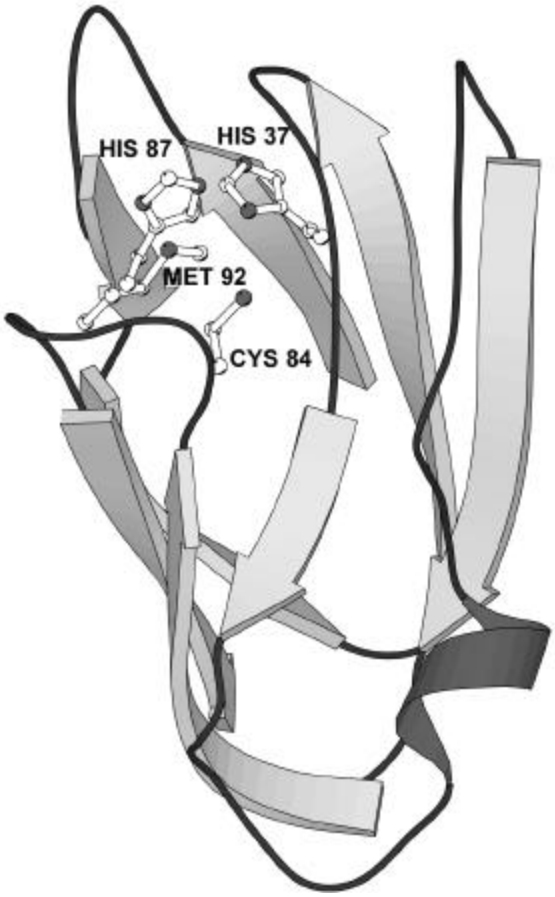
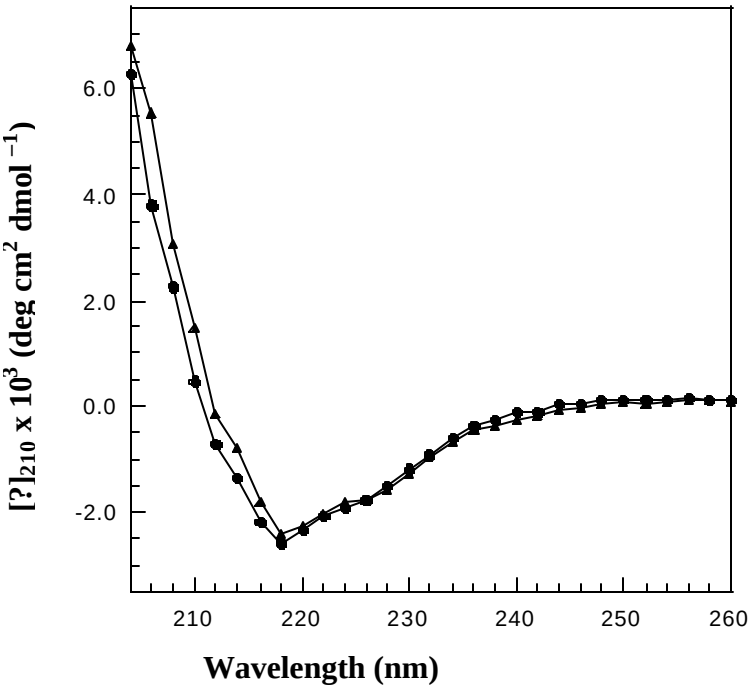
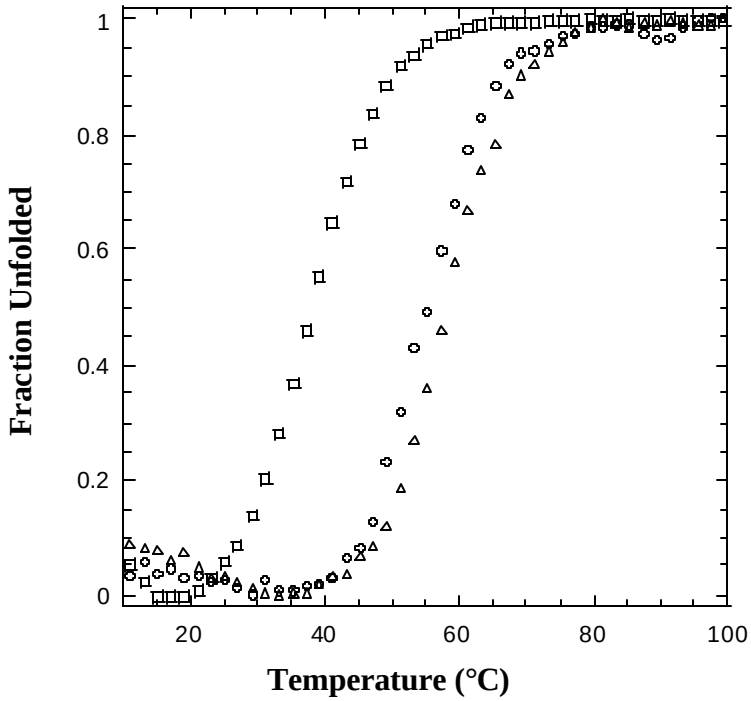

HIS 87
HIS 37
MET 92
CYS 84

Figure VI-2: (a) Wavelength scans of PCV in high salt conditions at 1 ºC before thermal denaturation (?) and after refolding (?). (b) Thermal unfolding curves monitored at 210 nm for apoPC in high salt (?), PCV in high salt (?), and PCV in low salt conditions (?).


6.0
4.0
2.0
0.0
-2.0
210
220
230
240
250
260
[?]210 x 10^3 (deg cm^2 dmol^-1)
Wavelength (nm)


1
0.8
0.6
0.4
0.2
0
20
40
60
80
100
Fraction Unfolded
Temperature (°C)

Chapter 7

Designing Protein b-Sheet Surfaces by *Z*-Score Optimization

Adapted from Street AG, Datta D, Gordon DB, Mayo SL. Phys Rev Lett 2000 May 22; 84(21): 5010-3

Abstract

Studies of lattice models of proteins have suggested that the appropriate energy expression for protein design may include non-thermodynamic terms in order to accommodate negative design concerns. A method has been developed to improve protein design in lattice model studies where enumeration of all possible sequences, and their ground state structures, is possible. The method maximizes a quantity known as the "Z-score," which compares the lowest energy sequence whose ground state structure is the target structure to an ensemble of random sequences. Here we show that, in certain circumstances, the technique can be applied to real proteins. The energy expression is then optimized using the assumption that the wildtype sequence is a low energy sequence (and its ground state is known to be the target structure). The new energy expression is used to design the β-sheet surfaces of two real proteins. We find experimentally that the resulting proteins are stable and well folded, and in one case, is even more thermostable than the wildtype.

Introduction

Much effort in the field of computational protein design is directed towards developing a potential function to rank the compatibility of amino acid rotamer sequences with a target structure (Gordon et al., 1999). In a "protein design cycle" (Dahiyat & Mayo, 1996; Street & Mayo, 1999), the potential function is developed by cycling between experiment and simulation, so that the computational potential ideally approaches nature's "true" potential. This technique has had some remarkable recent successes (Dahiyat & Mayo, 1997a; Malakauskas & Mayo, 1998).

The approach nevertheless rests on a controversial assumption. Rotamer sequences are threaded onto the target structure, and the sequence with the lowest energy (as determined by the potential function) is reported as the best sequence for that structure. It is conceivable, though, that in some circumstances this sequence will not adopt the desired ground-state structure. An extreme example is provided by imagining that the true potential function is one that only benefits hydrophobic contacts (and hydrophobic-polar and polar-polar interactions contribute zero energy) (Lau & Dill, 1989). Then, for any target structure, an all-hydrophobic sequence must be one of the best sequences. This sequence, of course, is not likely to fold specifically to the target structure — some polar residues ought to be included to characterize the surface of the molecule. Overcoming this problem involves introducing non-thermodynamic considerations to the design procedure, collectively known as "negative design" (Hellinga, 1997).

There are a number of schemes proposed to implement negative design, often specifically to solve the problem of the example in the last paragraph (or variations on it based on the Ising model of ferromagnetism). Perhaps the simplest is to use a fixed sequence

composition, that is, to hold the total number of hydrophobic and polar residues constant (Shakhnovich & Gutin, 1993). Even with this constraint, however, designed sequences are frequently found to fold to alternative structures of lower energy than the target structure (Shakhnovich, 1994; Yue et al., 1995). Alternatively, instead of minimizing the potential function, it is possible to choose a sequence to maximize the occupation probability of the target structure (Micheletti et al., 1998b; Seno et al., 1998).

Other approaches employed in lattice model studies involve adding non-thermodynamic terms to the potential function. One method is to introduce a "clamping potential" to force the molecule into the target structure, and then to minimize the difference between the clamping potential and the "true" potential (Kurosky & Deutsch, 1995; Deutsch & Kurosky, 1996). Another approach involves the addition of a penalty for exposing hydrophobic surface area (Sun et al., 1995).

Negative design is thus clearly important, at least in lattice model studies with simple potential functions and a limited set of amino acids (Crippen, 1996; Micheletti et al., 1998a). For real proteins and more physical potential functions, negative design can be necessary to guarantee the correct multimeric state of designed proteins (Harbury et al., 1993). A penalty for exposing hydrophobic surface area has also been shown to improve the designability of real proteins (Dahiyat & Mayo, 1997b; Malakauskas & Mayo, 1998).

We have taken yet another approach to determining the optimal potential function for protein design, in which we maximize the energy gap between a low energy sequence known to fold to the target structure, and the average energy of an ensemble of random sequences threaded onto a target structure (Chiu & Goldstein, 1998). In a cubic 3x3x3 lattice simulation, the desired "true" potential can be selected manually and the protein folding problem can be

solved. Thus a sequence *S*, whose ground state structure is the target structure, can be determined and its energy calculated. If the distribution of energies of the random sequences is assumed to be Gaussian, the success of the test potential for protein design is measured by the energy gap between the mean of the distribution and the energy of sequence *S*, normalized by the standard deviation of the distribution (Figure VII-1). This quantity is known as the *Z*-score of the sequence *S* on the target structure. The test potential is then adjusted to maximize the *Z*-score.

Chiu and Goldstein applied the method to a 3x3x3 lattice model, using statistically-derived pair potentials (Miyazawa & Jernigan, 1985) as the "true" potential. They found that the potential generated by maximizing the *Z*-score across many structures led to significantly better success at solving the protein design problem than the true potential. Here we show that the technique does not transfer readily to real proteins in their entirety. Nevertheless, we show that the technique can be applied to certain subsections of proteins. In particular we use it to design the β-sheet surfaces of the B1 immunoglobulin-binding domain of streptococcal protein G (GB1) and of a variant of poplar apoplastocyanin with the metal-binding site removed (PCV).

The *Z*-Score Applied to Real Proteins

One of the key assumptions of the lattice model method of Chiu and Goldstein (Chiu & Goldstein, 1998) is that the energies of random sequences threaded onto the target structure form a Gaussian distribution. It would be surprising if this assumption were to hold for real proteins. In particular, one would expect that placing random amino acid side chains in the core of a protein would typically lead to unresolvable steric clashes, especially since the modeled backbone of the target structure is held rigid. Indeed, Figure VII-2a shows the distribution of

potential energies of random sequences threaded onto the core of GB1. The distribution is clearly not Gaussian, with most sequences yielding enormous energies. A Gaussian distribution may be achievable by using a statistically-derived pair potential instead of an atomistic van der Waals potential, but designs using pair potentials have not yielded uniquely characterizable folded states (Isogai et al., 1999).

When only surface residues are considered, the situation is improved. For α-helix and β-sheet surface residues of GB1, the distribution of energies of random sequences is close to Gaussian, as shown in Figure VII-2b. Thus it appears that on the surface, even randomly selected amino acids are always able to find suitable rotamers that avoid severe steric interference. The *Z*-score analysis may therefore provide some insight into the appropriate potential function for α-helix and β-sheet surface design, provided one can find an appropriate sequence with which to calculate the *Z*-score. In lattice models, one knows the true potential function and can exhaustively search all conformations to solve the protein folding problem (Shakhnovich & Gutin, 1993). Hence the *Z*-score of a structure could be calculated using the lowest-energy sequence whose ground state is the target structure.

In contrast, in the lattice model study of Chiu and Goldstein (Chiu & Goldstein, 1998), the *Z*-score is actually calculated without knowledge of this lowest-energy sequence. One thousand 27-residue random amino acid sequences are constructed, which are found to correspond to 992 unique ground state structures. Eight sequences are discarded to yield a one-to-one correspondence between structures and sequences. The *Z*-score is calculated for each sequence in its ground state structure, using the 992 sequences to determine the energy distribution. The potential function is then modified to maximize an appropriately formed average

of the *Z*-scores. Thus, the reference sequence used to calculated the *Z*-score is not necessarily the lowest-energy sequence whose ground state structure is the target structure, but instead an arbitrary sequence whose ground state structure is the target structure. Nevertheless, the resulting potential function is significantly better for protein design than the "true" potential.

In our application of the theory to real proteins, we therefore expect that any arbitrary sequence known to fold to a target structure will suffice for calculating the *Z*-score of that structure. Given an experimentally determined structure, we can thus use the protein's wildtype sequence to calculate its *Z*-score. In essence, the method then chooses the potential function which locates the protein's wildtype sequence as far as possible down the tail of the distribution of energies.

Since a number of successful computational redesigns of a-helical surfaces have been reported (Dahiyat et al., 1997; Morgan, 2000), we chose to examine the *Z*-score technique on the β-sheet surface, where there have been few successful computational protein design efforts. Negative design issues are also expected to play a larger role in β-sheet design (Hecht, 1994). Rather than maximizing the *Z*-score of a large number of structures, as a first step we consider just one structure, so that the resulting potential function is optimized for protein design on that structure. This method should increase the possibility of the technique being successful for at least the one selected structure. The resulting potential function may then be applied to other proteins to test its generality, or a new potential function may be calculated by considering more protein structures. In particular, we chose to apply the technique to the eight β-sheet surface residues of GB1 which are not involved in stabilizing interactions with neighboring turns (Figure VII-3a), and to the seven β-sheet surface residues on one face of PCV (Figure VII-3c).

The computational potential function, E, included van der Waals interactions, E_{vdW} (Mayo et al., 1990; Dahiyat & Mayo, 1997b), electrostatics, E_{elec}, and a hydrogen bonding potential, E_{HB} (Dahiyat, et al., 1997), a bias for secondary structure propensity, E_{SS} (Dahiyat, et al., 1997), and solvation energies. The solvation energies were a benefit for burial of hydrophobic surface area, $A_{\mathrm{np}}^{\mathrm{buried}}$, a penalty for burial of polar surface area, $A_{\mathrm{polar}}^{\mathrm{buried}}$, and a penalty for exposure of hydrophobic surface area, $A_{\mathrm{np}}^{\mathrm{exposed}}$ (Street & Mayo, 1998), and a further penalty for polar hydrogen burial, E_{phb} (Dahiyat, et al., 1997).

$$E = vE_{\mathrm{vdW}} - s_{\mathrm{np}} A_{\mathrm{np}}^{\mathrm{buried}} + x_{\mathrm{np}} A_{\mathrm{np}}^{\mathrm{exposed}} + s_{\mathrm{p}} A_{\mathrm{polar}}^{\mathrm{buried}} + \frac{1}{e} E_{\mathrm{elec}} + DE_{\mathrm{HB}} + PE_{\mathrm{phb}} + E_{\mathrm{SS}}(N) \quad (1)$$

The magnitude of the van der Waals interactions, v, was held fixed and the relative magnitudes of the other seven energy terms (s_{np}, s_{np}, s_{p}, e, D, P, and N as shown, where E_{SS} is an exponential function of N) were allowed to vary individually until the Z-score was maximized.

Results and Discussion

The resulting potential functions are shown in Table 1. For GB1, the maximum Z-score is 2.6, i.e., the wildtype sequence is assigned an energy lower than 99.5% of all possible sequences. For PCV, the maximum Z-score is 2.2. Also shown in Table VII-1 is the potential function built up over many experiments using the protein design cycle, which has been successful in particular for core design and α-helix surface design (Street & Mayo, 1999). The Z-score optimized potential functions exhibit some interesting common features. The

hydrophobic burial benefit, which is the main embodiment of the hydrophobic effect (Wesson & Eisenberg, 1992), has disappeared. This reflects the relative lack of importance of hydrophobic burial on the surface of proteins (although there may be some role for small hydrophobic clusters on the surface of β-sheets (Tisi & Evans, 1995)). The other solvation parameters are broadly similar to the experimental potential function.

The most dramatic difference from the protein design cycle potential is the increased importance of electrostatic interactions. The value of the dielectric constant used in the protein design cycle is similar to that of water, and leads to electrostatic interactions being de-emphasized. This value was never experimentally tested, however. Although saltbridges are not encouraged, the hydrogen bonding potential from the protein design cycle is quite strong (an ideal hydrogen bond receives a benefit of 8.0 kcal/mol). The *Z*-score optimized dielectric constant is an order of magnitude smaller, closer to unity. This is justifiable because we are considering effects at the molecular level, where the assumptions behind the use of the dielectric constant break down. The screening effect of solvent is also approximated by using a distance attenuated Coulomb potential (Mayo, et al., 1990).

To determine if the *Z*-score technique may be useful, this potential function must be used for real protein design. We used a combination of dead-end elimination (Desmet et al., 1992; Gordon & Mayo, 1998) and branch-and-terminate (Gordon & Mayo, 1999) to find the lowest energy sequence for each β-sheet surface, using the new potential functions. (These minimization algorithms are guaranteed to produce the absolute lowest energy sequence, unlike stochastic algorithms such as Monte Carlo.)

The resulting GB1 variant, GB1-Z1, is a five-fold mutant of the wildtype protein. One can clearly see the impact of the electrostatic term in the potential function. The modeled side chain configurations are shown in Figure VII-3b, alongside those of the wildtype crystal structures (Gallagher et al., 1994). A cluster of threonines and an isoleucine have been replaced by two cross-strand saltbridges, Asp42 to Arg55, and Arg6 to Glu53. The wildtype saltbridge formed by Lys4 and Glu15 is maintained. Such cross-strand saltbridges might be expected to contribute to β-sheet formation and stability, and surface networks of saltbridges are postulated to be a stabilizing factor in hyperthermophilic proteins (Elcock, 1998; de Bakker et al., 1999).

The resulting PCV variant, PCV-Z1, is a three-fold mutant of the wildtype protein. The modeled side chain configurations are shown in Figure VII-3d, alongside those of the apoplastocyanin wildtype crystal structure (Garrett et al., 1984). Again, the impact of the electrostatic term is clear, with a saltbridge network formed by Glu18, Lys95, Lys97 and Glu79.

The designed proteins were made experimentally using standard molecular biology techniques and their properties measured. Their far UV circular dichroism spectra overlay those of the wildtype proteins. The melting temperature of GB1-Z1 was determined to be 71 °C (Figure VII-4a). The melting temperature of GB1 is 86 °C. Although the designed protein is not as stable as the wildtype protein, it appears to fold to the correct structure. Although the literature contains many examples of alterations to the β-sheet surface of GB1, we know of no instances resulting in greater than wildtype stability. This is the first example of a well formed, many-stranded β-sheet designed through purely computational means.

The results for PCV-Z1 were even more impressive. The melting temperature of PCV-Z1 was determined to be 64 °C, compared to the melting temperature of PCV of 56 °C (Figure VII-4b). The designed protein is thus even more stable than the natural one. To our knowledge, this is the first time a natural protein's stability has been increased by redesigning its β-sheet surface.

Materials and Methods

***Simulation*:** The core residues of GB1 are positions 3, 5, 7, 20, 26, 30, 34, 39, 52, and 54. The eight β-sheet surface positions of GB1 considered here are 4, 6, 15, 17, 42, 44, 53, and 55. The α-helix surface positions of GB1 are 24, 27, 28, 31, 32, 35, and 36. The seven β-sheet surface positions of PCV considered here are 18, 20, 79, 81, 93, 95, and 97. These follow from our residue classification algorithm (Dahiyat & Mayo, 1997a). The potential function used in Figure VII-1 is derived from the protein design cycle, shown in Table 1.

The *Z*-score maximization algorithm searched along each potential function basis vector (that is, varying the scale factor for each energy term in Eq. 1) individually to maximize the *Z*-score. The search was initiated at the potential function derived from the protein design cycle, from the van der Waals potential alone, and from other random potentials, and always converged to the same result. Further, the ordering of the search through basis vectors had no effect on the result. It was found that this optimization algorithm was sufficient to find the maximum *Z*-score.

The *Z*-score was calculated using 4000 random sequences to determine the energy distribution of the potential function on the structure, resulting in an uncertainty in the *Z*-score of

±0.04. The random sequences were composed of the polar amino acids Ser, Thr, Asp, Asn, Glu, Gln, Lys and Arg, as well as the hydrophobic amino acids Ala, Val and Ile. The results were surprisingly robust to changes in the set of amino acids considered. In particular, the results were not significantly different if Ala was removed from consideration, or if His, Met and Gly were included.

In contrast to the case in lattice models, real amino acids may adopt many different conformations, or rotamers. The energy of a given amino acid sequence on a structure is thus calculated by minimizing the energy across all possible rotamer configurations, using dead-end elimination. For this procedure a backbone-dependent rotamer library was used (Dunbrack & Karplus, 1993), in which the χ_1 angles of all hydrophobic amino acid rotamers were expanded ±1 standard deviation about the mean value (Dahiyat, et al., 1997).

***Experimental*:** A synthetic GB1 gene (Minor & Kim, 1994) was cloned into a pET11a vector (Novagen) and used as the template for QuikChange mutagenesis (Qiagen). A synthetic PCV gene was constructed by recursive PCR (Prodromou, et. al., 1994) The genes were confirmed by DNA sequencing. The expression and purification of the protein followed published procedures, and was verified by mass spectrometry. The 56-residue form of GB1 (with N-terminal methionine processed) and the 100-residue form of PCV (including the N-terminal methionine) were used. PCV was derived from wildtype poplar apoplastocyanin (Garrett, et al., 1984) by removing its metal-binding site through the mutations His37 to Val and Cys84 to Ala. These mutations are in the core of the molecule and are not expected to interact with changes to the surface of the protein. The melting temperature of PCV was observed to be 56 °C compared to 51 °C for unmodified apoplastocyanin.

Far UV circular dichroism spectra were measured on an Aviv 62DS spectrometer. The spectra of GB1 and GB1-Z1 were measured at pH 5.5, in 50 mM phosphate and 50 μM protein, using a 1 mm path length, with thermal melts performed at 218 nm using 2 °C temperature steps with an averaging time of 30 s and an equilibration time of 2 min. The spectra of PCV and PCV-Z1 were measured at pH 7.0, in 50 mM potassium phosphate, 0.5 M sodium sulfate, and 70 μM protein, with thermal melts performed at 210 nm. The melting temperatures were derived by evaluating the maximum of a $d\theta/dT$ versus T plot. Protein concentration was determined by UV spectrophotometry.

Conclusion

It is interesting that we have designed two stable protein β-sheet surfaces using different potential functions. Indeed, further application of the technique to other proteins suggests yet different potentials may be appropriate. This supports the belief that there may be alternative routes taken by nature to stabilize protein surfaces, and which may be taken in *de novo* design too (Cordes et al., 1996). Of course, one test of this proposal is to use the potential derived from one protein to design the β-sheet surface of another, and preliminary results in this regard appear promising (unpublished data). The potential derived for plastocyanin was applied to protein G and the mutant (PGPC) was 7 °C more stable than GB1-Z1. A further advantage of the approach outlined in this Letter is that it could lead to a faster turn-around time for protein design, since it optimizes the potential function with less frequent recourse to experiment.

References

Chiu TL, Goldstein RA. 1998. Optimizing potentials for the inverse protein folding problem. *Prot Eng 11*: 749-752.

Cordes MHJ, Davidson AR, Sauer RT. 1996. Sequence space, folding and protein design. *Curr Opin Struct Biol 6*: 3-10.

Crippen GM. 1996. Failures of inverse folding and threading with gapped alignment. *Proteins 26*: 167-171.

Dahiyat BI, Gordon DB, Mayo SL. 1997. Automated design of the surface positions of protein helices. *Prot Sci 6*: 1333-1337.

Dahiyat BI, Mayo SL. 1996. Protein design automation. *Prot Sci 5*: 895-903.

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

Dahiyat BI, Mayo SL. 1997b. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA 94*: 10172-10177.

de Bakker PIW, Hunenberber PH, McCammon JA. 1999. Molecular dynamics simulations of the hyperthermophilic protein Sac7d from *Sulfolobus acidocaldarius*: contributions of salt bridges to thermostability. *J Mol Biol 285*: 1811-1830.

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

Deutsch JM, Kurosky T. 1996. New Algorithm for Protein Design. *Phys Rev Lett 76*: 323-326.

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

Elcock AH. 1998. The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. *J Mol Biol 284*: 489-502.

Gallagher T, Alexander P, Bryan P, Gilliland GL. 1994. Two crystal structures of the β1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochem 33*: 4721-4729.

Garrett TPJ, Clingeleffer DJ, Guss JM, Rogers SJ, Freeman HC. 1984. The crystal structure of poplar apoplastocyanin at 1.8-Angstroms resolution. The geometry of the copper-binding site is created by the polypeptide *J Biol Chem 259*: 2822-5..

Gordon DB, Marshall SA, Mayo SL. 1999. Energy functions for protein design. *Curr Opin Struct* 1999;9 509-13. Review.

Gordon DB, Mayo SL. 1998. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J Comp Chem 19*: 1505-1514.

Gordon DB, Mayo SL. 1999. Branch and terminate: a combinatorial optimization algorithm for protein design. Structure Fold Des. 1999 ;7 :1089-98.

Harbury PB, Zhang T, Kim PS, Alber T. 1993. A switch between 2-stranded, 3-stranded and 4-stranded coiled coils in gcn4 leucine-zipper mutants. *Science 262*: 1401-1407.

Hecht MH. 1994. De novo design of β-sheet proteins. *Proc Natl Acad Sci USA 91*: 8729-8730.

Hellinga HW. 1997. Rational protein design: combining theory and experiment. *Proc Natl Acad Sci USA 94*: 10015-10017.

Isogai Y, Ota M, Fujisawa T, Izuno H, Mukai M, Nakamura H, Iizuka T, Nishikawa K. 1999. Design and synthesis of a globin fold. *Biochem 38*: 7431-7443.

Kurosky T, Deutsch JM. 1995. Design of copolymeric materials. *J Phys A 27*: L387-L393.

Lau KF, Dill KA. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules 22*: 3986-3997.

Malakauskas SM, Mayo SL. 1998. Design, structure and stability of a hyperthermophilic protein variant. *Nature Struct Biol 5*: 470-475.

Mayo SL, Olafson BD, Goddard WA, III. 1990. Dreiding - a generic force-field for molecular simulations. *J Phys Chem 94*: 8897-8909.

Micheletti C, Seno F, Maritan A, Banavar JR. 1998a. Design of proteins with hydrophobic and polar amino acids. *Proteins 32*: 80-87.

Micheletti C, Seno F, Maritan A, Banavar JR. 1998b. Protein design in a lattice model of hydrophobic and polar amino acids. *Phys Rev Lett 80*: 2237-2240.

Minor DL, Kim PS. 1994. Measurements of the β-sheet-forming propensities of amino acids. *Nature 367*: 660-663.

Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules 18*: 534-552.

Morgan CS. 2000. *Full sequence design of an alpha-helical protein and investigation of the importance of helix dipole and capping effects in helical protein design*. PhD thesis. California Institute of Technology.

Seno F, Micheletti C, Maritan A, Banavar JR. 1998. Variational approach to protein design and extraction of interaction potentials. *Phys Rev Lett 81*: 2172-2175.

Shakhnovich EI. 1994. Proteins with selected sequences fold into unique native conformations. *Phys Rev Lett 72*: 3907-3910.

Shakhnovich EI, Gutin AM. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc Natl Acad Sci USA 90*: 7195-7199.

Street AG, Mayo SL. 1998. Pairwise calculation of protein solvent accessible surface areas. *Folding and Design 3*: 253-258.

Street AG, Mayo SL. 1999. Computational protein design. *Structure 7*: R105-R109.

Sun S, Brem R, Chan HS, Dill KA. 1995. Designing amino acid sequences to fold with good hydrophobic cores. *Prot Eng 8*: 1205-1213.

Tisi LC, Evans PA. 1995. Conserved structural features on protein surfaces: small exterior hydrophobic clusters. *J Mol Biol 249*: 251-258.

Wesson L, Eisenberg D. 1992. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Prot Sci 1*: 227-235.

Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA. 1995. A test of lattice protein folding algorithms. *Proc Natl Acad Sci USA 92*: 325-329.

Table VII-1. Potential functions determined through different methods. The energy terms considered are shown in Eq. 1. The van der Waals energy scale factor *v* was held fixed. A potential function has been developed using the protein design cycle (Street & Mayo, 1999), and has been successful for core and α-helix surface design in particular. The *Z*-score method applied to the β-sheet surface of PCV and of GB1 yield new potential functions. Also shown are the ranges over which each parameter may be changed while keeping the *Z*-score within 5% of its maximum (when the other parameters are kept fixed).

Energy term	Design cycle	PCV	Range	GB1	Range
van der Waals v	1.0	1.0	n.a.	1.0	n.a.
np burial σ_{np} (kcal/mol/Å^2)	0.05	0.0	0.0 – 0.01	0.0	0.0 – 0.02
np exposure σ_{np} (kcal/mol/Å^2)	0.05	0.10	0.04 – 0.16	0.06	0.02 – 0.08
polar burial σ_p (kcal/mol/Å^2)	0.0	0.0	0.0 – 0.04	0.03	0.01 – 0.06
dielectric ε	40.0	4.0	2.0 – 6.0	4.0	2.0 – 6.0
H-bond D	8.0	1.0	1.0 – 8.0	6.0	1.0 – 8.0
polar H burial P	2.0	9.0	6.0 – 15.0	3.0	1.0 – 7.0
secondary structure bias N	n.a.	1.0	0.0 – 1.4	1.4	0.8 – 1.6

Figure VII-1. The assumed distribution of energies of sequences threaded onto the target structure. Sequence S_0 is the lowest energy sequence whose ground state structure is the target structure. Note that there may be sequences of lower energy that do not fold to the target structure. By altering the energy function non-thermodynamically, negative design seeks to move these sequences above S_0.

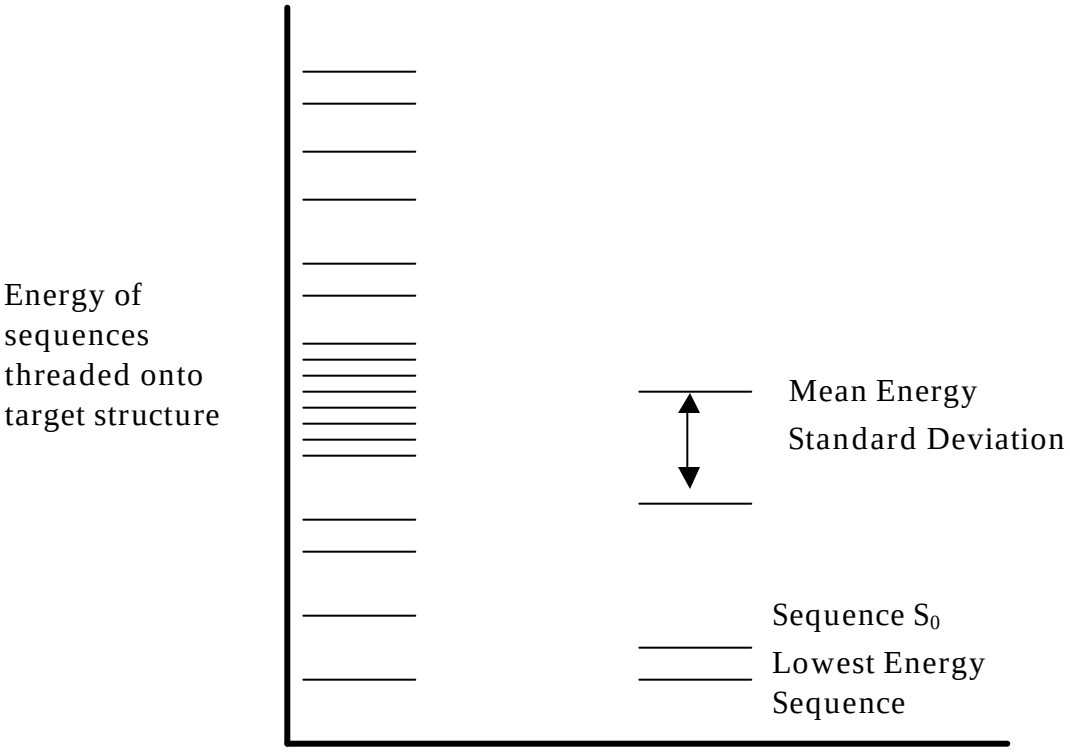
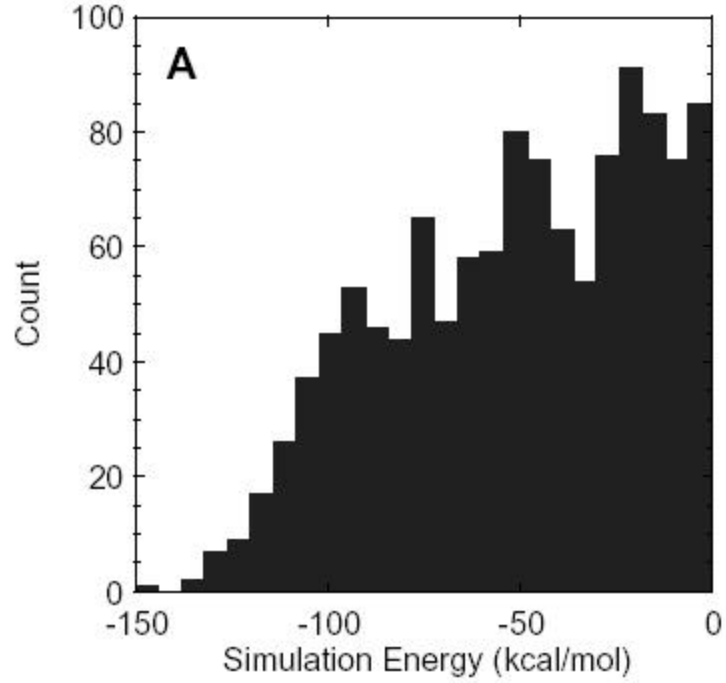

Energy of sequences threaded onto target structure
Mean Energy
Standard Deviation
Sequence S_0
Lowest Energy Sequence

Figure VII-2. The actual distribution of energies of various subsets of the real protein GB1, using the potential function derived from the protein design cycle (Table 1). a) The core (only the 2.5% lowest-energy sequences are shown), b) the β-sheet surface.


A
Count
100
80
60
40
20
0
-150
-100
-50
0
Simulation Energy (kcal/mol)

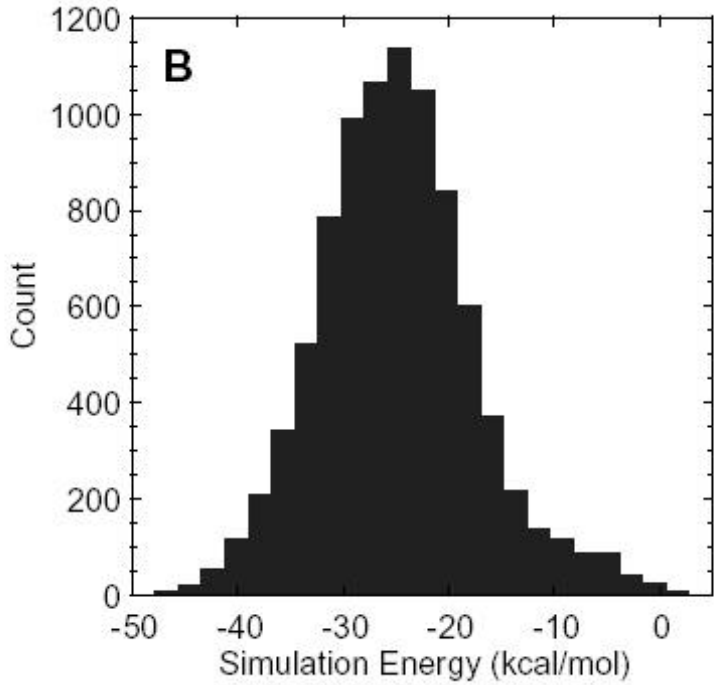

B
Count
1200
1000
800
600
400
200
0
-50
-40
-30
-20
-10
0
Simulation Energy (kcal/mol)

Figure VII-3. Views of the eight designed positions on the β-sheet surface of GB1 and PCV. a) the crystographically-determined wildtype GB1 side chain orientations. b) the orientations modeled using the *Z*-score-derived potential function on GB1. c) the crystographically-determined PCV wildtype side chain orientations. d) the orientations modeled using the *Z*-score-derived potential function on PCV.

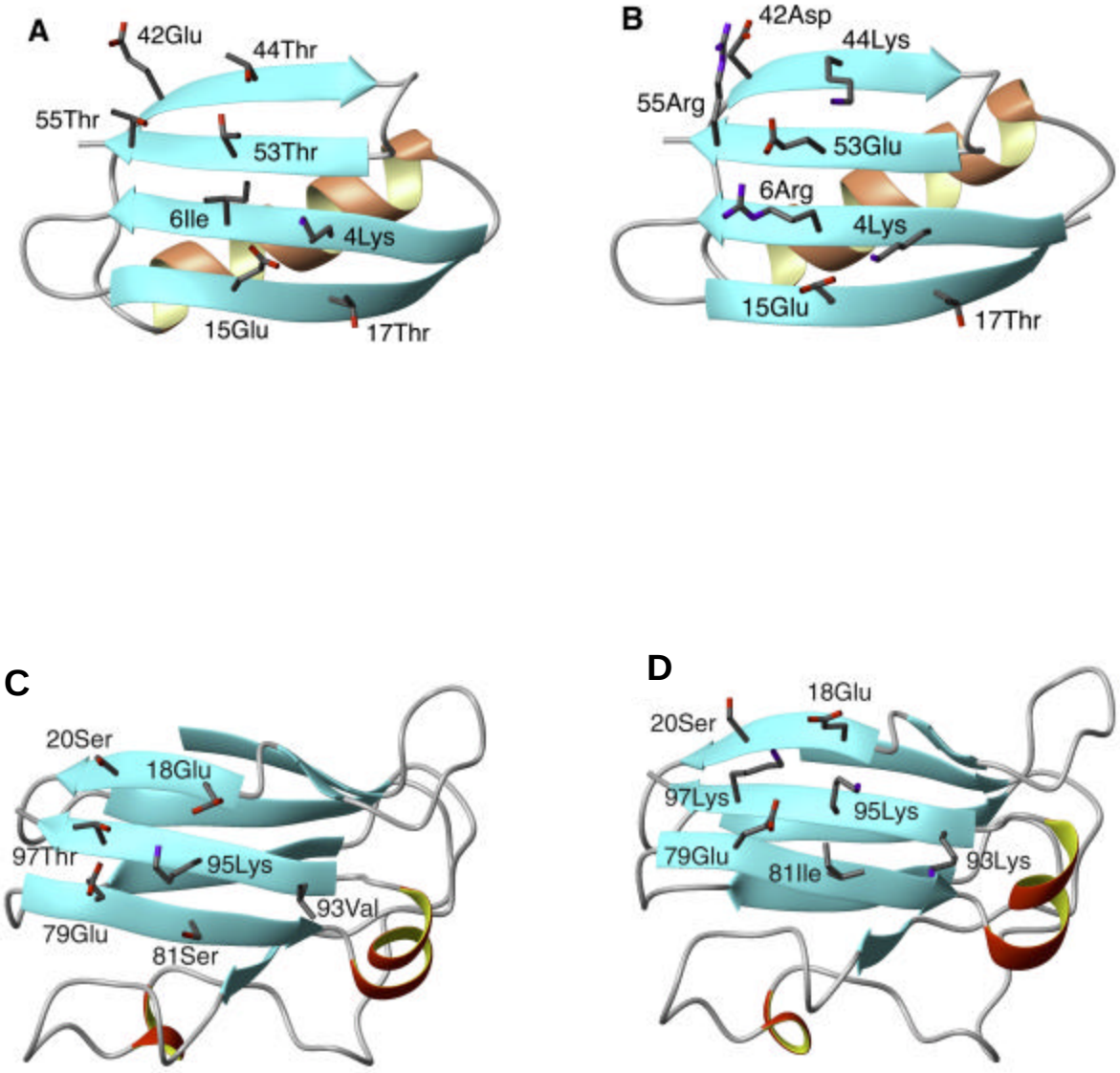
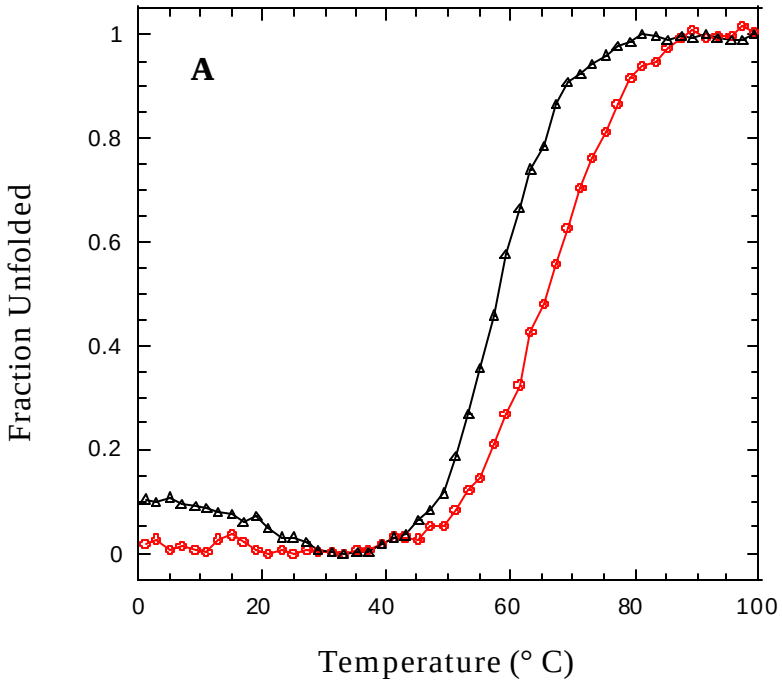
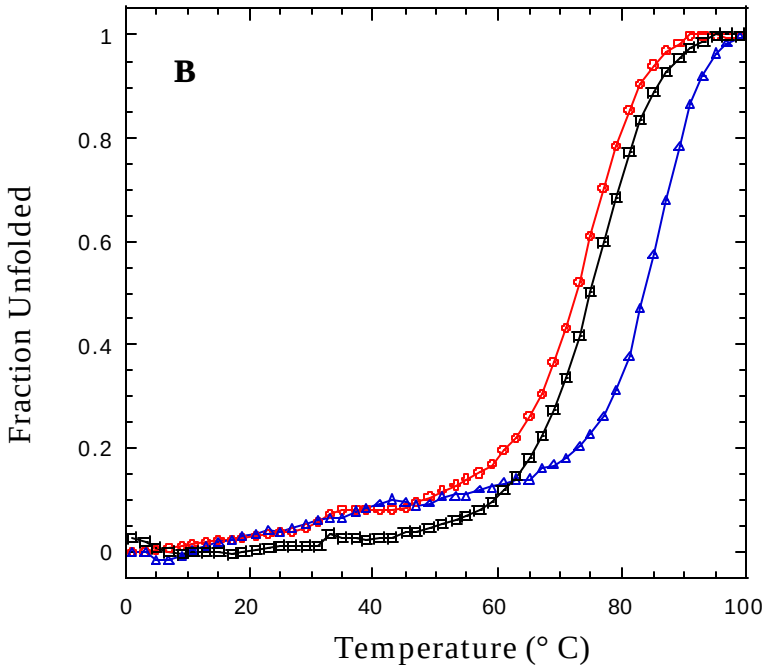

A
42Glu
44Thr
55Thr
53Thr
6Ile
4Lys
15Glu
17Thr
B
42Asp
44Lys
55Arg
53Glu
6Arg
4Lys
15Glu
17Thr
C
20Ser
18Glu
97Thr
95Lys
93Val
79Glu
81Ser
D
18Glu
20Ser
97Lys
95Lys
79Glu
81Ile
93Lys

Figure VII-4. Thermal denaturation data. a) Circular dichroism measurements of PCV (black) and PCV-Z1 (red) with temperature at 210 nm. b) Circular dichroism measurements of GB1 (blue), GB1-Z1 (red) and PGPC (black) with temperature at 218 nm.


A
Fraction Unfolded
Temperature (° C)
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100


B
Fraction Unfolded
Temperature (° C)
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100

Chapter 8

Evaluation of the Energetic Contribution of an Ionic Network to Beta-Sheet Stability

Adapted from Kirsten S. Lassila, Deepshikha Datta, and Stephen L. Mayo. Protein Sci. 2002 Mar;11(3):688-90.

Abstract

We have evaluated the interaction energy of a three residue ionic network constructed on the beta-sheet surface of protein G using double mutant cycles. Although the two individual ion pairs were each stabilizing by around 0.6 kcal/mol, the excess gain in stability for the triad was small (0.06 kcal/mol).

Introduction

The β-sheet surface of the protein G immunoglobulin-binding domain B1 (GB1) has been used as a model system for evaluating the β-sheet forming propensities of amino acids (Minor & Kim, 1994; Smith et al., 1994). These studies, in combination with statistical surveys of known structures and theoretical models of β-sheet propensity provide some general guidelines for amino acid selection in β-sheet design (Munoz & Serrano, 1994; Street & Mayo, 1999).

An important next step in understanding β-sheet stability is to define the role of side chain interactions such as hydrogen bonding and ionic interactions. In particular, the energetic effects of surface ionic interactions have been debated. Solvent exposed ion pairs have been found to stabilize folded proteins in a number of cases (Lyu et al., 1992; Horovitz et al, 1990; Spek et al, 1998; Takano et al, 2000; Serrano et al, 1990). In the context of the β-sheet surface environment, ion pairs have been reported to stabilize folded proteins by 0.4 – 1 kcal/mol (Merkel et al, 1999; Blasie & Berg, 1997; Smith & Regan, 1995). However, some surface ion pairs exhibit neutral or destabilizing effects (Strop & Mayo, 2000; Dao-pin et al, 1991). The high dielectric of the aqueous environment and the loss of side chain conformational freedom have been invoked to explain the marginal stabilizing effects of some pairwise electrostatic interactions.

Networks of charged surface residues have been observed in hyperthermophile proteins and have been proposed to offer an energetic advantage over single ion pairs due to the reduced entropic cost of fixing a third residue (Yip et al, 1995; Dao-pin et al, 1991). Indeed, two analyses of solvent exposed ionic triads in

α-helical regions have shown that three-residue networks offer a stabilizing effect greater than would be observed for the sum of the two individual pairwise interactions (Horovitz et al, 1990; Spek et al, 1998).

To test the effect of an ionic triad in the context of the β-sheet surface, we have evaluated the energetic contribution of a three residue triad constructed on the β-sheet surface of GB1. The network consists of Arg 6, Glu 53, and Arg 44, residues which lie on three adjacent strands of the β-sheet surface (Figure 1a). Double mutant cycle analysis was used to isolate the interaction energy of the triad (Horovitz & Fersht, 1990). Eight GB1 variants were constructed which represent all permutations of Arg or Ile at position 6, Glu or Ala at position 53, and Arg or Ala at position 44. In this three-residue thermodynamic cycle, the interaction energy of the ionic network is calculated as in equation 1.

$$???G^{RER}_{interaction} = \{(?G^{RER} - ?G^{RAA}) - [(?G^{REA} - ?G^{RAA}) + (?G^{RAR} - ?G^{RAA})]\} - \{(?G^{IER} - ?G^{IAA}) - [(?G^{IEA} - ?G^{IAA}) + (?G^{IAR} - ?G^{IAA})]\} \quad [\mathbf{1}]$$

$?G^{XYZ}$ is the free energy of unfolding for the GB1 mutant with amino acids X, Y, and Z at positions 6, 53, and 44, respectively. The interaction energy of an Arg-Glu ion pair in the presence of another residue X is calculated as in equation 2.

$$??G^{XER}_{interaction} = (?G^{XER} - ?G^{XAA}) - [(?G^{XEA} - ?G^{XAA}) + (?G^{XAR} - ?G^{XAA})] \quad [\mathbf{2}]$$

Free energies of unfolding (?G) were evaluated by two-state analysis of thermal denaturation curves monitored by circular dichroism (CD) (Figure 1b).

The variant containing both a single ion pair and isoleucine, $I_6E_{53}R_{44}$, had the highest T_m and ?G of unfolding (Table 1) while the variant with the ionic triad, $R_6E_{53}R_{44}$ was only slightly less stable. It is interesting to note that the addition of the third charged residue almost fully compensates for the loss of the beta-branched (and therefore β–sheet stabilizing) amino acid.

Interaction energies of the Arg6-Glu53 pair were 0.58 kcal/mol in the presence of Ala44 and 0.64 kcal/mol in the presence of Arg44 (Table 2). The Arg44-Glu53 pair had interaction energies of 0.51 kcal/mol (in the presence of Ile6) and 0.57 kcal/mol (with Arg6). This level of stabilization is consistent with other surface electrostatic interactions studied by double mutant cycles (Spek et al, 1998; Serrano et al, 1990; Merkel et al, 1999).

Although the pairwise electrostatic interactions are clearly favorable, the ionic network does not appear to significantly enhance GB1 stability any more than the simple sum of the individual pairs. As shown in Table 2, the interaction energy of unfolding for the ionic network, $???G_{interaction}$, determined at 75 °C (approximately the average T_m for the eight variants) was 0.06 kcal/mol. This very low interaction energy suggests that the contributions of the ion pairs are additive; there is no additional stabilization of one ion pair in the presence of a third charged residue. In contrast, previous studies of charged networks on α-helices using the double mutant cycle method showed stabilizing interaction energies of 0.77 kcal/mol for an Asp-

Arg-Asp triad (Horovitz et al, 1990) and 0.65 kcal/mol for an Arg-Glu-Arg triad (Spek et al, 1998).

The lack of a significant stabilizing interaction energy of the Arg6-Glu53-Arg44 triad may be due to a variety of factors. Previously reported factors such as desolvation, side chain entropy loss, and conformational strain may counteract the electrostatic benefits of the network. However, the local environment of the triad, including secondary structure and neighboring residues, may also affect the magnitude of the interaction energy of the triad. Further studies on β-sheet surface electrostatic interactions may help to clarify whether or not secondary structure influences the stabilizing effect of ionic networks.

Methods

Mutagenesis and protein expression

GB1 variants were constructed by inverse PCR mutagenesis and expressed using the T7 promoter system as previously described (Su & Mayo, 1997). Purification of 57-residue GB1 variants containing an N-terminal methionine was accomplished by reverse phase HPLC and verified by mass spectrometry.

Thermal denaturation

The increase in CD signal at 218 nm was followed during thermal unfolding from 1 °C to 99 °C using 50 μM protein in 50 mM sodium phosphate, pH 5.5. The midpoint of the thermal denaturation (T_m) and the enthalpy of unfolding (?H) were determined from a two-state analysis of each denaturation curve (Minor & Kim,

1994; Smith et al, 1994). The change in heat capacity upon unfolding (?C_p) was held constant at 0.621 kcal/K•mol, a value previously reported for wild-type GB1 (Alexander et al, 1992). ?G values were assigned using the Gibbs-Helmholtz relation with ?C_p = 0.621 kcal/K•mol (Minor & Kim, 1994; Smith et al, 1994). The average error in calculating ?G (as determined from curve fitting) was 0.06 kcal/mol.

References

Alexander, P., Fahnestock, S., Lee, T., Orban, J., and Bryan, P. 1992. Thermodynamic analysis of the folding of the streptococcal protein-G IgG-binding domains B1 and B2- why small proteins tend to have high denaturation temperatures. *Biochemistry* **31:** 3597-3603.

Blasie, C.A. and Berg, J.M. 1997. Electrostatic interactions across a beta-sheet. *Biochemistry* **36:** 6218-6222.

Dao-pin, S., Sauer, U., Nicholson, H., and Matthews, B.W. 1991. Contributions of engineered surface salt bridges to the stability of T4 lysozyme determined by directed mutagenesis. *Biochemistry* **30:** 7142-7153.

Horovitz, A., Serrano, L., Avron, B., Bycroft, M., and Fersht, A.R. 1990. Strength and cooperativity of contributions of surface salt bridges to protein stability. *J. Mol. Biol.* **216:** 1031-1044.

Horovitz, A. and Fersht, A.R. 1990. Strategy for analyzing the cooperativity of intramolecular interactions in peptides and proteins. *J. Mol. Biol.* **214:** 613-617.

Kraulis, P.J. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallography.* **24:** 946-950.

Lyu, P.C.C., Gans, P.J., and Kallenbach, N.R. 1992. Energetic contribution of solvent-exposed ion-pairs to alpha-helix structure. *J. Mol. Biol.* **223:** 343-350.

Merkel, J.S., Sturtevant, J.M., and Regan, L. 1999. Sidechain interactions in parallel beta sheets: the energetics of cross-strand pairings. *Structure* **7:** 1333-1343.

Minor, D.L. and Kim, P.S. 1994. Measurement of the β-sheet-forming propensities of amino acids. *Nature* **367:** 660-663.

Munoz, V. and Serrano, L. 1994. Intrinsic secondary structure propensities of the amino-acids, using statistical phi-psi matrices- comparison with experimental scales. *Proteins* **20:** 301-311.

Serrano, L., Horovitz, A., Avron, B., Bycroft, M., and Fersht, A.R. 1990. Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry* **29:** 9343-9352.

Smith, C.K., Withka, J.M., and Regan, L. 1994. A thermodynamic scale for the β-sheet forming propensities of the amino acids. *Biochemistry* **33:** 5510-5517.

Smith, C.K. and Regan, L. 1995. Guidelines for protein design- the energetics of β sheet side-chain interactions. *Science* **270:** 980-982.

Spek, E.J., Bui, A.H., Lu, M., and Kallenbach, N.R. 1998. Surface salt bridges stabilize the GCN4 leucine zipper. *Protein Sci.* **7:** 2431-2437.

Street, A.G. and Mayo, S.L. 1999. Intrinsic beta-sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc. Natl. Acad. Sci. USA* **96:** 9074-9076.

Strop, P. and Mayo, S.L. 2000. Contribution of surface salt bridges to protein stability. *Biochemistry* **39:** 1251-1255.

Su, A. and Mayo, S.L. 1997. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.* **6:** 1701-1707.

Takano, K., Tsuchimori, K., Yamagata, Y., and Yutani, K. 2000. Contribution of salt bridges near the surface of a protein to the conformational stability. *Biochemistry* **39:** 12375-12381.

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

Yip, K.S.P., Stillman, T.J., Britton, K.L., Artymiuk, P.J., Baker, P.J., Sedelnikova, S.E., Engel, P.C., Pasquo, A., Chiaraluce, R., Consalvi, V., Scandurra, R., and Rice, D.W. 1995. The structure of pyrococcus-furiosis glutamate-dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure* **3:** 1147-1158.

Table VIII-1 Stability data for GB1 variants

Variant	Tm[1] (°C)	?H_{Tm}[2] (kcal/mol)	?G (75 °C)[3] (kcal/mol)
$R_6E_{53}R_{44}$	79.1 ± 0.5	53.5 ± 2.0	0.61
$R_6E_{53}A_{44}$		76.1 ± 0.4	52.3 ± 2.0
	0.16		
$R_6A_{53}R_{44}$		69.6 ± 0.3	45.0 ± 1.6
	-0.74		
$R_6A_{53}A_{44}$	70.6 ± 0.3	46.6 ± 1.5	-0.62
$I_6E_{53}R_{44}$	80.1 ± 0.6	56.1 ± 2.5	0.79
$I_6E_{53}A_{44}$	77.2 ± 0.4	54.1 ± 1.9	0.34
$I_6A_{53}R_{44}$		75.5 ± 0.3	50.1 ± 1.5
	0.08		
$I_6A_{53}A_{44}$		76.0 ± 0.3	51.4 ± 1.6
	0.14		

[1] T_m, midpoint of thermal denaturation transition
[2] ?H_{Tm}, enthalpy of unfolding at T_m
[3] ?G (75 °C), free energy of unfolding calculated at 75 °C

Table VIII-2. Interaction energies for ion pairs and the three-residue network.

Interaction	??G (75 °C)[4] Kcal/mol	???G (75 °C)[5] Kcal/mol
R_6E_{53} (A_{44})	0.58	
R_6E_{53} (R_{44})	0.64	
$E_{53}R_{44}$ (I_6)	0.51	
$E_{53}R_{44}$ (R_6)	0.57	
$R_6E_{53}R_{44}$		0.06

[4]??G (75 °C) interaction energy (calculated at 75 °C) of the ion pair in the presence of the residue indicated in parentheses
[5]???G (75 °C) interaction energy (calculated at 75 °C) of the triad as described in the text

Figure 1. The β-sheet surface of GB1 showing possible orientations for side chains Arg 6, Glu 53, and Arg 44. In the positions shown, nitrogen-oxygen distances are 2.92 Å and 2.85 Å for residue pairs 6-53 and 44-53, respectively. Side chains were positioned with a dead-end elimination algorithm (Voigt et al, 2000) and the figure was created with MOLSCRIPT (Kraulis, 1991).

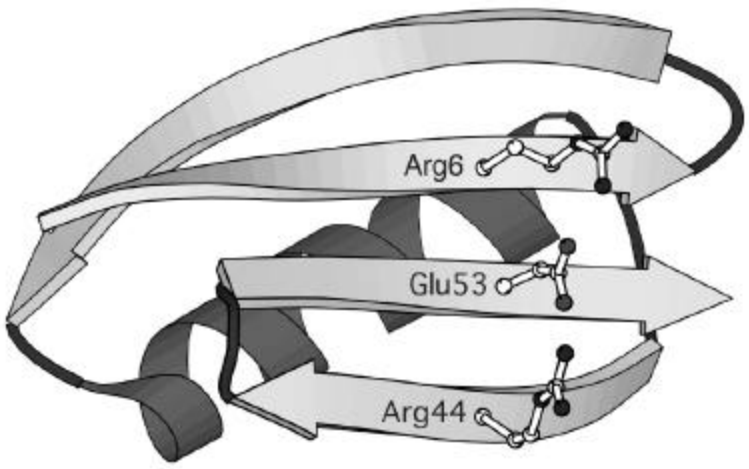
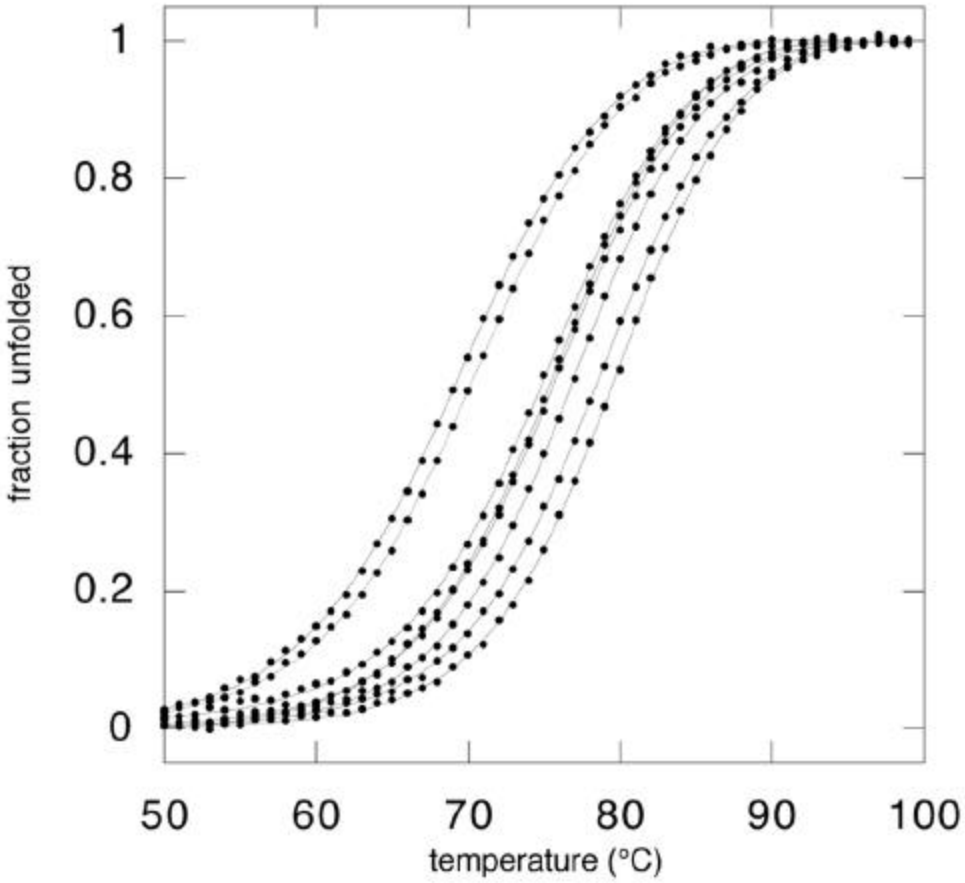

Arg6
Glu53
Arg44

Figure VIII-2 Thermal denaturation curves for GB1 variants. From left to right (at 50% unfolded): $R_6A_{53}R_{44}$, $R_6A_{53}A_{44}$, $I_6A_{53}R_{44}$, $I_6A_{53}A_{44}$, $R_6E_{53}A_{44}$, $I_6E_{53}A_{44}$, $R_6E_{53}R_{44}$, and $I_6E_{53}R_{44}$.


fraction unfolded
1
0.8
0.6
0.4
0.2
0
50
60
70
80
90
100
temperature (°C)

Chapter 9

Redesigning Aminoacyl-tRNA Synthetases for *in vivo* Incorporation of Non-natural Amino Acids

Adapted from Datta D, Wang P, Carrico IS, Mayo SL, Tirrell DA. J Am Chem Soc 2002; 124:5652-3

ABSTRACT

We present here a computationally designed new mutant *E. coli* phenylalanyl-tRNA synthetase, which allows efficient incorporation of aryl ketone functionality into protein *in vivo*. We demonstrate chemoselective modification of ketone-containing protein as a new means for selective modification of recombinant proteins. Given the mild reaction conditions, we envision that this chemoselective formation of hydrazone linkage can be widely utilized to *in vivo* label proteins of interest and post-translational protein engineering.

This study also demonstrates the usefulness of computational protein design in manipulating tRNA synthetases for incorporation of non-natural aminoacids. It can be potentially used to design both the synthetase for the incorporation amino acid analog and also the protein in which we would like to incorporate the analog to enhance or modify its structural, catalytic and binding properties. This dual design approach, in which we are not limited to the 20 natural amino acids, could be used as a very powerful protein-engineering tool to design biomolecules with novel structures and functions.

Introduction

Protein engineering, a powerful tool for structural and functional modification of proteins, relies on an efficient recognition mechanism for incorporating mutant amino acids in the desired protein sequences. Though this process has been very useful for designing new macromolecules with precise control of composition and architecture, a major limitation is that the mutagenesis is restricted to the 20 naturally occurring amino acids. For many applications of designed macromolecules, it would be desirable to develop methods for incorporating amino acids that have novel chemical functionality not possessed by the 20 natural amino acids. For example, the ability to synthesize large quantities of proteins containing heavy atoms would facilitate protein structure determination, and the ability to site specifically substitute fluorophores or photocleavable groups into proteins in living cells would provide powerful tools for studying protein functions *in vivo.* One might be able to enhance the properties of proteins by providing building blocks with new functional groups, such as an amino acid containing a keto-group.

Incorporation of novel amino acids in macromolecules has been successful to an extent. Biosynthetic assimilation of non-canonical amino acids into proteins has been achieved largely by exploiting the capacity of the wild type synthesis apparatus to utilize analogs of naturally occurring amino acids [1-5]. Nevertheless, the number of amino acids shown conclusively to exhibit translational activity *in vivo* is small, and the chemical functionality that has been accessed by this method remains modest. In designing macromolecules with desired properties, this poses a limitation since such designs may require incorporation of complex analogs that differ significantly from the natural substrates in terms of both size and chemical properties and hence, are unable to

circumvent the specificity of the synthetases. To further expand the range of non-natural amino acids that can be incorporated, the activity of the aminoacyl tRNA synthetases (AARS) needs to be manipulated.

The importance of the aminoacyl-tRNA synthetases (aaRS) in determining the success or failure of analogue incorporation has been recognized for decades. [5-16] Here, we report a computationally designed new variant of the *E. coli* phenylalanyl-tRNA synthetase (*e*PheRS), which allows efficient incorporation of aryl ketone functionality into protein *in vivo.* In this study we demonstrate computational protein design to an extremely useful tool in predicting mutations in synthetases that will allow incorporation of new analogs that do not get incorporated by the natural synthetases.

In 1991, Kast and coworkers [8, 10, 11] introduced a variant of *e*PheRS (termed *e*PheRS*), which bears an Ala294Gly mutation and which thereby acquires relaxed substrate specificity. We have recently shown that over-expression of *e*PheRS* can be exploited to effect efficient incorporation of *p*-bromo-, *p*-iodo-, *p*-ethynyl-, *p*-cyano- and *p*-azidophenylalanines into recombinant proteins in *E. coli* hosts[5, 7]. But similar experiments with *p*-acetylphenylalanine (DPA) failed (Figure IX-4); even in a host in which *e*PheRS* was over-expressed, phenylalanine-depleted cultures supplemented with DPA did not produce detectable yields of protein (Figure IX-5). Our interest in DPA arises from the chemical versatility of the side-chain ketone function, which can be chemoselectively ligated with hydrazide, hydroxylamino, and thiosemicarbazide group under physiological conditions[17-22]. Cornish and coworkers have accomplished site-specific incorporation of ketone functionality into recombinant proteins via *in vitro* translation; however, we are unaware of previous reports of *in vivo* methods of introducing

ketone functionality, Bertozzi and coworkers have reported that *N*-levulinoylmannosamine can be incorporated into cell surface oligosaccharides and that the associated ketone functionality can be captured through reaction with nitrogen nucleophiles [20, 22].

Methods

Computational Modeling

Model System

Phenylalanine –tRNA synthetase (PheRS): We have selected the PheRS as our model for this study. PheRS is an $(a\beta)_2$ enzyme with 350 amino acids in the a subunit and 785 in the ß subunit [23, 24]. The binding site for phenylalanine (Phe) is in the a subunit (Figure IX-2). There are a number of reasons for selecting this system. Although the crystal structure of E.coli PheRS is unavailable, the crystal structure of *Thermus aquaticus* PheRS (*t*PheRS) complexed with Phe is available [23]and there is 43% identity between *e*PheRS and *t*PheRS (Figure IX-1). An accurate description of the binding pocket is critical for the computational design approach since it depends on the crystal structure for the protein backbone descriptions. The crystal structure also defines the orientation of Phe in the binding pocket of the synthetase. The aromatic ring of Phe is buried in the hydrophobic pocket while the carbonyl and the amide groups of the backbone make extensive electrostatic contacts with the charged and polar residues at the mouth of the pocket. We attempted to design the binding pocket for the analogs so that they bind to PheRS in the same orientation as Phe since this orientation may be important for the adenylation step.

Another reason for using this system is that the structure of PheRS does not undergo any significant structural rearrangement on substrate binding as indicated by the

crystal structures of free PheRS and the PheRS-Phe complex. This makes the system well suited for our protein design algorithm which uses a fixed-backbone structure for side-chain selection. PheRS is an extensively studied system for the assimilation of artificial amino acids and has been used successfully in incorporation of a few Phe analogs. In fact, limited mutation analysis has also been done on this system to alter substrate specificity (Kast and Hennecke,1991). Moreover, molecular dynamics (MD) simulations have been performed in an effort to understand the binding behavior of PheRS towards various analogs.

Analogs: The Phe analogs selected for the study are shown in Figure IX-4. These analogs have been experimentally tested to check if they are readily incorporated in proteins by the natural PheRS. MD simulations were performed using these analogs and significant correlation was achieved between the predicted binding energies and the *in vivo* incorporation rates exhibited by these analogs (Wang *et al*., 2001).

Binding Site Design

We used a protein design algorithm [25, 26], ORBIT, to predict the optimal amino acid sequences of the binding pocket for binding to the different analogs. Selection of amino acids is performed using a very efficient search algorithm that relies on a discrete set of allowed conformations for each side chain and empirical potential energy functions that are used to calculate pairwise interactions between side chains and the between the side chains and backbone.

Surveys of protein structure database have shown that side chains exhibit marked conformational preferences, and that most side chains are limited to a small number of torsional angles. Thus, the torsional flexibility of most amino acids can be represented with

a discrete set of allowed conformations called rotamers. Rotameric preferences in side chains are observed that depend on the main-chain conformations. ORBIT accounts for the torsional flexibilities of side chains by providing rotamer libraries that are based on those developed by Dunbrack and Karplus [27, 28].

In our design, we performed optimization calculations by varying the torsional angles of the analogs and the side chains lining the pocket simultaneously. This required generating rotamer libraries for the analogs, since they are not included in our standard rotamer libraries. For all the natural amino acids, the possible ?1 and ?2 angles are derived from database analysis. Since this was not feasible in the case of artificial amino acids, the closest approximation for ?1 and ?2 angles for Phe analogs were taken to be the same as those for Phe. Moreover, our goal was to select for conformations that were as close as possible to the orientation of Phe in the binding pocket. So allowing similar torsional angles for the analogs as of Phe seems logical.

Since the residues in the pocket are buried in the protein structure, we used force field parameters similar to those used in previous protein core design calculations. The design algorithm uses energy terms based on a force field that includes van der Waals interactions, electrostatic interactions, hydrogen bonding, and solvation effects [29].

We generated backbone independent rotamer libraries for all the analogs shown in Figure IX-4. Both the ?1 and ?2 torsional angles were varied to match those of Phe rotamers in our standard backbone independent rotamer library. The torsional angles of Phe in the crystal structure (?1: –101º, ?2: –104º) were also included in the new rotamer libraries for both Phe and the analogs. Charges were assigned only to the heavy atoms of the analogs to be consistent with the way charges for the natural amino acids are

represented in ORBIT. The first analog selected for design was acetylphenylalanine (DPA).

Design calculations were run by fixing the identity of the substrate to be DPA and varying 11 other positions on PheRS (137, 184, 187, 222, 258, 260, 261, 286, 290, 294, 314). Positions 137, 184, 258, 260, 261, 286, 290, 294, 314 were allowed to be any of the 20 natural amino acids except proline, methionine and cysteine. Methionine was allowed at position 187 because its wild-type identity is Met and only hydrophobic amino acids were allowed at position 222. Most of these positions are buried in the core and a number of them pack against Phe in the crystal structure. Mutation analysis at position 294 has been shown to alter substrate specificity. The anchor residues (Glu 128, Glu 130, Trp 149, His 178, Ser 180, Gln 183, Arg 204) were held fixed both in identity and conformation in all the calculations. These make very important electrostatic interactions with the substrate and this interaction is probably equally critical for the analogs. From the crystal structure it appears that the anchor residues hold the Phe zwitterion in a way that the carbonyl group of the zwitterion is close to the ATP binding site. This proximity may be important for the aminoadenylation reaction. The aminoadenylation step is required for the incorporation of all the amino acids and hence, it seems important to make sure that the carbonyl and the amide groups of the analog zwitterions are also anchored the same way as the natural substrate at this site.

In the first design attempt we allowed all the DPA rotamers that were generated in the rotamer library. The DPA rotamer selected in the structure generated was not buried in the binding pocket and most of it is solvent-exposed. A second calculation was run allowing only those DPA rotamers that would pack into the binding pocket. These are the

rotamers with all possible combinations of ?1 of –101 (±20°) and ?2 of -104 (±20°) in increments of 5°. The structure generated in this calculation has the aromatic ring of DPA buried in the pocket and is almost completely superimposable with the Phe in the PheRS crystal structure.

Gene construction, protein expression and analysis of designed PheRS activity

E. coli PheRS* was amplified by the polymerase chain reaction (PCR) from vector pQE-FS [5] The amplified *e*PheRS* was subjected to PCR mutagenesis to create the desired Thr251Gly mutant, which we designate *e*PheRS**. To allow constitutive expression of the synthetase, a *tac* promoter with an abolished *lac* repressor binding site was inserted upstream of the start codon of the *e*PheRS** gene [30]. The constitutive expression cassette was then cloned into pQE15 (Qiagen), which encodes the marker protein mouse dihydrofolate reductase (mDHFR). The resulting plasmid was designated pQE-FS**. As a control, plasmid pQE-FS* containing *e*PheRS* under control of *tac* constitutive promoter was constructed similarly. AF-IQ, a phenylalanine auxotrophic cell strain carrying the repressor plasmid pLysS-IQ[5], was transformed with pQE15, pQE-FS*, or pQE-FS** to generate expression systems AF-IQ[pQE15], AF-IQ[pQE-FS*], and AF-IQ[pQE-FS**], respectively. The capacity of DPA to support protein synthesis in each expression system was determined by induction of mDHFR expression in phenylalanine-free minimal media supplemented with DPA. The histidine-tagged protein isolated from the latter culture (mDHFR-DPA) was purified by nickel-affinity chromatography. The isolated protein yield was about 20 mg/L, approximately 60% of that obtained from cultures supplemented with Phe. Incorporation of DPA was confirmed by MALDI-TOF mass spectrometry analysis and tryptic peptide digestion analysis of purified mDHFR

With ketone-modified mDHFR, we also investigated the reactivity of hydrazide with this protein. Purified mDHFR-wt and mDHFR-2 was dissoved in PBS buffer (pH=6.0) and treated with 5 mM biotin hydrazide (BH) or PBS buffer alone as a negative control. The reaction products were analyzed by western blot after being seperated by SDS-PAGE and stained with streptavidin HRP conjugate (Figure IX-7). The products were also examined for the presence of 6xHis tag of mDHFR to ensure existence of proteins and no degradation occuring after reaction.

Results and Discussion

The structure generated in the design calculation has the aromatic ring of DPA buried in the pocket and is almost completely superimposable with the Phe in the PheRS crystal structure. We also ran a control calculation where we fixed Phe as the substrate. In all calculations, with DPA, position 258 is mutated from Phe to Tyr. This position is slightly exposed and is not in direct contact with DPA. Position 258 is very close to the anchor residues so mutating it may affect the transfer of the amino acid to its RNA and therefore, we did not include this mutation. We also ignored mutations V184I (this position is far enough from the substrate binding site and therefore, may not affect binding) and L222A (this mutation was predicted because of a potential clash between methionine rotamer at 187 and leucine at 184 in the calculation).

We compared the sequence predicted for Phe and DPA and observed that most significant difference between the two sequences was the prediction of two important cavity-forming mutations in the case of DPA: Val261 (Thr251 in *E. coli*) to Gly and Ala314 (Ala294 in *E. coli*) to Gly (Figure IX-3).

These predictions are consistent with the findings of Reshetnikova and coworkers [23] who pointed out that Ala314 and Val261 provide steric hindrance to the binding of larger amino acids (e.g., tyrosine) into the active site of *t*PheRS. Further confidence in the prediction was engendered by the fact that Ala294Gly allows incorporation of an interesting set of para-substituted analogues of Phe. We were encouraged to test whether the additional Thr251Gly mutation would relax the specificity of *e*PheRS sufficiently to allow incorporation of DPA into protein *in vivo.*

As shown in SDS-PAGE analysis of whole cell lysates (Figure IX-5), neither AF-IQ[pQE15] nor AF-IQ[pQE-FS*] yields mDHFR in the negative control cultures (-phe) or in the cultures containing DPA. On the other hand, expression of mDHFR is evident in the AF-IQ[pQE-FS**] culture supplemented with DPA.

MALDI-TOF mass spectrometry analysis showed that molecular weight of mDHFR-2 was increased 307.65 kDa, indicating 81.4% substitution of Phe residues by DPA (mDHFR contains 9 Phe residues). Incorporation of DPA was also confirmed by tryptic peptide digestion analysis of purified mDHFR (Figure IX-6). For mDHFR-wt, two peptides within mass range from 1550 to 1750 Da was observed, which can be assigned to residues 34-47 and 93-106, respectively (Figure IX-6a, both fragments contain 1 phe residue). The corresponding peptides containing analogue 2 (Figure IX-6b) showed additional mass of 42 Da, which is consistent to increased mass of DPA relative to Phe. The chemoselective ligation of hydrazide with ketone-containing protein proceeded sucessfully and western blot using anti-His antibody confirmed that no side reactions occured (Figure IX-7b).

RBIAS – for enhancing protein-substrate interactions

The sequence generated in the first calculation has two important cavity forming mutations which are Val261 to glycine and Ala314 to glycine. Mutation to glycine at position 314 has been previously reported to have enhanced the incorporation of larger Phe analogs that are not incorporated by the wild type synthetase. All the mutations predicted in this calculation gear towards making enough space in the binding pocket for accommodating DPA, but besides van der Waals interactions, we do not see any specific interaction between DPA and the protein.

In an attempt to design specificity in protein-substrate interaction, we developed a program, RBIAS, which enhances the interactions between the substrate and the protein positions. This was achieved by scaling up the pairwise energies between the substrate and the amino acids allowed at the design positions on the protein in the energy calculations. In an optimization calculation where the protein-substrate interactions are scaled up compared to the intra-protein interactions, sequence selection will be biased toward selecting amino acids to be those that have favorable interaction with the substrate.

We performed multiple calculations by scaling up the substrate-protein interactions by factors of 2.0 to 20.0 in increments of 2.0. A scale factor of 4.0 generated an interesting mutation, Val 286 to Gln, which makes a hydrogen bond with the acetyl group at the distal end of DPA. The interaction between DPA and the PheRS in this sequence was enhanced by 2.12 kcal/mol although the complex is destabilized by 12.96 kcal/mol as indicated by the total energy. A bias scale factor of 18.0 generated a new mutation, Val 290 to lysine. We believe this mutation is not important for specificity since lysine at this position is not making significant interactions with DPA. Moreover, polar groups in the core, especially those that are not involved in a salt-bridge or a hydrogen bond may significantly destabilize

proteins. Therefore, we can trade off only some amount of overall protein stability in order to gain specificity between the protein and the substrate.

References

1. Deming, T.J., Fournier, M. J., Mason, T. L., & Tirrell, D. A., Biosynthetic Incorporation and Chemical Modification of Alkene Functionality in Genetically Engineered Polymers. J. Macromol. Sci. Pure Appl. Chem, 1997. **A34**: p. 2143-2150.

2. Budisa, N., Steipe, B., Demange, P. Eckerskorn, C., Kellermann, J., & Huber, R., High-Level Biosynthetic Substitution of Methionine in Proteins by its Analogs 2-Aminohexanoic Acid, Selenomethionine, Telluromethionine, and Ethionine in E. coli. Eur. J. Biochem, 1995. **230**: p. 788-796.

3. Duewel, H., Daub, E., Robinson, V., & Honek, J. F., Incorporation of Trifluoromethionine into a Phage Lysozyme: Implications and a New Marker for Use in Protein 19F NMR. Biochemistry, 1997. **36**(3404-3416).

4. van Hest, J.C. and D.A. Tirrell, Efficient introduction of alkene functionality into proteins in vivo. FEBS Lett, 1998. **428**(1-2): p. 68-70.

5. Sharma, N., Furter, R., Kast, P., Tirrell, D. A., Efficient introduction of aryl bromide functionality into proteins in vivo. FEBS Lett, 2000. **467**(1): p. 37-40.

6. Kiick, K.L., van Hest, J.C., Tirrell, D.A., Expanding the Scope of Protein Biosynthesis by Altering the Methionyl-tRNA Synthetase Activity of a Bacterial Expression. Angew Chem Int Ed Engl, 2000. **39**(2148-2152).

7. Kirshenbaum, K., I.S. Carrico, and D.A. Tirrell, Biosynthesis of proteins incorporating a versatile set of phenylalanine analogues. Chembiochem, 2002. **3**(2-3): p. 235-7.

8. Kast, P., Hennecke, H., Amino acid substrate specificity of Escherichia coli phenylalanyl-tRNA synthetase altered by distinct mutations. J Mol Biol, 1991. **222**(1): p. 99-124.

9. Kast, P., C. Wehrli, and H. Hennecke, Impaired affinity for phenylalanine in Escherichia coli phenylalanyl- tRNA synthetase mutant caused by Gly-to-Asp exchange in motif 2 of class II tRNA synthetases. FEBS Lett, 1991. **293**(1-2): p. 160-3.

10. Ibba, M., et al., Increased rates of tRNA charging through modification of the enzyme- aminoacyl-adenylate complex of phenylalanyl-tRNA synthetase. FEBS Lett, 1995. **358**(3): p. 293-6.

11. Ibba, M., P. Kast, and H. Hennecke, Substrate specificity is determined by amino acid binding pocket size in Escherichia coli phenylalanyl-tRNA synthetase. Biochemistry, 1994. **33**(23): p. 7107-12.

12. Ibba, M. and H. Hennecke, Relaxing the substrate specificity of an aminoacyl-tRNA synthetase allows in vitro and in vivo synthesis of proteins containing unnatural amino acids. FEBS Lett, 1995. **364**(3): p. 272-5.

13. Hamano-Takaku, F.e.a., A mutant Escherichia coli tyrosyl-tRNA synthetase utilizes the unnatural amino acid azatyrosine more efficiently than tyrosine. J. Biol. Chem., 2000. **275**: p. 40324-40328.

14. Doring, V.M., H. D.; Nangle, L. A.; Hendrickson, T. L.; de Crecy-Lagard, V.; Schimmel, P.; Marliere, P., Enlarging the amino acid set of Escherichia coli by infiltration of the valine coding pathway. Science, 2001. **292**: p. 501.

15. Behrens, C.N., J. J.; Fan, X. -J.; Doisy, X.; Kim, K. -H.; Praetorius-Ibba, M.; Nielsen, P. E.; Ibba, M., Development of strategies for the site-specific in vivo incorporation of photoreactive amino acids: p-azidophenylalanine, p-acetylphenylalanine and benzofuranylalanine. tetrahedron, 2000. **56**: p., 9443.

16. Kowal, A.K.K., C.; RajBhandary, U. L., Twenty-first aminoacyl-tRNA synthetase-suppressor tRNA pairs for possible use in site-specific incorporation of amino acid analogues into proteins in eukaryotes and in eubacteria. Proc. Natl. Acad. Sci. USA, 2001. **98**: p. 2268-2273.

17. Rose, K.J., FACILE SYNTHESIS OF HOMOGENEOUS ARTIFICIAL PROTEINS. J Am Chem Soc, 1994. **116**: p. 30.

18. Rideout, D.C., T, Synergism through direct covalent bonding between agents: a strategy for rational design of chemotherapeutic combinations. Biopolymers, 1990. **29**: p. 247.

19. Rideout, D., Self-assembling drugs: a new approach to biochemical modulation in cancer chemotherapy. Cancer Invest., 1994. **12**: p. 189.

20. Mahal, L.K.Y., K. J.; Bertozzi, C. R., Engineering chemical reactivity on cell surfaces through oligosaccharide biosynthesis. Science, 1997. **276**: p. 1125-1128.

21. Shao, J.T., J. P., UNPROTECTED PEPTIDES AS BUILDING-BLOCKS FOR THE SYNTHESIS OF PEPTIDE DENDRIMERS WITH OXIME, HYDRAZONE, AND THIAZOLIDINE LINKAGES. J Am Chem Soc, 1995. **117**: p. 3893.

22. Yarema, K.J.M., L. K.; Bruehl, R. E.; Rodriguez, E. C.; Bertozzi, C. R., Metabolic delivery of ketone groups to sialic acid residues. Application To cell surface glycoform engineering. J. Biol. Chem., 1998. **273**: p. 31168-31179.

23. Reshetnikova, L., Moor, N., Lavrik, O., Vassylyev, D. G., Crystal structures of phenylalanyl-tRNA synthetase complexed with phenylalanine and a phenylalanyl-adenylate analogue. J Mol Biol, 1999. **287**(3): p. 555-68.

24. Stepanov, V.G., Moor, N. A., Ankilova, V. N., Lavrik, O. I., Phenylalanyl-tRNA synthetase from Thermus thermophilus can attach two molecules of phenylalanine to tRNA(Phe). FEBS Lett, 1992. **311**(3): p. 192-4.

25. Dahiyat, B.I. and S.L. Mayo, Protein design automation. Protein Sci, 1996. **5**(5): p. 895-903.

26. Dahiyat, B.I., C.A. Sarisky, and S.L. Mayo, De novo protein design: towards fully automated sequence selection. J Mol Biol, 1997. **273**(4): p. 789-96.

27. Dunbrack, R.L. and M. Karplus, Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat Struct Biol, 1994. **1**(5): p. 334-40.

28. Dunbrack, R.L. and M. Karplus, Backbone-dependent rotamer library for proteins. Application to side- chain prediction. J Mol Biol, 1993. **230**(2): p. 543-74.

29. Gordon, D.B., S.A. Marshall, and S.L. Mayo, Energy functions for protein design. Curr Opin Struct Biol, 1999. **9**(4): p. 509-13.

30. Furter, R., Expansion of the genetic code: Site-directed p-fluoro-phenylalanine incorporation in Escherichia coli. Protein Sci., 1998. **7**: p. 419.

Table IX-1. RBIAS calculations for DPA. A big energy clash between DPA and the binding pocket in the wild type sequence (WT_dpa) indicates why DPA is not incorporated by the wild type synthetase

Sequence	128	130	137	149	178	180	183	184	187	204	222	258	260	261	286	290	294	314	Ligand Energy	Total Energy
WT_phe	E	E	L	W	H	S	Q	V	M	R	L	F	F	V	V	V	V	A	-16.91	-240.71
No bias_phe	E	E	L	W	H	S	Q	I	M	R	A	Y	F	V	L	I	I	A	-16.87	-242.14
WT_dpa	E	E	L	W	H	S	Q	V	M	R	L	F	F	V	V	V	V	A	-103314	51669.81
No bias_dpa	E	E	L	W	H	S	Q	I	M	R	L	Y	F	G	L	I	V	G	-21.40	-225.13
Bias2.0	E	E	L	W	H	S	Q	I	M	R	L	Y	F	G	L	I	V	G	-21.40	-225.13
Bias4.0	E	E	L	W	H	S	Q	I	M	R	A	Y	F	G	Q	L	V	G	-23.52	-212.17
Bias6.0	E	E	L	W	H	S	Q	I	M	R	A	Y	F	G	Q	L	V	G	-23.52	-212.17
Bias8.0	E	E	L	W	H	S	Q	I	M	R	A	Y	F	G	Q	L	V	G	-23.52	-212.17
Bias10.0	E	E	L	W	H	S	Q	I	M	R	A	Y	F	G	Q	L	V	G	-23.52	-212.17
Bias12.0	E	E	L	W	H	S	Q	I	M	R	A	Y	F	G	Q	L	V	G	-23.52	-212.17
Bias14.0	E	E	L	W	H	S	Q	I	M	R	L	Y	F	G	Q	L	V	G	-23.63	-209.34
Bias16.0	E	E	L	W	H	S	Q	I	M	R	L	Y	F	G	Q	L	V	G	-23.63	-209.34
Bias18.0	E	E	L	W	H	S	Q	I	M	R	A	Y	F	G	Q	K	V	G	-23.96	-198.83
Bias20.0	E	E	L	W	H	S	Q	I	M	R	A	Y	F	G	Q	K	V	G	-23.96	-198.83

Figure IX-1. Sequence alignment between *E.coli* and *T. thermophilus* PheRS a subunits. Conserved residues are colored in red and conservative replacements are indicated by blue + symbol.

```
Query:    61
XXXXXXXXXXXXXXXXXXXXXXXXVDVSLPGASLFSGGLHPITLMERELVEIFRALGYQ 120
                          +DVSLPG  + +GGLHP+T     +   F  LG+
Sbjct:    71
VQQALNAAKAELESAALNARLAAETIDVSLPGRRIENGGLHPVTRTIDRIESFFGELGFT 130

Query:   121
AVEGPEVESEFFNFDALNIPEHHPARDMWDTFWLTGEGFXXXXXXXXXXXXXXXXXTHTS 180
   GPE+E ++ NFDALNIP HHPAR   DTFW                        T TS
Sbjct:   131
VATGPEIEDDYHNFDALNIPGHHPARADHDTFWFD---------------TTRLLRTQTS 175

Query:   181
PMQVRYMVAHTPPFRIVVPGRVFRFEQTDATHEAVFHQLEGLVVGEGIAMAHLKGAIYEL 240
 +Q+R M A  PP RI+ PGRV+R +  D TH  +FHQ+EGL+V   I+  +LKG +++
Sbjct:   176
GVQIRTMKAQQPPIRIIAPGRVYRNDYDQTHTPMFHQMEGLIVDTNISFTNLKGTLHDF 234

Query:   241 AQALFGPDSKVRFQPVYFPFVEPGAQFAVWWPEG 274
              +  F  D ++RF+P YFPF EP A+  V    G
Sbjct:   235 LRNFFEEDLQIRFRPSYFPFTEPSAEVDVMGKNG 268
```

Figure IX-2. Residues involved in the binding site of PheRS. Phenylalanine in shown in magenta and the anchor residues are labeled in bold.

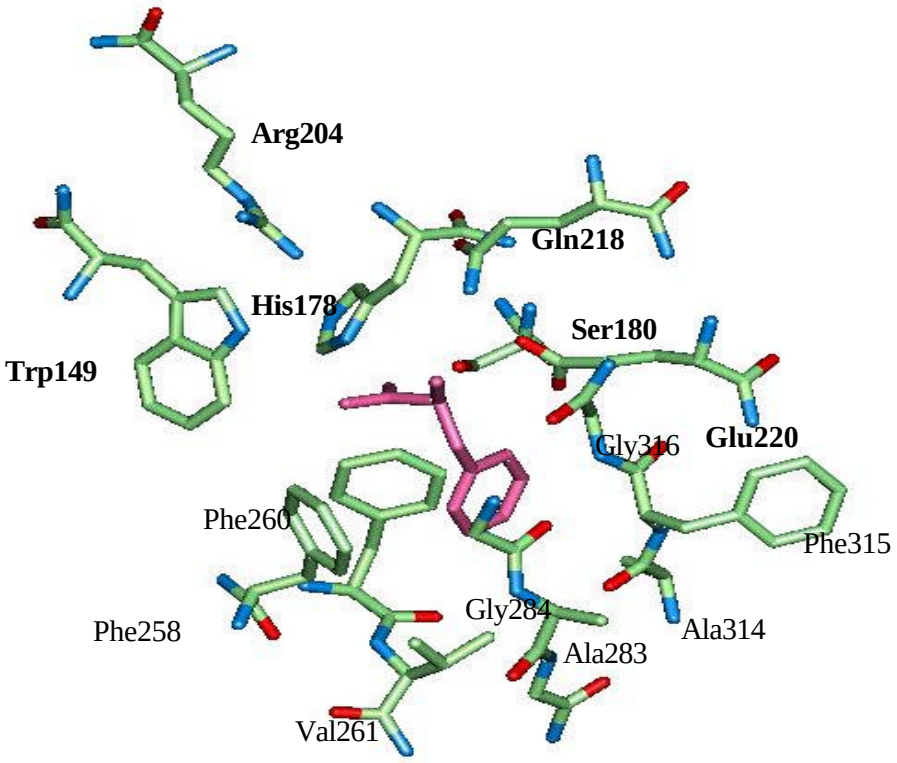
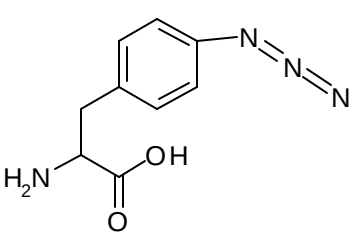

Arg204
Gln218
His178
Ser180
Trp149
Gly316
Glu220
Phe260
Phe315
Gly284
Phe258
Ala314
Ala283
Val261

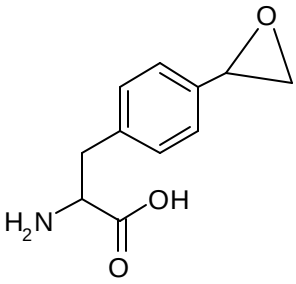
Figure IX-3. The redesigned binding site of PheRS showing the orientation of DPA in bound conformations and the two cavity forming mutations.


V261G
A314G

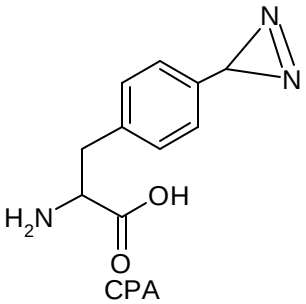
Figure IX-4. Phenylalanine analogs considered as interesting because of their unique chemical functionalities.



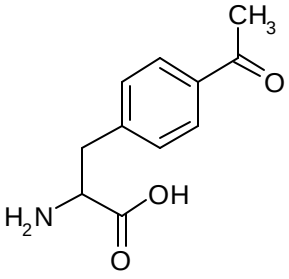
APA



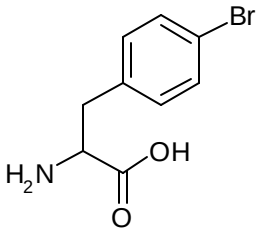
BPA



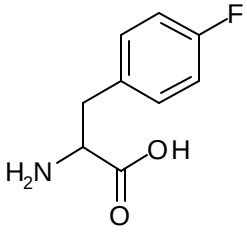
CPA



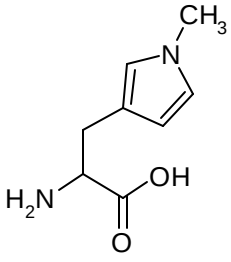
DPA



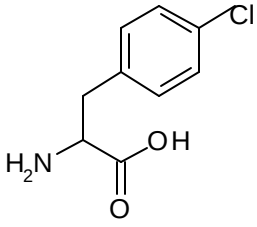
BRP



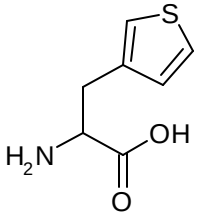
FPA



NPA



CLP



SPA

Figure IX-5. SDS-PAGE of cell lysates of different expression systems with 4 hr post-induction with 1 mM IPTG demonstrated *e*PheRS** allows incorporation of 2 into protein *in vivo.* Concentration of Phe(**1**)=20mg/l; DPA(**2**)=250mg/l.

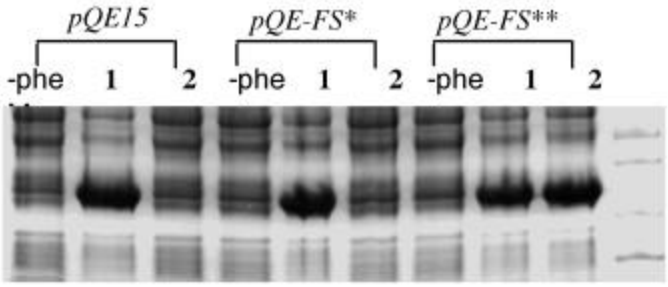

pQE15
pQE-FS*
pQE-FS**
-phe
1
2
-phe
1
2
-phe
1
2

Figure IX-6. MALDI TOF mass spectra of tryptic peptides digested from mDHFR expressed in media supplemented with Phe (a) or DPA (b).

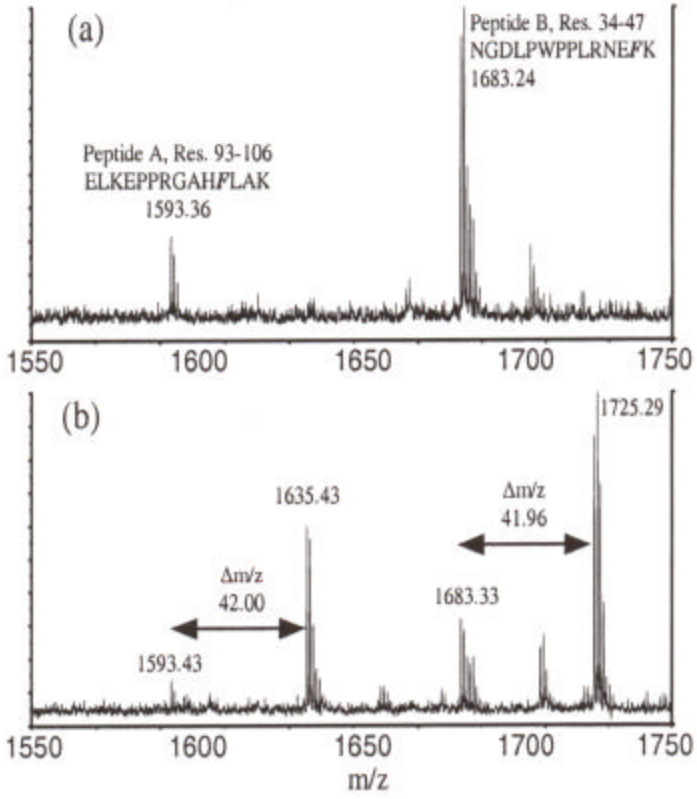
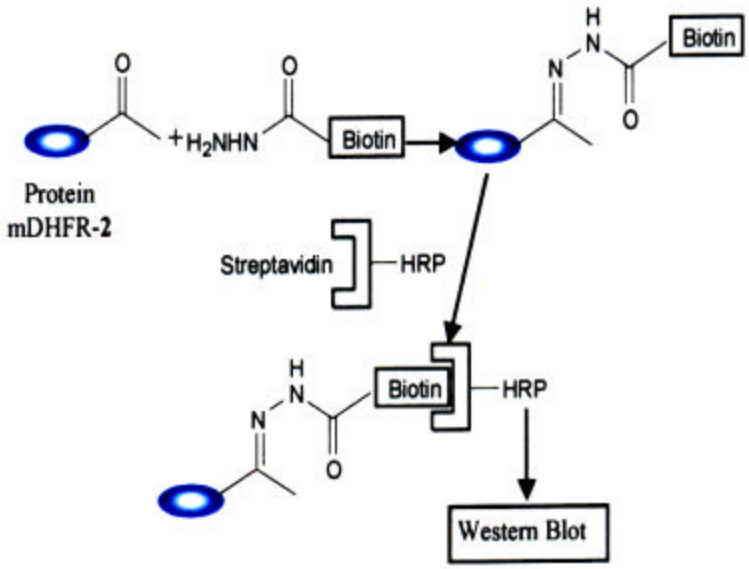
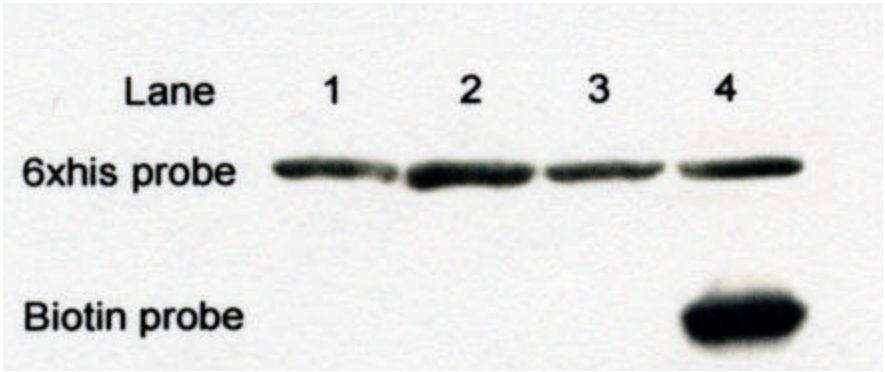

(a)
Peptide B, Res. 34-47
NGDLPWPPLRNEFK
1683.24
Peptide A, Res. 93-106
ELKEPPRGAHFLAK
1593.36
1550
1600
1650
1700
1750
(b)
1725.29
1635.43
Δm/z
41.96
Δm/z
42.00
1683.33
1593.43
1550
1600
1650
1700
1750
m/z

Figure IX-7. Western blot for the detection of chemoselective formation of hydrazone linkage. (a) Modified protein was treated with biotin hydrazide (BH), stained with HRP conjugated streptavidin and analyzed by western blot. (b) Western blot analysis of the products. Lane 1: mDHFR-wt + buffer; Lane 2: mDHFR-DPA + buffer; Lane 3: mDHFR-wt + BH; Lane 4: mDHFR-DPA + BH.



(a)

Phe + buffer	DPA + Buffer	Phe + biotin hydrazide	DPA + biotin hydrazide



(b)

Chapter 10

An Experimental and Computational Approach for Designing a Conformation Switch in Proteins

The entropy profile calculations in this study were generated by Christopher Voigt.

Abstract

In this study we have combined methods from mean-field theory (derived from statistical mechanics) and computational protein design to understand protein sequence space and eventually design a protein sequence capable of switching between two completely different protein folds – a ß-sheet fold to an a-helical fold. It has been suggested that switch sequences are evolutionary bridges that serve as intermediates in the pathway for the evolution of new folds, as evolutionary end points in the development of allosteric systems, or as hazardous dead ends, as in protein misfolding diseases. An important determinant for switching conformation in proteins/peptides is the external perturbation that induces the switch. In our study, we are attempting to achieve this by using metal binding as an external force to direct the sequence towards the a-helical conformation over the ß-sheet conformation. To design an amino acid sequence capable of adopting both the conformations, we have used a strategy of first determining positions on both the proteins (protein G and engrailed homeodomain) that are highly tolerant to substitutions to common residues. This was done by using a mean-field approach to identify the conserved and mutable amino acids on the two proteins. In the next step, the mutable positions on both the folds were designed using a protein design algorithm to select for identical amino acids and consequently, to bring the two conformations closer in sequence space.

Introduction

Localized conformation changes in secondary structure are physiologically important for the correct functioning of many proteins. The native forms of such proteins are metastable and this property is often critical to their biological function. Mutational analysis and structural examination of metastable proteins have revealed unusual interactions, such as side-chain over-packing, buried polar groups, and cavities (Bullough *et al.*, 1994; Chen *et al*,. 1998). These structural defects are likely to be the design principle of metastable native proteins to regulate conformation changes. Some examples of such metastable proteins are the plasma serpins (serine protease inhibitors) (Huber *et al.*, 1989), the spring-loaded structure of the membrane fusion protein of influenza virus (Bullough *et al.*, 1994; Carr *et al.*, 1993), heat shock transcription factors (Orosz *et al.*, 1996), and G-protein EF-Tu complex (Abel *et al.*, 1996). It has been suggested that the conformational switches exhibited by these structures may represent a general mechanism to mediate activation in metastable proteins.

An important determinant for switching conformation in proteins/peptides is the external environment that induces the switch. For example, up to heptapeptide long sequences have been identified that adopt either a a-helical or ß-sheet conformation, depending on the protein in which they are found (Mezei *et al.*, 1998). The secondary structure of such chameleon sequences is determined by their tertiary environment. The dual nature of an 11-residue switch peptide was experimentally demonstrated (Minor and Kim, 1996) by placing it in two different regions within a small molecule, the ß1 domain of Protein G (Gronenborn *et al.*, 1991). This indicated that context is an important factor governing the secondary structures of stretches of amino acid sequences in a protein.

The structure adopted by peptides and proteins also depends on external environmental conditions such as pH, temperature, and the ionic strength of the solvent. Two 17-residue peptides display switching between a-helix and ß-sheet conformations when these solvent conditions are varied (Cerpa *et al.*, 1996). In another example, a designed variant of Rop protein converts from a-helical to a fibrillar ß-sheet form when the pH of the solution is changed (Dalal and Regan, 2000).

Conformation changes in proteins are also seen when proteins bind to other molecules. A segment of MATa2, a DNA binding protein, converts from a-helical to ß-sheet conformation when it binds to DNA (Tan and Richmond, 1998). The folding state of an artificially designed peptide has been shown to be regulated by cofactor binding. A 17-amino acid peptide can be prevented from forming ß-sheet aggregates by binding heme to it, which facilitates the formation of an a-helix tetramer (Sakamoto *et al.*, 1999). In calmodulin, a peptide binding protein, the structural flexibility of the central a-helical tether is believed to be an essential element in the calcium-dependent recognition of target peptides (Yap *et al.*, 1999). An interesting example of a designed conformation switch has been demonstrated by introducing a single mutation on a surface position of the arc repressor protein (Cordes *et al.*, 2000). The designed arc repressor protein is able to adopt both the ß-sheet and a-helical folds with the ß-sheet form being stabilized by DNA binding.

Conformation switching has been proposed to occur during protein folding. A transient population of a non-native a-helical intermediate has been observed in the folding pathway of a predominantly ß-sheet protein, ß-lactalbumin (Hamada *et al.*, 1996). A number of protein misfolding diseases are associated with protein conformation change. In amyloid diseases, fibril formation is a result of the conversion of soluble

protein into regular ß-sheet aggregates (Cohen *et al.*, 1999; Harrison *et al.*, 1997). It has been proposed that all proteins have the potential to convert to such misfolded and aggregated assemblies under appropriate conditions (Chiti *et al.*, 1999; Gross *et al.*, 1999).

It has been suggested that switch sequences are evolutionary bridges that serve as intermediates in the pathway for the evolution of new folds, as evolutionary end points in the development of allosteric systems, or as hazardous dead ends, as in protein misfolding diseases. Predictions of such bridges have also emerged from theoretical lattice models of protein folding (Bornberg-Bauer *et al.*, 1997; Bornberg-Bauer and Chan, 1999). Besides providing an insight into some fundamental protein evolution questions and sequence-structure relationships, understanding and designing a protein conformation switch can have a tremendous impact in the biotechnology industry for creating genetically engineered protein biosensors.

In this study, we have attempted to design a sequence capable of switching conformation between two completely different structures – the fold adopted by protein G (PG) and that seen in engrailed homeodomain. We have tried to achieve this by using metal binding as the external perturbation to direct the sequence towards the a-helical (engrailed homeodomain) conformation over the ß-sheet (PG) conformation. To design an amino acid sequence capable of adopting both the conformations, we have used a strategy of first determining positions on both the proteins that are highly tolerant to substitutions to common residues. Using a mean-field approach, the conserved and mutable amino acids on the two proteins were identified. Several positions on the proteins were mutated to histidines; these were selected such that they would collectively coordinate a metal ligand in the engrailed structure but not in the PG structure. The

mutable positions were designed using our protein design algorithm, ORBIT, to select for common amino acids and consequently, to bring the two conformations closer in sequence space. The design strategy is based on the assumption that if a designed protein sequence that folds spontaneously into PG is close in energy to the engrailed structure, then addition of metal could facilitate a conformation switch to the engrailed fold.

Results and Discussion

Model system

The two proteins that we have chosen for this study are the ß1 domain of PG and a variant of engrailed homeodomain, SC1. SC1 is a redesigned variant of engrailed homeodomain with 29 core and surface positions mutated from the wild type protein (Morgan, 2000). Both are compact globular proteins and are small enough to attempt experimental characterization of a number of hybrid sequences and to monitor the change in the structural properties of the sequences as they diverge from their native sequences towards a mutual sequence. Both proteins follow a simple two-state folding transition and are easy to express *in vivo*. The sizes of the proteins are comparable ? PG has 56 amino acids, while the engrailed homeodomain variant (SC1) has 51 amino acids. SC1 has a T_m of 91 ºC that is in the same range as that of PG (86 °C). Regardless of these similarities, the conformations of the two proteins are very different. SC1 is an all-helical protein, while PG has a well-structured ß-sheet with a central helix packed against it. The circular dichroism (CD) signals of both the structures are very distinct and changes in the structures can be easily followed by observing the changes in their CD signals.

Designing metal binding sites

Metal binding sites are often formed from amino acids that are distant in the primary sequence but are brought into spatial proximity in the folded protein structure. Identifying key features for metal recognition and modeling the scaffold provided by the protein for metal site design is difficult and complicated. Instead, we have adopted a simpler approach where protein secondary structure provides the rigid framework necessary for metal recognition. This approach has been used in constructing metal binding sites by placing two histidines as His-X3-His on exposed positions on the helices (Todd *et al.*, 1991). This motif forms part of the metal coordination site in a number of metalloproteins, including zinc-finger proteins (Berg, 1988), thermolysin (Holmes and Matthews, 1982), and hemocyanin (Volbeda and Hol, 1989). The His-X3-His metal-binding site engineered into the surface of the protein confers a unique affinity for forming complexes of metals bound to solid supports or soluble polymers (Suh *et al.*, 1991). In this orientation, the e-nitrogens of the imidazole groups from both histidines can coordinate a single metal. Using this simple geometric consideration, surface *i, i+4* positions of the helices on SC1 can be selected to study the effect of metal binding.

Six metal binding sites were designed on SC1 on the surface positions of the C-terminal and the N-terminal helices. Histidines were substituted on the surface residues in *i* and *i+4* positions to get His-X3-His geometry that is capable of chelating metal. Three such sites were created on the C-terminal helix (SCCH1, SCCH2, SCCH3) and three more on the N-terminal helix (SCNH1, SCNH2, SCNH3) (Figure X-1). A total of six proteins were created and tested for a change in stability in the presence of $NiCl_2$ (Figure X-2). SC1 was used as a control since it does not have a His-X3-His designed

site. Five of the six His-X3-His proteins showed a gain in stability in the presence of metal (Table X-1). Variants that showed the maximum gain in stability are SCNH1 and SCCH3 with a stabilization of greater than 1.0 kcal mol^{-1} in the presence of nickel. This result can be explained by the fact that in these two proteins the metal binding occurs near the C and N termini of the protein. The termini of proteins are generally more disordered than the internal positions. Metal binding increases the stability of these molecules by stabilizing the frayed ends of the protein to a well-defined helical structure. We selected positions 5 and 9 (SCNH1 metal-binding site) and positions 46 and 50 (SCCH3 metal-binding site) to be the sites that we expected would trigger the conformation switch on chelating metal.

Sequence alignment

To make the design strategy simpler, we aligned the sequences of the two molecules so that their existing similarities were maximized. The simplest similarities to look for are sequence identity and binary pattern. There are six possible ungapped alignments between SC1 and PG. Selection of the best frame can be determined by minimizing binary pattern mismatches and maximizing identical residues. Binary pattern is an important consideration because it is a key determinant in defining the topology of a protein fold (Dill *et al.*, 1995). Based on the best alignment, positions on PG that correspond to the selected histidine positions of SC1 need to be analyzed. This analysis is important because the corresponding histidine positions on the PG structure must be solvent accessible and mutating them to histidines should not significantly destabilize PG.

SC1 has 51 residues while PG has 56. Hence, there are six possible ways SC1 and PG can be aligned with each other in an ungapped manual alignment (Figure X-3a). Every position on both the proteins was classified as core, surface or boundary using a residue classification program as described previously (Dahiyat and Mayo, 1997). The boundary positions were further reclassified as core and surface. The wild type binary pattern was used to reclassify PG and the binary pattern of the B7 molecule (a boundary redesigned variant of SC1) was used to reclassify SC1 (Marshall and Mayo, 2001). The number of identical residues and the number of binary pattern mismatches between the two sequences were noted for each alignment. Alignment 1 was selected as the best alignment since it has the least number of binary pattern mismatches and the highest number of identical residues (Figure X-3a). In this alignment, the histidine positions of SCNH1 and SCCH3 correspond to positions 10 and 14, and positions 51 and 55, respectively, on the PG sequence. All four positions are solvent-accessible surface positions. Positions 14, 51 and 55 are on the ß-strands and position 10 is on the ß-turn between strand 1 and strand 2 (Figure X-4). In this alignment, the central helix of PG can be almost completely aligned with the central helix of SC1 so that the conformation switch is designed primarily to occur in the ß-sheet region, converting the two ß-hairpins (two anti-parallel ß-strands connected by a turn) to two a-helices (Figure X-3b).

Position 10 is a glycine in the native sequence of PG and we were concerned that replacing it with a histidine may be highly unfavorable for the molecule because a histidine may not be compatible with the torsional angles of the backbone at this site (phi 177.87º, psi 178.41º). A point mutant, G14H, shows that replacing glycine with a histidine is a neutral mutation and the mutant has the same thermal stability as the wild type (Figure X-5). A PG variant (PG4H) was made that has histidines at positions 10, 14,

51 and 55, and this molecule was tested for thermal stability (Figure X-5). Another molecule (SC4H) that has the histidine mutations of both SCCH3 and SCNH1 was also created and tested for stability. CD analysis indicates that SC4H is a well-folded and stable molecule (experimental data not shown).

SC4H and PG4H became our starting molecules that we have used to generate the first two entropy profiles. In both calculations, the identities of positions 10, 14, 51 and 55 on PG, and, positions 5, 9, 46 and 50 on SC1, are fixed to be histidines. The identities of five other residues that are identical in alignment1 (positions 12, 13, 19, 42 and 47 on PG and positions 7, 8, 14, 37, and 42 on SC1) were held fixed to wild type identities in further calculations. Residues 1 to 5 in PG were also fixed as wild type identities since there are no corresponding amino acids for them in SC1 in alignment 1.

Selecting mutable paired positions on PG and SC1

Many proteins maintain their native structures while undergoing single and double mutations at many different sites. The sequences that fold into a particular native fold form neutral networks that percolate through the sequence space. A neutral net is defined as a collection of unique sequences that are interconnected by single-point mutants and encode for the same native structure (Bornberg-Bauer and Chan, 1999). An important question asked in understanding sequence-structure relationships is how close do the neutral networks of the two structures come towards each other. In other words, how similar can sequences be that fold into different structures? As a “rule of thumb”, if two natural proteins have 30% or greater sequence identity, one has confidence that they share the same fold (Sander and Schneider, 1991). On the other hand, this assumption is challenged in designed proteins, where it has been shown that two sequences with 50%

sequence identity fold to two completely different structures (Dalal *et al.,* 1997). Computational simulations support an even stronger claim that sequences that fold into two completely different native structures need not differ by more than a few crucial amino acids (Babajide *et al.,* 2001). This indicates that neutral sets form extensive neutral networks that make them suitable for efficient protein evolution. Empirical evidence for functional neutrality in protein space is indeed observed (Martinez *et al.*, 1996). In some cases seemingly unrelated sequences have essentially the same fold (Holm and Sander, 1997; Murzin *et al.*, 1996). In fact, it is proposed that only some information in a protein sequence is necessary to specify a fold, with the rest just stabilizing that specified structure (Lattman and Rose, 1993; Rose and Creamer, 1994).

In this study, we have tried to explore the possibility of to designing a single sequence to have more than one conformation. It has been proposed that a sequence folds to a single unique conformation if there exists an energy gap between the native state and any other possible conformation, including the unfolded state (Karplus and Sali *et al.*, 1995); (Bryngelson *et al.*, 1995). This theory suggests that a particular sequence specifies a single fold. However, examples of sequences that can adopt multiple conformations described earlier indicate that distinct protein folds need not be isolated islands in a sequence space but can be linked by evolutionary bridges where multiple native conformations coexist (Cordes *et al.*, 2000).

In designing a sequence that can convert from PG to the SC1 conformation on binding to metal, we attempted to converge the sequences as much as possible towards a common sequence. At the same time, we wanted make sure that the native structures of both the proteins were maintained. This is possible if we design only the positions that are tolerant to mutations while keeping the critical residues the same as the starting

sequence. The critical residues are expected to maintain the correct fold of the protein. The rationale for our design is that if we have a sequence that is highly biased towards a metal binding variant of SC1, but folds to the PG structure due to the presence of only a few conserved residues that maintain the PG structure, the molecule is going to be unstable enough that in the presence of metal it will change to the SC1 structure. We expect that metal binding will increase the propensity of the sequence to adopt a helical conformation since metal bound SC1 variants have been shown to stabilize the helical structure (described later in the chapter). Our design strategy includes the following steps:

1. Identifying the two sets of positions: those that are highly mutable and those that are highly conserved on both structures.
2. Redesigning as many positions as possible on both structures, starting with the most mutable positions and changing to mutually acceptable amino acids, while making sure that the redesigned proteins fold to their native structures.

The tolerance of protein structures to mutations can be calculated using a mean-field approach for recognizing the high entropy positions or sites (Voigt *et al.*, 2001). These sites are classified as highly tolerant to amino acid substitutions. The site entropy is determined by the variability of the amino acid identity at a given site among the sequences that fold to the native conformation. The amino acid probabilities are calculated as the sum of the amino acid's rotamer probabilities, as determined by the mean-field theory. The probabilities of the existence of all allowed amino acids at all positions are tabulated and condensed into site entropy. Therefore, site entropy is a measure of the number of amino acid substitutions that can be made at each position without disrupting the overall protein structure. Sites with low entropy are intolerant to

mutations and are classified as conserved residues, whereas the positions with high entropy are tolerant to mutations and are selected for design. A tabulation of the entropy at each position produces an entropy profile for the protein. By comparing the entropy profiles of PG and SC1, we can determine the positions that can be selected for design. We are using the entropy profiles as a guide for designing the switch sequence. The entropy profile is dependent on the force field used for energy calculations and the amino acids allowed at each site.

Generating entropy profiles

Entropy profile I. To select high entropy positions for design, we generated entropy profiles of PG and SC1. In our first try, we generated profiles by allowing only two amino acids at each position on both the molecules, one from the PG sequence and one from the SC1 sequence. The positions showing the highest entropy on both the structures were the common surface positions. The positions with the lowest entropy were the positions that showed binary pattern mismatches in the sequence alignment. Since we wanted to bias the sequence as much as possible towards SC1, we selected only the high entropy positions on PG and mutated them to amino acids on the SC1 sequence in two sets; these were SW2 and SW4 (Figure X-6).

We were unable to express both SW2 and SW4. Looking at the structures closely, it was obvious why these molecules were probably destabilized to a point that they could not be expressed. The sites on PG that show the highest entropy are predominantly the ß-sheet surface positions and the helix capping positions (Figure X-7). The force field used to generate the entropy profile does not capture helix capping unless we set it to do so. This can be achieved by scaling up a term in the force field that selects

for hydrogen bonds between the side chain and the backbone. Also, the force field does not take secondary structure propensity of the amino acids into consideration unless this term is turned on. As a result, all the ß-sheet surface positions on PG that are predominantly made of high ß-sheet propensity residues were selected as high entropy positions, indicating that these residues could be substituted by the corresponding residues from SC1.

Entropy profile II – using secondary structure propensity and scaling up the remote backbone H-bond scale factor. To overcome the problems in the first entropy profile, a new profile was generated by including a secondary structure propensity term and a term to accentuate hydrogen bonds between the side chain and the backbone in the force field. Some of the helix capping positions and ß sheet surface positions were now calculated to be low entropy positions (Figure X-8).

Entropy profile III – allowing more than two amino acids. One of the biggest concerns in designing the two structures towards a mutual sequence is that allowing only two amino acids at a given position is an extremely limited search of the sequence space between the two structures. And, as a result, we may not be able to capture the "switch" sequence. An easy way to overcome this would be to allow all amino acids at each position to generate the entropy profile. However, a simple entropy profile generated by allowing all amino acids at each position will give us the high entropy positions on both molecules but will not give any information on the possibility that the two linked positions can be mutated to a common amino acid (Figure X-9a and X-9b). Also, some corresponding positions that have high entropies may have tolerance for very different sets of amino acids. For example, a high entropy core position in one molecule corresponding to a high

entropy surface position on the other molecule will have a very low probability of tolerating common amino acids.

The tolerance of the paired positions on the two structures to substitution by common amino acids can be extracted from the probabilities calculated using mean-field theory. The probabilities of all amino acids, except for histidine, proline and cysteine, at each paired position were calculated and tabulated (Figure X-11). This tabulation along with the joint entropy profile (Figure X-10) provides us with the essential information we need to run the design calculations. First, it gives us information on which positions have high entropy in both structures. Second, we can now choose from paired positions on both structures that not only have high entropy, but more specifically, a higher probability for tolerating a common amino acid mutation (Figure X-10). Third, it gives us a set of amino acids that are allowed at each paired position. This information is very useful for subsequent design calculations because reducing the number of allowed amino acids at each linked position will make the calculations run much faster.

Experimental analysis of the designed SC1 and PG variants: Using the information from the joint entropy profile four variants of PG (PG50, PG45, PG40 and PG35) and four variants of SC1 (SC50, SC45, SC40 and SC35) were designed using ORBIT. They were tested for secondary structure content and stability in the presence and absence of $NiCl_2$ and copper iminodiacetic acid (Cu(II)IDA). Metals with vacant coordination sites bind to ligating atoms exposed on protein surfaces. However, since nickel has six vacant coordination sites and the nickel binding sites on the proteins are designed as di-histidine binding sites, $NiCl_2$ in the solution caused most of the variants to aggregate because of non-specific chelation of the metal. To address this issue, we used Cu(II)IDA as the chelating agent. We expected it to limit non-specific binding since it has only two vacant

coordination sites. Moreover, it is a five-membered ring that has been shown to have a very large chelating effect mainly attributed to a low entropic cost on bi-dentate binding (Kellis, 1991; Suh, 1991).

While none of the engrailed mutants (SC50, SC45, SC40 andSC35) showed a significant change in structure on binding to Cu(II)IDA as indicated by their wavelength spectra, dramatic changes were observed in the spectra of the PG mutants (PG50, PG45, PG40 and PG35) in the presence of metal (Figures X-15 to X-18). Moreover, binding to Cu(II)IDA also increased the thermal stabilities of the engrailed mutants, while it appears to have to destabilized the PG mutants. PGWT and SC1, the starting structures, do not have any designed histidines and therefore, did not show a change in structure or stability in the presence of metal (Figure X-13). However, when the divalent histidine sites were engineered into these structures to form PG4H and SC4H56, binding to copper increased the stability of one (SC4H56) and decrease the stability of the other (PG4H). These changes are marginal, and based on the wavelength spectra, indicate that metal binding does not cause a structural change in either molecule. This implies that designing the divalent histidines on the ß-sheet surface of PG is not sufficient to cause a structural change. The striking changes that we observe in the PG variants are because these sequences are designed to be closer in sequence space to SC1 fold.

PG50 shows reduced ß-sheet content in the presence of copper; however, it still appears folded. Conversely, PG45 and PG35 both appear to unfold on adding copper. PG40 is unfolded without the metal, but on adding copper, it appears to be partially folded. PG35 has a 41% sequence identity with SC35 (Table X-2). At such high sequence identities, one would expect them to have the same structure. But their response metal binding indicates otherwise. PG35 completely unfolds on binding to metal, while

SC35 gains stability. Further characterization of these molecules with NMR could be useful in understanding these characteristics better.

1D NMR spectra of PG50, PG45, PG35, SC50, SC45, SC40, and SC35 were obtained without metal at pH 7.6 and at a temperature of 5° C. PG35 is folded but indicates molten globule-like features. This is not surprising since, as we link more positions on both the molecules to be identical, we are moving them closer in sequence space and thereby, moving them further away from their own global minima.

Materials and Method

Gene construction, protein expression, and purification: Genes for the metal binding variants of SC1 (SCCH1, SCCH2, SCCH3, SCNH1, SCNH2, SCNH3, SC4H) were constructed by inverse PCR on the SC1 gene, which was cloned in the pET-11a (Novagen) vector. The two PG variants (G14H and PG4H) were also made by inverse PCR on the wild type PG gene, which was also cloned into the pET-11a vector. SW2, SW4, SC50, SC45, SC40, SC35, PG50, PG45, PG40 and PG35 genes were synthetically constructed using recursive PCR and cloned into a pET-11a variant. Sequences for all constructs were confirmed by DNA sequencing. All proteins were expressed in *E. coli* BL21 (DE3) hosts (Stratagene) by IPTG induction and proteins were isolated from the cells using a freeze-thaw protocol followed by purification by HPLC on a reverse phase column. Protein masses were determined by mass spectrometry.

Circular dichroism (CD): CD data were obtained on an Aviv 62A DS spectropolarimeter equipped with a thermoelectric cell holder and an autotitrator. Gain in stability (??G) on metal binding was experimentally measured for SC1 and all the mutants of SC1 containing the His-X3-His site by measuring the difference in the free energy of

unfolding ($?G_u$) of the proteins in the presence and absence of $NiCl_2$. $?G_u$ was measured by guanidinium chloride denaturation of protein samples at pH 7.5 at 25 °C (Figure X-2 2). The protein concentrations were 5 µM in 10 mM HEPES buffer. A 1:100 ratio of protein to $NiCl_2$ was used for denaturations done in the presence of metal. Data was acquired at a wavelength of 222 nm every 0.2 M from 0.0 M to 6.0 M GdmCl using a mixing time of nine minutes and averaging time of 100 seconds. $?G_u$ was obtained from the chemical denaturation data assuming a two-state transition and using the linear extrapolation model. Wavelength scans and thermal denaturation data for the PG and SC1 variants were also obtained in the presence of 1 mM copper(II)iminodiacetic acid (Cu(II)IDA). Cu(II)IDA crystals were obtained by following a procedure decribed by Roman-Alpiste and coworkers. (Roman-Alpiste *et al.*, 1999).

Thermal denaturation data for G14H and PG4H mutants were obtained using samples containing 50 µM protein in 50 mM potassium phosphate buffer at pH 5.5. Thermal denaturations were performed by increasing the temperature from 1 °C to 99 °C with a step size of 2 °C, an equilibration time of 90 seconds, and an averaging time of 30 seconds. Melting temperatures were calculated by evaluating the maximum of a d?/dT versus T plot.

Nuclear magnetic resonance studies:

1D ^1H NMR spectra were obtained on a Varian 600 MHz spectrometer using a Varian triple resonance probe. All samples except PG35 and SC35 contained 250 µM protein and 50 mM sodium phosphate in a 10% ^2H_2O buffer at pH 7.6. PG35 and SC35 were not soluble at high concentrations; therefore, lower protein concentrations were used for their analysis.

Structural entropy calculations:

Force field and rotamer library: The energy term used in ORBIT consists of two contributions – the rotamer-backbone energies $e(i_r)$ and the rotamer-rotamer energies $e(i_r,j_s)$,

$$E = \sum_{i=1}^{N} e(i_r) + \sum_{i=1}^{N-1} \sum_{j>1}^{N} e(i_r, j_s) \qquad (1)$$

where N is the number of residues and i_r is rotamer r at position i. The total energy for a rotamer is the sum of the van der Waals, hydrogen bonding, electrostatic, secondary structure propensity and atomic solvation energies (Dahiyat and Mayo, 1996,1997). The parameters for these potentials are described in previous work. We use Dreiding force field parameters for the atomic radii and internal coordinate parameters (Mayo *et al*., 1990). The van der Waals energies are modeled using a 6-12 Leonard-Jones potential with an additional 0.9 scale factor applied to the atomic radii to soften the lack of flexibility implied by using a fixed backbone model. All rotamer-rotamer and rotamer-backbone energies are calculated and stored prior to mean-field calculations.

The rotamer library used in mean field calculations is our expanded version of the backbone dependant library described by Dunbrack and Karplus (1993, 1994). Rotamers that interact with the backbone with energies greater than 20 kcal mol^{-1} are eliminated from the calculation.

Mean-field theory: The mean field solution for equation (1) is

$$e_{mf}(i_r) = e(i_r) + \sum_{j=1}^{N} \sum_{s=1}^{K_j} e(i_r, j_s) p(j_s) \qquad (2)$$

where $e_{mf}(i_r)$ is the mean-field energy felt by rotamer r at position i and K_j is the total number of rotamers at residue j (Koehl and Delarue, 1994, 1996; Lee *et al.*, 1994). The probability vector $p(j_s)$ is calculated at some temperature T using Gibbs equation

$$p(js) = \frac{e^{-\beta e_{mf}(j_s)}}{\sum_{s=1}^{K_j} e^{-\beta e_{mf}(j_s)}} \qquad (3)$$

where $ß = 1/k_BT$, where k_B is the Boltzmann constant. The algorithm iterates between equations (2) and (3) until self-consistency is achieved.

Entropy profiles: The entropy can be calculated from the probability distribution of allowed amino acid substitutions (Voigt *et al*., 2001). The site entropy is calculated by determining the variability of amino acid identity among sequences consistent with an energy. It is calculated from the probability $p_i(a)$ that an amino acid identity i_a exists at site i,

$$s_i(F) = -k_b \sum_{a}^{A} p_i(a) \ln p_i(a)$$

where A is the total number of amino acids and k_b is chosen to be 1. The amino acid probability, $p_i(a)$ is calculated as the sum of amino acid's rotamer probabilities as determined by the mean-field theory.

Design calculations using ORBIT based on joint entropy profile: The joint entropy profile was used to select positions for design on both PG and SC1. We used cutoffs (0.5, 0.45, 0.40 and 0.35) as indicated by the dotted lines in Figure X-10 to select for design positions. All the positions that are above the cutoff were selected on both the molecules and were forced in the design protocol to be identical amino acids. The positions that were not designed were allowed to maintain their wild type identities but were varied in their rotameric conformations. The designed histidine sites, positions 10,

14, 51 and 55 on SC4H56 and PG4H were fixed to be histidine in all calculations. The amino acids allowed at the paired design positions were based on the combined amino acids probabilities calculated for those positions. As the cutoff was lowered from 0.5 to 0.35, the number of identical amino acids on both the structures increased. The SC1 variants made by this method were SC50, SC45, SC40 and SC35 and the PG variants were PG50, PG45, PG40 and PG35.

Selecting common amino acids at the high entropy positions: Design calculations were run so that the corresponding high entropy positions on the two structures are forced to have the same amino acid identity. This is achieved by specifying these positions and linking them in our protein design algorithm, ORBIT. As the optimization calculation runs, the linked positions can be forced to maintain amino acid symmetry by setting a high penalty energy for pairs of amino acid combinations that break the symmetry at the linked positions. Typically, the penalty energy is set at 100 kcal/mol. If this penalty energy is set to zero, the design calculation will run as two independent calculations on the two structures.

When the two positions are linked in a calculation, the selection of a common amino acid is based on the sum of the interaction energies for the amino acids at these two sites. If position A on SC1 is linked to position B on PG, the linked energy of having a residue *i* at these two positions (E_iAB) is given by the sum of the total energy of residue *i* in the SC1 structure at position A ($E_i SC1_{A)}$ and the total energy of residue *i* in the PG structure at position B ($E_i PG_B$):

$$E_i AB = E_i SC1_A + E_i PG_B$$

If there exists another amino acid, *j*, that has a more favorable linked energy at these two positions than residue *i*, i.e., if

$$E_jAB < E_iAB$$

j will be selected as the common amino acid.

Conclusions

We have developed a procedure that is potentially useful in characterizing protein sequence-structure relationships. Using this method we are able to generate PG variants that showed significant changes in structure on binding to metal. Most of the conformation changes were from folded to unfolded structures and *vice versa*. However, none of the changes caused a complete switch to a a-helical fold from the ß-sheet fold, as we had hoped. The primary reason for not obtaining a distinct structural switch could be that the model systems we selected for this study are not close enough in sequence space to allow it. It is also possible that the energy contributed through binding Cu(II)IDA is not enough to completely tip the energy balance towards the helical structure. Future efforts using this procedure with more appropriate model systems and a more effective external perturbation to facilitate the switch could be potentially useful understanding and designing ligand-induced conformation changes in proteins.

References

Abel K, Y. M., Hilgenfeld R, Jurnak F. (1996). An alpha to beta conformational switch in EF-Tu. *Structure* 4, 1153-1159.

Berg, J. M. (1988). Proposed structure for the zinc-binding domains from transcription factor IIIA and related proteins. *Proc Natl Acad Sci U S A* 85(1), 99-102.

Bornberg-Bauer, E. (1997). How are model protein structures distributed in sequence space? *Biophys J* 73(5), 2393-403.

Bornberg-Bauer, E. & Chan, H. S. (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci U S A* 96(19), 10689-94.

Bryngelson, J. D., Onuchic, J. N., Socci, N. D., Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* 21(3), 167-95.

Bullough, P. A., Hughson, F.M., Skehel, J.J., Wiley, D.C. (1994). Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature* 371, 37-43.

Carr, C. M., Kim, P.S. (1993). A spring-loaded mechanism for the conformational change of influenza hemagglutinin. *Cell* 73, 823-832.

Cerpa, R., Cohen, F. E., Kuntz, I. D. (1996). Conformational switching in designed peptides: the helix/sheet transition. *Fold Des* 1(2), 91-101.

Chen. J, Lee , K.H., Steinhauer DA, Stevens DJ, Skehel JJ, Wiley DC. (1998). Structure of the hemagglutinin precursor cleavage site, a determinant of influenza pathogenicity and the origin of the labile conformation. *Cell* 95, 409-417.

Chiti, F., Webster, P., Taddei, N., Clark, A., Stefani, M., Ramponi, G., Dobson, C. M. (1999). Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proc Natl Acad Sci U S A* 96(7), 3590-4.

Cohen, F. E. (1999). Protein misfolding and prion diseases. *J Mol Biol* 293(2), 313-20.

Cordes, M. H., Burton, R. E., Walsh, N. P., McKnight, C. J., Sauer, R. T. (2000). An evolutionary bridge to a new protein fold. *Nat Struct Biol* 7(12), 1129-32.

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

Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science* 278(5335), 82-7.

Dalal, S., Balasubramanian, S. & Regan, L. (1997). Protein alchemy: changing beta-sheet into alpha-helix. *Nat Struct Biol* 4(7), 548-52.

Dalal, S. & Regan, L. (2000). Understanding the sequence determinants of conformational switching using protein design. *Protein Sci* 9(9), 1651-9.

Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P.,Thomas, P. D., Chan, H. S. (1995). Principles of protein folding--a perspective from simple exact models. *Protein Sci* 4(4), 561-602.

Dunbrack, R. L. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side- chain prediction. *J Mol Biol* 230(2), 543-74.

Dunbrack, R. L. & Karplus, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat Struct Biol* 1(5), 334-40.

Gronenborn, A. M., Filpula, D. R., Essig, N. Z., Achari, A., Whitlow, M., Wingfield, P. T., Clore, G. M. (1991). A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* 253(5020), 657-61.

Gross, M., Wilkins, D. K., Pitkeathly, M. C., Chung, E. W., Higham, C., Clark, A., Dobson, C. M. (1999). Formation of amyloid fibrils by peptides derived from the bacterial cold shock protein CspB. *Protein Sci* 8(6), 1350-7.

Hamada, D., Segawa, S., Goto, Y. (1996). Non-native alpha-helical intermediate in the refolding of beta- lactoglobulin, a predominantly beta-sheet protein. *Nat Struct Biol* 3(10), 868-73.

Harrison, P. M., Bamborough, P., Daggett, V., Prusiner, S. B., Cohen, F. E. (1997). The prion folding problem. *Curr Opin Struct Biol* 7(1), 53-9.

Holm, L. & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* 25(1), 231-4.

Holmes, M. A. & Matthews, B. W. (1982). Structure of thermolysin refined at 1.6 A resolution. *J Mol Biol* 160(4), 623-39.

Huber R, C. R. (1989). Implications of the three-dimensional structure of alpha 1-antitrypsin for structure and function of serpins. *Biochemistry* 28, 8951-66.

Karplus, M. & Sali, A. (1995). Theoretical studies of protein folding and unfolding. *Curr Opin Struct Biol* 5(1), 58-73.

Kellis, J. T., Todd, R. J., Arnold F. H. (1991). Protein stabilization by engineeried metal chelation. *Biotechnology* 9, 994-995.

Koehl, P. & Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol* 239(2), 249-75.

Koehl, P. & Delarue, M. (1996). Mean-field minimization methods for biological macromolecules. *Curr Opin Struct Biol* 6(2), 222-6.

Lattman, E. E. & Rose, G. D. (1993). Protein folding--what's the question? *Proc Natl Acad Sci U S A* 90(2), 439-41.

Lee, C. (1994). Predicting protein mutant energetics by self-consistent ensemble optimization. *J Mol Biol* 236(3), 918-39.

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

Martinez, M. A., Pezo, V., Marliere, P., Wain-Hobson, S. (1996). Exploring the functional robustness of an enzyme by in vitro evolution. *Embo J* 15(6), 1203-10.

Mezei, M. (1998). Chameleon sequences in the PDB. *Protein Eng* 11, 411-414.

Minor, D. L. & Kim, P. S. (1996). Context-dependent secondary structure formation of a designed protein sequence. *Nature* 380(6576), 730-4.

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

Murzin, A. G. (1996). Structural classification of proteins: new superfamilies. *Curr Opin Struct Biol* 6(3), 386-94.

Orosz, A., Wisniewski, J., Wu, C. (1996). Regulation of Drosophila heat shock factor trimerization: global sequence requirements and independence of nuclear localization. *Mol Cell Biol* 16, 7018-7030.

Roman-Alpiste, M. J., Martin-Ramos, J. D., Castineiras-Campos, A., Bugella-Altamirano, E., Sicilia-Zafra, A.G., Gonzalez-Perez, J. M., Niclos-Gutierrez, J. (1999). Synthesis, XRD structures and properties of diaqua(iminodiacetato)copper(II), [Cu(IDA)(H2O)(2)], and aqua(benzimidazole)(iminodiacetato)copper(II), [Cu(IDA)(HBzIm)(H2O)]. *POLYHEDRON* 18, 3341-3351.
Rose, G. D. & Creamer, T. P. (1994). Protein folding: predicting predicting. *Proteins* 19(1), 1-3.

Sakamoto, S., Obataya, I., Ueno, A., Mihara, H. (1999). Regulation of alpha/Beta-folding of a designed peptide by haem binding. *Chem Comm*, 1111-1112.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* 9(1), 56-68.

Suh, S. S., Haymore, B. L., Arnold, F. H. (1991). Characterization of His-X3-His sites in alpha-helices of synthetic metal-binding bovine somatotropin. *Protein Eng* 4(3), 301-5.

Tan, S. & Richmond, T. J. (1998). Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. *Nature* 391(6668), 660-6.

Todd, R. J., Van Dam, M. E., Casimiro, D., Haymore, B. L., Arnold, F. H. (1991). Cu(II)-binding properties of a cytochrome c with a synthetic metal- binding site: His-X3-His in an alpha-helix. *Proteins* 10(2), 156-61.

Voigt, C. A., Mayo, S. L.,Arnold, F.,Wang Z-G. (2001). Computational method to reduce the search space for directed protein evolution. *Proc Natl Acad Sci*. 98(7), 3778-83

Volbeda, A. & Hol, W. G. (1989). Crystal structure of hexameric haemocyanin from Panulirus interruptus refined at 3.2 A resolution. *J Mol Biol* 209(2), 249-79.

Yap, K. L., Ames, J.B., Swindells, M.B., Ikura, M. (1999). Diversity of conformational states and changes within the EF-hand protein superfamily. *Proteins* 37, 499-507.

Table X-1. Guanidinium chloride denaturation data. (+) indicates that chemical denaturation was performed in the presence of 500µM $NiCl_2$, and (-) indicates the absence of $NiCl_2$.

Mutants	**? G_u (kcal mol^{-1})[a]**	**m value[b] (kcal mol^{-1} M^{-1})**	**? ? G (kcal mol^{-1})[c]**
SCCH1 (-)	2.471	1.455	0.07
SCCH1 (+)	2.537	1.289	
SCCH2(-)	3.088	1.575	0.728
SCCH2(+)	3.816	1.259	
SCCH3(-)	2.580	1.530	1.55
SCCH3(+)	4.130	1.481	
SNH1(-)	2.814	1.670	1.106
SNCH1(+)	3.920	1.807	
SNH2(-)	3.596	1.885	-0.662
SNH2(+)	2.934	1.7164	
SNH3(-)	2.820	1.587	0.472
SNH3+)	3.292	1.619	
SC1(-)	3.428	1.749	0.20
SC1(+)	3.628	1.739	

[a]Free energy of unfolding at 25 °C
[b]Slope of ?G_u versus denaturant concentration.
[c]Change in ?G_u in the presence of 500 µM $NiCl_2$

Table X-2: Thermal denaturation data of all the relevant protein G and engrailed homeodomain mutants constructed using joint entropy profiles. For comparison, the stabilities of the starting sequences of both the structures, PGWT and SC1, were also measured in the presence and absence of Cu(II)IDA. Sequence identities between pairs of molecules are also noted.

Protein	Tm Without Cu(II)IDA (° C)	Tm 1mM Cu(II)IDA (° C)	Sequence Identity (%)
SC1	87	85	8.9
PGWT	81	79	
SC4H	71	83	
SC4H56	79	91	25
PG4H	57	47	
PG50	49	31	31
SC50	65	77	
PG45	41	Unfolded	33
SC45	67	79	
PG40	Unfolded	33	37.5
SC40	77	85	
PG35	25	Unfolded	41
SC35	61	71	

Figure X-1: SC1 structure showing the designed metal binding sites. The amino acids labeled represent the helical surface positions that were considered for metal site design. *i, i+4* combinations of these positions were tested for gain in stability on chelating nickel. SCCH1, SCCH2, and SCCH3 were three such combinations for the C-terminal helix, and SCNH1, SCNH2 and SCNH3 were the combinations for the N-terminal helix.

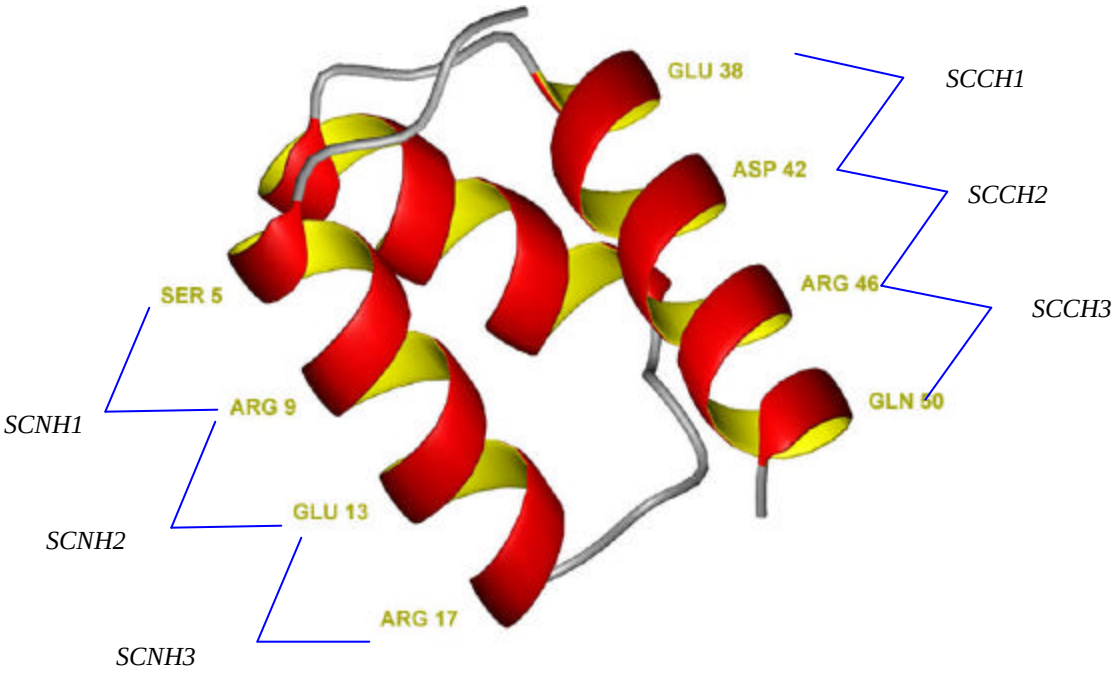
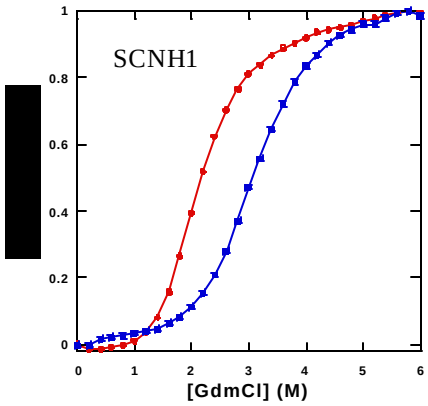
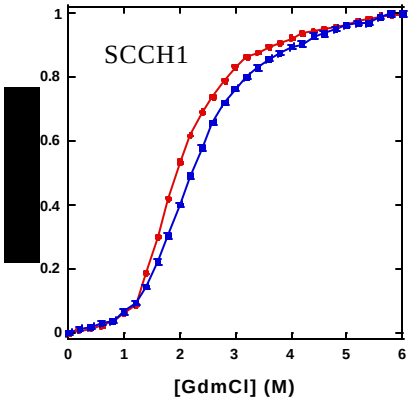
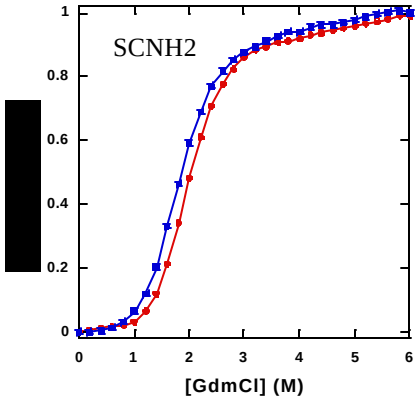
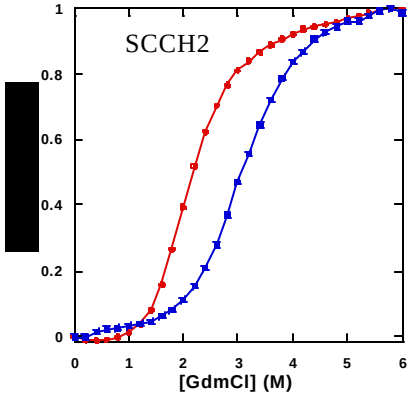

GLU 38
SCCH1
ASP 42
SCCH2
ARG 46
SCCH3
SER 5
GLN 50
SCNH1
ARG 9
GLU 13
SCNH2
ARG 17
SCNH3

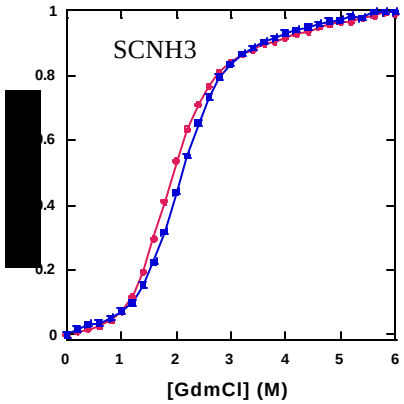
Figure X-2: Chemical denaturation curves monitored at 222 nm to study the effect of metal binding in metal site designed mutants of SC1. Guanidinium chloride denaturation was performed at 25 °C in the presence (shown in red) and absence of nickel (shown in blue). Maximum gain in stability is seen for SCCH3 and SCNH1. SC1 was used as a control since it does not have histidines on the helix surface positions that can chelate metals. The denaturation curves show that its stability is not affected by the presence of metal.

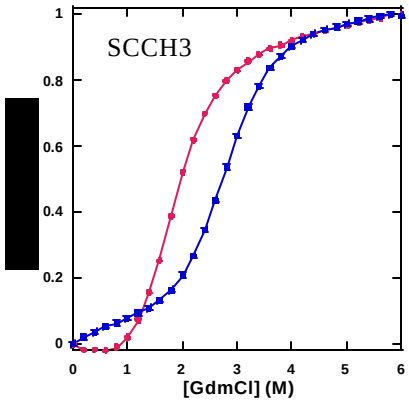

SCNH1
[GdmCl] (M)


SCCH1
[GdmCl] (M)


SCNH2
[GdmCl] (M)


SCCH2
[GdmCl] (M)


SCNH3
[GdmCl] (M)


SCCH3
[GdmCl] (M)

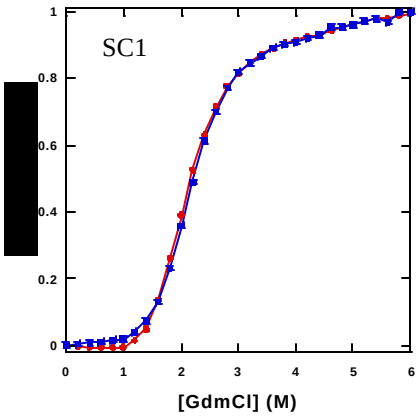
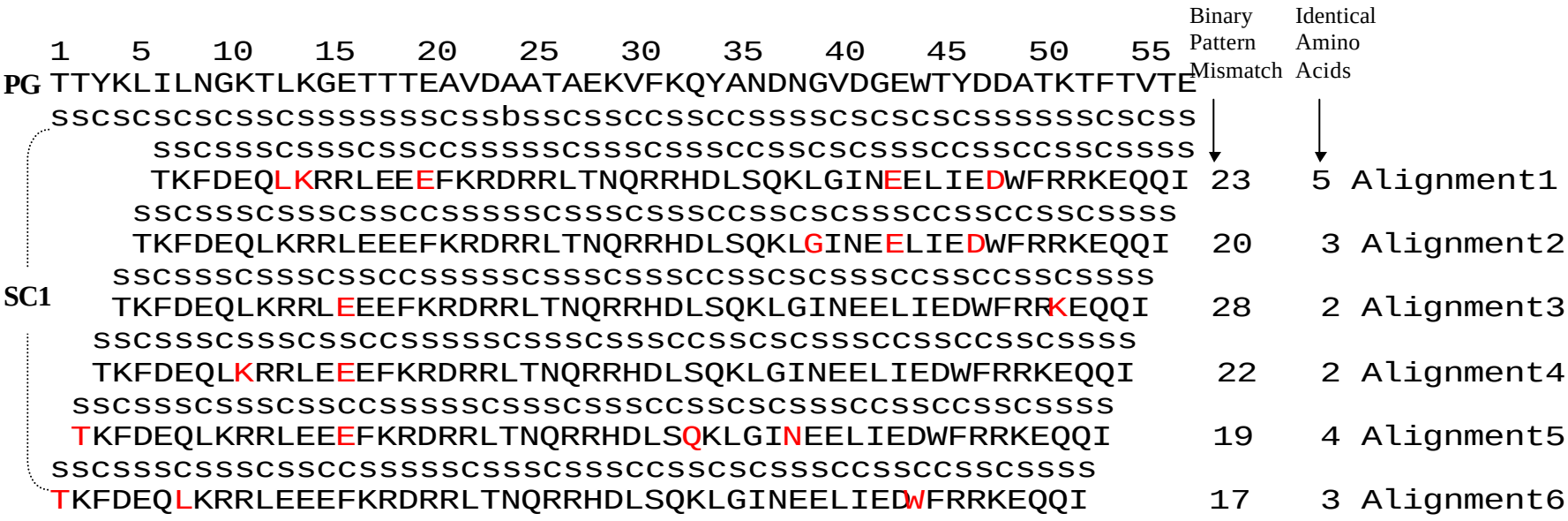
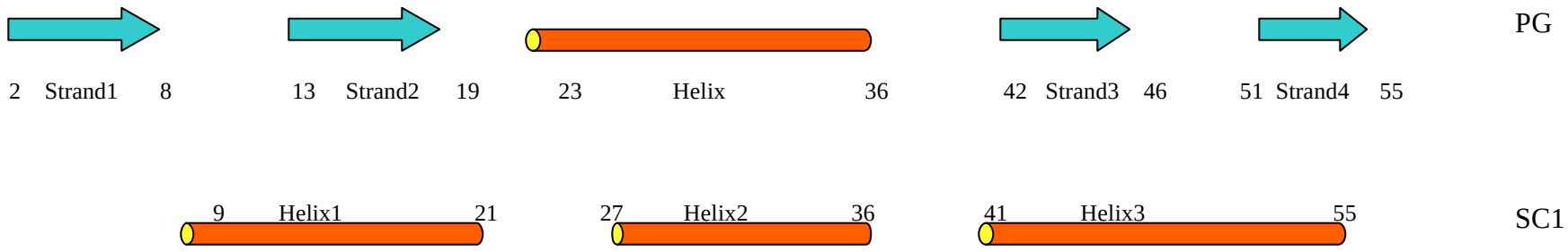

SC1
[GdmCl] (M)

Figure X-3: Sequence alignment between PG and SC1. (a) There are six possible alignment frames. “s” indicates surface positions and “c” indicates the core positions. Alignment 1 was selected as the most favorable because this has the least number of binary pattern mismatches and has five residues with common identities. (b) The secondary structure alignment corresponding to Alignment1 shows that the Helix2 of SC1 overlaps with the central helix of PG.

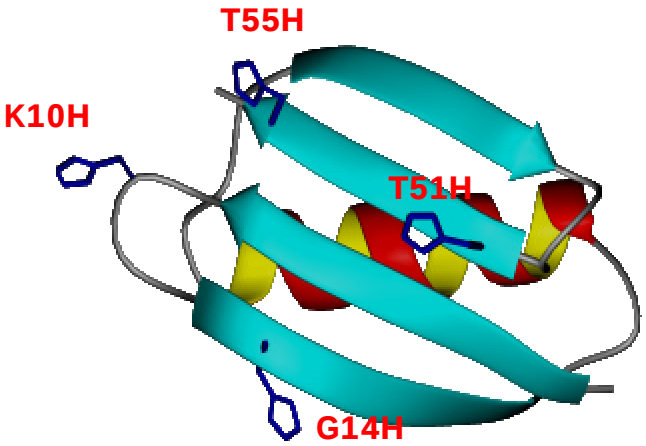

Binary Pattern Mismatch
Identical Amino Acids
1 5 10 15 20 25 30 35 40 45 50 55
PG TTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
sscscscscsscsssssssscssbsscsscccsscccsssscscscscsssssscscss
sscsssscsssscsscccsssssscsssscsssccsscscsssccsscccsscssss
TKFDEQLKRRLEEEFKRDRRLTNQRRHDLSQKLGINEELIEDWFRRKEQQI 23 5 Alignment1
sscsssscsssscsscccsssssscsssscsssccsscscsssccsscccsscssss
TKFDEQLKRRLEEEFKRDRRLTNQRRHDLSQKLGINEELIEDWFRRKEQQI 20 3 Alignment2
sscsssscsssscsscccsssssscsssscsssccsscscsssccsscccsscssss
SC1 TKFDEQLKRRLEEEFKRDRRLTNQRRHDLSQKLGINEELIEDWFRRKEQQI 28 2 Alignment3
sscsssscsssscsscccsssssscsssscsssccsscscsssccsscccsscssss
TKFDEQLKRRLEEEFKRDRRLTNQRRHDLSQKLGINEELIEDWFRRKEQQI 22 2 Alignment4
sscsssscsssscsscccsssssscsssscsssccsscscsssccsscccsscssss
TKFDEQLKRRLEEEFKRDRRLTNQRRHDLSQKLGINEELIEDWFRRKEQQI 19 4 Alignment5
sscsssscsssscsscccsssssscsssscsssccsscscsssccsscccsscssss
TKFDEQLKRRLEEEFKRDRRLTNQRRHDLSQKLGINEELIEDWFRRKEQQI 17 3 Alignment6

(a)

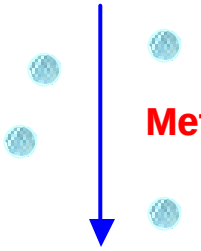

PG
2 Strand1 8
13 Strand2 19
23 Helix 36
42 Strand3 46
51 Strand4 55
SC1
9 Helix1 21
27 Helix2 36
41 Helix3 55

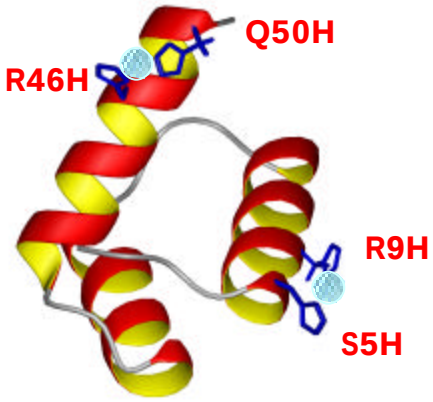
(b)

Figure X-4: The four histidine mutations on the surface of SC1 and PG are shown. The positions on PG correspond to the *i, i+4* histidine positions of SCCH3 and SCNH1 metal chelating mutants of SC1.



Protein G fold





Engrailed homeodomain fold

Figure X-5: Thermal denaturation of PG mutants monitored by CD at 218 nm. G14H mutant is represented by red closed circles and PG4H by blue triangles.

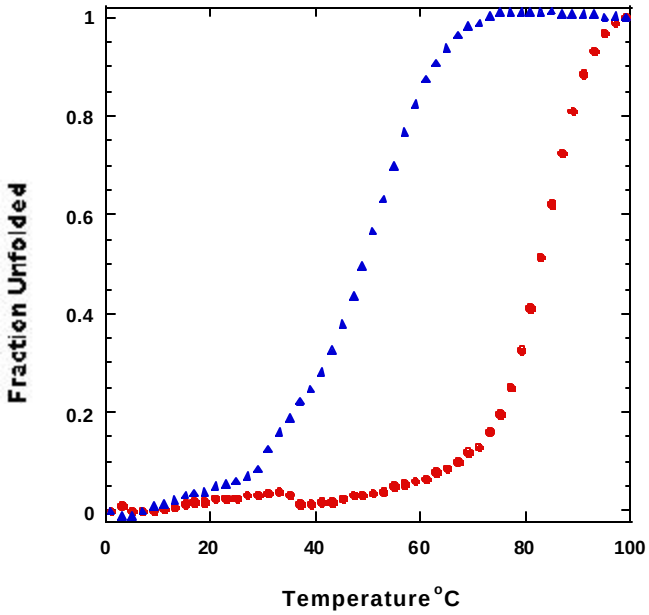
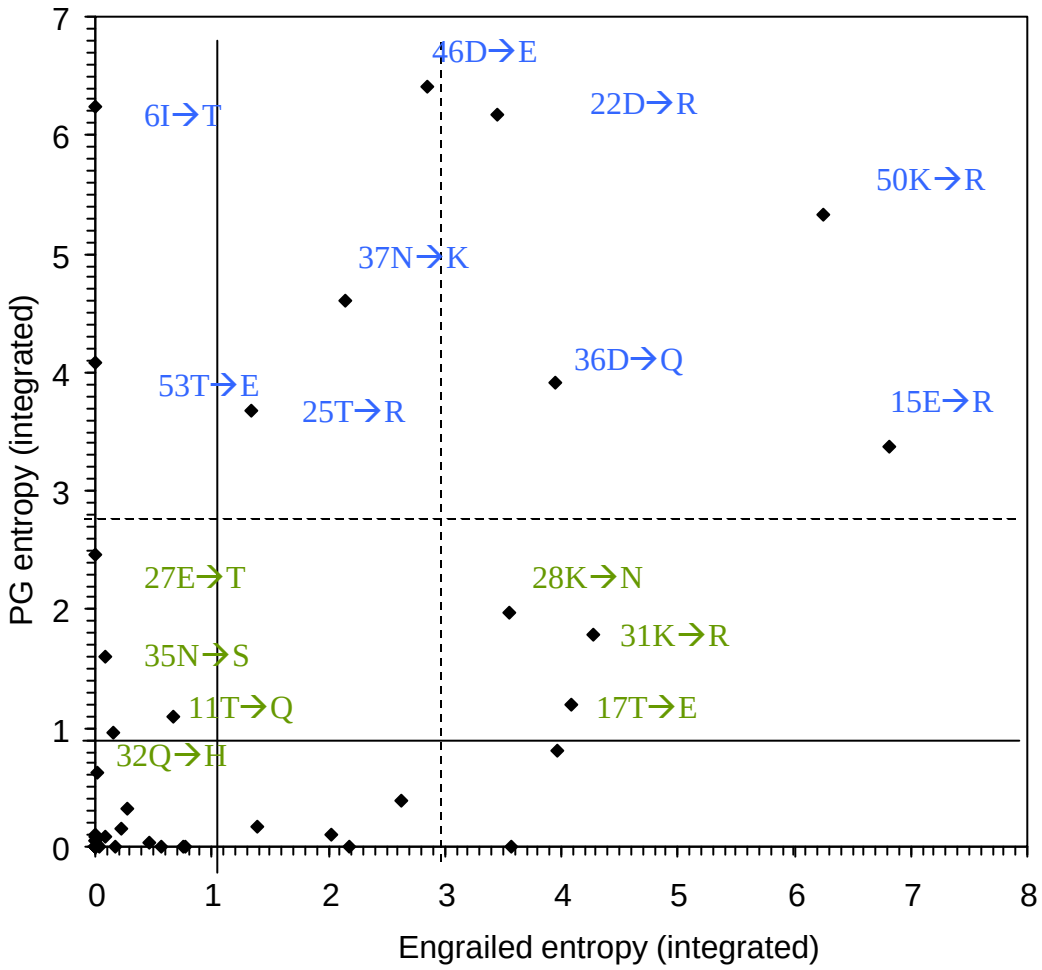
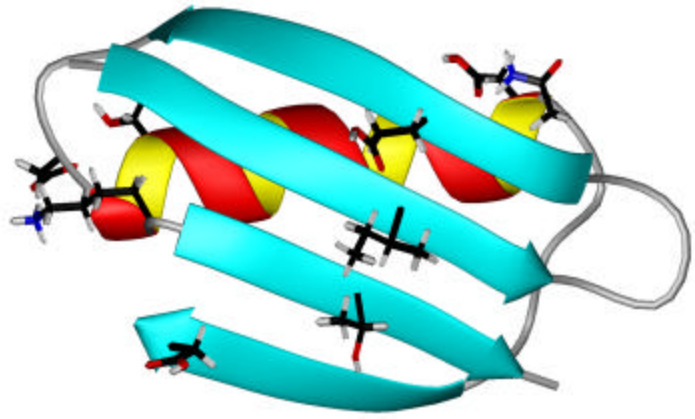

Fraction Unfolded
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
Temperature °C

Figure X-6: Entropy profile I. A comparison of site entropies is based on allowing only the two corresponding amino acids from Alignment 1 (Figure X-3a) at each position. The dotted lines represent the mean of the distribution for the entropies of both the molecules and the solid lines represent the standard deviations. Positions close to the Y-axis have a low tolerance for substitution in the SC1 structure and those close to the X-axis have a low mutability in the PG structure. The positions on the upper right quadrant of the graph have high entropies on both the molecules and can be substituted with each other's corresponding amino acids. The sequences at the bottom are of mutants constructed based on this profile. Molecule SW2 is a PG mutant that has mutations presented in the graph (in blue) up to the standard deviation line of the PG distribution. SW4 has additional mutations, marked in green in the graph, that are between the average line and the standard deviation line of PG entropy distribution in the graph.

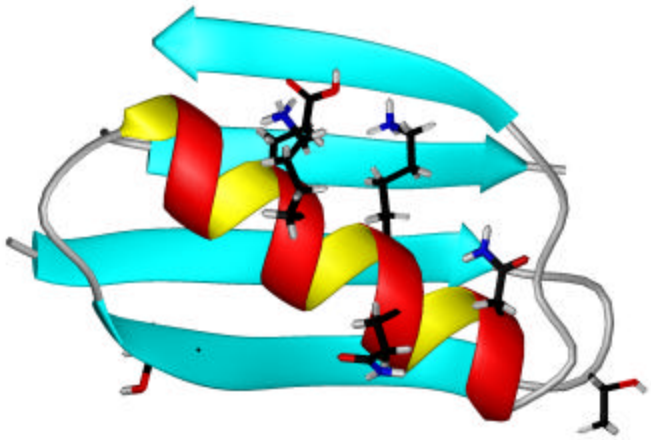



```
T T Y K L I L N G H T L K H E T T T E A V D A A T A E K V F K Q Y A N D N G V D G E W T Y D D A T K H F T V H E   PG_4H
T T Y K L T L N G H T L K H R T T T E A V R A A R A E K V F K Q Y A N K Q G V D G E W T Y E D A T R H F E V H I   SW2
T T Y K L T L N G H Q L K H R T E T E A V R A A R A T N V F R H Y A S Q Q G V D G E W T Y E D A T R H F E V H I   SW4
```

Figure X-7: Mutation distribution of SW2 and SW4 on PG. (a) SW2 mutations are predominantly clustered on the ß-sheet surface. (b) The additional mutations of SW4 are clustered on the helical surface.



(a)



(b)

Figure X-8: Entropy profile II: Some of the high entropy positions from entropy profile I are now reclassified as low entropy positions.

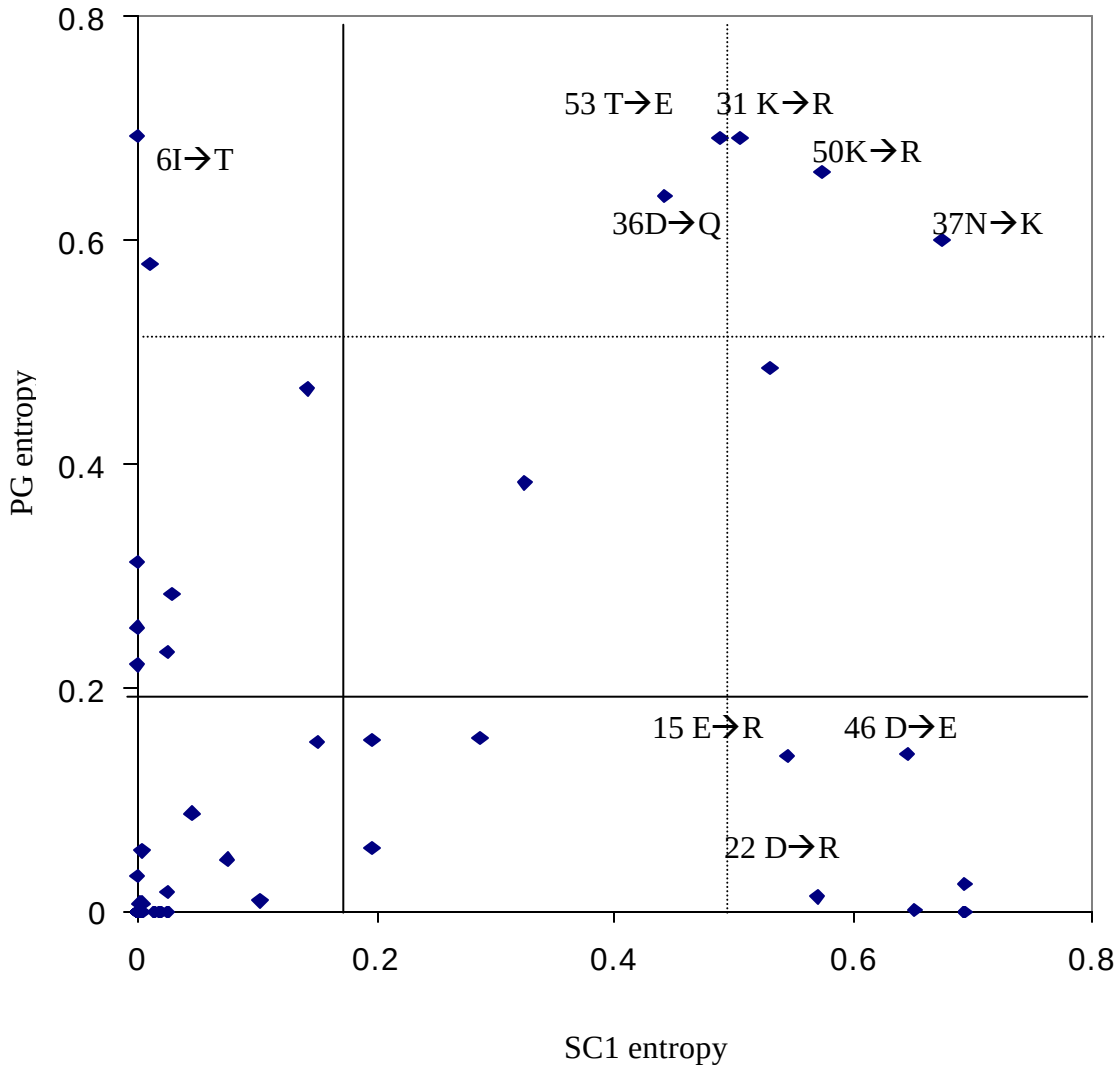
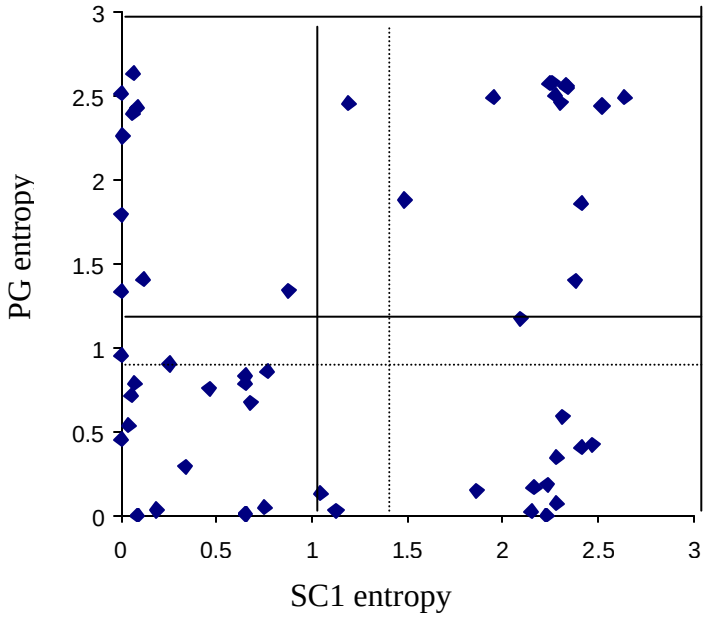
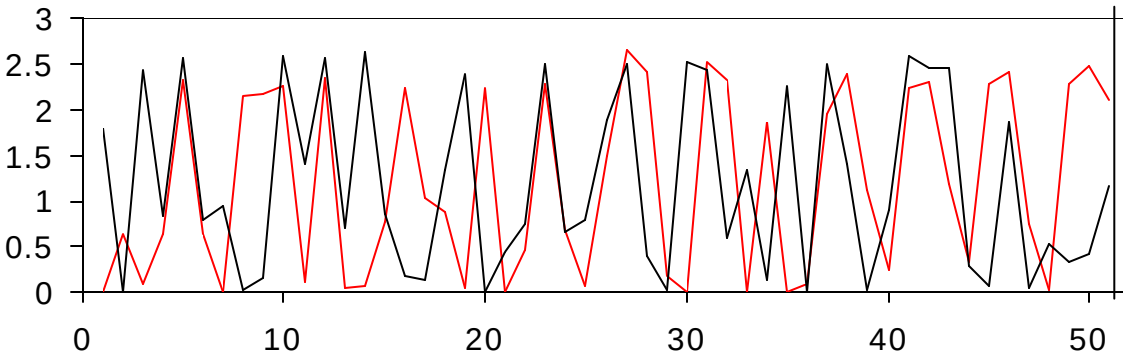

0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
PG entropy
SC1 entropy
53 T→E
31 K→R
50K→R
6I→T
36D→Q
37N→K
15 E→R
46 D→E
22 D→R

Figure X-9: Entropy profile III. (a) Comparison of the site entropies of all the positions on the two structures when all 20 amino acids are considered in the entropy profile calculations. (b) The entropy profiles of SC1 (red) and PG (black).


PG entropy
SC1 entropy
0
0.5
1
1.5
2
2.5
3

(a)


0
0.5
1
1.5
2
2.5
3
0
10
20
30
40
50

(b)

Figure X-10: Joint entropy profile of both the molecules derived from the combined probabilities of having common amino acids at the corresponding positions. The amino acid probabilities, $p_i(a)$ are the combined probabilities calculated as the sum of the amino acids rotamer probabilities at the paired positions on PG and SC.

$$p_i(a) = p_i(a)_{PG} \, p_i(a)_{SC}$$

The dotted lines represent the cutoffs used to select for design positions.

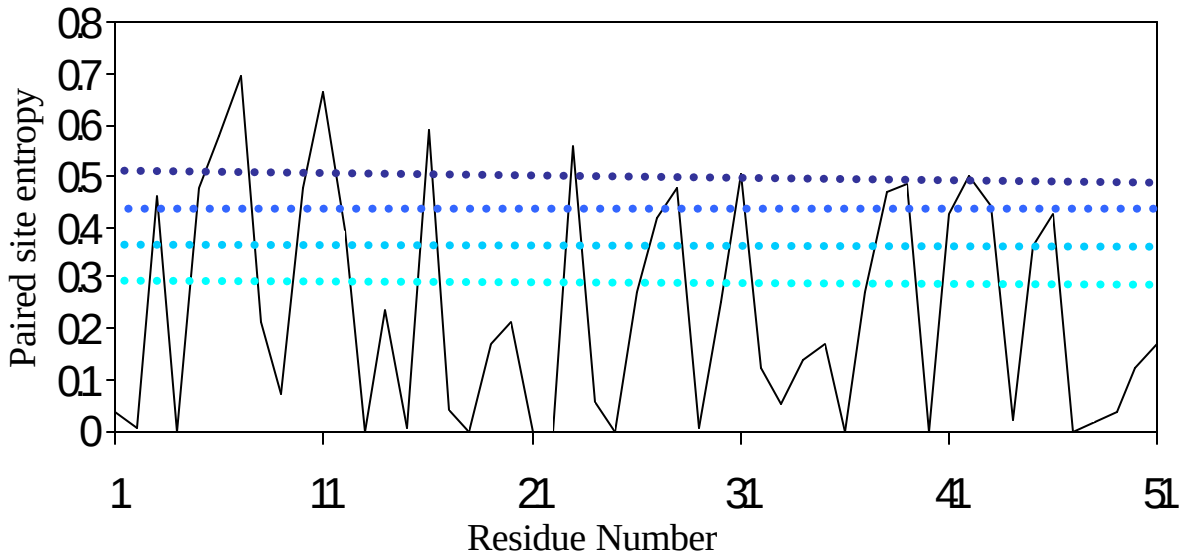

0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Paired site entropy
1
11
21
31
41
51
Residue Number

Figure X-11: Probability grid for determining common amino acids at the linked positions. The rows are the linked positions on the two structures and the columns are the 17 amino acids used in the study. An "X" in a box indicates that it may be possible to substitute the linked positions with the amino acid corresponding to that box. Positions considered for design upto a cutoff of 0.35 are highlighted

SC1	PG	G	A	V	L	I	M	F	Y	W	S	T	D	E	N	Q	K	R
1	6																X	
2	7																	
3	8			X	X	X						X						
4	9																	
5	10	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X
6	11						X				X	X	X	X	X	X	X	X
7	12															X		
8	13																X	X
9	14						X					X					X	
10	15			X	X	X	X				X	X	X	X	X	X	X	X
11	16				X	X								X		X		X
12	17				X		X	X	X	X	X	X	X		X		X	
13	18																	
14	19													X		X		
15	20																	
16	21						X								X	X	X	X
17	22												X		X			
18	23																	
19	24										X		X		X	X		X
20	25													X				
21	26																	
22	27																	
23	28		X	X		X	X	X	X	X	X	X	X	X	X	X	X	X
24	29												X	X	X	X	X	X
25	30																	
26	31		X								X	X	X		X			
27	32		X		X		X	X	X	X	X	X	X	X	X	X	X	X
28	33													X	X	X		
29	34																	
30	35										X							
31	36		X	X	X		X	X	X	X	X	X	X	X	X	X	X	X
32	37				X										X			
33	38				X													
34	39											X						
35	40				X													
36	41																	
37	42			X			X	X			X	X		X	X	X	X	X
38	43										X	X	X	X	X	X		X
39	44															X		
40	45																	
41	46			X	X	X	X				X	X	X	X	X	X	X	X
42	47			X		X	X				X	X	X	X	X	X	X	X
43	48						X						X	X		X	X	
44	49										X				X			
45	50															X		X
46	51						X				X	X	X	X	X	X	X	X
47	52																	
48	53														X			
49	54				X	X												
50	55										X	X						
51	56										X	X	X	X	X	X		
		G	A	V	L	I	M	F	Y	W	S	T	D	E	N	Q	K	R

Figure X-12: Structure of copper(II)iminodiacetic acid is shown with a di-histidine site chelating the copper center. The geometry of *i, i+4* positions on a helix is ideal for maximum chelating effect.

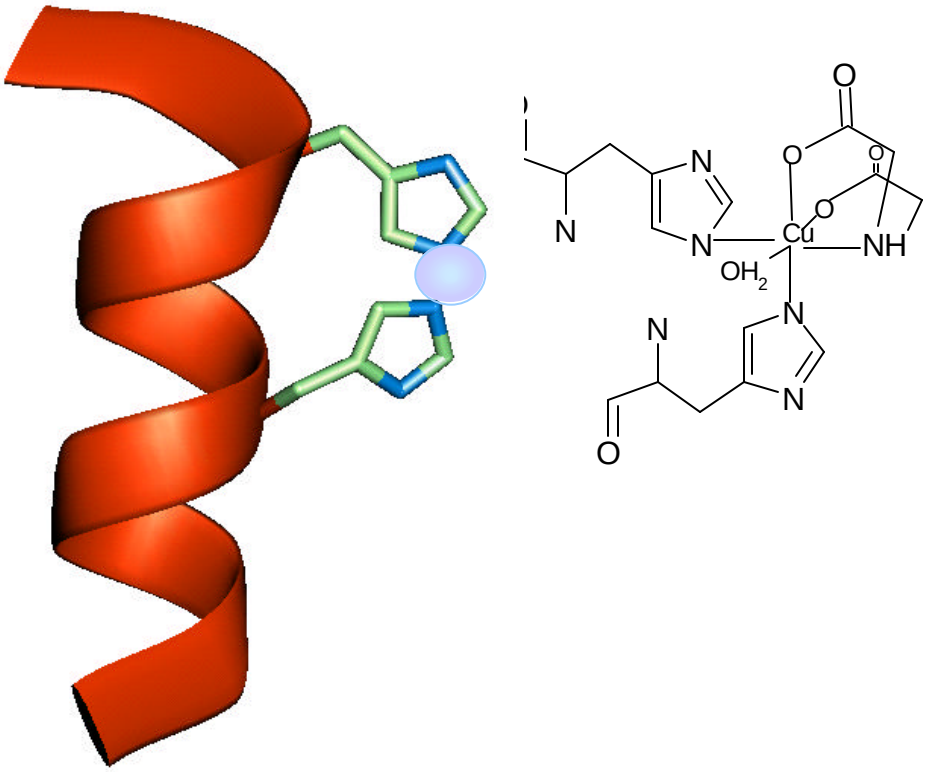

O
N
O
O
N
O
N
Cu
NH
OH2
N
N
N
O

Figure X-13: (a) and (c) are the wavelength scans and (b) and (d) are the thermal denaturation plots of PGWT and SC1. In all plots, red indicates measurements taken in the absence of Cu(II)IDA, and blue represents those taken in the presence on Cu(II)IDA. The wavelength scans indicate that the structures of the molecules do not undergo any significant conformational change in the presence of metal. These molecules do not have designed di-histidine sites and are not able to chelate metal ions.

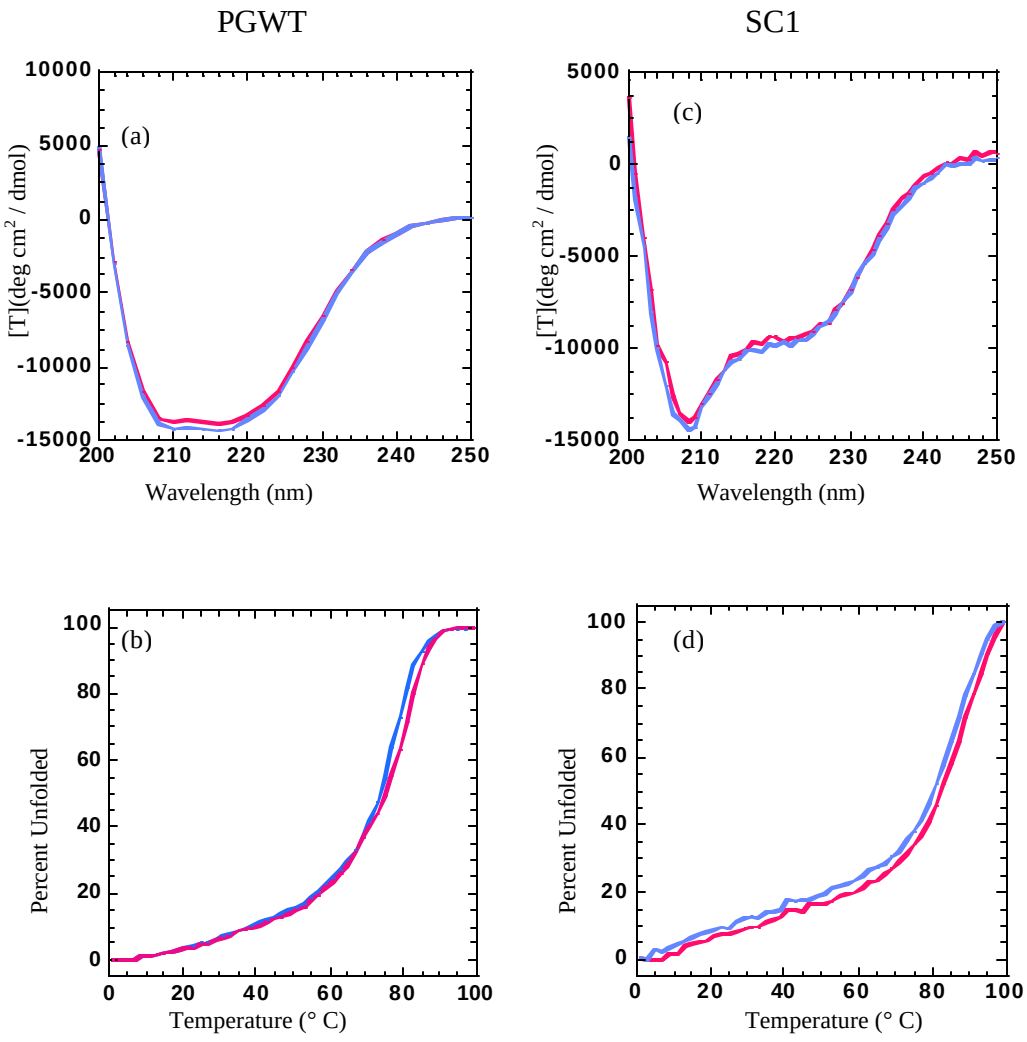

PGWT
SC1
(a)
(b)
(c)
(d)
[T](deg cm^2 / dmol)
Wavelength (nm)
Percent Unfolded
Temperature (° C)
10000
5000
0
-5000
-10000
-15000
200
210
220
230
240
250
100
80
60
40
20
0
0
20
40
60
80
100

Figure X-14: (a) and (c) are the wavelength scans and (b) and (d) are the thermal denaturation plots of PG4H and S4H56. In all plots red indicates measurements taken in the absence in Cu(II)IDA and blue indicates those taken in the presence on Cu(II)IDA. Wavelength scans indicate no significant structural changes. PG4H destabilizes by 10 °C in the presence of copper while SC4H56 gains 12 °C in thermostability.

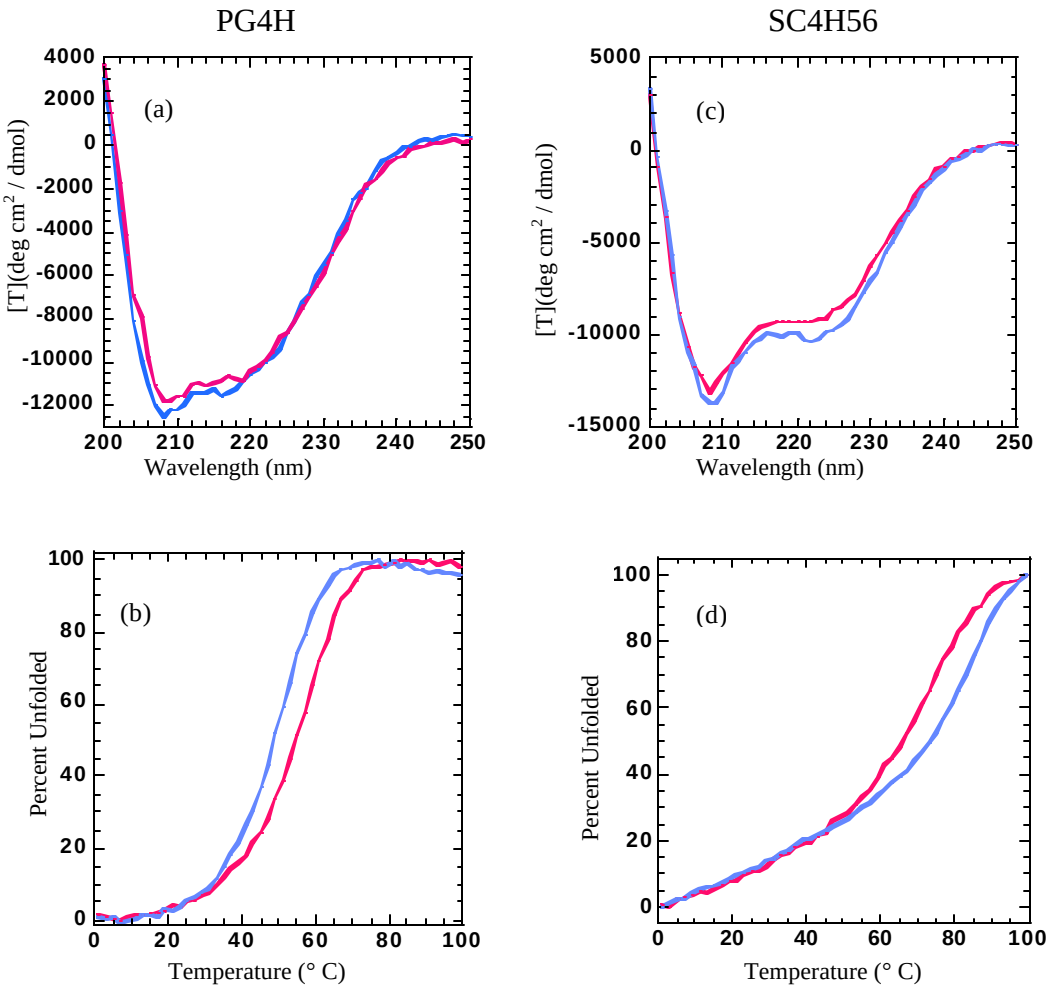
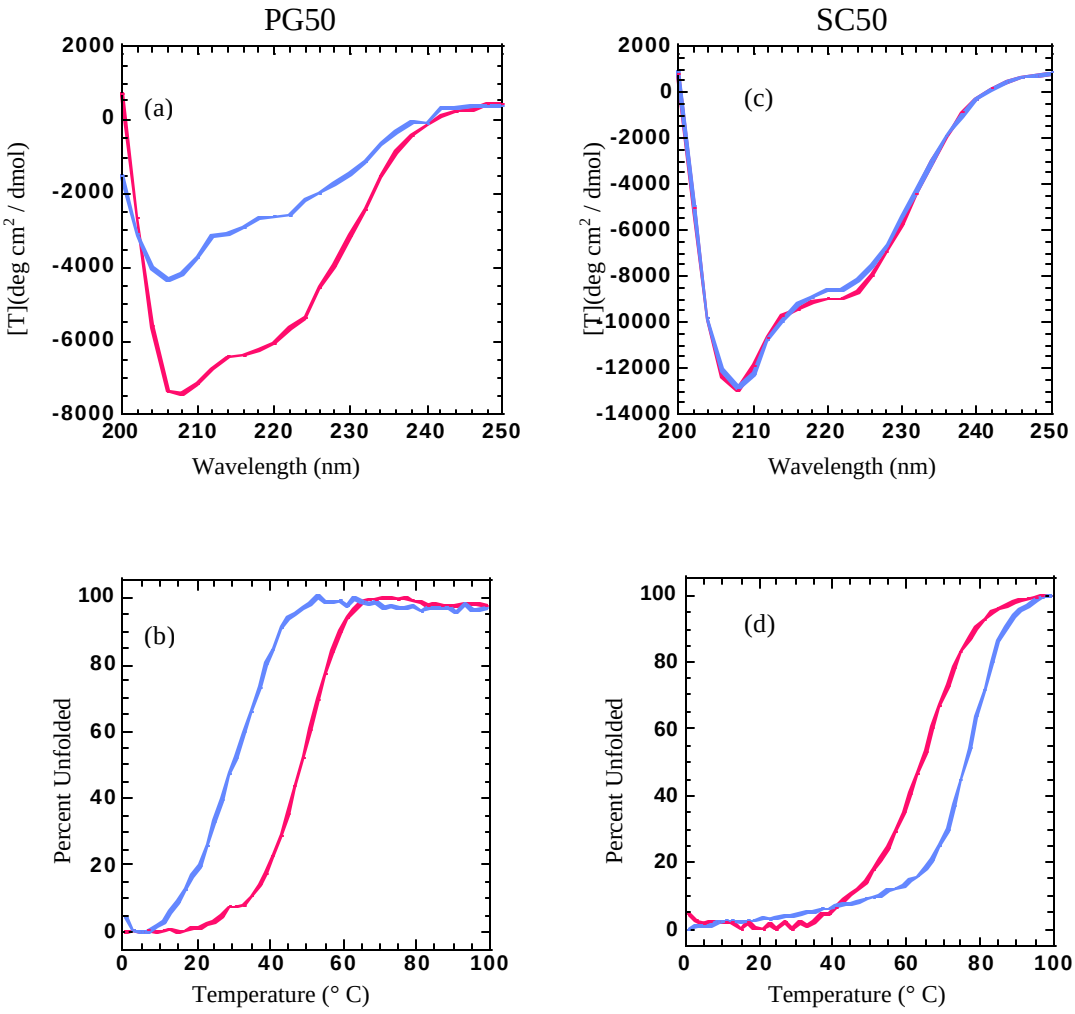
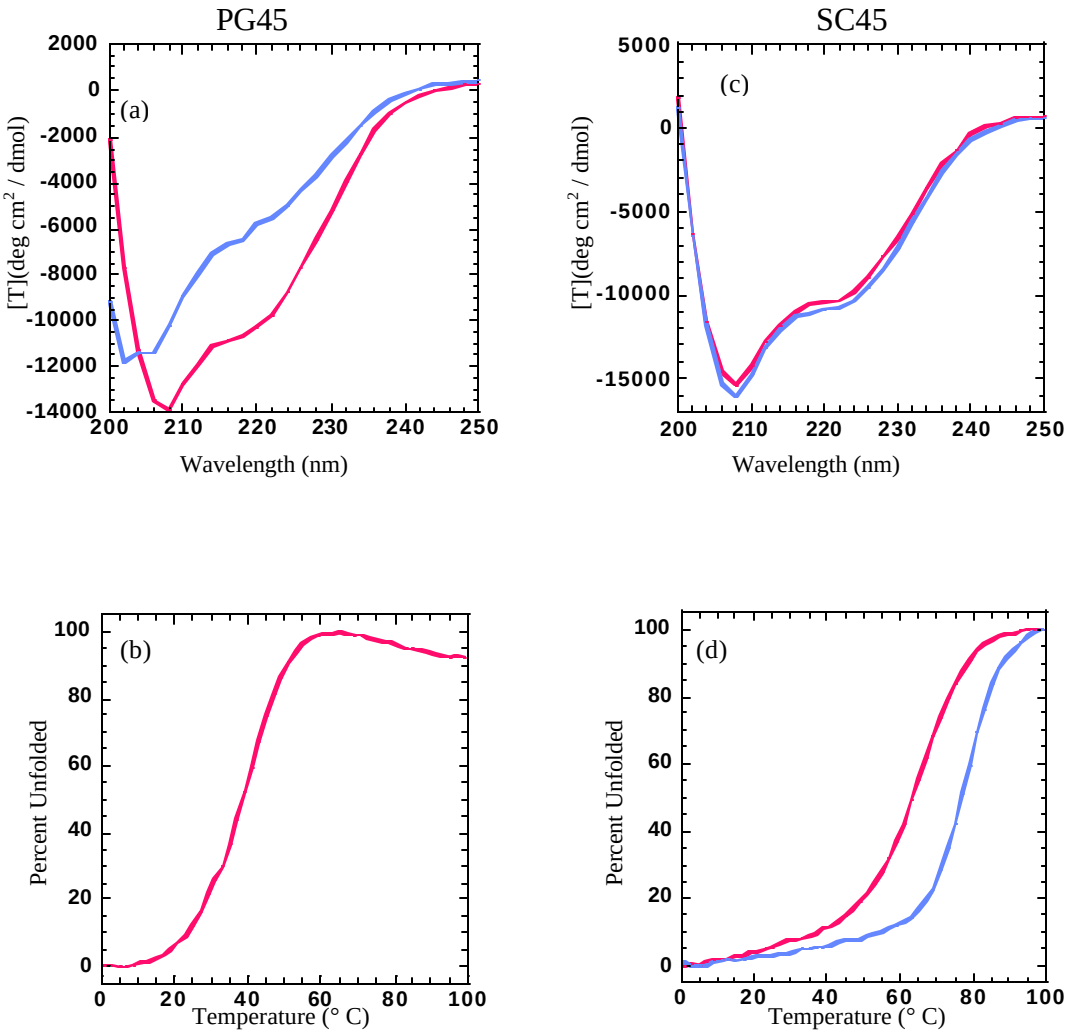

PG4H
SC4H56
(a)
(b)
(c)
(d)
[T](deg cm^2 / dmol)
Percent Unfolded
Wavelength (nm)
Temperature (° C)
4000
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
5000
-5000
-15000
200
210
220
230
240
250
20
40
60
80
100

Figure X-15: (a) and (c) are the wavelength scans and (b) and (d) are the thermal denaturation plots of PG50 and SC50. In all plots, red indicates measurements taken in the absence of Cu(II)IDA and blue indicates those taken in the presence of Cu(II)IDA. PG50 shows a significant shift in wavelength spectrum in the presence of Cu(II)IDA and also a loss in thermostability of 18 °C. Wavelength spectra of SC50 in the presence and absence of metal are identical, but a 12 °C gain in thermostability is observed on binding to metal. (e) represents the amino acid sequences of the two molecules. The identical amino acids are colored in red.



PG50:TTYKLILNGHKVKHETTTEAEDAATAENVFKQYANENGVDGEWTYDDATKHFTVHE
SC50:TTYKLTKFDHKVKHRLEEEFERDRRLTNQRRHDLSEKLGINEELIEDWFRHKEQHI
(e)

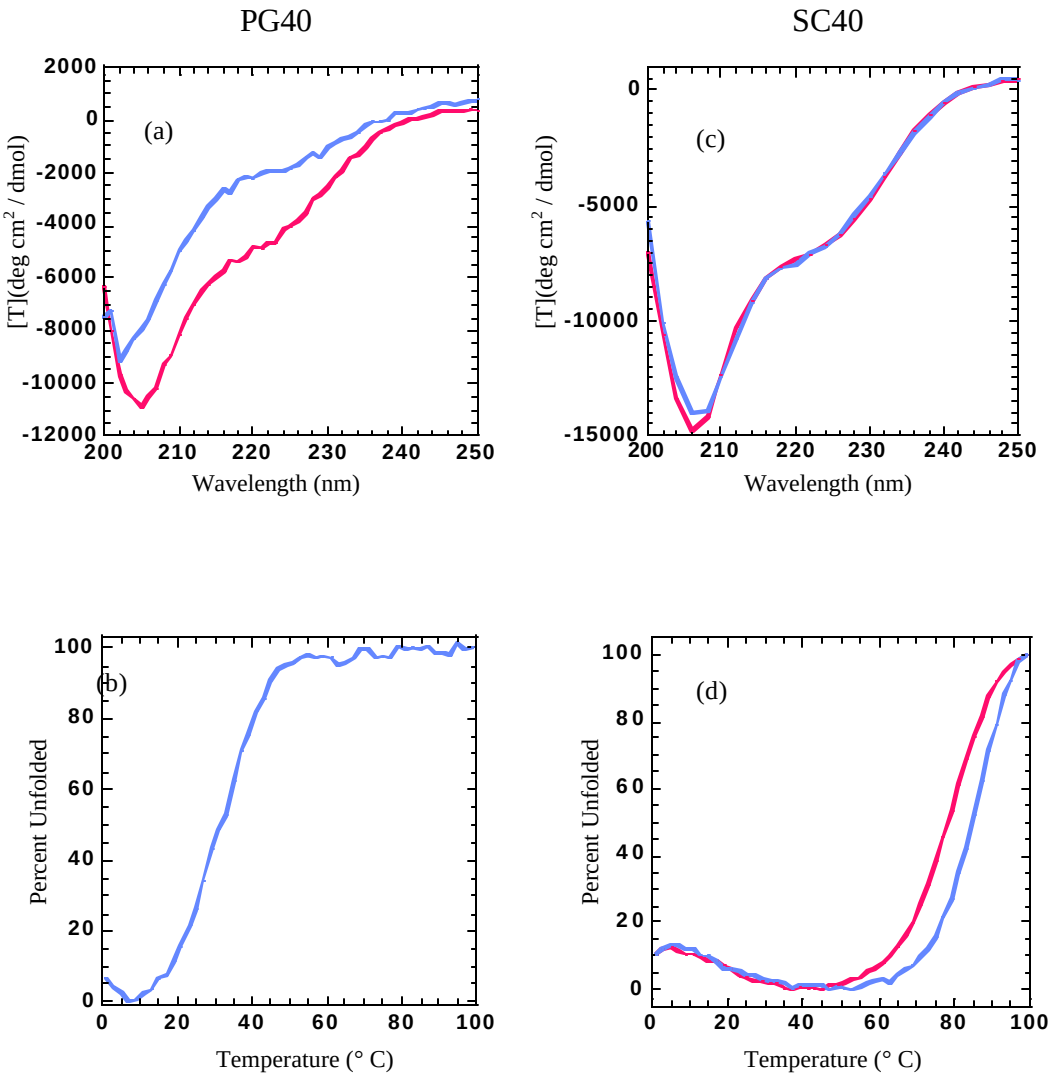
Figure X-16: (a) and (c) are the wavelength scans and (b) and (d) are the thermal denaturation plots of PG45 and SC45. In all plots, red indicates measurements taken in the absence of Cu(II)IDA and blue indicates those taken in the presence of Cu(II)IDA. In the presence of copper PG45, undergoes a structural change from a folded to an unfolded protein. (e) represents the amino acid sequences of the two molecules. The identical amino acids are colored in red.



PG45: TTYKLILNGHKVKHKTTTEAKDAATAENVFKQYANENGVDGEWTYDDATKHFTVHE
SC45: TTYKLTKFDHKVKHKLEEEFKRDRRLTNQRRHDLSEKLGINEELIEDWFRHKEQHI

(e)

Figure X-17: (a) and (c) are the wavelength scans and (b) and (d) are the thermal denaturation plots of PG40 and SC40. In all plots, red indicates measurements taken in the absence of Cu(II)IDA and blue indicates those taken in the presence of Cu(II)IDA. In the absence of copper, PG40 is unfolded but gets partially folded in the presence of copper. (e) represents the amino acid sequences of the two molecules. The identical amino acids are colored in red.

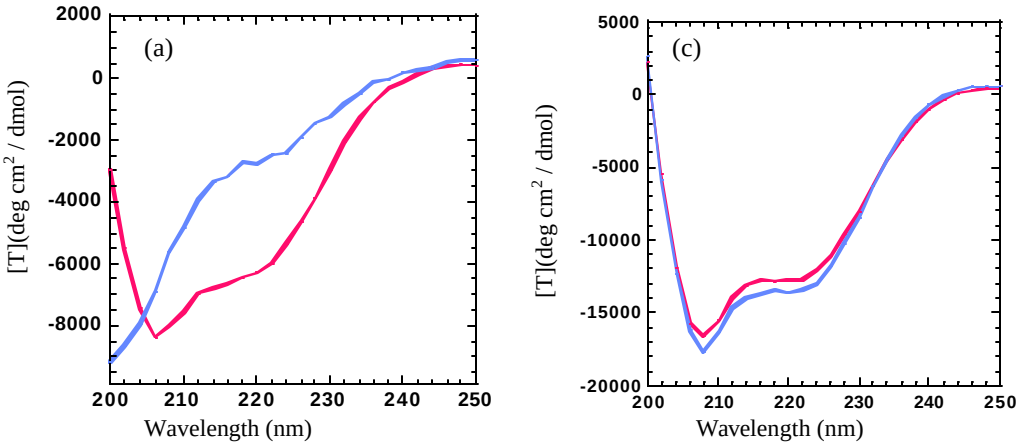


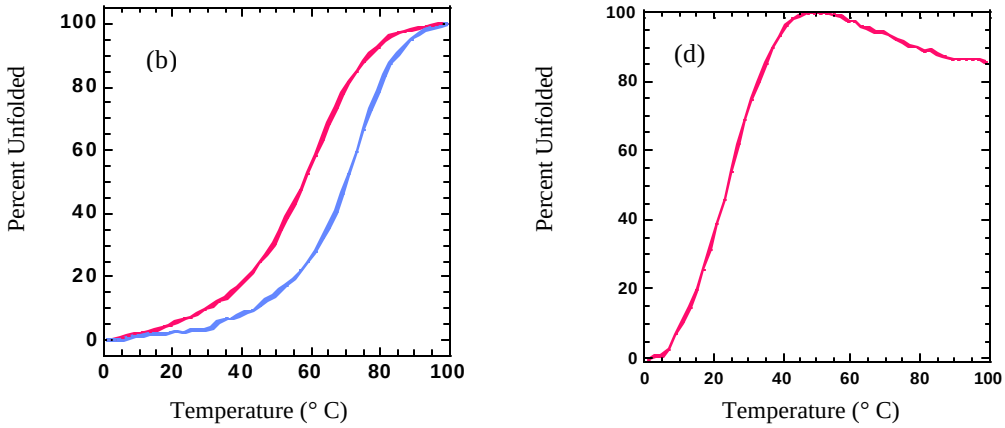
PG40:TTYKLILNGHKVKHKTTTEAQDAATAERVFKRYANENGVDGEWTYDDATKHFTVHE
SC40:TTYKLTKFDHKVKHKLEEEFQRDRRLTRQRRRDLSEKLGINEELIDDWFRHKEQHI

(e)

Figure X-18: (a) and (c) are the wavelength scans and (b) and (d) are the thermal denaturation plots of PG40 and SC40. In all plots, red indicates measurements taken in the absence of Cu(II)IDA and blue indicates those taken in the presence of Cu(II)IDA. In the presence of copper, PG35 undergoes a structural change from a folded to an unfolded protein. (e) represents the amino acid sequences of the two molecules. The identical amino acids are colored in red.

PG35 SC35





PG35: TTYKLILNGHKVKHKTTTEAEDAATAERVFKRYANENGVDGEWTYDRATEHFTVHE
SC35: TTYKLTKFDHKVKHKLTEEFERDRRLTRQRRRDLSEKLGINEELIDRWFEHKEQHI

(e)

Figure X-19: 1-D NMR spectra of engrailed homeodomain mutants. Spectra of all variants except SC35 indicate folded structures.

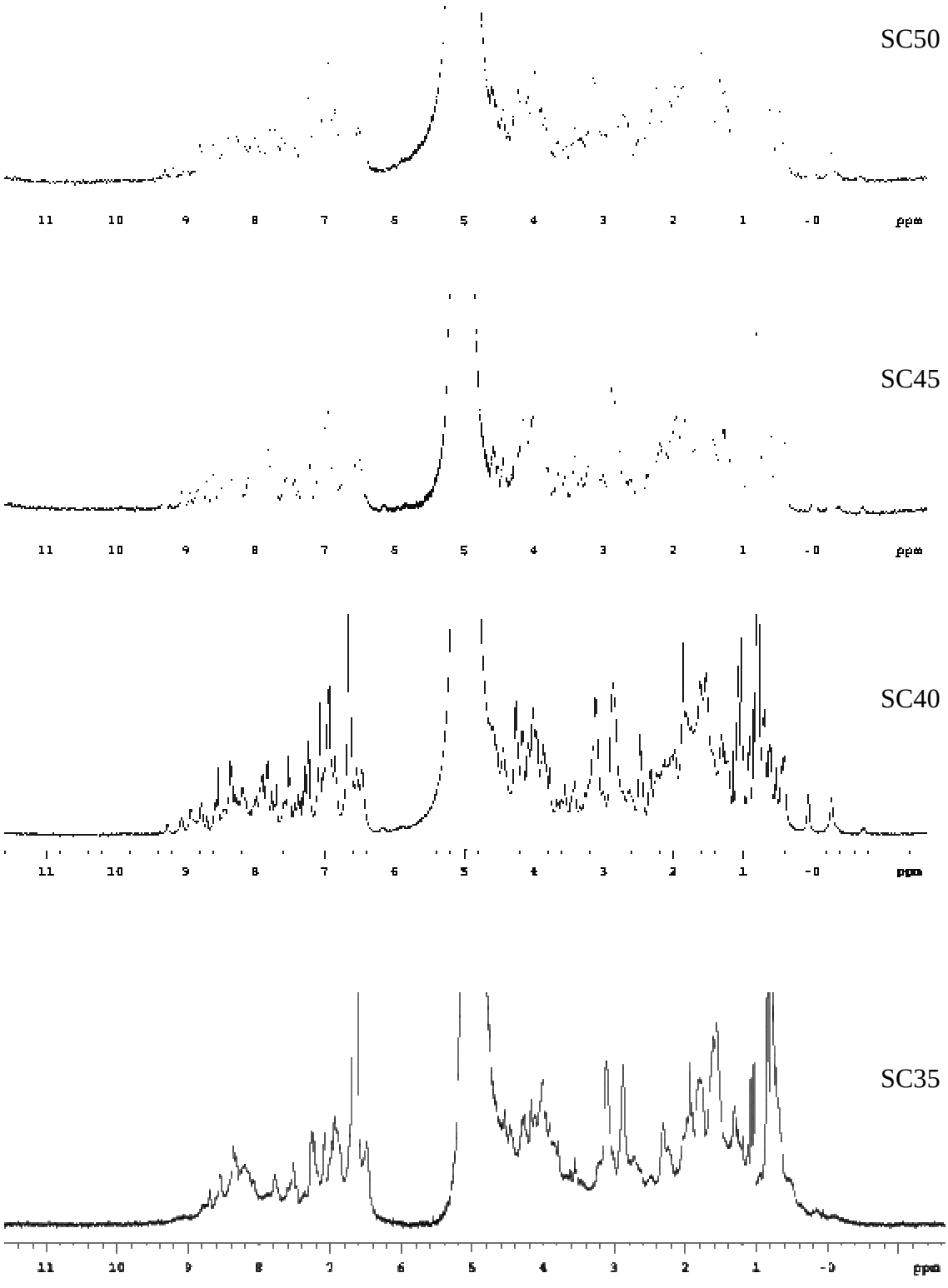
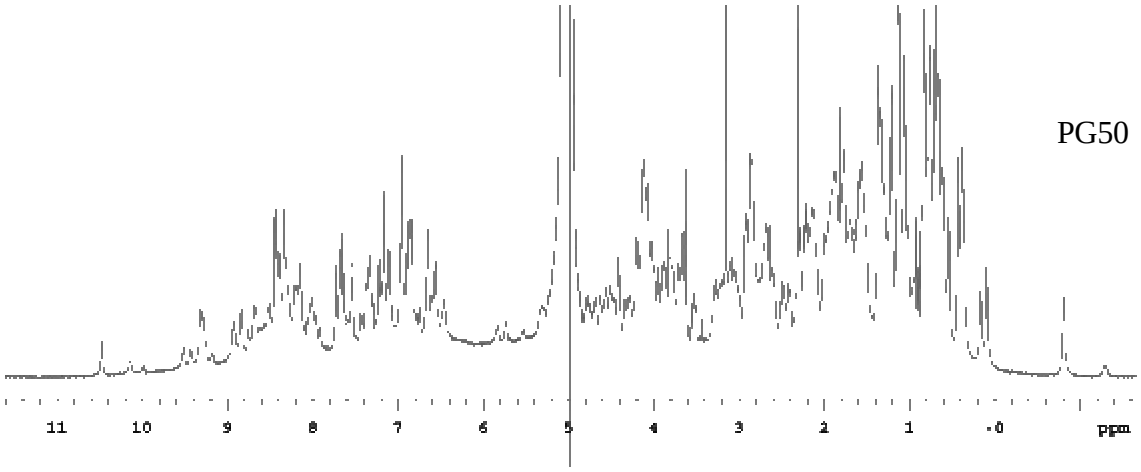
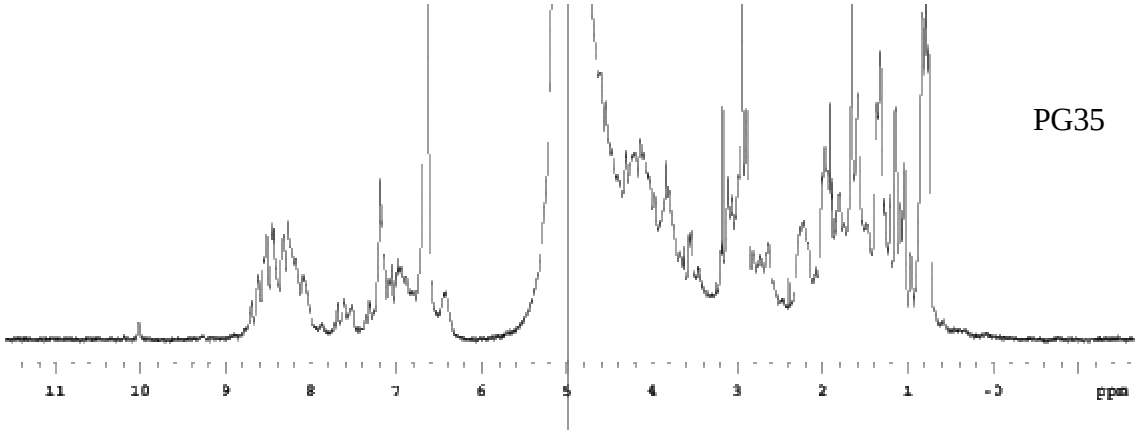

SC50
SC45
SC40
SC35
11 10 9 8 7 6 5 4 3 2 1 -0 ppm

Figure X-20: 1-D NMR spectra of Protien G variants. PG50 and PG45 are folded proteins while PG35 appears more like a molten globule.


PG50
11
10
9
8
7
6
5
4
3
2
1
-0
ppm


PG45
11
10
9
8
7
6
5
4
3
2
1
-0
ppm


PG35
11
10
9
8
7
6
5
4
3
2
1
-0
ppm

Appendix A

Repacking the Core of T4 Lysozyme by Automated Design

This appendix is adapted from an unpublished manuscript that was coauthored with Blaine H. M. Mooers, Walter A. Baase, Professor Stephen L. Mayo and Professor Brian W. Matthews. This was a collaborative effort between Caltech and University of Oregon. D. D. did the computational part of this study.

ABSTRACT

Automated protein redesign, as implemented in the program ORBIT, was used to redesign the core of T4 lysozyme. Twenty-six buried or partially buried sites in the C-terminal domain were allowed to vary both their sequence and side-chain conformation while the backbone and non-selected side-chains remained fixed. A variant with seven substitutions ("Core-7") was identified as having the most favorable energy. The redesign experiment was repeated with a penalty for the presence of methionines. In this case the redesigned protein ("Core-10") had ten amino acid changes. The two designed proteins, as well as the constituent single mutants, and several single-site revertants were over-expressed in *E. coli*, purified, and subjected to crystallographic and thermal analyses. The thermodynamic and structural data show that some repacking was achieved although neither redesigned protein was more stable than the wildtype protein. The use of the methionine penalty was shown to be effective. Several of the side-chain rotamers in the predicted structure of Core-10 differ from those observed. Rather than changing to new rotamers predicted by the design process, side-chains tend to maintain conformations similar to those seen in the native molecule. In contrast, parts of the backbone change by up to 2.8 Å relative to both the designed structure and wildtype.

Water molecules that are present within the lysozyme molecule were removed during the design process. In the redesigned protein the resultant cavities were, to some degree, reoccupied by side-chain atoms. In the observed structure, however, water molecules were still bound at or near their original sites. This suggests that it may be preferable to leave such water molecules in place during the design procedure. The results emphasize the specificity of the packing that occurs within the core of a typical protein. While point substitutions within the core are tolerated they almost always result in a loss of stability. Likewise, combinations of substitutions may also be tolerated but usually destabilize the protein. Experience with T4 lysozyme suggests that genuine core repacking with retention or enhancement of stability is difficult if not impossible to achieve without provision for

shifts in the backbone.

INTRODUCTION

The cores of proteins are generally well-packed (Richards, 1986; Richards & Lim, 1994). They have shown a remarkable ability to accommodate changes in buried hydrophobic residues although generally with some loss of stability (Baldwin et al., 1993, 1996; Gassner et al., 1996). It has been suggested that protein core packing is not like a jigsaw puzzle. Rather, it is more like nuts and bolts in a jar (Liang & Dill, 2001). If this is the case there may be opportunities to improve the stability of native proteins by optimizing the packing of buried amino acids. An early test with T4 lysozyme showed that the effectiveness of doing so by single amino acid substitutions seemed limited (Karpusas et al., 1989). A more general and possibly more powerful approach is by using automated design procedures that permit the consideration of multiple substitutions with alternative side-chain packing arrangements.

Several side-chain packing algorithms have been developed in which core redesign has been simplified by placing the side-chains on a rigid template. The side-chain conformations are usually varied by selecting from a library of rotamers, which are defined as statistically significant combinations of dihedral angles of a side-chain (Janin et al., 1978). One of the earliest attempts at automated side-chain repacking was implemented in the program, known as propack, developed by Ponder and Richards (1987). Hurley et al. (1992) used a modification of this program to redesign the C-terminal domain of T4 lysozyme. They considered several hundred promising sequences and energy minimized the best candidates. These redesigned proteins folded into native-like structures, but their stabilities were lower than that of the wildtype protein.

Programs such as propack make a direct attack on the combinatorial problem of finding the globally optimal arrangement of side-chains on a fixed template. The astronomical number of possible rotamer combinations limits the size of the rotamer library

and the number of positions that are allowed to vary in sequence. In addition, these algorithms have no guarantee of finding the structure with the lowest energy.

A different approach that has been developed recently is to iteratively eliminate the so-called dead-ending rotamers, i.e. those rotamers that cannot be part of the lowest energy structure (Desmet et al., 1992; Goldstein, 1994). This improvement allowed the extremely rapid testing of the 10^{40} to 10^{60} possible rotamer sequences in a reasonable amount of time, thereby permitting the use of more detailed rotamer libraries and the consideration of larger numbers of sites for repacking.

The ORBIT (Optimization of Rotamers By Iterative Techniques) protein redesign program allows use of several alternative versions of the dead end elimination theorem (Dahiyat & Mayo, 1996, 1997a, 1997b; Dahiyat et al., 1997). Several optional terms in the forcefield and alternative design strategies were developed using feedback from the redesign of two small proteins: the 56 residue β1 domain of streptococcal protein G (Dahiyat & Mayo, 1997a, 1997b; Su & Mayo, 1997) and 33 residue peptides that form homodimeric coiled-coils based on GCN4-p1 (Dahiyat & Mayo, 1996). By implementing these strategies, the β1 domain of streptococcal protein G was successfully redesigned with substantially enhanced thermal stability. One variant had a melting temperature in excess of 100 °C and an increase in thermal stability of 4.3 kcal mol^{-1} at 50 °C (Malakauskas & Mayo, 1998).

For several reasons, it was unclear whether the success of ORBIT with small proteins would be directly transferable to larger ones. For example, the change in exposed surface area on unfolding, as well as the change in heat capacity on unfolding both increase essentially linearly with protein size (Myers et al., 1995). Thus a given number of substitutions is likely to have a larger effect on stability when the total number of residues is small. Also, a larger proportion of residues are buried in larger proteins compared to smaller ones (Janin, 1979; Miller et al., 1987). This may require the design process to be more stringent.

To test the applicability of ORBIT to a larger protein, we used it to redesign the C-terminal domain of T4 lysozyme. Two designs were developed: one without and one with a penalty for the incorporation of methionine. The proteins were constructed, their thermal stabilities measured and their crystal structures determined. To determine the contributions made by individual substitutions, we studied proteins with constituent single mutations as well as proteins with the designed sequences but with a single site changed back to the wildtype sequence.

RESULTS

Redesigned T4-Lysozyme

The coordinates of the starting model were from the atomic resolution crystal structure of the cysteine-free pseudo-wildtype T4 lysozyme, referred to as WT* (Matsumura & Matthews, 1989). To obtain the highest possible accuracy the X-ray diffraction data were collected to 1.05Å resolution at 100K (Eriksson et al., 1993; B.H.M.M. & B.W.M., unpublished). After removal from the coordinate file of the solvent molecules and the alternative side-chain conformations, the crystal structure was energy minimized to relieve possible van der Waals clashes and steric strain before its use as the starting model in the redesign exercise. The discrepancy between the backbone atom positions in the crystal structure and in the energy minimized structure was 0.21 Å which is less than that between the 100K and 293K crystal structures (0.30 Å) (data not shown). Thus the energy minimization resulted in only small changes in the crystal structure.

T4 lysozyme has a N-terminal and a C-terminal domain. The latter is composed of a tightly packed α-helical bundle and includes residues 1-11 plus 70-164. It includes the most extensive and well-defined hydrophobic core and the redesign was in this part of the molecule. Twenty-six buried or largely buried residues were selected as contributing to the core (Figure A-1(a)). The amino acids at these positions were allowed to vary with regard to both their amino acid identity and their side-chain conformation while the remaining

residues were held fixed. The amino acids allowed at each site were those with hydrophobic side-chains, namely alanine, isoleucine, leucine, methionine, phenylalanine, tyrosine, tryptophan, and valine. Proline, glycine and cysteine were omitted from consideration to avoid possible disruption of secondary structure and the formation of disulfide bonds. This resulted in about 6×10^{26} amino acid sequences. By also allowing for different side-chain conformations the overall number of possible combinations increased to about 4×10^{59}. Based on the most favorable calculated energy the optimal design selected by ORBIT had seven substitutions (I78V, V87M, L118I, M120Y, V149I and T152V). It is referred to as Core-7.

This design protocol has been found to lead to an over-representation of methionines compared to the occurrence of methionine in natural protein cores [unpublished results]. The larger number of possible rotameric states for methionine leads to a proportionately over-representation of methionine in the rotamer library in comparison to other amino acids. It is also known that methionine-to-leucine substitutions at geometrically appropriate sites can enhance stability (Lipscomb et al., 1998). To take these factors into account, the design procedure was repeated with a penalty of 10 kcal/mol for each methionine included. With this penalty in place, ORBIT selected the ten-fold mutant ("Core-10") which has the mutations shown in Figure A-1(b). In the present instance the effect of the penalty was to both prevent the selection of new methionines and eliminate methionines present in the wildtype protein.

In order to obtain calculated energies for the various single, double and other mutants that had been constructed, the same procedure was applied without allowing amino acid sequence variation at the 26 sites. Energies were determined in the presence and absence of the methionine penalty (Table A-1).

Thermal Stability

Table A-1 includes the thermodynamic data for Core-7, Core-10, and the other variants.

Neither Core-7, Core-10, nor any of the revertants is as stable as WT*. The pH of maximum stability for both Core-7 and Core-10 is between pH 5 and 5.5 (data not shown). This is similar to WT* (Anderson et al., 1990) and suggests that the strong salt bridges, especially that between His31 and Asp70, are not significantly perturbed by either set of mutations.

Crystal Structures

Structures were determined for almost all of the proteins that had not been analyzed previously (Table A-2). Most crystallized isomorphously with WT* in space group $P3_221$. Diffraction data were generally to high resolution with an estimated uncertainty in the main-chain atom positions of 0.1 Å. Although the diffraction data were collected at 100K, the crystal structures are assumed to be accurate representations of the structure at room temperature. This is supported by comparisons of pairs of 100K and 293K crystal structures for the wildtype protein, WT*, and several mutants not included in this study (B.H.M.M. & B.W.M., unpublished results).

Core-7 crystallized in space group F222 with two or three molecules in the asymmetric unit and diffracted to 2.4 Å resolution, but it has not been possible to use molecular replacement to solve the structure. Crystals of M87V/Core-7 were also non-isomorphous with WT*. In this case there were three molecules per asymmetric unit and it was possible to determine the structure to 1.56 Å resolution. Crystals of the single-site revertant I118L/Core-7 were isomorphous with WT* and the structure was determined to high resolution (Table A-2). As will be apparent from the behavior of Core-10 revertants, however, the structure of M87V/Core-7 and I118L/Core-7 cannot be reliably used to infer the structure of Core-7 itself.

The redesigned protein Core-10 crystallized isomorphously with WT* and its structure was determined to 1.65 Å resolution. The structure is generally similar to WT* but also has some distinct differences in both the backbone structure and the side-chain

conformations (Figure A-2(a)). The average discrepancy between the main-chain atoms of residues 81-161 in Core-10 and WT* is 0.49 Å (Table A-4) which is about three times the combined uncertainty in the positions of the backbone atoms in each structure. When sites 106-123 are excluded from the least-squares superimposition to avoid incorporating the effect of the shifts in helices F and G, the discrepancy is 0.21 Å (Table A-3). This shows that the backbone structure of most of the C-terminal domain is well conserved, but within helices F and G some atoms move substantially (up to about 2.8 Å) (Figure A-3 (a)). The shift in helix F is associated with the breaking of the hydrogen bond between Thr109 O and Gly113 N. This distance increases from 3.0 Å to 4.2 Å. The breaking of this hydrogen bond was also observed in the crystal structure of the single mutant Val111→Ile (Hurley et al., 1992). The outward shift of helix F creates a cavity to which a water, HOH310, binds and is within hydrogen bonding distance of Ala111 O (2.9 Å with a C—O...HOH angle of 100°).

The temperature factors for the side-chain atoms at the ten sites of mutation in the crystal structures of Core-10 and of WT* are quite similar, indicating that these side-chains are not disordered; nor is there any indication of a molten globular state (Table A-4).

Comparison of the crystal structures of WT* and Core-10 reveals that the side-chain rotameric states are completely conserved at all of the non-substituted sites (Table A-5). Conservation also occurred at all but one of the substitution sites, the single exception being Met102→Leu, where both χ_1 and χ_2 changed (Table A-5; Figure A-5(a)).

DISCUSSION

The overall objective of the present experiments was to use ORBIT to identify variants of T4 lysozyme that had repacked cores and were more stable than wildtype. The most promising variant identified by the design process, Core-7, was found to be a functional lysozyme but with melting temperature reduced by 9.8°C, which corresponds to a destabilization of 3.5 kcal/mol relative to WT*. Change of the design procedure to

include a penalty for methionine residues led to a modified design, Core-10, which was 1.1 kcal/mol more stable than Core-10 but still not equal to WT*. In the following sections we discuss these findings in more detail with their implications for future design initiatives.

Energetics of the Designed Variants

As noted above, neither of the designed variants was as stable as wildtype lysozyme. This is at variance with the success of ORBIT in predicting stabilized variants of the β1 domain of protein G and coiled-coils based on GCN4 (Dahiyat & Mayo, 1996, 1997a, 1997b; Su & Mayo, 1997). It is, however, in agreement with earlier experiments on T4 lysozyme. Hurley et al. (1992) used a computational procedure to identify combinations of amino acids that would repack the core. Some possible combinations were suggested but their stability was, at best, slightly less than the native molecule. Also Baldwin et al. (1993) used a genetic approach to select variants that had repacked cores. Again, a large number of variants were identified, but none had stability greater than that of WT*.

One possible inference of these results is that it may be energetically more costly to repack larger proteins than smaller ones. In a very small protein most side-chains may be at least partly in contact with solvent. This may allow them freedom to be substituted, or to adjust their positions in response to substitutions at nearby sites. Within the core of a larger protein the side-chains tend to be tightly packed by their neighbors and it is more difficult for the structure to relax in response to introduced changes.

Calculated and Observed Stabilities

The stabilities of the various T4 lysozymes predicted by ORBIT are compared with those determined experimentally in Table A-1 and Figure A-4. These two energy terms do not have the same definition, but they are expected to correlate. For T4 lysozyme the experimental ΔΔG is traditionally defined to be the free energy of unfolding relative to the WT* protein (Elwell & Schellman, 1975; Grütter et al., 1987). ΔΔG refers to the free

energy of unfolding and a positive value indicates that the protein is more stable than WT*. The ORBIT score is the sum of the calculated energies of interactions for the side-chains that are allowed to vary. The energetically more favorable ORBIT scores are in the negative direction. For the individual mutations, excluding sites 102 and 103, there is a possible correlation between the calculated and observed energies (Figs. 4(c), 4(d)). This is equally apparent whether or not the ORBIT score includes the methionine penalty. However, when all sites and all constructs are considered no clear-cut correlation emerges (Figs. 4(a), 4(b)). This lack of agreement between the experimental and the predicted energies could be due to a number of factors including the following. (1) Some of the mutant proteins experience significant changes in the main-chain (see below). These may invalidate the rigid template assumption. (2) The rotameric states of some of the side-chains in the calculated structures (in particular, in Core-10) do not agree with those in the actual proteins leading to inaccurate energies. (3) The forcefield, which was based in part on experience with smaller proteins, may not be appropriate for proteins of the size of T4 lysozyme.

Predicted and Observed Structure of Core-10

Figure A-2(b) compares the backbone of the predicted and observed structure of Core-10. For residues 81-105 plus 124-161 the backbone agreement is generally good but in the remaining region there are shifts up to 2.8 Å. Likewise, some but not all of the side-chain conformations are correctly predicted. Eight of the ten modified side-chains adopt the rotameric state that was predicted (Table A-5). The two exceptions are Ile87 and Ile149 in which cases the differences are restricted to the χ_2 torsion angle. At another site (Val1036Ile) the χ_2 angle differs from the predicted structure more than 30°. Of the ten non-alanine residues that were included in the design process but did not change identity nine had correctly-predicted rotamers. The three incorrect predictions plus the prediction that is somewhat in error are discussed briefly below.

(a) Val87→Ile ($D\chi_1$ = 25°, $D\chi_2$ = -96°)

With reference to the crystal structure of WT*, the introduction of the CD1 carbon atom of Ile87 is associated with the outward movement of the side-chains of Leu118 and Glu122 as well as other shifts (Figure A-5(a)). Notwithstanding these shifts, potential steric clashes appear to cause the CD1 methyl of Ile87 to adopt a rotameric state that is fairly uncommon (frequency of 14%; Blaber et al., 1994). Before the design process, the starting model was energy minimized. Comparison of the design to the crystal structure of WT* shows that the distal part of the side-chain of Gln122 has moved outward in the designed structure. At least in part this suggested that an isoleucine in a common rotameric state could be accommodated at this site. (Gln122 is a surface residue that was held fixed during the design process. Thus, its outward movement is the result of the energy minimization step and not the rotamer selection step.)

(b) Leu91 ($D\chi_1$ = 24°, $D\chi_2$ = 139°)

Leu91 was included in the design process but its identity remained unchanged. It was predicted, however, that the two methyl groups at the end of the leucine side-chain would flip by about 180°. This change is not observed. Rather, the conformation of Leu91 in Core-10 is essentially identical to WT*. The change in conformation in the designed structure presumably occurs in concert with the introduction of an isoleucine at site 87. As mentioned above, Ile87 is predicted to have an altered conformation in Core-10. To avoid close contact with this residue Leu91 in the designed structure Ile87 adopts a different rotamer. Thus the error in prediction at the two sites seems to be coupled.

(c) Val103→Ile ($D\chi_1$ = 1°, $D\chi_2$ = -32°)

The side-chain of Ile103 adopts a rotameric state which has a frequency of only 3% among proteins in general. This is essentially as predicted although the observed χ_2 is 32°

from that anticipated. The distal methyl groups of the side-chain adopt positions that are close to those predicted (Figure A-5(b)). This coincidence occurs in spite of the change in the side-chain torsion angle and extensive shifts in several of the surrounding residues (especially 106-111). The superimposition of the crystal structure of WT* on the crystal structure of Core-10 suggests that these shifts may be caused in part by the need for Val111 to avoid a close contact with the CD atom of Ile103. Since the design process assumes a rigid framework such backbone shifts are not anticipated.

(d) Val149Ile ($D\chi_1$ = -2°, $D\chi_2$ = 111°)

Ile149 was predicted to adopt the most common rotameric state for leucine which has a frequency of 57%. Instead, it adopts a rotameric state which has a frequency of 14%. The design procedure deleted the four water molecules that are bound within the T4 lysozyme molecule (Weaver & Matthews, 1987). The removal of one of these resulted in a cavity into which the CD methyl group of Ile149 was predicted to occupy. In actuality, the water molecule remains bound to Core-10 and forces the isoleucine to adopt an alternative rotamer. [The water HOH197 shifts by 0.7 Å but retains its hydrogen bonding partners (Figure A-5(c)).]

Internal Water Molecules

Four buried waters occupy three cavities in WT* (Weaver & Matthews, 1987; Xu et al., 2001). These four waters were removed from the coordinate file during the design process. Two of the cavities are in the C-terminal domain and were therefore available for repacking by side-chain atoms. The first of these two cavities is next to site 149 and has already been discussed. The second cavity decreases slightly in the designed structure following the replacement of Thr152 with a valine. In the crystal structure, however, the water molecule (HOH173) still appears in the cavity although it is displaced towards the surface of the protein by about 1 Å.

Core-10 Revertants

Selected single-site revertants were constructed to address, both energetically and structurally, how the different sites interact with each other. The sites chosen for reversion were those where the point mutant had the largest effect on the stability of WT* (Table A-1). The single-site revertants of Core-10 are discussed briefly below.

(a) Leucine102→Methionine/Core-10

In the revertant L102M/Core-10, the leucine at site 102 in Core-10 was changed back to a methionine as in the wild-type sequence. The structure, however, remains very similar to that of Core-10 (rmsd of 0.14 Å). Thus, the amino acid change at site 102 back to that of the wildtype sequence does not recover the backbone atoms positions of the WT* structure. Met102 in the revertant adopts a side-chain conformation that is very similar to that of Met102 in WT* but that differs from that of Leu102 in Core-10 by a rotation of 85° about the χ_1 torsion angle.

In the L102M/Core-10 revertant the sum of the $\Delta\Delta G$'s of the remaining nine constituent mutants is essentially the same as the measured $\Delta\Delta G$ for the revertant (Table A-1, Figure A-6). This suggests that each of these nine sites is acting independently and that there is no interaction between them.

(b) Isoleucine103→Valine/Core-10

The discrepancy between revertant I103V/Core-10 and Core-10 for the main-chain atoms from sites 81-161 is 0.47 Å while it is only 0.22 Å relative to WT*. Thus the change of this single site back to the wildtype sequence is sufficient to revert the Cα positions in Core-10 essentially back to those of WT* (Figure A-3(b)). [It should be noted that the discrepancy between Core-10 and V103I is 0.68 Å, showing that the introduction of this single mutation is not sufficient to cause all the structural changes seen in Core-10. At the

same time, the single mutant V103I crystallized in a different space group and has a hinge-bending motion relative to WT*. This results in shifts in the C-terminus of helix C which makes detailed structure comparison more difficult.]

The change back to a valine from an isoleucine at site 103 removes a buried methyl group. This is correlated with Ala111 moving into a position similar to that occupied by Val111 in WT* and with helix F reverting to its wildtype conformation. It appears that the potential clash between the Ile103 CD1 methyl group and the CB methyl of Ala111 causes helix F to move outwards. The I103V revertant resulted in an 0.8 kcal/mol increase in stability relative to Core-10. This is notwithstanding the decrease in hydrophobicity resulting from the Ile→Val substitution and clearly suggests that the original V103I replacement introduces strain in the Core-10 structure.

The I103V/Core-10 revertant shows the largest non-additivity in $\Delta\Delta G$ of all the variants studied (Figure A-6). This also suggests that the remaining nine sites have the greatest degree of repacking and synergistic interaction.

(c) Revertant Protein, Alanine111→Valine/Core-10

In the Core-10 revertant A111V/Core-10, the alanine at position 111 in the Core-10 background is changed back to a valine as in the wildtype sequence. If the vicinity of site 111 in Core-10 was tightly packed, it would be expected that the introduction of two methyl groups would result in large structural changes. This, however, is not the case. The observed changes are actually modest. Val111 moves closer to the core by about 0.3 Å compared to Ala111 in Core-10, and atoms surrounding the reintroduced valine side-chain move by at most a few tenths of an Angstrom (Figure A-7(a)). The two methyl groups of the valine essentially refill the cavity that was created by the Val111→Ala substitution in Core-10. The most dramatic change in atomic position in the revertant is a 2 Å movement of the CD1 atom in the side-chain of Ile103. This movement occurs largely by a rotation of about the χ_2 angle to an energetically unfavorable rotameric state which places the CD1

atom at a distance of 2.7 Å from Ile103 CG2 atom (as opposed to 3.8 Å in Core-10).

The A111V reversion increases the stability of Core-10 by 0.6 kcal/mol (Table A-1). The fact that this is an increase rather than a decrease also suggests that the valine side-chain occupies a preformed cavity and does not introduce any serious steric clashes.

Evidence for Synergy Between the Mutation Sites

One can ask whether the ORBIT procedure results in genuine repacking of the core or, conversely, the individual substitutions act independently. In the case of Core-10 none of the constituent point mutations causes a large change in stability. Six of the ten mutations change the melting temperature by less than 1.0°C and the largest effect is for V111A for which the change is 2.9°C (Table A-1). If there were to be large synergistic effects one would anticipate that at least some of the point mutations would be quite destabilizing and that these effects would be compensated in the multiple mutant. This is not obviously the case. The hallmark of synergistic interaction is non-additivity of the $\Delta\Delta G$'s. If each of the substitutions acts independently of the others the change in stability of the multiple mutant should equal the sum of the $\Delta\Delta G$'s of its single-site constituents. As can be seen in Table A-1 and Figure A-6, the sum of the $\Delta\Delta G$'s for Core-10 is numerically 1.0 kcal/mol greater than the observed $\Delta\Delta G$. This shows that there is some favorable interaction among the redesigned sites, although the effect is modest. By way of comparison, in the "size switch" mutant in which the sizes of adjacent residues were switched by the substitutions Leu216Ala and Ala129Leu, the thermodynamic compensation was substantially larger (2.5 kcal/mol) (Baldwin et al., 1996).

Cooperativity between substitutions at different sites can also be evaluated structurally. Using a cut-off distance of 4.0 Å the average number of residues among the 26-residue set that are in contact or almost in contact with any given residue is 2.4 (or 1.4 residue-residue contacts if the threshold is reduced to 3.5 Å). Thus, even though the 26 residues are all within the most pronounced hydrophobic core of T4 lysozyme there do not

tend to be multiple close contacts between each residue and a multitude of neighbors. This separation of the sites may make cooperativity difficult to achieve. In the present case the design algorithm assumes that selected variants will retain the same backbone structure as the parent molecule. As noted above this is true for much of the C-terminal domain of Core-10, but not in the vicinity of the F and G helices.

In this context it is instructive to contrast the behavior of the Core-10 revertant L102M/Core-10 with that of I103V/Core-10. When the single-site reversion I103V is made in Core-10 the structure reverts much closer to that of WT* (Figure A-3(b)). Also the stability of the protein is increased by 0.8 kcal/mol and, in addition, the non-additivity of the $\Delta\Delta G$'s increases by 0.3 kcal/mol (Table A-1, Figure A-A-6). Thus, the I103V/Core-10 revertant is, in all respects, a more superior design than Core-10 itself (it is more stable, more synergistic and has a structure more like WT*). When the V103I mutation is included in the full Core-10 construct, the addition of the CD1 methyl group introduces a steric clash which is not compensated by the other replacements and, therefore, leads to a relatively large change in the structure.

In contrast, the behavior of the L102M/Core-10 revertant is quite different. Here the reversion of Leu102 to Met causes almost no change in the Core-10 structure. At the same time (as judged by the equivalence of the $\Delta\Delta G$'s; Figure A-6), it eliminates any synergistic interaction between the remaining nine sites. The L102M/Core-10 structure seems "poised" to accept the M102L substitution without structural perturbation, and, in so doing, the Leu102 side-chain contributes to the synergistic interaction that is observed in Core-10.

Since the L102M revertant in Core-10 eliminates synergistic interaction between the remaining nine sites it implies that the M102L substitution does contribute to cooperativity in Core-10. There is some structural evidence for this. When the M102L mutation is made in WT* it results in a rotation of the side-chain of Phe114 by almost 70° into a strained conformation. (This rotation appears to be mediated indirectly *via* Trp138

and possibly other residues as well) In Core-10 (and in M102L/Core-10), however, the combination of substitutions allows the side-chain of Phe114 to revert to the angle seen in WT* (Figure A-7(b)), relaxing the strain that had been introduced.

Success of the Methionine Penalty

Because of their conformational adaptability methionine side-chains tend to be more readily accommodated within a designed protein. At the same time incorporation of multiple methionines can result in a loss of stability (Gassner et al., 1996). Conversely, under favorable circumstances substitutions from methionine to leucine can increase stability (Lipscomb et al. 1998). For these reasons it would seem desirable to avoid the introduction of methionines into the designed protein.

In the present case the imposition of a methionine penalty resulted in four positions in Core-7 being retained in Core-10 while I78V and I118L were lost and V87M was replaced with V87I. Meanwhile, five new positions were added, resulting in the loss of two methionines: I100V, M102L, V103I, M106I, and V111A. In total, Core-10 has three fewer methionines than Core-7. The M102L substitution is known to introduce steric clashes (Hurley et al., 1992) and it could be that the additional sites of substitution in Core-10 arise from the need to minimize this steric interference. In any event, the incorporation of the methionine penalty did increase the stability of the protein by 1.1 kcal/mol.

Conclusions

One of the main findings of this work is that the introduction of the designed core-repacking mutations resulted in changes of the backbone up to 2.8 Å. Also both of the designed variants were less stable than the wildtype protein. Taken together these results suggest that genuine core repacking with retention or enhancement of stability may be difficult if not impossible to achieve without provision for shifts in the backbone.

A second finding is that the rotamer angles that occur in WT* are strongly conserved in the

mutant. For the substituted and non-substituted sites in Core-10 there is only one case (Met102 -> Leu) where there is a change of rotamer (Table A-5, Figures A-2(a), A-5(c)). Conservation of rotamers was also observed in genetically selected core-repacking variants of T4 lysozyme (Baldwin et al., 1993). This suggests that core redesign might be improved by favoring models that maintain the side-chain rotamers present in the reference structure.

If, as was the case with the Core-10 design, a total of 26 sites were allowed to vary, the overall number of possibilities is astronomical. At a given site, however, the packing is typically determined by the side-chain itself plus two or three neighbors. Here the number of choices is more limited. Also since the number of hydrophobic amino acids is fairly small, and each amino acid is restricted to distinct rotamers, the choice of substitutions is "quantized" (Karpusas et al., 1989). On the other hand, if the backbone were allowed to move it would allow a wider range of substitutions to be considered.

MATERIALS AND METHODS

Redesign by ORBIT

All residues of cysteine-free pseudo-wildtype T-4 lysozyme, referred to as WT*, were classified as surface, core, or boundary, using a residue classification program, RESCLASS (Dahiyat and Mayo 1997) RESCLASS classifies the residues based on their Ca and Cß distances from a solvent accessible surface, which is the calculated using the Connolly algorithm (Connolly, 1983).

We selected 26 core positions located in the C-terminal domain of WT* for design. The selected positions were I3, M6, A74, I78, L84, V87, Y88, L91, A97, A98, L99, I100, M102, V103, M106, V111, I118, Y120, L121, A129, L133, A146, V149, I150, T152 and F153. Positions 3 and 91 were classified as boundary residues but were nevertheless included in the core calculations as visual inspection showed them to be significantly buried. Position 6 and 3 belong to the N-terminal domain but were considered for design

because of their close proximity to the C-terminal core residues. The hydrophobic amino acids allowed at all 26 positions were Ala, Val, Phe, Leu, Ile, Phe, Tyr, Trp and Met. An expanded version of the backbone dependent rotamer library of Dunbrack and Karplus was used for the calculations (Dunbrack R. L. and Karplus M., 1993). For aromatic residues, the expansions included the mean χ values $\pm$ 1 standard deviation about $\chi 1$ and $\chi 2$ torsional angles. For other hydrophobic groups, a similar expansion was performed, but was limited only to the $\chi 1$ torsional angle. Energies for the point mutants were calculated by fixing the identities of amino acids at all 26 positions while allowing their rotameric conformations to vary based on the rotamer library. The design calculations were run using an optimization procedure based on Dead-End Elimination algorithm (Desmet 1992, Pierce 2000).

Mutagenesis, protein expression, and purification

The two redesigns of the C-terminal core of bacteriophage T4 lysozyme, Core-7 and Core-10, were made by iterative two stage PCR (Landt et al., 1990) using the gene for the cysteine-free (C54T/C97A) pseudo-wildtype (WT*) T4 lysozyme as the template. The BamHI/HindIII digested PCR products were ligated into the vector PH1403. The single (where they did not previously exist), double, and revertant mutants were made by the inverse PCR (Hemsley et al., 1989). The gene for WT*, Core-10, or Core-7 in the vector PH1403 was used as the template. The individual single site mutants (relative to WT*) were drawn from existing stocks except for I78V, V87I, I100V, V103I, M106I, L118I, M120Y, and L133F. The double mutant V149/T152V was made in the WT* background. The DNA sequences of the new constructs were confirmed by automated methods incorporating the polymerase chain reaction (Perkin-Elmer ABI PRISM 377 DNA sequencer). The vectors were transformed into *E. coli* RR1 cells for over-expression. The mutant proteins were over-expressed and purified by standard methods (Alber & Matthews, 1987; Muchmore et al., 1989; Poteete et al., 1991). The molecular mass of the mutant

proteins were checked with a Perspective Biosystems Voyager-DE MALDI/TOF mass spectrometer. The buffer used for protein storage was 0.1 M sodium phosphate pH 6.5, 0.55 M NaCl, and 0.02% NaN_3. As judged by the fact that each lysozyme caused cell lysis and behaved similarly during purification we assume that all have activity similar to that of WT*.

Thermal Unfolding

Circular dichroism monitored thermal stability data were collected at 223 nm using a JASCO model J-600 spectropolarimeter and the Hewlett-Packard model HP89101 thermal control system (Eriksson et al., 1993). The buffer was 0.10 M sodium chloride, 1.4 mM acetic acid, 8.6 mM sodium acetate, pH 5.35, with protein concentrations of 0.01 to 0.03 mg/ml as determined from optical density at 280 nm (Elwell & Schellman, 1975). Unfolding profiles were analyzed by means of the two-state model to determine the temperature of melting (T_m) and the van't Hoff enthalpy at the melting temperature (ΔH) (Zhang et al., 1995). At least three independent trials were done for each mutant. Averaged values of T_m and ΔH were used to calculate $\Delta G°$ at 61°C by means of an integrated form of the Gibbs-Helmholtz equation (Hawkes et al., 1984) assuming a ΔCp of 2.5 kcal/mol-K. $\Delta\Delta G$ values were computed as $\Delta G°$(mutant) - $\Delta G°$(WT*).

Crystallization

It was possible to crystallize both of the two designed proteins, selected single mutant back revertant proteins, and the previously unpublished single mutants. Thirteen of the fifteen new proteins were crystallized in space group $P3_221$ isomorphously with the wildtype protein in 2 M K/Na phosphate buffers as previous described (Eriksson et al., 1993). Core-7 crystallized in space group C2 in 100 mM Na/K phosphate buffer pH 6.7 and 20% MPD. V103I crystallized in space group $P2_12_12_1$ in solutions of 0.1 M Hepes pH 7.5, 20% PEG3400, and 5% isopropanol. M87V/Core-7 crystallized in space group C2 in

25% PEG3400, 5% PEG600, 200 mM NaCl, 100 mM Na/K phosphate, pH 6.7 and Fos-choline 12 at its critical micelle concentration.

X-ray Data Collection

Since the 100 K structure of the pseudo wildtype had been used as the template in the design process, X-ray data of the new proteins were collected at 100 K. Crystals of proteins grown from the high salt solutions were mounted in paratone and flash cooled. Crystals of Core-7 and of V103I were flash cooled in rayon loops containing cryogenic reservoir solutions. X-ray data for Core-7 and I103V/Core-10 were collected at beamline 7-1 at SSRL with monochromatic radiation having a wavelength of 1.06 Å and a MAR image plate. X-ray data for the remaining structures were collected in-house with 1.54 Å radiation and a Rigaku RAXIS4 image plate. The data were integrated with Mosflm and scaled with Scala (Leslie, 1992; Evans, 1994)

Structure Determination

The structures of V103I and M87V/Core-7 were solved by molecular replacement using the program EPMR (Kissinger et al., 1999) while the remaining structures were determined by molecular substitution using the coordinates of WT* (Table A-2) as the starting model.

Structure Refinement

The crystal structures were refined using the refinement package TNT (Tronrud et al., 1987; Tronrud, 1997) following the procedures described previously (Eriksson et al., 1993). The Xfit molecular graphics module of XtalView was used for model rebuilding (MacRee ,1992).

Acknowledgments

We thank Hong Xiao, Leslie Gay, and Andy Fields for making the mutant proteins and for crystallizing them, Cathy Sarisky for her explanation of the methionine penalty , Doug Juers for collecting preliminary X-ray data for the mutant proteins V103I and Core-7, and the user support staff at SSRL and ALS for their assistance. This work was supported in part by grants from the NIH (GM21967 to B.W.M.)

REFERENCES

Alber, T., & Matthews, B. W. (1987) Temperature-sensitive mutation of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. Biochemistry 26, 3754-3758.

Anderson, D. E., Becktel, W. J., & Dahlquist, F. W. (1990) pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. Biochemistry 29, 2403-2408.

Baldwin, E. P., Hajiseyedjavadi, O., Baase, W. A., & Matthews, B. W. (1993) The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. Science 262, 1715-1718.

Baldwin, E. P., Xu, J., Hajiseyedjavadi, O., Baase, W. A., & Matthews, B. W. (1996) Thermodynamic and structural compensation in "size-switch" core repacking variants of bacteriophage T4 lysozyme. J. Mol. Biol. 259, 542-559.

Blaber, M, Zhang, X-J., Lindstrom, J.D., Pepiot, S.D., Baase, W.A., & Matthews, B.W. (1994) Determination of ✂-helix propensity within the context of a folded protein: Sites 44 and 131 in bacteriophage T4 lysozyme. J. Mol. Biol. 235, 600-624.

Connolly M.L. (1983) Analytical Molecular-Surface Calculation. *Journal Of Applied Crystallography.* 16: 548-558

Dahiyat, B. I. & Mayo, S. L. (1996) Protein design automation. Protein Sci. 5, 895-903.
Dahiyat, B. I. & Mayo, S. L. (1997a) Probing the role of packing specificity in protein design. Proc. Natl. Acad. Sci. USA 94, 10172-10177.

Dahiyat, B. I. & Mayo, S. L. (1997b) De novo protein design: fully automated sequence selection. Science 278, 82-87.

Dahiyat, B. I., Sarisky, C. A. & Mayo, S. L. (1997) De novo protein design: towards fully automated sequence selection. J. Mol. Biol. 273, 789-796.

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

Dunbrack RL, Karplus M. (1993). Backbone dependent rotamer library for proteins - an application to side-chain prediction. *J Mol Biol,* 230: 543-574.

Elwell, M., & Schellman, J. A. (1975) Phage T4 lysozyme. Physical properties and reversible unfolding. Biochim. Biophys. Acta 386, 309-323.

Eriksson, A. E., Baase, W. A., Zhang, X-J., Heinz, D. W., Blaber, M., Baldwin, E. P. &

Matthews, B.W. (1991) A cavity-containing mutant of T4 lysozyme is stabilized by buried benzene. Science 255, 178-183.

Eriksson, A. E., Baase, W.A., & Matthews, B. W. (1993) Similar hydrophobic replacements of Leu99 and Phe153 within the core of T4 lysozyme have different structural and thermodynamic consequences. J. Mol. Biol. 229, 747-769.

Evans, P. R. (1994) Scala. Joint CCP4 and ESF-EACBM Newsletter 33, 22-24.

Gassner, N. C., Baase, W. A., & Matthews, B. W. (1996) A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. Proc. Natl. Acad. Sci. USA 93, 12155-12158.

Gassner, N. C., Baase, W. A., Joel D. Lindstrom, J.D., Lu, J., Dalqusit, F. W., & Matthews, B.W. (1999) Methionine and alanine substitutions show that the formation of wild-type-like structure in the carboxy-terminal domain of T4 lysozyme is a rate-limiting step in folding. Biochemistry 38, 14451-14460.

Goldstein, R. F. (1994) Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys. J. 66, 1335-1340.

Grütter, M. G., Gray, T. M., Weaver, L. H., Alber, T., Wilson, K., & Matthews, B. W. (1987) Structural studies of mutants of the lysozyme of bacteriophage T4. The temperature sensitive mutant protein Thr 157 6 Ile. J. Mol. Biol. 197, 315-329.

Hawkes, R., Grutter, M. G., & Schellman, J. (1984) Thermodynamic stability and point mutations of bacteriophage T4 lysozyme. J. Mol. Biol. 175, 195-212.

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. Nucl. Acids Res. 17, 6545-6551.

Hurley, J. H., Baase, W. A., & Matthews, B. W. (1992) Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. J. Mol. Biol. 224, 1143-1159.

Janin, J. (1979) Surface and inside volumes in globular proteins. Nature 277, 491-492.

Janin, J., Wodak, S., Levitt M. & Maigret, B. (1978) Conformation of amino acid side-chains in proteins. J. Mol. Biol. 125, 357-386.

Karpusas, M., Basse, W.A., Matsumura, M. & Matthews, B.W. (1989) Hydrophobic packing in T4 lysozyme probed by cavity-filling mutants. Proc. Natl. Acad. Sci. USA 86, 8237-8241.

Kissinger, C. R., Gehlhaar, D. K., & Fogel, D. B. (1999) Rapid automated molecular

replacement by evolutionary search. Acta Cryst. D55, 484-491.

Landt, O., Grunert, H. & Hahn, U. (1990) A general method for rapid site-directed mutagenesis using the polymerase chain reaction. Gene 96, 125-128.

Leslie, A. G. W. (1992) Recent changes to the MOSFLM package for processing film and image plate data. Joint CCP4 and ESF-EAMCB Newsletter on Protein Crystallography No. 26.

Liang, J. & Dill, K. A. (2001) Are proteins well-packed? Biophys. J. 81, 751-766.

Lipscomb, L. A., Gassner, N. C., Snow, S. D., Eldridge, A. M., Basse, W. A., Drew, D. I., & Matthews, B. W. (1998) Context-dependent protein stabilization by methionine-to-leucine substitution shown in T4 lysozyme. Protein Sci. 7, 765-773.

MacRee, D. E. (1992) A visual protein crystallographic software system for X11/Xview. J. Mol. Graphics 10, 44-46.

Malakauskas, S. M., & Mayo, S. L. (1998) Design, structure and stability of a hyperthermophilic protein variant. Nat. Struct. Biol. 5, 470-475.

Matsumura, M., & Matthews, B. W. (1989) Control of enzyme activity by an engineered disulfide bond. Science 243, 792-794.

Miller, S., Janin, J., Lesk, A. M., Chothia, C. (1987) Interior and surface of monomeric proteins. J. Mol. Biol. 196, 641-656.

Muchmore, D. C., McIntosh, L. P., Russell, C. B., Anderson, D. E., Dahlquist, F. W. (1989) Expression and 15N labelling of proteins for proton and nitrogen-15 NMR. Methods Enzymol. 177, 44-73.

Myers, J. K., Pace, C. N., & Scholtz, J. M. (1995) Denaturant *m* values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. Prot. Sci. 4, 2138-2148.

Ponder, J. W. & Richards, F. M. (1987) Tertiary templates for proteins--use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193, 775-791.

Poteete. A. R., Doa-pin, S., Nicholson, H., & Matthews, B. W. (1991) Second-site revertants of an inactive T4 lysozyme mutant restore activity structuring the active site cleft. Biochemistry 30, 1425-1432.

Richards, F. M. (1986) Protein design: Are we ready? Prot. Struct. Fold. Des., 171-196.

Richards, F. M., & Lim, W. A. (1994) An analysis of packing in the protein folding

problem. Quart. Rev. Biophys. 26, 423-498.

Su, A. & Mayo, S. L. (1997) Coupling backbone flexibility and amino acid sequence selection in protein design. Protein Sci. 6, 1701-1707.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987) An efficient general-purpose least-squares refinement program for macromolecular structures. Acta Cryst. A43, 489-501.

Tronrud, D. E. (1997) TNT refinement package. Methods Enzymol. 277, 306-319.

Weaver, L.H., & Matthews, B. W. (1987) Structure of bacteriophage T4 lysozyme refined at 1.7Å resolution. J. Mol. Biol. 193, 189-199.

Xu, J., Baase, W. A., Quillin, M. L., Baldwin, E. P. & Matthews, B. W. (2001) Structural and thermodynamic analysis of the binding of solvent at internal sites in T4 lysozyme. Protein Sci. 10, 1067-1078.

Zhang, X-J., Baase, W.A., Shoichet, B. K., Wilson, K. P., & Matthews, B. W. (1995) Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive. Prot. Eng. 8, 1017-1022.

Table A-1. Stabilities of mutant lysozymes. The first two columns give the score calculated by ORBIT, respectively, without and with a penalty for incorporation of methionines (see text). ΔT_m is the change in melting temperature relative to WT* which is 65.5°C under these conditions. ΔH is the enthalpy of unfolding at T_m. $\Delta\Delta G$ is the change in the free energy of unfolding relative to WT*. "Non-additivity of $\Delta\Delta G$" is the difference between $\Delta\Delta G$ measured for the multiple construct and the sum of the $\Delta\Delta G$'s for the constituent single mutants. Uncertainties in ΔT_m are about $\pm$0.2°C, in ΔH about $\pm$5% and in $\Delta\Delta G$ about 0.15 to 0.4 kcal/mol (increasing from the more-stable to the least-stable mutants). As is also explained in the text more negative ORBIT scores correspond to proteins that are predicted to be more stable whereas more negative $\Delta\Delta G$ values correspond to proteins that are of lower stability.

Mutant	Score (kcal/mol)	ORBIT score with penalty (kcal/mol)	ΔT_m (°C)	ΔH (kcal/mol)	$\Delta\Delta G$ (kcal/mol)	Non additivity of $\Delta\Delta G$ (kcal/mol)
I78V	-364	-322	-2.1	127	-0.8	--
V87M	-349	-299	-6.3[a]	113[a]	-2.3[a]	--
L118I	-352	-321	-3.1	123	-1.2	--
V87I	-362	-330	-0.8	127	-0.3	--
I100V	-365	-333	-1.1	129	-0.4	--
M102L	-316	-292	-2.3	118	-1.0	--
V103I	-295	-263	-1.5	130	-0.5	--
M106I	-362	-338	0.6	132	0.2	--
V111A	-354	-322	-2.9	121	-1.1	--
M120Y	-365	-335	-0.1	126	-0.1	--
L133F	-368	-336	-0.7	130	-0.3	--
V149I	-362	-323	-0.3	128	0.0	--
T152V	-365	-333	0.8[b]	127[b]	0.2[b]	--
Core-7	-382	-341	-9.8	103	-3.5	1.0
M87V/Core-7	-368	-344	-5.0	117	-3.0	-0.8
I118L/Core-7	-381	-350	-9.5	103	-3.3	0.0
Core-10	-371	-363	-6.4	97	-2.4	1.1
L102M/Core-10	-372	-356	-7.2	101	-2.6	-0.1
I103V/Core-10	-370	-363	-4.0	110	-1.6	1.3
A111V/Core-10	-269	-261	-4.8	106	-1.8	0.5
WT*	-345	-305	0.0	132	0.0	--

[a]From Gassner et al. (1999).

[b]From Xu et al. (2001). Note that ΔH is a corrected value.

Table A-2. Crystal and refinement statistics.

	Cell dimensions							
Protein	***a,b*** **(Å)**	***c*** **(Å)**	**Resolution (Å)**	**R_{merge} (%)**	**Completeness (%)**	**R-factor (%)**	**D_{bonds} (Å)**	**D_{angles} (°)**
WT*	60.1	95.5	1.05	3.4	95.3 (88)	12.85	0.019	2.7
I78V	60.0	95.23	1.58	4.6	94.6 (88)	19.6	0.017	2.5
V87I	59.6	95.3	1.58	5.5	94.1 (77)	17.2	0.016	2.3
I100V	59.8	95.6	1.45	6.0	97.9 (86)	19.1	0.015	2.4
V103I	(a)	(a)	1.5	5.6	95.3 (78)	19.0	0.018	2.6
M106I	60.1	95.6	1.67	4.6	96.9 (84)	18.5	0.018	2.5
L118I	60.2	95.9	1.65	4.9	94.0 (91)	20.1	0.017	2.6
M120Y	60.3	95.3	1.54	5.1	97.7 (96)	18.7	0.017	2.5
L133F	60.1	96.2	1.62	4.4	96.5 (79)	18.9	0.016	2.3
V149I/T152V	59.8	95.4	1.52	5.8	93.5 (71)	17.6	0.016	2.5
M87V/Core-7	(b)	(b)	1.56	5.2	96.0 (91)	18.6	0.020	2.9
I118L/Core-7	60.0	95.6	1.56	4.9	91.1 (91)	19.8	0.016	2.7
Core-10	60.0	96.6	1.65	7.6	96.9 (85)	17.8	0.016	2.7
L102M/Core-10	59.5	96.2	1.57	6.1	90.6 (75)	17.7	0.018	2.5
I103V/Core-10	60.0	95.9	1.55	6.0	96.7 (97)	18.7	0.015	2.3
A111V/Core-10	59.5	95.5	1.90	6.1	99.0 (100)	18.8	0.016	2.6

(a) V103I crystallized in space group $P2_12_12_1$ with cell dimensions a = 30.8 Å, b = 54.9 Å and c = 88.4 Å

(b) M87V/Core-7 crystallized in space group C2 with cell dimensions a = 156.5 Å, b = 61.9 Å, c = 67.4 Å, β = 112.3°.

Table A-3. Backbone shifts in designed and mutant T4 lysozymes.Each entry in the table gives the root-mean-square difference between the main-chain atoms of the specified structure and WT*. The column labeled "C-terminal domain" gives the rms shift for essentially the whole C-terminal domain (i.e. for residues 81-161). The column labeled "C-terminal domain without helices F and G" gives the rms shifts for residues 81-105 plus 124-161. Superpositions were carried out using EDPDB (Zhang & Matthews, 1995).

Protein	Shift, C-terminal domain (Å)	Shift, C-terminal domain without helices F and G (Å)
Core-10 design	0.19	0.23
Core-10 crystal	0.49	0.21
L102M/Core-10	0.49	0.22
I103V/Core-10	0.22	0.18
A111V/Core-10	0.55	0.26
M87V/Core-7		
Molecule A	0.40	0.36
Molecule B	0.49	0.50
Molecule C	0.44	0.46
I118L/Core-7	0.28	0.25

TableA- 4. Comparison of temperature factors at mutated sites in WT* and Core-10. The Wilson B-value is 14.4 Å² for WT* and 17.9 Å² for Core-10.

	Main-chain B (Å²)		Side-chain B (Å²)	
Residue	WT*	Core-10	WT*	Core-10
87	13.1	20.0	17.7	26.0
100	11.6	17.2	13.0	15.4
102	12.3	17.0	13.5	17.3
103	13.5	19.9	15.7	23.3
106	17.6	21.4	16.6	21.5
111	18.1	33.5	16.7	27.8
120	12.1	15.9	16.4	19.4
133	11.1	13.9	12.3	14.4
149	10.2	16.3	11.4	15.1
152	11.3	19.2	11.4	14.1

Table A-5. Comparison of the side-chain torsion angles at the 26 sites open to modification in T4 lysozyme. The torsion angles are listed for the crystal structure of WT*, the energy minimized model of WT* used in the design process, the predicted structure of Core-10, and the observed crystal structure. The five sites that started as alanine and remained alanine (sites 74, 97, 98, 129, 146) are not shown. The IUPAC conventions for determining χ_1 and χ_2 were followed except for the following two changes which were made to simplify comparison of unlike side-chains. (1) Following Blaber et al. (1994) the χ_1 torsion angle of valine was measured using the CG2 carbon atom rather than CG1 as in the standard IUPAC nomenclature. This is about the same as increasing χ_1 by 120° and makes the *gauche-*, *trans* and *gauche+* conformations for valine the same as for the other amino acids. (2) For the phenylalanine and tyrosine side-chains marked with an asterisk the χ_2 value was decreased by 180°. This change essentially corresponds to a renaming of the ring atoms. χ_3 values are not shown but in general agree fairly well at any given site. At sites 6, 102, 106 and 120 the maximum discrepancy in χ_3 among the structures being compared is, respectively, 21°, 2°, 19° and 3°.

	WT* (c_1,c_2)	**Energy minimized WT*** (c_1,c_2)	**Core-10 design** (c_1,c_2)	**Core-10 crystal structure** (c_1,c_2)
3	Ile (183, 57)	Ile (186, 52)	Ile (185, 62)	Ile (186, 60)
6	Met (184, 201)	Met (188, 197)	Met (185, 191)	Met (191, 213)
78	Ile (285, 163)	Ile (290, 161)	Ile (285, 168)	Ile (277, 163)
84	Leu (302, 175)	Leu (305, 175)	Leu (300, 174)	Leu (304, 173)
87	Val (307, --)	Val (309, --)	Ile (282, 49)	Ile (307, 313)
88	Tyr (184, 82)	Tyr (185, 82)	Tyr (178, 93)	Tyr (186, 85)
91	Leu (297, 168.5)	Leu (291, 173.1)	Leu (268, 32.1)	Leu (292, 170.8)
100	Ile (296, 163)	Ile (305, 158)	Val (303, --)	Val (292, --)
102	Met (293, 186)	Met (297, 186)	Leu (193, 63)	Leu (178, 68)
103	Val (293, --)	Val (293, --)	Ile (282, 49)	Ile (283, 17)
106	Met (76, 182)	Met (72, 179)	Ile (65, 172)	Ile (78, 183)
111	Val (303, --)	Val (305, --)	Ala (--, --)	Ala (--, --)
118	Leu (293, 169)	Leu (288, 167)	Leu (289, 176)	Leu (291, 167)
120	Met (300, 175)	Met (294, 171)	Tyr (276, 153*)	Tyr (283, 124*)
121	Leu (290, 172)	Leu (295, 170)	Leu (282, 174)	Leu (289, 177)
133	Leu (282, 164)	Leu (286, 161)	Phe (271, 104*)	Phe (267, 114*)
149	Val (296, --)	Val (299, --)	Ile (298, 169)	Ile (296, 280)
150	Ile (292, 171)	Ile (289, 168)	Ile (286, 169)	Ile (288, 176
152	Thr (307, --)	Thr (311, --)	Val (296, --)	Val (297, --)
153	Phe (280, 303)	Phe (280, 299)	Phe (275, 324)	Phe (282, 308)

Figure A-1. (a) C^{α} trace of the WT* T4 lysozyme backbone showing, in red, the 26 sites that were allowed to vary during the design process. The sites are identified at the left.

V87I
I100V
M102L
V103I
M106I
V111A
M120Y
L133F
V149I
T152V

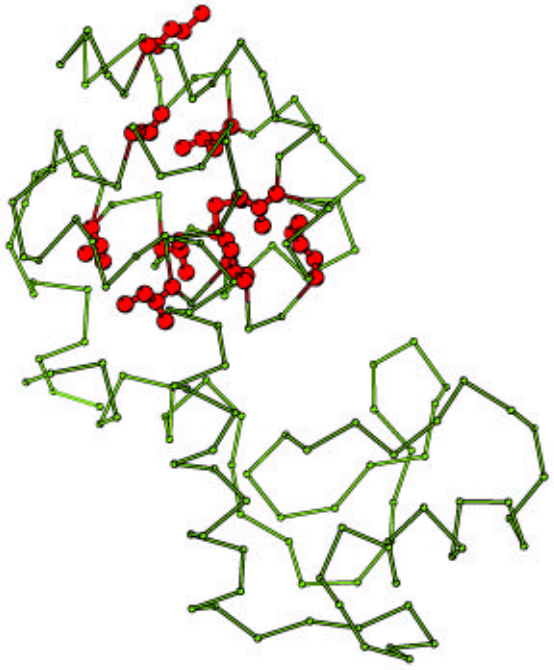


Figure A-1(b) Structure of T4 lysozyme showing the ten sites that were substituted in Core-10.

I3
M6
A74
I78
L84
V87
Y88
L91
A97
A98
L99
I100
M102
V103
M106
V111
L118
M120
L121
A129
L133
A146
V149
I150
T152
F153

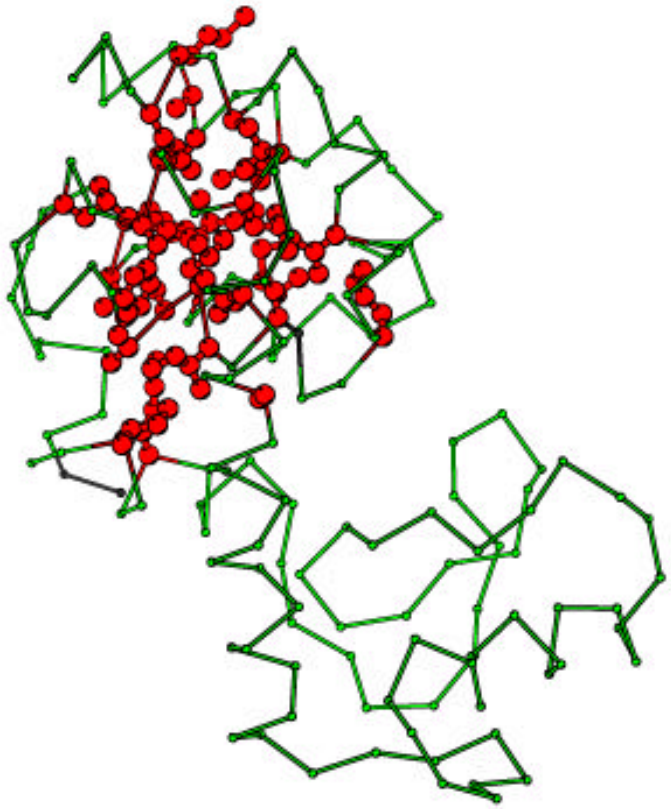
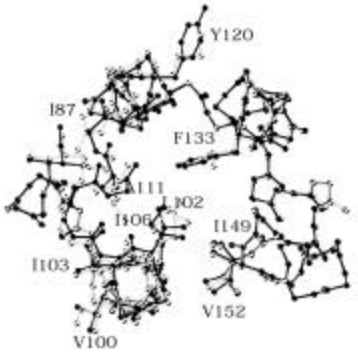
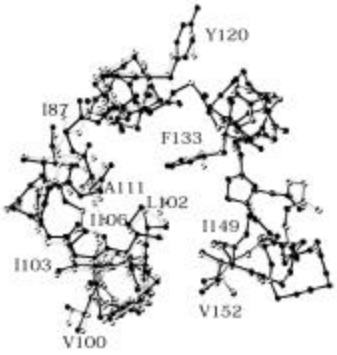
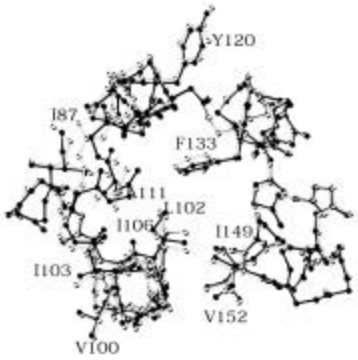


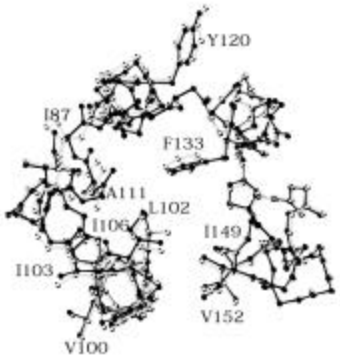
Figure A-2. (a) Stereo view showing the superposition of the crystal structure of Core-10 (open bonds) on the crystal structure of WT* (solid bonds). For clarity only the side-chains of the ten substituted residues are shown. (b) Superposition of designed Core-10 (solid bonds) on to the observed crystal structure (open bonds).


Y120
I87
F133
I111
L102
I106
I149
I103
V152
V100


Y120
I87
F133
A111
L102
I106
I149
I103
V152
V100

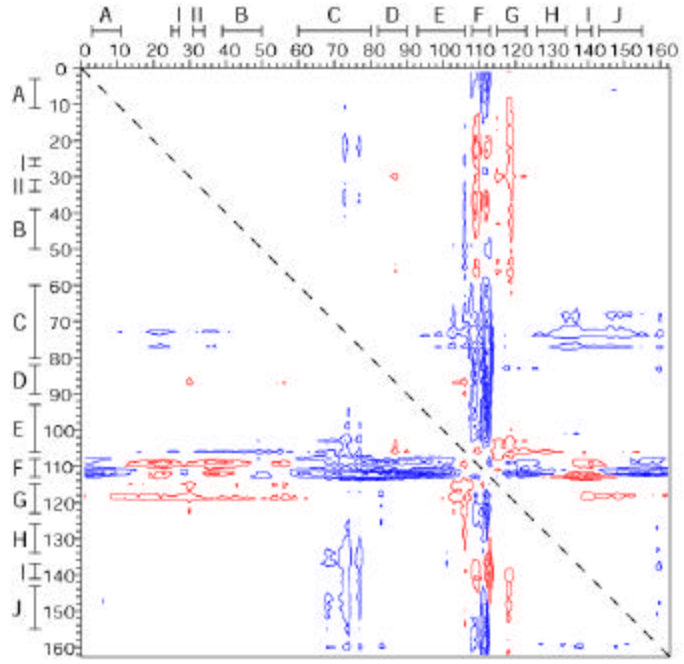
(a)


Y120
I87
F133
A111
L102
I106
I149
I103
V152
V100

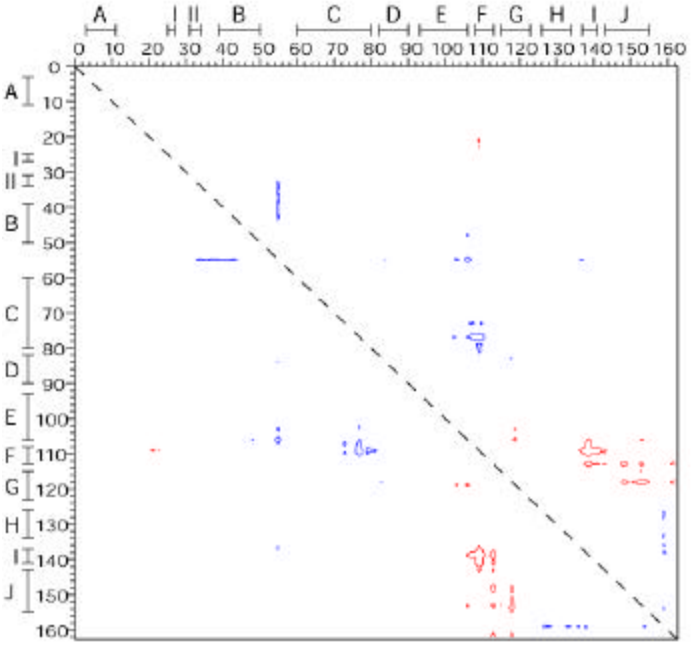

Y120
I87
F133
A111
L102
I106
I149
I103
V152
V100

(b)

Figure A-3. Plots showing differences in C^{α}-C^{α} separation in different crystal structures. The contours start at $\pm$0.5 Å and have 0.5 Å intervals. The red contours correspond to decreased separation and the blue contours correspond to an increase in distance. (a) Core-10 *versus* WT*. (b) I103V/Core-10 *versus* WT*.

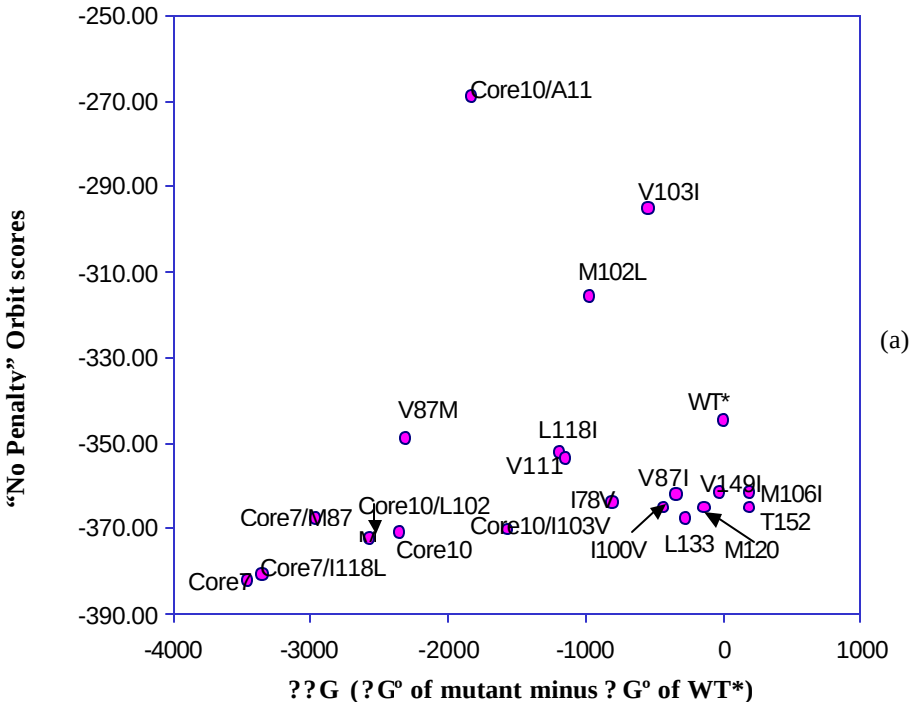

A I II B C D E F G H I J
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160

(a)

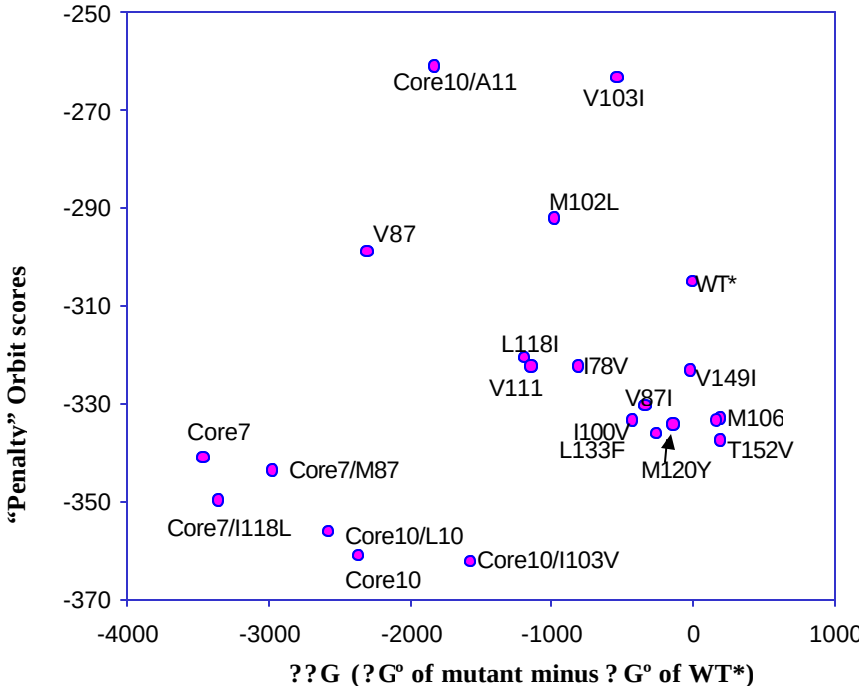

A I II B C D E F G H I J
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160

(b)

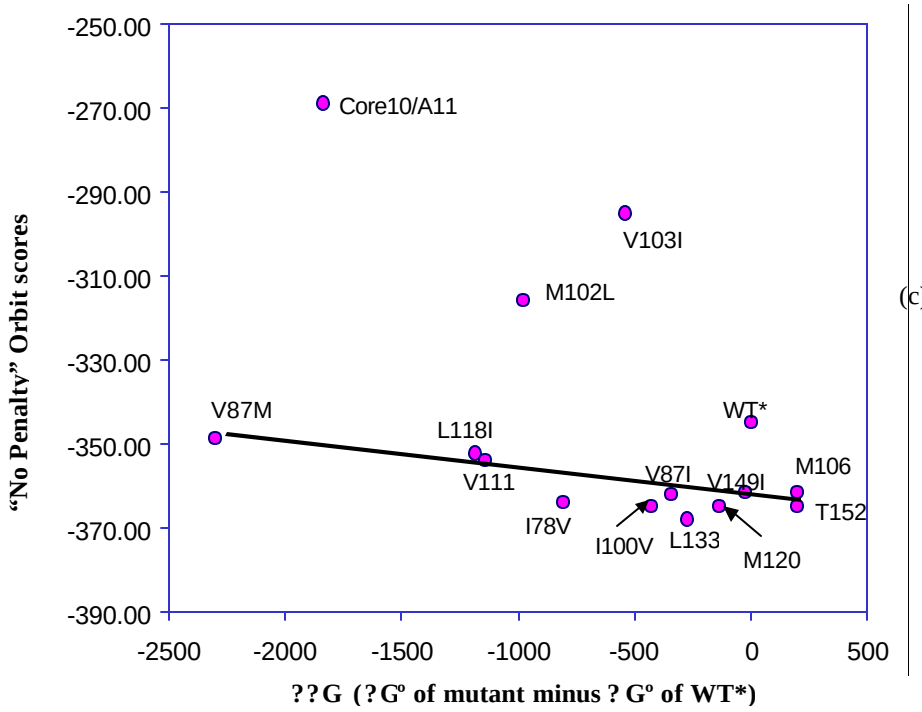
Figure A-4. Comparison of the energies calculated using ORBIT with the observed protein stability (Table A-1). (a) Comparison of single and multiple mutants with the ORBIT score determined without a penalty for incorporation of methionine. (b) Comparison of single and multiple mutants with the ORBIT score determined with a penalty for incorporation of methionine. (c) Comparison of single mutants with the ORBIT score determined without a penalty for incorporation of methionine. The straight line is the best fit to the data excluding M102L, V103I and WT*. (d) Comparison of single mutants with the ORBIT score determined with a penalty for incorporation of methionine. The straight line is the best fit to the data excluding M102L, V103I and WT*.


-250.00
-270.00
-290.00
-310.00
-330.00
-350.00
-370.00
-390.00
-4000
-3000
-2000
-1000
0
1000
"No Penalty" Orbit scores
??G (?Gº of mutant minus ?Gº of WT*)
Core10/A11
V103I
M102L
V87M
WT*
L118I
V111
V87I
V149I
M106I
I78V
T152
Core7/M87
Core10/L102
Core10/I103V
I100V
L133
M120
Core10
Core7/I118L
Core7

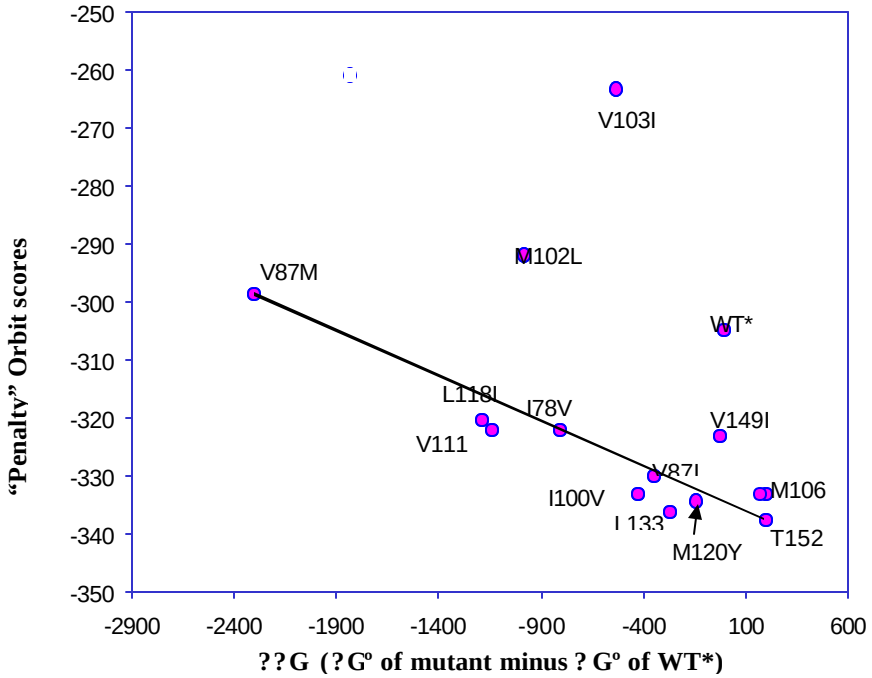
(a)


-250
-270
-290
-310
-330
-350
-370
-4000
-3000
-2000
-1000
0
1000
"Penalty" Orbit scores
??G (?Gº of mutant minus ?Gº of WT*)
Core10/A11
V103I
M102L
V87
WT*
L118I
I78V
V149I
V111
V87I
M106
I100V
T152V
L133F
M120Y
Core7
Core7/M87
Core7/I118L
Core10/L10
Core10/I103V
Core10

(b)

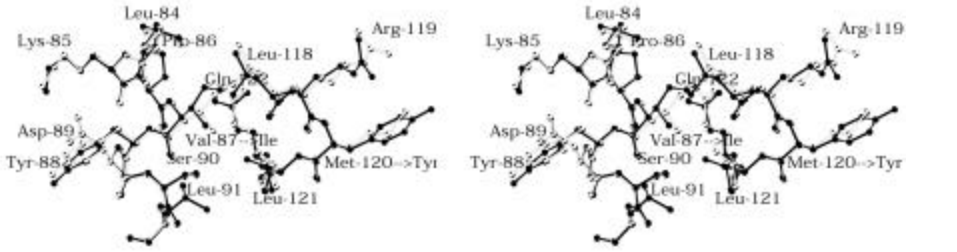

-250.00
-270.00
-290.00
-310.00
-330.00
-350.00
-370.00
-390.00
-2500
-2000
-1500
-1000
-500
0
500
"No Penalty" Orbit scores
??G (?Gº of mutant minus ?Gº of WT*)
Core10/A11
V103I
M102L
V87M
L118I
V111
WT*
V87I
V149I
M106
T152
I78V
I100V
L133
M120

(c)

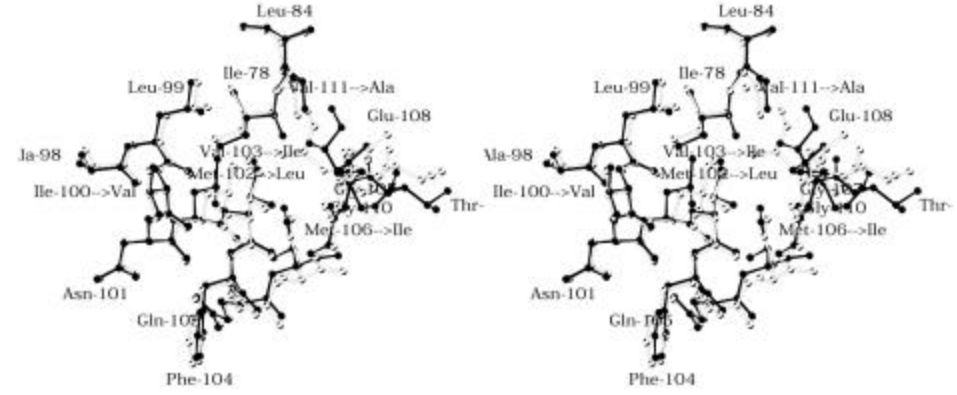

-250
-260
-270
-280
-290
-300
-310
-320
-330
-340
-350
-2900
-2400
-1900
-1400
-900
-400
100
600
"Penalty" Orbit scores
??G (?Gº of mutant minus ?Gº of WT*)
V103I
M102L
V87M
WT*
L118I
I78V
V149I
V111
M106
I100V
T152
M120Y

(d)

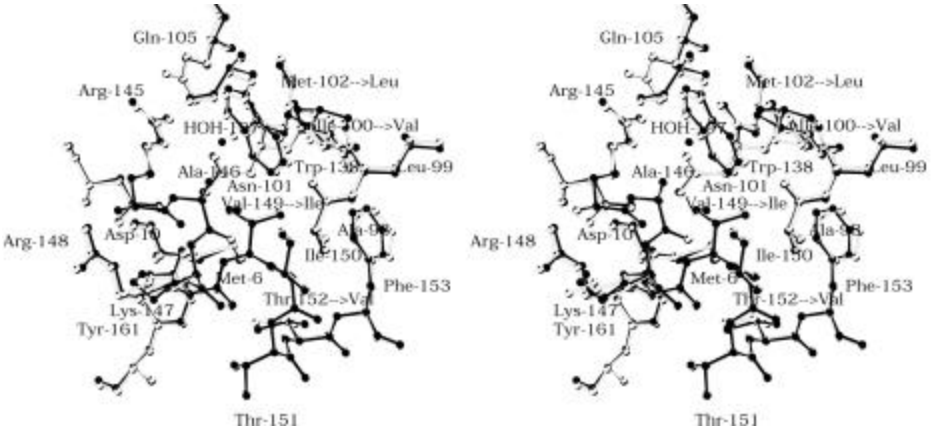
Figure A-5. Stereo views of the sites of discrepancies between the predicted and observed structure of Core-10. (a) Crystal structure of WT* (open bonds) superimposed on the crystal structure of Core-10 (closed bonds) in the vicinity of site 87. (b) Design of Core-10 (open bonds) superimposed on the crystal structure of Core-10 (closed bonds) in the vicinity of site 102. (c) Crystal structure of WT* (open bonds) superimposed on the crystal structure of Core-10 (closed bonds) in the vicinity of site 149.


Leu-84
Lys-85
Pro-86
Arg-119
Leu-118
Asp-89
Val-87-->Ile
Tyr-88
Ser-90
Met-120-->Tyr
Leu-91
Leu-121

(a)


Leu-84
Ile-78
Leu-99
Val-111-->Ala
Glu-108
Ala-98
Val-103-->Ile
Met-102-->Leu
Ile-100-->Val
Met-106-->Ile
Asn-101
Phe-104

(b)


Gln-105
Met-102-->Leu
Arg-145
Ile-100-->Val
Trp-138
Leu-99
Asn-101
Val-149-->Ile
Arg-148
Ile-150
Phe-153
Met-6
Thr-152-->Val
Lys-147
Tyr-161
Thr-151

(c)

Figure A-6. Comparison of the observed stabilities of the multiple mutants ($\Delta\Delta G$, Table A-1, red bars) with the sums of the $\Delta\Delta G$'s of the constituent single mutants (purple bars.)

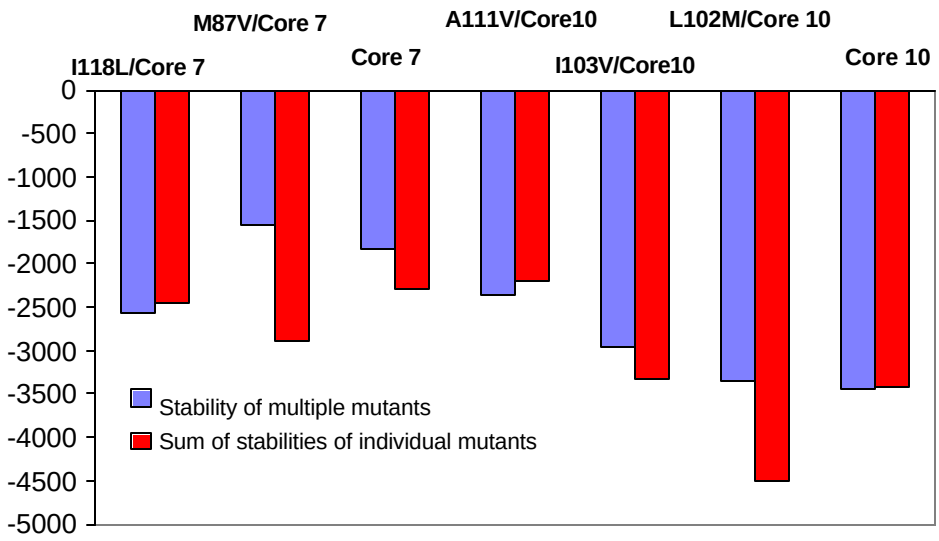
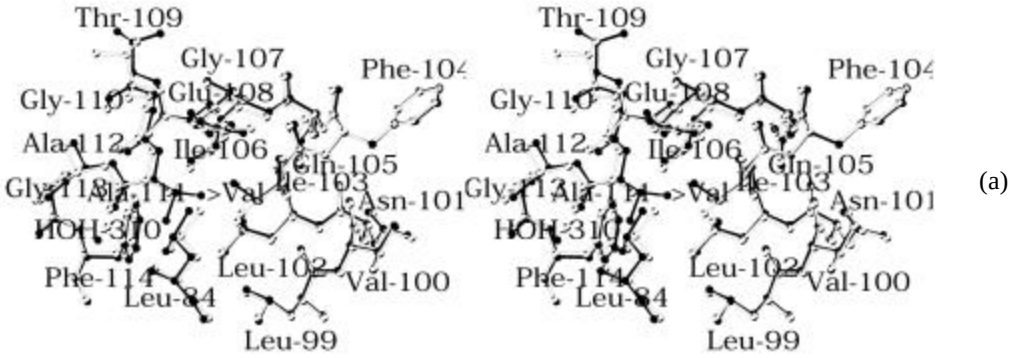
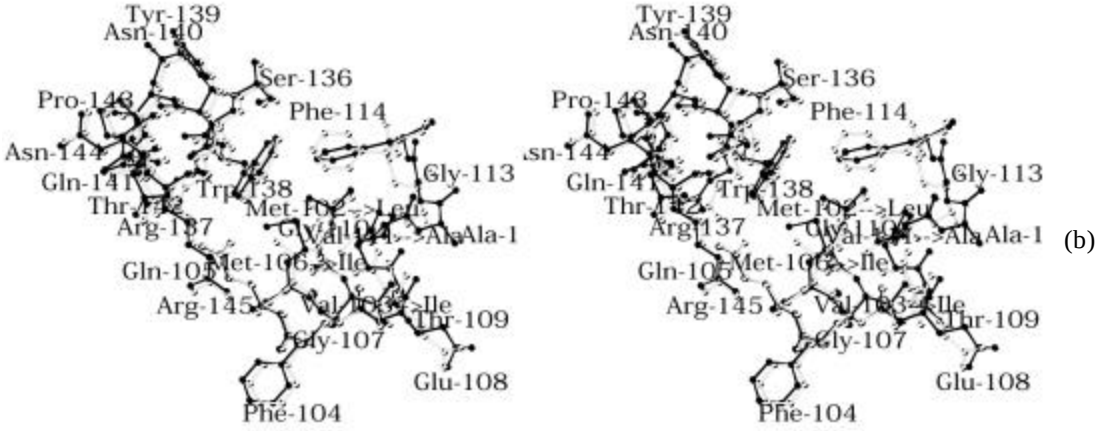

I118L/Core 7
M87V/Core 7
Core 7
A111V/Core10
I103V/Core10
L102M/Core 10
Core 10
0
-500
-1000
-1500
-2000
-2500
-3000
-3500
-4000
-4500
-5000
Stability of multiple mutants
Sum of stabilities of individual mutants

Figure A-7. (a) Stereo diagram showing the superposition of the single-site revertant A111V/Core-10 (solid bonds) on Core-10 (open bonds). (b) Superposition of the structure of the single mutant, M102L (Hurley et al., 1992) (open bonds) on Core-10 (solid bonds).


Thr-109
Gly-107
Phe-104
Gly-110
Glu-108
Ala-112
Ile-106
Gln-105
Gly-111
Ala-114
->Val
Ile-103
Asn-101
HOH-310
Phe-114
Leu-102
Val-100
Leu-84
Leu-99

(a)


Tyr-139
Asn-140
Ser-136
Pro-148
Phe-114
Asn-144
Gln-141
Gly-113
Trp-138
Thr-142
Met-106->Leu
Arg-137
Val-104->Ala
Ala-1
Gln-105
Met-106->Ile
Arg-145
Val-103->Ile
Thr-109
Gly-107
Glu-108
Phe-104

(b)

Appendix B

Using Positional Bias for Minimizing Surface Charge of Ubiquitin

This work was done in collaboration with Andy Robertson's group in University of Iowa, The predicted ubiquitin variants are being made and tested in his group.

Our principal interest in minimizing charge on ubiquitin is to reduce or remove the effects of coulombic interactions on pKs. The idea is that we could explore ionization behavior of isolated residues in a protein background where we could identify how factors such as hydrogen-bonding and solvation affect the properties of ionizable side chains. In "normal" proteins, charge-charge interactions probably dominate ionization behavior and it is thus difficult to identify the other effects with confidence. The broader significance of all this is gaining a better understanding of the chemical and physical properties of protein surfaces.

Ubiquitin residues were classified as core, surface and boundary using a residue classification program, RESCLASS. We considered only the 12 charged surface positions (6, 16, 18, 24, 39, 51, 52, 63, 64, 68, 72, 74) for design (Figure B-1). Amino acids allowed at each design position were Thr, Ser, Gln, Asn, Ala and the wild type charged amino acid. The non-design surface residues were fixed to have the wild type amino acid identity but were allowed to vary in their rotameric conformations. We used our most expanded rotamer library for all calculations. A Secondary structure propensity term was applied to beta sheet surface positions (4, 16, 64, 68, 72) and calculations were run with propensity scale factors of 0.0, 1.0 and 2.0. Using this term in our calculations helps in selection of good ß-sheet forming residues. We used a sequence bias program (SBIAS) to rank order the mutations predicted in our calculations. This program can be used to direct the amino selection towards the wild type sequence or force the amino acid to be different from the wild type. This is accomplished by adding a bias energy to the design positions. A charged position that requires a very high bias energy to select for the wild

type identity is considered less important for the structural stability of the molecule than a position that requires a very low energy and hence, a predicted neutral residue can replace it more easily. We have ranked all the 12 charged positions considered in the calculations by varying the bias energy from high to low. The mutations predicted in calculations run with different secondary structure propensity scale factors are listed in tables B-2, B-3 and B4. Table B-1 represents the minimum number of changes suggested by the calculations to neutralize the charged residues on the protein surface. It is possible that the protein may tend lose stability after a few mutations. At that point, it may be necessary to redesign some of the other surface positions that could help to accommodate more charged to neutral mutations.

Table B-1. Mutations predicted by SBIAS. The mutations are cumulative as we go down each column. Mutations predicted by using scale factors of 1.0 and 2.0 seem to be more interesting than those predicted with scale factor of 0.0 since there is a higher selection of threonines on the ß-sheet surface positions.

Secondary structure propensity 0.0	**Secondary structure propensity 1.0**	**Secondary structure propensity 2.0**
H68Q	H68Q	H68T
E18N	E18N	E16T
E16N	E16N	E64Q
D39T	E64Q	E18N
E64Q	R72T	R72T
R72Q	E24Q	K6T
R74Q	D39T, K63Q, R74Q	E24Q
K6N, E24Q	K6Q	K63Q
K63Q	D52Q	D39N, R74Q
D52Q	E51N	D52Q
E51N		E51N

Table B-2. Sequences predicted from calculations with varying bias energies. A secondary structure propensity scale factor of 0.0 was used for the charged positions.

	2	4	6	8	9	12	14	16	18	20	22	24	39	49	51	52	57	60	62	63	64	66	68	70	71	72	73	74
Wildtype	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	H	V	L	R	L	R
Bias10.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias9.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias8.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias7.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias6.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias5.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias4.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias3.0	Q	F	K	L	T	T	T	N	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias2.0	Q	F	K	L	T	T	T	N	N	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias1.0	Q	F	K	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	E	T	Q	V	L	Q	L	R
NOBIAS	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	E	T	Q	V	L	Q	L	Q
Bias-1.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-2.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-3.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-4.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-5.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-6.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	Q	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-7.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	N	Q	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q

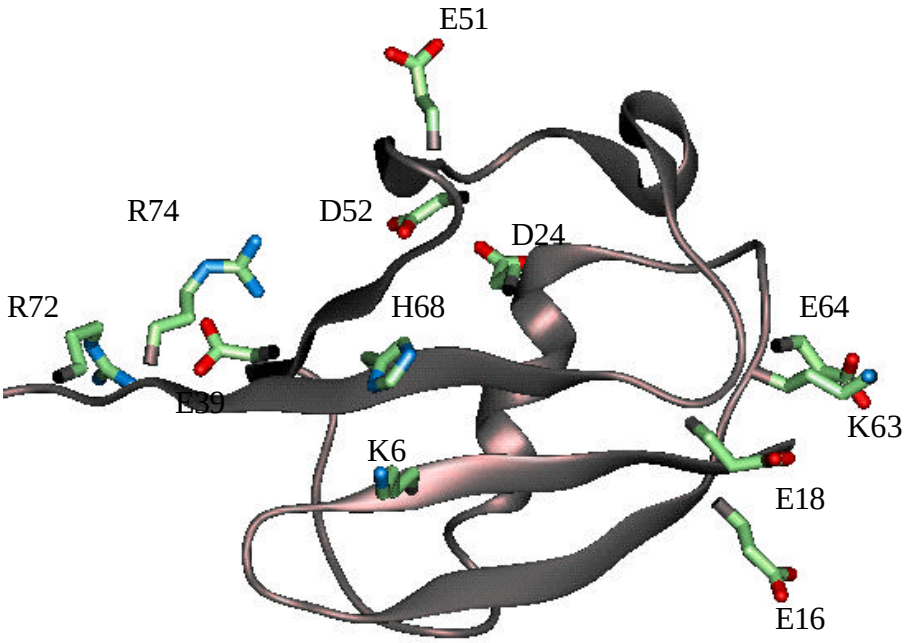
Table B-3. Sequences predicted from calculations with varying bias energies. A secondary structure propensity scale factor of 1.0 was used for the charged positions.

	2	4	6	8	9	12	14	16	18	20	22	24	39	49	51	52	57	60	62	63	64	66	68	70	71	72	73	74
Wild type	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	H	V	L	R	L	R
Bias10.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias9.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias8.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias7.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias6.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias5.0	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias4.0	Q	F	K	L	T	T	T	N	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias3.0	Q	F	K	L	T	T	T	N	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias2.0	Q	F	K	L	T	T	T	N	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	Q	V	L	R	L	R
Bias1.0	Q	F	K	L	T	T	T	N	E	S	T	E	D	Q	E	D	S	N	Q	Q	E	T	Q	V	L	T	L	R
NO BIAS	Q	F	Q	L	T	T	T	N	N	S	T	Q	D	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	T	L	R
Bias-1.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-2.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-3.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-4.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-5.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-6.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	E	D	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q
Bias-7.0	Q	F	Q	L	T	T	T	N	N	S	T	Q	T	Q	N	Q	S	N	Q	Q	Q	T	Q	V	L	Q	L	Q

Table B-4. Sequences predicted from calculations with varying bias energies. A secondary structure propensity scale factor of 2.0 was used for the charged positions.

	2	4	6	8	9	12	14	16	18	20	22	24	39	49	51	52	57	60	62	63	64	66	68	70	71	72	73	74
Wild type	Q	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	H	V	L	R	L	R
Bias10.0	T	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	T	V	L	R	L	R
Bias9.0	T	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	T	V	L	R	L	R
Bias8.0	T	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	T	V	L	R	L	R
Bias7.0	T	F	K	L	T	T	T	E	E	S	T	E	D	Q	E	D	S	N	Q	K	E	T	T	V	L	R	L	R
Bias6.0	T	F	K	L	T	T	T	T	E	S	T	E	D	Q	E	D	S	N	Q	K	T	T	T	V	L	R	L	R
Bias5.0	T	F	K	L	T	T	T	T	E	S	T	E	D	Q	E	D	S	N	Q	K	T	T	T	V	L	R	L	R
Bias4.0	T	F	K	L	T	T	T	T	E	S	T	E	D	Q	E	D	S	N	Q	K	T	T	T	V	L	R	L	R
Bias3.0	T	F	K	L	T	T	T	T	E	S	T	E	D	Q	E	D	S	N	Q	K	T	T	T	V	L	T	L	R
Bias2.0	T	F	K	L	T	T	T	T	E	S	T	E	D	Q	E	D	S	N	Q	K	T	T	T	V	L	T	L	R
Bias1.0	T	F	T	L	T	T	T	T	E	S	T	E	D	Q	E	D	S	N	Q	K	T	T	T	V	L	T	L	R
NO BIAS	T	F	T	L	T	T	T	T	N	S	T	Q	D	Q	E	D	S	N	Q	K	T	T	T	V	L	T	L	R
Bias-1.0	T	F	T	L	T	T	T	T	N	S	T	Q	N	Q	E	D	S	N	Q	Q	T	T	T	V	L	T	L	Q
Bias-2.0	T	F	T	L	T	T	T	T	N	S	T	Q	N	Q	E	D	S	N	Q	Q	T	T	T	V	L	T	L	Q
Bias-3.0	T	F	T	L	T	T	T	T	N	S	T	Q	N	Q	E	D	S	N	Q	Q	T	T	T	V	L	T	L	Q
Bias-4.0	T	F	T	L	T	T	T	T	N	S	T	Q	N	Q	E	D	S	N	Q	Q	T	T	T	V	L	T	L	Q
Bias-5.0	T	F	T	L	T	T	T	T	N	S	T	Q	N	Q	E	D	S	N	Q	Q	T	T	T	V	L	T	L	Q
Bias-6.0	T	F	T	L	T	T	T	T	N	S	T	Q	N	Q	E	D	S	N	Q	Q	T	T	T	V	L	T	L	Q
Bias-7.0	T	F	T	L	T	T	T	T	N	S	T	Q	N	Q	N	Q	S	N	Q	Q	T	T	T	V	L	T	L	Q

Figure B-1: The twelve charged residues on Ubiquitin surface considered for design are shown


E51
R74
D52
D24
R72
H68
E64
E39
K63
K6
E18
E16