Title
Investigating Protein Folding and Function by Manipulating Rare and Partially-Folded Conformations

Author
Horner, Geoffrey Ashworth

Publication Date
2010

Peer reviewed|Thesis/dissertation

Investigating Protein Folding and Function by Manipulating Rare and Partially-Folded Conformations

by

Geoffrey Ashworth Horner

A dissertation submitted in partial satisfaction of the requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Susan Marqusee, Chair

Professor Michael Marletta

Professor David Wemmer

Professor Jasper Rine

Spring 2010

Abstract

Investigating Protein Folding and Function by Manipulating

Rare and Partially-Folded Conformations

By

Geoffrey Ashworth Horner

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Susan Marqusee, Chair

This thesis includes work from three major projects. In the first chapter I describe work on the structural heterogeneity of the folding intermediate of RNase H. In this project we were able to populate the kinetic intermediate of RNase H at equilibrium with a mutation that strategically disrupted the native state. By populating this intermediate at equilibrium, we were able to characterize it by NMR and show that it is a highly dynamic conformation. The second chapter presents work using hydrophobic core repacking to manipulate protein function. We used a constrained directed evolution approach to generate novel function in the transcriptional activator MarA. We created libraries of core mutations and selected for core mutants that could stimulate transcription with a novel promoter sequence. Our results demonstrated that reorganization of the core alone can be sufficient to drive the evolution of novel function.

Finally, in the appendix, I describe my work in trying to isolate and characterize a class of mutations in ligand binding proteins which are vitamin remedial. Remedial mutations are those which disrupt protein function, but can be reversed with elevated levels of cofactor. Vitamin remediation is particularly interesting for its therapeutic benefits in the case of mutations linked to heritable disease. We hypothesized that vitamin remedial mutations might be simply derived from shifts in protein stability. To characterize the vitamin remedial effects of mutations in folate-binding proteins, we coupled *in vivo* evidence for folate-responsive growth to biophysically measured changes in stability and binding.

Table of Contents

Acknowledgements

There have been many people who have contributed to the work in this thesis, and I am thankful to all of them for their help and effort. All of the projects here have included some collaborative component. The work in the second chapter was done in concert with Katelyn Connell in the Marqusee Lab, and was helped along with work from Erik Miller and Nate Fernhoff. The second chapter contains work that was an ebullient collaboration with Dr. Katherine Tripp in the Marqusee Lab. A great deal of work, which is described briefly in the appendix, was a collaboration with Jasper Rine and Nick Marini who were always positive in the face the most negative of results.

Unofficial scientific contributions are sometimes as important as the contributions from co-authors, and I want to thank all of the Marqusee Lab members who have pondered and frowned with me over puzzling data for many years. It made working in the lab that much more enjoyable. In particular, Philip Elms and I joined the Marqusee Lab together and have shared the scientific and unscientific trials of graduate school. No one has done more to help me make this thesis what it is than Susan Marqusee, my advisor, who has been counselor and motivator, and who without question has made me a better scientist. My most important thanks go to my advisory committee outside the lab. My parents, whose support and encouragement allowed me the opportunity to come to Berkeley, have always been interested in my projects without always knowing quite what they were. Finally, I would like to thank my wife, Elli, who has shared with me the highs and lows of every datum and who has demonstrated that the smallest model system can have the most extraordinary phenotype.

Chapter 1. Introduction

1.1 Introduction

Proteins are natural organic heteropolymers which drive the chemistry of living things. The diversity of protein function is remarkable in itself, but it is all the more striking considering that all proteins are composed of the same 20 amino acid subunits. Within the order of amino acids in each sequence is the chemical information that precisely arranges functional groups to carry out biological activity. This breadth of function comes from the combinatorial scale that gives evolution a nearly limitless number of possible sequences to choose from. For most proteins, positioning the many functional groups into the native state, protein folding, is a fundamental, intrinsic property of the polypeptide.

How does the native chain direct the formation of protein structure? Currently, there is no predictive understanding of how a protein reaches its native state. The challenge lies partly in the tremendous diversity of conformations available to any polypeptide. Even in the native state, small local fluctuations and partial unfolding events create microscopic diversity. The unfolded state is even more complex with a much larger number of possible conformations available to each polypeptide. Therefore, even under folded conditions, a protein is best modeled as a Boltzmann distribution of ensembles of different conformations. Work in the folding field has focused on characterizing the small populations and microstates that make up the ensembles populated during protein folding. This thesis comprises several projects with diverse objectives, but each uses the manipulation and characterization of protein conformational ensembles to better understand protein function, structure, and folding.

1.2 Protein Folding

1.2.1 The protein folding problem

Unlike typical polymers, proteins behave more like small organic molecules: under native conditions they populate a very narrow range of their potential conformations. This was illustrated with the solution of the first protein crystal structure fifty years ago [1]. Populating this native ground state is a spontaneous, reversible reaction of the protein polypeptide. Christian Anfinsen demonstrated this in one of the first protein folding experiments. Anfinsen was investigating the behavior of oxidized ribonuclease A (RNase A) and observed that by moving the protein from denaturing to native conditions, it would spontaneously recover activity [2, 3]. This demonstrated that protein folding is reversible, inherent to the sequence of amino acids, and, importantly, that in folded conditions the structure of a native protein is its lowest energy state.

However, the question remains: how do proteins find this lowest energy conformation? Cyrus Levinthal considered this problem in one of protein folding's most famous thought experiments. He estimated the number of conformations available to a protein given the degrees of freedom for a polypeptide chain. Levinthal concluded that even for a small protein, the conformational space is so immense that if the native structure was found by a 'random walk', no protein would ever fold within observed folding timescales [4]. 'Levinthal's paradox' therefore concludes that proteins must reach their native state through a directed, non-random process. The observations of Anfinsen and Levinthal largely define the protein folding problem. If proteins fold to their lowest-energy native state through a bias or directed 'pathway', what defines that trajectory?

1.2.2 Intermediates as clues to pathways

One of the best ways to characterize a folding pathway is to identify partially folded states that may be populated along this trajectory. From work in the early 70s it was known that proteins do indeed transiently populate partially-folded conformations or intermediates [5]. Early conceptions of intermediates described them as points along the sequence of conformations leading to the native state. We now know that intermediates are a much more diverse class of structures. Proteins may fold through sequential intermediates or through parallel pathways [6, 7]. Intermediates may be on-pathway or off-pathway [8-10]. They may form before or after the rate limiting step [11, 12]. Regardless, the observation of transient intermediates demonstrates that the native state is not the only low-energy conformation of a protein.

Proteins that transiently populate an intermediate before their rate-limiting step are known as three-state folders. In some cases, however by slightly altering experimental conditions partially folded intermediates may be populated at equilibrium making them amenable to detailed characterization. The apo form of myoglobin was one of the first three-state proteins to be characterized this way. At low pH (~pH 4) apo-myoglobin populates an intermediate at equilibrium that is distinct from both the native and unfolded state. This equilibrium species resembles the transient intermediate formed in native conditions and therefore serves as an excellent model for its folding intermediate [13]. Compared to the native state it has an expanded hydrodynamic radius and loose tertiary interactions, but in comparison to the unfolded state it is more compact and contains more native secondary structure. The case of apo-myoglobin suggests that protein folding begins with a collapse of the polypeptide to a compact state that is missing many specific native contacts but excludes solvent.

What distinguishes these intermediates from low-energy alternative conformations? Some direct evidence comes from experiments using the fluorescent dye 8-anilino-1-naphthalene sulfonate (ANS). ANS binds and fluoresces in hydrophobic environments not exposed to aqueous solvent. In many proteins, monitoring folding by ANS fluorescence reveals an initial increase in signal during folding that diminishes

with formation of the native state [14]. Such experiments suggest that intermediates have a high degree of non-polar burial, but without the close packing that excludes ANS from binding the native state. Furthermore, experiments using far UV circular dichroism (CD), which is regarded as a direct measure of secondary structure, can often detect secondary structure forming within the dead time of stopped-flow initiated folding experiments [15]. As a result, many intermediates are modeled as a collapsed state driven by hydrophobic burial containing some secondary structure but lacking the close packing characteristic of native conformations. Such a structure is often described as a 'molten globule' [16].

Are intermediates essential for folding? This question is answered in part by the observation that some proteins fold without a discernable intermediate structure. These proteins, such as chymotrypsin inhibitor 2 (CI2), fold through what appears to be a single kinetic barrier [17, 18]. Two-state proteins suggest that the formation of stable detectable intermediates is not required for folding, but they do not, however, rule out the possibility of intermediate formation altogether. Instead intermediates could be too unstable or exist after the rate-limiting step making detection difficult. Experiments with two-state proteins also suggest that in small proteins where collapse leads to near native conformations, the folding process can be very fast (i.e. less than a millisecond).

When monitoring the folding process using ensemble-based methods, intermediates are only observed when they precede the rate limiting step, or transition state. There are many examples of partially folded intermediates that cannot be detected by these kinetic measurements. High-resolution equilibrium experiments have revealed that even two-state folding proteins can populate high-energy, partially-unfolded forms. For instance, native state hydrogen exchange has revealed partially unfolded structures populated after the rate-limiting step of folding in cytochrome c, often called hidden intermediates [12, 19-21]. Similar partially-unfolded structures have been observed in model systems like barnase, T4 lysozyme, and apocytochrome b_{562} [22-24]. Although they form after the kinetic barrier to folding, such structures may be important for the folding process by providing consecutive folding units as the protein progresses to the native state [25]. So-called hidden intermediates, occuring after the rate-limiting step, demonstrate that even within the native state the energy landscape is much rougher than a simple rate-limited kinetic analysis suggests.

1.2.3 The landscape model

As discussed above, intermediates are much more diverse and dynamic than a strictly sequential model can fully explain. Instead, the predominant view of protein folding takes a more global perspective of protein conformations and conceives of folding as a landscape of high and low energy conformations, providing multiple routes to the folded state. In this way, many pathways can co-exist, allowing conformationally distinct but energetically similar routes to the native state. Within a landscape model, intermediates collect in local energy minima along contours in the landscape. The

landscape can be visualized graphically as a rough funnel where the energy of any conformation on the z-axis is a function of its degrees of freedom in the x and y plane. By broadening the pathway both conceptually and graphically, the landscape model allows for much more diversity in the path any individual protein will take to the folded state.

The landscape model of folding is founded in a statistical mechanics perspective. Each state in the reaction coordinate is an ensemble of microstates occupying similar energetic wells. Each unfolded polypeptide is gyrating through potential conformations and encountering opposing constraints from sterics, hydrophobic collapse, dipole interactions, and hydrogen bonding pairs. An unfolded protein under native conditions is 'frustrated' by contradictory inputs, and thus occupies a very high energy state. A random heteropolymer would never do more than make transient interactions with itself and its neighbors. Proteins on the other hand are encoded by evolution with a native conformation that is lower energy because it resolves some opposing forces that make the unfolded state so high-energy. For this reason the native state is sometimes referred to as a structure of 'minimal frustration' [26].

1.3 Conformational Ensembles

In characterizing protein conformational landscapes, a great deal of effort has been spent investigating the very bottom of the funnel, the native well. While the native conformation is often depicted as a single structure, in reality it is not. The native state describes a tight ensemble of interconverting structures. Such dynamics and heterogeneity within a conformational state are essential for both function and folding. These fluctuations are not limited to those within the native well but can include large scale fluctuations to rare high-energy states. Max Perutz noted in the first structure of hemoglobin that there was no route for ligand to gain access to the active site [27]. In order for the protein to bind oxygen, structural flexibility was required. Similarly, allosteric alterations caused by ligand binding suggest that ligand can preferentially bind a rare high-energy conformation [28]. These and other studies probing the role of rare populations in protein function reinforce the importance of diversity within the native ensemble.

A surprising implication of recent experiments on protein flexibility is that the inherent populations of protein conformations may be intricately linked to protein function. The enzyme cyclophilin A (CypA) catalyzes the isomerization of prolyl peptide bonds between cis and trans conformations. Using specialized NMR relaxation dispersion experiments, Kern and colleagues were able to identify and monitor different conformations during binding and catalysis of CypA [29]. Interestingly, they could identify active and bound conformational states populated by the apo-protein in the absence of ligand [30]. Furthermore, the proportions of those populations correlated with the turnover values of CypA, implying that the high-energy conformations (i.e. the conformational landscape) of CypA are an intrinsic part of protein function. Improved

experimental tools, including NMR, have provided a more detailed characterization of low population conformations within protein ensembles.

Other experimental techniques have also highlighted the diverse nature of protein energy landscapes. Optical tweezer experiments and atomic force microscopy can probe single molecules with force [7], and developments of single molecule FRET techniques have expanded landscape analyses to larger proteins [31, 32]. Advanced techniques in crystallography (traditionally viewed as a technique to probe static structure), such as Laue crystallography, also shed light on conformational dynamics and structural heterogeneity [33].

The work presented here focuses on the role of conformational heterogeneity in protein structure, folding, and function. How heterogeneous and dynamic are high energy conformations such as folding intermediates? Can a functionally disrupted native state ensemble recover function through ligand binding? Can shifting native ensembles using mutations in the protein core generate new function?

1.4 The Hydrophobic Core – Contribution to Protein Structure and Stability

1.4.1 The hydrophobic effect

The hydrophobic effect, proposed by Walter Kauzmann in 1959, is broadly agreed to be the major driving force of protein folding. The hydrophobic effect in relation to protein folding states that folding is driven by the burial of non-polar surfaces within the interior of the protein [34]. Experimental observations consistently support this theory. Three-dimensional models from crystallographic and NMR studies generally reveal an arrangement of non-polar residues buried within the core sequestered away from the solvent-exposed surface. Non-polar solvents generally denature proteins, and protein stability shows a direct correlation with the free energies of transfer for non-polar groups between non-polar and aqueous solvent [35, 36]. Mutations in protein cores are frequently much more disruptive than alterations of polar residues [37].

At room temperature the hydrophobic effect is driven by entropy. This entropy is often attributed to the clathration of water molecules around non-polar surface in order to optimize hydrogen bonds, particularly at low temperatures. The entropic aversion to this ordering substantially contributes to the hydrophobic effect [38-40]. This entropic effect decreases with temperature, however, and at the point of strongest free energy of transfer (a quantitative measure of the hydrophobic effect) the entropic contribution is zero. At this temperature, the enthalpic component of hydrophobicity comes from the energetically favorable nature of water's interactions with itself. The result is that hydrophobicity has very non-ideal behavior, including a temperature

dependence and a change in the heat capacity of mixing that an ideal mixing system would never display [41].

A remarkable feature of the hydrophobic effect is that while it drives the association of non-polar groups it lacks fidelity or steric specificity. That is to say, unlike hydrogen bonds or charge attractions, hydrophobic clustering is very non-specific. The result is that mutations of non-polar residues which preserve hydrophobicity can preserve much of the energetic driving force for folding. These variations are, however, at the expense of close packing. This feature, the ability of protein cores to stably repack, may play an important role in conformational heterogeneity.

1.4.2 Core packing

The two most defining characteristics of protein cores are non-polar burial and very dense packing. Early work by Richards and coworkers revealed that protein cores arrange atoms so densely they are almost crystalline [42]. This packing density contributes to the stability of folded proteins. Conservative mutations which preserve side-chain volume and hydrophobic composition are still destabilizing to proteins, illustrating that packing itself contributes to the energy of folding [43]. On the other hand, close packing is not so significant that all poorly packed proteins cannot fold [44]. The general pattern of polar and non-polar residues, coupled with hydrophobic burial, is much more important than packing density for defining a protein's topology.

How tolerant are proteins to disruptions in the core? While mutations in the core are more disruptive than those on the surface, they are tolerated well enough that single mutations seldom denature proteins under physiological conditions. Systematic mutations of core residues show that mutations are most disruptive when introducing polar or charged groups to the protein interior [45, 46]. Structural studies reveal that core mutants often rearrange to optimize packing very effectively, though the new orientation is almost never as well packed as the native core [47]. While a single mutation may disrupt local packing, it rarely affects the overall burial of non-polar groups. Perhaps more surprising is the observation that proteins are tolerant of multiple mutations in the core although these often are accompanied by subtle movements of the backbone [48].

Given that proteins are well packed, Ponder and Richards investigated how multiple permutations of sequences can pack a given core geometry. They used a computational algorithm that, given side chain geometry and rotamer preferences, evaluated the combination of residues compatible with a specific structural core [49]. Their analysis revealed that many sequences could function in a given core, indicating that the composition of protein cores may be surprisingly plastic given their importance to stabilizing the native structure.

Perhaps the best experimental example of core rearrangement comes from early work by Lim and Sauer who used a genetic approach to investigate the compositional plasticity of the core of the N-terminal domain of λ-repressor [50]. They created a combinatorial library of mutants of the repressor's hydrophobic core varying three to five residues at a time and then selected for active functional proteins using both stringent and tolerant thresholds [50]. By isolating and sequencing selected variants, they were able to assemble the sets of active, partially active and inactivating core sequences. Consistent with the work on T4 lysozyme, Lim and Sauer found that the core of λ-repressor is surprisingly plastic. As many as 70% of the possible core arrangements were tolerated at the least stringent threshold [51]. While many λ-repressor mutants were able to demonstrate some binding activity, very few of the identified mutants had dissociation constants as low as wild-type and none had tighter DNA binding than wild-type.

1.5 Probing Folding and Function

1.5.1 Mutagenesis

My thesis covers a range of topics using diverse experimental techniques, but they all share the common approach of altering a protein's sequence. Site-directed mutagenesis can probe the role of specific residues with minimal disruption to structure and function. This approach has confirmed and revealed some of the fundamental principles of folding and stability [52]. Fersht and coworkers developed a framework for using a mutational analysis to infer structural information about the high energy transition state in protein folding, which they applied to proteins such as barnase and chymotrypsin inhibitor [53, 54]. Matthews and Alber used extensive mutation of the phage protein T4 lysozyme to probe details of folding and structure. They found that the most disruptive mutations were in the rigid core regions of the protein [55, 56].

Mutation is the vehicle of change in evolution and has broad impacts on human health. In Appendix A, I describe a series of experiments that attempt to link simple destabilizing mutations in enzymes to clinically manifested health disorders, and the potential for remediation through controlling cofactor concentration. The growing ability to sequence quickly and cheaply has illuminated the role that genetic variation has on human health [57]. Single nucleotide polymorphisms (SNPs) have become a standard information source for disease risk factors [58]. As personal genomic information becomes available, the demand for more personal, polymorphism-specific therapies for genetic diseases will grow. An example of this is the enzyme methylenetetrahydrofolate reductase, which has a polymorphism in humans, C677T, that is associated with elevated risk for higher homocysteine levels in the body that can lead to neural tube defects and other health disorders [59, 60].

1.5.2 Engineering by directed evolution

Mutation has recently been applied as a laboratory tool to generate novel protein function. Directed evolution is an experimental method which seeks to mimic and accelerate natural selection to find novel protein function. Directed evolution begins by generating libraries of mutants with large sequence diversity. These libraries are then filtered by experimental assessment, eliminating non-functional molecules and isolating those with desired function, to identify new proteins with novel characteristics. Directed evolution exploits the efficiency of evolutionary selection to generate or engineer new proteins for both functional benefit and to better understand the evolutionary constraints on proteins.

Library generation is a relatively simple process. The vast number of unique sequences available means that most methods of introducing mutations will produce a diverse library within a short time in the lab. A sequence of only 100 amino acids can encompass a mind boggling 20^{100} possible proteins. Techniques to introduce random mutations such as using mutagenic strains of bacteria, UV irradiation of DNA or error-prone PCR are all used successfully. It is possible with current DNA synthesis methods to generate libraries with diversity of up to 10^{15} distinct members, but smaller libraries are used more often.

The largest challenge in directed evolution is identifying the tiny population of interesting mutants in a very diverse library. Two general approaches are used depending on the rarity of the behavior being assessed. In experiments where a large number of the library members will show a positive effect a screen can be used, wherein each individual member is analyzed separately. For enzymes this often relies on the detection of a downstream product with spectroscopy or the use of HPLC [61]. Some recent screens have increased throughput by using fluorescence activated cell sorting [62]. In the case where the targeted mutants are much less abundant, it is better to identify them by a selection. In a selection, desirable behavior in a library member will rescue an expressing cell from environmental or chemical growth inhibitors. In this way a culture dish can easily test 10^8 or more library clones [61]. By adjusting the selective pressure the stringency of the selection can be modulated to capture a range of activities within the library.

Directed evolution is essentially the generation of desired protein function. Generating enzymes that catalyze entirely novel reactions is still beyond what most experiments can achieve. Recent successful applications have focused on taking existing protein activities and evolving novel specificity for new substrates [63]. This approach, closer to protein redesign than *de novo* protein design, has yielded a great deal of information about how protein function can be manipulated. A surprising lesson from this work is how effective small changes can be. In one recent experiment, a point mutation was sufficient to alter substrate specificity of a type II restriction endonuclease [64]. Cordes et al. have shown that a small number of strategic mutations can be sufficient to reengineer even the secondary structure of the native fold [65].

When novel function requires more than one mutation it is often approached stepwise through evolutionary 'intermediates' [63]. In some cases, directed evolution is first used to select for a protein that shows broad specificity and then, in later rounds, this selected protein is evolved for new specificity. Recent examples of such an approach can be seen in the directed evolution of novel endopeptidases by Varadarajan et al. [66] and a D-xylose-specific xylose reductase by Nair et al. [67]. Selecting for the desired activity can sometimes be made more efficient by compensating for the destabilizing effects of accumulated mutations. Since stable proteins are more tolerant of mutation, introducing mutations that stabilize the protein before beginning rounds of directed evolution can produce more mutants with novel function [68]. Another strategy to increase the chances of success is to introduce negative selection for the native enzyme function. This approach is especially useful in the cases where the experimentalist wants to push the evolutionary path to a general binder and then select for novel function [66]. These developments are accelerating the pace of protein engineering and bringing us closer to a point where novel proteins can be evolved routinely.

1.6 Model Systems Used in this Thesis

1.6.1 RNase H

Ribonuclease H (RNase H) is a ribonuclease responsible for the degradation of RNA from RNA:DNA hybrids. RNase H is a small and monomeric protein that hydrolyzes RNA in a Mg^{2+}-dependent manner. RNase H has homologues in many prokaryotic organisms, and RNases H from a variety of taxonomic classes have been studied, including thermophilic bacteria and even HIV. RNase H from *E. coli* has been one of the primary model systems for folding studies in the Marqusee Lab for more than 15 years. It provides an excellent model system for investigating the protein folding landscape, it is small, soluble, tractable for experimentation, and its landscape is complex enough to be an informative model for folding.

RNase H folds in a three state manner through a kinetic folding intermediate which is not well populated at equilibrium [69-71]. This kinetic folding intermediate forms in the dead time of a stopped-flow mixer and can be detected as rollover or curvature in a chevron plot (the natural log plot of observed folding rates vs. denaturant concentration). The folding intermediate, as determined by pulsed-labeling hydrogen exchange, consists of a core set of residues encompassing the central part of the sequence made up of helices A and D and strand 4. Expressed as a protein fragment, this core folds autonomously [72]. From kinetic analysis, the intermediate is determined to have free energy of folding (ΔG_{UI}) of -3.55 kcal/mol (the stability of the full protein, ΔG_{UN}, is -9.7 kcal/mol) [69]. Single molecule studies of RNase H have confirmed that this intermediate is on pathway and obligatory [7]. These results

indicate a protein landscape for RNase H in which an autonomously folded core forms early, in less than five milliseconds, while a periphery folds around the core, bringing together residues with lower contact order.

1.6.2 MarA

MarA is a transcriptional activator that triggers gene expression in response to superoxide stress or antibiotic exposure. MarA is a member of the AraC family of transcription regulators which share homology within a 100 amino acid DNA binding domain. The structure of this 17 kD protein is composed entirely of α-helices and has been solved in a bound state with its cognate DNA [73]. The protein forms an elongated structure which allows two helices to directly contact positions in the major groove. MarA binds a highly degenerate sequence upstream of its target genes, but a consensus sequence from various promoters has been identified [74]. Alanine scanning across the structure has shown that alteration of some core sites completely disrupt function across all possible promoters while some mutations are differentially disruptive [75]. MarA is unusual for a member of the AraC family as it functions as a monomer in the cell. It shares function with two other regulators SoxS and Rob with which it has more than 60% sequence similarity [76]. MarA activates downstream antibiotic and superoxide responses by binding to an upstream promoter and recruiting RNA polymerase.

Activation of MarA-regulated genes begins with binding to a 20 base pair sequence known as the marbox. MarA binds to marbox sequences in vitro with K_ds that range from micromolar to nanomolar. Interestingly, in crystal structures of MarA solved with a cognate DNA, there are relatively few hydrogen bond interactions between sequence-specific residues [73, 77]. Instead, MarA appears to determine specificity more by van der Waals contacts and steric occlusion. MarA binds marboxes which are positioned upstream of the -35 promoter region and recruits RNA polymerase (RNP) to begin transcription through interactions with the C-terminal domain of RNP.

MarA is an appealing model system for investigating protein specificity and binding. The details of its interactions both with marbox sequences and RNP have been thoroughly described [76, 78, 79]. It is an unusual transcription factor in that it functions as a monomer and binds in a directional fashion, simplifying interpretations of mutations. Transcription factors are an attractive target for function assays as downstream transcription provides an easy means to measure functional proteins. In several experiments probing the surface interactions of MarA and its transcript, Shultzaberger et al. coupled MarA binding to tetracycline resistance in order to monitor the requirements and evolutionary limits of sequence specificity [80]. In all MarA is a proven target for both DNA binding and selection experiments.

1.7 Summary of Thesis Work

This thesis describes results from two major studies. The first is an investigation of the structural heterogeneity of the folding intermediate of RNase H (Chapter 2). This work was published in the Journal of Molecular Biology 2009 [81] with myself as a co-first author. The second is a recent investigation of the role of hydrophobic packing in determining protein function. Specifically, we evaluated the core packing of the protein MarA and asked if we could alter the hydrophobic core of the protein and generate MarA variants with novel DNA-binding specificity (Chapter 3). While the initial studies revealed the surprising result that several combinations of core variants alter MarA binding specificity, the structural details of these proteins are still being determined. Finally, in the appendix, I describe my work in trying to isolate and characterize mutated versions of folate-binding proteins in an attempt to understand the potential for cofactor remediation in disease-associated variants. While several years were spent on these studies, the protein proved refractory to biophysical analysis and no publishable conclusions were obtained.

1.7.1 Equilibrium populated intermediate of RNase H

In this work, a single mutation in the periphery of RNase H was shown to selectively destabilize the native state and populate the folding intermediate under native conditions. As described above, the protein ribonuclease H (RNase H) has been studied extensively as a model system for protein folding and transiently populates an intermediate, formed by the association of several core helices. The final native state is reached after collapse of a periphery region around this core. At the center of this periphery is the residue isoleucine 25. In studying the folding pathway we mutated isoleucine 25 to alanine to destabilize the periphery. This mutation dramatically destabilized RNase H. Initial data suggested that I25A folded stably to the intermediate state without reaching the native state. Two dimensional NMR measurements, however, contradicted this conclusion by revealing an almost completely native spectrum with well dispersed, sharp peaks. By creating a double mutant that combined a stabilizing periphery mutation (D10A) with I25A, we were able to isolate the energetic effect of I25A on native state stability. By monitoring NMR HSQCs in increasing concentrations of denaturant we were able to identify the signals of the intermediate. Surprisingly few signals from the intermediate were detected, which we interpreted as evidence for an intermediate ensemble that was highly dynamic.

1.7.2 Refolded core mutants selected for novel function

Engineering of protein function frequently focuses on amino acids which directly contact the ligand or substrate. Previous work in protein folding has shown that while

the core of a protein is densely packed, it can be rearranged and accommodate new compositions while preserving function. We wondered if the hydrophobic core could influence function beyond serving simply as a "scaffold" for an enzyme active site. We have taken advantage of a system developed to investigate transcription factor DNA-binding specificity and used it to probe the role of core repacking in evolving protein function. As mentioned above, MarA is a small transcriptional activator which binds a range of promoters that can trigger expression of genes involved in antibiotic resistance. We have made three libraries of MarA mutants which contain in total 24,000 unique variants of the core of the protein, allowing up to three residues to vary at a time. We screened our library against promoter sequences with strong affinity to wild-type MarA and also to promoter sequences that wild-type could not bind. Our results suggest that even minimal reorganization in the core can generate novel protein binding specificity.

1.7 Citations

1. Kendrew, J.C., *Structure and function in myoglobin and other proteins.* Fed Proc, 1959. **18**(2, Part 1): p. 740-51.
2. Anfinsen, C.B., et al., *The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain.* Proc Natl Acad Sci, 1961. **47**: p. 1309-14.
3. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(96): p. 223-30.
4. Levinthal, C., *Are there pathways for protein folding.* J Chim Phys, 1968. **65**(1): p. 44-45.
5. Tsong, T.Y., R.L. Baldwin, and E.L. Elson, *The sequential unfolding of ribonuclease A: detection of a fast initial phase in the kinetics of unfolding.* Proc Natl Acad Sci, 1971. **68**(11): p. 2712-5.
6. Radford, S.E., C.M. Dobson, and P.A. Evans, *The folding of hen lysozyme involves partially structured intermediates and multiple pathways.* 1992.
7. Cecconi, C., et al., *Direct observation of the three-state folding of a single protein molecule.* Science, 2005. **309**(5743): p. 2057.
8. Horwich, A., *Protein aggregation in disease: a role for folding intermediates forming specific multimeric interactions.* Journal of Clinical Investigation, 2002. **110**(9): p. 1221-1232.
9. Dobson, C.M., *Protein folding and misfolding.* Nature, 2003. **426**(6968): p. 884-890.
10. Matouschek, A., et al., *Transient folding intermediates characterized by protein engineering.* Nature, 1990. **346**(6283): p. 440-445.
11. Roder, H., G.A. Elöve, and S.W. Englander, *Structural characterization of folding intermediates in cytochrome c by H-exchange labelling and proton NMR.* Nature, 1988. **335**(6192): p. 700-704.
12. Bai, Y., et al., *Protein folding intermediates: native-state hydrogen exchange.* Science, 1995: p. 192-192.
13. Griko, Y.V., et al., *Thermodynamic study of the apomyoglobin structure.* J Mol Biol, 1988. **202**(1): p. 127-38.
14. Matthews, C.R., *Pathways of protein folding.* Annual Review of Biochemistry, 1993. **62**(1): p. 653-683.
15. Kuwajima, K., et al., *Rapid formation of secondary structure framework in protein folding studied by stopped-flow circular dichroism.* FEBS letters, 1987. **221**(1): p. 115-118.
16. Privalov, P.L., *Intermediate states in protein folding.* J Mol Biol, 1996. **258**(5): p. 707-725.
17. Itzhaki, L.S., et al., *Search for nucleation sites in smaller fragments of chymotrypsin inhibitor 2.* J Mol Biol, 1995. **254**(2): p. 289-304.
18. Jackson, S.E. and A.R. Fersht, *Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition.* Biochemistry, 1991. **30**(43): p. 10428.
19. Hoang, L., et al., *Cytochrome c folding pathway: Kinetic native-state hydrogen exchange.* Proc Natl Acad Sci, 2002. **99**(19): p. 12173.

20. Milne, J.S., et al., *Experimental study of the protein folding landscape: unfolding reactions in cytochrome c.* J Mol Biol, 1999. **290**(3): p. 811-822.
21. Xu, Y., L. Mayne, and S.W. Englander, *Evidence for an unfolding and refolding pathway in cytochrome c.* Nature Structural & Molecular Biology, 1998. **5**(9): p. 774-778.
22. Feng, H., Z. Zhou, and Y. Bai, *A protein folding pathway with multiple folding intermediates at atomic resolution.* Proc Natl Acad Sci U S A, 2005. **102**(14): p. 5026-31.
23. Kato, H., et al., *The folding pathway of T4 lysozyme: An on-pathway hidden folding intermediate.* J Mol Biol, 2007. **365**(3): p. 881-891.
24. Vu, N.D., H. Feng, and Y. Bai, *The folding pathway of barnase: the rate-limiting transition state and a hidden intermediate under native conditions.* Biochemistry, 2004. **43**(12): p. 3346-3356.
25. Bai, Y., *Hidden intermediates and levinthal paradox in the folding of small proteins.* Biochemical and Biophysical Research Communications, 2003. **305**(4): p. 785-788.
26. Bryngelson, J.D. and P.G. Wolynes, *Spin glasses and the statistical mechanics of protein folding.* Proc Natl Acad Sci, 1987. **84**(21): p. 7524.
27. Perutz, M.F. and F.S. Mathews, *An x-ray study of azide methaemoglobin.* J Mol Biol, 1966. **21**(1): p. 199.
28. Kern, D. and E.R.P. Zuiderweg, *The role of dynamics in allosteric regulation.* Current Opinion in Structural Biology, 2003. **13**(6): p. 748-757.
29. Eisenmesser, E.Z., et al., *Enzyme dynamics during catalysis.* Science, 2002. **295**(5559): p. 1520.
30. Eisenmesser, E.Z., et al., *Intrinsic dynamics of an enzyme underlies catalysis.* NATURE-LONDON-, 2005. **438**(7064): p. 117.
31. Diez, M., et al., *Proton-powered subunit rotation in single membrane-bound F 0 F 1-ATP synthase.* Nature Structural & Molecular Biology, 2004. **11**(2): p. 135-141.
32. Schuler, B. and W.A. Eaton, *Protein folding studied by single-molecule FRET.* Current Opinion in Structural Biology, 2008.
33. Bourgeois, D. and A. Royant, *Advances in kinetic protein crystallography.* Current Opinion in Structural Biology, 2005. **15**(5): p. 538-547.
34. Kauzmann, W., *Some factors in the interaction of protein denaturation.* Adv. Protein Chem, 1959. **14**: p. 1-63.
35. Pace, C.N. and J. Hermans, *The Stability of Globular Protein.* Critical Reviews in Biochemistry and Molecular Biology, 1975. **3**(1): p. 1-43.
36. Privalov, P.L., *Stability of proteins: small globular proteins.* Adv. Protein Chem, 1979. **33**: p. 167.
37. Matthews, B.W., *Structural and genetic analysis of protein stability.* Annual Review of Biochemistry, 1993. **62**(1): p. 139-160.
38. Dill, K.A., *Dominant forces in protein folding.* Biochemistry, 1990. **29**(31): p. 7133-7155.
39. Stillinger, F.H., *Water revisited.* Science, 1980. **209**(4455): p. 451.
40. Southall, N.T., K.A. Dill, and A.D.J. Haymet, *A view of the hydrophobic effect.* J. Phys. Chem. B, 2002. **106**(3): p. 521-533.

41. Privalov, P.L. and S.J. Gill, *Stability of protein structure and hydrophobic interaction.* Adv. Protein Chem, 1988. **39**(1): p. 191.
42. Richards, F.M., *Areas, volumes, packing, and protein structure.* Annual Review of Biophysics and Bioengineering, 1977. **6**(1): p. 151-176.
43. Hurley, J.H., W.A. Baase, and B.W. Matthews, *Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme.* J Mol Biol, 1992. **224**(4): p. 1143.
44. Behe, M.J., E.E. Lattman, and G.D. Rose, *The protein-folding problem: the native fold determines packing, but does packing determine the native fold?* Proc Natl Acad Sci, 1991. **88**(10): p. 4195.
45. Eriksson, A.E., W.A. Baase, and B.W. Matthews, *Similar hydrophobic replacements of Leu99 and Phe153 within the core of T4 lysozyme have different structural and thermodynamic consequences.* J Mol Biol, 1993. **229**(3): p. 747.
46. Eriksson, A.E., et al., *Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect.* Science, 1992. **255**(5041): p. 178-183.
47. Lim, W.A., et al., *The crystal structure of a mutant protein with altered but improved hydrophobic core packing.* Proc. Natl Acad. Sci. USA, 1994. **91**: p. 423-427.
48. Baldwin, E.P., et al., *The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme.* Science, 1993. **262**(5140): p. 1715.
49. Ponder, J.W. and F.M. Richards, *Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.* J Mol Biol, 1987. **193**(4): p. 775.
50. Lim, W.A. and R.T. Sauer, *Alternative packing arrangements in the hydrophobic core of lambda repressor.* Nature, 1989. **339**(6219): p. 31-6.
51. Lim, W.A. and R.T. Sauer, *The role of internal packing interactions in determining the structure and stability of a protein.* J Mol Biol, 1991. **219**(2): p. 359.
52. Alber, T., *Mutational effects on protein stability.* Annual Review of Biochemistry, 1989. **58**(1): p. 765-792.
53. Matouschek, A., et al., *Mapping the transition state and pathway of protein folding by protein engineering.* Nature, 1989. **340**(6229): p. 122-126.
54. Neira, J.L., et al., *Towards the complete structural characterization of a protein folding pathway: the structures of the denatured, transition and native states for the association/folding of two complementary fragments of cleaved chymotrypsin inhibitor 2. Direct evidence for a nucleation-condensation mechanism.* Folding and Design, 1996. **1**(3): p. 189-208.
55. Alber, T. and B.W. Matthews, *Structure and thermal stability of phage T 4 lysozyme.* Methods in Enzymology, 1987. **154**: p. 511-533.
56. Alber, T., et al., *Temperature-sensitive mutations of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein.* Biochemistry, 1987. **26**(13): p. 3754.
57. Wang, D.G., et al., *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.* Science, 1998. **280**(5366): p. 1077.
58. Cargill, M., et al., *Characterization of single-nucleotide polymorphisms in coding regions of human genes.* Nature Genetics, 1999. **22**(3): p. 231-238.

59. Kirke, P.N., et al., *Impact of the MTHFR C677T polymorphism on risk of neural tube defects: case-control study.* British Medical Journal, 2004. **328**(7455): p. 1535.
60. Mtiraoui, N., et al., *MTHFR C677T and A1298C gene polymorphisms and hyperhomocysteinemia as risk factors of diabetic nephropathy in type 2 diabetes patients.* Diabetes Research and Clinical Practice, 2007. **75**(1): p. 99-106.
61. Jäckel, C., P. Kast, and D. Hilvert, *Protein design by directed evolution.* 2008.
62. Aharoni, A., A.D. Griffiths, and D.S. Tawfik, *High-throughput screens and selections of enzyme-encoding genes.* Current opinion in chemical biology, 2005. **9**(2): p. 210-216.
63. Tracewell, C.A. and F.H. Arnold, *Directed enzyme evolution: climbing fitness peaks one amino acid at a time.* Current Opinion in Chemical Biology, 2009. **13**(1): p. 3-9.
64. Saravanan, M., K. Vasu, and V. Nagaraja, *Evolution of sequence specificity in a restriction endonuclease by a point mutation.* Proc Natl Acad Sci, 2008. **105**(30): p. 10344.
65. Cordes, M.H.J., et al., *An evolutionary bridge to a new protein fold.* Nature Structural & Molecular Biology, 2000. **7**(12): p. 1129-1132.
66. Varadarajan, N., et al., *Highly active and selective endopeptidases with programmed substrate specificities.* Nature chemical biology, 2008. **4**(5): p. 290.
67. Nair, N.U. and H. Zhao, *Evolution in reverse: Engineering a D-xylose-specific xylose reductase.* ChemBioChem, 2008. **9**(8): p. 1213-1215.
68. Bloom, J.D., et al. *Protein stability promotes evolvability*. 2006.
69. Raschke, T.M. and S. Marqusee, *The kinetic folding intermediate of ribonuclease H resembles the acid molten globule and partially unfolded molecules detected under native conditions.* Nature Structural & Molecular Biology, 1997. **4**(4): p. 298-304.
70. Chamberlain, A.K., T.M. Handel, and S. Marqusee, *Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH.* Nature Structural & Molecular Biology, 1996. **3**(9): p. 782-787.
71. Raschke, T.M., J. Kho, and S. Marqusee, *Confirmation of the hierarchical folding of RNase H: a protein engineering study.* nature structural biology, 1999. **6**: p. 825-830.
72. Chamberlain, A.K., et al., *Folding of an isolated ribonuclease H core fragment.* PRS, 1999. **8**(11): p. 2251-2257.
73. Rhee, S., et al., *A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator.* Proc Natl Acad Sci U S A, 1998. **95**(18): p. 10413-8.
74. Martin, R.G., et al., *Autoactivation of the marRAB multiple antibiotic resistance operon by the MarA transcriptional activator in Escherichia coli.* J Bacteriol, 1996. **178**(8): p. 2216-23.
75. Gillette, W.K., R.G. Martin, and J.L. Rosner, *Probing the Escherichia coli transcriptional activator MarA using alanine-scanning mutagenesis: residues important for DNA binding and activation.* J Mol Biol, 2000. **299**(5): p. 1245-55.
76. Jair, K.W., et al., *Purification and regulatory properties of MarA protein, a transcriptional activator of Escherichia coli multiple antibiotic and superoxide resistance promoters.* J Bacteriol, 1995. **177**(24): p. 7100-4.
77. Dangi, B., et al., *Structure and dynamics of MarA-DNA complexes: an NMR investigation.* J Mol Biol, 2001. **314**(1): p. 113-27.

78. Martin, R.G., W.K. Gillette, and J.L. Rosner, *Promoter discrimination by the related transcriptional activators MarA and SoxS: differential regulation by differential binding.* Mol Microbiol, 2000. **35**(3): p. 623-34.
79. Martin, R.G., et al., *Complex formation between activator and RNA polymerase as the basis for transcriptional activation by MarA and SoxS in Escherichia coli.* Mol Microbiol, 2002. **43**(2): p. 355-70.
80. Shultzaberger, R.K., *Functional Variability in Transcriptional Initiation Complexes.* Unpublished Doctoral Thesis, University of California Berkeley, 2009.
81. Connell, K.B., G.A. Horner, and S. Marqusee, *A Single Mutation at Residue 25 Populates the Folding Intermediate of E. coli RNase H and Reveals a Highly Dynamic Partially Folded Ensemble.* J Mol Biol, 2009. **391**(2): p. 461-470.

Chapter 2. A Single Mutation at Residue 25 Populates the Folding Intermediate of *E. coli* RNase H and Reveals a Highly Dynamic Partially Folded Ensemble

2.1 Abstract

This work was published in the Journal of Molecular Biology, 2009, volume 391 issue 2. Understanding the nature of partially folded intermediates transiently populated during protein folding is important for understanding both protein folding and misfolding. These ephemeral species, however, often elude direct experimental characterization. The well-characterized protein ribonuclease H from *E. coli* populates an on-pathway intermediate identified in both bulk studies and single molecule mechanical unfolding experiments. Here, we set out to trap the transient intermediate of RNase H at equilibrium by selectively destabilizing the region of the protein known to be unfolded in this species. Surprisingly, a single change at Ile 25 (I25A) resulted in the equilibrium population of the intermediate under near-native conditions. The intermediate was undetectable in a series of HSQC's, revealing the dynamic nature of this partially unfolded form on the timescale of NMR detection. This result is in contrast to studies in which the structures of trapped intermediates are solved by NMR, indicating that they are well-packed and native-like. The dynamic nature of the RNase H intermediate may be important for its role as an on-pathway, productive species that promotes efficient folding.

2.2 Introduction

Partially folded intermediates are known to play an important role in the mechanism of protein folding. For some proteins, intermediates appear to play a productive role, aiding in the formation of the native fold, while for others they constitute an off-pathway species. In addition to their importance in folding mechanisms, partially folded forms may also play crucial biological roles. There are many examples of proteins with unstructured regions whose disordered-ordered folding transitions are important for binding and other regulatory events [1, 2]. For example, the active form of the steroidogenic acute regulatory protein appears to be a molten globule [3], a state usually associated with early folding intermediates. Such intermediates have also been implicated in the formation of aggregation-prone species [4]. In spite of their clear importance in both folding and function, these partially folded intermediates are usually not amenable to detailed structural and biophysical studies due to their low populations at equilibrium and transient nature.

Here, we set out to trap the intermediate of ribonuclease H from *E. coli* (RNase H), a small (17.5 kDa), monomeric enzyme known to fold through such a partially folded kinetic intermediate [5-7] (throughout this manuscript, wild type RNase H refers to the *E. coli* variant with all three cysteines replaced by alanine). Recent single molecule experiments conclusively demonstrated that this species is on-pathway and obligatory [8]. Characterization of the intermediate by phi-value analysis and quenched flow hydrogen exchange suggest that the intermediate is composed of native-like topology in

the core and a relatively unstructured periphery (Figure 2-1) [5, 6]. This partially folded ensemble appears to play a robust and important role in the folding trajectory of RNase H, dominating the folding trajectory even in variants that do not transiently populate the state [9]. Equilibrium native state hydrogen exchange experiments detect a rare high energy partially unfolded form whose structure and energetics mirror this kinetic intermediate, suggesting that the most stable region of the protein is the first to fold [5, 10].

We have used rational design to selectively destabilize the native state of the protein, enriching the population of the high-energy species. The response of RNase H to mutations highlights the differential distribution of stability in the core and periphery that makes this approach possible. Isoleucine 53, located in the core, was mutated to alanine (I53A), destabilizing the protein by ~2 kcal/mol ($\Delta\Delta G_{UN}$ ~ 2 kcal/mol) [11]. Native state hydrogen exchange demonstrated that this destabilizing effect was localized to the core, while residues in the periphery were unaffected, resulting in destabilization of the intermediate. A similar study was carried out in the periphery with the stabilizing mutation D10A [12], where, although all secondary elements showed an increase in the free energy of unfolding as measured by native state hydrogen exchange, the difference between unfolding the periphery and the core remained unchanged, suggesting that the two regions are energetically independent. This selective increase in stability causes a decrease in the population of the partially folded form. For both of these variants, the effects in the equilibrium energetics were mirrored in the transient intermediate observed during the folding process. In fact, a severely destabilizing mutation in the core of the protein (I53D) destabilizes the intermediate such that the protein folds in an apparent two-state manner without the obvious accumulation of the intermediate [7].

In principle, then, we should be able to use mutations that destabilize the periphery of the protein without perturbing the stability of the intermediate to selectively destabilize the native state, enhancing the population of the high-energy intermediate such that we can study it more easily. This strategy has been demonstrated by Bai [13-18] who, based on native state hydrogen exchange data, created mutations to selectively destabilize the regions predicted to be unfolded in the intermediate, thereby selectively destabilizing the native state. Similarly, Radford and coworkers destabilized regions of the immunity protein Im7 based on results from phi-value analysis (monitoring the folding kinetics for engineered site-specific mutations) [19].

In an alternative approach, Bai and coworkers also recently designed a fragment of *T. thermophilus* RNase H modeled, in part, on our native state hydrogen exchange data. By removing two strands believed to be disordered in the intermediate, they obtained a well folded species amenable to high resolution NMR spectroscopy. Their results reveal a fragment that folds into a well-packed native-like species, suggesting that the structure of the partially folded intermediate may be a subset of the native structure. These results are at odds with studies that suggest the intermediate state of RNase H is a heterogeneous ensemble with the hallmarks of a molten globule [20, 21]. While the fragment approach has the advantage of creating a fragment amenable for

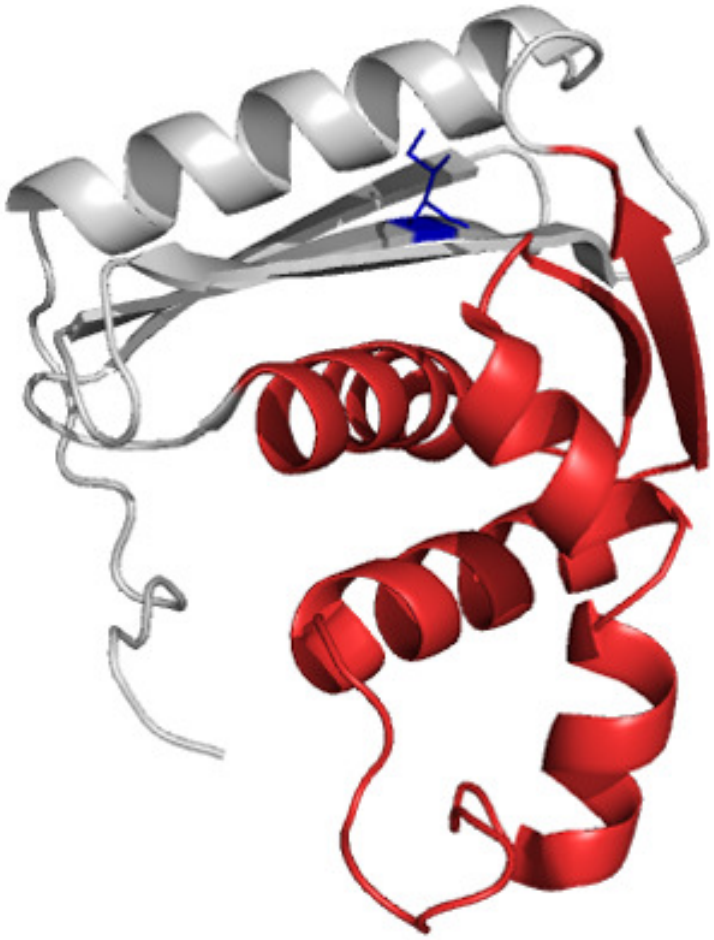

Figure 2-1 Ribbon representation of RNase H (pdb 1F21) with the core of the protein colored red and the periphery colored grey. Residue I25 is shown as sticks and colored blue.

high-resolution studies, there are drawbacks. The subjective assumptions about the structure, based on the limited hydrogen exchange data, may omit important structural components of the intermediate. All interpretations are based on the underlying assumption that the fragment is indeed a representative mimic of the high-energy ensemble. Questions therefore remain as to the exact nature of the contacts formed in the high-energy intermediate of the protein, the extent to which tertiary structure has been fixed and the breadth of this species, and which regions contribute to the 'molten' nature of the intermediate.

Here, we set out to trap the intermediate of RNase H by rational design of destabilizing mutants believed to selectively to the native state stability. We make no assumptions as to the structure of the species, but instead use mutagenesis to selectively destabilize the native state to enrich the population of the high-energy species. This approach is completely analogous to Bai's protein engineering approach to uncover hidden intermediates, only in our case, the intermediate is not hidden, but rather a detectable transiently populated species.

To our surprise, the first target for mutation, isoleucine 25, located in Strand II of the periphery and completely buried in a hydrophobic pocket (Figure 2-1), converted to alanine was almost sufficient to accomplish our task. This variant, which is part of a series of mutations used to investigate the robustness of the folding trajectory in two-state and three-state versions of the protein [9], results in a protein that detectably populates the kinetic intermediate at equilibrium. Although we could not obtain high-resolution structural information, with this mutation we were able to gain valuable insight into the dynamic properties of the intermediate.

2.3 Results

2.3.1 I25A populates the intermediate.

Equilibrium urea denaturation of I25A was monitored by circular dichroism (CD) at 222 nm and fit with the standard two-state linear extrapolation method (Figure 2-2) [22]. This analysis results in surprisingly low values for both ΔG and the m-value (the denaturant dependence of ΔG_{UN}). The calculated ΔG_{UN} is 3.4 kcal/mol, indicating an abnormally large destabilization from the wild type protein (wt ΔG_{UN} = 9.7 kcal/mol, see Table 2-1), and the m-value is 1.3 kcal mol^{-1} M^{-1}, much lower than the wild type m-value of 2.1 kcal mol^{-1} M^{-1} (Table 2-1). The m-value is believed to correlate with the burial of surface area upon folding [23]; therefore, if the overall fold is retained, this parameter should remain unchanged. The m-value we obtained for I25A, while clearly different from that expected based on the native structure, is strikingly similar to that determined for the transient folding intermediate of RNase H [5], suggesting that the native state was unexpectedly destabilized to the extent that the intermediate is now the most stable species.

	wt[6]	I53A[5]	D10A[5]	I25A		D10A/I25A
				Urea melt	Model	
ΔG_{UN} (kcal mol^{-1})	9.7	7.6	13	3.4 (0.3)	4.5	7.86 (0.07)
m_{UN} (kcal mol^{-1} M^{-1})	2.1	2.1	2.1	1.3 (0.1)	2.1[a]	2.1
ΔG_{UI} (kcal mol^{-1})	3.55	1.66	4.38	ND	3.55[a]	2.77[b]
m_{UI} (kcal mol^{-1} M^{-1})	1.24	1.24	1.45	ND	1.24[a]	1.15[b]

Table 2-1 Stability measurements of RNase H mutants. The values in parentheses represent the standard deviation calculated from at least three individual denaturant melts [5, 6]. The reported ΔG_{UN} for D10A/I25A was obtained by fixing the m-value to 2.1 kcal mol^{-1} M^{-1}. a) wildtype-values assumed for the model, b) values determined in Connell and Marqusee [9]

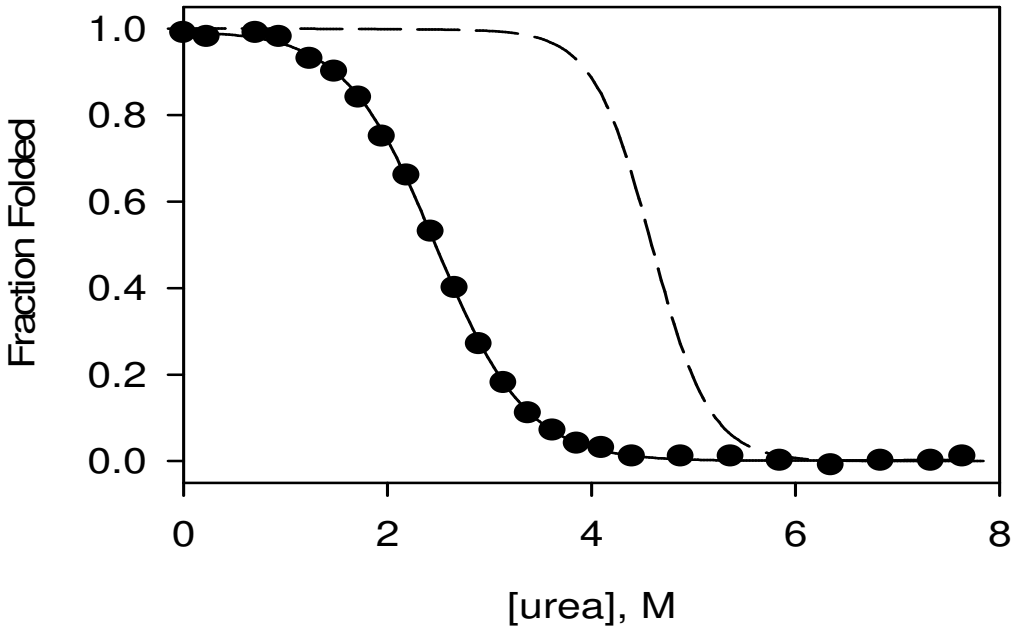

Figure 2-2 Fraction folded of I25A (filled circles) determined from urea melts overlaid with the wild type curve (dashed line). The solid line represents the two-state fit for I25A.

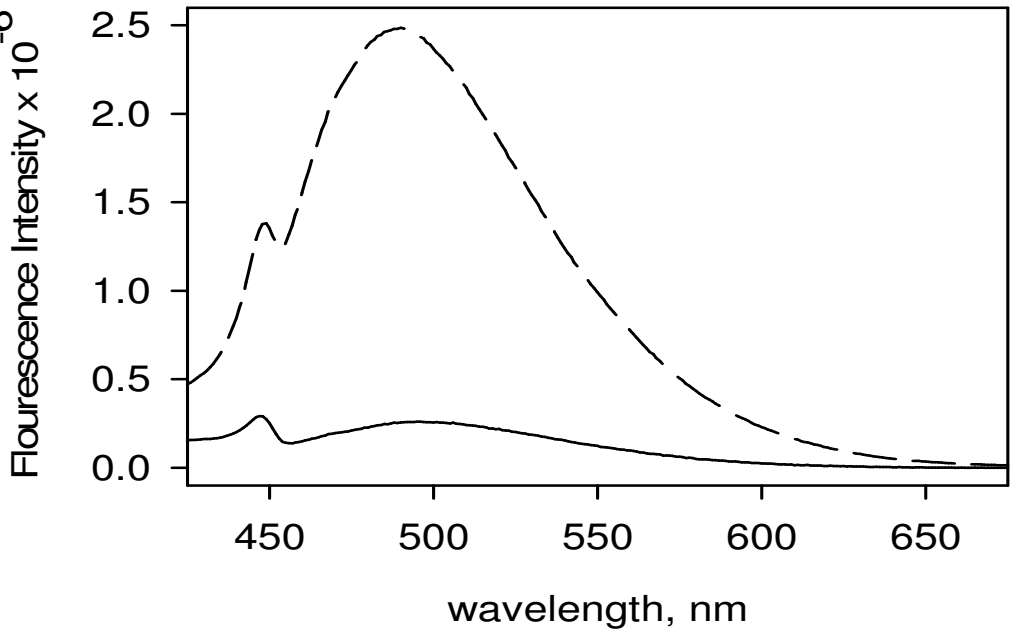

Figure 2-3 Fluorescence emission spectra of ANS in the presence of WT (solid line) and I25A (dashed line). The fluorescence emission of ANS in buffer alone has been subtracted from the data.

To further investigate the possibility that the I25A variant populates the intermediate state, 1-anilino-8-napthalene sulphonic acid (ANS) binding was monitored by fluorescence (Figure 2-3). The RNase H kinetic intermediate resembles its equilibrium molten globule [6], and one of the trademarks of molten globules is binding of the hydrophobic dye ANS. When ANS binds exposed hydrophobic surface area, its emission maximum blue-shifts and its fluorescence intensity increases [24]. Under native conditions, I25A incubated with ANS shows a large increase in fluorescence compared to the wild type enzyme; this increase is similar in magnitude to what is generally observed for molten globules.

2.3.2 I25A RNase H populates the native fold

In contrast to the evidence presented above, several studies demonstrate that I25A retains a native-like fold. The CD spectrum of I25A under native conditions is indicative of a folded, α/β protein and resembles that of wild type RNase H (Figure 2-4). I25A RNase H is also active, assessed by the loss of the hypochromic effect as the RNA strand is cleaved from a DNA-RNA hybrid (Figure 2-5). Preservation of the active enzyme, albeit impaired in comparison to wild type, implies that the protein can access the native fold under these conditions. To obtain higher resolution structural information, a native state ^{15}N-^{1}H HSQC NMR spectrum was obtained (Figure 2-6). I25A shows clearly dispersed peaks characteristic of a well-folded protein; in contrast, the spectra of molten globules and unfolded proteins exhibit a collapsed set of peaks in the ^{1}H dimension. Furthermore, the I25A peaks overlay well with those of wild type RNase H, clearly demonstrating that I25A populates the native structure under the conditions of the NMR experiments.

2.3.3 The two-state assumption is not valid for I25A

In light of the conflicting behavior of I25A RNase H, intrinsic tryptophan fluorescence was used to further probe equilibrium behavior via urea denaturation. RNase H has six Trp residues. Fluorescence intensity at 340 nm as a function of urea concentration was fit using the two-state approximation allowing a direct comparison between CD and fluorescence. In the case of wild-type RNase H, the curves representing the fraction folded as calculated by CD and fluorescence are indistinguishable [21]. For I25A, however, the fluorescence and CD data are non-coincident. The fit from the fluorescence results in an m-value 1.5 kcal mol^{-1} M^{-1} and ΔG_{UN} of 4.1 kcal/mol, which, while again significantly lower than wild-type values, do not match with those calculated by CD. This non-coincidence of results from different probes suggests that for I25A RNase H the two-state assumption breaks down. In support of this, the emission spectra do not have an isosbestic point, also indicating that more than two species contribute to the signal.

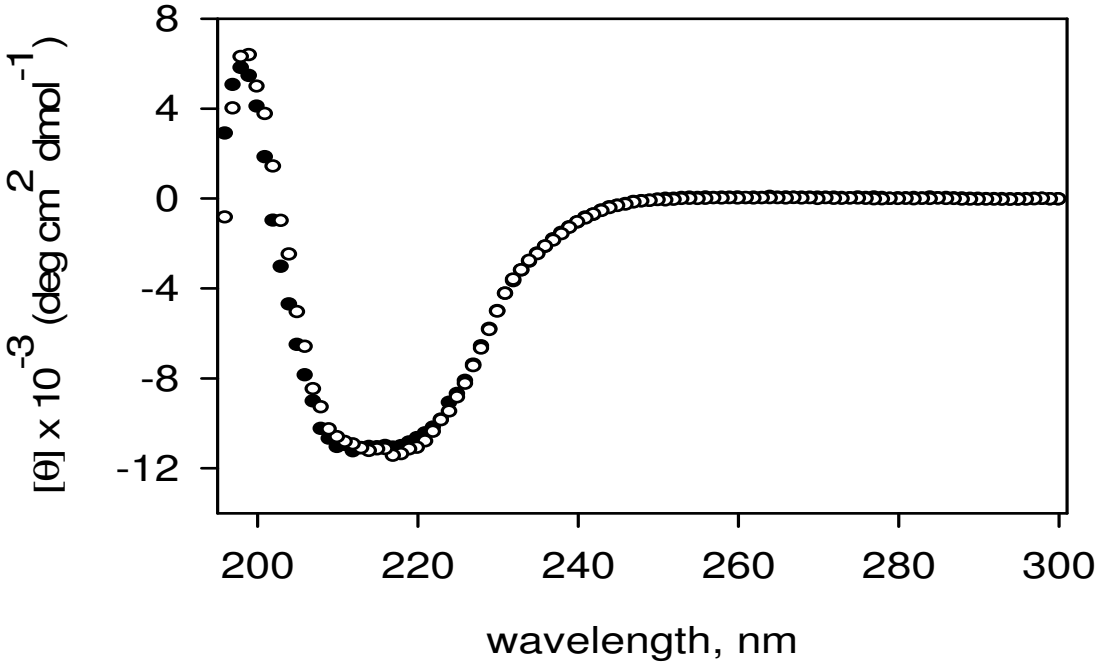

Figure 2-4 CD spectrum of I25A (filled circles) overlaid with wild type RNase H (open circles).

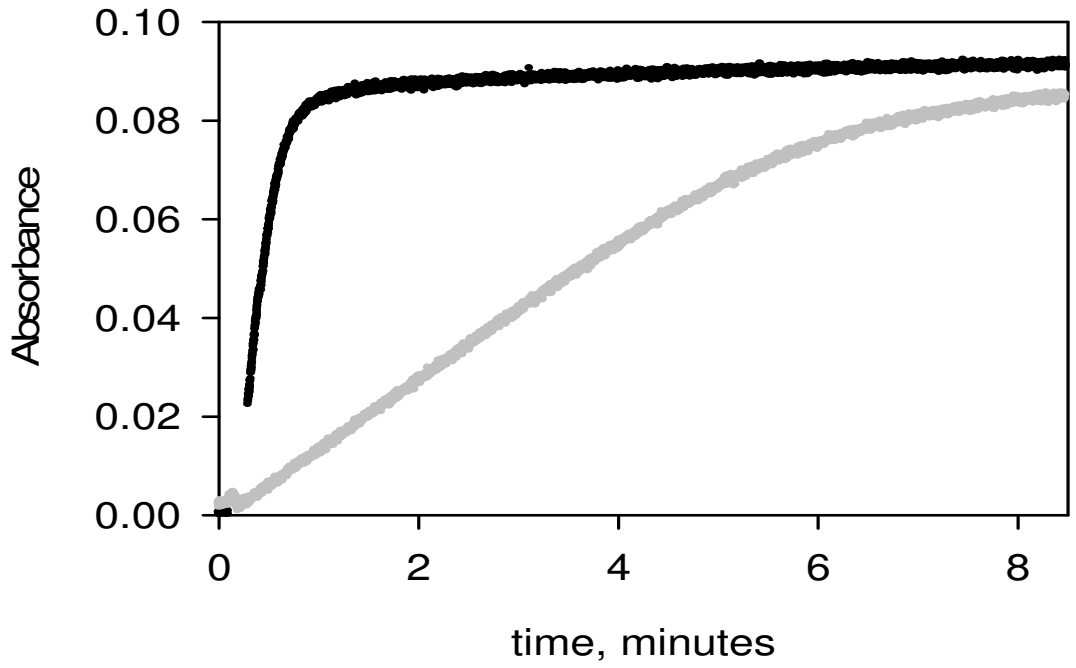

Figure 2-5 Activity of I25A (grey) compared with that of WT (black).

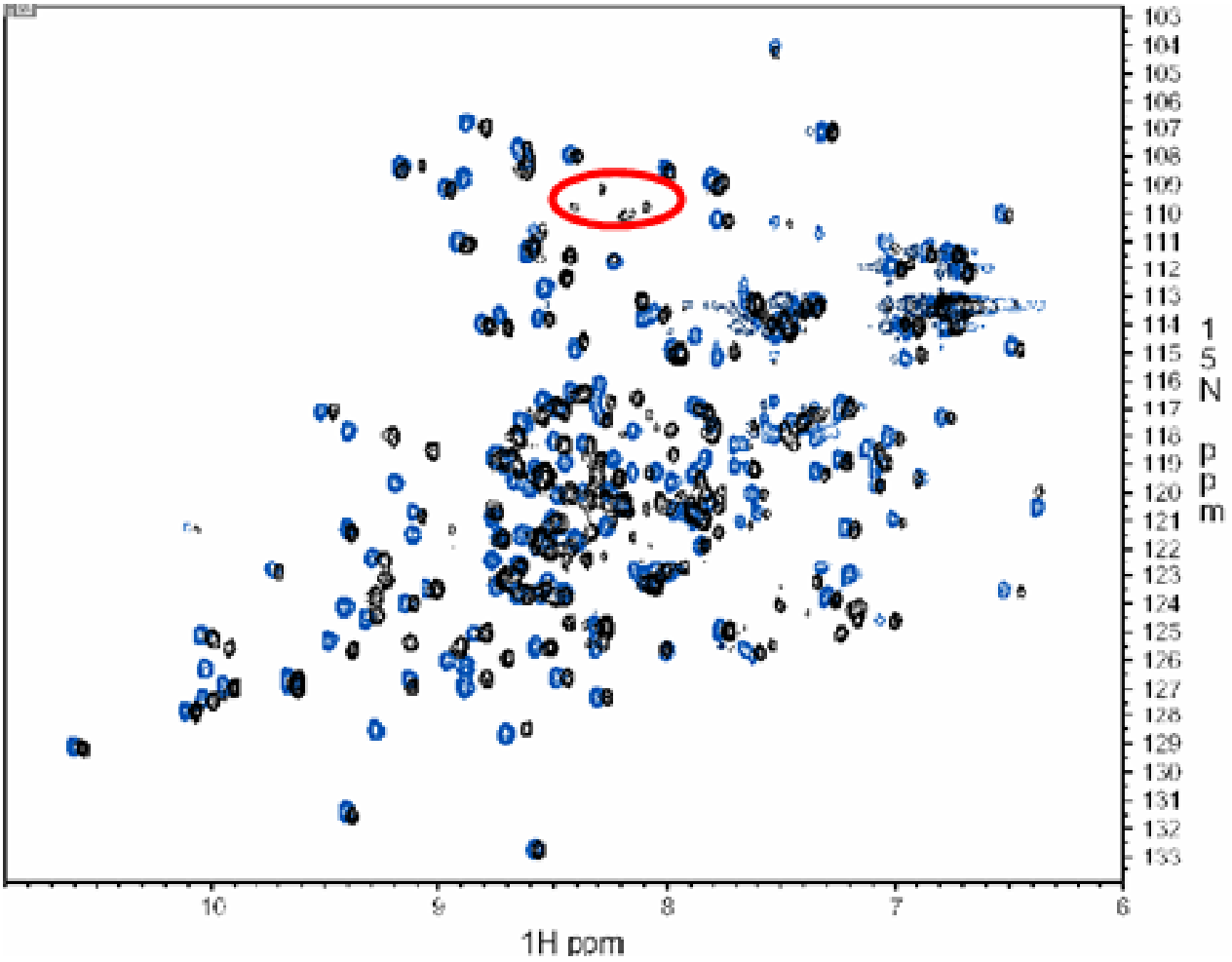

Figure 2-6 ^{1}H-^{15}N HSQC of I25A RNase H (black) overlaid with that of WT (blue) in 20mM NaOAc, pH 5.5, 50mM KCl, and 10% D_2O. Glycines region highlighted in red.

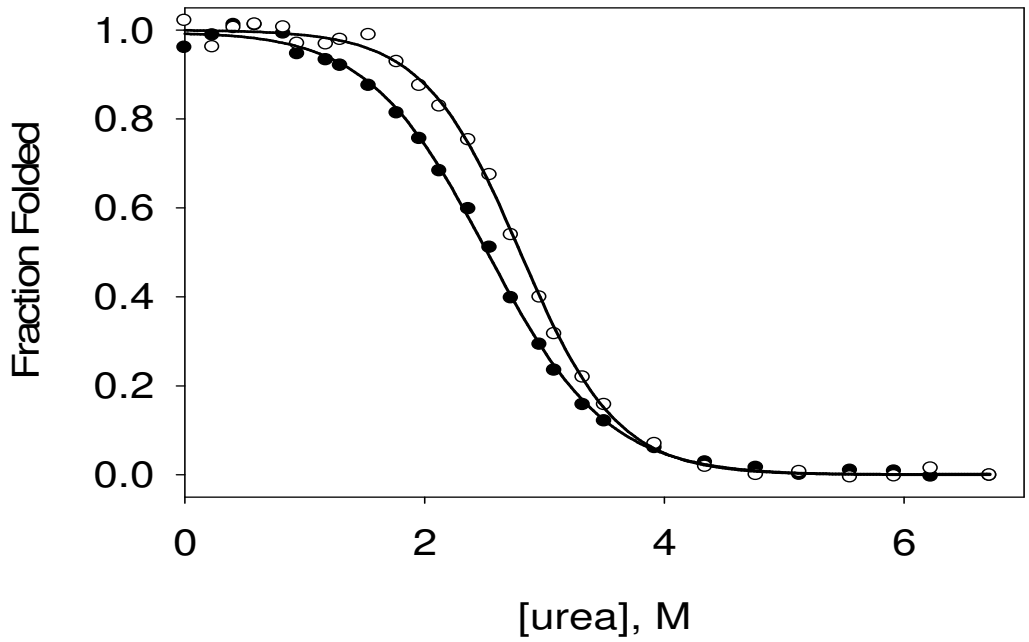

Figure 2-7 I25A folding lacks coincidence with different probes. Shown above, the fraction folded of I25A by CD (closed circles) and fluorescence (open circles) determined from urea melts fit to a two-state approximation. Solid lines represent each fit.

To obtain an estimate of the true global stability of the native conformation without assuming a two-state model, hydrogen exchange was carried out on the I25A variant. By monitoring the rate at which amide hydrogens exchange with solvent deuterium by NMR, we calculated the free energy of exchange for 63 residues ($\Delta G = -RT\ln(k_{obs}/k_{rc})$) [25]. The ΔG_{HX} values obtained for I25A, shown in Figure 2-8, reveal that the regions with the highest protection are the A and D helices, located in the core. These values suggest a global stability of approximately 6 kcal/mol. Since it is known that D_2O has an effect on protein stability [26], we repeated the denaturant melt in D_2O in order to compare the stability calculated by hydrogen exchange to that obtained by standard urea denaturation (data not shown). The stability of I25A in D_2O, assuming a two-state model, is increased by less than 1 kcal/mol, and the m-value remains the same. A large discrepancy, therefore, remains between the stability of I25A calculated assuming a two-state fit and that by hydrogen exchange (Figure 2-8). This incongruity is not evident for wild type RNase H. In this case the values for the global stability of the protein determined by native state hydrogen exchange and a two-state fit to the equilibrium denaturation profile are closely matched [10], confirming the validity of the two-state assumption. The inconsistency observed for I25A is additional evidence that the two-state assumption does not hold for this variant.

2.3.4 I25A populates the intermediate under native conditions

We determined the effect of I25A in the background of D10A, a mutation in the periphery known to stabilize RNase H by 3.3 kcal/mol (ΔG_{UN} = 13.0 kcal/mol) [6]. In a more stable background, there is less potential for a destabilizing mutation to affect the apparent two-state equilibrium transition. Equilibrium urea denaturation of D10A/I25A was monitored by CD and fit using the two-state approximation (Figure 2-9). The introduction of I25A into D10A destabilizes the protein by 5.1 kcal/mol (ΔG_{UN} = 7.9 kcal/mol) and results in an m-value of 2.1 kcal mol^{-1} M^{-1}, identical to that of wild type and D10A (Table 2-1). Assuming additivity provides a theoretical value for the stability of I25A of 4.6 kcal/mol. We constructed a model of the populations of native, intermediate, and denatured I25A RNase H as a function of urea using this value for ΔG_{UN} and the wild-type values for m_{UN}, ΔG_{UI} and m_{UI}. This model, shown in Figure 2-10, reveals that the stability of the native state of I25A is sufficiently close to the stability of the intermediate that a significant population of intermediate exists under native conditions and becomes the dominant species around 2 M urea (Figure 2-11).

2.3.5 NMR spectra in denaturant reveal two sets of peaks

Confident that I25A does in fact populate the intermediate state, we obtained HSQCs at varying urea concentrations in the hopes of obtaining structural information about the intermediate at the residue level (Figure 2-12). In low urea, the spectrum closely resembles that in no denaturant. Upon closer inspection, especially around the glycine region, however, there is a clear cluster of peaks that is not present in the wild type spectrum under native conditions. These peaks are circled in red in Figure 2-6 and Figure 2-12. As urea is added, these peaks intensify and persist to high concentrations of denaturant, indicating that they correspond to residues in the unfolded state.

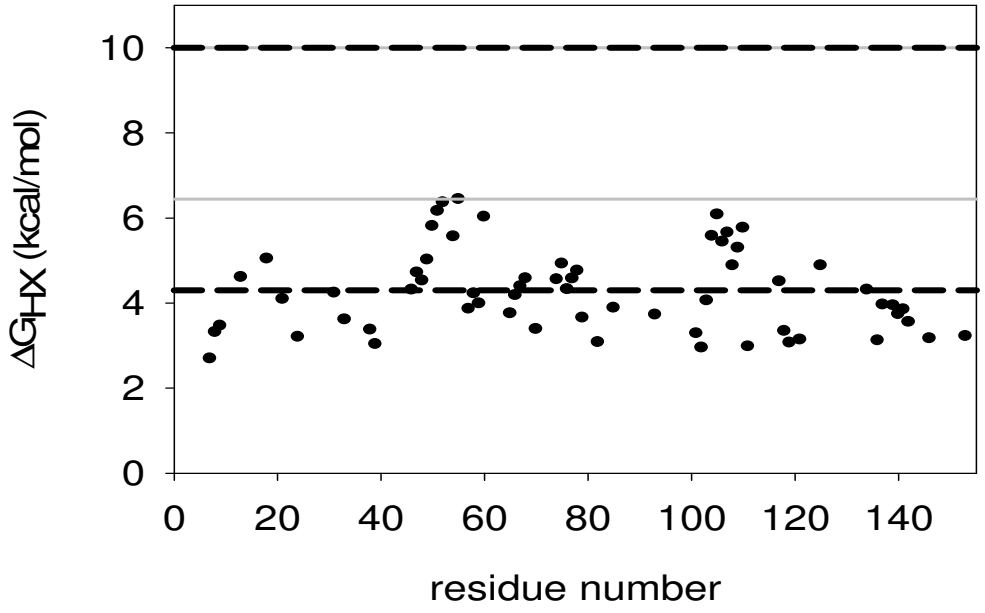

Figure 2-8 Hydrogen exchange data for I25A. The free energy of exchange of I25A is plotted as a function of residue. The grey continuous lines represent the stability of WT and I25A by hydrogen exchange methods, as indicated, and the dotted black lines represent the stability of each protein calculated from the two-state fits to melts acquired in deuterated solvent.

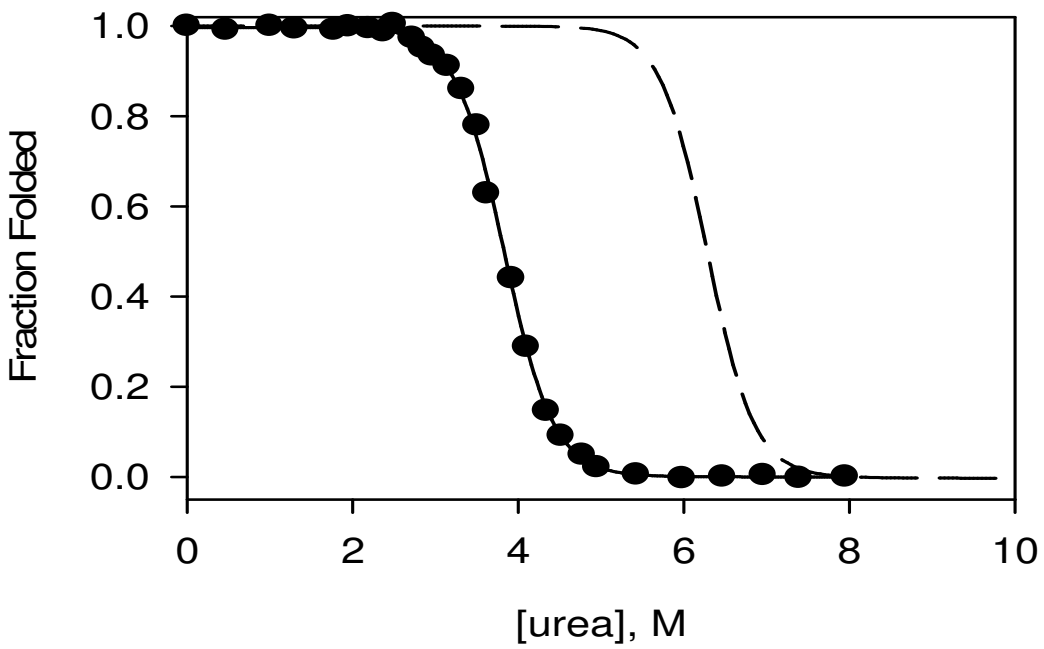

Figure 2-9 Determining the effect of I25A. Fraction folded of D10A/I25A determined from urea melts (filled circles) overlaid with curves for wild type (dashed line). The solid line represents the two-state fit to the D10A/I25A melt.

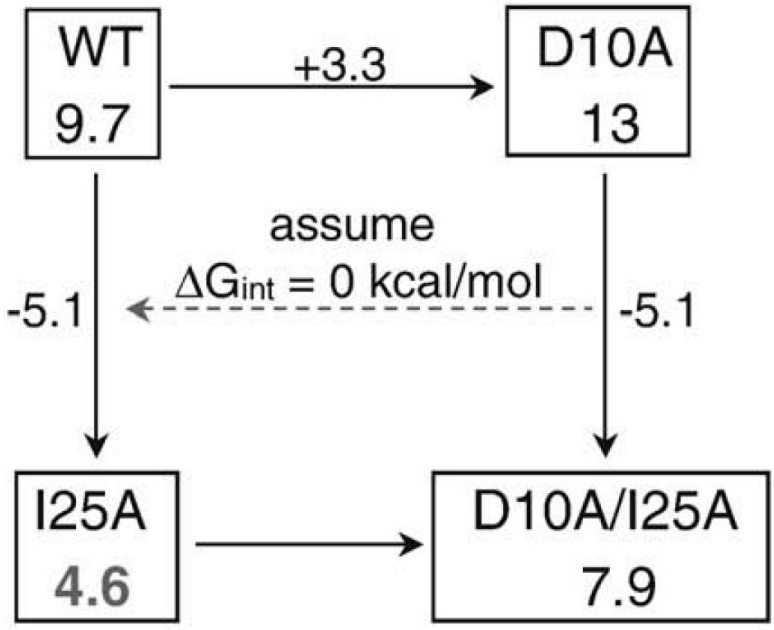

Figure 2-10 A thermodynamic cycle assuming no interaction between residues 10 and 25 was used to calculate the expected stability of I25A. Numbers in each box represent the global stability, ΔG_{UN} (kcal/mol), and the numbers along the arrows indicate the change in stability upon making the indicated mutation.

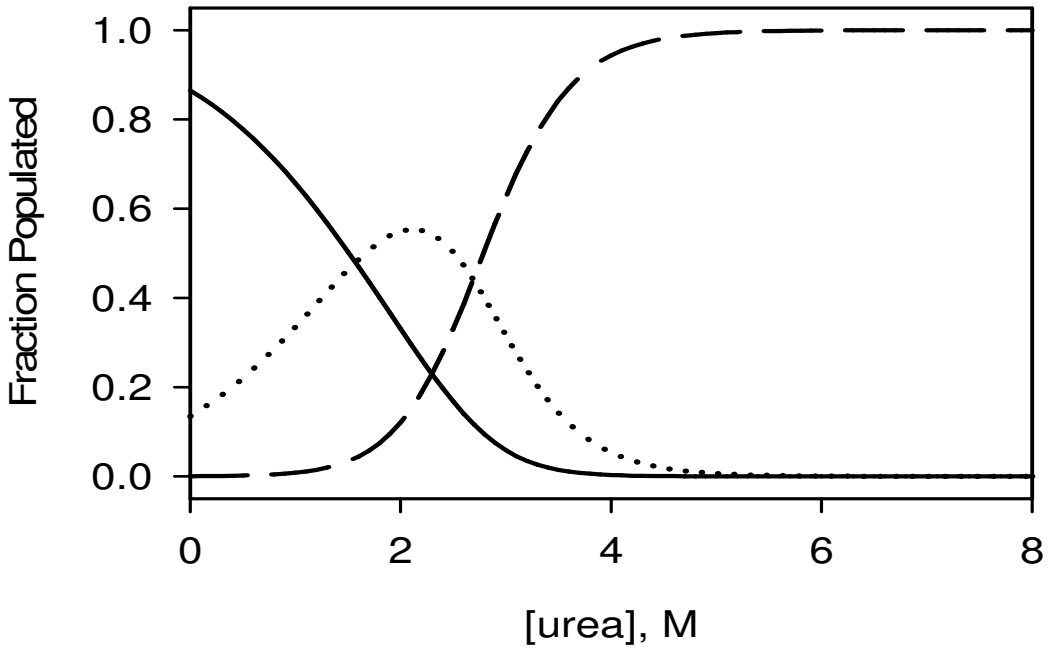

Figure 2-11 A graph of the theoretical populations for the native (solid), intermediate (dotted), and unfolded (dashed) forms of I25A as a function of urea concentration. The model assumes the native form of I25A is 4.6 kcal/mol stable with respect to the denatured form and that an m-value of 2.1 kcal mol^{-1} M^{-1} is associated with this transition. The intermediate is assumed to be 3.5 kcal/mol stable with respect to the denatured form, as it is in the wild type protein, with an m-value of 1.3 kcal mol^{-1} M^{-1}.

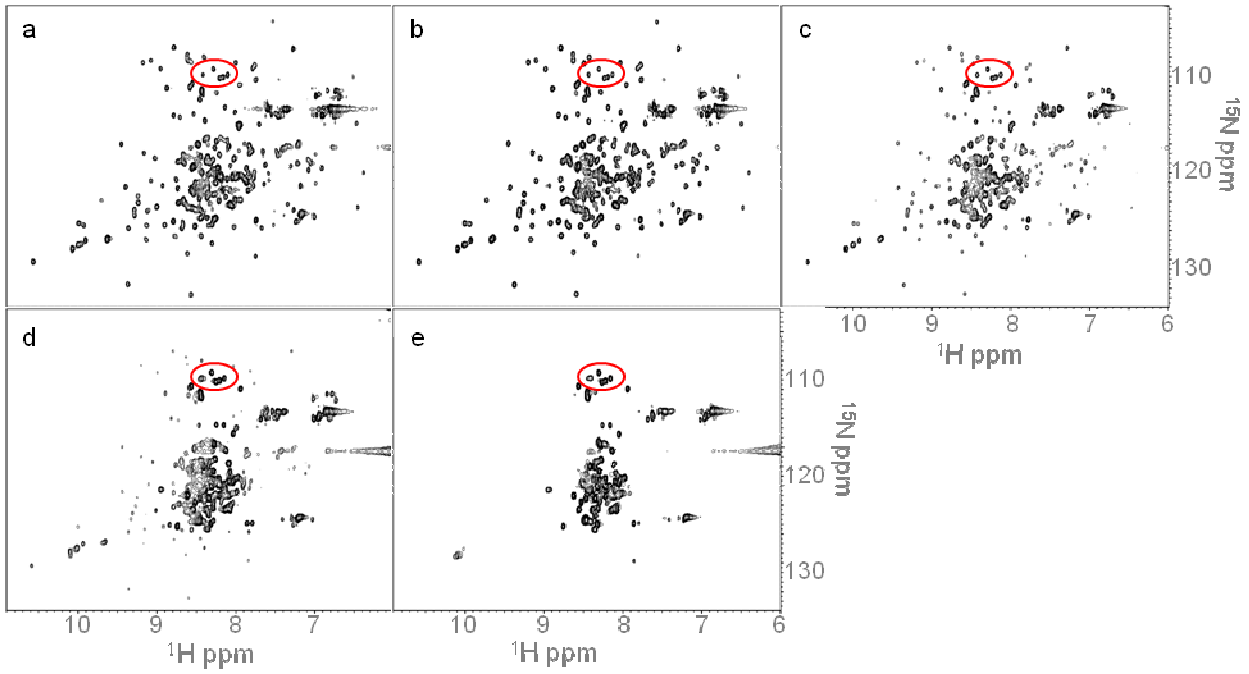

Figure 2-12 A series of HSQC's in 20 mM NaOAc, pH5.5, 50 mM KCl, 10 % D_2O and a) 0.5 M, b) 1.0 M, c) 1.5 M, d) 2.0 M, and e) 2.5 M urea.

Peaks in other regions of the spectrum also emerge at increasing urea concentrations. By 1.5 M urea (Figure 2-12), there are clearly two sets of peaks. One set corresponds to the dispersed set expected for the native conformation, and these decrease in intensity as denaturant increases. A second set, which builds in intensity with urea concentration, is characterized by a lack of dispersion in the ^{1}H dimension and appears to arise from residues in the unfolded state. This is in stark contrast to what we observe for the wild-type protein, where we see one set of peaks at all concentrations of urea.

2.4 Discussion

Protein engineering when combined with prior knowledge of the energy landscape of a protein, particularly from hydrogen exchange data, can be a useful tool to manipulate the population of protein conformations. Here, by creating a single destabilizing mutation in the periphery of RNase H (I25A) we show that it can be used to populate the partially folded intermediate known to be important in the folding trajectory of the protein, and thereby allowing us to study this experimentally elusive species at equilibrium.

2.4.1 Modeling the population of each species in I25A

Initial characterization of I25A provided conflicting views of the effect of the mutation. The CD spectrum, activity assay, and ^{15}N-^{1}H HSQC indicate that the overall structure of the protein is not notably perturbed from that of wild type. On the other hand, the significantly lower m-value obtained by the two-state fit of the denaturant melts and the ANS binding assay suggested that I25A resembles the intermediate conformation of RNase H. The noncoincidence of CD and fluorescence and the estimate of global stability obtained by hydrogen exchange indicate that the two-state approximation cannot be applied in this case. Examining the effect of the isoleucine to alanine mutation at position 25 in a more stable background (D10A) allowed us to obtain a better estimate of the effect of this mutation on the native state of RNase H. Surprisingly, this single mutation appears to destabilize the protein by more than 5 kcal/mol. Since the stability of the intermediate of RNase H is approximately 3.5 kcal/mol, a mutation of this magnitude that selectively destabilizes the native state (~10 kcal/mol) will result in a native state stability close to that of the intermediate and in a significant accumulation of the intermediate at equilibrium. The two-state assumption used to fit the denaturant melt is therefore invalid in this case, resulting in an inaccurate determination of the stability of the protein standard urea induced denaturation profiles.

Why does this single amino acid change have such a drastic effect on the stability of RNase H? By introducing alanine at position 25, we expected to destabilize the periphery by creating a small cavity in the hydrophobic pocket. Based on the change in hydrophobic surface area, we expect this mutation to destabilize the protein by ~ 3 kcal/mol [27]. At this point, we do not fully understand why the change is so much greater, although such large changes are not without precedent. For example, in T4

lysozyme, small perturbations (leucine to alanine) at buried sites have been known to result in large stability changes when coupled to the formation of a cavity [28]. In our case, this effect may also be exacerbated the presence of a nearby salt bridge network. Creating a cavity in the vicinity of these salt bridges may decrease the strength of the interaction, allowing the periphery to unfold more readily.

Our three-state equilibrium model successfully accounts for all of our experimental data. To model our data, the intermediate was assigned a CD signal 80% that of the native state based on kinetic studies on the wild-type protein [6]. Fitting the simulated melt to the two-state model reproduces the observed data with a calculated free energy of unfolding (3.5 kcal/mol) and the m-value (1.3 kcal mol^{-1} M^{-1}) of I25A RNase H.

The stability estimate obtained by hydrogen exchange is also consistent with our conclusions. This method should allow us to measure the free energy of hydrogen exchange of individual residues with an upper limit of the stability of I25A in D_2O. By looking at the most protected residues in the protein, those in helices A and D, we would estimate the stability of the protein to be ~ 5 - 6 kcal/mol. Studies of RNase H in the presence of D_2O suggest that the isotope effect increases the stability by just over 1 kcal/mol, which would put our estimate of the stability of I25A at ~4.5 kcal/mol, consistent with the value derived from assuming additivity with the mutation D10A. We attempted to repeat the urea melts used in this study in D_2O to get a more exact number, but urea is not stable for the long equilibration times required for the variants containing D10A. Nonetheless, our expectation of 4.5 kcal/mol is consistent with both the hydrogen exchange and mutant cycle analysis.

The HSQC spectra in varying urea concentrations provide further corroborating evidence that I25A populates the intermediate. The cluster of peaks circled in Figure 2-6 is not present in the wild type spectrum. They fall in the glycine region, and twelve of the fifteen glycines in RNase H are located in the periphery, a region presumed to be unstructured in the intermediate. These peaks persist in the higher urea concentrations (see peaks circled in red, Figures 2-6 and 2-12), where two sets of peaks are evident, one corresponding to folded protein and one characterized by the lack of dispersion typical of unfolded proteins. At these relatively low levels of denaturant, there should be no detectable unfolded protein, suggesting that these peaks arise from the intermediate state. In addition, similar spectra of the wild-type protein in varying amounts of urea do not exhibit two sets of peaks. We therefore attribute the glycine peaks and the appearance of the collapsed second set to unfolded regions of the intermediate.

No peaks appear to correspond to structured areas of the intermediate. The dispersed set of peaks appears to arise from the native state. At 2 M urea, where the intermediate should be maximally populated, the dispersed set of peaks overlays well with the native spectrum, and disappear together with the glycine peaks believed to arise from the native periphery. Together, these data suggest that the peaks from the folded core arise from the native structure, not the folded region of the intermediate. Although this set of peaks does exhibit some shift from the native spectrum, these shifts may arise from the addition of urea and are not of the magnitude we would expect if they were to originate from the intermediate. The gradual shift of peaks can also

indicate two species in fast exchange, the observed peak being an average of the separate signals from each of the populations rapidly interconverting. The intermediate and native structure are unlikely to be in fast exchange because the rate of the I•N step in H_2O is 0.74 s^{-1}, far slower than the NMR fast exchange regime. Instead we attribute the loss of peaks for the structured region of the intermediate to dynamic line broadening.

2.4.2 Insights into the nature of the folding intermediate

The HSQC spectra of molten globules commonly show no peaks that report on the structured regions, with signal building in only as structure is lost [29]. Fluctuations throughout the molten globule on the millisecond timescale are cited as the cause of peak broadening. We believe that, likewise, the absence of peaks arising from the intermediate in these experiments indicates that it is a dynamic species. This exchange between conformers suggests that the barriers between them are low and that they lack fixed tertiary interactions; contacts in the core, though native-like, are most likely weak.

This model of the folding intermediate of RNase H as a dynamic species agrees with studies on equilibrium molten globules that are thought to mirror the intermediates formed in the kinetic folding trajectory. The intermediate of apomyoglobin is also thought to be a heterogeneous and dynamic ensemble [30]. Our results, however, differ from those obtained on other systems where mutations were made to selectively populate the intermediate, such as Im7 and redesigned-apocytochrome b_{562}. Both of these proteins are four-helix proteins with one or more helices unstructured in their intermediates. NMR studies on the models of these intermediates indicate that the folded regions are quite rigid [15, 16, 19, 31], although the Im7 intermediate does show increased heterogeneity over the native state. In the cases of both Im7 and redesigned-apocytochrome b_{562}, the authors stress the existence of non-native contacts in the intermediate that need to be broken upon folding to the native state. Instead of utilizing non-native tertiary contacts to stabilize the intermediate, RNase H appears to fluctuate, forming weak native-like interactions that do not persist.

Our results are also at odds with recent NMR studies on a fragment of *T. thermophilus* RNase H that forms a well-folded native-like subdomain. The fragment was generated by removing regions from the periphery: two internal strands and the final helix of the protein, both of which appear unprotected by hydrogen exchange, while leaving one strand from the periphery (strand 1) which, based on native state hydrogen exchange and mutagenesis studies, also does not appear to be structured in the folding intermediate of *E. coli* RNase H. The inclusion of this strand may stabilize the fragment, allowing for a well-behaved sample amenable to high resolution NMR. The high-resolution structure reveals only native-like interactions, questioning the nature of the barrier in the progression of folding from the intermediate to the native state and at odds with the observation that mutations in this region affect the transition state stability more than the intermediate [5, 7]. The difference between what we find here and the NMR study on this fragment may also be attributed to a difference in the interactions that stabilize the intermediates from these two proteins (*E. coli* and *T.*

thermophilus). Another explanation might be that the well-folded fragment is actually a mimic of another identified high-energy intermediate of the *T. thermophilus* protein where just the last helix appears to be unfolded.

The dynamic nature of the *E. coli* RNase H intermediate may have important implications for efficient and productive folding in vivo. The intermediate of RNase H is populated prior to the rate-limiting step and its presence appears to aid in folding. Destabilizing the intermediate slows overall folding, suggesting that the interactions that stabilize the native state are present and important in the rate-limiting transition state. If interactions that form as the reaction progresses stabilize the transition state more than the intermediate, the folding reaction will be accelerated. This may be the case for RNase H, for which all evidence points to a productive, on-pathway, obligatory intermediate.

2.5 Materials and Methods

2.5.1 Materials

Deuterium oxide, ^{15}N ammonium chloride, deuterated buffers, acids, and bases were purchased from Isotec. All other buffer reagents were purchased from Sigma or Fischer.

2.5.2 Protein expression and purification

E. coli BL21 pLysS cells were transformed with the appropriate plasmid and grown at 37°C in LB medium with 200 μg/ml ampicillin. Induction was initiated by the addition of 1 mM IPTG to cells at OD600 ~0.6. Cells were harvested by centrifugation 3–4 hours after induction. I25A expressed in inclusion bodies, and purification was carried out on cell pellets as previously described for other RNase H variants [7]. D10A/I25A was expressed solubly and was purified as previously described [21]. Purity and molecular weights of all variants were verified by mass spectrometry (data not shown).

To express ^{15}N-labeled protein, log-phase cells were transferred to M9 medium containing ^{15}N ammonium chloride as described [32].

All experiments were carried out in 20 mM sodium acetate and 50 mM KCl at pH 5.5 unless otherwise noted. Protein concentrations were determined based on the extinction coefficient, calculated according to the number of Trp and Tyr residues [33].

2.5.3 Equilibrium CD experiments

Circular dichroism measurements were carried out on an Aviv 62DS spectrometer at 25°C. For denaturant melts, individual samples containing 40-50 μg/mL protein at varying urea concentrations were equilibrated overnight. For each sample, CD signal was monitored at 222 nm, and the signal was averaged over a 60 second time period. Data were fit assuming a two-state model and linear dependence of ΔG_{UN} on urea concentration.

2.5.4 ANS binding

Samples containing 500 μM 1-anilino-8-napthalene sulphonic acid in buffer with and without 2 μM protein were prepared and equilibrated overnight. Fluorescence

emission spectra were collected from 425 to 675 nm with an excitation wavelength of 405 nm. The spectrum of ANS in buffer alone was subtracted from those of ANS containing protein.

2.5.5 Activity assay

Activity of RNase H is monitored by the loss of the hypochromic effect as the RNA strand is cleaved from a DNA-RNA hybrid. The reaction was carried out in 50 mM Tris, 50 mM NaCl, 10 mM $MgCl_2$, and 10 μg/mL rA-dT and was initiated by the addition of enzyme to a final concentration of 5 nM. Absorbance at 260 nm was followed over time on a Cary UV-Vis spectrometer.

2.5.6 Tryptophan fluorescence measurements

Urea denaturation was monitored by tryptophan fluorescence using a Fluoromax 3 fluorimeter (JYHoriba) at 25oC. Individual samples of 40–50 μg/mL in varying urea concentrations equilibrated overnight. Excitation was at 295 nm, and emission spectra were recorded with both slits at 4 nm. Fluorescence at 340 nm as well as the center of mass were analyzed and fit using a two-state approximation and a linear dependence of ΔG_{UN} on urea concentration.

2.5.7 Hydrogen exchange

Amide hydrogen exchange was initiated by exchanging protonated ^{15}N I25A RNase H into deuterated buffer (20 mM sodium acetate and 50 mM KCl at pDr 5.6) using a polypropylene spin column (Pierce) packed with Sephadex resin. The sample was immediately transferred to an NMR tube and placed in the instrument; time between initiation of exchange and start of data collection was approximately 25 minutes. ^{15}N-^{1}H HSQC spectra were recorded on a Bruker 600 MHz at 25°C as an average of 16 scans with 1024 points in the direct dimension and 256 complex points in the indirect dimension. HSQC's (~1 hour each) were collected consecutively for 10 hours and then increasingly spaced out to two weeks. Spectra were processed using Felix 97.0 (Accrelys), and peak height as a function of time was fit to a single exponential decay in SigmaPlot (SSI) to obtain a value for k_{obs}. Wild type peak assignments were used to assign peaks for I25A. The free energy of exchange was calculated as: $\Delta G = -RT \ln (k_{obs}/k_{rc})$ where k_{rc} is the intrinsic rate of exchange for that residue in a random coil [25].

2.5.8 HSQC's at varying concentrations of urea

Two-dimensional, gradient-enhanced HSQC's were recorded on a Bruker 600 MHz at 25°C. 32 scans were collected with 1024 points in the direct dimension and 128 complex points in the indirect dimension. These data were processed using NMRPipe and viewed in NMRDraw.

2.6 Citations

1. Dyson, H.J. and P.E. Wright, *Coupling of folding and binding for unstructured proteins.* Curr Opin Struct Biol, 2002. **12**(1): p. 54-60.
2. Wright, P.E. and H.J. Dyson, *Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm.* J Mol Biol, 1999. **293**(2): p. 321-31.
3. Bose, H.S., et al., *The active form of the steroidogenic acute regulatory protein, StAR, appears to be a molten globule.* Proc Natl Acad Sci U S A, 1999. **96**(13): p. 7250-5.
4. Jahn, T.R., et al., *Amyloid formation under physiological conditions proceeds via a native-like folding intermediate.* Nat Struct Mol Biol, 2006. **13**(3): p. 195-201.
5. Raschke, T.M., J. Kho, and S. Marqusee, *Confirmation of the hierarchical folding of RNase H: a protein engineering study.* nature structural biology, 1999. **6**: p. 825-830.
6. Raschke, T.M. and S. Marqusee, *The kinetic folding intermediate of ribonuclease H resembles the acid molten globule and partially unfolded molecules detected under native conditions.* Nature structural & molecular biology, 1997. **4**(4): p. 298-304.
7. Spudich, G.M., E.J. Miller, and S. Marqusee, *Destabilization of the Escherichia coli RNase H kinetic intermediate: switching between a two-state and three-state folding mechanism.* Journal of molecular biology, 2004. **335**(2): p. 609-618.
8. Cecconi, C., et al., *Direct observation of the three-state folding of a single protein molecule.* Science, 2005. **309**(5743): p. 2057.
9. Connell, K.B., E.J. Miller, and S. Marqusee, *The Folding Trajectory of RNase H Is Dominated by Its Topology and Not Local Stability: A Protein Engineering Study of Variants that Fold via Two-State and Three-State Mechanisms.* Journal of molecular biology, 2009. **391**(2): p. 450-460.
10. Chamberlain, A.K., T.M. Handel, and S. Marqusee, *Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH.* Nature structural & molecular biology, 1996. **3**(9): p. 782-787.
11. Spudich, G., S. Lorenz, and S. Marqusee, *Propagation of a single destabilizing mutation throughout the Escherichia coli ribonuclease HI native state.* Protein Sci, 2002. **11**(3): p. 522-8.
12. Goedken, E.R. and S. Marqusee, *Native-state energetics of a thermostabilized variant of ribonuclease HI.* J Mol Biol, 2001. **314**(4): p. 863-71.
13. Bai, Y., *Energy barriers, cooperativity, and hidden intermediates in the folding of small proteins.* Biochemical and Biophysical Research Communications, 2006. **340**(3): p. 976-983.
14. Bai, Y., H. Feng, and Z. Zhou, *Population and structure determination of hidden folding intermediates by native-state hydrogen exchange-directed protein engineering and nuclear magnetic resonance.* Methods in molecular biology (Clifton, NJ), 2007. **350**: p. 69.
15. Feng, H., N.D. Vu, and Y. Bai, *Detection and structure determination of an equilibrium unfolding intermediate of Rd-apocytochrome b562: native fold with non-native hydrophobic interactions.* J Mol Biol, 2004. **343**(5): p. 1477-85.
16. Feng, H., Z. Zhou, and Y. Bai, *A protein folding pathway with multiple folding intermediates at atomic resolution.* Proc Natl Acad Sci U S A, 2005. **102**(14): p. 5026-31.
17. Kato, H., H. Feng, and Y. Bai, *The folding pathway of T4 lysozyme: The high-resolution structure and folding of a hidden intermediate.* Journal of molecular biology, 2007. **365**(3): p. 870-880.
18. Kato, H., et al., *The folding pathway of T4 lysozyme: An on-pathway hidden folding intermediate.* Journal of molecular biology, 2007. **365**(3): p. 881-891.
19. Spence, G.R., A.P. Capaldi, and S.E. Radford, *Trapping the on-pathway folding intermediate of Im7 at equilibrium.* J Mol Biol, 2004. **341**(1): p. 215-26.

20. Chamberlain, A.K. and S. Marqusee, *Molten Globule Unfolding Monitored by Hydrogen Exchange in Urea†.* Biochemistry, 1998. **37**(7): p. 1736-1742.
21. Dabora, J.M. and S. Marqusee, *Equilibrium unfolding of Escherichia coli ribonuclease H: Characterization of a partially folded state.* Protein Science, 1994. **3**(9): p. 1401-1408.
22. Greene, R.F., Jr. and C.N. Pace, *Urea and guanidine hydrochloride denaturation of ribonuclease, lysozyme, alpha-chymotrypsin, and beta-lactoglobulin.* J Biol Chem, 1974. **249**(17): p. 5388-93.
23. Myers, J.K., C.N. Pace, and J.M. Scholtz, *Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding.* Protein Sci, 1995. **4**(10): p. 2138-48.
24. Semisotnov, G.V., et al., *Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe.* Biopolymers, 1991. **31**(1): p. 119-28.
25. Bai, Y., et al., *Primary structure effects on peptide group hydrogen exchange.* Proteins, 1993. **17**(1): p. 75-86.
26. Makhatadze, G.I., G.M. Clore, and A.M. Gronenborn, *Solvent isotope effect and protein stability.* Nat Struct Biol, 1995. **2**(10): p. 852-5.
27. Xu, J., et al., *The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect.* Protein Sci, 1998. **7**(1): p. 158-77.
28. Eriksson, A.E., et al., *Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect.* Science, 1992. **255**(5041): p. 178-183.
29. Redfield, C., *Using nuclear magnetic resonance spectroscopy to study molten globule states of proteins.* Methods, 2004. **34**(1): p. 121-32.
30. Nishimura, C., H. Jane Dyson, and P.E. Wright, *The apomyoglobin folding pathway revisited: structural heterogeneity in the kinetic burst phase intermediate.* Journal of molecular biology, 2002. **322**(3): p. 483-489.
31. Gsponer, J., et al., *Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein Im7.* Proc Natl Acad Sci U S A, 2006. **103**(1): p. 99-104.
32. Marley, J., M. Lu, and C. Bracken, *A method for efficient isotopic labeling of recombinant proteins.* J Biomol NMR, 2001. **20**(1): p. 71-5.
33. Edelhoch, H., *Spectroscopic Determination of Tryptophan and Tyrosine in Proteins*.* Biochemistry, 1967. **6**(7): p. 1948-1954.

Chapter 3. Novel Protein Function Engineered by Selectively Reorganizing a Protein Core

3.1 Abstract

Protein design and directed evolution are moving closer to the *de novo* design of novel protein function. This process is far from routine, however, and it is worth assessing new approaches to deriving novel function in proteins. We have attempted to use constrained directed evolution to select for novel binding specificity by exclusively evolving the hydrophobic core of the transcriptional activator MarA. To restrict the evolution of MarA to hydrophobic rearrangements, we generated mutant libraries of MarA in three buried clusters of nonpolar residues. We selected these libraries for novel activity by coupling MarA binding of an engineered promoter sequence to expression of tetracycline resistance in *E. coli*. We identified a number of mutants with altered specificity, demonstrating that core rearrangements can be sufficient to alter protein function.

3.2 Introduction

The promise of designed protein function has tremendous potential. Novel enzymes could combine the efficiency of catalysis with the facility of biological molecules. Computational *de novo* design [1, 2] and directed evolution [3-5] experiments are advancing our ability to generate novel proteins and protein functions. Many challenges still remain, and many design experiments succeed by limiting their scope and redesigning an existing protein activity for novel binding specificity rather than novel catalysis (for review see [6-8]). Current efforts in protein engineering are focused on refining methods that improve success and better understand evolutionary pathways. Approaches such as negative selection, modulating folding stability, and sequential selection thresholds have shown promise in evolving novel proteins more efficiently [7]. But a detailed understanding of what works in generating novel function is still missing. Here we have asked if evolutionary rearrangements restricted by structural localization can select for novel function. Specifically, we have selected for novel function by exclusively mutating the hydrophobic core.

Protein cores are densely packed and nonpolar. Mutations within the core, particularly when they introduce charge, are more destabilizing than mutations elsewhere in the structure [9]. Core residues are rarely involved directly in enzyme catalysis or binding as they are sequestered away from the protein surface. As a result, many design projects begin with a folded core "scaffold" and design sequences with novel function onto it [2, 3]. There is reason to believe that core mutations are overlooked in directed evolution of protein function. Many proteins with redesigned

function do not accumulate mutations in the active site but at distant sites elsewhere in the protein [7, 10]. Furthermore, while mutations in protein cores are more destabilizing, the native structure will often fold, as cores are able to rearrange and accommodate new residues, particularly if they are non-polar [11].

The repacking of hydrophobic cores was investigated elegantly in an experiment by Lim and Sauer that probed the limits of core rearrangement in λ-repressor [12]. Lim et al. made a library of core mutants by doing saturated mutagenesis on the seven amino acid residues that make up the core of λ-repressor. They then selected for mutants with functioning cores by plating transformed cells on media containing phage lacking a repressor gene. Cells that could express a functional repressor protein were not killed by phage and produced colonies. The mutations that allowed growth consistently preserved hydrophobic character, volume composition, and avoided steric clash. By lowering the stringency of their selection, they could define the minimal threshold for functional repacking and found as much as 70% of repacked cores had at least partial activity. The mutants that grew demonstrated that many repacked cores could remain folded and functional.

We have chosen as a model system the DNA binding protein, MarA. MarA, shown in Figure 3-1, is a transcriptional activator of the Mar operon and is expressed in response to antibiotic and superoxide exposure in prokaryotes. It activates transcription of many stress response genes, and as many as 100 such genes have been identified in *E. coli* [13]. MarA activates genes by binding a degenerate 20 base pair region of an upstream promoter and recruiting RNA polymerase to bind upstream of the transcription start site [14]. MarA binds as a monomer at the recognition sequence, called the 'marbox', in a directional fashion [15]. It is unusual to find transcription factors which do not bind symmetrical sequences of DNA as multimers. Even λ-repressor, described above, binds its operator sequence as a dimer. This makes MarA an attractive model system for repacking. Its function can be assessed by monitoring downstream gene expression, and its folding and stability are not complicated by oligomerization and concentration effects. Furthermore, work by Rosner and colleagues have detailed much of the biochemistry of this protein [16, 17]. Both a crystal structure and a solution structure of MarA (in complex with DNA) have been solved [18, 19]. While the protein is not well behaved in the absence of its DNA substrate, it is experimentally tractable. Structural characterization, particularly by NMR, indicates that MarA interacts with its DNA substrate in a highly dynamic state. We attempted to reengineer the sequence specificity of MarA by repacking its hydrophobic core.

3.3 Results

3.3.1 The selection System

We attempted to select novel binding specificity from a library of reorganized MarA core variants. To do this, we took advantage of phylogenic analysis of MarA

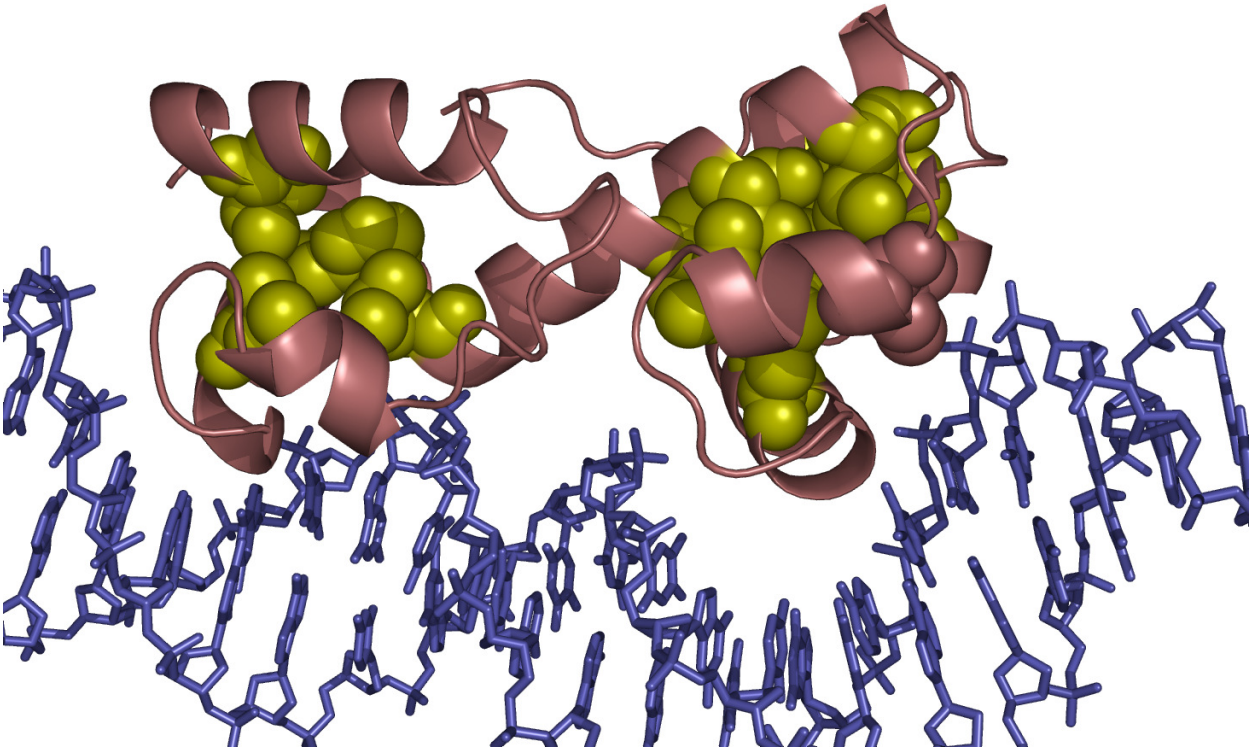

Figure 3-1 Structure of MarA (1BL0). Shown bound to DNA substrate, with helices three and six contacting the major groove at two points. Buried residues that make up hydrophobic core libraries are shown as spheres in yellow [18].

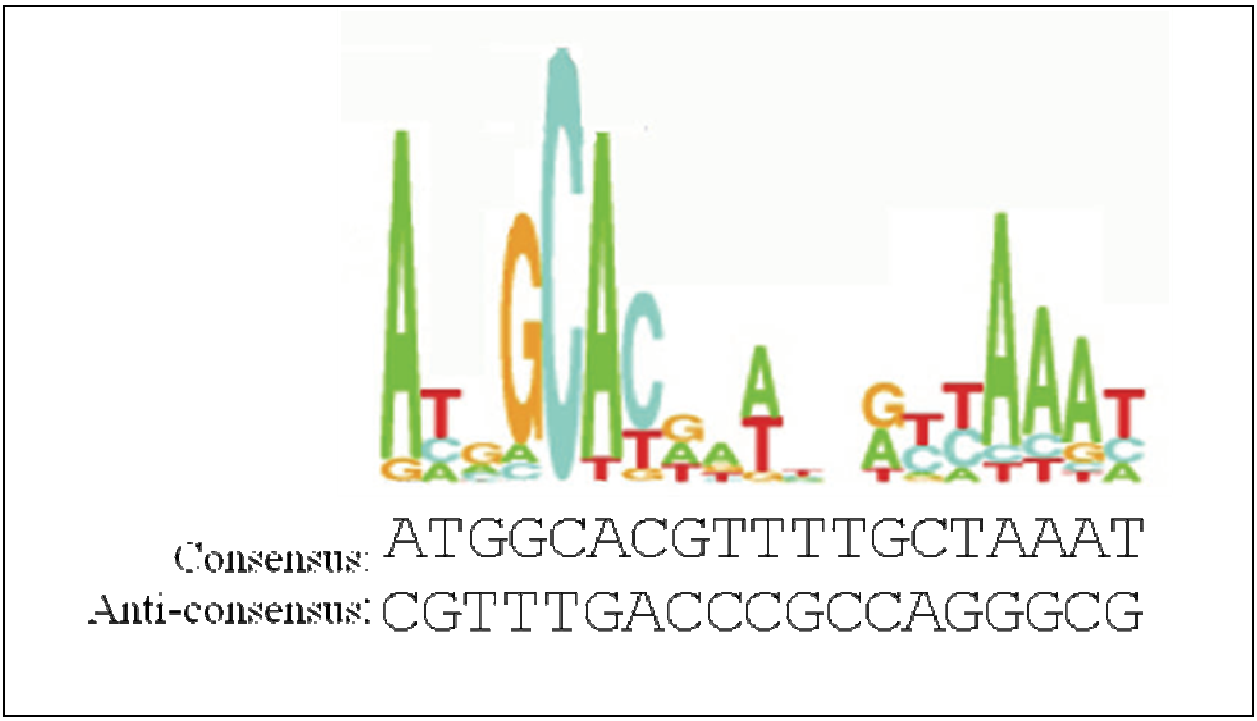

Figure 3-2a Design of selection plasmid sequences. Marbox sequences from phylogenic analysis shown in sequence logo [20]. Consensus and Anti-consensus sequences are shown below derived from the most and least frequent bases at each position.

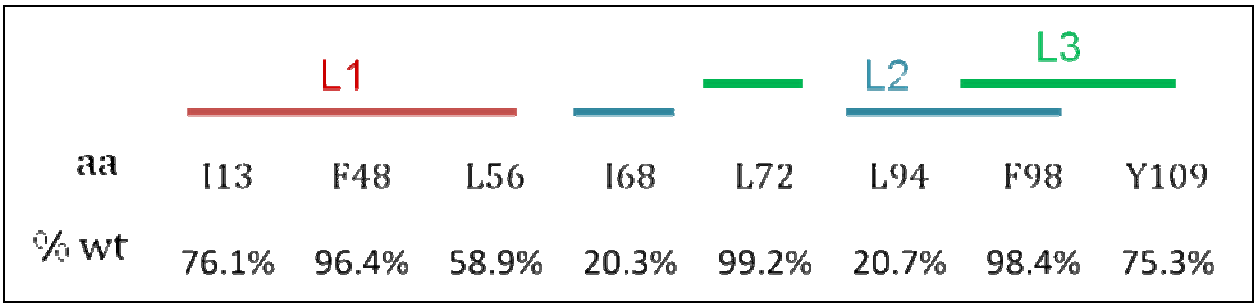

Figure 3-2b MarA residues chosen for the three libraries with their conservation in MarA homologs shown as percentages below. Library 2 and Library 3 share one residue and exist in the same cluster above helix six.

binding in prokaryotes and derived consensus and anti-consensus sequences based on base pair frequencies of many MarA promoters (see Figure 3-2a) [20, 21]. A consensus sequence was derived from the most frequent base pairs at each position, while the anti-consensus sequence was derived from the least frequent base-pairs. Wild-type MarA binds to the consensus sequence at nanomolar K_d, but does not bind the anti-consensus sequence at detectable levels based on previous selections [20]. We selected for variants that would bind to this anti-consensus sequence with the consensus promoter used as a control. In order to monitor binding of MarA to this novel anti-consensus binding sequence we used a plasmid designed by Shultzaberger et al. [20], which put the TetA gene downstream of the marbox initiating site. In this system (shown in Figure 3-3) binding of MarA at the anti-consensus sequence will result in tetracycline resistance.

To restrict evolution of MarA exclusively to the core we rationally designed three libraries which spanned the hydrophobic core of MarA. Using saturated mutagenesis at each site, we created three libraries which each contained all possible combinations of mutations at three residues, termed Library 1 (residues Ile13, Phe48, and Leu56), Library 2 (residues Ile68, Leu94 and Phe98), and Library 3 (residues L72, Phe98, and Y109) (shown in Figure 3-2b). This produced more than 10^4 testable sequences for selection. MarA binds DNA with two helices inserted into the major groove, and as a result its structure is elongated with two hydrophobic cores clustered against each DNA-contacting helix (see Figure 3-1). Our three libraries were chosen to include both the N-terminal and C-terminal hydrophobic core and to contain buried residues that were closely packed based on solved structures of MarA. Tyrosine 109 had a very small amount of exposed surface area from the hydroxyl group but was largely buried and part of a central hydrophobic cluster. Our mutant libraries were cloned into the selection plasmids which carried the same MarA promoter and tetracycline resistance, with MarA under an L-arabinose inducible promoter. With this plasmid system we could monitor MarA DNA sequence affinity with *E. coli* growth on tetracycline.

3.3.2 Library confirmation

With mutant libraries ligated into a selection plasmid system, we transformed into electrocompetent knockout *E. coli* missing endogenous MarA as well as its paralogs SoxS and Rob. We induced MarA expression by plating overnight on LB amp plates in the presence of 0.1% L-arabinose. Transformed colonies were collected and grown in liquid culture (containing ampicillin and L-arabinose) for six hours and then diluted and plated on 30 μg/ml tetracycline with 0.1% L-arabinose. Selected colonies were picked and evaluated further for novel binding function. A positive control plasmid containing wild type MarA protein sequence but with the consensus marbox promoter grew a lawn of cells on tetracycline-containing plates. A negative control plasmid, which contained a wild type MarA sequence in a selection plasmid with the anti-consensus marbox promoter, was selected for and after 36 hours a few colonies were seen. This was expected due to the ability of some *E. coli* in culture to acquire

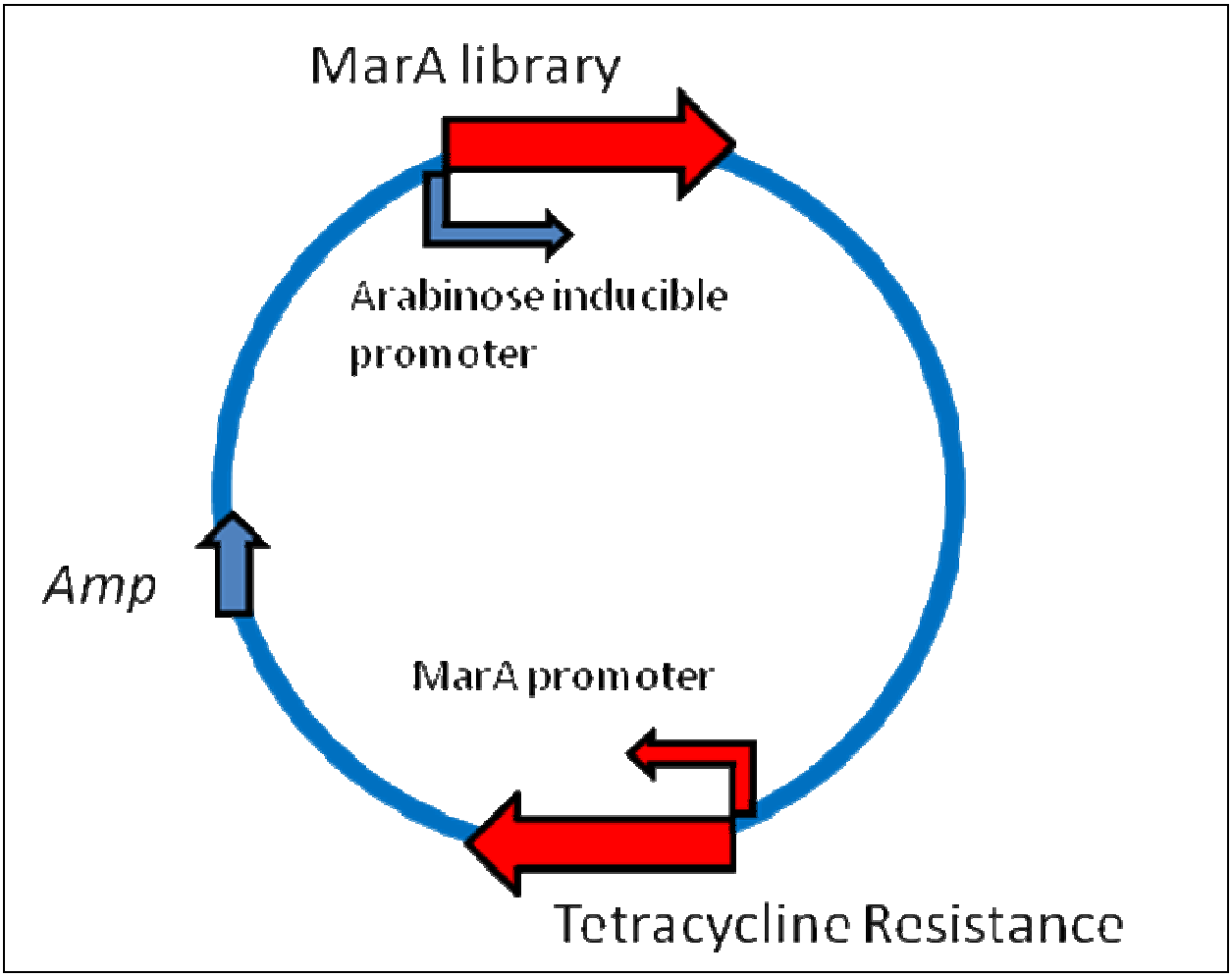

Figure 3-3 Diagram of selection plasmid. MarA is induced by the presence of L-arabinose. TetA is induced by MarA binding. This plasmid also has common components of pBR322 vector design, including ampicillin resistance and a low-copy origin of replication [20].

L72	F98	Y109
-	-	-
-	-	-
-	-	-
-	-	-
A	E	-
A	L	-
A	V	D
A	L	F
A	S	I
A	I	K
A	E	L
A	D	S
A	D	V
A	R	V
D	P	C
D	S	H
D	K	K
D	D	R
E	E	A
E	H	S
E	G	T
F	S	A
F	C	L
F	S	R
G	N	D
G	R	E
G	G	I
G	F	L
G	R	L
G	H	S
G	N	S
G	T	S
G	M	T
H	K	L
I	P	I
I	P	I
K	G	C
K	N	E
L	-	-
L	K	A
L	P	C
L	C	F
L	F	F
L	L	F
L	E	G
L	L	I
L	L	I
L	D	S
L	L	S
L	P	S
L	R	V
L	F	Y
L	G	Y
M	V	S
P	P	I
P	A	P
P	T	R
P	W	S
P	R	Y
Q	S	T
Q	Y	T
R	-	-
R	E	E
R	P	E
R	Y	F
R	S	Y
R	Y	Y
S	G	G
S	G	K
S	H	L
S	R	L
S	T	S
S	S	V
T	S	L
T	T	R
T	W	S
T	L	Y
V	I	-
V	D	I
V	F	I
V	H	I
V	M	P
V	C	V
V	L	V
V	Y	Y
W	K	-
W	K	C
W	D	L
W	P	R
Y	P	F
Y	V	K

Table 3-1 Sequences from Library 3 with no selection. Colonies were grown after transformation and plated on ampicillin. A demonstration of wide diversity in our selected library.

tetracycline resistance over time. These few colonies could not grow at elevated tetracycline levels. To avoid sequencing false positives we restreaked all colonies at 50μg/ml tetracycline. Those streaks that grew after 24 hours at elevated tetracycline concentrations were sequenced.

To prove that the density of plated cells was even and dispersed, we also plated on ampicillin in parallel with tetracycline. Colonies from this mixture were picked and sequenced to ensure that the library had appropriate diversity. The sequences from Library 3 are shown in Table 3-1, and display wide variation with very few duplicates.

3.3.3 Selection for binding to the consensus promoter

We first transformed libraries in selection plasmids with the consensus marbox promoter, thus selecting for wild-type behavior in the mutant library similar to Lim et al. [12]. Many colonies grew in these conditions and, as an initial analysis, we sequenced over 100 from Libraries 2 and 3. The mutations that were identified are very conservative (see Table 3-2). Similar to the results with λ-repressor, they conserve hydrophobicity in every instance and often conserve volume and packing. The number of mutants correlated strongly with residue conservation. Library 2, which contains residues with much more variability by homology to other MarA and AraC family activators, had far more mutants than Library 3. Library 3, which has two residues which are conserved in more than 98% of homologous sequences (see Figure 3-2b), only had two mutants identified. At the individual residue level, phenylalanine 98, which is present in Library 2 and 3 and strongly conserved in homologs, did not vary in either library, indicating its importance for a functional structure.

3.3.4 Selection results in anti-consensus

We were primarily interested in mutants which could generate novel function. Compared to the number of mutants which grew when selected for binding to the consensus marbox sequence, we observed far fewer colonies when we selected for the anti-consensus marbox promoter sequence and incubation times were often longer. We were encouraged to see colonies, which when restreaked were not false positive effects, suggesting that our selection would yield results. Initial selections found far more colonies in Library 3, and so we focused our initial analysis on this library. The sequences of picked colonies from Library 3 are found in Table 3-3. Due to sequencing errors, redundancies, and spontaneous mutations, to date, we have only catalogued a list of 12 core mutants that were identified in five experiments. Generally the sequences preserve hydrophobic character, though to a far lesser degree than consensus-binding sequences. There are several examples in which polar residues, and even charged residues, are introduced into the core. It was interesting that Library 3, which was highly conserved in the consensus library selection, generated the most colonies in the

Library 2				Library 3			
I68	L94	F98	(frequency)	L72	F98	Y109	(frequency)
L	L	F	17	*L*	*F*	*Y*	16
L	V	F	7	L	F	F	17
M	L	F	6	V	F	Y	1
V	L	F	6				
V	F	F	6				
L	M	F	5				
I	M	F	3				
V	M	F	2				
L	F	F	2				
F	L	F	1				
A	F	F	1				

Table 3-2 Sequences identified by selecting for core mutants in Library 2 and 3 that can bind to the consensus marbox sequence. The residues in italics under Library 3 represent the number of wild-type sequences we found. Wild type was never identified in selection of Library 2.

Library 3		
L72	F98	Y109
D	T	M
I	D	Q
I	S	R
P	M	D
S	E	R
S	P	V
S	P	L
S	C	R
T	I	S
W	P	V
W	T	L
Y	M	Y

Table 3-3 Selection of MarA binding with novel specificity. Sequences were identified by selecting for core mutants in Library 3 that can bind to the anti-consensus marbox sequence.

MarA sequence	Promoter	[Tet] µg/ml				
		20	**25**	**30**	**40**	**50**
Library 3	Anti-consensus	1000s	20	5	0	0
	Consensus	lawn	lawn	lawn	lawn	1000s
wt MarA	Anti-consensus	1000s	3	1	0	0
	Consensus	lawn	lawn	lawn	lawn	1000s
3DTM	Anti-consensus	lawn	lawn	1000s	200	50
	Consensus	500	100	50	0	0
3SCR	Anti-consensus	lawn	lawn	1000s	750	50
	Consensus	500	200	8	1	0

Table 3-4 Number of *E. coli* colonies found from transformation of MarA sequences identified in library selection. Cloned sequences are from Library 3 and read L72D, F98T, Y109M and L72S, F98C, Y109R. Colony numbers are compared to wild type and the library colony numbers at different tetracycline concentrations in plasmids that select for binding to the consensus or anti-consensus marbox sequence. Where cell growth was so dense that it could not be counted *lawn* is listed. Numbers are approximate averages of at least three plates.

anti-consensus marbox selection. Also the most conserved residues, leucine 72 and phenylalanine 98, were never mutated in anti-consensus selected cores.

3.3.5 Selection confirmation

To confirm that these novel-binding mutants were not the result of changes elsewhere in the genome of our knockout *E. coli* strain, we purified plasmid from two mutants from Library 3 that showed the strongest growth in tetracycline. The mutants, L72D, F98T, Y109M (3DTM) and L72S, F98C, Y109R (3SCR) were then sub-cloned into a selection vector with the consensus marbox. We transformed both vectors in parallel into MarA knockout cells and selected on tetracycline as before. We counted colonies grown in selective conditions and used these numbers as a crude confirmation of novel binding. As shown in Table 3-4, these mutants grew many colonies on tetracycline in the anti-consensus plasmid background; importantly, these variants did not rescue growth when combined with the consensus promotor (over background). Wild type MarA in this system can only resist high tetracycline with the consensus promoter. Unlike selection from the library of mutants where these mutants were a fraction of the sequences present, these transformations are direct comparisons with wild type MarA *in vivo*. The robust improvement in tetracycline resistance is an affirmation that these sequences bind with a novel specificity.

3.4 Discussion

We can conceive of two mechanisms by which MarA could evolve new DNA affinity. In one instance, binding would represent entirely novel affinity for the anti-consensus sequence alone, novel specificity, shifting binding capacity from the consensus sequence to the anti-consensus. Alternatively, MarA mutants could lose sequence specificity and become general DNA binders which would have affinity for all sequences. The difference in growth on anti-consensus and consensus in the subcloned mutants confirmed that at least two of our most robust mutants have novel specificity and not indiscriminate DNA affinity.

We cannot be sure that binding is specific for the anti-consensus sequence, but further characterization of anti-consensus binding indicates that it is likely. With the help of Ryan Shultzaberger, we assessed our promoter sequence upstream of TetA with an algorithm to search for MarA binding sites. We did not identify any likely alternative binding sites within 100 base-pairs of the transcription start site (data not shown). Martin et al. observed tight regulation of the spacing between a marbox and the transcription start site. Shifting the marbox just three residues reduced transcription by fivefold [15]. Given the strict positional requirements for transcription of wild-type MarA *in vivo*, and the lack of strong alternative binding sites nearby, we are confident that the MarA induced transcription represents novel binding affinity.

We first assessed the output of our experiment by selecting from our core libraries with the consensus sequence promoter. By selecting for wild-type binding function, we were performing an experiment similar to Lim and Sauer, probing the constraints of the functional core [12]. Our results reproduced the observations seen with λ-repressor nicely. The mutations we find are very conservative. All observed mutations preserve hydrophobic character and most preserve sterics and volume. It was interesting that the propensity of a residue to be mutated in our libraries correlated with phylogenic conservation of other MarA homologs. This is especially notable in residue phenylalanine 98, which is highly conserved in other MarA proteins (>99%) and was never altered in any mutant despite being in both Library 2 and 3. Conservation overall seemed to be an inverse predictor of library mutability. Library 3 is the most conserved of the three core clusters we chose; its average conservation is 89% for three residues. We only identified two mutants out of 59 sequences in the consensus selection of Library 3. Alternatively Library 2, which is less conserved, showed much higher mutability. Detailed statistics of these data await high throughput sequencing studies using next-generation sequencing technology.

This pattern of conservation seemed to be reversed when we attempted to select for non-native function. When we selected our libraries for binding to the anti-consensus sequence we found more colonies in Library 3. For this reason we focused on collecting sequences from Library 3. The mutations that we collected were much less conservative than those found selecting for consensus promoter binding. Many in fact introduce charge residues into the core. Residue phenylalanine 98 is mutated in all of the mutants we collected, whereas it had been conserved very strictly in the consensus binders. The severity of the mutations seen in mutants binding the anti-consensus sequence indicates a much larger change in the structure of the protein.

The close relation of conservation to mutation rates indicates that our selection is reflecting real patterns of evolution. While our results would suggest that mutations within the core are sufficient to generate novel function, they would also suggest that the most efficient method to evolve new function will be one that combines repacking of the core with mutations at the surface. Nature almost certainly makes use of this broader strategy in evolutionary selection, optimizing distant as well as near effects in concert to generate the fittest protein.

Many proteins would not be amenable to generating novel binding through rearrangements at a distant site in the core. MarA possesses a number of unique qualities that may make it easier to redesign. Being a monomer which binds directionally, MarA doesn't require preservation of a dimerization interface or symmetry along the DNA sequence for binding as would be required for many other DNA-binding proteins. With more degrees of freedom to evolve, it is easier to find a working structure. MarA is relatively poorly behaved in the absence of its substrate, indicating that its folding is closely coupled to the presence of cognate DNA. MarA binds with remarkably non-specific interactions. NMR data have shown that very few of the sequence specific interactions with DNA involve specific bonds, such as

hydrogen bonds [19]. Instead, MarA appears to bind its substrate through shape specificity and van der Waals interactions. Thus if we have rearranged the interior of the protein the corresponding fluctuations at the surface are less likely to be breaking up specific coding interactions and more likely to find a new surface that can bind a novel sequence. NMR analysis also revealed that the bound state of the protein is quite dynamic and populated by more than one conformation even when coupled with a DNA sequence to which it had high specificity. Like work on cyclophilin A by Kern and colleagues, apo MarA may populate many conformations with different binding affinities [22]. This makes novel specificity all the more accessible. Instead of creating novel specificity for the entire native ensemble, we could be shifting the population of a previously-rare conformation within the native ensemble. To what extent these results are MarA-specific we cannot know. Whether such core-driven binding changes are a general phenomenon will require similar experiments on other proteins. To our knowledge, this is the only example of such an attempt.

We set out to determine whether we could generate mutants in the hydrophobic core with novel ligand binding specificity. Our results suggest that this is clearly possible. Current efforts are working to confirm these results with higher numbers and the confidence of statistics. While some of our mutants appeared in more than one screen, many were only found once. As more mutants are generated we will be able to make comparisons between them and perhaps even use the selection protocol in competition, determining which resist tetracycline most robustly and therefore which have the highest affinity for the new sequence. Once we have generated a large set of novel binding core mutants, it will be important to confirm, via biochemical analysis (specifically a gel-shift assay), that these mutants behave *in vitro* as they have *in vivo*. MarA has already been subjected to analysis of this kind and so this should not be a technical barrier [23]. Finally, once we have isolated a set of the most robust mutants we would determine their ligand-binding preferences. We intend to create of library of MarA promoters and select for binding using each of our mutants to confirm that we have evolved a novel binding landscape.

These data serve as a proof of principle that the core repacking can drive formation of novel function. Hydrophobic burial drives folding, and core arrangements are central to the presentation of residues on the surface of the protein. The MarA mutants we have identified do not bind as tightly as wild type MarA, but they do display marked differences in sequence affinity. This may suggest that if directed evolution and design experiments are going to become more adept at generating novel protein function, explicit attention to core repacking should be considered.

3.5 Materials and Methods

3.5.1 Plasmids, libraries and strains

The selection plasmid used in these experiments was provided by Ryan Shultzaberger in the lab of Michael Eisen. Briefly, this plasmid represents a combination of an arabinose inducible plasmid containing MarA (pBAD18-hisMarA), and a pBR322 derived segment containing the Tetracycline A (TetA) gene under the control of a MarA binding site [20]. MarA binding sites contained either a strong binding site, the consensus marbox site, or an anti-consensus site representing the least frequent base pair at each sequence position.

Core libraries were prepared by inspecting solved structures of MarA and determining buried, contacting residues that spanned the core of MarA. Saturated mutagenesis at the sites in the three libraries was done by GeneArt. Libraries were PCR amplified and then were cloned into the selection vectors listed above, which contained TetA controlled by consensus and anti-consensus marboxes. This was begun with restriction digest with XmaI and BsrGI, of plasmid and backbone. The backbone was phosphatase treated to reduce reannealing. Ligation was performed overnight at 16 °C with T4 ligase. Ligations were heat inactivated at 65 °C for 20 minutes and then ethanol precipitated with carrier yeast tRNA. This ligated DNA preparation was transformed into electrocompetent XL10 Gold cloning cells and plated on LB Ampicillin. Colonies were checked and approximately counted the next day and enough ligation was plated to exceed 40,000 colonies. These colonies were scraped, combined and pelleted. Plasmid was extracted from these pellets using a Qiagen Midipred kit.

The knockout *E. coli* strain N8453 (Δmar, Δsoxs::cat, Δrob::kan) used in selection experiments was obtained from Ryan Shultzaberger in the lab of Michael Eisen and originally prepared by Judah Rosner.

All reagents and enzymes were purchased from Fisher.

3.5.2 Selection protocol

Novel specificity mutants were isolated by first transforming electrocompetent N8453 *E. coli* with library plasmid in duplicate. Electroporation was done at 2.0 V, 25 μF, and 200 Ohms in a Bio Rad Gene Pulser. Transformations were then incubated in LB for 1 hour with shaking, and then plated on LB agar plates containing 30 μg/ml Ampicillin and 0.1% L-arabinose. Plates grew overnight, and consistently grew to dense lawns. Cells were scraped and grown in liquid culture with the same concentrations of Amp and L-arabinose for 4-6 hours. From these cultures, 0.3 μl was diluted to 300 ml of LB and plated on LB agar plates with 30 μg/ml tetracycline and 0.1% L-arabinose. Plates were grown for 36 hours.

3.5.3 Sequencing and colony confirmation

To avoid the occurrence of false positives we restreaked all colonies of interest from 30 μg/ml to 50 μg/ml tetracycline and 0.1% L-arabinose. Streaks that grew after 24 hours were picked into 20% sterile glycerol and stored. From these sterile glycerol stocks we did colony PCR with primers designed to complement the MarA sequence of the gene. Colony PCRs were cleaned up to eliminate primers and enzyme and then submitted for sequencing at the UC Berkeley Sequencing Facility. Confirmed sequences were also sequenced in their promoter region to confirm that they were in fact selecting for the correct promoter sequence. Sequences after collection were aligned with the ClustalW multiple alignment algorithm and sequences without errors or spontaneous mutations elsewhere in the sequence were used for further characterization.

3.6 Citations

1. Jiang, L., et al., *De novo computational design of retro-aldol enzymes.* Science, 2008. **319**(5868): p. 1387.
2. Röthlisberger, D., et al., *Kemp elimination catalysts by computational enzyme design.* Nature, 2008. **453**(7192): p. 190-195.
3. Park, H.S., et al., *Design and evolution of new catalytic activity with an existing protein scaffold.* Science, 2006. **311**(5760): p. 535.
4. Lodeiro, S., T. Schulz-Gasch, and S.P.T. Matsuda, *Enzyme redesign: two mutations cooperate to convert cycloartenol synthase into an accurate lanosterol synthase.* J. Am. Chem. Soc, 2005. **127**(41): p. 14132-14133.
5. Fasan, R., et al., *Engineered alkanehydroxylating cytochrome P450 (BM3) exhibiting nativelike catalytic properties.* Angew Chem Int Ed, 2007. **46**: p. 8414-8418.
6. Gerlt, J.A. and P.C. Babbitt, *Enzyme (re) design: lessons from natural evolution and computation.* Current opinion in chemical biology, 2009. **13**(1): p. 10-18.
7. Tracewell, C.A. and F.H. Arnold, *Directed enzyme evolution: climbing fitness peaks one amino acid at a time.* Current opinion in chemical biology, 2009. **13**(1): p. 3-9.
8. Jäckel, C., P. Kast, and D. Hilvert, *Protein design by directed evolution.* 2008.
9. Alber, T., *Mutational effects on protein stability.* Annual review of biochemistry, 1989. **58**(1): p. 765-792.
10. Umeno, D., A.V. Tobias, and F.H. Arnold, *Diversifying carotenoid biosynthetic pathways by directed evolution.* Microbiology and molecular biology reviews, 2005. **69**(1): p. 51.
11. Baldwin, E.P., et al., *The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme.* Science, 1993. **262**(5140): p. 1715.
12. Lim, W.A. and R.T. Sauer, *Alternative packing arrangements in the hydrophobic core of lambda repressor.* Nature, 1989. **339**(6219): p. 31-6.
13. Barbosa, T.M. and S.B. Levy, *Differential expression of over 60 chromosomal genes in Escherichia coli by constitutive expression of MarA.* Journal of Bacteriology, 2000. **182**(12): p. 3467.
14. Martin, R.G., et al., *Complex formation between activator and RNA polymerase as the basis for transcriptional activation by MarA and SoxS in Escherichia coli.* Mol Microbiol, 2002. **43**(2): p. 355-70.
15. Martin, R.G., et al., *Structural requirements for marbox function in transcriptional activation of mar/sox/rob regulon promoters in Escherichia coli: sequence, orientation and spatial relationship to the core promoter.* Molecular Microbiology, 1999. **34**(3): p. 431-441.
16. Jair, K.W., et al., *Purification and regulatory properties of MarA protein, a transcriptional activator of Escherichia coli multiple antibiotic and superoxide resistance promoters.* J Bacteriol, 1995. **177**(24): p. 7100-4.
17. Martin, R.G., et al., *Autoactivation of the marRAB multiple antibiotic resistance operon by the MarA transcriptional activator in Escherichia coli.* Journal of Bacteriology, 1996. **178**(8): p. 2216.
18. Rhee, S., et al., *A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator.* Proc Natl Acad Sci U S A, 1998. **95**(18): p. 10413-8.
19. Dangi, B., et al., *Structure and dynamics of MarA-DNA complexes: an NMR investigation.* J Mol Biol, 2001. **314**(1): p. 113-27.
20. Shultzaberger, R.K., *Functional Variability in Transcriptional Initiation Complexes.* Unpublished Doctoral Thesis, University of California Berkeley, 2009.

21. Martin, R.G., et al., *Autoactivation of the marRAB multiple antibiotic resistance operon by the MarA transcriptional activator in Escherichia coli.* J Bacteriol, 1996. **178**(8): p. 2216-23.
22. Eisenmesser, E.Z., et al., *Intrinsic dynamics of an enzyme underlies catalysis.* NATURE-LONDON-, 2005. **438**(7064): p. 117.
23. Martin, R.G. and J.L. Rosner, *Binding of purified multiple antibiotic-resistance repressor protein (MarR) to mar operator sequences.* Proc Natl Acad Sci U S A, 1995. **92**(12): p. 5456-60.

Appendix A. Folate Remedial Phenotypes Derived by Destabilized Polymorphisms in Folate Binding Proteins

The work in this section was carried out in collaboration with Nick Marini and Jasper Rine. I carried out all the protein biochemistry.

A.1 Abstract

Using high doses of vitamins to treat hereditary diseases has been shown to be clinically beneficial. In some cases, the remedial effects of vitamins are attributed to vitamin concentrations that compensate for deleterious mutations. We attempted to investigate the role that destabilizing mutations might play in such remedial effects by characterizing a number of clinically significant polymorphisms in folate binding proteins. We hoped to couple *in vivo* effects of enzyme dysfunction to *in vitro* measurements of protein stability. We used yeast-based growth rates as *in vivo* assessments of cofactor remediation in polymorphism phenotypes by generating knockout strains and monitoring complementation by transformed mutants. Work in a number of folate binding proteins attempted to match *in vivo* yeast based growth to biophysical characterization of the stability impacts of these mutations in folding. Unfortunately, several target proteins that were tractable in yeast genetics were not tractable *in vitro* and vice versa.

A.2 Introduction

Vitamins are essential molecules that facilitate normal metabolism, aid in development, and maintain normal cell function. Vitamin deficiency can have severe consequences, ranging from seizures due to riboflavin deficiency [1] to spinal cord degeneration in cases of inadequate cobalamin levels (vitamin B12) [2]. Treatment of deficiency cases is straightforward: simply provide the missing nutrient. But can vitamins be beneficial beyond cases of deficiency? Clinical successes suggest that in some cases the administration of vitamins significantly above RDA levels is useful in treating hereditary disorders [3]. Pyridoxine (vitamin B6) and thiamine (vitamin B1) have been used as so called "remediative vitamin therapeutics", but other vitamins are less widely used [4-6]. A review of clinical literature has suggested that many vitamins and nutrients can have therapeutic effects when used at high doses [3]. Few therapies are as simple and accessible as vitamin administration: we wondered if vitamin remedial behavior is more widespread than previously acknowledged and whether it might be reduced to a simple molecular description.

A.2.1 A mechanism for vitamin remedial mutants

While the benefit of such large doses has been proposed before, less attention has been paid to the mechanism of this remediative effect. Linus Pauling, long a proponent of therapeutic administration of vitamins, conjectured that vitamin supplementation would benefit patients when a mutation had affected the ability of an enzyme to bind substrate or coenzyme. These mutants would be differentiated from wild type enzyme

by demonstrating raised Michaelis-Menten constant, K_m , values. He wrote that for these mutants "the catalyzed reaction could be made to take place at or near its normal rate by an increase in the substrate concentration" [7].

Single nucleotide polymorphisms (SNPs) appear on average four times in every gene or 1 per 1000 base pairs [8]. A number of metabolic proteins have been identified with SNPs that measurably increase K_m values. In the enzyme ornithine aminotransferase, polymorphisms A226V and V332M both raise the K_m and have responded to higher levels of pyridoxine, its cofactor [9]. In folate enzymes, methylenetetrahydrofolate reductase (MTHFR) has SNP A222V, which has been associated with hyperhomocysteinemia, and is remediated by higher doses of folate [10].

The Michaelis-Menten constant, K_m, describes the substrate concentration for an enzyme at its half maximal rate of activity, and when product formation is rate limiting it can be interpreted as a measure of propensity to form an enzyme-substrate complex. A mutation could impede complex formation, and raise the K_m, by more than one mechanism. We concluded that mutations could affect enzyme K_m in two ways. Any ligand-binding enzyme is in equilibrium with both its native and unfolded state, as well

$$U + nL \rightleftharpoons N + nL \rightleftharpoons NL_n$$

as with its ligand-bound and unbound state. The scheme shown above makes the reasonable assumption that ligands bind preferentially to the native form of the enzyme. A mutation that disrupts cofactor binding at the active site, decreasing its K_d, would shift this equilibrium and make the enzyme appear less active. For example, thymidylate synthase mutation T51S has a measurably higher K_m for folate with no change in apparent stability or structure [11]. Poor binding can be compensated by increasing cofactor concentrations, and wild-type levels of enzyme activity can be recovered.

An alternative mechanism to alter an enzyme's K_m could lie, not in disrupting the binding site, but the stability of the enzyme itself. Enzymes are only catalytic once folded to their lowest energy, native conformation [12]. Mutations can shift a protein's energy landscape thereby changing the equilibrium of $U \rightleftharpoons N$. Mutants could be destabilized such that they promoted unfolding or partial folding, decreasing the fraction of protein capable of binding cofactor. In the case of proteins which bind a ligand, the ligand concentration significantly affects the folding equilibrium, and the energetics of folding are impacted according to $\Delta G_{binding} = -RT\ln(1 + [L]/K_d)$ [13]. Excess cofactor will shift the equilibrium (assuming preferential binding of the folded form) of folded protein towards the bound native state. Because the equilibrium of folding is dependent on the logarithm of the free energy, small changes in free energy of folding may not significantly change the amount of unfolded protein. Destabilized mutants will predominantly populate a folded form, but these mutants may have an increased propensity to unfold or partially unfold which may lead to higher levels of degradation or misfolding. Mutational analyses show that even destabilizing mutations that do not lead to higher proportions of unfolded protein will reduce the total active enzyme *in vivo* [14]. SNP analysis has shown that 85% of disease causing SNPs are likely to affect enzyme stability [8, 15]. As more of the protein sequence contributes to stability than to binding directly, it seems that vitamin remedial mutations are more likely to be derived

from shifts in stability. We set out to demonstrate that disease associated mutations correlate with disrupted populations of the native state and that remediative effects can be traced to simple shifts in equilibrium.

A.3 Characterizing Vitamin Remedial Mutations: Thymidylate Synthase

Thymidylate synthase (TS) catalyzes the transfer of a methyl group from 5,10-methylene-5,6,7,8-tetrahydrofolate (CH_2H_4folate) to dUMP. It is the only source of thymine in most organisms and thus is essential for DNA synthesis and survival [16]. It is among the most highly conserved genes known; the human and *E. coli* sequences are 75% identical [16]. Due to its role in DNA synthesis, TS has also been implicated in a number of cancers, particularly Human Colorectal Cancer. A number of chemotherapeutic inhibitors have been designed for TS [11], and it has been the subject of extensive scrutiny concerning its structure, mechanism, and activity [16]. X-ray structures are available for TS from numerous organisms including human and *E. coli*. The structure, shown in Figure A-1b, is an obligate homodimer, and structures have been determined for TS with and without cofactor and substrate, as well as with analogues and single-site mutations [17-19]. A number of characterized TS mutants have characteristics that suggest it may be cofactor responsive. Two of these, T51S and Y258F, have been particularly well described and have reduced capacity to bind tetrahydrofolate and lowered stability [20, 21]. Many mutations, even those of conserved residues in the active site, do not abrogate activity [11]. TS is an excellent model system has demonstrated clinical relevance, a number of identified polymorphisms and has been used explicitly in folding studies [22]. As shown in Figure A-1a, TS folding is easy to measure experimentally. It is soluble and stable and its folding can be monitored by circular dichroism. We set out to establish both an *in vivo* and *in vitro* system to study TS. *In vitro* the protein worked well, it could be expressed and its ΔG_{UN} of folding was straightforward to measure. However, to correlate the appearance of novel mutations with *in vitro* work we were committed to generating *in vivo* data in yeast. Despite extensive effort, we could not produce a knockout Δts yeast strain for complementation experiments. Because our intent was to connect the behavior *in vivo* with data *in vitro*, biophysical characterization by itself was not worth pursuing.

A.4 Methylenetetrahydrofolate reductase

An excellent alternative prospect to TS is 5,10-methylenetetrahydrofolate reductase (MTHFR). MTHFR catalyzes the reduction of 5,10 methylenetetraydrofolate to methyltetrahydrofolate. As part of the folate pathway, its function is important in supporting the chemistry of one-carbon transfers in the cell. Methyltetrahyrdofolate is used to synthesize methionine and dysfunction in MTHFR leads to build up of the precursor homocysteine [23]. MTHFR is an excellent model system for us because it has a well known polymorphism in humans, A222V, which has been studied extensively in the clinic (see spheres in structure of Figure A-2). A222V is present in approximately

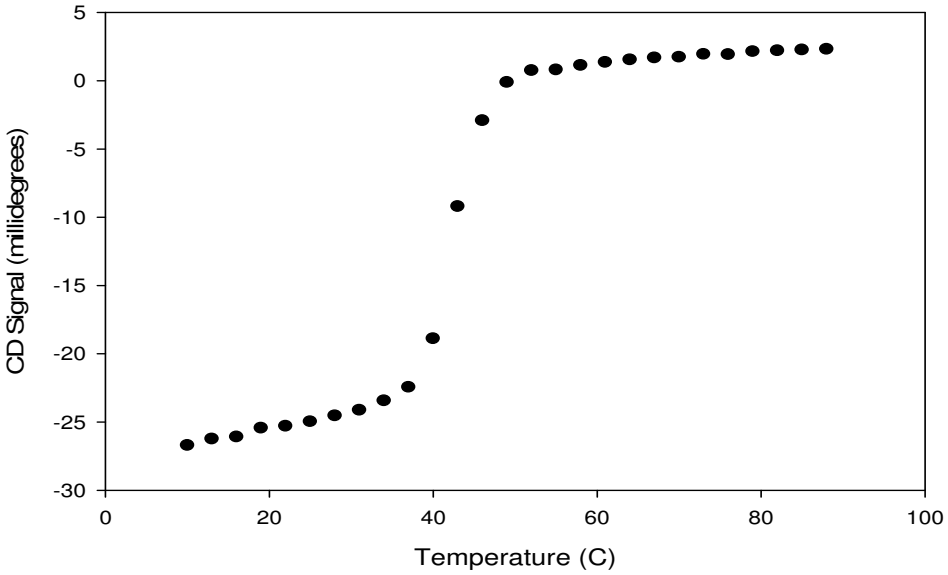

Figure A-1a Thymidylate synthase temperature melt

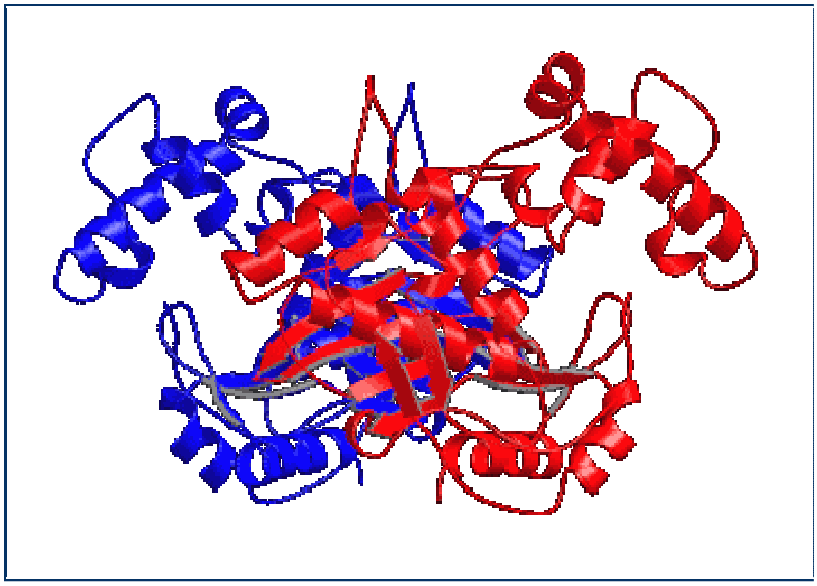

Figure A-1b The structure of thymidylate synthase

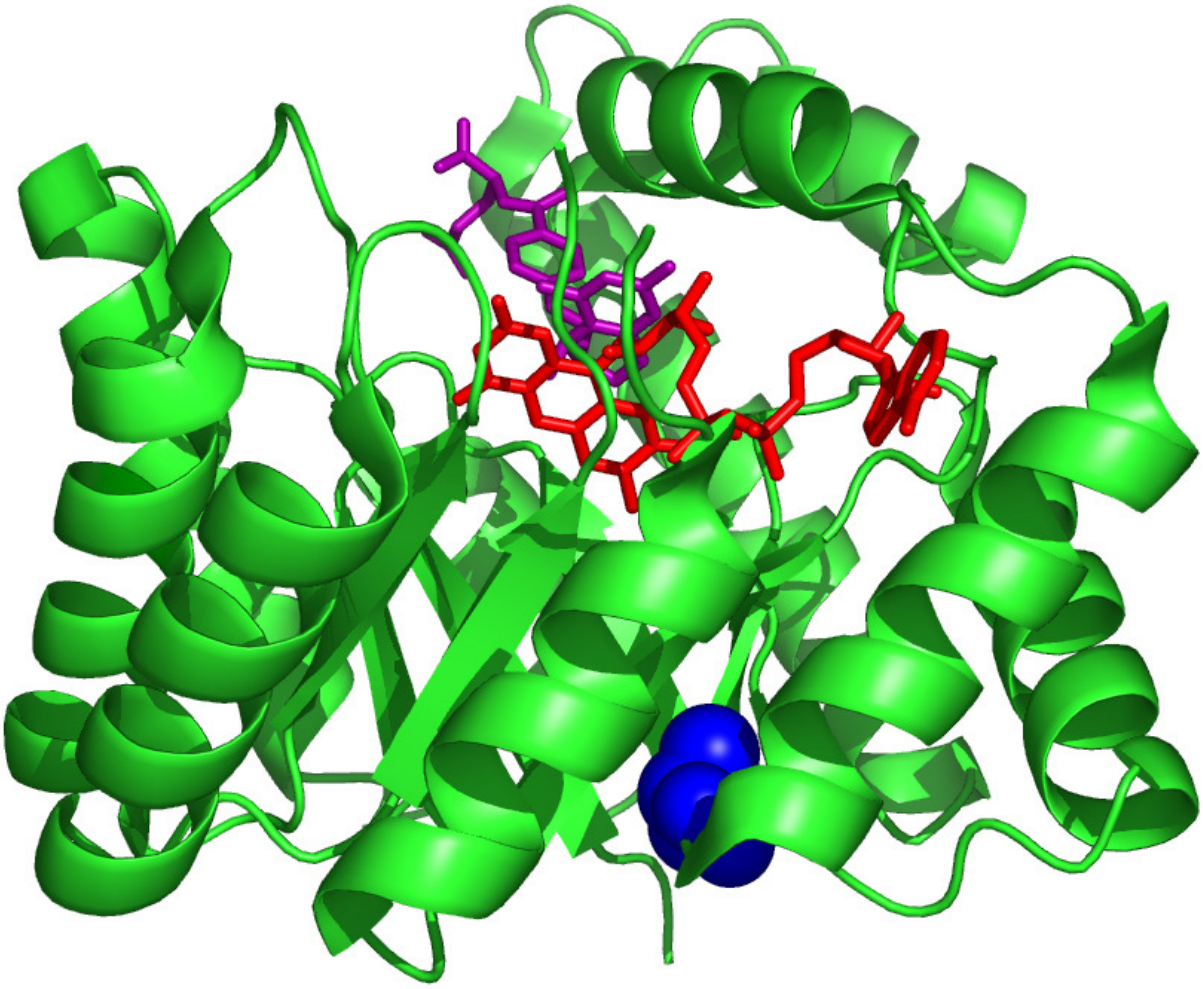

Figure A-2 The *E. coli* structure of MTHFR with ligands bound and alanine 177 (homologous to A222 in the human protein) highlighted.

30% of the population [24]. Homozygous individuals with A222V carry elevated homocysteine levels in their blood and have elevated risk for cardiovascular disease, colon cancer, and neural tube defects [25]. Elevated folate intake, however, mitigates these risks. MTHFR is thought to be destabilized by A222V to the extent that its function can be recovered at higher folate levels, but is defective when concentrations are low [26].

Work by Nick Marini in Jasper Rine's Lab not only produced a viable yeast knockout strain (Δmthfr), but was able to test the ability of expressed MTHFR mutants to complement *in vivo* behavior. As seen in Figures A-3 and A-4, a number of genetically derived mutations disrupt MTHFR at low levels of folate, but at high levels return the cells to wild-type growth rates, including A222V [24]. These mutants with differential behavior are quintessential vitamin remedial polymorphisms.

My objective was to determine the role that stability might play in these mutations. By measuring the folding stability of each mutant with and without ligand present we hoped to find evidence for stability defects in vitamin remedial mutants. Unfortunately, MTHFR was not as biophysically tractable as TS (for example see Figure A-5). MTHFR is a much larger protein and is prone to aggregation. This difficulty is exacerbated by the presence of 13 cysteines in the sequence, making it highly prone to oxidation. The stability melt of a well-folded protein is characterized by a sigmoidal cooperative transition. When a protein does not populate a well-folded state due to aggregation or misfolding, as in MTHFR, its stability cannot be measured.

We used a number of techniques to bias the population of MTHFR to the folded state including osmolytes, ligands, buffer conditions, as well as truncations of the protein. We were unable to find a working system that involved the human MTHFR protein. To get biophysical data we instead decided to use the *E. coli* version of MTHFR, which conserves many of the residues of interest (see Figure A-6a). The *E. coli* protein (shown in Figure A-2) is much more tractable in folding experiments (see Figure A-6b). While we were encouraged to be able to collect folding stability values, the prokaryotic MTHFR has a relatively low native stability of folding, only 2.4 kcal/mol in the absence of substrate (see Figure A-6b). As a result the mutations we were investigating were simply too unstable to measure in solution, and the protein again would simply aggregate during experimentation. The most prominent example is A177V, which is the analog of A222V in the human protein. It did not remain folded long enough for equilibration or measurements (Figure A-7). Without being able to compare the stability of mutations to wild type, the characterization of *E. coli* alone was not informative.

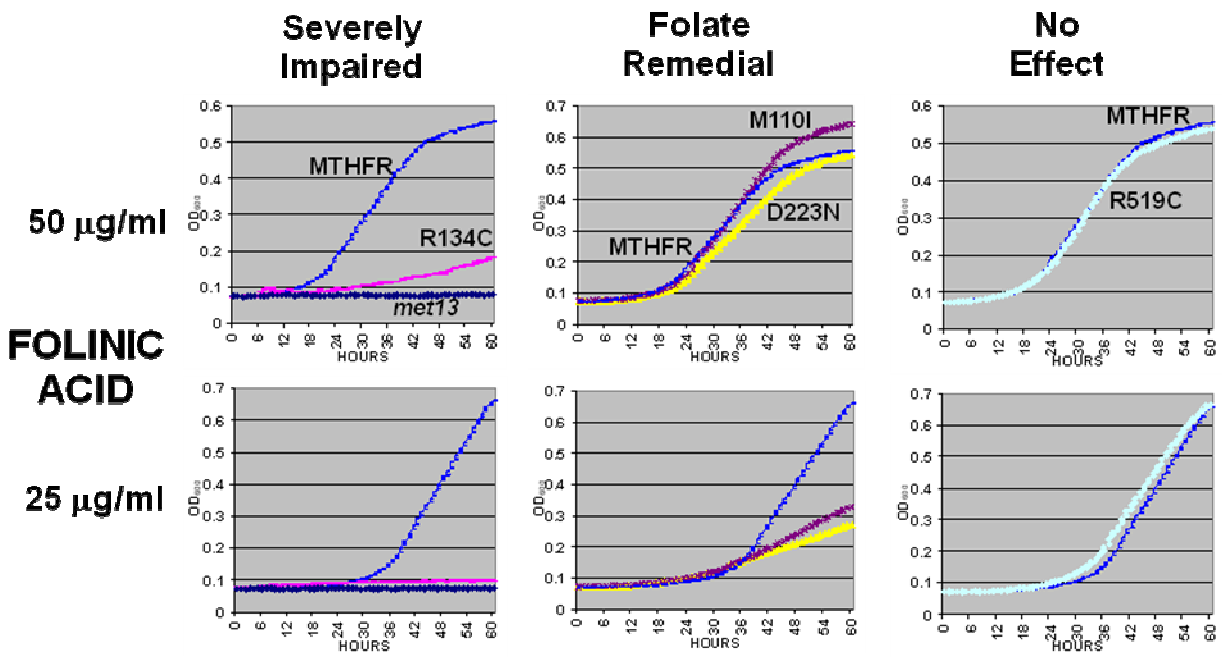

Figure A-3 Yeast based growth assays of knockout yeast lines transformed with putative MTHFR mutants and grown at varied folate levels. Wild type is shown as a dark blue trace.

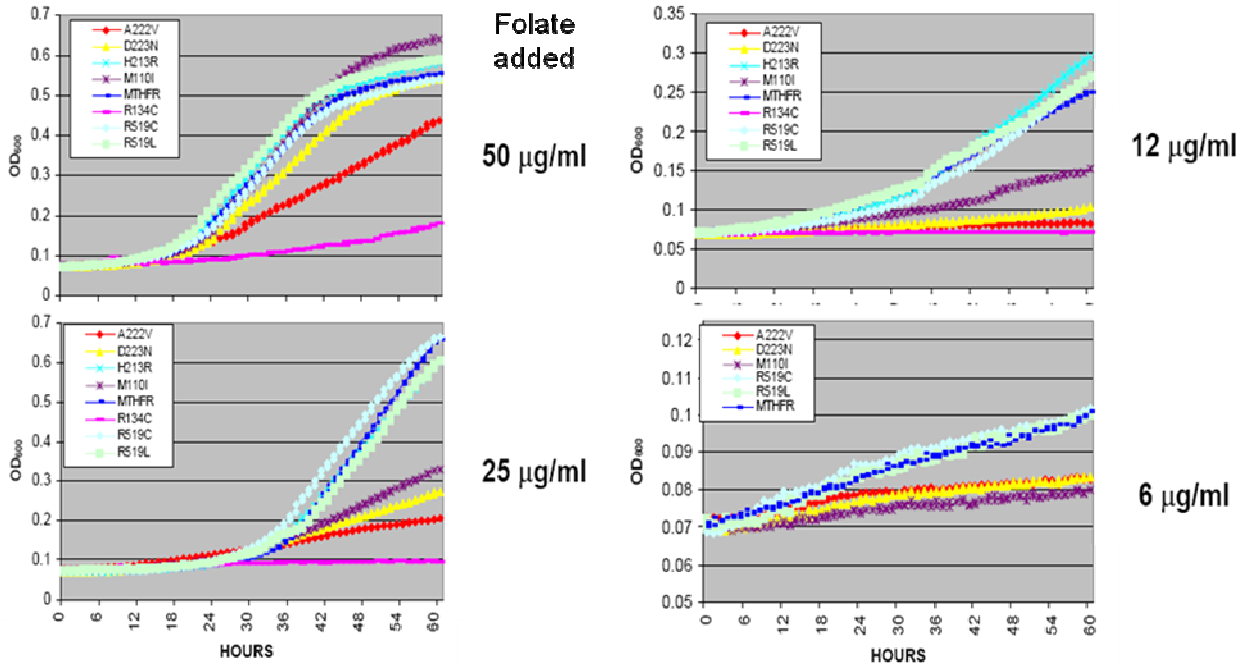

Figure A-4 Further characterization of a number of MTHFR mutants. A222V, is clearly deficient at low levels of folate but recovers at elevated concentrations

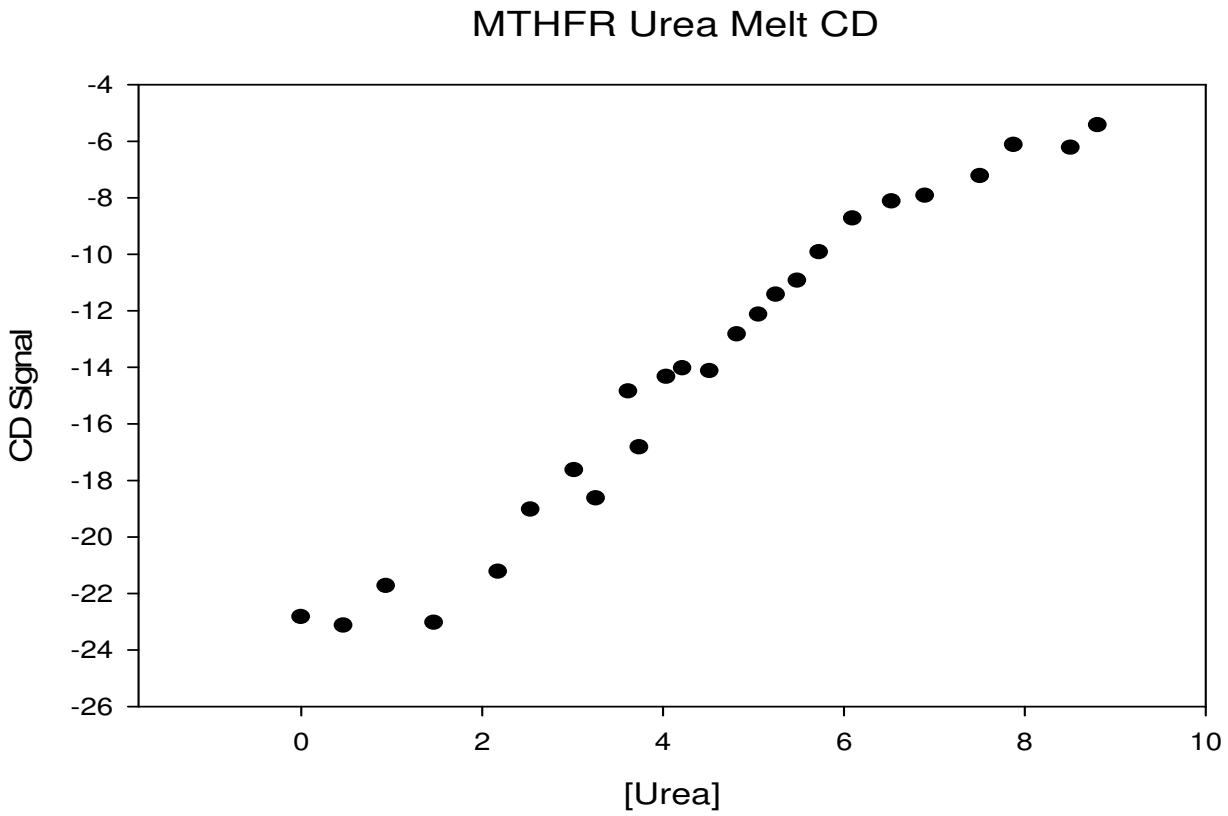

Figure A-5 A urea melt of MTHFR as monitored by CD

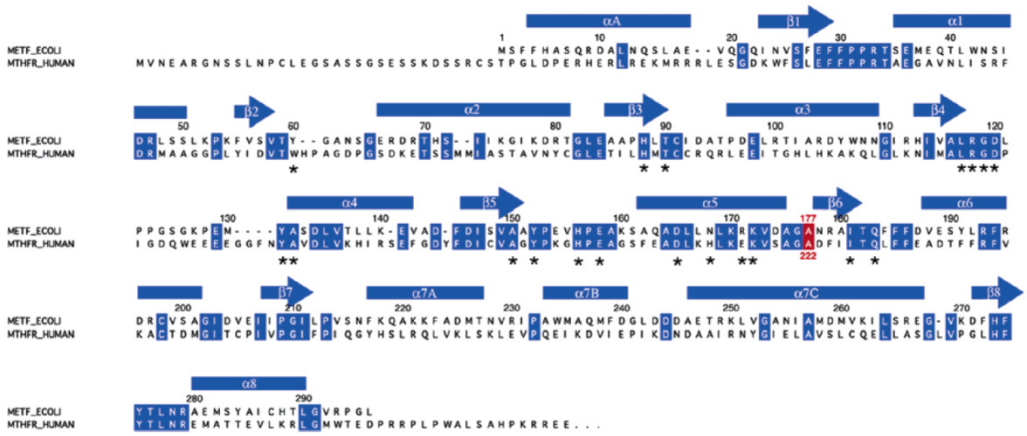

Figure A-6a An alignment of *E. coli* and *H. sapiens* MTHFR sequences, showing high similarity including A222V a primary vitamin remedial variant highlighted in red

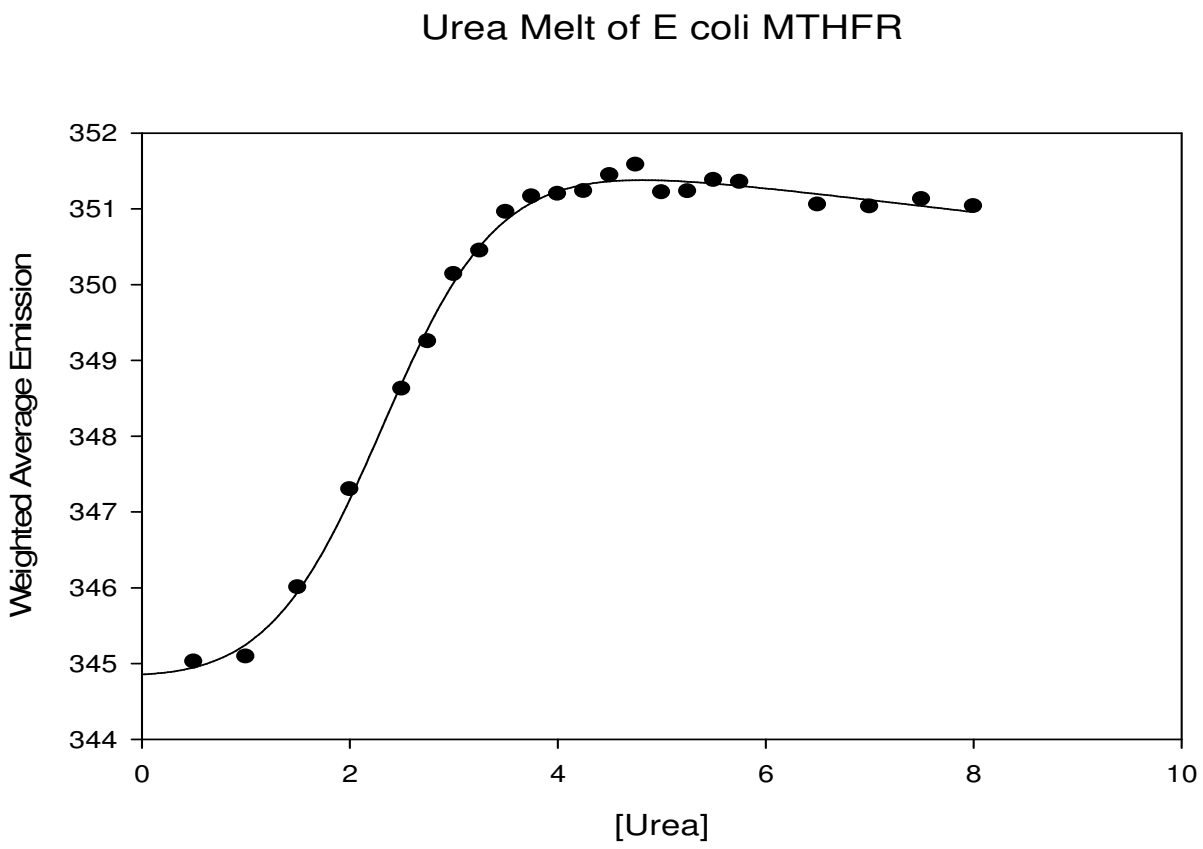

Figure A-6b A urea melt of *E. coli* MTHFR

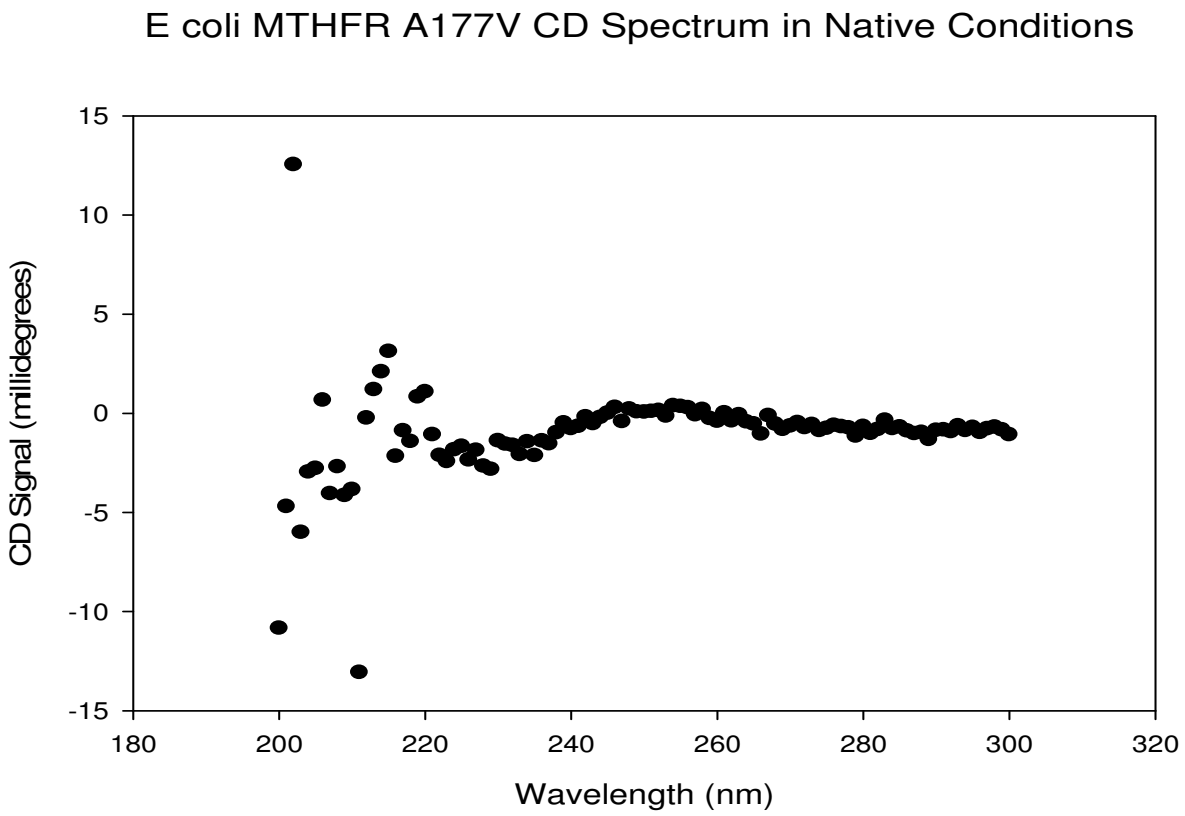

Figure A-7 A circular dichroism spectrum of the A177V variant of *E. coli* MTHFR. Protein aggregated very quickly as this variant was too destabilized to be experimentally tractable

A.5 Dihydrofolate Reductase

Dihydrofolate reductase (DHFR) is a small soluble protein that has been used in folding studies extensively for many years. DHFR catalyzes the reduction of dihydrofolate to tetrahydrofolate (see Figure A-8b). Work with DHFR in the folding field has used both the human and *E. coli* versions of the protein with great success [27, 28]. DHFR is a member of folate binding family of proteins and a good candidate for a genetic screen. We explicitly pursued DHFR because of literature-cited work with measurable stability characteristics (shown in Figure A-8a). Unfortunately, similar to thymidylate synthase, no *in vivo* data could be collected about DHFR. DHFR appears to be too essential for knockout yeast strains, even those grown in supplemented media, to survive. Without *in vivo* data available, we could not identify DHFR mutants that were vitamin remedial.

A.6 Conclusion

Controlling diet to treat disease has been used in medicine for many years. Known instances where diet affects health often hinge on genetic abnormalities. The condition phenylketonuria for example requires removing all phenylalanine from the diet. But what of milder phenotypes? Most missense mutations do not disrupt activity of an enzyme, but instead reside outside the active site. Many of these mutations will affect protein stability, and consequently, will be inactive or degraded in the cell. We hypothesized that vitamin-remedial mutations are likely to result from a disruption in the proteins ability to fold stably in the cell. As such, the binding of a ligand, cofactor, or substrate can act as a stabilizing force within the cellular environment. While this hypothesis is theoretically sound, proteins are part of a complex system *in vivo* and ideal experimental targets are not always the most experimentally facile.

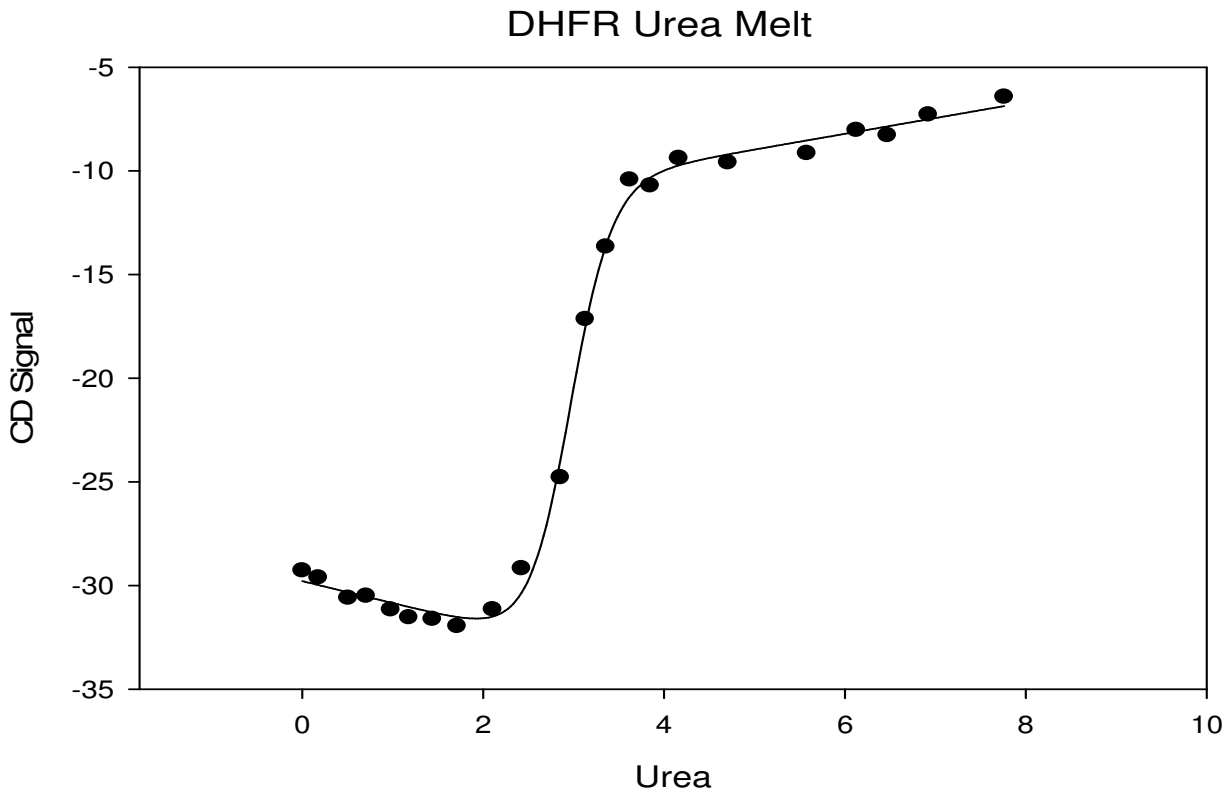

Figure A-8a A urea melt of human DHFR

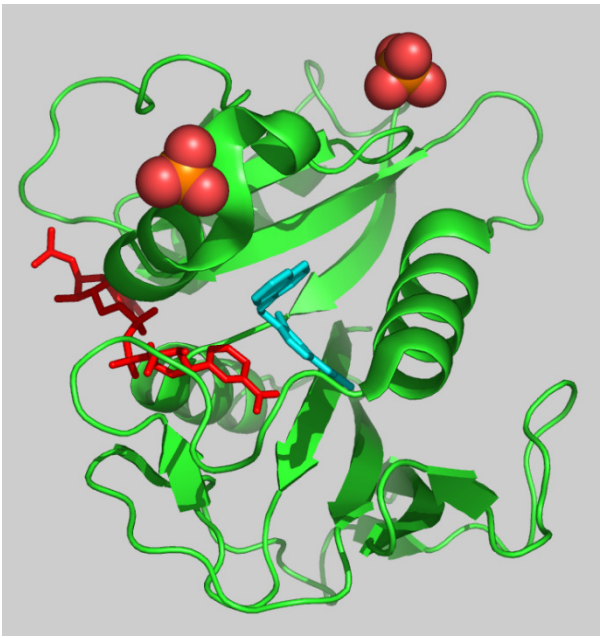

Figure A-8b The structure of DHFR with ligands bound

A.7 Citations

1. Powers, H.J., *Riboflavin (vitamin B-2) and health.* American Journal of Clinical Nutrition, 2003. **77**(6): p. 1352.
2. Allen, R.H., et al., *Diagnosis of cobalamin deficiency I: usefulness of serum methylmalonic acid and total homocysteine concentrations.* Am J Hematol, 1990. **34**(2): p. 90-98.
3. Ames, B.N., I. Elson-Schwab, and E.A. Silver, *High-dose vitamin therapy stimulates variant enzymes with decreased coenzyme binding affinity (increased Km): relevance to genetic disease and polymorphisms.* American Journal of Clinical Nutrition, 2002. **75**(4): p. 616.
4. McCarty, M.F., *High-dose pyridoxine as an 'anti-stress' strategy.* Medical hypotheses, 2000. **54**(5): p. 803-807.
5. Duran, M. and S.K. Wadman, *Thiamine-responsive inborn errors of metabolism.* Journal of inherited metabolic disease, 1985. **8**: p. 70-75.
6. Wang, T., et al., *Correction of ornithine accumulation prevents retinal degeneration in a mouse model of gyrate atrophy of the choroid and retina.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(3): p. 1224.
7. Pauling, L., *Orthomolecular psychiatry: varying the concentrations of substances normally present in the human body may control mental disease.* Journal of Nutritional and Environmental Medicine, 1995. **5**(2): p. 187-198.
8. Wang, D.G., et al., *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.* Science, 1998. **280**(5366): p. 1077.
9. Michaud, J., et al., *Pyridoxine-responsive gyrate atrophy of the choroid and retina: clinical and biochemical correlates of the mutation A226V.* American journal of human genetics, 1995. **56**(3): p. 616.
10. Guenther, B.D., et al., *The structure and properties of methylenetetrahydrofolate reductase from Escherichia coli suggest how folate ameliorates human hyperhomocysteinemia.* Nature structural & molecular biology, 1999. **6**(4): p. 359-365.
11. Kawate, H., D.M. Landis, and L.A. Loeb, *Distribution of mutations in human thymidylate synthase yielding resistance to 5-fluorodeoxyuridine.* Journal of Biological Chemistry, 2002. **277**(39): p. 36304.
12. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(96): p. 223-30.
13. Schellman, J.A., *Macromolecular binding.* Peptide Science. **14**(5): p. 999-1018.
14. Pakula, A.A. and R.T. Sauer, *Genetic analysis of protein stability and function.* Annual review of genetics, 1989. **23**(1): p. 289-310.
15. Cargill, M., et al., *Characterization of single-nucleotide polymorphisms in coding regions of human genes.* Nature genetics, 1999. **22**(3): p. 231-238.
16. Carreras, C.W. and D.V. Santi, *The catalytic mechanism and structure of thymidylate synthase.* Annual review of biochemistry, 1995. **64**(1): p. 721-762.
17. Finer-Moore, J., et al., *Refined structures of substrate-bound and phosphate-bound thymidylate synthase from Lactobacillus casei.* Journal of molecular biology, 1993. **232**: p. 1101-1101.
18. Forsthoefel, A.M., et al., *Structural Determinants for the Intracellular Degradation of Human Thymidylate Synthase†.* Biochemistry, 2004. **43**(7): p. 1972-1979.

19. Berger, S.H., F.G. Berger, and L. Lebioda, *Effects of ligand binding and conformational switching on intracellular stability of human thymidylate synthase.* BBA-Proteins and Proteomics, 2004. **1696**(1): p. 15-22.
20. Zhang, J., et al., *Association of the thymidylate synthase polymorphisms with esophageal squamous cell carcinoma and gastric cardiac adenocarcinoma.* Carcinogenesis, 2004. **25**(12): p. 2479.
21. Pedersen-Lane, J., et al., *High-level expression of human thymidylate synthase.* Protein expression and purification, 1997. **10**(2): p. 256-262.
22. Perry, K.M., et al., *Reversible dissociation and unfolding of the dimeric protein thymidylate synthase.* Protein Sci, 1992. **1**(6): p. 796-800.
23. Weisberg, I., et al., *A second genetic polymorphism in methylenetetrahydrofolate reductase (MTHFR) associated with decreased enzyme activity.* Molecular genetics and metabolism, 1998. **64**(3): p. 169-172.
24. Marini, N.J., et al., *The prevalence of folate-remedial MTHFR enzyme variants in humans.* Proceedings of the National Academy of Sciences, 2008. **105**(23): p. 8055.
25. Ueland, P.M., et al., *Biological and clinical implications of the MTHFR C677T polymorphism.* Trends in pharmacological sciences, 2001. **22**(4): p. 195-201.
26. Mtiraoui, N., et al., *MTHFR C677T and A1298C gene polymorphisms and hyperhomocysteinemia as risk factors of diabetic nephropathy in type 2 diabetes patients.* Diabetes research and clinical practice, 2007. **75**(1): p. 99-106.
27. Touchette, N.A., K.M. Perry, and C.R. Matthews, *Folding of dihydrofolate reductase from Escherichia coli.* Biochemistry, 1986. **25**(19): p. 5445-5452.
28. Horst, R., et al., *Folding trajectories of human dihydrofolate reductase inside the GroEL–GroES chaperonin cavity and free in solution.* Proceedings of the National Academy of Sciences, 2007. **104**(52): p. 20788.