Title
Detailed characterization of the earliest events in protein folding

Permalink
https://escholarship.org/uc/item/5rm0r75g

Author
Rosen, Laura

Publication Date
2014

Peer reviewed|Thesis/dissertation
Detailed characterization of the earliest events in protein folding

By

Laura Elizabeth Rosen

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

Doctor of Philosophy

in

Biophysics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Susan Marqusee, chair
Professor David Wemmer
Professor Andreas Martin
Professor Danielle Tullman-Ercek

Spring 2014
Abstract

Detailed characterization of the earliest events in protein folding

by

Laura Elizabeth Rosen

Doctor of Philosophy in Biophysics

University of California, Berkeley

Professor Susan Marqusee, Chair

A major goal in biology is to understand how the amino acid sequence encodes all aspects of a protein’s structure and dynamics. A protein spends most of its time in its fully folded native state but it also transiently populates high-energy, partially folded states. These high-energy states can be important for function and regulation, in addition to their importance to protein folding (the process by which a protein folds to its native state from the fully unfolded state). Together, all of these conformational states and the dynamics within and between them constitute a protein’s energy landscape. Only by fully characterizing protein energy landscapes of model systems can we make progress on understanding all of the information encoded in an amino acid sequence. The present work makes a significant contribution to this effort by providing a detailed characterization of the high-energy, partially folded states that are most difficult to access – those populated earliest in protein folding – for an important protein folding model system, ribonuclease H (RNase H).

The earliest events in the folding of *E. coli* RNase H were studied using two complementary approaches: a pulse-labeling hydrogen exchange mass spectrometry (HX MS) method, and ultrarapid mixing coupled with sensitive detection of tryptophan fluorescence. The HX MS method monitors the formation of secondary structure in the early folding process and tryptophan fluorescence monitors the formation of tertiary structure. By HX MS, we observe that RNase H folds through a single dominant pathway by the sequential acquisition of cooperative units of native secondary structure. This work identifies three intermediates populated prior to the rate-limiting barrier to folding (the last of which is the previously characterized I_{core} intermediate). The fluorescence experiments provide detailed information about the barrier heights in the early folding steps and also demonstrate that significant non-native structure is formed early in the process, likely productively stabilizing the earliest folding intermediate. Based on both studies, we present a detailed model for the early folding pathway of RNase H, on a timescale and size-scale amenable to comparison with atomistic folding simulations.

The RNase H intermediate that transiently accumulates prior to the rate-limiting barrier to folding (I_{core}) was further investigated by manipulating the energy landscape to make I_{core} the ground-state conformation at equilibrium. Mimics of I_{core} were made using two approaches: a
truncation to generate a fragment and selective destabilization of the native state using point mutations. Characteristics of the $I_{\text{core}}$ mimics are consistent with the transient kinetic intermediate, indicating that they are good models for $I_{\text{core}}$. The structural properties of the mimics indicate that $I_{\text{core}}$ likely has closely packed side chains in its structured region, in contrast to a long-standing hypothesis that it has molten tertiary packing. This result is important for understanding the nature of the subsequent rate-limiting barrier to folding. Additionally, we performed similar studies with an RNase H homolog from *T. thermophilus*, yielding similar results. The homolog comparison also reveals that *T. thermophilus* RNase H populates a partially folded intermediate ($I_{\text{core+1}}$) that is not populated by *E. coli* RNase H. This difference sheds light on the interplay between native topology and sequence in defining the features of a protein energy landscape. Finally, preliminary kinetic experiments performed with *T. thermophilus* RNase H directly probe the nature of the rate-limiting barrier.
# Table of Contents

for

**Detailed characterization of the earliest events in protein folding**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>1</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>i</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
</tbody>
</table>

## 1 Introduction

1.1 Goals and challenges in the field of protein folding | 1
1.2 Experimental methods for characterizing energy landscapes | 3
  1.2.1 Detection of very rare states at equilibrium | 3
  1.2.2 Using relaxation kinetics to populate rare states | 4
  1.2.3 Equilibrium mimics of high-energy states | 8
1.3 The RNase H model system | 9
  1.3.1 The RNase H folding pathway | 9
  1.3.2 The nature of side chain packing in the I_{core} intermediate | 10
  1.3.3 Comparison between RNase H sequences | 11
1.4 Summary of thesis work | 12
1.5 References | 13

## 2 The stepwise folding of *E. coli* RNase H observed at near amino acid resolution using hydrogen exchange and mass spectrometry

2.1 Introduction | 16
2.2 Results | 17
  2.2.1 Folding by spectrophotometry | 17
  2.2.2 HX MS experiment | 17
  2.2.3 Data interpretation | 18
  2.2.4 MS results | 19
  2.2.5 Partially folded structure at high resolution | 21
  2.2.6 Anomalously slow exchange for β-edge amides | 23
2.3 Discussion | 24
  2.3.1 Experimental work to the contrary | 25
  2.3.2 Theoretical studies | 26
  2.3.3 Principles of protein folding | 26
2.4 Materials and Methods | 27
2.5 Acknowledgments | 27
2.6 References | 28

## 3 Direct observation of the earliest events in the folding of *E. coli* RNase H using ultrarapid mixing and intrinsic fluorescence

3.1 Introduction | 32
3.2 Results | 33
5.2.3 NMR suggests the interior of the ttRNH I_{core} fragment is closely packed 67
5.2.4 EcRNH does not populate the I_{core+1} intermediate 69
5.2.5 Mutational analysis to locate ttRNH I_{core+1} on the folding pathway 69
5.2.6 Strand I docking on the alpha helical core monitored directly by FRET 72

5.3 Discussion 74
5.4 Materials and Methods 75
5.5 Acknowledgments 76
5.6 References 76

A Appendix: Preliminary characterization of the earliest events in the folding of T. thermophilus RNase H and the ancestral RNase H Anc1 using ultrarapid mixing and intrinsic fluorescence
A.1 Introduction 79
A.2 Preliminary results 80
A.3 Materials and Methods 82
A.4 References 82
Acknowledgments

I owe thanks to many people for my successful navigation of these last six years. My advisor Susan Marqusee has provided unflagging support and encouragement. I have learned so much from my time in her lab and have benefited so much from her constant advocacy. All of the members of the Marqusee Lab have created a wonderful work environment, both scientifically and socially. In particular, an enormous thank you to Sabriya Rosemond and Diana Koulechova: I can’t imagine what this process would have been like without them and I am so grateful for their friendship. Former members of the Marqusee Lab have also had a big impact on my time in graduate school. In particular, Katelyn Connell was a wonderful collaborator (not to mention cycling partner). And Rachel Bernstein has been an excellent companion during the process of writing this dissertation (as well as a great climbing partner).

Many friends outside the lab have been invaluable sources of support – too many to name them all, but a few deserve special thanks. My former roommates Katie Harding and Lowry Kirkby have always been there with advice and encouragement (and also wine). Anna Schneider and John Chodera have reassured and inspired me, and have been delightful companions whether at lunch in Berkeley or during overseas travels. And my friendship with Samantha Rudolph has been an incredible source of understanding, compassion, and laughter, extending back to our college days.

I would have never made it to UC Berkeley without the support and guidance of numerous people. In particular, thank you to my undergraduate advisor, Len Seligman, for introducing me to science research and setting me up so well to succeed in grad school. And, of course, thank you to my parents – Carolyn Hayek and Steven Rosen – for always cheering me on and for providing such a supportive foundation off which to launch.

Finally, a huge thank you to Kyle Vinson for his constant support, patience, and good humor. I am so lucky to have him as my partner – he is the best discovery I made in graduate school.
Chapter 1
Introduction to
Detailed characterization of the earliest events in protein folding

1.1 Goals and challenges in the field of protein folding

Proteins are the primary movers and shakers in the cell. They catalyze chemical reactions from metabolism to DNA synthesis, transduce signals within cells and across cell membranes, and play essential organizational and structural roles. Most critical to advances in human health, proteins are our primary drug targets, and may also be engineered to provide new medical therapies and new technologies. Understanding proteins in order to realize these advances requires knowledge of their structure and dynamics. Such knowledge is slowly acquired when studying individual proteins one by one. What we really want is a deep understanding of how information is encoded by a protein’s amino acid sequence, so that sequence information – which is so abundant – can be translated into functional insight.

Since Anfinsen’s seminal experiments with RNase A in 1961, it has been understood that the amino acid sequence contains all the information needed for a protein to fold to its ground state, or “native” structure. (Even for those proteins that require chaperones to fold, the blueprint for folding is encoded in the protein’s sequence.) For a long time, the protein folding field was defined by the problem of how sequence encodes native structure. Indeed, huge advances have been made in our ability to predict structure from sequence, and some consider this a solved problem. However, it is increasingly clear that a static picture of native structure does not begin to capture all the aspects of a protein’s behavior that are important for function, and that amino acid sequence encodes much more than just the native state.

A more modern view envisions the amino acid sequence as encoding an entire protein “energy landscape.” An energy landscape refers to all possible conformations of the polypeptide chain, their relative free energies, and how the conformations are kinetically connected. Key features of the energy landscape (as depicted in Figure 1.1) are the global minimum (i.e. the native state), high-energy, local minima (i.e. partially folded intermediates) and barriers that regulate dynamics between states. Conformational dynamics within each state are also included in the landscape view, and can be thought of as the breadth of the wells. A protein can sample all of these different states according to Boltzmann statistics, so while it will spend the majority of its time in the native state, it will periodically sample the other high-energy states. These high-energy states may be important points of regulation or may be directly important for activity. Therefore, understanding the energy landscape is important for truly understanding protein function.

The energy landscape view gives us a framework for interpreting all aspects of protein behavior and structural properties. For example, it has become increasingly clear that some proteins do not have a global free energy minimum – they do not populate a native state – and are better thought of as an ensemble of many different interconverting conformations. (These “intrinsically
disordered proteins” – or “IDPs” – are thought to make up ~25% of the mammalian proteome, which increases to ~50% when proteins containing intrinsically disordered domains are considered. For proteins that do fold to a native state, the landscape view helps us to understand how two proteins that have the same native-state structure can have very different functional properties, such as different conformational dynamics and thermal stabilities. The landscape view also provides important context for understanding mutations. Typically, small changes in sequence do not affect the native state structure, but sequence-change effects may be large on partially folded intermediate states, which could dramatically affect function.

Lastly, the energy landscape view provides a framework to understand that fundamental question (first posed by Levinthal in 1968) of how a protein finds its native state, and so quickly, among all possible conformations. Protein folding can be envisioned as diffusion on the energy landscape from the high-energy, unfolded state conformational space to the native state energy minimum. Partially folded states of successively lower energy may guide the protein to its final folded structure. So in addition to any direct role in function, partially folded states are important for shepherding a protein to its native state.

Overall, the major goal of the protein folding field is to understand how the amino acid sequence encodes the entire energy landscape. This is an extremely complex undertaking: an amino acid sequence is a diverse polypeptide chain with many degrees of freedom, and it is a very subtle balance of free energies that dictates all of the features of the landscape. It may be that computational predictions are our best hope for reliably translating sequence information into behavioral and functional insight on a large scale. However, right now detailed simulations can only handle very small proteins and fast timescales. And simulations are only as good as their experimental validation. So what we currently need are detailed experimental characterizations of the energy landscapes of protein model systems, to provide general principles as well as data for comparison to simulation.

Figure 1.1 Visual representation of an energy landscape. In this depiction, the horizontal plane represents all possible conformations of the polypeptide chain and the vertical axis may be thought of as the free energy associated with each conformation. Notable features are the global minimum (fully-folded native state), local minima depicting partially folded states, and high-energy barriers. Two possible folding pathways are depicted with yellow arrows, showing how a protein may reach the native state basin from the large unfolded conformational space. This image was generated by the Dill Group at Stony Brook University.
To characterize a protein energy landscape experimentally means characterizing the high-energy, partially folded states on that landscape (as well as how proteins reach one state from another). The specific questions are: What partially folded intermediates are present on a protein energy landscape and what is the nature of their structure? For transitions between states, especially folding to the native state, is there one dominant folding pathway, or many parallel pathways? What is the nature of the barriers between states?

Studying high-energy states is very difficult because they are rarely populated and transient. This is especially true for the highest-energy states that are close to the unfolded state and are the first intermediates populated when a protein folds: the very high-energy states are the least well-characterized aspects of protein energy landscapes. Therefore, the work in this thesis makes a significant contribution to the protein folding field by providing detailed characterization of the early folding pathway of an important protein folding model system.

1.2 Experimental methods for characterizing energy landscapes

There are many types of experimental methods used to characterize the high-energy, partially folded states on a protein energy landscape. Here, I give a brief treatment of these approaches.

1.2.1 Detection of very rare states at equilibrium

High-energy, partially folded states can be directly detected at equilibrium using techniques that are sensitive to rare structural fluctuations. Typically, such a technique relies on a chemical probe that will label a region of the protein if it becomes unfolded in a fluctuation away from the native state. The most common such technique is hydrogen-deuterium (HD) exchange. Hydrogens that are bonded to nitrogen will exchange with solvent on a timescale of milliseconds to minutes (depending on the pH) when freely exposed to solvent, but when that hydrogen is participating in structure, such as formation of a hydrogen bond, this exchange can slow to the order of days or weeks. Therefore, isotope exchange can be used to label any region of the protein that has become transiently unfolded, marking the presence of a high-energy partially folded intermediate. Most of the amine protons in a protein are in the backbone, so this technique primarily monitors protein secondary structure. This method is discussed in further detail in section 1.2.2.

A technique orthogonal to the labeling of unfolded secondary structure by isotope exchange is the labeling of unfolded side chains by covalent modification. This is most easily done using a series of single-cysteine variants of a protein. A label can be used to selectively modify a thiol as it becomes transiently exposed to solvent under equilibrium conditions. Monitoring side chain exposure permits observation of partially folded states with disrupted tertiary structure, even if they have native-like secondary structure. With both HD exchange and thiol labeling, sophisticated kinetic models are needed to interpret the data, but they have both been demonstrated as very powerful techniques capable of revealing detailed insight about partially folded structures on the energy landscape.
Lastly, single molecule detection can also be used to observe rare fluctuations to high-energy states. Examples include single molecule FRET\textsuperscript{9} or force spectroscopy using optical tweezers operated in equilibrium mode.\textsuperscript{10}

1.2.2 Using relaxation kinetics to populate rare states

Another powerful way to learn about partially folded states is to populate them transiently using relaxation kinetics. Typically in such an experiment, the whole population of protein molecules is biased to the unfolded state and then quickly switched to conditions promoting folding to the native state. The population is monitored as it relaxes to the ground state, and observations are made of whether the population accumulates in any high-energy, partially folded states as it folds. This experiment provides observation of partially folded states on the energy landscape, but crucially also monitors the folding pathway – how a protein reaches the native state from the unfolded state, which is a particularly important feature of the energy landscape.

\textit{Initiating relaxation kinetics}

The most universal way to initiate refolding is by dilution of chemical denaturant. Typically, protein in denaturant is diluted into refolding buffer in a cuvette and folding is monitored spectroscopically. For many proteins, a stopped-flow mixing apparatus is necessary in order to mix fast enough to observe folding. In stopped-flow mixing, two solutions are mixed very fast into a cuvette using precision syringes (Figure 1.2A). The fastest mixing time achievable with this experiment is about one millisecond, but in practice the mixing time is closer to 5 or 10 milliseconds (because of the large change in solvent viscosity, and the size of the cuvette required for sufficient signal).

Some proteins fold entirely to their native state within milliseconds. And even the slower proteins that take seconds to completely fold often have early folding events that occur within milliseconds. Therefore, methods to study fast folding events have been desired. Early studies of really fast folding required using techniques such as temperature jumps, pressure jumps or photo-excitation in order to initiate the reaction quickly. However, only very specific model proteins are amenable to study by these methods. A more universal approach is to continue to use dilution of chemical denaturant to initiate folding, but to try to perform this mixing much faster than possible in the stopped flow.\textsuperscript{11}

Ultrarapid mixing can be achieved by using continuous-flow instead of stopped-flow mixing. In this technique, very small volumes of denatured protein and refolding buffer are mixed continuously using a steady flow of the two solutions, where the solution flows along a channel after the point of mixing (Figure 1.2B). The small volumes allow for very rapid mixing (the fastest theoretical mixing time is 10 microseconds). But despite the small volume, a detectable signal can be obtained by measuring along the channel. The detector can sit at one position in the channel and build up enough signal-to-noise, then slowly scan along the channel to build up the entire kinetic trace. The position in the channel is converted to folding time based on the cross-sectional area of the channel and the flow speed. This technique is the cornerstone of Chapter 3.
**Figure 1.2** Initiating protein refolding by dilution of chemical denaturant. A. In stopped-flow mixing, precision syringes rapidly push the solutions through the mixer and into the cuvette. A stop valve at the far end of the cuvette engages concurrent with syringes stopping, whereupon signal evolution is monitored in the cuvette. The theoretical dead time of this experiment is one millisecond, but in protein folding is typically 5-10 milliseconds. B. Continuous-flow mixing is performed by flowing the solutions through a microfluidic channel (diameter on the order of 100 microns). Turbulent mixing is produced by the large momentum change where the solutions meet. The solutions flow continuously during detection. Signal is measured along the channel by slowly moving the detector relative to the channel. Channel distance is converted to folding time using the flow speed and cross-sectional area of the channel. The theoretical dead time is 10 microseconds.

**Spectroscopic monitoring of folding**

There are a few options for monitoring protein folding spectroscopically. Circular dichroism (CD) is a good tool for observing folding because it reports on the formation of secondary structure. CD also has the benefit that it is monotonically correlated with increasing amounts of structure. Tryptophan fluorescence is another intrinsic property of a protein that may be monitored in order to follow folding. Tryptophans can be used as probes of folding because their fluorescent properties are sensitive to their environment: namely, what chemical moieties nearby (or diffusing in the solvent) are capable of quenching fluorescence. Typically, the largest quenching occurs when the tryptophan is solvent-exposed, so increased fluorescence signal is often associated with more folding (though this is not always the case, for example if there is significant quenching by neighboring residues in the folded structure). Tryptophan fluorescence also has the benefit that it is probing tertiary structure. This is used extensively in Chapter 3.

Other techniques using fluorescence to monitor folding require labeling the protein. Förster resonance energy transfer (FRET) can be used by labeling a protein at two positions with a FRET dye pair and observing the change in fluorescence of the donor or acceptor dye, which is related to the distance between the two dye molecules. This can be a powerful tool, but requires site-specific labeling which adds a level of difficulty and also adds a caveat, since it is no longer exactly the wild-type protein being monitored. A version of FRET that is slightly easier uses a single dye to quench intrinsic tryptophan fluorescence. This method works if the tryptophans in the protein
are sufficiently confined to one region, and it is easier because it requires only one labeling site. This technique is used in Chapter 5.

**Interpreting kinetic data**

In all kinetic experiments, the observed signal as a function of time is fit with the minimum number of exponential terms (i.e. Figure 1.3C). Frequently in protein folding this is a single exponential but if additional exponential terms are required it may indicate the formation of an intermediate (as observed in Chapter 3). The observed rate constants extracted from the fit tell you the timescales of the observed processes. (To obtain a microscopic rate constant for the folding reaction requires fitting the data to a model, as in Figure 1.3A,B.) It may also happen that an intermediate can be inferred if different folding rates are observed using different probes.

An intermediate may also be inferred from the amplitude of the observed signal. If the observed signal extrapolated to zero-time is very different from the expected starting signal, it may indicate a fast kinetic process has occurred within the mixing time of the experiment (i.e. Figure 1.3C). With sufficient separation of timescales, the fast kinetic process may be treated as a pre-equilibrium and the extrapolated observed amplitude can map out equilibrium properties of the early folding intermediate (i.e. Figure 1.3D).

Finally, performing kinetic experiments at different final concentrations of chemical denaturant can be very helpful for distinguishing kinetic models. For a two-state folding mechanism (meaning no intermediates are significantly populated) the logarithm of the observed folding and unfolding rate constants vary linearly with denaturant (Figure 1.3A). A deviation from linearity in such a plot is another signature of an intermediate (Figure 1.3B).

**Hydrogen-deuterium exchange used to probe folding intermediates**

Structural information about partially folded states that are populated along a folding pathway can be obtained by selectively labeling the unfolded region at different points in time after folding is initiated. (This is very similar to the techniques discussed in section 1.2.1 for detecting rare states at equilibrium.) The most common way to perform such an experiment is to use a “pulse” of hydrogen-deuterium exchange to label unstructured amides. Quench flow mixing can be used to: 1. Initiate refolding, 2. Initiate hydrogen exchange using high pH, and 3. Quench the reaction. The quench is important so that isotope exchange occurs only during a brief window of time, capturing the protein’s conformation at that point in folding. This experiment provides a residue-level picture of secondary structure formation along the folding pathway.

Traditionally, the amount of isotope present at each amide position is determined by NMR spectroscopy with the native protein (which works because deuterium does not have a signal in traditional NMR experiments). However, there are a few drawbacks with using NMR detection. It is a very time-consuming process to assign peaks in an NMR spectrum to each specific amide in the protein, which must be laboriously repeated with each new protein studied. Additionally, not all amides in the protein can be probes of structure, because only those that are well protected from exchange in the native state (while the NMR spectrum is being measured) can be probes. Lastly, NMR only reports on a population average, and correlated exchange at different residues cannot be readily detected.
Figure 1.3 Analysis of kinetic protein folding data. A. Chevron plot for a two-state folder. The plotted rate constants are obtained by single-exponential fits of kinetic data. The logarithm of each microscopic rate constant varies linearly with denaturant. B. Chevron plot for a three-state folder. Characteristic rollover in the folding limb indicates the presence of an early intermediate. The on-pathway kinetic model is shown, where \( f \) is the fraction of I out of total U and I. C. Observed folding kinetics for a three-state folder. Observed signal is plotted versus refolding time and fit to a single-exponential decay. The fit extrapolated to zero time does not match the expected signal, indicating a fast kinetic phase occurred in the dead time of the experiment, i.e. a burst phase. D. Amplitudes for a three-state folder. Burst phase amplitudes are plotted as a function of final denaturant concentration (filled symbols), together with final kinetic amplitudes (open squares and circles) and equilibrium amplitudes (open triangles). Squares represent data measured using stopped-flow mixing; circles represent data measured using manual mixing. The burst phase and final/equilibrium amplitudes map out the denaturation-induced equilibrium unfolding of the intermediate and the native state, respectively. Data can be fit using a two-state, linear extrapolation model to obtain the free energy of unfolding (\( \Delta G_{\text{unf}} \)) for each state.\(^{13}\) (Panel A shows data from the D10A/I53D variant of RNase H, figure adapted from Connell et al.\(^{13}\) Panels B-D show wild-type RNase H data, figures adapted from Raschke et al.\(^{14}\)
Another option for analysis uses mass spectrometry to identify the location of each isotope. This is performed by proteolyzing the protein into many different peptides and analyzing the individual peptides. This technique has the benefits that essentially every amide in the protein can be used as a probe, it is not population averaged, and it is relatively high-throughput to determine which peptide corresponds to which region of the protein. In theory, the results of the overlapping peptides can be compared to get residue-level resolution, but in the past the methods for generating peptides with stable populations of isotopes have not been robust enough to achieve good resolution. However, new methodological advances have achieved this resolution and made mass spec detection a very powerful tool, as shown in Chapter 2.

**Mutagenesis**
Mutagenesis can be a powerful way to characterize energy landscapes when used in conjunction with techniques for monitoring folding. Typically, a single side chain is “removed” by making a mutation to alanine. Comparing the effect of the mutation on the folding kinetics versus the native state stability can provide insight into whether that side chain is structured in a folding intermediate, in the transition state of a folding step, or not appreciably structured until formation of the native state. The quantification of these effects is known as ϕ-value analysis, which is demonstrated in Chapter 5.

1.2.3 **Equilibrium mimics of high-energy states**
Another way to study partially folded intermediates is to bias the energy landscape so that a high-energy state becomes the ground state and therefore is the dominant population at equilibrium. This can be done using mutagenesis, or by changing solution conditions.

One of the earliest examples of an equilibrium partially folded state is the molten globule. This is a type of partially folded intermediate with features of both well-folded and unfolded states. It is defined as a compact protein conformation with well-formed secondary structure but only loosely defined tertiary contacts. A hallmark is usually the binding and fluorescence of the dye ANS. Many proteins are known to adopt molten globule states under extreme solution conditions such as moderate denaturant concentration (e.g. α-lactalbumin) and low pH (e.g. apomyoglobin and RNase H). It is thought that these equilibrium molten globules may mimic high-energy molten globule conformations on the energy landscape under native conditions.

Another way to populate a partially folded state is to use mutagenesis. If there is prior knowledge about which region of the protein is unfolded in a given intermediate, that region can be prevented from participating in structure by removing it (creating a fragment) or putting mutations in that region. The resulting construct mimics a high-energy intermediate, but can be studied using equilibrium tools. This technique is employed extensively in Chapters 4 and 5.
1.3 The RNase H model system

Understanding how protein sequences encode their energy landscapes is a huge undertaking. Progress on this goal requires picking model systems for extensive characterization and looking for generalizable principles. This thesis centers around work on the model system ribonuclease H (RNase H). RNase H has been an important model system in protein folding for many years. It is a small (~150 amino acid), monomeric, mixed alpha/beta protein found in all kingdoms of life. Its biological function is to cleave the RNA strand from an RNA-DNA hybrid. The homolog that has been best studied and characterized is from *E. coli*, and, in particular, the most studied version of this protein is a variant where conservative mutations have been used to remove the three native cysteines. Any reference to RNase H in this work refers to the cysteine-free version of the *E. coli* homolog, unless otherwise stated.

1.3.1 The RNase H folding pathway

RNase H is one of the best-characterized “three-state” folders, meaning that it clearly populates one intermediate species as it folds to the native state. This intermediate has been observed and characterized in a myriad of ways. In one of the earliest experiments, the folding kinetics of RNase H were monitored by circular dichroism after folding was initiated by dilution from chemical denaturant in a stopped flow instrument. Though only single exponential kinetics are observed, biphasic kinetics can be inferred because of the large signal change that occurs within the mixing time of the instrument, termed a “burst phase” (Figure 1.3C). This indicates that the protein completely populates an intermediate within the mixing time of the experiment (12 milliseconds) and that the observed folding is from the intermediate to the native state.

The structured region of the intermediate was studied using pulsed hydrogen exchange monitored by NMR as well as by mutational analysis. Both studies agreed that the region of the protein that is structured in the intermediate is generally comprised of the alpha helical “core” region of the protein, encompassing helices A through D (as well as strand IV). A computational algorithm for predicting autonomously folding units of protein structure successfully identified the same region of RNase H. The algorithm offered a more detailed prediction for the exact boundaries of this region than could be determined definitively from the experimental data, most notably by the inclusion of strand V. These boundaries have been used as the definition of the folded region of the intermediate, as highlighted in Figure 1.4A. This intermediate is referred to as Icore.

Other experiments have demonstrated that Icore is an on-pathway, obligate intermediate. Optical tweezers were used to monitor the folding and unfolding of a single RNase H molecule (observed by the end-to-end extension change). This experiment directly observed that formation of Icore precedes formation of the native state, indicating that the intermediate is both obligate and on-pathway. The robustness of this folding pathway was demonstrated by the study of a single amino acid variant of RNase H that does not populate Icore. The I53D mutation completely destabilizes Icore so that RNase H folds in a two-state manner and does not obviously populate an intermediate. However, a mutational analysis indicates that even in these conditions when Icore
is destabilized, RNase H still folds through the same pathway, with the alpha helical core region folding first.\textsuperscript{13}

Despite all of this work on the folding pathway of RNase H, a big gap remains in our knowledge: How does \( I_{\text{core}} \) form? Does \( I_{\text{core}} \) form in one cooperative step, or are there early intermediates preceding its formation? Are there multiple ways to reach \( I_{\text{core}} \) or a single dominant pathway? The earliest events in a folding pathway are really important determinants of the overall folding reaction, but they are the most difficult to access experimentally because they involve the highest-energy intermediates; consequently, the earliest steps in RNase H folding have remained uncharacterized.

1.3.2 \textbf{The nature of side chain packing in the \( I_{\text{core}} \) intermediate}

Though we know the region of the protein that is structured in the \( I_{\text{core}} \) intermediate, it is still an open question what the nature of that structure is: is it well-folded, or does it have only loosely associated hydrophobic contacts? Such questions about tertiary structure are difficult to address for a very high-energy, transiently populated state. And because \( I_{\text{core}} \) is populated immediately preceding the rate-limiting barrier to folding, knowledge about its side chain packing is critical to understanding the nature of the rate-limiting barrier. If \( I_{\text{core}} \) is not well folded, then perhaps the barrier involves the formation of specific tertiary contacts in the core and/or squeezing out solvent. If \( I_{\text{core}} \) is well folded, the barrier is more likely related to docking and assembly of the beta sheet region of the protein.

\textbf{Figure 1.4} RNase H structures. \textit{A.} Ribbon diagram of the \textit{E. coli} RNase H crystal structure (PDB ID: 2RN2).\textsuperscript{23} Helices are labeled with letters and strands with Roman numerals. The structured region of the \( I_{\text{core}} \) intermediate is highlighted in blue. Residue 125 is shown in stick in strand II, and residue 153 is shown in stick in helix A. \textit{B.} Ribbon diagram of the \textit{T. thermophilus} RNase H crystal structure (PDB ID:1RIL).\textsuperscript{24} The structured region of the \( I_{\text{core}} \) intermediate model is highlighted in light red and strand I is highlighted in dark red. \textit{C.} Ribbon diagram of the NMR structure of the \textit{T. thermophilus} \( I_{\text{core+1}} \) fragment (PDB ID: 2RPI, one representative structure shown).\textsuperscript{25}
Several results have been interpreted to support the hypothesis that $I_{\text{core}}$ has only loosely associated hydrophobic contacts, but the evidence is not clear-cut. The first result is that RNase H populates an equilibrium molten globule (MG) under acidic conditions ($pH < 2$), and the equilibrium MG bears resemblance to the $I_{\text{core}}$ kinetic folding intermediate using probes of secondary structure: the CD spectra and hydrogen exchange protection patterns are very similar between the acid-state MG and $I_{\text{core}}$.\textsuperscript{14} If they share the same secondary structure, perhaps they also share the same tertiary structure and are both molten globules.

Additionally, results from mutagenesis experiments can also be interpreted as evidence that $I_{\text{core}}$ is a molten globule. The I53A mutation is destabilizing to $I_{\text{core}}$ to a similar extent as in the native state, supporting the fact that residue 53 is in the structured region of $I_{\text{core}}$.\textsuperscript{20} However, the I53L mutation does not perturb $I_{\text{core}}$ nearly as much as the native state.\textsuperscript{22} This could be because tertiary contacts are not well formed in the hydrophobic core of $I_{\text{core}}$, and therefore the hydrophobicity of residues in the core matters much more than the actual geometry. However, this result would also be observed if the core can repack to accommodate the mutation in the intermediate.

Lastly, a single point mutation, I25A, was shown to cause extensive population of the $I_{\text{core}}$ intermediate under equilibrium conditions.\textsuperscript{20} In the context of this mutation and in the presence of $\sim 2$ M urea, $I_{\text{core}}$ is the dominant species at equilibrium. In these conditions, no peaks corresponding to the structured region of the intermediate could be observed in NMR HSQC experiments, suggesting that the structured region of $I_{\text{core}}$ is very dynamic. One way for the structured region of $I_{\text{core}}$ to be dynamic would be if it is not well folded and the hydrophobic side chains are only loosely associated. But another way it would appear dynamic is if $I_{\text{core}}$ has close side chain packing but folds and unfolds rapidly. How can we distinguish the nature of the side chain packing?

### 1.3.3 Comparison between RNase H sequences

One benefit of RNase H as a protein folding model system is how many homologs of this protein there are, from many different organisms. Comparing homologous proteins shows us to what extent native-state topology dictates a protein’s energy landscape versus what is mutable by sequence changes. Such a comparison also helps us understand how sequence encodes properties such as the thermal stability, which can be different between homologs.

Other RNase H homologs that have been characterized besides the $E.\ coli$ homolog are RNases H from *Thermus thermophilus*\textsuperscript{27} and *Chlorobium tepidum*.\textsuperscript{28} In particular, a lot of work has been performed with the *T. thermophilus* homolog in an attempt to determine the origin of its high thermostability. In many ways the *T. thermophilus* RNase H (ttRNH) energy landscape appears similar to that of *E. coli* RNase H (ecRNH). They both populate a partially folded state within milliseconds that is structured in a similar region, based on hydrogen exchange experiments.\textsuperscript{29} And they both have similar distributions of stability across their structures.\textsuperscript{30} The main observed difference is that ttRNH has residual structure formed in its unfolded state.\textsuperscript{31,32}

Another possible difference in their energy landscapes is suggested by work from Yawen Bai’s lab at the National Cancer Institute.\textsuperscript{25} In this work, they constructed a fragment of ttRNH that consists of the structured region of the $I_{\text{core}}$ intermediate plus strand I (added through a non-
native junction, see Figure 1.4B,C). They suggest that this fragment is a mimic of the kinetic folding intermediate. Their inclusion of strand I was based on a few ambiguous hydrogen exchange probes (in both kinetic and equilibrium HX experiments).\textsuperscript{29,30} They solved the NMR structure of the fragment and showed that it is well folded and looks like a subset of the native state structure (Figure 1.4C). This is in contrast to our knowledge that strand I is not involved in the ecRNH $I_{\text{core}}$ folding intermediate. Is their fragment truly a mimic of the ttRNH folding intermediate populated prior to the rate-limiting step to folding? If so, this would be a big difference between the energy landscapes of these two homologous proteins.

1.4 Summary of thesis work

In this thesis, I present an extensive analysis of the early folding pathway of RNase H. I address the questions: How does RNase H fold to the $I_{\text{core}}$ intermediate? What is the nature of side chain packing in the $I_{\text{core}}$ intermediate? And, are there differences in the early folding pathway of RNase H between the $E. \coli$ and $T. \text{thermophilus}$ homologs? The chapters are summarized below.

Chapter 2
Secondary structure formation in the folding pathway of $E. \coli$ RNase H is monitored at unprecedented resolution using a pioneering hydrogen exchange technique monitored by mass spectrometry. This work is a collaboration with the lab of S. Walter Englander at the University of Pennsylvania. I generated the protein, helped with data collection and interpretation, and helped write the manuscript. We observe two novel intermediates early in the folding pathway of RNase H, in addition to the previously characterized $I_{\text{core}}$ intermediate. The key result is that the early folding of RNase H follows a single dominant pathway consisting of stepwise assembly of native units of secondary structure. The assembly of structure occurs in the following order: first, helix A and strand IV; second, helix D and strand V; third, helices B and C; lastly, in the rate-limiting step, strands I-III and helix E. This chapter has been adapted from a manuscript that was published in 2013 in the Proceedings of the National Academy of Sciences.\textsuperscript{33}

Chapter 3
The first several milliseconds of $E. \coli$ RNase H folding are directly observed by spectroscopy for the first time, using a probe of tertiary structure. A continuous flow microfluidic mixing device was used to initiate refolding in 50 microseconds and folding was monitored by the change in tryptophan fluorescent lifetime. This work was performed in collaboration with Osman Bilsel and Sagar Kathuria in the lab of C. Robert Matthews at the University of Massachusetts Medical School. (I was the lead on this work.) We observe the formation of two folding intermediates, the second of which is the $I_{\text{core}}$ intermediate. The other, earlier intermediate forms structure in a region of the protein that has no hydrogen exchange protection at the earliest folding times (as shown in Chapter 2), indicating it involves non-native structure. Other evidence suggests that this may be the same species as the earliest intermediate observed in Chapter 2 (helix A + strand 4). If helix A is stabilized by non-native structure in the earliest intermediate, it would be a rare example of productive non-native structure formation and would explain why helix A is the first native helix to form despite not having the highest intrinsic helicity. Additionally, we use a fragment of RNase H (characterized in Chapter 4) to demonstrate that only half of the protein is
significantly involved in early structure formation. Our results provide detailed folding information on both a timescale and a size-scale accessible to all-atom molecular dynamic simulations of protein folding.

Chapter 4
The side chain packing in the structured region of the *E. coli* RNase H *I*core intermediate is evaluated using mimics of *I*core that are populated at equilibrium. An open question has been whether the structured region of *I*core is a molten globule or rather has a distinct, well-folded conformation. *I*core is the intermediate populated prior to the rate-limiting barrier to folding, so understanding its structural properties is important for understanding the nature of the rate-limiting barrier. We take two approaches to making mimics, based on our knowledge of the unfolded region of *I*core (i.e. the region unprotected from hydrogen exchange): 1. Remove the unfolded region to create a fragment mimic, and 2. Make mutations in the unfolded region to selectively destabilize the native state relative to *I*core, creating a full-length mimic. The second approach was originated by Katelyn Connell in her thesis work in the Marqusee Lab. We show that the two constructs have very similar properties, and are similar to the *I*core kinetic intermediate in ways that suggest they are indeed good mimics. Additionally, both mimics have closely packed side chains, a result that we demonstrate is consistent with previous evidence that had been used to support molten-ness. We conclude that the *I*core intermediate of RNase H is likely not a molten globule.

Chapter 5
In the final chapter, I study the *T. thermophilus* RNase H homolog in comparison to the *E. coli* homolog in work that highlights how two proteins with identical native state topology can have differences in their folding pathways. I use equilibrium mimics to determine that *T. thermophilus* RNase H populates two different partially folded intermediates on its energy landscape: one equivalent to the *E. coli* RNase H *I*core intermediate, and a second one with a larger structured region. Additionally, I use multiple kinetic techniques to place these intermediates on the energy landscape relative to the rate-limiting barrier to folding. This work is ongoing, but preliminary results suggest that both intermediates are populated prior to the rate-limiting step to folding. This difference in the folding pathways of *E. coli* and *T. thermophilus* RNases H is one of the most dramatic differences observed between their energy landscapes to date, and changes our understanding of the rate-limiting barrier to folding for this well-studied protein.

1.5 References

13
23. Katayanagi, K., Miyagawa, M., Matsushima, M., Ishikawa, M., Kanaya, S., Nakamura, H.,
from Escherichia coli as refined to an atomic resolution. *Journal of Molecular Biology*
**223**, 1029–1052

HB8 refined at 2.8 Å resolution. *Journal of Molecular Biology* **230**, 529–542

of the Early Folding Intermediate of the Thermus thermophilus Ribonuclease H. *Journal
of Molecular Biology* **384**, 531–539

Populates the Folding Intermediate of E. coli RNase H and Reveals a Highly Dynamic

Thermophilic Ribonucleases H. *Biochemistry* **38**, 3831–3836

Ribonuclease H1 from the Moderately Thermophilic Chlorobium tepidum: Comparison

and E. coli ribonucleases H. *Journal of Molecular Biology* **316**, 327–340


structure in the unfolded state of a thermophilic protein. *Proceedings of the National
Academy of Sciences* **100**, 11345–11349

(2003). Energetic evidence for formation of a pH-dependent hydrophobic cluster in the

(2013). Stepwise protein folding at near amino acid resolution by hydrogen exchange and

34. Connell, K.B. (2010). Beyond the native state: Exploring the role of partially folded
conformations on the protein energy landscape. *UC Berkeley Electronic Theses and
Dissertations*
Chapter 2

The stepwise folding of *E. coli* RNase H observed at near amino acid resolution using hydrogen exchange and mass spectrometry

This chapter is adapted from the paper:

2.1 Introduction

Do proteins fold through varied and multiple tracks, or do they fold through predetermined intermediates according to understandable biophysical principles? This question is fundamental for the interpretation of a large amount of biophysical and biological research. The question could be resolved if it were possible to define the intermediate structures and pathways that unfolded proteins move through on their way to the native state. Unfortunately, transient intermediates cannot be studied by the usual crystallographic and NMR methods. A wide range of kinetic and spectroscopic methods has been applied to many proteins, but these methods often do not yield the necessary structural information.

We used a developing technology, hydrogen exchange pulse labeling measured by MS (HX MS), to study the folding of a cysteine-free variant of *Escherichia coli* ribonuclease H1 (RNase H), a mixed α/β protein that has served as a major protein folding model. Previous studies showed that RNase H folds in a fast, unresolved burst phase (15 ms dead time) to an intermediate termed “I(core)” and then much more slowly (in seconds) to the native state. HX pulse-labeling and equilibrium native-state HX experiments monitored by NMR showed that I(core) comprises a continuous region of the protein between helix A and strand 5 and that β-strands 1, 2, and 3 and helix E acquire protection much later, consistent with mutational analysis. Single-molecule and mutational studies indicated that the intermediate is obligatory, on-pathway, and folds first even when I(core) is not observably populated.

The HX MS technique used here is able to follow the entire folding trajectory of RNase H in considerable structural and temporal detail. The analysis monitors every amide site, evaluates the folding cooperativity between them, and describes the separate folding steps. The results identify at near amino acid resolution the formation and stepwise incorporation of native-like units of secondary structure in four sequential events that progressively assemble the native structure. A comparison with other experimental and theoretical observations suggests that this pathway behavior is the prevalent mode for protein folding and that it is dictated by two straightforward biophysical principles.
Figure 2.1 The folding of RNase H monitored by circular dichroism. A. A burst-phase intermediate ($I_{core}$) is formed within the dead time of the experiment (manual mixing) followed by slower folding (30 s) to the native state (0.6 M urea, 10 °C). B. The equilibrium urea melt (black circles), the amplitude of the burst phase (diamonds), and the final kinetic amplitude (white circles). C. Folding and unfolding rates (chevron plot) at 10 °C (black trace) compared with the published fit at 25 °C (gray trace).

2.2 Results

2.2.1 Folding by spectrophotometry

Figure 2.1 shows results for RNase H folding monitored by circular dichroism under the conditions used in the HX MS studies (10 °C, pH 5). Folding is very similar to the observations in previous studies (25 °C, pH 5.5) but with slower final folding to the native state. The results fit to a three-state model [unfolded (U), intermediate (I), and native (N)] with the following parameters at 0 M urea. The free energy of unfolding for the intermediate ($\Delta G_{UI}$) is 4.1 kcal/mol, the free energy of global unfolding ($\Delta G_{UN}$) is 10.1 kcal/mol, and the rate constant for folding from the intermediate to the native state ($k_{IN}$) is 0.07 s$^{-1}$. These spectrophotometric data provide population-averaged kinetic and thermodynamic folding parameters with little structural detail or information about pathway steps.

2.2.2 HX MS experiment

Figure 2.2 diagrams the HX MS pulse-labeling experiment. To begin the folding process, urea-unfolded protein fully deuterated at amide sites in D$_2$O is diluted into H$_2$O folding buffer under conditions in which D-to-H back exchange is very slow (pH 5.0, 10 °C; ~40-s HX time constant). After trial folding times, a brief labeling pulse to high pH is imposed (10 ms, pH 10, 10 °C), equivalent to ~25 times the average HX time constant. When the protection factor (Pf) of an amide site at the time of the labeling pulse is very low (<10), it will appear unprotected and exchange to H. It will appear protected and remain almost fully D-labeled when Pf >100 or $k_{op} <$ 20 s$^{-1}$, and it will be fractionally H/D labeled in the narrow range of intermediate Pf values. As a result, any amide position switches from fully unprotected when it is unfolded to almost fully protected when folding occurs. We associate the gain of HX protection with the formation of H-bonded secondary or tertiary structure.$^8$
The subsequent HX MS analysis is designed to detect which residues are protected and which are not at each folding time. To preserve the H/D labeling reached during the labeling pulse, the protein sample is quenched immediately to a slow exchange condition (pH 2.5, 0 °C) and is proteolyzed quickly into many overlapping peptide fragments. The fragments then are separated and analyzed for carried deuterium by HPLC and mass spectrometry. Previous papers describe how to maximize the number of peptide fragments obtained, how to minimize the back exchange of D label during sample preparation, and a program (ExMS) for identifying and characterizing the many peptides in the final MS spectra. These methods provided 326 useful peptide fragments; 228 unique fragments for this 155-residue protein are indicated in Figure 2.2B. Of these, 156 unique peptides passed the ExMS autocheck tests (no operator intervention) for all folding experiments out to 3 s and were used for the time-dependent analysis. For residue-resolved analysis, the entire peptide set was used.

2.2.3 Data interpretation

Each HX MS experiment at each folding time point records ~1,000 mass spectra. The ExMS analysis separately identifies and characterizes each of the hundreds of peptide fragments detected. Each peptide monitors the folding history of the segment of the protein it represents. As folding proceeds and amides become protected at the time of the labeling pulse, each peptide converts from its unfolded, lighter state to a state that is heavier by a mass increment equal to the number of D-labeled sites protected.

Different folding scenarios will produce recognizably different results. At the individual peptide level, protection that occurs one or a few residues at a time will be seen as a continuous slide in time from an unprotected state to more protected states. If a folding step protects many amides in a given segment all together, the conversion will be seen as a bimodal envelope with occupancy...
shifting in time from the lighter to the heavier envelope.

For the whole protein, if folding is two-state, all segments will transition to native-like protection in the same concerted single-exponential step. In a multistep pathway, different regions fold at different times, and peptides that represent those regions will acquire protection accordingly. For example, in a classical stepwise pathway, 100% of any given segment will convert to its protected state in a concerted step, and other segments will convert in other steps, defining the pathway sequence. If a protein folds through several parallel pathways, any given segment will convert in fractional steps, earlier in some molecules of the protein population and later in others. If structure is formed and then lost, whether native-like or nonnative, the timed measurements of peptides that cover that region will reveal that behavior.

In summary, when many peptides are seen individually, their HX MS behavior can depict the spatially resolved and the time-resolved development of protection of segments throughout the protein. During data analysis, the many overlapping peptide fragments provide multiple internal consistency checks. In an ultimate step, multipeptide folding data can be deconvolved to near amino acid resolution. This information can specify the detailed protein folding behavior.

2.2.4 MS results

HX MS results for the 156 overlapping RNase H peptides measured as a function of folding time are given in the Supplementary Information of Hu et al. Four are shown in Figure 2.3. The top panel of MS data shows the spectrum obtained for the unfolded protein (6 M urea, protonated, lower mass). The bottom panel shows the spectrum obtained from the native protein where the D label at H-bonded amide sites is protected and retained through the D-to-H labeling pulse. The time-dependent data provide a graphic snapshot of the fraction of the protein population that already is protected (heavier) and the fraction that is not yet protected (lighter) at each segmental position at the time of the labeling pulse.

Figure 2.3A shows a peptide that monitors the C-terminal turn of helix A and most of β-strand 4. This segment forms structure that protects a sum of approximately eight deuterons before the first measured folding time point (9 ms), which appears as a shift of the peptide envelope to higher mass. The same state is adopted by 100% of the protein population; no lighter population exists. The segment remains folded at all later times in the folding process, and peptides that monitor other positions add on in subsequent folding steps. The same is true in all cases.

The peptide in Figure 2.3B monitors most of helix D plus β-strand 5. At the earliest measured time only a small population fraction is not yet folded, indicating a folding rate slightly slower than for the A/4 region.

The peptide in Figure 2.3C monitors the kinked B/C helix plus the long connecting loop to helix D. The measured data for this peptide capture the last part of its transition to a folded and protected state. About 7% of the population remains unprotected until later in the folding process. One explanation would be a mis-isomerized proline creating a barrier for a small portion of the molecules. To investigate this, the one proline present in this region of sequence (P97) was mutated to glycine and the HX-MS experiment performed with the single-site variant.
Figure 2.4 shows the P97G HX-MS results at a folding time of 720 ms for the equivalent peptide as in Figure 2.3C. The unprotected population persists in the P97G variant. Further, the unprotected population becomes much larger when a longer pulse length is used (27 ms vs 10 ms). This indicates that the unprotected population is not due to a proline block and likely arises due to the dynamic behavior of this segment after initial folding.

Finally, the peptide in Figure 2.3D monitors helix E and a long C-terminal protein segment. Protection develops with a halftime of ~30 s. Other peptides detect equally slow folding for the N-terminal β-strands 1, 2, and 3. This final folding step completes the native state.

Figure 2.3 Illustrative MS spectra versus folding time. The ribbon diagrams on top illustrate the location of each peptide in the native structure (colored region). The top and bottom panels show control experiments in which the unfolded and native proteins were subjected to the same labeling pulse and analysis. Fitted envelopes separate the fractional populations of the unfolded, intermediate, and native state present at the time of the labeling pulse. Deuterons on side chains and the first two residues of each peptide are lost during sample preparation. The subpeaks within each isotopic envelope are caused by the natural abundance of 13C (~1%) convolved with the carried number of deuterons. A leftward drift in folded peptide mass at long folding times occurs because not-yet-protected sites are exposed to D-to-H exchange during the pre-pulse folding period (pH 5, 10 °C).
The D-label protected in each earlier step does not yet match the native state – not because of the protection of fewer sites in the already-formed structure but because of less than complete protection at numerous positions. This is likely because the folding intermediates, although “native-like” structurally, are not yet fully native and therefore are less stable and more dynamic.

Figure 2.5 plots the time dependence for folding of the different protein segments. The ordinate refers to the fraction of the protein population that has reached the protected state at specified regions of the protein. The Inset is renormalized to show that the green (helix D/strand 5) segment folds detectably faster than the yellow (helix BC/loop) segment.

These results identify four kinetically separable steps in RNase H folding.

2.2.5 Partially folded structure at high resolution

The HX MS data for the large data set of many overlapping peptides at various folding times can be deconvolved to near amino acid resolution (using the HDsite program). The results (Figure 2.6) identify the structural regions that fold in each of the steps just described. They represent secondary structural elements of the native protein.

To guide the interpretation of residue-resolved pulse-labeling results, native RNase H was passed through the same pulse-labeling and analysis procedure. We expect that sites protected by H-bonding will resist exchange during the D-to-H labeling pulse, regardless of their exposure to solvent.8,9 The results match this expectation.

To obtain the earliest possible kinetic information, a competition experiment14,15 was done in which unfolded deuterated protein was diluted into folding and labeling conditions in the same mixing step. In this experiment, protection competes kinetically with labeling. We find measurable protection in the helix A segment (~50%) and more evanescent protection in segments in β-strand 4 and helices C and D but none in helix B and E or other β-strands (Figure 2.6, Top). The protection develops as rapidly as the mixing process (perhaps 0.1 ms). Although
helix A and strand 4 may fold on this time scale, helix C and D fold later, suggesting either some nascent pre-folding protection when the still unfolded protein is placed into native conditions or their interaction with already formed helix A and strand 4 early in structure formation. The protection pattern does not correlate with intrinsic helix propensity, estimated by AGADIR16 at 50% for helix E (see also ref.17), 30% for helix A, and as negligible for helices B, C, and D.

By the 9-ms folding time point, the helix A and β-strand 4 segments (blue in Figure 2.7A) have achieved protection in 100% of the refolding population. The residue-resolved pattern of protection is similar but not identical to the native protein. For example, the N terminus of helix A is less protected, as might be expected because of end fraying in the partly folded intermediate. The far N- and C-terminal segments remain wholly unprotected.

At this time point, peptides that monitor the segments shown in green and yellow in Figure 2.7A exhibit bimodal envelopes (Figure 2.3). The bimodal MS transition documents their cooperative folding behavior and measures the unfolded and folded fraction of the population at each time point. Bimodal data, noted in gray in Figure 2.6, cannot be analyzed to site resolution but this limitation does not hinder the analysis of other peptides.

At the 720-ms folding time, helices A through D, their local β-strand segments, and even the connecting C-to-D loop (blue, green, and yellow in Figure 2.7A) have achieved protection patterns similar to the native protein. By 20 s, the extensive N- and C-terminal regions have begun to fold in a bimodal, cooperative way.

Interestingly, the C-terminal Val155 amide remains deuterated in all cases. This apparent anomaly is caused by HX chemistry during the pulse, namely an expected 60-fold slowing factor for the C-terminal amide multiplied five-fold by the valine side-chain effect.18

Figure 2.5 Time dependence for the protection of different protein regions, color coded to match Figure 2.7. Inset: The folding phase of the yellow curves is renormalized to 100% to allow direct comparison with the folding time of the green segment. For this comparison, the experiment was replicated in triplicate, and only the highest-precision peptides were used. The green and yellow segments fold in detectably different phases. Peptides are identified in the SI Appendix of Hu et al13.
2.2.6 Anomalously slow exchange for β-edge amides
The residue-level analysis at the early folding time points finds that some residues in strand 4 are moderately protected even when their native H-bond acceptors on strand 1 are not yet in place. Although this protection might be thought to indicate nonnative interactions, several amide hydrogens on the unprotected solvent-exposed edges of β-strands 3 and 5 in the native protein similarly avoid D-to-H exchange in the labeling pulse. Analogous behavior has been seen before.8,9,19
2.3 Discussion

Pulse-labeling HX MS results obtained during the kinetic folding of the single-domain RNase H protein graphically display a time-resolved sequence of four concerted steps in structure formation. Each step corresponds to the folding of one or more secondary structural elements of the native protein. The residue-resolved pattern of HX protection mimics that seen for the native protein, implying native-like features for each added unit of structure and the intermediate states that they produce. Once formed, these units remain in place as subsequent units are added, demonstrating a sequential stepwise buildup of the native structure. Each step involves essentially the entire protein population, indicating a single dominant route.

The helix A + strand 4 segment (blue in Figure 2.7A) folds within the first millisecond. Helix D and strand 5 (green in Figure 2.7A) add on at ~5 ms. In the native protein these two elements pack together to form a major hydrophobic core of the protein. Shortly thereafter (~9 ms), the more tenuously associated yellow segment (Figure 2.7A) folds to complete the I\textsubscript{core} intermediate previously identified by pulse-labeling HX NMR.\textsuperscript{3,20} Finally, in a much slower reaction (~30 s) the elements shown in gray in Fig. 7A (strands 1, 2, and 3 and helix E) fold concertedly, even though they are drawn from the most distant protein termini.

It is hard to picture this sequence of segmental folding events as anything other than a classical pathway that constructs the native protein by the sequential incorporation of native-like units of structure. If any significant fraction of protein (>4%) formed a protecting structure that differed in any significant way, structurally or kinetically (e.g., an alternative parallel pathway), it would be seen clearly, as we see for the 7% unprotected population in the yellow segment. The folding behavior observed is well represented by the pathway in Figure 2.7B.
2.3.1 Experimental work to the contrary
The folding model pictured in Figure 2.7B is analogous to the extensively worked out case of cytochrome c for which a series of HX NMR studies defined four native-like intermediates in a well-ordered pathway.21-23 It also is consistent with the finding of individual native-like intermediates in many other proteins. However, some experimental studies have been interpreted in terms of more complex models. Those studies have depended mainly on spectroscopic measurements that provide little structural information and therefore are subject to ambiguity; prominent suggestions are that proteins may fold through many alternative intermediates, or through none at all, and that observed intermediates are nonnative and therefore nonproductive, or even obstructive.

In respect to the suggestion that proteins may fold through a number of alternative, independent, parallel routes, we have shown that the kinetic complexity observed in such spectroscopic studies can be explained with fewer fitting factors and equivalent or better $\chi^2$ fit by a single pathway in which a probabilistic barrier insertion affects some fraction of the population and not another.23,24 Given only kinetic phase data and no structural information, it is not possible to distinguish a multiple pathway interpretation from a single pathway with an optional barrier that slows one population fraction and not another.23,24

Many small proteins are known to fold in an apparent two-state manner without observable intermediates. Folding intermediates, however, will be kinetically visible only under certain limited conditions. They must occupy a free-energy well that is relatively low compared with all prior wells and be blocked by a forward barrier that is sufficiently high relative to all prior barriers. Otherwise, intermediates will not accumulate noticeably even when they are present and important.

The requirement that the population of an intermediate depends on a subsequent large barrier has led to the idea that folding intermediates may be obstructive because visible intermediates and slowed folding go together. Rather it is the barrier, whether optional or intrinsic, that slows folding and correlates it with the population of an otherwise invisible intermediate. In the present experiments, the ability to distinguish and characterize RNase H folding intermediates depended on the presence of barriers between each of the folding steps. However, it seems clear that the sequential intermediates observed are on-pathway and constructive.

Do nonnative interactions imply that an intermediate is non-obligatory and off-pathway? An alternative is that the intermediate is on-pathway and productive but in addition has some nonnative character, which may or may not tend to slow folding. In a partly folded intermediate, energy-minimizing nonnative interactions can be expected25 and have been seen in intermediates that are clearly on-pathway and constructive.26-30

In summary, although a quantity of ambiguous protein folding data has led to different interpretations, it appears that the great majority of experimental protein-folding observations can be understood in terms of the reaction scheme in Figure 2.7B.
2.3.2 Theoretical studies

In 1992 Zwanzig et al.\textsuperscript{31} showed that, in principle, protein folding need not require any particular pathway. A small energetic bias toward native-like interactions would allow the native state to be found quickly through multiple independent routes. The formulation was phrased in terms of native-centric selectivity at the amino acid level. Theoretical studies before and since, focused at the amino acid level, naturally emphasize the role of an ensemble of microscopically diverse structures, but this diversity is not at odds with higher-level, more deliberate pathway behavior. It now appears that distributed, residue-level searching does not lead directly to the native protein but rather functions to sequentially construct small units of native-like structure. Here the Levinthal paradox\textsuperscript{32} is reduced to a manageable problem, because a random search for a small structural unit can be accomplished rapidly. This microscopic behavior is out of the reach of macroscopic experiment. One must look forward to the promise of theoretical approaches for ultimately providing a deep understanding of protein folding from this basic level and up. The present work contributes to this effort by demonstrating that RNase H folds via a well-ordered pathway where native-like units of structure assemble to form intermediates and, finally, the native protein.

Physically based computational folding simulations can, in principle, discern folding mechanisms from the fundamental residue-level steps and up, but severe challenges exist. These include the immense computer power required to reach folding timescales, the great accuracy required for the force fields, and the need to extract a descriptive trace of the progression of the folding reaction. Recent advances have begun to transcend these limitations, bridging the divide between the micro- and macroscopic views of protein folding. In a computational tour de force, the D.E. Shaw group simulated the folding of 12 small proteins, observing that all appeared to fold through one dominant pathway by sequential formation of native structural elements.\textsuperscript{33} Another recent approach explicitly uses sequential stabilization in an iterative search strategy.\textsuperscript{34} Some other theoretical efforts lead in this same direction.\textsuperscript{35} These scenarios closely resemble the reaction scheme in Figure 2.7B. Other funnel and network models are conceptually different but are not inconsistent with this conclusion.\textsuperscript{36-41}

2.3.3 Principles of protein folding

The results and considerations described here support a determinate mechanism for protein folding. Two straightforward biophysical principles appear to underlie this mechanism. The first is the inherent cooperativity of subunits of protein structure. The unfolding and refolding of cooperative structural units and their role in forming intermediates and pathways has been detected, in greater or lesser detail, in many proteins by HX methods,\textsuperscript{20-22,42-49} by sulfhydryl labeling,\textsuperscript{50,51} and by relaxation dispersion NMR.\textsuperscript{52}

The second principle orders the pathway and determines its sequential nature. In the initial on-pathway step, a limited structure is assembled by a relatively unguided amino acid search.\textsuperscript{53} In subsequent steps in which amino acid searching can be informed by existing structure, an assisted folding mode emerges called "sequential stabilization."\textsuperscript{22,54} Like the process of coupled folding upon binding,\textsuperscript{55-58} the prior structure will selectively guide and stabilize regions with which it interacts in the native protein. When prior structure can support more than one incoming region, the sequential stabilization principle can lead to pathway branching.\textsuperscript{59}
The ordered buildup of native structure observed here must depend on the summed free energy of multiple native-like interactions acting together to select each correct step among the many competing alternatives. It seems likely that the above two factors—cooperative structure formation and sequential stabilization—could provide the required selectivity to clearly define a folding pathway whereas individual residue-level interactions would not. When the bias toward native interactions is insufficient, probabilistic misfolding is likely to occur.22

2.4 Materials and Methods

Protein
The version of *E. coli* RNase H [Protein Data Bank (PDB) ID: 1F21] used here is the wild-type protein with its three cysteines changed to alanine. It was expressed and purified as described previously.60

Kinetic folding and pulse labeling
Previously described methods10–12 were used to obtain the many peptide fragments studied here. Of these, 156 unique peptides (203 with different charge states) were found to pass the ExMS autocheck tests (no operator intervention) and were used in the timed 9-ms to 3-s folding experiments. The HX MS experiment is diagrammed in Figure 2.2 and detailed in the Appendix of Hu et al.13

Sample Analysis
The selectively H/D-labeled reaction mixture flowed into a home-built cooled chamber containing an online flow analysis system (low pH, 0 °C) in which the protein sample was proteolyzed (immobilized tandem pepsin and fungal protease columns), caught on a trap column, washed, eluted, roughly separated (HPLC column, acetone, acetonitrile gradient), and continuously electrosprayed into the mass spectrometer for a second dimension of separation by mass (~1,000 MS spectra). Previous papers describe methods for obtaining and analyzing many fragments11 and for minimizing back exchange of the D-label during sample preparation.12

Mass spectra were processed by the home-written ExMS program10 to identify and characterize all of the individual peptide isotopic peaks and envelopes and to read out the placement of protecting structure and the time course for its formation during refolding. To quantify folded (HX-protected) and not-yet folded populations present during the H–D exchange-labeling pulse, bimodal MS envelopes were fit by a home-written program (HDpop), using binomial fitting with appropriate weighting. To obtain H–D labeling information at amino acid resolution, MS results for many overlapping peptides were analyzed using in-house software (HDsite).

2.5 Acknowledgments

We thank Y. Bai, D. Barrick, N. R. Kallenbach, G. D. Rose, J. J. Skinner, T. R. Sosnick, and A. J. Wand for helpful suggestions on the manuscript.
2.6 References

35. Weinkam, P., Zong, C. & Wolynes, P.G. (2005). A funneled energy landscape for cytochrome c directly predicts the sequential folding route inferred from hydrogen...
exchange experiments. *Proc Natl Acad Sci USA* **102**, 12401–12406


Chapter 3
Direct observation of the earliest events in the folding of *E. coli* RNase H using ultrarapid mixing and intrinsic fluorescence

3.1 Introduction

Protein folding spans a wide range of timescales, from microseconds to minutes. Even for the slower folders, many reach partially structured intermediate states on microsecond or millisecond timescales. These early folding events are important determinants of the overall reaction and are the ones accessible to detailed simulations of folding, but are the hardest to directly observe. The improvement of folding simulations – important for our ability to predict protein dynamics and behavior from sequence – requires detailed experimental information for comparison and validation. In this work, we use a microfluidic continuous flow mixer to directly observe the earliest folding of an important protein folding model system, *E. coli* RNase H, at a timescale – and size-scale – amenable to simulation.

*E. coli* RNase H populates an obligate, on-pathway, partially folded intermediate within several milliseconds of folding, with subsequent folding to the native state occurring in seconds. (All work on *E. coli* RNase H discussed here refers to a cysteine-free variant.) The intermediate, termed I\textsubscript{core}, has been characterized using pulse-labeling hydrogen exchange monitored by NMR and also by mutational analysis, and contains native-like secondary structure in approximately half of the protein (Figure 3.1). Although very well characterized, folding to this intermediate has never been observed directly, as it occurs within the dead time of a stopped-flow instrument.

Recently, the formation of the I\textsubscript{core} intermediate was monitored using pulse-labeling hydrogen exchange and a novel mass spectrometry technique (HX MS) (Chapter 2). Two new early folding intermediates were identified. (The experiment was conducted at 10°C instead of the 25°C conditions of previous experiments, slowing early folding events so they were accessible in a quench-flow instrument.) This work was a landmark for detailed structural characterization of early folding events, and critical for our understanding of RNase H folding. However, the data provided only a rough sense of the rates associated with the early steps, and were blind to tertiary structure formation.

In the present work, we monitor the first 9 milliseconds of RNase H folding spectroscopically (at 25°C) using ultra-rapid continuous flow mixing with a 50 microsecond dead time, allowing us to characterize early folding kinetics with high temporal resolution. Further, we use intrinsic tryptophan fluorescence to monitor the progress of the folding reaction, providing a window into tertiary structure formation. RNase H has six tryptophans clustered in two regions (Figure 3.1). Both regions are within the structured portion of I\textsubscript{core}, providing a good opportunity to observe the formation of this intermediate. Our results clearly demonstrate two kinetic steps in the first few milliseconds of RNase H folding, revealing the formation of a new early intermediate (I\textsubscript{early}) in addition to the formation of I\textsubscript{core}. Mutational analysis and comparison with the HX MS data suggests that I\textsubscript{early} is an on-pathway intermediate containing non-native structure. Additionally,
we use a fragment of RNase H (characterized at equilibrium in Chapter 4) to confirm that only half the protein is significantly involved in the early folding steps. These results, together with the HX MS data, result in a detailed model for the early folding of RNase H on both a timescale and size-scale amenable to comparison with atomistic folding simulations.

3.2 Results

3.2.1 Direct observation of two kinetic phases in the first nine milliseconds of folding
Folding of *E. coli* RNase H was initiated using a 6 M to 0.6 M urea concentration jump in a microsecond-resolved continuous flow (CF) mixing device with a 50 microsecond dead-time. Folding was monitored by the change in average fluorescence lifetime of the tryptophans, determined using time-correlated single photon counting (TCSPC). Plotting the average lifetime versus folding time shows that we are capturing two kinetic phases, clearly distinguishable by the opposite directions of their change in amplitude (Figure 3.2A, inset). The second kinetic phase, however, was poorly determined in these initial experimental conditions where we could only monitor folding up to one millisecond.

In order to better resolve the second kinetic phase, we repeated the experiment using a larger CF mixing device that has a ~one millisecond dead time, with folding observed up to ~nine milliseconds. The second kinetic phase is well resolved in this time range. The data from both experiments were combined, and the change in quantum yield-weighted average fluorescence lifetime as a function of folding time was fit globally to a two-exponential decay (Figure 3.2A). The fit yielded time constants of 120 +/- 11 microseconds and 3.1 +/- 0.12 milliseconds for the first and second kinetic phases, respectively.

![Figure 3.1 Structure of *E. coli* RNase H. Helices are labeled with letters and β-strands with Roman numerals. The region that is structured in the Icore intermediate is colored blue. Tryptophan residues are shown in stick (in the 4Trp variant, the two green tryptophans are mutated to phenylalanine, leaving only the four orange tryptophans).](image_url)
3.2.2 Equilibrium unfolding monitored by average fluorescent lifetime
We performed an equilibrium urea titration monitored by TCSPC to relate the average fluorescent lifetimes observed in the kinetic experiment with the signal of the unfolded and native states. The equilibrium data shows a cooperative folding transition (Figure 3.2B) with a $\Delta G_{\text{unf}}$ of 9.5 +/- 0.2 kcal/mol, and an m-value of 2.13 +/- 0.04 based on a two-state assumption and linear-extrapolation model, consistent with previous work on RNase H.2

The native state at 0.6 M urea has an average fluorescence lifetime of 2.36 ns, smaller than the final amplitude of our observed millisecond kinetics (2.60 ns). Extrapolation of the unfolded baseline back to 0.6 M urea results in an expected average lifetime of 2.58 ns for the unfolded state at the start of the kinetic experiment. The amplitude of the observed kinetics extrapolated to zero time is 2.69 ns. These values are just outside of our error (error is approximately +/- 0.05 ns, based on an error of 15 microseconds – one time step – for determining the zero time). The difference could be due to the inaccuracy of such a long extrapolation of the unfolded baseline or could indicate that there is a very early kinetic phase occurring within our 50 microsecond mixing time.
To further characterize the two kinetic phases, we repeated the ultra-rapid mixing experiments with higher final concentrations of denaturant. Final urea concentrations were picked to span the accessible range for each mixer: the accessible range is determined by the pressure in the mixer during the experiment, which increases with more urea, and is higher in the smaller mixer. The results are plotted in Figure 3.3. These data are compared with possible kinetic models in section 3.3.1.

**3.2.4 The effect of the I53D mutation on early folding**

The single amino acid substitution I53D is enough to fully destabilize the I\textsubscript{core} intermediate.\textsuperscript{11} The I53D variant still folds to the native state and follows the same general folding pathway as wild type,\textsuperscript{12} but overall folding is greatly slowed, and the I\textsubscript{core} intermediate is never detectably populated. To test whether we are monitoring the formation of the I\textsubscript{core} intermediate in our experiment, we performed the CF mixing experiment with I53D.

Refolding was performed in both mixers using a 6 M to 0.6 M urea dilution. No kinetics were observed in the larger mixer; however, interestingly, we still observe a fast kinetic phase in the smaller mixer (Figure 3.4). (Data from the two mixers are displayed separately since we cannot rigorously combine them without an overlapping kinetic process.) Overall, the I53D mutation abolishes the second kinetic phase observed in the wild-type experiments, suggesting that the second phase is indeed formation of I\textsubscript{core}.

![Figure 3.3 Urea dependence of wild type E. coli RNase H early folding kinetics. Refolding of wild type RNase H was monitored at multiple final urea concentrations, spanning the accessible range of urea concentrations based on the pressure limits of the mixers (pressure increases with urea concentration). The data from experiments performed in both mixing devices (0.6 M and 1.5 M urea) were combined, fit to two-exponentials and both observed rate constants plotted. Experiments performed in only the fastest mixing device (0.9 M and 1.2 M urea) were fit to two-exponentials but only the fastest observed rate constant is plotted. Experiments performed in the slower mixing device (1 M and 2 M urea) were fit to one-exponential. The 95% confidence intervals were calculated using the standard deviation of the fits and likely underestimate the true error.](image-url)
3.2.5 Structural information from monitoring fewer tryptophans

To obtain further structural information about the two kinetic phases, we reduced our number of tryptophan reporters and asked how the observed kinetics changed. We made conservative substitutions for two of the six tryptophans, W118F and W120F, creating a variant with just four tryptophans, all in one region of the protein (herein referred to as the 4Trp variant) (Figure 3.1)

Refolding of the 4Trp variant was monitored as previously using a 6 M to 0.6 M urea concentration jump in both mixers (Figure 3.5A). The observed kinetics look qualitatively very similar to wild type, indicating that the remaining region of tryptophans is the primary reporter on both kinetic phases observed in the wild-type experiments. Quantitatively, a global fit of the 4Trp data fits best to a three-state rather than a two-state model. The time constants associated with these three phases are: 90 +/- 1.1 microseconds, 300 +/- 23 microseconds and 2.3 +/- 0.11 milliseconds. The observation of three kinetic phases may indicate that the mutations have altered the early folding pathway, or else may indicate that the wild-type protein also folds in three steps but one step is obscured by signal from the additional tryptophans.

We performed an equilibrium urea titration with the 4Trp variant, monitored by average fluorescence lifetime (Figure 3.5B). The equilibrium data can be fit to a two-state model, with a ΔG_{unf} of 8.5 +/- 0.2 kcal/mol and an m-value of 2.19 +/- 0.04, indicating that these mutations destabilize the native state by only 1 kcal/mol. Comparing amplitudes between the equilibrium titration and the kinetic experiment suggests a possible small burst phase (the predicted signal of U at 0.6 M based on extrapolation of the unfolded baseline is 2.53 ns versus 2.63 ns predicted by extrapolation of the kinetics to zero time) and also suggests that the region containing the remaining four tryptophans is not natively folded in the I_{core} intermediate: the final signal of the millisecond kinetics is 2.34 ns while the signal of the native state in 0.6 M urea is 1.90 ns.
3.2.6 Structural information from a truncation mutant

The current model of the I\textsubscript{core} folding intermediate does not involve any structure in the region encompassing strands I-III and helix E in the native state (Figure 3.1). In the present kinetic experiment we do not have any probes in this region of the protein, and are not able to evaluate what role it may play in the early folding pathway. However, we have recently made a fragment of RNase H with the I-III/E regions removed and have shown that this fragment folds and may be a mimic of I\textsubscript{core} (Chapter 4). Thus, carrying out CF folding experiments with this fragment and comparing the results to wild type will allow us to determine whether the I-III/E regions impact the early folding pathway.

The first nine milliseconds of the folding of the fragment were monitored as previously, using a 5 M to 0.5 M urea concentration jump in both mixers (Figure 3.6A). We observe qualitatively very similar folding to that observed in the wild-type protein: two kinetic phases with very similar amplitude change. When the data is globally fit to a two exponential model, the time constants for the two phases are 80 +/- 4.9 microseconds and 1.11 +/- 0.32 milliseconds, faster than observed for wild type (even when an extrapolation to 0.5 M is considered). These observations fit with a model where the I-III/E regions of the protein are uninvolved during the early stages of folding, and are essentially “dead weight” as the rest of the protein folds.
Figure 3.6 The I$_{core}$ fragment. A. The fragment refolding into 0.5 M urea was monitored by average fluorescent lifetime and two kinetic phases are observed. (The data are combined from two ultrarapid mixing devices.) The extrapolated signal of the unfolded state (U) and the signal of the final equilibrium state (F) at these conditions are indicated. B. Equilibrium urea denaturation of the fragment monitored by average fluorescent lifetime. Inset: The same experiment monitored by total fluorescence intensity demonstrates that the fragment does in fact exhibit cooperative equilibrium unfolding when monitored by fluorescence, even though this is not visible when monitoring average fluorescent lifetime.

An equilibrium urea titration with the fragment, monitoring average fluorescence lifetime, reveals a steep dependence on urea concentration in the folded and unfolded baselines, preventing us from fitting these data to a two-state model (Figure 3.6B). (Global stability of the fragment is determined using circular dichroism in Chapter 4, and the urea titration monitored by total fluorescence intensity is shown in the inset of Figure 3.6B, demonstrating that the fragment does show folding cooperativity when monitored by fluorescence.) Nevertheless, the equilibrium average lifetime data are useful to make comparisons to the kinetic amplitudes. The final amplitude in the kinetic experiment (2.66 ns) matches closely with the equilibrium signal of the fragment at 0.5 M urea (2.60 ns), suggesting that the fragment folds to its equilibrium conformation within the time window of our experiment. Additionally, when the equilibrium signal of the unfolded state is extrapolated to 0.5 M urea, we observe a close correspondence to the zero-time extrapolated signal from the kinetic experiment: 2.64 ns versus 2.58 ns, respectively.
3.3 Discussion

Using a continuous-flow rapid mixing device, we directly observe the first several milliseconds of *E. coli* RNase H folding (starting from 50 microseconds after refolding is initiated) monitoring the change in tryptophan fluorescence. Two kinetic phases are clearly observed: one on the hundred-microsecond timescale and one an order of magnitude slower, on the millisecond timescale. These kinetics represent folding to intermediates, with final folding to the native state occurring on a slower timescale (seconds), beyond the range of these experiments. Experiments with the I53D variant demonstrate that the millisecond-timescale kinetic phase is the formation of the previously characterized I\(_{\text{core}}\) intermediate. Therefore, our data indicate we are observing the transient population of one new early intermediate (which we will term I\(_{\text{early}}\)) in addition to formation of I\(_{\text{core}}\).

3.3.1 Fitting a kinetic model to the data

It has been previously established that the I\(_{\text{core}}\) intermediate is obligate and on-pathway to the native state. There are three ways to incorporate I\(_{\text{early}}\) into this folding pathway without invoking any unjustified complexity: I\(_{\text{early}}\) is on-pathway to I\(_{\text{core}}\) (Model 1), I\(_{\text{early}}\) is off-pathway from the unfolded state (Model 2), or I\(_{\text{early}}\) is off-pathway from I\(_{\text{core}}\) (Model 3). These three models are depicted below (the associated microscopic rate constants are labeled; in all cases, formation of the new early intermediate, I\(_{\text{early}}\), is assigned \(k_1\), formation of I\(_{\text{core}}\) is assigned \(k_2\), and formation of the native state is assigned \(k_3\)).

![Kinetic model diagrams](image)

We want to determine which models are consistent with our data. Model 3 can be discounted because of the results for the I53D variant. We observe that the I53D variant still populates the hyper-fluorescent I\(_{\text{early}}\) intermediate, despite the fact that the I\(_{\text{core}}\) intermediate is never populated. This renders Model 3 very unlikely.

We performed fits with Models 1 and 2 to determine how consistent these models are with our measurements, and to determine possible values for the microscopic rate constants at each kinetic step. The models were fit to the urea-dependent rate constants (\(k_{\text{obs}}\)) determined in the
present experiments simultaneously with previously-determined $k_{obs}$ for folding to the native state. Figure 3.7 shows all of these $k_{obs}$ plotted as a function of urea. The models are fit using a scheme where the first two steps are treated as a pre-equilibrium relative to the third step (see section 3.4). This treatment is justified since there are three orders of magnitude separating the second kinetic process from the final kinetic process.

Models 1 and 2 are both consistent with the observed kinetics. Table 3.1 shows two different sets of parameters for Model 1 that result in a reasonable fit to all of the data, visualized in Figure 3.7A. Table 3.2 and Figure 3.7B display two parameter sets for Model 2. The fitting errors associated with parameters defining the first two kinetic steps are very large (larger than the values themselves) so they should only be interpreted as ballpark values. The tables also give the stabilities of the two intermediates and the native state predicted by the fit parameters; for all, the data are reasonably consistent with previously published values. We cannot distinguish between the two models with these data, though a comparison with previous work lends support to an on-pathway kinetic model (see section 3.3.2).

**Figure 3.7** Four-state models fit to wild type RNase H kinetic data. The urea-dependent kinetic data from Figure 3.3 and from Raschke et al. are plotted together. The former corresponds to the formation of early folding intermediates and the latter to the formation of the native state. A. The on-pathway kinetic model (Model 1) fits to the data with multiple sets of parameters (Table 1). Solid lines correspond to parameter set A and dashed lines to parameter set B. B. An off-pathway kinetic model (Model 2) also fits the data with multiple sets of parameters (Table 2). Solid lines correspond to parameter set C and dashed lines to parameter set D.
Table 3.1  Model 1 parameters that fit the observed data

<table>
<thead>
<tr>
<th>Parameter set A</th>
<th>Parameter set B</th>
<th>Published^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$ (sec$^{-1}$)</td>
<td>11,000</td>
<td>400</td>
</tr>
<tr>
<td>$k_1$ (sec$^{-1}$)</td>
<td>100</td>
<td>500</td>
</tr>
<tr>
<td>$k_2$ (sec$^{-1}$)</td>
<td>400</td>
<td>11,000</td>
</tr>
<tr>
<td>$k_2$ (sec$^{-1}$)</td>
<td>20</td>
<td>80</td>
</tr>
<tr>
<td>$k_3$ (sec$^{-1}$)</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>$k_3$ (sec$^{-1}$)</td>
<td>$1.9 \times 10^{-5}$</td>
<td>$1.6 \times 10^{-5}$</td>
</tr>
<tr>
<td>$m_1$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>$m_1$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>-0.5</td>
<td>-0.2</td>
</tr>
<tr>
<td>$m_2$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>$m_2$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>-0.6</td>
<td>-0.6</td>
</tr>
<tr>
<td>$m_3$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>$m_3$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>-0.43</td>
<td>-0.44</td>
</tr>
<tr>
<td>$\Delta G_{U-Iearly}$ (kcal/mol)</td>
<td>-3</td>
<td>0.1</td>
</tr>
<tr>
<td>$\Delta G_{U-Icore}$ (kcal/mol)</td>
<td>-5</td>
<td>-3</td>
</tr>
<tr>
<td>$\Delta G_{U-N}$ (kcal/mol)</td>
<td>-11</td>
<td>-9</td>
</tr>
</tbody>
</table>

Table 3.2  Model 2 parameters that fit the observed data

<table>
<thead>
<tr>
<th>Parameter set C</th>
<th>Parameter set D</th>
<th>Published^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$ (sec$^{-1}$)</td>
<td>5,000</td>
<td>100*</td>
</tr>
<tr>
<td>$k_1$ (sec$^{-1}$)</td>
<td>800</td>
<td>300</td>
</tr>
<tr>
<td>$k_2$ (sec$^{-1}$)</td>
<td>6,000</td>
<td>12,000</td>
</tr>
<tr>
<td>$k_2$ (sec$^{-1}$)</td>
<td>8</td>
<td>100</td>
</tr>
<tr>
<td>$k_3$ (sec$^{-1}$)</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>$k_3$ (sec$^{-1}$)</td>
<td>$1.6 \times 10^{-5}$</td>
<td>$1.5 \times 10^{-5}$</td>
</tr>
<tr>
<td>$m_1$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>0.1</td>
<td>0.05*</td>
</tr>
<tr>
<td>$m_1$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>0.2</td>
<td>-0.07</td>
</tr>
<tr>
<td>$m_2$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>$m_2$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>-1</td>
<td>-0.9</td>
</tr>
<tr>
<td>$m_3$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>$m_3$ (kcal mol$^{-1}$ M$^{-1}$)</td>
<td>-0.44</td>
<td>-0.45</td>
</tr>
<tr>
<td>$\Delta G_{U-Iearly}$ (kcal/mol)</td>
<td>-1</td>
<td>-0.3</td>
</tr>
<tr>
<td>$\Delta G_{U-Icore}$ (kcal/mol)</td>
<td>-4</td>
<td>-3</td>
</tr>
<tr>
<td>$\Delta G_{U-N}$ (kcal/mol)</td>
<td>-10</td>
<td>-9</td>
</tr>
</tbody>
</table>

*These parameters are very poorly constrained in the background of the rest of parameter set D.
The parameters for the Model 1 fits suggest two possible scenarios in the event that I\textsubscript{early} is on-pathway. Parameter set A corresponds to a scenario where the first kinetic step (from U to I\textsubscript{early}) is faster than the second kinetic step (from I\textsubscript{early} to I\textsubscript{core}) and where I\textsubscript{early} is a stable intermediate ($\Delta G_{\text{unf}} < 0$). In contrast, parameter set B defines a scenario where I\textsubscript{early} is an unstable intermediate ($\Delta G_{\text{unf}} > 0$) and the second kinetic step (from I\textsubscript{early} to I\textsubscript{core}) is faster than the first kinetic step (from U to I\textsubscript{early}). A comparison with previous work suggests parameter set A is more likely, if the on-pathway model is correct (see section 3.3.2). But even if parameter set B is not accurate, it is interesting that in these kinetic conditions it is possible to observe the transient population of an unstable intermediate.

### 3.3.2 Detailed structural model of the early folding pathway

What does the I\textsubscript{early} intermediate look like? Experiments performed with the 4Trp variant indicate that the remaining tryptophan region (orange in Figure 3.1) is the primary reporter on the observed kinetic phases. Therefore, this region undergoes the most dramatic change in environment during formation of I\textsubscript{early}, and is structured differently in I\textsubscript{early} and I\textsubscript{core}.

Any model for early folding must be reconciled with previous experiments. A high-resolution pulse-labeled hydrogen exchange experiment monitored by a new mass spectrometry technique (HX MS) was recently used to study the early folding pathway of RNase H\textsuperscript{7} (Chapter 2). This work observed that early folding occurred in three steps by sequential addition of native units of secondary structure: helix A and strand 4, then helix D and strand 5, then finally helices B and C, forming the previously characterized I\textsubscript{core} intermediate. In view of this work, our present results are most readily explained by the model depicted in Figure 3.8.

According to the scheme in Figure 3.8, I\textsubscript{early} is the same species as the on-pathway, HX MS A/4 intermediate. Non-native structure (orange) stabilizes the first native secondary structure elements (blue). This interaction would explain the mystery of why helix A is the first native helix to form, when it is not the helix with the highest intrinsic helicity (that would be helix E). It is also possible that the I\textsubscript{early} intermediate observed by fluorescence is actually an average of two states: with and without helix A docked onto the “orange” non-native structure. Either way, this model makes a case for non-native structure formation that is on-pathway, supporting the folding of native structural units.

In subsequent folding steps, the non-native structure unfolds and is replaced by native-like packing between helices A and D, and then finally helices B and C assemble to form I\textsubscript{core}. This model is consistent with our observation of three-state early folding in the 4Trp variant, suggesting that in the wild-type protein, the strand V tryptophans (green in Figure 3.1) obscure observation of the latter two folding events as discrete steps. Additionally, the time dependence of the structure formation observed by HX indicates that formation of I\textsubscript{early} does occur faster than subsequent folding steps, suggesting that Parameter set A is more accurate than Parameter set B (if, indeed, our proposed kinetic model is correct).

The scheme in Figure 3.8 also predicts that the I-III/E regions (gray) are not involved in the early folding pathway, and only become structured in the rate-limiting step to folding of the native state. This is supported by our observation that a fragment of the protein encompassing the
contiguous sequence from helix A to strand V folds in a manner very similar to the wild-type protein. Because the fragment folds and appears to mirror the early folding of wild-type RNase H, the fragment could be a good target for atomistic folding simulations, which are generally limited to small proteins. (Typical proteins studied by detailed folding simulations with explicit solvent are no larger than ~80 amino acids in size; the fragment is 81 amino acids.)

Overall, we have presented a detailed model for the early folding pathway of RNase H, at a timescale and size-scale that is currently accessible by atomistic simulations of folding. Most interestingly, our model makes the prediction that productive non-native structure stabilizes an on-pathway intermediate.

Figure 3.8 A detailed model for the early folding pathway of RNase H. This model incorporates the results presented here for the early folding monitored by fluorescence as well as the HX MS results presented in Chapter 2.
3.4 Materials and Methods

Materials
Cysteine-free *E. coli* RNase H was expressed and purified as described previously, with the following modifications. Heparin column fractions containing RNase H were diluted 1:1 with water and adjusted to a pH of 5.5. Precipitated protein was removed by centrifugation and the supernatant further purified using cation-exchange chromatography (HiTrap Capto S column). Purified protein was dialyzed against 50 mM ammonium bicarbonate, lyophilized and resuspended into the designated buffer before use.

The I53D variant was constructed previously. The 4Trp variant was generated using the QuikChange mutagenesis protocol, using pSM101 as a template, and then subcloned into a pET27 vector. The fragment was generated by subcloning from pSM101 and insertion into a pET27 vector.

The 4Trp variant was purified like the wild type. The I53D variant and the fragment express insolubly as inclusion bodies and were purified according to the following protocol. (Expression was performed in the Rosetta2(DE3)pLysS strain.) Cells were lysed by sonication and centrifuged. The pellet was washed by stirring for an hour in 10 mM Tris pH 8, 100 mM NaCl, 1 mM EDTA, 1% Triton x100, then centrifuged. The pellet was washed again by stirring for an hour in the same buffer but without Triton, then centrifuged. The pellet was solubilized by stirring for an hour in 7 M guanidinium chloride, 20 mM Tris pH 8, and 5 mM DTT. After centrifugation, protein in the soluble fraction was refolded dropwise into 20 mM NaOAc pH 5.5. Centrifugation was used to remove aggregated protein, and the supernatant was purified by cation-exchange chromatography (HiTrap Capto S). In the case of the fragment, S column fractions were further purified by size chromatography (Superdex 75, running buffer: 20 mM NaOAc, pH 5.5, 400 mM NaCl, 0.1 mM EDTA).

The purity and molecular mass of all the variants were verified by SDS-PAGE and mass spectrometry (data not shown).

Experimental conditions
For all experiments, the buffer conditions were 20 mM sodium acetate, pH 5.5, and 50 mM potassium chloride. All experiments were performed at room temperature. Final protein concentration after mixing was 1.2 μM for wild type and fragment, and 1.5 μM for the 4Trp variant.

Time-resolved tryptophan fluorescence
Details of the TCSPC apparatus equipped with a very similar microsecond CF mixer have been described. In this work, we use mixers that are fused silica microfluidic chips from Translume. The smaller one has custom, arrow-shaped geometry, with a 70 μm x 100 μm cross-section and the larger one has a T-shaped geometry with a 300 μm x 300 μm cross section. Both are 3 cm in length after the point of mixing. Prior to use, both mixers were covalently modified using silanated PEG in order to passivate the surfaces and prevent protein aggregation on the channel.
This was done by slowly flowing 2-[methoxy(polyethylenedioxy)propyl]trimethoxysilane, tech-90 through the mixer for over three hours followed by rinsing with methanol.

Flow to the microchannel mixer was provided by two syringe pumps (Isco) operating at a combined flow rate of 8 mL/min in the smaller mixer, corresponding to a ~50 μs dead time, and 15 mL/min in the larger mixer, corresponding to a ~2 ms dead time. Excitation at 293 nm with a repetition rate of 3.8 MHz was provided by the vertically polarized third harmonic of a Ti:sapphire laser. Typical excitation power was 500 uW. Tryptophan emission was measured using a 357 nm filter with 44 nm bandwidth from Semrock. The variation in excitation intensity along the flow channel was corrected using a standard (N-acetyl-L-tryptophanamide), as described previously. Separate instrument responses were recorded for each channel by recording of scattered light signal or by numerical deconvolution from the N-acetyl-L-tryptophanamide decay curve. The photon counts in the peak channel of a decay curve were typically 7 x 10^4 to 1 x 10^5. Approximately two to four decay curves were summed at each folding time in the CF experiment.

The quantum yield-weighted excited-state lifetime, τ_ave, was determined by calculating the center of mass for the time-resolved fluorescence distribution, based on the following equality:

$$\tau_{ave} = \frac{\sum_i a_i \tau_i^2}{\sum_i a_i \tau_i} = \frac{\sum_j I_j \tau_j}{\sum_j I_j}$$

where a_i and τ_i are the amplitude and lifetime of the ith phase in a Trp fluorescence decay curve, and I_j is the intensity at the t_j point in decay time.

**Fits to the kinetic data**
The kinetic data were globally fit to one- or two-exponentials according to

$$Signal = A \exp(-k_{obs} t) + C,$$

or

$$Signal = A_1 \exp(-k_{obs,1} t) + A_2 \exp(-k_{obs,2} t) + C,$$

to determine the rate constants for the observed kinetic phases (in the text, we report time constants of the kinetic phases: the time constant is the reciprocal of the observed rate constant). Fits were performed using Savuka.

**Equilibrium titrations**
Equilibrium unfolding experiments were set up using a titrator and measured using a homemade autosampler assembled from Zaber stages and a Hamilton Microlab 500 syringe pump. Signal was measured using the same TCSPC apparatus as used in the kinetic experiments. For comparison to kinetic data, signal from the equilibrium titration was adjusted to match equilibrium endpoints measured within the mixing device.
Fitting kinetic models to the Chevrons
The following equations were used to globally fit the Chevron data ($k_{\text{obs},1}$, $k_{\text{obs},2}$, and $k_{\text{obs},3}$ are represented by red, cyan and purple, respectively, in Figure 3.7).

The equations used for Model 1 are:

\[ k_{\text{obs},1} = \frac{(p + q)}{2}, \]
\[ k_{\text{obs},2} = \frac{(p - q)}{2}, \] and
\[ k_{\text{obs},3} = f_1 k_3 + k_3, \]
where
\[ p = k_1 + k_{-1} + k_2 + k_{-2}, \]
\[ q = \left[ \frac{p^2 - 4(k_1 k_2 + k_{-1} k_{-2} + k_3 k_2)}{2} \right]^{1/2}, \]
\[ f_1 = \frac{(k_2 / k_{-2}) (k_1 / k_{-1})} {1 + (k_1 / k_1) + (k_2 / k_{-2}) (k_1 / k_{-1})}, \]
and each parameter $k_i$ has the form given above, describing how it varies with denaturant concentration.

The equations used for Model 2 are:

\[ k_{\text{obs},1} = \frac{(p + q)}{2}, \]
\[ k_{\text{obs},2} = \frac{(p - q)}{2}, \] and
\[ k_{\text{obs},3} = f_1 k_3 + k_3, \]
where
\[ p = k_1 + k_{-1} + k_2 + k_{-2}, \]
\[ q = \left[ \frac{(k_1 + k_{-1} - k_2 - k_{-2})^2 + 4k_1 k_2}{2} \right]^{1/2}, \]
\[ f_1 = \frac{(k_2 / k_{-2})} {1 + (k_1 / k_1) + (k_2 / k_{-2})}, \]
and each parameter $k_i$ has the form given above, describing how it varies with denaturant concentration.

These equations are obtained from solutions to the rate laws for a two-step, reversible, unimolecular reaction and, separately, a one-step, reversible, unimolecular reaction with a pre-equilibrium, where the two-step reaction is either sequential\(^\text{14}\) (Model 1) or parallel (Model 2). The equations for the two-step reaction can be derived from a general treatment of reversible, unimolecular reactions.\(^\text{15}\) Fits were performed using IgorPro.

3.5 Acknowledgments
We thank K. Connell for help with initial experiments and purification of protein.

3.6 References


Chapter 4
Revisiting the early folding intermediate of RNase H: regions of the protein are closely packed before the rate-limiting step

4.1 Introduction

While many proteins populate intermediates early in the folding process, the role of such intermediates is still unclear. To what extent are they obligatory? And what is the nature of the barrier that allows for the transient build up of these states? Answering such questions requires detailed knowledge of the structure and dynamics of folding intermediates.

Many larger proteins (>100 amino acids) populate a partially-folded intermediate state within the burst-phase of stopped flow experiments (msecs) that is thought to be a molten globule: a compact structure with a high degree of secondary structure that lacks the tight tertiary interactions that are the hallmark of natively folded proteins. The molten globule has been proposed to be a general folding intermediate, where the exclusion of water and formation of close tertiary packing would be the last step in folding. To evaluate this claim, it’s important to clearly demonstrate the nature of the tertiary packing in early folding intermediates.

Molten globules were first observed in equilibrium studies where they can be populated under extreme solution conditions (i.e. low pH, in the presence of chemical denaturant, and/or by removal of a cofactor such as heme). Typical probes such as circular dichroism (CD), NMR, and ANS fluorescence revealed a conformational ensemble not well described by a native or unfolded state and therefore termed molten globule intermediates. Landmark hydrogen-deuterium exchange studies on the acid molten globule of apomyoglobin revealed secondary structure in a subset of helices from the native protein (A,G and H). Later, using pulse-labeling hydrogen exchange, these same helices were found to be protected in the transient early (msecs) folding intermediate under native condition. The protein ribonuclease H also populates a molten globule under acidic conditions and again the acid molten globule and transient kinetic intermediate have similar patterns of protection, indicating they form a similar subset of the native secondary structure. Consequently, the dominant hypothesis has been that the kinetic folding intermediates are themselves molten globules.

It is particularly difficult, however, to assay the structural details of these transient folding intermediates to conclusively determine whether the protected region is molten or well folded. Structural probes such as CD or fluorescence typically yield only very global information. While HD exchange is a powerful way to gain site-specific information, it only directly monitors secondary structure. Orthogonal techniques such as phi-value analysis and recently alkyl-proton exchange have been used to probe the role of specific side chain interactions in folding intermediates, but ultimately these techniques cannot conclusively evaluate the side chain packing in the structured region of a transient intermediate. In such experiments, evidence for molten-ness does not rule out other possible hypotheses, such as a closely packed region with rapid unfolding and refolding or with non-native tertiary structure.
To address the question of tertiary packing in folding intermediates, we have designed and interrogated different mimics of the kinetic folding intermediate of RNase H, an important protein folding model system with a well-characterized folding intermediate. The folding process of *E. coli* RNase H has been characterized using CD, HD exchange, mutagenesis, single molecule force spectroscopy (optical tweezers) and a computational prediction algorithm.\(^4\)\(^{12}\)\(^{15}\) All of these studies lead to a picture of a folding intermediate (termed I\(_{\text{core}}\)) populated before the rate-limiting barrier to folding, with native-like secondary structure in a contiguous region of sequence from helix A to strand V with the rest of the protein unfolded (Figure 4.1).

In addition to similarities to the acid state molten globule, other observations have been interpreted to support the hypothesis that the structured region of I\(_{\text{core}}\) is molten. Amino acid substitutions at residue 53 in the core that preserve hydrophobicity were found to destabilize the native state but not notably affect the stability of I\(_{\text{core}}\).\(^16\) This could be due to loose association of hydrophobic side chains in the intermediate, but could also be explained if I\(_{\text{core}}\) had non-native tertiary structure. Additionally, a single destabilizing mutation (I25A) was shown to cause extensive population of the I\(_{\text{core}}\) intermediate under equilibrium conditions.\(^17\) In the context of this mutation, and in the presence of \(\sim 2\) M urea, I\(_{\text{core}}\) is the dominant species at equilibrium but no peaks corresponding to the structured region of the intermediate could be observed in NMR HSQC experiments. This suggests that the structured region of I\(_{\text{core}}\) is very dynamic. The dynamics could be due to a poorly packed, heterogeneous structured region, or due to fast folding and unfolding of a closely packed structure.

Here, we probe the structured region of the RNase H I\(_{\text{core}}\) intermediate by creating equilibrium mimics of the folding intermediate under native conditions. We generate a fragment of *E. coli* RNase H comprised of the sequence from helix A to strand V (Figure 4.1). The fragment folds in isolation and mimics the known behavior of the kinetic intermediate. We find, however, that the fragment is not molten and forms a well-folded structure. Furthermore, in a second approach, by creating several destabilizing mutations in the (presumably) unfolded region of I\(_{\text{core}}\), we populate the intermediate under native conditions in the context of the full-length protein and find that the structured region is again well folded. Finally, we show that our results are compatible with the previous data that were interpreted to support the hypothesis of a molten structure.

In sum, our data imply that the structured region of the *E. coli* RNase H kinetic intermediate (I\(_{\text{core}}\)) is well folded, with closely packed side chains. It remains possible that the region lacking protection in the kinetic HD exchange experiments, previously assumed to be unstructured in the folding intermediate, may itself have molten globule character, and/or that the protein forms a molten intermediate before this well-folded intermediate. This work suggests that the rate-limiting step in the folding of *E. coli* RNase H is not the packing down of existing tertiary interactions in the I\(_{\text{core}}\) intermediate, but more likely related to assembly of the rest of the protein or a reorganization of the folded region. These results also suggest that the evidence for molten-ness in the folding intermediates of other proteins is worth re-visiting.
4.2 Results

4.2.1 A peptide mimic of the transient folding intermediate, I\textsubscript{core}, from RNase H
Previous pulse-labeling hydrogen exchange, computational and protein engineering studies have identified residues 43-122 as the structured region of I\textsubscript{core}. The sequence of \textit{E. coli} RNase H comprising this region was cloned, expressed and purified from \textit{E. coli}. (In this study, the reference WT \textit{E. coli} RNase H sequence is a cysteine-free variant.) The resulting fragment is soluble upon refolding from inclusion bodies, and is shown to adopt a stable fold; the circular dichroism (CD) spectrum appears to show double minima close to 208 nm and 222 nm, consistent with helix formation (Figure 4.2A). Equilibrium urea-induced denaturation of the fragment, monitored by the CD signal at 222 nm, shows a cooperative folding transition (Figure 4.2B) with a $\Delta G_{\text{unf}}$ of 3.0 +/- 0.30 kcal/mol, and an $m$-value of 1.16 +/- 0.047 kcal/mol/M (95% confidence intervals are based on the average of three experiments) using a two-state assumption and linear-extrapolation model.\textsuperscript{19} The stability of transient folding intermediate I\textsubscript{core} is approximately 3.5 kcal/mol (with an $m$-value of 1.2 kcal/mol/M) as determined by stopped-flow CD studies fit to an on-pathway three–state model.\textsuperscript{4} Thus, both the CD spectrum and the energetics are consistent with that measured for the transient kinetic folding intermediate.\textsuperscript{4}

4.2.2 The fragment folds as a monomer with a $K_d$ for dimerization of $\sim$75 $\mu$M
In order to confirm that the folded fragment is a monomer under our conditions, we used equilibrium analytical ultracentrifugation (AUC). The AUC data fit well to a monomer-dimer model, with a dissociation constant of $\sim$75 $\mu$M (Figure 4.3). Based on this $K_d$, the above CD studies (3-4 $\mu$M) contain $\sim$95% monomer, indicating we are effectively measuring properties of the monomer. Additionally, since dimerization will be attenuated by urea, the stability reported from the urea denaturation study should approximate the monomeric stability. To underscore this, the melt was repeated with a 10-fold higher protein concentration (36 $\mu$M), where $\sim$75% monomer is expected in the 0M urea sample (Figure 4.2B), and yielded a $\Delta G$ and $m$-value of 3.1 kcal/mol and 1.19 kcal/mol/M, within error of the result at 3.6 $\mu$M protein concentration.
Figure 4.2 The RNase H I<sub>core</sub> fragment is well folded. 
A. CD spectrum of the fragment measured at 4 μM protein concentration. B. Representative equilibrium denaturation curves of the fragment monitored by CD and normalized to fraction folded, at 3.6 μM (black) and 36 μM (red). C. Overlay of <sup>1</sup>H-<sup>15</sup>N HSQC spectra of the fragment measured at 25 μM (black) and 200 μM (red). The red peaks were shifted slightly to the right for easier comparison.
Figure 4.3 Equilibrium analytical ultracentrifugation of the RNase H I\textsubscript{core} fragment. Absorbance at 250 nm versus radius for a 50 μM sample at rotor speeds of 24,500 rpm (blue), 30,000 rpm (gray), and 37,000 rpm (red) was globally fit to a monomer-dimer equilibrium together with equivalent data from 25 μM and 100 μM samples.

4.2.3 **NMR suggests the interior of the fragment is closely packed**

We used \textsuperscript{1}H-\textsuperscript{15}N heteronuclear single quantum coherence (HSQC) NMR studies to evaluate the structural heterogeneity of the folded fragment. These experiments require a minimum concentration near our \(K_d\) of dimerization, so to distinguish the monomer signal from dimer signal, spectra were measured at two different protein concentrations: 200 μM and 25 μM, predicted to contain ~50% monomer and ~80% monomer, respectively (Figure 4.2C).

Both HSQC spectra exhibit sharp, well-dispersed peaks. In each spectrum, many more peaks are observed than expected if only one molecular species were present, but the peaks can be categorized into two groups that can be attributed to monomer and dimer. In the 200 μM sample, both sets of peaks have similar intensities; however, in the 25 μM sample, the monomer peaks dominate and the dimer peaks persist only at low intensities. The fact that all peaks are sharp and well dispersed indicates that both the monomer and dimer are unique well-folded structures. Fortuitously, dimerization kinetics are slow enough to allow clear resolution of peaks from both species.

4.2.4 **Mutational analysis of the fragment**

We used mutational analysis to make a direct comparison between the fragment mimic and the kinetic intermediate. In the native state of RNase H, helix A forms buried hydrophobic contacts with helix D – interactions that are also predicted to be present in the kinetic intermediate (Figure 4.1). The effect of various mutations at residue 53 on the folding and stability of RNase H have been previously characterized.\textsuperscript{16} We chose to construct and analyze a subset of these mutations in the fragment – I53A, I53V, and I53L – and determine their effect on the stability of the fragment. All three single-site variants fold and show similar CD spectra as the wild-type fragment (Figure 4.4A). All were evaluated by equilibrium urea-induced denaturation and show cooperative unfolding transitions (Figure 4.4B). The denaturation curves for fragments I53L and I53V were fit using the standard two-state assumption. For the fragment I53A, which was too destabilized to observe a folded baseline, an approximate \(C_m\) was determined by assuming the protein is 100% folded at 0 M denaturant. Then, assuming it shares the same m-value as the other fragment variants, an approximate stability was also determined. Thus, this value should be
considered an upper estimate on the stability. The results are summarized in Table 4.1 and demonstrate that the fragment mirrors the transient kinetic intermediate in that the effects of these mutations in the fragment are consistent with their effects in the transient kinetic intermediate (keep in mind that there is substantial error in the stabilities calculated from burst phase intermediates, so the trends are more meaningful than the exact numbers).

Table 4.1  Stabilities of the $I_{\text{core}}$ fragments compared to the kinetic intermediates

<table>
<thead>
<tr>
<th></th>
<th>Fragment equilibrium denaturation</th>
<th>Kinetic burst phase intermediate$^{16}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\Delta G_{\text{unf}}$ (kcal/mol)</td>
<td>$m_{\text{unf}}$ (kcal/mol/M)</td>
</tr>
<tr>
<td>WT</td>
<td>3.0</td>
<td>1.2</td>
</tr>
<tr>
<td>I53A</td>
<td>$\leq 1.4^*$</td>
<td>-</td>
</tr>
<tr>
<td>I53L</td>
<td>2.7</td>
<td>1.2</td>
</tr>
<tr>
<td>I53V</td>
<td>3.3</td>
<td>1.2</td>
</tr>
</tbody>
</table>

*Value calculated by estimating the $C_m$ and assuming a WT $m$-value

4.2.5  NMR analysis of a single-site variant suggests that the variant is also well folded

As seen in the transient kinetic intermediate, I53L has no significant effect on fragment stability, even though I53L destabilizes the native state by 1.5 kcal/mol. (The fragment is also not destabilized at all by I53V but neither is the native state.) One interpretation of the I53L result could be a loosely packed core without steric constraints in the fragment. To gain structural insight into the effect of the I53L mutation on side-chain packing in the core of the fragment, we determined the HSQC spectrum of this variant.

The spectrum of the I53L variant shows many well-dispersed peaks at similar (but not identical) chemical shifts to the wild-type fragment (Figure 4.4C). Measuring spectra at two protein concentrations again demonstrated two sets of peaks with relative intensities dependent on protein concentration. These data suggest that the I53L variant has a similar monomer-dimer equilibrium as the wild-type fragment, and that both the monomer and dimer are well folded.

4.2.6  The urea dependence of the HSQC spectrum

Previous NMR studies on the single-site RNase H variant I25A, which preferentially populates the $I_{\text{core}}$ intermediate in the presence of ~2 M urea, show no folded peaks corresponding to the structured region of $I_{\text{core}}$. To investigate whether or not these results could be consistent with our fragment mimic, we measured HSQC spectra of the wild-type fragment as a function of urea (Figure 4.5).

The HSQC spectrum of the wild-type fragment in 1 M urea looks essentially identical to the spectrum measured in 0 M urea. However, in the 2 M urea spectrum, only unfolded state peaks are visible, and only at very low signal intensity. Based on our CD studies, we expect greater than 80% of the molecules are folded in these conditions (Figure 4.2B). Therefore, these data suggest that at 2 M urea the rate of fragment unfolding increases sufficiently to cause significant exchange broadening of the folded peaks. Thus, the lack of folded $I_{\text{core}}$ peaks in the I25A variant NMR spectrum does not have to arise from a molten heterogeneous structure.
Figure 4.4 The effect of mutations at residue 53 in the RNase H I<sub>core</sub> fragment. A. CD spectra of I<sub>core</sub> fragment single-site variants I53L (orange circles) and I53V (white circles) compared to WT (line), measured at ~3.5 μM protein concentration. B. Representative equilibrium denaturation curves of the I<sub>core</sub> fragment variants I53A (black circles), I53L (orange circles) and I53V (white circles) normalized to fraction folded, compared to WT (line alone). Two-state fits for the I53L and I53V variants are shown. All experiments were measured at ~3.5 μM. C. Overlay of HSQC spectra of the I53L fragment variant measured at 50 μM (black) and 420 μM (orange). The orange peaks were shifted slightly to the right for easier comparison.
4.2.7 Full-length mimics of I\textsubscript{core} made via selective destabilization of the native state

While the fragment mimic allows us to evaluate properties of the structured region of the transient intermediate, it lacks the region predicted to be unstructured. To evaluate the role of this region, we took a second approach to making an equilibrium mimic of I\textsubscript{core}, starting from the full-length protein.\textsuperscript{20}

Single-site amino acid changes (to glycine or alanine) were generated in the unstructured region of the I\textsubscript{core} model. For each single point mutant, the equilibrium urea-induced denaturation was monitored by its CD signal at 222 nm, and fit as previously to calculate its effect on the stability of the protein.\textsuperscript{20} To create full-length intermediate mimics, these point mutations were combined such that the native state would be less stable than the intermediate state assuming additive destabilizations (Figure 4.6A inset). Two different full-length mimics were created using different combinations of mutations. The full-length mimic 1 (FL1) contains the mutations I25A/R27A/S36G and full-length mimic 2 (FL2) contains F8A/S12G/R27A/S36G/E135G (Figure 4.1). Additionally, a hexahistidine tag was added to the C-terminus to allow purification from unwanted proteolysis products.

Both full-length mimics display CD spectra notably different from the native protein (Figure 4.6A). Like the fragments, the spectra appear to have double minima close to 222nm and 208nm, whereas the full-length native protein has a single minimum around 215 nm. Equilibrium urea-induced denaturation was monitored by CD signal at 222nm and showed cooperative transitions with a $\Delta G_{unf}$ of 3.2 +/- 0.36 kcal/mol and m-value of 1.23 +/- 0.099 kcal/mol/M for FL1 and $\Delta G_{unf}$ of 3.3 +/- 0.46 kcal/mol and m-value of 1.2 +/- 0.17 kcal/mol/M for FL2 (95% confidence intervals are based on the average of four experiments). The m-value of 1.2 kcal/mol/M is much lower than the m-value of the native protein, 2.0 kcal/mol/M, indicating that the full-length mimics bury much less solvent exposed surface area than the native states. These values mimic those measured in the equilibrium denaturation of the fragment discussed above.
Figure 4.6 Full-length $I_{\text{core}}$ mimics made by selective destabilization of the native state are well folded. A. CD spectra of the FL1 mimic (black circles) and FL2 mimic (white circles) measured at ~1.5 μM protein concentration, compared to full-length WT RNase H (line). Inset: The native state is selectively destabilized so that $I_{\text{core}}$ becomes the ground state. B. Equilibrium denaturation curves of the FL1 mimic (black circles) and FL2 mimic (white circles) measured at ~2 μM protein concentration, normalized to fraction folded, compared to full-length WT RNase H (line alone). Two-state fits are shown. C. HSQC spectrum of the FL1 mimic measured at ~100 μM.
4.2.8 The full-length mimics are structurally similar to the fragment
The HSQC spectra of the full-length mimics show well-dispersed peaks and very similar
chemical shifts as the fragment HSQC spectrum, as illustrated by the spectrum of FL1 (Figure
4.6C). Specifically, the fragment HSQC peaks match the chemical shifts of the peaks identified
previously as belonging to the monomer, indicating that the structured regions of the full-length
mimics are similar to the fragment mimic. The most notable difference is the presence of a large
number of collapsed peaks along the hydrogen axis in the full-length mimic spectra as compared
to the fragment spectrum. We interpret these collapsed peaks as corresponding to the unfolded
region of the full-length mimics. Interestingly, peaks at chemical shifts matching those of the
fragment dimer are visible at low intensity in the full-length mimic spectra.

4.2.9 The intermediate mimics do not bind and fluoresce with ANS
Our last experiment to evaluate the tertiary packing of the Icore mimics was to monitor binding to
1-anilino-8-naphthalene sulfonic acid (ANS) by fluorescence, a traditional hallmark of molten
globules. Under acid-state conditions (i.e. the previously identified molten globule), the WT
protein shows a large increase in ANS fluorescence consistent with it being a molten globule. In
contrast, under native conditions, we observe that the fragment and full-length mimics all exhibit
low ANS fluorescence similar to native WT, supporting that they adopt a well-folded structure
(Figure 4.7). These experiments were performed using 2 μM protein and 50 μM ANS. Molten
globule-like fluorescence with the fragment and full-length mimics (as well as I25A17) could be
induced by increasing the ANS concentration to 500 μM. However, under this condition of high
ANS concentration, observed fluorescence intensity correlates with precipitation of a protein-
ANS aggregate and is not a true measure of monomer tertiary packing.

Figure 4.7 The Icore mimics do not bind the hydrophobic dye ANS. Fluorescence emission spectra of ANS
in the presence of (A) the Icore fragment (red) and (B) the Icore full-length mimics FL1 (solid red) and FL2
(dashed red). For comparison, spectra in the presence of full-length WT RNase H at pH 1.2 (top black
curve), pH 5.5 (middle black curve) and 6 M urea (bottom black curve) are shown in both plots. All
experiments were performed with 2 μM protein and 50 μM ANS. The fluorescence emission of ANS in
buffer alone has been subtracted from the data.
4.3 Discussion

In this work, we set out to investigate the claim that the RNase H transient folding intermediate (I_{core}) is a molten globule. Tertiary packing is very difficult to probe in transient intermediates so we used protein engineering to create mimics of I_{core} that can be studied with equilibrium tools. In contrast to the long-standing hypothesis, the I_{core} mimics exhibit all the hallmarks of well-folded proteins, including cooperative equilibrium denaturation and good dispersion in NMR HSQC experiments. Additionally, we demonstrate that previous evidence used to support a molten globule hypothesis is compatible with well-folded structure. We conclude that the RNase H I_{core} folding intermediate is likely not a molten globule.

It is possible that RNase H populates a molten globule state prior to the well-folded intermediate state, but we know that the well-folded state is formed prior to the rate-limiting step to folding because of recent kinetic studies we have performed with the fragment mimic (Chapter 3). We used an ultrarapid mixing technique to observe folding of the fragment directly. Not only do we see that the fragment folding looks very similar to the early folding of the full-length wild-type protein, we observe that the final signal in the folding kinetics of the fragment matches its equilibrium signal. This shows that the fragment is reaching its closely packed equilibrium state within a few milliseconds and is a good mimic for the intermediate populated before the rate-limiting barrier.

What is the nature of the barrier if not the packing down of tertiary interactions? Our mutagenesis results suggest that the rate-limiting barrier could be related to a reorganization of the folded core. The I53L mutation destabilizes the native state, but not the transient I_{core} intermediate nor the fragment mimic, even though the fragment is well folded. This may indicate that the folded region reorganizes upon formation of the native state, i.e. the folded core is not quite native-like in the I_{core} intermediate and therefore mutations have a different effect on stability than in the native state. Non-native interactions in on-pathway kinetic intermediates have been observed for other model systems.\textsuperscript{21-23} Alternatively, it is also possible that the folded region of I_{core} is natively packed in the wild-type protein, but can repack to accommodate the I53L mutation with no penalty in stability – the impact on stability occurs only when the rest of the protein assembles onto the repacked core to form the native state. In this situation, the folding barrier may simply be the unfavorable conformational entropy for assembly of the beta sheet. More work will be needed to distinguish between these scenarios.

Generally, this work extends previous observations that acid molten globules are not accurate models for transient folding intermediates under native conditions. Extensive experiments on apomyoglobin have shown that the hydrogen-exchange protection in the acid state molten globule follows a slightly different pattern and has a different stability distribution than observed in the kinetic intermediate.\textsuperscript{24,25} In a related protein, apoleghemoglobin, there are substantial structural differences between the protection in the acid state molten globule and the kinetic intermediate.\textsuperscript{26} Additionally, while significant conformational heterogeneity has been observed in the kinetic intermediate of apomyoglobin, such observations are not incompatible with a region of structure that is well folded.\textsuperscript{23,25,27}
Lastly, there is precedent for proposing that RNase H has a folding intermediate with closely packed side chains. In a study from Yawen Bai's lab, the authors made a fragment mimic of the *T. thermophilus* RNase H folding intermediate – albeit using a slightly different model for the structured region of the folding intermediate – and observed that their fragment is well folded.\textsuperscript{28} This fragment did not have the complication of dimerization and thus they were able to solve the NMR structure.\textsuperscript{28} They observe that the overall topology of the fragment looks completely native-like, though a close analysis reveals deviations in side chain packing compared to the crystal structure of wild-type *T. thermophilus* RNase H.\textsuperscript{29} However, it is not clear whether these deviations are significant or indeed whether this fragment truly mimics the intermediate prior to the rate-limiting barrier to folding.

Overall, the RNase H data suggest that many observed protein folding intermediates may have regions of closely packed structure, contrary to a widespread assumption of molten globule kinetic intermediates. Experimental data interpreted to support molten-ness should be examined carefully for other possible interpretations.

### 4.4 Materials and Methods

*Construction and purification of RNase H variants*

Cloning and purification of the wild type fragment was described in Chapter 3. Single-site variants of the fragment were created using Quikchange, and expressed and purified similarly (from inclusion bodies). Cloning, expression and purification of the full-length intermediate mimics were described previously.\textsuperscript{20} Additionally, in some cases the full-length intermediate mimics were purified using the same protocol as used for the fragment, but with the addition of a Ni column prior to the Capto S column (and no EDTA or DTT present), and without the sizing column. After purification, proteins were dialyzed into experiment buffer conditions: 20 mM sodium acetate, pH 5.5, and 50 mM potassium chloride.

Expression of \textsuperscript{15}N labeled protein was done by initial growth in LB with a switch to M9 media with \textsuperscript{15}NH\textsubscript{4}Cl as the sole nitrogen source prior to induction for three hours by IPTG. The labeling efficiency was \textasciitilde90\% as evaluated by mass spectrometry.

*CD experiments*

All CD experiments were measured on an Aviv 410 CD spectropolarimeter. All experiments were performed in a cuvette with a 1-cm pathlength, except the urea melts with 10x higher protein concentration were performed in a 1-mm pathlength cuvette. For each fragment and full-length intermediate mimic, at least one equilibrium denaturation melt was performed after incubating individual samples overnight. The rest of the melts were performed with shorter incubations, and some were performed using a titrator with a five minute equilibration time between samples. The results were consistent at all incubation times. For all, signal at 222 nm was averaged over 60 seconds for each sample. Protein concentrations were determined based on the extinction coefficient, calculated according to the number of Trp and Tyr residues.\textsuperscript{30}
Equilibrium analytical ultracentrifugation

Sedimentation equilibrium experiments on the wild type fragment were performed with a Beckman XL-I analytical ultracentrifuge using the absorbance optics system to detect the protein at 250 nm. Protein samples were at 25, 50, and 100 μM in 20 mM sodium acetate (pH 5.5) and 50 mM potassium chloride; sample volume was 110 μL in a 6-channel centerpiece. Data were acquired at 25 °C with a radial step size of 0.001 cm and 5 replicates, at speeds of 24,500, 30,000 and 37,000 revolutions per minute. Absorbance scans were collected at four-hour intervals and successive scans were compared graphically using SEDFIT v. 14.1 to ensure that the sample reached equilibrium.

Global analysis of the nine data sets was accomplished using SEDPHAT v. 10.58d. Data were fit using mass conservation with rotor stretch, using an extinction coefficient of 13,940 M⁻¹ cm⁻¹ (which was measured and calibrated against the theoretical extinction coefficient at 280 nm, calculated according to the number of Trp and Tyr residues⁴⁰) and buffer density was calculated using SEDNTERP v. 1.09. Fitting the data to a single species model only yielded good fits (judged by examination of the residuals) if the mass of the monomer was allowed to vary from the theoretical mass of 9.581 kD to 13-14 kD (increasing with higher protein concentration). The data was best fit by a monomer-dimer equilibrium model, with a Kd of 75 μM. The exact value of the dissociation constant is only slightly sensitive (+/- 10 μM) to parameters such as the boundaries of the fitted data and the theoretical extinction coefficient.

HSQC spectra

Two-dimensional $^1$H-$^{15}$N HSQCs were recorded on a Bruker Avance II 900 MHz spectrometer equipped with a TCI cryoprobe at 25 °C. For samples with protein concentration ~100 μM or above, 16 or 32 scans were collected, and for samples with protein concentration less than 100 μM, 64 or 128 scans were collected, in all cases with 1024 points in the direct dimension and 256 points in the indirect dimension. The data were processed and viewed using either mNOVA or NMRpipe and CARA.

ANS binding

Samples containing 50 μM ANS in buffer with and without 2 μM protein were prepared and equilibrated for several hours. Fluorescence emission spectra were collected from 430 to 650 nm with an excitation wavelength of 370 nm. The spectrum of ANS in buffer alone was subtracted from the samples containing ANS and protein.

4.5 Acknowledgments

We thank J. Pelton for guidance in NMR data collection and analysis and D. Wemmer for additional help with interpretation of NMR data. We thank the QB3 MacroLab (and C. Jeans) for help with equilibrium analytical ultracentrifugation data collection and analysis, as well as K. Fleming and C. Kimberlin for additional advising about the AUC experiments.
4.6 References

Chapter 5
Comparing RNase H homologs from *E. coli* and *T. thermophilus*: Differences on the energy landscape and what they reveal about the rate-limiting barrier to folding

5.1 Introduction

A fundamental goal in biology is to understand how the amino acid sequence encodes all aspects of a protein’s structure and dynamics, referred to as a protein’s energy landscape. Progress on this complex undertaking is enabled by characterizing folding pathways for many protein folding model systems. However, it can be difficult to determine principles for sequence-encoded properties when drawing conclusions between very different model systems. Therefore, a powerful tool for understanding the subtleties of how information is encoded by sequence is to compare homologous proteins. Such studies yield important insights into what aspects of an energy landscape are dictated by native state topology versus can be modulated by sequence.

There are a number of examples of homologous proteins with the same native topology having differences in their folding pathways. Examples include large differences in folding rates,\(^2\) the presence or absence of folding intermediates,\(^3\) and differences in structural details of folding intermediates.\(^4,5\) But it is particularly interesting to relate such differences to biophysical properties known to be of functional importance. For example, mesophilic and thermophilic protein homologs have very different stabilities with respect to temperature, a property very important for their function. How do differences in their energy landscapes relate to thermal stability? Interesting models for such a comparison are the RNase H homologs from *E. coli* and from the thermophilic bacteria *Thermus thermophilus*.

*E. coli* RNase H (ecRNH) and *T. thermophilus* RNase H (ttRNH) have very similar native state topology\(^6,7\) (Figure 5.1), but different thermodynamic properties: ttRNH is more stable than ecRNH across a wide range of temperatures, and ttRNH is active at a temperature (65°C) at which half of ecRNH molecules are unfolded (\(\Delta G_{\text{unf}} = 0\)).\(^8\) To investigate the source of ttRNH thermostability, both proteins have been studied using native-state (equilibrium) hydrogen exchange, and also kinetic experiments monitored by spectroscopy and pulse-labeling hydrogen exchange (all at 25°C).\(^9-12\) (In both cases, the “wild-type” protein is a cysteine-free version of true wild type.\(^8,13,14\) All references to these proteins in this text refer to the cysteine-free versions.) The results indicate that both proteins have the same distribution of stability across their structures, and both populate a similar partially folded intermediate before the rate-limiting barrier to folding. This intermediate contains secondary structure in the contiguous region of the protein from helix A to strand V, and is referred to as I\(_{\text{core}}\) (Figure 5.1).
One big difference observed between the two homologs is that ttRNH has a lower $\Delta C_p$ (change in heat capacity between the unfolded and folded state). This serves to broaden the stability curve ($\Delta G_{\text{unf}}$ as a function of temperature), increasing the stability of ttRNH at all temperatures. It can be inferred that a lower $\Delta C_p$ is due to residual structure in the unfolded state of ttRNH, which was confirmed using protein engineering studies and observed directly using calorimetry.$^{15,16}$ However, this result does not explain all of the difference in thermostability, and it is still very possible that there are other major differences in the energy landscapes of the two proteins.

One possible difference between the ecRNH and ttRNH energy landscapes was suggested by a 2008 paper from the lab of Yawen Bai at the National Cancer Institute.$^{17}$ Their paper proposes that the ttRNH folding intermediate includes strand I in the structured region, which would be a big difference compared to the ecRNH folding intermediate. (In the present work, we will refer to the intermediate populated immediately prior to the rate-limiting barrier to folding as “the” folding intermediate, though we know there are other intermediates on the folding pathway.$^{18}$) In their 2008 work, the authors noted that the ttRNH pulse-labeled hydrogen exchange data (monitored by NMR) includes only one probe in strand I, which shows partial protection 14 milliseconds after refolding is initiated.$^{12}$ Additionally, there are three strand I probes in the native-state hydrogen exchange experiment, at least one of which shows protection similar to other residues in the alpha helical core.$^{10}$ They proposed a model where strand I is structured in the ttRNH folding intermediate, and they made a fragment mimic of the structured region of this
model (to include strand I they made a non-native junction where strand I is directly linked to the N-terminus of helix A, see Figure 5.1). We will call this intermediate model I$_{\text{core+1}}$. Their I$_{\text{core+1}}$ fragment is well folded, and they solved its NMR structure, showing that it looks like a subset of the native state structure.

Strand I together with the region from helix A to strand V is well folded on its own, indicating that I$_{\text{core+1}}$ is a stable partially folded state of ttRNH. But the question is: is I$_{\text{core+1}}$ truly the folding intermediate or is it populated elsewhere on the energy landscape? Additionally, is this a universal RNase H intermediate, i.e. does ecRNH populate an I$_{\text{core+1}}$ intermediate anywhere on its energy landscape? In this work, we investigate these questions by making and characterizing fragment mimics of the putative I$_{\text{core}}$ and I$_{\text{core+1}}$ intermediates for both homologs (Figure 5.1).

By studying protein fragments, we determine that ttRNH populates both an I$_{\text{core}}$ and an I$_{\text{core+1}}$ intermediate on its energy landscape, whereas ecRNH populates only the I$_{\text{core}}$ intermediate. The ttRNH I$_{\text{core}}$ intermediate is likely well packed, similar to the ttRNH I$_{\text{core+1}}$ intermediate (and also similar to the ecRNH I$_{\text{core}}$ intermediate investigated in Chapter 4). Additionally, we perform preliminary work to determine whether I$_{\text{core}}$ or I$_{\text{core+1}}$ is the ttRNH folding intermediate. The answer hinges on whether or not strand I is structured prior to the rate-limiting barrier to folding. We use mutagenesis and also FRET to address this question. This work is ongoing, but preliminary results suggest that strand I does in fact gain structure prior to the rate-limiting barrier to folding, changing our understanding of the folding barrier for the RNase H model system.

5.2 Results

5.2.1 Truncation mutants of T. thermophilus RNase H reveal multiple partially folded states on the energy landscape

Work from Yawen Bai’s lab demonstrated that ttRNH populates the I$_{\text{core+1}}$ intermediate. To determine whether ttRNH also populates the I$_{\text{core}}$ intermediate, we created a fragment consisting of residues 42 to 122 (in this work, numbering for the ttRNH sequence is based on an alignment with the ecRNH sequence). For comparison, we also re-created the Bai lab’s I$_{\text{core+1}}$ fragment but without a C-terminal hexahistidine tag that was used in the original study. (A hexahistidine tag was still used for purification, but was removed via a Tev cleavage, leaving a non-native glycine-histidine at the N-terminus of our I$_{\text{core+1}}$ fragment.) Both fragments were expressed and purified from E. coli.

Circular dichroism (CD) studies indicate that both fragments are folded. The CD spectra show two minima near 208 nm and 222 nm, consistent with helix formation, and equilibrium urea-induced denaturation monitored by the CD signal at 222 nm show cooperative folding transitions (Figure 5.2A,B). The denaturation curves can be fit using a two-state assumption and linear extrapolation model, yielding a $\Delta G_{\text{unf}}$ of 2.5 kcal/mol and an m-value of 1.0 kcal/mol/M for the I$_{\text{core}}$ fragment and a $\Delta G_{\text{unf}}$ of 6.1 kcal/mol and an m-value of 1.3 kcal/mol/M for the I$_{\text{core+1}}$ fragment. The I$_{\text{core}}$ fragment folds on its own, and has a very different stability than when strand I
is present (in the I$_{\text{core}+1}$ fragment) indicating that the I$_{\text{core}}$ partially folded intermediate is also present on the ttRNH energy landscape.

![Graphs and diagrams related to Figure 5.2 T. thermophilus RNase H populates two partially-folded states.](image)

**Figure 5.2** T. thermophilus RNase H populates two partially-folded states. **A.** CD spectra of the ttRNH I$_{\text{core}}$ (black) and I$_{\text{core}+1}$ (white) fragments measured at ~3.5 μM protein concentration. **B.** Equilibrium denaturation curves of the ttRNH I$_{\text{core}}$ fragment at 3.6 μM (black) and 36 μM (red), and the I$_{\text{core}+1}$ fragment (white), monitored by CD and normalized to fraction folded. **C.** Overlay of $^1$H-$^{15}$N HSQC spectra of the ttRNH I$_{\text{core}}$ fragment measured at 54 μM (black) and 430 μM (red). The red peaks were shifted slightly to the right for easier comparison. **D.** Wire representation of the NMR structure of the hexahistidine-tagged ttRNH I$_{\text{core}+1}$ fragment. Highlighted with spheres are the residues whose $^1$H-$^{15}$N chemical shifts overlap with peaks present in the ttRNH I$_{\text{core}}$ fragment HSQC spectrum at 54 μM protein concentration.
5.2.2 The tTRNH I$_{\text{core}}$ fragment folds as a monomer with a K$_d$ for dimerization of ~150 μM

The ecRNH I$_{\text{core}}$ fragment dimerizes with a K$_d$ of ~75 μM (Chapter 4), so we carried out equilibrium analytical ultracentrifugation (AUC) with the tTRNH I$_{\text{core}}$ fragment to assess self-association. (It was previously determined that the I$_{\text{core+1}}$ fragment does not dimerize.) The AUC data was fit well by a monomer-dimer model, with a dissociation constant of ~150 μM (Figure 5.3). Based on this K$_d$, the CD samples (at 3–4 μM) are estimated to contain ~97% monomer. Therefore, we are effectively measuring properties of the monomer.

Additionally, we believe that the stability reported from the urea denaturation should accurately reflect monomer stability, since dimerization will be weakened by urea. To support this, the melt was repeated with a 10-fold higher protein concentration, at 36 μM, where ~85% monomer is expected in the 0 M urea sample (Figure 5.2B). This melt was fit as before, and yielded a ΔG$_{\text{unf}}$ of 2.3 kcal/mol and an m-value of 0.9 kcal/mol/M, consistent with results from the lower-concentration experiment.

5.2.3 NMR suggests the interior of the tTRNH I$_{\text{core}}$ fragment is closely packed

The I$_{\text{core+1}}$ fragment was shown to be a well-folded structure using NMR, and in fact its structure was solved. We wanted to determine whether the I$_{\text{core}}$ fragment is also well folded, especially since it is less stable than the larger I$_{\text{core+1}}$ fragment. We measured the $^1$H-$^{15}$N heteronuclear single quantum coherence (HSQC) NMR spectrum of the I$_{\text{core}}$ fragment at 430 μM protein (~50% monomer expected). The spectrum shows peak dispersion, but also a cluster of significantly broadened peaks in the center. To distinguish the signature of the monomer from the dimer, the HSQC spectrum was measured at a lower protein concentration (54 μM, 80% monomer expected). We observe that when the protein concentration is decreased, the broad, poorly-dispersed peaks disappear and the spectrum is dominated by sharp, well-dispersed peaks (Figure 5.2C). This suggests that the monomer is well folded but that the kinetics of dimerization are on the right timescale to cause exchange broadening in peaks associated with the dimer.

A comparison between the I$_{\text{core}}$ fragment spectrum and the assigned spectrum of the hexahistidine-tagged I$_{\text{core+1}}$ fragment shows some overlapping chemical shifts. In particular,
residues furthest from strand I retain the same chemical shifts in the absence of strand I, whereas residues near strand I in the 6His-I_{core+1} fragment are shifted when strand I has been removed in the I_{core} fragment (Figure 5.2D). This suggests that the region furthest from strand I retains the same structure in both fragments. The movement of peaks in the region closest to strand I could be interpreted as a change in structure or just a change in chemical environment due to the absence of strand I. Additionally, based on analysis by electrospray ionization mass spectrometry, approximately half of the I_{core} fragment has had its N-terminal methionine cleaved in vivo, which may result in peak doubling for residues nearest to the N-terminus (data not shown).

**Figure 5.4** *E. coli* RNase H only populates the I_{core} intermediate. **A.** CD spectrum of the ecRNH I_{core+1} fragment (black circles) compared to the ecRNH I_{core} fragment (line) measured at ~3.5 μM protein concentration. **B.** Equilibrium denaturation of the ecRNH I_{core+1} fragment (black circles) compared to the ecRNH I_{core} fragment (line only), measured at ~3.5 μM protein concentration, normalized to fraction folded. **C.** Overlay of HSQC spectra of the ecRNH I_{core+1} fragment (red) and the ecRNH I_{core} fragment (black) measured at ~100 μM.
5.2.4 EcRNH does not populate the \( I_{\text{core+1}} \) intermediate

Having established that ttRNH populates two similar, well-folded intermediates, we wondered if this was true for the *E. coli* homolog. We constructed the analogous \( I_{\text{core+1}} \) fragment for the ecRNH sequence (residues 1-20, 42-122). The fragment was expressed and purified from *E. coli*.

Analysis by circular dichroism illustrates that this fragment folds to a helical structure, with minima near 208 nm and 222 nm (Figure 5.4A). Equilibrium denaturation with urea shows a cooperative transition, which can be fit as previously, yielding \( \Delta G_{\text{unf}} \) of 3.2 kcal/mol and an m-value of 1.1 kcal/mol/M (Figure 5.4B). These results are within error of the equilibrium denaturation result for the ecRNH \( I_{\text{core}} \) fragment in Chapter 4. This is in sharp contrast to the large stability difference observed for ttRNH \( I_{\text{core}} \) and \( I_{\text{core+1}} \) fragments, suggesting that the strand I residues are not structured in the ecRNH \( I_{\text{core+1}} \) fragment and therefore ecRNH does not populate an \( I_{\text{core+1}} \) intermediate. This is further supported by the full-length ecRNH \( I_{\text{core}} \) mimics in Chapter 4, where the presence or absence of a mutation in strand I did not affect the stability of the mimics.

Additional evidence is provided by the NMR HSQC spectrum of the ecRNH \( I_{\text{core+1}} \) fragment. Comparison to the spectrum of the ecRNH \( I_{\text{core}} \) fragment (originally characterized in Chapter 4) shows that the majority of peaks are identical between the two spectra (Figure 5.4C). The major difference is found in the \( I_{\text{core+1}} \) fragment spectrum: a set of high intensity peaks with minimal dispersion along the 'H axis (centered at a ppm of ~8.3). This is the exact signature we expect to observe if strand I is unstructured in the \( I_{\text{core+1}} \) fragment.

5.2.5 Mutational analysis to locate ttRNH \( I_{\text{core+1}} \) on the folding pathway

The fragment results indicate that ttRNH populates both \( I_{\text{core}} \) and \( I_{\text{core+1}} \) intermediates, but where do they lie on the energy landscape? Previous pulse-labeled hydrogen exchange experiments show that \( I_{\text{core}} \) must be populated prior to the rate-limiting step to folding, but what about \( I_{\text{core+1}} \)? The answer lies in whether or not strand I is structured prior to the rate-limiting step to folding.

To address this, we performed a mutational analysis (\( \phi \)-value analysis) on the full-length, wild-type protein to determine where on the folding pathway particular residues become structured.\(^{20}\) In this analysis, a single residue is mutated, the effect of the mutation on (un)folding kinetics and equilibrium stability is determined, and the results are used to quantify the effect of the mutation on different regions of the folding pathway according to the following equations:

\[
\phi_1 = \frac{\Delta \Delta G_{\text{U1}}}{\Delta \Delta G_{\text{UN}}}, \text{ and }
\]

\[
\phi_{\text{TS}} = 1 - \frac{\Delta \Delta G_{\text{TS-N}}}{\Delta \Delta G_{\text{UN}}}.
\]

where \( \Delta \Delta G_{\text{U1}} \) is the difference in stability of the folding intermediate between the wild-type protein and the single-site variant (determined using folding kinetics), \( \Delta \Delta G_{\text{TS-N}} \) is the difference in transition state stability (determined using unfolding kinetics), and \( \Delta \Delta G_{\text{UN}} \) is the difference in equilibrium stability. (\( \phi_1 \) is the intermediate \( \phi \)-value and \( \phi_{\text{TS}} \) is the transition state \( \phi \)-value.) For both cases, the \( \phi \)-values are expected to fall between 0 and 1, quantifying the degree to which the residue is structured in that region of the folding pathway.
Ideally we would determine $\phi_I$ to answer our question about the ttRNH folding intermediate. Unfortunately, we are limited to calculating $\phi_{TS}$ because we cannot quantitatively analyze folding kinetics, only unfolding kinetics: wild-type ttRNH exhibits two kinetic phases when folding, making a mutational analysis in the folding direction untenable.\textsuperscript{12} Though a high $\phi_{TS}$ (near 1, meaning a residue is structured in the transition state) does not distinguish whether or not that residue is also structured in the folding intermediate, a low $\phi_{TS}$ (near 0, residue not structured in TS) can possibly be interpreted as absence of structure in the folding intermediate, since it is unlikely for structure to be lost going from the kinetic intermediate to the transition state.

Transition state $\phi$-values were determined for three strand I residues – L7, T9 and A13 – as well as for strand II residue L26 and helix A residue I53 (Figure 5.1). In all cases the residues were mutated to alanine, except for A13 which was changed to a glycine. To avoid too large of an extrapolation (since unfolding kinetics were determined in high concentrations of urea) $\phi$-values were calculated at 6 M urea. The results are tabulated in Table 5.1, and partial Chevrons are shown in Figure 5.5. Though only unfolding kinetics were used to calculate $\phi$-values, a limited range of folding kinetics are plotted as a qualitative comparison for the range of urea concentrations where the wild-type protein exhibits single exponential folding kinetics.

The results of the $\phi$-value analysis are consistent with the hypothesis that I$_{core}$ is the ttRNH folding intermediate. Residue 53 has a very high $\phi_{TS}$, consistent with expectation since it is in the structured region of both I$_{core}$ and I$_{core+1}$. The strand II residue L26 has a medium $\phi_{TS}$, despite pulse-labeling hydrogen exchange revealing little protection in strand II in the kinetic intermediate.\textsuperscript{12} This suggests that residue 26 gains partial structure in the transition state. Results for the three variants with mutations in strand I are consistent with strand I being unstructured in the folding intermediate. Residue 13 has a low $\phi_{TS}$, a clear result that suggests A13 is not structured in the transition state or the preceding folding intermediate. Residue 9 has a medium $\phi_{TS}$, which could be interpreted similarly as for residue 26 where structure is not present in the folding intermediate but is gained in the transition state. Finally, the third strand I residue, L7, has a $\phi_{TS}$ greater than 1, a dramatic result indicating that this residue is structurally very important in the transition state, much more so than in the native state.

From the L7 $\phi_{TS}$ we cannot determine whether or not L7 is structured in the ttRNH folding intermediate. However, these results are consistent with studies of ecRNH where the I7 side chain is not structured in the folding intermediate but is the lynchpin of the transition state (i.e. ecRNH residue 7 also has a $\phi_{TS} > 1$).\textsuperscript{21} Further, ecRNH is observed to have a polarized transition state, where the $\phi_{TS}$ values get smaller as spatial distance away from residue 7 increases (spatial distance based on position in the native structure). The same result is observed here for ttRNH (i.e. residue 13 has a lower $\phi_{TS}$ value than residues 9 and 26).
The effect of mutations on *T. thermophilus* RNase H folding and unfolding kinetics. **A.** The T9A and A13G mutations in strand I primarily affect the unfolding rate. **B.** The I53A mutation in helix A primarily affects the folding rate. **C.** The L7A mutation in strand I and the L26A mutation in strand II significantly affect both folding and unfolding. The L7A mutation effect is particularly rare because it slows unfolding.

**Figure 5.5**
Table 5.1  Unfolding kinetics, stability and φ-values for ttRNH single-site variants

<table>
<thead>
<tr>
<th></th>
<th>At 0 M urea</th>
<th>At 6 M urea</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ln (k_{unf}, sec^{-1})</td>
<td>ΔG_{UN} (kcal/mol)*</td>
</tr>
<tr>
<td>WT</td>
<td>-10.2</td>
<td>12.8</td>
</tr>
<tr>
<td>L7A</td>
<td>-11.6</td>
<td>11.1</td>
</tr>
<tr>
<td>T9A</td>
<td>-9.4</td>
<td>12.5</td>
</tr>
<tr>
<td>A13G</td>
<td>-8.8</td>
<td>11.9</td>
</tr>
<tr>
<td>L26A</td>
<td>-9.3</td>
<td>11.3</td>
</tr>
<tr>
<td>I53A</td>
<td>-9.7</td>
<td>10.7</td>
</tr>
</tbody>
</table>

*ΔG_{UN} were calculated using a fixed m-value of 2.0 kcal/mol/M

5.2.6  Strand I docking on the alpha helical core monitored directly by FRET

The φ-value analysis – particularly for residue 13 – suggests that strand I is not structured in the folding intermediate. However, the interpretation of φ-values requires many assumptions, such as that all structure formed along the folding pathway is native-like. Stronger evidence would be to directly monitor strand I contacting the alpha helical core; if strand I is not structured in the folding intermediate, we should observe that the strand I/helical core interaction occurs on the same timescale as global folding to the native state. To test this prediction, we used FRET to monitor the distance between strand I and the helical core during folding.

Our FRET experiment was performed using intrinsic tryptophan fluorescence as the fluorescence donor and a thionitrobenzoate (TNB) label in strand I to quench fluorescence. Wild-type ttRNH has five tryptophans: four in one group in the alpha helical core and one in strand II (Figure 5.6A). We made a conservative substitution (W22Y) in order to remove the tryptophan in strand II so that all the tryptophan fluorescence is present in the alpha helical core. Additionally a cysteine was engineered at position 6 (in place of a wild-type alanine) in order to attach the TNB label on the N-terminal end of strand I.

The TNB-A6C/W22Y construct was evaluated using CD to determine that the native structure and stability had not been too perturbed by the mutations and the TNB label. The CD spectrum has a very similar shape as wild type, and equilibrium urea-denaturation yielded a ΔG_{unf} of 10.8 kcal/mol compared to the wild-type ΔG_{unf} of 12.8 kcal/mol, indicating that the protein likely has the same overall native structure and is only slightly destabilized (Figure 5.6B,C). Additionally, we confirmed that the TNB label quenches fluorescence in the native structure by measuring fluorescence emission spectra in both folded and unfolded conditions (Figure 5.6D). The difference in signal is strongest near 360 nm, at which wavelength the unlabeled A6C/W22Y ttRNH protein shows almost no difference in fluorescence between folded and unfolded. Therefore, this is an ideal wavelength at which to monitor folding kinetics by FRET.

Re-folding of TNB-A6C/W22Y was monitored by both circular dichroism (a global probe of structure) and fluorescence at 360 nm (specifically monitoring formation of a strand I contact). Results are shown in Figure 5.6E,F. Observed kinetics are very different using the two probes of structure. Global folding, monitored by CD, occurs on a timescale of approximately one minute,
while strand I docking, monitored by fluorescence at 360 nm, occurs on a timescale of tens of milliseconds. (For this comparison, experiments were performed at 3 M final urea concentration so that kinetics by fluorescence were slow enough to observe clearly.) Also, as expected based on studies of the wild-type protein, a large burst phase is observed by CD. These preliminary results suggest that the I{sub core} intermediate may form first, but strand I docks onto the alpha helical core prior to the rate-limiting barrier to folding, indicating that I{sub core+1} may be the folding intermediate.

Figure 5.6 Using FRET to monitor the formation of strand I contacts in the folding of *T. thermophilus* RNase H. A. Ribbon diagram of *T. thermophilus* RNase H native structure with tryptophan side chains shown in red stick. (Residue W22 is shown in light red.) Residue A6 is shown in blue stick. B. CD spectrum of TNB-labeled A6C/W22Y ttRNH. C. Equilibrium denaturation of TNB-labeled A6C/W22Y ttRNH, normalized to fraction folded. D. Fluorescence emission spectra of TNB-labeled A6C/W22Y ttRNH in folded and unfolded conditions (black filled circles and open circles, respectively) and unlabeled A6C/W22Y ttRNH in folded and unfolded conditions (gray filled circles and open circles, respectively). The excitation wavelength is 280 nm. E. Folding of TNB-labeled A6C/W22Y ttRNH into 3 M urea conditions monitored by CD. F. Folding of TNB-labeled A6C/W22Y ttRNH into 3 M urea conditions monitored by fluorescence at 360 nm (with excitation at 280 nm).
5.3 Discussion

In this work, we used fragment mimics of partially folded intermediate models to determine that *T. thermophilus* RNase H (ttRNH) populates the \( I_{\text{core}} \) and \( I_{\text{core}+1} \) intermediates, but *E. coli* RNase H (ecRNH) only populates an \( I_{\text{core}} \) intermediate. This is one of the clearest differences that has been observed between the energy landscapes of these homologous proteins. We then sought to determine whether the ttRNH \( I_{\text{core}} \) intermediate is the folding intermediate, analogous to ecRNH, (with the \( I_{\text{core}+1} \) intermediate populated elsewhere on the energy landscape) or whether the ttRNH \( I_{\text{core}+1} \) intermediate is the folding intermediate. If it was the latter case, this would be a particularly dramatic difference between the two homologs. Determining where the ttRNH \( I_{\text{core}+1} \) intermediate is populated will shed light on the nature of the rate-limiting barrier for RNase H folding.

There are multiple hypotheses as to what is the most difficult step in RNase H folding, i.e. what is the nature of the rate-limiting barrier. Historically, it had been thought that the barrier might be the packing down or tightening of molten tertiary interactions present in the folding intermediate. However, from work in Chapter 4 (as well as this chapter) it appears that the structured region in the RNase H intermediate is likely well folded, invalidating this hypothesis. Another possibility is that the barrier is assembly of the beta sheet onto the well-folded alpha helical core (overcoming unfavorable conformational entropy). In this model, contact between strand I and the helical core would accompany global folding and the \( I_{\text{core}+1} \) intermediate would not be the ttRNH folding intermediate. Another model is that the barrier is a rearrangement of the structured region of the folding intermediate. In this model, the structured region of the folding intermediate has non-native structure that must be resolved. (There is precedent for non-native side chain interactions being present in productive, on-pathway intermediates).\(^{22-24}\)

Demonstrating whether \( I_{\text{core}+1} \) is the ttRNH folding intermediate could help us determine which of these two models – assembly of the beta sheet or rearrangement of the helical core – may be the folding barrier for RNase H. We took two approaches to addressing this question. We used \( \phi \)-value analysis and a FRET experiment to evaluate whether strand I is structured in the folding intermediate of ttRNH. Preliminary results are contradictory. The \( \phi \)-value analysis is consistent with a model where no strand I contacts are made until the transition state; in particular, it clearly predicts that residue 13, at the C-terminal end of strand I, is not structured prior to the rate-limiting step. However, directly monitoring strand I contacting the helical core using FRET observes that this contact occurs much faster than global folding, suggesting that \( I_{\text{core}+1} \) may be the ttRNH folding intermediate.

More work is needed to make sense of these contradictory data. However, a model where rearrangement of non-native structure in the folding intermediate is the barrier may explain all of the observations. This model is consistent with formation of strand I/helical core contacts prior to the rate-limiting barrier, as observed by FRET. Additionally, this model directly predicts non-native structure in the intermediate, and non-native structure in a folding pathway is the Achilles’ heel of \( \phi \)-value analysis, rendering the \( \phi \)-values un-interpretable.\(^{20}\) However, one problem with this model is that we might expect to learn about the non-native structure in the
folding intermediate by looking at the NMR structure of the ttRNH I\textsubscript{core+1} fragment,\textsuperscript{17} but the structure of this fragment appears to have generally native-like packing (though it is possible that the fragment represents a partially-folded species lower in energy than the I\textsubscript{core+1} folding intermediate).

A final possible model that would explain all of the data is if regions of the beta sheet sequence make contact with the helical core prior to the rate-limiting barrier. In this model, the beta sheet would loosely assemble prior to the barrier, and the rate-limiting step would be packing down of these interactions. (This is analogous to the first model discussed above, but the molten contacts are in the region of the folding intermediate unprotected from hydrogen exchange, rather than the protected region.) This could explain why we observe both early contact of strand I, as well as that residue 13 is unstructured in the transition state: it may be that only part of strand I forms an early contact with the helical core. (Also, this model is not mutually exclusive with rearrangement of non-native structure in the helical core.)

One thing is clear: more work is needed to resolve the nature of the RNase H folding barrier. Despite many years of work on this protein folding model system, mystery still surrounds important features of its energy landscape. This underscores what a complex undertaking it is to determine how all aspects of a protein’s behavior and function are encoded by the amino acid sequence.

5.4 Materials and Methods

Construction of RNase H variants
The ttRNH I\textsubscript{core} and I\textsubscript{core+1} fragments and the ecRNH I\textsubscript{core+1} fragment were subcloned from pJH109 and pSM101, respectively. For the cloning of the I\textsubscript{core+1} fragments, the N-terminal region of the sequence containing strand I was encoded on a primer. These constructs were all cloned into a pET27 vector, except for the ttRNH I\textsubscript{core+1} fragment which was cloned into a modified pET28 vector with a TEV-cleavable hexahistidine tag.

All of the full-length ttRNH variants were created in the pSV272 vector, which contains a TEV-cleavable hexahistidine-tagged MBP fusion at the N-terminus of the RNase H gene. Mutations of the cysteine-free “wild type” gene were created using Quikchange.

Protein expression and purification
Expression and purification of the fragments is as described for the ecRNH I\textsubscript{core} fragment in Chapter 3, with the following exception. The ttRNH I\textsubscript{core+1} fragment was purified from both the soluble and insoluble fractions, in both cases using a Ni column and then TEV cleavage prior to purification with a Capto S column.

Expression of \textsuperscript{15}N labeled protein was done by initial growth in LB with a switch to M9 media with \textsuperscript{15}NH\textsubscript{4}Cl as the sole nitrogen source prior to induction for three hours by IPTG. The labeling efficiency was \textasciitilde90\% as evaluated by mass spectrometry.
Expression of the MBP-ttRNH fusion proteins was performed as previously except using Rosetta2(DE3)pLysS cells and kanamycin. For purification, cell pellets were lysed by sonication, cell debris was removed by centrifugation and the soluble fraction was first purified using a Ni column. After overnight TEV-cleavage, fractions containing ttRNH are purified in a final step using a Heparin column (which removes the free MBP very efficiently).

**TNB labeling**
Labeling of the single cysteine in the A6C/W22Y ttRNH variant was accomplished by incubating protein in 6 M GdmCl, 20 mM Tris, pH 8.3, and 250 μM EDTA with a 50x molar excess of DTNB at room temperature for 30 minutes (the protein had been prepared with a PD-10 column to remove reducing agent). Another PD-10 column was used to exchange the labeled protein into unfolding buffer for kinetic experiments (6 M urea, 20 mM sodium acetate, pH 5.5, and 50 mM potassium chloride) and simultaneously remove free dye. Mass spectrometry indicated that the labeling efficiency was ~100%.

**Equilibrium experiments**
CD experiments, equilibrium ultracentrifugation experiments, and NMR experiments were performed as described in Chapter 4, with the following exception: all samples for CD melts of full-length ttRNH variants were incubated overnight.

All experiments were performed at room temperature, with the following buffer conditions: 20 mM sodium acetate, pH 5.5, and 50 mM potassium chloride.

**Kinetic experiments**
Kinetics monitored by CD at 222 nm were performed on an Aviv 410 CD spectropolarimeter with a 1-cm pathlength cuvette. The dead time for the manual mixing experiment was ~15 seconds.

Kinetics monitored by fluorescence were performed on a Biologic SFM-400 stopped flow instrument. Kinetics were initiated using a 10-fold dilution into the FC-15 cuvette, with a 250 uL shot volume and 7.5 mL/sec total flow speed. In these conditions the dead time is 5 milliseconds. Data were recorded with a sampling rate of 50 us.

**5.5 Acknowledgments**

We thank J. Pelton for guidance in NMR data collection and analysis and D. Wemmer for additional help with interpretation of NMR data. We thank C. Jeans, K. Fleming and C. Kimberlin for help with equilibrium analytical ultracentrifugation data collection and analysis. We thank S. Kumar Jha for guidance with the FRET experiment.

**5.6 References**


Appendix A
Preliminary characterization of the earliest events in the folding of *T. thermophilus* RNase H and the ancestral RNase H Anc1 using ultrarapid mixing and intrinsic fluorescence

A.1 Introduction

How much is a protein’s energy landscape dictated by the topology of the native state versus how much is mutable by sequence? In particular, among protein homologs that fold to the same native state structure, to what extent do their folding pathways look the same? Ribonuclease H (RNase H) is an excellent model system for asking such questions. *E. coli* and *T. thermophilus* RNases H are one of the best-studied homolog pairs in the field of protein folding. Further, in recent work, phylogenetic tree reconstruction was used to infer possible ancient RNase H homolog sequences, including the putative last common ancestor between *E. coli* and *T. thermophilus* RNases H (called Anc1). These ancestor sequences were synthesized and studied experimentally. (This technique is called ancestral protein resurrection, or APR.) The new proteins obtained through APR provide a great opportunity for exploring the sequence space of RNase H.

Folding of the *E. coli*, *T. thermophilus*, and Anc1 RNase H homologs has been monitored and characterized starting from 18 milliseconds after folding is initiated. Additionally, the first several milliseconds of *E. coli* RNase H (ecRNH) folding have been characterized in this dissertation. However, the earliest events in the folding of *T. thermophilus* RNase H (ttRNH) and Anc1 RNase H (ancRNH) have not yet been observed. In this preliminary work, we use a microfluidic continuous flow mixer and time-correlated single photon counting (TC-SPC) to monitor early events in the folding of ttRNH and ancRNH by tryptophan fluorescence. (This work is analogous to experiments performed in Chapter 3 with ecRNH.) These experiments mark the first time that folding on the millisecond timescale has been observed for either of these proteins.

The preliminary data indicate that both proteins fold to an intermediate on the same timescale as is observed for ecRNH. Additionally, like the *E. coli* homolog, both ttRNH and ancRNH appear to form an earlier intermediate in the ~2 ms dead time of the kinetic experiment. Lastly, the three homologs have different amplitude changes associated with their millisecond folding monitored by fluorescence. It is not clear whether this is due to differences in the folding pathways or due to differences in the number and placement of the tryptophans being monitored (Figure A.1A).
A.2 Preliminary results

Folding of *T. thermophilus* RNase H (ttRNH) was initiated using a 8 M to 0.8 M urea concentration jump in a microsecond-resolved continuous flow (CF) mixing device with a ~2 ms dead time. Folding was monitored by the change in average fluorescence lifetime of the tryptophans, determined using time-correlated single photon counting (TC-SPC). TC-SPC data without instrument response deconvolution were averaged to obtain a raw average fluorescence lifetime at each point in folding time. (In this first-pass analysis, the “average lifetime” includes the time between the emission of the excitation pulse and its arrival at the sample – about 12 ns – so is not the true fluorescence lifetime. Therefore, these values can only be directly compared to other data measured with the exact same set-up.) Plotting the raw average fluorescence lifetime versus folding time shows that we are observing folding on the millisecond timescale (Figure A.1C). Fitting the kinetics to a single exponential yields a lifetime of ~6 milliseconds for ttRNH folding to an intermediate species. Folding to the native state occurs on a timescale of seconds, beyond the range of this experiment.

Folding of Anc1 RNase H (ancRNH) was initiated using a 7.5 M to 0.75 M urea concentration jump. Kinetics were monitored and analyzed as above. Plotting the raw average lifetime versus folding time, we again observe kinetics on the millisecond timescale (Figure A.1D). Fitting the kinetics to a single exponential yields a lifetime of ~3 milliseconds for ancRNH folding to an intermediate species.

For comparison, folding of *E. coli* RNase H (ecRNH) into 0.6 M urea is displayed in Figure A.1B. This data is from Chapter 3, but has been analyzed in the same way as above. Fitting to a single exponential yields a lifetime of ~3 milliseconds for ecRNH folding to an intermediate.

For all three homologs, equilibrium data were measured to provide comparison to the kinetic data (Figure A.1B-D). The raw average fluorescence lifetime of each natively-folded protein was determined in the same conditions as the kinetic experiment (0.6 M urea for ecRNH, 0.8 M urea for ttRNH and 0.75 M urea for ancRNH). Additionally, an estimate of the raw average fluorescence lifetime of the unfolded state was obtained by measuring the fluorescence of the unfolded state in conditions of high urea and then using linear extrapolation back to the low urea condition of the kinetic experiment. (Fluorescence of unfolded ecRNH was measured at 6 M, 7 M and 8 M urea; fluorescence of unfolded ttRNH was measured at 8 M and 9.5 M; fluorescence of unfolded ancRNH was measured at 7.5 M, 8.5 M and 9.5 M.) In all cases, the extrapolated unfolded state fluorescence lifetime is well below the fluorescence lifetime predicted by extrapolating the observed kinetics back to zero time, indicating a burst phase. In the case of ecRNH, this “burst phase” was observed directly in a faster kinetic experiment in Chapter 3.
Figure A.1 Millisecond folding of three RNase H homologs monitored by tryptophan fluorescence. A. Crystal structures of *E. coli* RNase H\(^4\) (left) and *T. thermophilus* RNase H\(^5\) (right) shown in ribbon with tryptophan residues highlighted in stick. The Anc1 protein shares all tryptophans shown in green: the conserved cluster of four tryptophans in the C/D region, the strand II tryptophan from the *T. thermophilus* homolog and one of the strand V tryptophans from the *E. coli* homolog. B-D. Refolding kinetics monitored by tryptophan fluorescence for (B) *E. coli* RNase H at 0.6 M urea, (C) *T. thermophilus* RNase H at 0.8 M urea, and (D) Anc1, the putative last common ancestor of *E. coli* and *T. thermophilus* RNases H, at 0.75 M urea. For all, gray dots indicate the fluorescence lifetime of the native state (N) at equilibrium under the conditions of the kinetic experiment. Additionally, for all, the fluorescence lifetime of the unfolded state (U) in the conditions of the kinetic experiment was estimated by measuring equilibrium unfolded state fluorescence lifetimes at multiple high concentrations of urea and using a linear extrapolation.
A.3 Materials and Methods

**Materials**

Cysteine-free *E. coli* and cysteine-free Anc1 RNases H were expressed and purified as described for *E. coli* RNase H in Chapter 3. Cysteine-free *T. thermophilus* RNase H was expressed and purified as described in Chapter 5, with a TEV-cleavable MBP domain.

**Experimental conditions**

Experimental conditions are as described in Chapter 3. Final protein concentration after mixing is 1.2 μM for *E. coli* and Anc1 RNases H and 1.3 μM for *T. thermophilus* RNase H.

**Time-resolved tryptophan fluorescence**

Kinetic experiments were performed as described in Chapter 3, though using only the larger 300 μm x 300 μm mixer with the ~2 ms dead time.

Data in Chapter 3 is presented after the instrument response has been deconvoluted. Here, data has been analyzed by calculating the center of mass of the raw TC-SPC data, without instrument response deconvolution.

**Equilibrium samples**

Fluorescence of the equilibrium samples was measured inside the mixer to enable direct comparison of the raw average fluorescence lifetime of the equilibrium samples with the kinetic data. Equilibrium samples were slowly flowed through the microfluidic mixer at 0.5 mL/min; fluorescence was measured across the channel and then averaged across channel position. Samples were incubated for the appropriate amount of time before measuring, to ensure that equilibrium had been reached.

A.4 References