PROTEIN AND RNA FOLDING: FROM BULK TOWARDS HIGH THROUGHPUT SINGLE MOLECULE EXPERIMENTS

BY

KRISHNARJUN SARKAR

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemistry in the Graduate College of the University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:
Professor Martin Gruebele, Chair & Director of Research
Professor J. Douglas McDonald
Professor James M. Lisy
Assistant Professor Yann R. Chemla
Abstract

The folding energy landscapes of RNA and proteins have been studied using experiments and molecular dynamics (MD) simulations. Firstly, the folding of an RNA hairpin has been looked at with the fluorescent probe 2-aminopurine. The effect of stem and loop dynamics have been separately analyzed. Motivated by MD simulations, a global four state energy landscape for both the stem and loop mutants was found to be an ideal fit to the observed experimental data. Secondly, kinetic and thermodynamic predictions based on very long MD simulations of a variant of the β-sheet WW domain protein Fip35, were experimentally verified to be true. This provided an atomistic detail understanding of the folding landscape and will act as a key benchmark in unraveling the protein folding problem. Thirdly, protein and RNA interactions were looked at with MD simulations of the ribosomal signatures S4 and h16 along with fluorescence experiments to characterize the nature of the binding interaction between them. The results point toward a fly casting mechanism whereby the presence of the h16 helps the S4 signature to adopt its structure. Finally, a design and implementation of a high throughput time correlated single photon counting experiment is presented. Drops of ~10 μm diameter are optically trapped and interrogated by a femtosecond probe laser beam. Fluorescence photons emitted by the sample inside the drop are collected in all 4π steradians with polarization sensitivity. The instrument performs close to the single molecule limit, illustrated by the detection of ~100 Cerulean molecules on average in each drop. Further modifications that would possibly allow high throughput single molecule detection and its corresponding implications are discussed.
To my family
Acknowledgements

First of all, I would like to thank my advisor Martin Gruebele for being a source of constant encouragement in difficult times. His inexhaustible patience and unending source of insight were invaluable to me. He has been an incredible resource as an advisor and mentor. Along with Martin I would like to thank Professor Douglas McDonald for his extensive support during the construction of the high throughput single molecule instrument. His experience with electronics, data acquisition and overall experimental know-how was invaluable.

I thank the members of my thesis and preliminary examination committee for taking the time from their busy lives to improve my thesis with suggestions and advice. I am thankful to Martin, Jim Lisy (Jim), J Douglas McDonald (Doug) and Yann Chemla for serving on my thesis committee. I would also like to thank Klaus Schulten for serving on my preliminary exam committee.

During my years here as a graduate student I have had the privilege of working with many diversely talented individuals. I would like to thank Seung Joong Kim for working with me on the single molecule project. His effortless ability to simulate any physical system has mesmerized me. I thank Feng Liu for invaluable experience on the T-Jump instrument and molecular biology protocols. I thank Praveen Chowdary for insightful discussions in various topics on physical chemistry. His dedication to science is supreme. I thank two wonderful visiting summer undergraduate students, Konrad Meister and Duc Nguyen for their support in the RNA folding experiments. I thank Charles Dumont, Daniel Weidinger, Sharlene Denos, Simon Ebbinghaus, Apratim Dhar and Eduardo Berrios for very thoughtful discussions on a varied range of topics. Minghao Guo and Maxim Prigozhin, two of the newest Gruebele group members have been wonderful to work with.

I have collaborated with various groups during my stay here as a graduate student. I would like to thank Elizabeth (Betsy) Komives for many thought provoking discussions on research. I thank Zan Luthey-Schulten, Ke Chen, John Eargle and Anurag Sethi for providing me a detailed glimpse of Molecular Dynamics simulations. Anurag’s contribution for the RNA folding work was very valuable. I thank the members of D. E. Shaw Research, Stefano and Kresten for their elaborate and extensive simulations and analysis for the Fip35 WW domain folding project.
Departmental resources provided a strong backbone to my research effort. I have had the fortune to work with a very talented machinist Dave Hire for the construction of the single molecule instrument. His patience, ability to understand my needs without any difficulty and willingness to try out and deliver on challenging tasks lead to a successful fabrication of the instrument. I would like to add that all the machine shop staff have contributed to a number of my fabrication needs and hence deserve my utmost thanks. I thank Mike Thompson at the SCS Electronic services for extensive support in the assembly of time correlated single photon counting electronics box. His attention to detail is exemplary. I also thank Jim Wentz for design and fabrication of other electronic component.

Finally I would like to thank my wife for being patient and supporting me through my challenges. I thank our parents for their constant encouragement and support while being thousands of miles away.
# Table of Contents

List of Abbreviations ........................................................................................................... vii

Chapter 1. Introduction .......................................................................................................... 1

Chapter 2. RNA hairpin folding using 2-aminopurine as a fluorescence probe ............... 12

Chapter 3. Stem and loop dynamics in RNA hairpin folding ........................................... 34

Chapter 4. Pushing β-sheet proteins towards the folding speed limit with predictions from very long MD simulations .................................................................................. 52

Chapter 5. Binding of ribosomal signatures S4 and h16 .................................................... 70

Chapter 6. Design and fabrication of a high throughput fluorescence detector ............... 93

Chapter 7. Conclusion ........................................................................................................ 121

Appendix A. Supplementary Information for Chapter 2 .................................................... 123

Appendix B. Supplementary Information for Chapter 3 ..................................................... 130

Author’s Biography ............................................................................................................ 134
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>MD</td>
<td>Molecular Dynamics</td>
</tr>
<tr>
<td>2AP</td>
<td>2-aminopurine</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein Data Bank</td>
</tr>
<tr>
<td>VMD</td>
<td>Visual Molecular Dynamics</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic Acid</td>
</tr>
<tr>
<td>NMR</td>
<td>Nuclear Magnetic Resonance</td>
</tr>
<tr>
<td>FRET</td>
<td>Fluorescence Resonance Energy Transfer</td>
</tr>
<tr>
<td>CD</td>
<td>Circular Dichroism</td>
</tr>
<tr>
<td>EDTA</td>
<td>Ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>YAG</td>
<td>Yttrium Aluminium Garnet</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise</td>
</tr>
<tr>
<td>LB</td>
<td>Lysogeny Broth</td>
</tr>
<tr>
<td>IPTG</td>
<td>Isopropyl β-D-1-thiogalactopyranoside</td>
</tr>
<tr>
<td>SDS</td>
<td>Sodium dodecyl sulfate</td>
</tr>
<tr>
<td>ESI</td>
<td>Electrospray Ionization</td>
</tr>
<tr>
<td>GST</td>
<td>Glutathione S-transferase</td>
</tr>
<tr>
<td>TSE</td>
<td>Transition State Ensemble</td>
</tr>
<tr>
<td>SSU</td>
<td>Small Subunit</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal Ribonucleic Acid</td>
</tr>
<tr>
<td>RMSD</td>
<td>Root Mean Square Deviation</td>
</tr>
<tr>
<td>SMS</td>
<td>Single Molecule Spectroscopy</td>
</tr>
<tr>
<td>PMT</td>
<td>Photomultiplier Tube</td>
</tr>
<tr>
<td>DC</td>
<td>Direct Current</td>
</tr>
<tr>
<td>CMOS</td>
<td>Complementary Metal Oxide Semiconductor</td>
</tr>
<tr>
<td>LED</td>
<td>Light Emitting Diode</td>
</tr>
<tr>
<td>CW</td>
<td>Continuous Wave</td>
</tr>
<tr>
<td>FWHM</td>
<td>Full Width at Half Maximum</td>
</tr>
<tr>
<td>LBO</td>
<td>Lithium triborate</td>
</tr>
<tr>
<td>BBO</td>
<td>Beta Barium Borate</td>
</tr>
<tr>
<td>UV</td>
<td>Ultra Violet</td>
</tr>
<tr>
<td>GLMT</td>
<td>Generalized Lorentz-Mie Theory</td>
</tr>
<tr>
<td>IR</td>
<td>Infra Red</td>
</tr>
<tr>
<td>TAC</td>
<td>Time to Amplitude Converter</td>
</tr>
<tr>
<td>DAB</td>
<td>Data Acquisition Box</td>
</tr>
<tr>
<td>NI</td>
<td>National Instruments</td>
</tr>
<tr>
<td>RTSI</td>
<td>Real Time System Integration</td>
</tr>
<tr>
<td>GFP</td>
<td>Green Fluorescent Protein</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The Protein and RNA folding problem

Protein (1) and RNA (2) are among two of the most important building blocks of life. They are two of the three components in the central dogma of molecular biology: “DNA makes RNA makes protein” (3). It therefore becomes essential to understand the physical aspects of their folding behavior. Proteins are polymers of amino acids while RNAs are polymers of nucleotides (3). Almost fifty years ago Christian Anfinsen showed that the protein Bovine Pancreatic Ribonuclease completely refolds back to its original structure from the denatured state after the denaturing condition is changed to the native one (4). The ability of the protein to do this \textit{in vitro} in a buffer solution made him conclude that all the information about the protein’s structure is contained in the amino acid sequence itself. This discovery defined the protein folding problem (5). Can we predict the structure of a protein when only provided with its sequence? Furthermore, can we predict all the possible routes and their likelihoods that an unfolded protein would take as it makes its journey from the unfolded to the folded state? The same questions hold true for RNA. They too fold back to their native structure as the denaturing condition is changed to the native one. Since the pioneering work by Anfinsen the understanding of protein folding has seen a transformative change (5-19). RNA folding too has come to a much better light (20-29). Our understanding of protein and RNA folding will be key in unraveling the complex nature of interactions between these two kinds of molecules both \textit{in-vitro} and \textit{in-vivo}.

There are some key differences between the folding of proteins and RNA (30, 31). Proteins have a set of twenty natural building blocks called amino acids that make up their sequence while RNA only has a set of four nucleosides. Thus proteins are able to form more complex folds with a wider range of secondary and tertiary structures than RNA. The difference in the number of building blocks also makes the RNA energy landscape more rugged compared to those of proteins. RNA having a negatively charged backbone has an energy landscape that is strongly dependent on the concentration of positively charged ions in solution like magnesium, sodium and calcium. This dependence is not as severe for proteins in general. Also, for RNA, the folding process is known to be more hierarchical in nature. Secondary structures form first
followed by tertiary structures. On the other hand such a neat separation is usually not feasible for proteins. It is important to note here that for protein folding inside living cells, in certain cases, other protein molecules called chaperones may play an important role in the folding process (32-36). Chaperones are also known to be involved in the folding of RNA inside a cell (37).

**The Energy Landscape Theory**

Initially protein folding was explained through simple kinetic descriptions as used for small organic molecules. Typically experimental data would be fit to simple two state models or to models with a series of obligatory intermediates. It was forty years after Anfinsen’s pioneering experiments that a general theory on protein folding emerged. It was proposed by Bryngelson and Wolynes (10) and is called the “Energy Landscape Theory”. Cyrus Levinthal pointed out in 1969 that given the length of a typical protein (~ 300 amino acids), the time required by an unfolded protein to go through all the possible configurations so as to eventually reach its native structure is astronomically large ($10^{10}$ years). Thus random search is not a feasible solution for proteins to get to their native structure. The Energy Landscape theory proposes that proteins are natural biopolymers that have an inherent bias towards folding to the native state encoded into their sequence. This results in the so called “funnel” like energy landscape for the folding of proteins. The same theory holds true for RNA although the landscape is expected to be more rugged because of the fewer number of building blocks involved.

A typical energy landscape for proteins is shown in Fig.1.1 At the top of the funnel are the unfolded states of the protein. The width of the funnel represents the conformational entropy. As the protein travels down the funnel, through different paths, they encounter different energy barriers $\Delta E$ which represents the “roughness” of the landscape. The overall energy bias towards the native state is shown as $\partial E$ in Fig.1.1. On its journey down the funnel the unfolded protein collapses and then proceeds to a transition state ensemble. Along this path it may encounter traps resulting in misfolded states. At the bottom of the funnel it finally reaches the native state at the lowest energy. $Q$ is the fraction of correctly made native contacts. The energy landscape for a RNA is expected to have more roughness ($\Delta E$) than a protein for reasons discussed before.
Protein and RNA folding Experiments

Equilibrium and non-equilibrium experiments have both been used to study folding of proteins and RNA. NMR (12, 38-41), fluorescence (42-46), circular dichroism (47, 48) and small angle X-Ray scattering (49) are some of the probes that have been extensively used to study protein and RNA folding in equilibrium and non-equilibrium. Specific non-equilibrium methods include stopped flow (41, 50-53), temperature jump (54) and pressure jump (42).

Protein and RNA folding Simulations

Molecular dynamics simulations have had a significant contribution to understanding protein and RNA folding at atomistic detail. With the advent of faster processors and larger computer clusters, simulations can now be run for milliseconds for reasonably large biomolecules. This timescale was not accessible until recently (55, 56). On the other hand, experimental studies of protein folding for timescales as short as a few μs have been around for some time (57). The overlap of experimental and simulation timescales have finally allowed experimental validation of simulation results. The ability to do so is providing valuable feedback for improving the potential parameters of the MD force fields that were originally developed from small molecules. As force field parameters keep getting better, in-silico approaches will yield more quantitative results.

Bulk versus Single Molecule Experiments

Protein folding has commonly been performed in bulk experiments. These provide us with an ensemble averaged result. As an example multiple folding time scales may be indicative of a single process going through these timescales or many processes occurring simultaneously while resulting in the same average effect of showing multiple timescales. Thus a model based on either approach would agree with the observed data. Alleviating this problem, single molecule techniques allow us to access the entire distribution of the observed phenomenon. The application of single molecule techniques to protein folding have only started recently (58). They provide a more detailed picture of the heterogeneities of the energy landscape. The primary
single molecule techniques currently in use for the study of protein folding involve FRET based measurements (59) and single molecule force spectroscopy experiments (60).

**Structure of this thesis**

In the second and third chapters of this thesis, we describe our study of the folding of a small RNA hairpin. RNA hairpins have been shown to have a rugged energy landscape (23). MD simulations of RNA hairpins also point in the same direction (61). We incorporated 2-aminopurine [2AP/A*/a*], an adenine analog, as a fluorescence probe within a RNA hairpin sequence. 2AP is known to be sensitive to base stacking (62). In the second chapter the RNA hairpin sequence studied is ga*cUUCGgac. This was chosen because the same system was studied earlier by Ma et al. and Stancik et al. by ultra-violet and infra red absorption respectively. We show that by using 2AP as a complementary probe we are able to get a better picture of the energy landscape for the small hairpin. In the third chapter we study the effect of 2AP as a probe for the loop and the stem in the RNA hairpin, independently. The sequences studied are ga*cUACGguc for the stem mutant and gacUA*CGguc for the loop mutant. Surprisingly, we find that the thermodynamics and kinetics for both the loop and the stem mutants can be fit to the same four state model with concurrent parallel and sequential schemes.

In the fourth chapter, we show how long trajectories of MD simulations were analyzed to predict mutations that would enhance the observed folding rate and also further stabilize the protein. This chapter highlights how MD simulations are coming closer to experimental protein folding timescales and helping us to better predict and understand the folding behavior of proteins in atomistic detail.

In the fifth chapter, we detail the nature of interaction between the signature regions of the ribosomal RNA h16 and the ribosomal protein S4 through experiments and MD simulations. Both S4 and h16 are known to interact early on in the progressive build up of the ribosomal complex. The S4 fragment is natively unfolded. Circular dichroism measurements along with fluorescence quenching experiments provide conclusive proof of a binding interaction. Dilution experiments performed show a low micromolar binding constant. Elaborate MD simulations indicate a fly-casting mechanism whereby the S4 gains structure upon binding to the h16 RNA.
In the sixth chapter, we detail our efforts to build a high throughput single molecule detector. The setup consists of a droplet generator that automatically generates ~ 10 micron sized droplets which are trapped by counter propagating infrared laser beams in the middle of a cube made of fused silica lenses. A femtosecond laser capable of exciting chromophores from the ultra-violet to the visible range probes the droplet. These droplets contain only a few labeled protein molecules which fluoresce upon excitation. The fluorescence is collected in all $4\pi$ steradians using a custom built time correlated single photon counting setup. Possible improvements for obtaining single molecule sensitivity are discussed.
Fig. 1.1: The folding energy landscape
References


protein folding: The time evolution of primary processes. Annual Review of Physical 
Opinion in Structural Biology 9:189-196.
Chemistry 50:485-516.
Biophysics and Biomolecular Structure 29:327-359.
77.
National Academy of Sciences of the United States of America 97:646-651.
5:303-310.
Herschlag. 2002. Exploring the folding landscape of a structured RNA. Proceedings of 
293:271-281.


Chapter 2

RNA hairpin folding using 2-aminopurine as a fluorescence probe*

Introduction

Nucleic acids were the first biopolymers to have their folding and unfolding dynamics investigated systematically by fast temperature jump methods (1). As experimental work has moved towards smaller structures to make direct connections with simulation, it has become clear that the energy landscapes of even the smallest RNA fragments, such as short tetraloop motifs, are highly rugged compared to small proteins (2-7). Multiphasic kinetics occur on timescales from picoseconds to milliseconds (8, 9), and it can become difficult to disentangle loss of hydrogen bonding, base stacking or compactness as RNA unfolds.

Four experimental and computational studies have recently focused on the UNCG tetraloop motif with a short stem (2-4 base pairs in the stem). Proctor et al. (6) showed that incorporation of bromine to make a UUCG bromo-guanosine loop sterically increases the unfolded state free energy, favoring the folded state. Later Ma et al. (7) observed three unfolding relaxation phases (fast, medium, slow) of the eight nucleotide hairpin gcUUCGgc. Thus at least four states were required to model its unfolding kinetics. A simple lattice model with an entropic correction factor accurately reproduced the experimental data. In accordance with the model, one of the four states, “S,” was assigned as a kinetic trap, but based on its high fitted entropy in a four state model, it was speculated that it could be a single-stranded state with highly variable base stacking instead. The question of base stacking was taken up quantitatively by Stancik et al. (10), who detected gcUCCGgc unfolding induced by a T-jump by using structurally sensitive infrared fingerprints. They concluded that the fastest phase monitors mainly base unstacking dynamics, while the medium (ca. 5-10 µs) phase monitors mainly hydrogen bond breakage. In silico, Garcia and Paschek have studied the folding/unfolding equilibrium thermodynamics of the

gcUUCGgc hairpin as a function of temperature and pressure using replica exchange molecular dynamics simulations, showing that the RNA exists in a diverse ensemble of conformations (11).

To investigate further the relative role of hydrogen bond loss and base unstacking during RNA unfolding, we study the ga\*cUUCGguc tetraloop by fluorescence-detected T-jump kinetics and molecular dynamics. We insert a fluorescent 2-aminopurine residue a\* or A\* into the sequence as a probe with stacking-sensitive fluorescence lifetime. Previously, a temperature jump study was done on the GNRA type (R: purine) RNA tetraloop with a 2-aminopurine substitution in the loop region showing single and double exponential kinetics for two different substitutions: GA*AA and GAA*A. As shown by Proctor et al., stems with two paired bases (8-mer) vs. three paired bases (10-mer) do not show substantially different kinetics (6), although the longer stem hairpin is slightly more stable.

We find that the 2-aminopurine probe again yields three kinetic phases. The fast and medium phases have the same time scales as the corresponding phases observed by Ma et al. and Stancik et al. The slow phase has the same time scale as the slow phase observed by Ma et al. at high temperature.

The molecular dynamics simulations and experiment yield a picture that is in general agreement with the previous data, but does not neatly segregate hydrogen bonding and unstacking into separate phases. We distinguish two types of base stacking dynamics: increased stacking fluctuations, when bases continually move in and out of planar contact, and true base unstacking, where the distance between bases increases clearly (>3.5 Ångstrom) beyond the π-π interaction distance. Based on the simulations, we assign the fast experimental phase to increased stacking fluctuations and fraying at the end of the stem. We assign the medium phase to concerted unfolding where the native hydrogen bond network is entirely lost, and simultaneously some unstacking within single strands occurs. Finally we assign the slow phase observed at the highest temperature to a single stranded state in which most stacking interactions are lost.

**Methods**

**RNA design and sample preparation**

In order to study the folding of a small fluorescent RNA tetraloop hairpin similar to the UNCG motifs studied previously, we chose the sequence 5’-ga\*cUUCGguc-3’ (stem nucleotides
in lower case and loop nucleotides in upper case). The “\(a^*\)” represents 2-aminopurine (2AP) an analog of adenine which we use as a fluorescence reporter in our folding and unfolding measurements (12-14). The 2AP fluorescence is known to be very sensitive to base-stacking (13, 14). It is also mildly sensitive to base pairing (14). The sensitivity to base pairing seemingly comes about indirectly through an effect on base stacking.

The 2AP labeled oligonucleotide was purchased from Thermo Scientific (Waltham, MD). It was dissolved and stored in a low salt \(\text{P}_{10}\text{E}_{0.1}\) buffer (10 mM sodium phosphate, 0.1 mM \(\text{Na}_2\)-EDTA, pH 7.1). This buffer is known to favor hairpin conformation (6). The RNA concentration was measured by absorption, using an extinction coefficient of 83,400 \(\text{M}^{-1}\text{cm}^{-1}\) at 260 nm supplied by Thermo Scientific.

**Thermal denaturation**

Thermal unfolding was measured with a spectro-polarimeter (JASCO Inc, Easton, MD). The spectro-polarimeter detected circular dichroism (CD), integrated fluorescence, and absorption. A thermoelectric cooler accompanied by a water bath was used to scan the temperature of the sample. Temperature melting curves were repeated several times with the same sample, and with different samples. The results were reproduced within measurement uncertainty. For direct comparison of the thermodynamic fluorescence denaturation experiments to the temperature jump measurements, fluorescence was excited at 280 nm and the same Hoya B-370 (Hoya Corp., Fremont, CA) filter was used in the fluorescence melt and the T-Jump measurements.

**Temperature jump experiment**

The relaxation kinetics of the RNA were obtained at final temperatures in the 313-353 K range by jumping the temperature of the RNA solutions by 10 degrees. The jump duration was 10 ns using a YAG (Continuum, Santa Clara, CA) laser pulse Raman-shifted to 1.9 \(\mu\)m. Relaxation to the original temperature took > 50 ms after the jump, much longer than the 0.5 ms observation window reported here. A fused silica sample cell of 200 \(\mu\)m path length was used. 2AP was excited at 281±2 nm by a mode-locked tripled Ti:Sapphire laser beam. The repetition rate of the laser yielded 14 ns intervals between excitation pulses. The fluorescence decays were digitized using 500 ps resolution oscilloscope (Tektronix), yielding 28 data points per
fluorescence decay. The data were downloaded to a computer and analyzed by program written in LabWindows (National Instruments). A more detailed description of the setup is found in (15). RNA concentrations of 160 µM and 500 µM were used. Thirty temperature jumps were averaged to obtain better SNR for the 160 µM data and eleven were averaged for the 500 µM data. No difference was found between the two data sets within the signal-to-noise ratio, so RNA aggregation is below the detection level. The size of the temperature jump was calibrated using the known tryptophan fluorescence lifetime change with temperature, obtained with a separate amino acid sample in the same cell configuration.

**Data analysis and fitting method**

Fluorescence lifetime (χ-analysis) and fluorescence intensity analyses were initially performed on the kinetics data taken for each final temperature using the Igor Pro software package (15). χ-analysis fits fluorescence decays to a linear combination of initial (just after the T-jump) and final (0.5 ms after the T-jump) decays, producing a relative lifetime shift versus time. Intensity analysis integrated the 28 data points per fluorescence decay. The time zero was determined from Raman scatter of the heating pulse in the sample cell, and decays near time zero (≈ 5 decays, or 70 ns of data) were not included in the analysis. The data was binned logarithmically, so that each phase and the baseline would contribute a comparable number of data points to the fit.

A four state kinetic and thermodynamic model was developed to fit all steady-state and kinetic fluorescence data simultaneously using the Levenberg–Marquardt algorithm:

\[
N \leftrightarrow E \leftrightarrow U \leftrightarrow S
\]

We label the states with single letters (shown in bold) “Native”, “FrayEd”, “Unfolded” and “UnStacked” to keep the nomenclature from ref. (7), although the interpretation of the last state, as suggested by its name, will be somewhat different. A simultaneous fit allowed us to use few parameters to account for all of the data. The free energies of the four states and of the three free energy barriers connecting them were fitted to a first order temperature dependence, which sufficed over the full temperature range studied:

\[
\Delta G_i = \Delta G_i^{(0)} + \Delta G_i^{(1)} (T - T_0) \\
\Delta G_{i \rightarrow j}^\dagger = \Delta G_{i \rightarrow j}^{(0)\dagger} + \Delta G_{i \rightarrow j}^{(1)\dagger} (T - T_0)
\]

The subscript “i” and “j” runs over states and \( T_0 \) is a reference temperature.
For each state, a single signal baseline linear in temperature was fit simultaneously for the intensity-detected kinetics and thermodynamics of each state. This reduction in parameters was possible because the same detection filter was used in steady state and kinetics experiments. A separate signal baseline, also linear in temperature, was necessary for the lifetime analysis.

\[ S_i(\text{Intensity or lifetime}) = S_i^{(0)} + S_i^{(1)}(T - T_0) \]  

(3)

The fitting program first calculated the free energies at the pre-jump temperature using equation (2) above. Each population was then assigned by calculating the partition function for each state at the pre-jump temperature. This population was then placed on a new free energy surface, also computed from eq. (2) for the higher temperature just after the jump. The relaxation kinetics were simulated by matrix multiplication,

\[ P(t + \Delta t) = R \times P(t)\Delta t + P(t) \]  

(4)

where \( R \) is the rate coefficient matrix. For example in the sequential scheme:

\[ N \xleftarrow{k_1} E \xrightarrow{k_2} U \xleftarrow{k_3} S \]

where only the forward rates are shown. The corresponding reverse rates have a “-“ sign in the subscript e.g. \( k_{-1} \). Equation (4) can then be explicitly expressed as,

\[
\begin{bmatrix}
P_N(t + \Delta t) \\
P_E(t + \Delta t) \\
P_U(t + \Delta t) \\
P_S(t + \Delta t)
\end{bmatrix} =
\begin{bmatrix}
-k_1 & k_{-1} & 0 & 0 \\
k_1 & -(k_{-1} + k_2) & k_{-2} & 0 \\
0 & k_2 & -(k_{-2} + k_3) & k_{-3} \\
0 & 0 & k_3 & -k_3
\end{bmatrix} \otimes \begin{bmatrix}
P_N(t) \\
P_E(t) \\
P_U(t) \\
P_S(t)
\end{bmatrix} \times \Delta t + \begin{bmatrix}
P_N(t) \\
P_E(t) \\
P_U(t) \\
P_S(t)
\end{bmatrix}
\]

(5)

A pre-exponential factor \( k_m(T) = k_m \eta(298K)/\eta(T) \) with a temperature-dependent viscosity from ref. (16) was also fitted, and instead the average N to E activation energy \( (\Delta G_N^{(0)} + \Delta G_E^{(0)})/2 \) was held constant. \( k_m \) thus assumes solvent friction in the normal Kramers regime (17). The actual viscosity dependence may be different due to internal friction or local hydrodynamic effects, but the simplicity of eq. (2) currently does not warrant a more detailed treatment of the prefactor.

**Molecular dynamics simulation setup**
The molecular dynamics simulations of the solvated complexes were performed using NAMD2 (18) with the CHARMM27 force field (19, 20). Parameters for 2-aminopurine were adapted from (21) and were further checked using GAMESS (22). The structure of the 10-mer was obtained from pdb id of 1K2G. All hydrogen atoms were added to the RNA molecule using the Psfgen plugin in VMD (23). The RNA molecules were explicitly solvated with TIP3 water molecules (24). Twenty potassium ions and ten chloride ions were initially placed at the minima of the Coulombic electrostatic interaction energy 6.5 Å away from the RNA molecules, to simulate experimental concentrations of buffer ions. The minima were calculated on a uniform grid using the program Ionize (25). The models were solvated in a 2-step process. In the first step, Solvate 1.0 (26) was used with 2 Gaussians to add 2 layers of water molecules to the system. This placed water molecules in the solvation shell of the RNA and ions. Then, the Solvate 1.2 plugin to VMD was used to place the bulk water, resulting in a box size of $55 \times 55 \times 70$ Å and a system size of about 18000 atoms.

Simulation Protocols

The native state simulations were done with periodic boundary conditions in an NPT ensemble. A pressure of 1 atmosphere was achieved by using a Langevin piston, and the temperature was set to 273 K using Langevin dynamics. Electrostatics were calculated with the particle mesh Ewald method (27). The van der Waals interactions were calculated using a switching distance of 10 Å and a cutoff of 12 Å. Time steps for updates of bonded, van der Waals, and electrostatic calculations were 1, 2, and 4 fs, respectively. The system free energy was initially minimized using a 4-step protocol in which the water molecules were allowed to associate with the macromolecule before allowing the macromolecule to move. These steps were: 1) all heavy atoms fixed (2,000 steps), 2) heavy atoms fixed excluding water and ions (3,000 steps), 3) macromolecule backbone atoms fixed (5,000 steps), and 4) all atoms free to move (20,000 steps). During the initial equilibration, the system was gradually heated from 100 K to 273 K according to the protocol in (28) during which different parts of the system were harmonically constrained. After this 4-ns equilibration, each system was run for 6 ns at 273 K. Under these condition, the hairpin was stable.

Simulations at 298 K resulted in an unfolded RNA hairpin within 10 ns, indicating a much lower unfolding transition temperature of the molecular dynamics force field compared to
experiment. We performed 298 K simulations with three different salt compositions in the simulation box to monitor the effect of salt concentration on the RNA hairpin’s stability: 1) ten potassium ions; 2) twenty potassium ions and ten chloride ions; and 3) three magnesium ions, fourteen potassium ions, and ten chloride ions. None of these salt conditions shifted the melting point above 298 K. The disparity between the unfolding temperature of the RNA hairpin in experiments and MD simulations indicates that the CHARMM parameters of RNA molecules have to be improved. Similar shifts in melting point have been observed in previous RNA and protein molecular dynamics simulations (11, 29).

Two different conformations from the native state simulation were taken as the starting point for temperature jump simulations from 273 K to three different temperatures: 300 K and 333 K at 1 atmosphere, and 498 K at 216 atmospheres pressure. The temperature was increased uniformly from 273 K to the target temperature during each of these simulations in ten steps over a period of 250 ps. The pressure was increased from 1 atmosphere to 216 atmospheres in a single step.

Analysis of MD runs

Two variables were monitored in the analysis of the MD simulations: the number of native hydrogen bonds, and the integrity of native base stacking interactions. The general results obtained are independent of the exact definition of a broken hydrogen bond or lost stacking interaction. We describe the choices used in the figures in more detail.

The number of native hydrogen bonds between bases in the stem (maximum of 8) is calculated during each frame of the simulation. The maximum number of hydrogen bonds is eight in the RNA hairpin stem. We also looked at hydrogen bonding in the loop. A hydrogen bond is formed under two conditions: the hydrogen bond acceptor (A) atom and the hydrogen atom are within a cutoff distance of 3.5 Å from each other; and the angle formed by the donor atom, hydrogen atom, and the acceptor atom is between 180±30 degrees.

The number of native base stacking interactions between bases in the stem portion of the RNA hairpin (maximum of 4) is counted during each frame of the simulation. The maximum number of base stacking interactions in the native structure is six in the RNA hairpin including two loop to strand base stacking interactions. Our RNA crystal structure taken from pdb id 1K2G shows no base stacking in the loop and hence a base stacking analysis in the loop was not
performed. A base stacking interaction is formed under three conditions: two bases have non-hydrogen atoms that are within 3.5 Å of each other, the normal vectors to the aromatic ring planes are at an angle of less than 30 degrees of each other, and the bases are not hydrogen bonded to each other.

**Results**

**Thermal denaturation**

The fluorescence melting curve of the RNA obtained with a filter identical to the T-jump experiments is shown in Fig.2.1A. The native baseline persists up to 310 K after which unfolding begins. The unfolding transition lasts to about 340 K, after which a strong unfolded baseline takes over. The data can be fitted by a two-state scenario with midpoint at 330 K. Circular dichroism data (Appendix A) has a few degree different midpoint when fitted by a two-state model with linear baselines (Data in Appendix A shows few degrees of deviations). Like the gcUUCGgc system studied earlier, the thermodynamic melting behavior of gc*UUCGgc alone is amenable to a deceptively simple two-state fit, if one is willing to neglect the small difference in CD and fluorescence melting points.

**Temperature jump**

The temperature jump relaxation kinetics at three representative final temperatures are shown in Fig.2.1B-G. All temperature jumps were 10±1 degrees, and final temperatures in the range from 313 K to 353 K were monitored. The full set of data is included in the Appendix A. Fluorescence lifetime and integrated fluorescence intensity were used as two different probes. The lifetime-detected kinetics in Fig. 2.1B-D originated primarily from a fast component of the fluorescence decay ($\tau_{\text{short}} \approx 3$ ns) that accounted for $\approx 70\%$ of the fluorescence decay amplitude. This component relaxed to a longer lifetime (less quenching) as the RNA unfolded after the T-jump. A longer fluorescence lifetime component of $\tau_{\text{long}} \approx 6$ ns that accounted for $\approx 30\%$ of the fluorescence amplitude switched within the T-jump dead time ($< 70$ ns). We assign it to the intrinsic temperature dependence of unquenched 2AP fluorescence, although we cannot exclude the possibility of a fourth very fast kinetic phase due to increased stacking fluctuations (**vide infra**).

Both the lifetime and intensity data show three resolvable kinetic components. At low
temperature the relaxation is bi-exponential showing a fast phase of a ≈2 µs along with a medium phase of 5-10 µs (depending on temperature). As the temperature increases, the fast phase amplitude decreases. Near the melting temperature, the kinetics appear nearly single exponential. At even higher temperatures the medium phase speeds up to a few µs while a new slow phase of ~ 50 µs appears. In the intensity data, the slow phase has an amplitude with the same sign as the faster phases; in the lifetime data, its amplitude has the opposite sign.

**Four state model fit**

A four-state model is minimally required to explain the presence of the three resolved phases. We fitted a model that used the same signal baselines for thermodynamic melt and fluorescence intensity-detected kinetics, and separate signal baselines for the lifetime data. Four free energies, two barrier free energies, and a Kramers prefactor $k_m$ were fitted; one barrier was held constant because of barrier-prefactor correlation in the fit. The model that best fits the data has parameters shown in Table 2.1. The fits are laid over the experimental data in Fig. 2.1, and account for the data nearly within measurement uncertainty. Slightly better fits could be obtained by floating more parameters (e.g. quadratic terms in eq. (2) ), but we refrained from introducing additional parameters to avoid excessive parameter correlation and over-interpretation of the simple free energy model.

The model in Fig. 2.1 invokes four states \( N \) (Native), \( E \) (FrayEd), \( U \) (Unfolded) and \( S \) (UnStacked). These states are sequentially more unfolded as evidenced by an increasingly negative slope of free energy with temperature, indicative of increased state entropy (Table 2.1a, Fig. 2.2). With this model, a linear free energy dependence on temperature was sufficient to account for the data.

The data could also be fitted to an alternative four-state model, which includes transition from the unfolded state to a trapped state that is not populated at the highest temperature. However, this model required more fitting parameters (quadratic free energy term) and resulted in an inferior fit of the data (Appendix A). Thus we favor the sequential model as the simplest one needed to describe the data. This does not rule out the existence of additional states. Four states are simply the minimum we needed to fit the resolved kinetic phases within measurement uncertainty.
**Molecular Dynamics Simulations**

To aid in the structural interpretation of the four states, we complemented the experiment with MD simulations. The RNA was simulated at 273 K, as well as two jumps each from 273 K to the final temperatures of 300, 333 and 498 K. The MD results could be naturally grouped into at least four macrostates based on hydrogen bonding pattern (Fig. 2.3B), although sequestration of populations into further sub-ensembles cannot be ruled out with a small number of trajectories. Before we interpret the data with the aid of molecular dynamics trajectories, it should be cautioned that *in silico*, RNA is not as thermally stable as observed experimentally, and also unfolds much faster.

Representative microstate examples of the four most prominent macrostates are shown in Fig. 2.3A. The native state (N) has on average 65% of the native hydrogen bonds formed. On average 65% of the stem bases are stacked. Yet even the equilibrium trajectory in Fig. 2.4A shows one major excursion to a frayed state with fewer hydrogen bonds and less base stacking. The frayed state (E) is populated more extensively than the U and the S states in the smallest simulated T-jumps from 273 to 300 K (Fig. 2.4).

Immediately after the T-jump the RNA shows increased stacking and hydrogen bond fluctuations, even though these native interactions are not immediately broken. A concomitant loss of base stacking and hydrogen bonding then occurred in both 300K T-jump simulations. Afterwards, the frayed (E) state sampled in the simulations has only 20% of native hydrogen bonds left in the stem, while it still has 35% of base stacking. A much more unfolded state U is reached in the T-jump to 333 K (Fig. 2.4C). Most of the native hydrogen bonding is lost at this stage, although 20% base stacking persists. Finally, in the temperature jump to 498 K, even the base stacking is reduced to 10%, with large fluctuations as transient interactions are formed and broken during the simulation. 10% may actually be a slight overestimate, based on our relatively lax criterion for base stacking (Methods). The hydrogen bond and stacking patterns in Fig. 2.4 are the average of two trajectories at each temperature. We have also included an analysis of the hydrogen bonding in the loop region. However we have only shown the analysis for the equilibrium run. The other runs do not show any general trends and would not add to a better understanding of our system. The individual simulations can be found in Appendix A.

On the whole, the loss of hydrogen bonding with increasing temperature is much more facile than the loss of single-stranded base stacking, but increased stacking fluctuations occur earlier
than loss of hydrogen bonding. Based on our limited sampling, the MD simulations support the following sequence of events as temperature is increased:

Native $\rightarrow$ Frayed $\rightarrow$ Unfolded $\rightarrow$ Unstacked

The native state has most hydrogen bonds and stacking interactions formed at any given instant (Fig. 2.3B). The frayed state has increased stacking fluctuations, suffers disruption of hydrogen bonds, but retains considerable single-stranded stacking. The unfolded state has lost most hydrogen bonding, but retains some single-stranded stacking. The unstacked state maintains no significant hydrogen bonding, and little single-stranded stacking. Thus stacking fluctuations and loss of stacking bracket the loss of hydrogen bonding.

**Discussion**

We use 2AP as an unfolding probe for RNA because it is mainly sensitive to base stacking (14, 30, 31). Stacking causes a reduction of the fluorescence intensity in two stages. When a stacking interaction first forms, static quenching results. Quenching is enhanced when 2AP is stacked with a purine compared to a pyrimidine. Once two bases are stacked, the fluorescence quenching still depends on fluctuations of the stacked geometry. More coplanar geometries with fewer fluctuations result in more quenching, referred to as dynamic quenching (13). The reverse of these processes would be increased stacking fluctuations, followed by true unstacking of the bases beyond a critical distance. A loss of rigidity from breakage of hydrogen bonding also has been shown to cause a small change in 2AP fluorescence (14). On the other hand, quantum calculations on 2AP claim a minimal effect of hydrogen bonding on its fluorescence characteristics (32).

Our RNA stem is longer than the one used by Ma *et al.* and Stancik *et al.* by one base pair, and has a different loop base than the latter. Based on previous measurements, these differences are unlikely to change the general conclusions we make here. Proctor *et al.* studied kinetics as a function of stem length, and found only small changes in rate (33). Ma *et al.* studied a very different loop (UUUU), and still found four phases, albeit in slightly different proportions from the UUCG loop or UCCG loop (10).

In the study by Ma *et al.* (7), a four state model like the one in Fig. 2.1 fitted three phases with a temperature dependence similar to the one observed here. The favored interpretation of state S was a misfolded trap, based on its structural assignment in a lattice simulation. It was also speculated that state S could be a single stranded state with little or no hydrogen bonding.
and variable single-strand stacking. Our best model fit is in better agreement with the latter interpretation. The slope $\partial \Delta G / \partial T$ becomes increasingly negative in the progression $\text{N-E-U-S}$ (Fig. 2.2B, Table 2.1a), indicating again that $S$ has the highest entropy of the four states. A state with low hydrogen bonding and stacking satisfies this requirement, but a misfolded compact state does not. Attempts to fit a trap that is depopulated in favor of $U$ at the highest temperatures produces an inferior fit despite requiring more fitting parameters (Appendix A). The molecular dynamics simulation shows that while stacking persists at moderate temperatures (300 K in Fig. 2.4), it is almost completely lost at higher temperatures.

In the study by Stancik et. al. (10), base stacking rearrangements are assigned to the earliest phase, and hydrogen bond breaking follows thereafter based on structure-specific IR transient absorption data. Our fluorescence unquenching fast and medium phases are in general agreement with the phases observed by Stancik et al., except that the two coordinates (stacking, hydrogen bonding) are not as neatly assignable to separate phases. Our data at 323 K have significant amplitudes for both the fastest (1-2 $\mu$s) and medium (6-8 $\mu$s) phases fitted by the four state model. Since 2AP is stacking sensitive, changes in stacking occur in both phases. The molecular dynamics simulation supports loss of stacking on three different time scales. Immediately after the temperature jumps in Fig. 2.4 (time > 0), larger fluctuations of the native stacking are observed. The stacking is then reduced in concert with loss of hydrogen bonding, and continues to fluctuate more after the hydrogen bonding is broken. Only at the highest temperature is native stacking completely lost even within single strands. By analogy, we assign our fast experimental phase to increased fluctuations of base stacking after the temperature jump, the medium phase to coordinated loss of hydrogen bonding and stacking, while the slow phase observed only at high temperature is due to further unstacking of the single RNA strand.

In the study by Menger et al. the presence of multiple states in the loop region is brought forth by the different phases observed in their T-Jump experiments. For the GA*AA mutant a single 22 $\mu$s phase is seen whereas for the GAA*A a biexponential is observed with lifetimes of 5 and 41 $\mu$s.

The RNA free energy landscape is thus much more rugged than the free energy landscape of small peptides. With a prefactor of $k_m \approx 1.3 \mu s^{-1}$ in table 2.1, similar to that measured for beta sheet peptides (34), the barriers separating the four states range from 8 to 20 kJ/mole at the reference temperature $T_0$ of each state. For small peptide hairpins with aromatic stacking
interactions simulations estimate a root mean square roughness of only \( \approx 2.5 \) kJ/mole (35), and measurements of relaxation rate fluctuations yield a root mean square value of 2 kJ/mole (36), also corroborated by single molecule experiments (37). Although four different probes (ultraviolet absorption, infrared absorption at two frequencies, and 2AP quenching) have so far resolved three phases, it is possible that more than four states exist. Both our fluorescence data and the infrared data (10) show unresolvable phases, which may be intrinsic temperature dependences of the chromophores, or they may signal additional transitions involving smaller structural changes, or fast relaxation down steep parts of the free energy surface accessed by the RNA population right after the temperature jump. Picosecond temperature jumps have resolved additional fast dynamics (8). It would be interesting to see if future experiments with different or faster probes detect additional phases, or if extensive simulations allow classification of more states.

Our four state model with 8-20 kJ/mole free energy barriers should be viewed as a minimal model for roughness of the RNA free energy landscape, even though it reproduces the data nearly within the signal to noise ratio. Together with the kinetic data, the molecular dynamics simulation indicates that stacking fluctuations contribute to unfolding on the shortest time scales, while concerted unstacking and hydrogen bond loss contributes at intermediate times, and unstacking within single strands becomes important at higher temperature. The currently available data can be explained by such a sequential model, but additional resolvable states or traps cannot be ruled out. The fast time scale of the experiments makes RNA hairpins amenable to simulation with future improved force fields that better reproduce stability and absolute kinetic time scales than our simulations here. Unfortunately, the <100 \( \mu \)s relaxation times observed for small RNAs are not resolvable by current single molecule experiments. Even for the simplest RNA secondary structures, the energy landscape may turn out to be more complex yet.
### Tables

Table 2.1: Equilibrium thermodynamic parameters and activation energies of the four state model summarized in equation (2) and Fig. 2.1.

<table>
<thead>
<tr>
<th>State</th>
<th>$T_0$ (K)</th>
<th>$\Delta G^{(0)}$ (kJ mol$^{-1}$)</th>
<th>$\Delta G^{(1)}$ (kJ mol$^{-1}$ K$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native</td>
<td>328</td>
<td>-0.2599</td>
<td>0.4297</td>
</tr>
<tr>
<td>FrayEd</td>
<td>325</td>
<td>-1.4599</td>
<td>0.3713</td>
</tr>
<tr>
<td>Unfolded</td>
<td>328</td>
<td>-0.7925</td>
<td>0.0000</td>
</tr>
<tr>
<td>UnStacked</td>
<td>344</td>
<td>0.0000</td>
<td>-0.7251</td>
</tr>
</tbody>
</table>

Table 2.1a.

<table>
<thead>
<tr>
<th>Transition state</th>
<th>$\Delta G^{(0)*}$ (kJ mol$^{-1}$)</th>
<th>$\Delta G^{(1)*}$ (kJ mol$^{-1}$ K$^{-1}$)</th>
<th>$k_m$ (µs$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native $\rightarrow$ FrayEd</td>
<td>8.000</td>
<td>0.1916</td>
<td></td>
</tr>
<tr>
<td>FrayEd $\rightarrow$ Unfolded</td>
<td>12.1862</td>
<td>-0.0501</td>
<td>1.323</td>
</tr>
<tr>
<td>Unfolded $\rightarrow$ UnStacked</td>
<td>20.3012</td>
<td>0.2074</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1b.
Figures

Fig. 2.1: Experimental thermal denaturation and temperature jump data fitted by a four state model. (A) Thermodynamic data and model. (B-D) Relative fluorescence lifetime change and model fit at three of the five temperatures measured (see also Appendix A). (E-G) Relative fluorescence intensity change and model fit.
Fig 2.2: State populations and free energies: $N$ = native state; $E$ = frayed state; $U$ = unfolded state without native hydrogen bonding but residual stacking; $S$ = single stranded unstacked denatured state. (A) Populations from the 4 state model in Fig. 2.1. (B) Free energies of the four states and activation energies between connected states at three temperatures (see Appendix A for an alternative fit where state $S$ is a trap populated at intermediate temperatures only).
Fig 2.3: (A) Representative conformations of the native state $N$, frayed state $E$, unfolded state $U$, and fully unstacked state $S$. Arrows in Fig. 2.4 indicate the times at which these configurations were sampled. Blue: loop; red: stacked stem bases; ochre: unstacked stem bases; blue dotted lines: native hydrogen bonds. (B) Conformational probability plot of hydrogen bonds and stacking interactions color coded by temperature: dark blue: 273 K equilibrium trajectory; blue-green: after jumps to 300 K; orange: after jumps to 333 K; red: after jumps to 498 K. The outermost contour for each ensemble is at 50% of the highest probability contour.
Fig 2.4: Molecular dynamics at 273 K (top left) and for three temperature jumps starting at $t=0$. Native base stacking (red/light) and native hydrogen bonding (blue/darker) are shown, with each time point averaged over a 0.5 ns window. The arrows refer to the conformations shown in Fig. 2.3A. The T-jumps are each an average of two trajectories with different random initial conditions. The individual trajectories are shown in the Appendix A.
References


Chapter 3

Stem and loop dynamics in RNA hairpin folding*

Introduction

Motivated by our studies of 2AP as a fluorescence probe in Chapter 2 we looked at the thermodynamics and kinetics from the stem and loop of an RNA hairpin. Ideally, substituting 2AP at different adenine positions would yield structurally localized information about RNA folding, without changing the overall folding mechanism. Our starting point is the gacUACGguc 10-nucleotide hairpin. We study its stem mutant ga*cUACGguc and its loop mutant gacUA*CGguc by thermal denaturation and by fast (nanosecond) temperature jump experiments.

We find that the stem and loop mutants yield opposite signals for thermal melting and folding/unfolding relaxation kinetics. Nonetheless, these signals can be fit by a single global free energy landscape. The differences between the stem and loop mutants are due to structurally localized information about the folding process: the stem loses base stacking upon unfolding, whereas the loop gains some base stacking upon unfolding. 2AP is thus a minimally intrusive local probe for base stacking even in the smallest RNA molecules.

Using entropy and fluorescence intensity as reaction coordinates, we construct a quantitative minimalist energy landscape for gacUACGguc. The free energy surface features four states along two reaction coordinates. Analyzing the surface by restricting dynamics to parallel or sequential population flow through the four states, we find that a 4 state sequential model provides a very good approximation to the full free energy landscape. Equilibrium molecular dynamics (MD) simulations are consistent with this minimal free energy landscape extracted from experiment: they reveal between 4 to 5 states sequentially populated during individual equilibrium trajectories started in the folded state conformation, but run near the unfolding temperature. MD simulation does not account for the increased base stacking in the loop observed in the experiment.

* Parts of this chapter have been submitted for publication by K. Sarkar, D.A. Nguyen and M. Gruebele.
Methods

RNA design and sample preparation

RNA oligonucleotides with 2-aminopurine modifications were purchased from Thermo Scientific (Waltham, MA). The following sequences were purchased: gacUA*CGguc, ga’cUACGguc, ga’c, UA*C. The capital letters indicate bases from the loop region, while lower case indicates those from the stem. The four RNAs will be referred to as loop mutant, stem mutant, stem control and loop control, respectively. The first two mutants are identical with only the position of the 2-aminopurine (a* or A* or 2AP) swapped. This was done to have the 2AP probe the stem and loop dynamics separately. The two 3-nucleotide sequences were used as controls. They form a fragment of the stem and loop sequences with the 2AP in the middle. 2AP itself was used as an additional control. The samples were dissolved in a low-salt P10E0.1 buffer (10 mM sodium phosphate and 0.1 mM Na2EDTA, pH 7.1). This buffer is known to favor hairpin conformation (15) in small RNA fragments. The RNA concentrations were measured by absorption spectroscopy using extinction coefficients supplied by Thermo Scientific.

Thermal denaturation

Thermal unfolding measurements were performed using a Carey-Eclipse Spectrofluorimeter (Varian Inc., Palo Alto, CA) for fluorescence and spectropolarimeter (JASCO, Easton, MD) for absorbance. Fluorescence was excited at 280 nm and the emission was scanned from 300 nm to 550 nm. The sample was heated in intervals of 5 °C. Data from these measurements are shown in Fig. 3.1A and Fig. B.1. Absorbance was monitored at 260 nm. A thermoelectric cooler along with a water bath was used to heat the sample in steps of 2 °C. Data from these measurements is shown in Fig. B.2. Integrated fluorescence melts were also performed with the spectropolarimeter. The optical filter Hoya B-370 (Hoya, Fremont, CA) used in the T-Jump experiments was used in this measurement of the fluorescence melt curves. This was done to check if the use of the filter changes the melt curves significantly. No significant change was observed. The results were reproduced within measurement uncertainty and checked for reversibility.

Temperature jump experiment

Temperature jump relaxation kinetics were measured with a custom-built instrument (16). The RNA sample was held in a 200 µm path length fused silica cuvette. Two counter-
propagating, 10 ns duration, Nd:YAG laser pulses Raman-shifted to a wavelength of 1.9 µm were used to jump the temperature of the sample. The energy of the laser pulse is dissipated as heat into the solution within 10 ps (17). The solution remains at the new elevated temperature for > 50 ms before diffusional cooling occurs. We use only the first 0.5 ms after the T-jump for our observation.

The relaxation of RNA population towards more unfolded populations was observed by exciting the 2AP chromophore every 14 ns at 281 ± 2 nm by a mode-locked, frequency-tripled Ti:Sapphire laser beam. The fluorescence decays were captured by a fast photomultiplier tube and digitized with a 500-ps resolution oscilloscope (Tektronix, Beaverton, OR). The data was downloaded to a computer and analyzed by a program written in LabWindows. (National Instruments, Austin, TX). The time \( t=0 \) position was determined from the Raman scattering of the T-Jump pulse. The first five decays after the T-Jump pulse were not considered in the data analysis.

Thus there are 28 data points at \( \Delta t = 0, 0.5 \ldots 13.5 \) ns per 14-ns fluorescence decay, and there are >35,000 such decays at \( t = 0, 0.014 \ldots 500 \) µs. To reduce the data volume of the raw kinetics traces, we performed “\( \chi \)-analysis” (16). Briefly, each normalized 28-point fluorescence decay \( f(t, \Delta t) \) was fitted to a linear combination of the initial decay profile \( f_1(\Delta t) \) just after the jump, and the final decay profile \( f_2(\Delta t) \) after 0.5 ms: \( f(t, \Delta t) = \chi(t)f_1(\Delta t)+(1-\chi(t))f_2(\Delta t) \). The coefficient \( \chi(t) \) can then be plotted to show how the RNA fluorescence decay changes from the initial fluorescence decay to the final fluorescence decay. The \( \chi(t) \) data was further logarithmically binned so that phases with different lifetimes contribute a comparable number of points to the overall kinetics.

**Global data analysis and fitting**

The thermal titration signal changes when the RNA is heated to populate different states with different 2AP fluorescence intensities and lifetimes. Likewise, the kinetic data relaxes after the T-jump as the RNA molecules populate different states on their way to a new equilibrium after the temperature jump. We performed a global non-linear least squares analysis of all thermodynamics and kinetics of both RNA mutants, using a single four state thermodynamic and kinetic model (4). The overall scheme is as follows:
Sequential \( N \leftrightarrow E \leftrightarrow U \leftrightarrow S \) and parallel models were also tested separately, by deleting the appropriate rate coefficients. The four states (N=native, E=on-path intermediate, U=maximally unfolded, S=completely unstacked) are discussed in detail in the Results section.

**Molecular dynamics simulation**

Molecular dynamics simulations of the stem and loop mutants were carried out using NAMD2 (18) with the CHARMM27 force field nucleic acid parameters (19). Parameters of 2AP were adapted from Sarzynska et al. (20). The starting structure of the RNA was derived from the pdb structure of 1Z31. Equilibrium simulations were performed with periodic boundary conditions in an NPT ensemble. A pressure of 1 atm was maintained using a Langevin piston. The particle-mesh Ewald method was used for the calculation of electrostatic forces. van der Waals interactions were calculated using a switching distance of 10 Å and a cutoff of 12 Å. Bonded, van der Waals and electrostatic interactions were updated at time steps of 1, 2, and 4 fs, respectively. A detailed description of the setup of the simulations can be found in (4).

A structure close to the native state was chosen for both the stem and the loop mutant from an equilibration run at 0 ºC. The temperature was then increased to 32 ºC and the gradual unfolding of the RNA hairpin monitored. The simulation for both the loop and the stem mutant were run for 50 ns at 32 ºC. It was previously noted that the CHARMM27 force field has a tendency of unfolding RNA hairpins at a temperature lower than their experimental folding temperature. In total more than 0.5 \( \mu \)s of simulations were run to gain insight into the folding landscape of the RNA hairpin.

The MD simulated RNA structures were monitored for base stacking in the stem and loop, and for hydrogen bonding in the stem, following the procedure described in reference (4). All possible pairwise base stacking interactions between the four bases in the loop were monitored. For the stem, only the base stacking interactions present in the native state were monitored.
Results

Thermal denaturation

2AP fluorescence is sensitive to base stacking (14). Fluorescence melts were obtained for the stem mutant ga*cUACGguc, the loop mutant gacUA*CGguc, the short loop and stem controls, and 2AP. Fig. 3.1A shows the fluorescence intensity melts of the two mutants divided by their corresponding controls (see Appendix B Fig. B.1 for full data). At 20 °C, the stem mutant has 23 times lower fluorescence intensity than the loop mutant. Base stacking in the native stem severely quenches the fluorescence of 2AP, whereas base stacking is nearly absent in the native loop. The fluorescence melt for the stem mutant is similar to that obtained for the sequence ga*cUUCGguc from ref (4).

As the RNA is melted, 2AP fluorescence from the loop decreases monotonically (Fig. 3.1A), while stem fluorescence increases. Thus melting the RNA reduces native base stacking in the stem, but increases non-native base stacking in the loop. At the highest temperatures, the loop is slightly more quenched than the stem.

We also studied the ultraviolet absorbance and fluorescence emission lifetime upon melting of the two mutants (Figs. B.2 and B.3, Appendix B). The fitted melting temperatures of the two RNA mutants obtained from absorbance are identical within measurement uncertainty (±1 °C), indicating that the introduction of 2AP into the loop or stem does not differentially perturb the stability of the hairpin. The fluorescence lifetime is long for the loop mutant, which has little base stacking. It is short for the stem mutant because base stacking quenches the fluorescence. The temperature trends upon RNA melting are opposite for loop and stem: the fastest part of the fluorescence decay speeds up for the loop because quenching increases upon melting, consistent with gain of base stacking. The fluorescence decay slightly slows down in the stem because quenching decreases upon melting, consistent with loss of base stacking.

Temperature jump kinetics

Temperature jump relaxation kinetics for the loop and stem mutants are shown in Fig 3.1 (bottom panels). Laser T-jumps were performed at 10 °C intervals with 14 ns time resolution. 2-AP was excited by a UV pulse every 14 ns, and its fluorescence decays were collected to monitor the progress of the kinetics. In Fig. 3.1, the shape of the fluorescence decays evolves from the initial state immediately after the T-jump (χ = 1) to the final state observed at 0.5 ms (χ
In order to improve the signal-to-noise the data over 280 ns were binned together.

As can be seen in Fig. 3.1, the stem mutant produced several positive kinetic phases, while the loop mutant also produced a slow negative kinetic phase at some temperatures. The appearance of multiple phases indicates several stages of folding with interconversion between more than two thermodynamic states.

Analysis of the fluorescence decay shapes (Fig. B.3, Appendix B) shows that only the fastest part of the fluorescence decay reports on kinetics: it speeds up and overall intensity decreases when base stacking increases. The slow part of the fluorescence decay (> 4 ns) does not report on resolvable kinetics. It simply switches in < 10 ns during the temperature jump.

The data for the stem mutant had the best signal-to-noise ratio at a T-jump size of 19 °C and concentration of 430 μM. In the case of the loop mutant, 10 °C T-jumps provided good enough signal-to-noise at a concentration of 165 μM. Data were obtained for multiple T-jump sizes and concentrations. No jump size dependence or concentration dependence could be observed, in agreement with previous findings for other RNA hairpins (2, 4). T-Jumps for the short stem and loop controls were also performed. They resulted in instantaneous responses with no resolvable kinetic phase (Fig. B.3, Appendix B). Thus the resolved kinetics are due to (un)folding relaxation of the RNA hairpin.

**Global thermodynamic and kinetic model**

A global four state model simultaneously fitted all the experimental data. Thus the stem and loop mutants can be described by the same free energy landscape. We obtained an excellent fit with a full model that allows interconversion between all four states $N, E, U$ and $S$ (eq. (1)). A simpler sequential model connecting only pairs of states $N \leftrightarrow E \leftrightarrow U \leftrightarrow S$ did nearly equally well, and is shown in Fig. 3.1. Based on MD simulation results, the four states are classified as native ($N$), an on-path intermediate frayed in the stem ($E$), an unfolded state with only some base stacking left ($U$), and a completely unstacked off-pathway intermediate ($S$).

We fitted other models with variable baselines, numbers of states, thermodynamic parameters, and barriers between states. Fewer than 4 states did not fit the experimental data within measurement uncertainty. Fitting the stem and loop mutants separately also yielded excellent fits, but no better than the global fit.
The quantitative results of the global model are shown in Table 3.1 (sequential) and Table 3.2 (full).

In our model, each state “j” = “N, E, U, S” is assigned a temperature-dependent signal baseline for fluorescence intensity $I_j$ and fluorescence decay shape $\chi_j$: $I_j(T) = I_j(T_0) + I_j^{(1)}(T - T_0)$ and $\chi_j(T) = \chi_j(T_0) + \chi_j^{(1)}(T - T_0)$. Each state is also assigned an adjustable melting temperature $T_m$ and a free energy given by $\Delta G_j(T) = \Delta G_j^{(1)}(T - T_m)$. A linear expansion of the free energy was sufficient. Finally, states are connected by free energy barriers $\Delta G_{ij}^\dagger(T) = \Delta G_{ij}^\dagger(0) + \Delta G_{ij}^\dagger(1)(T - T_0)$. Populations and kinetics for the model were evaluated from the equilibrium constants, $K_{ij} = \exp[-(\Delta G_i - \Delta G_j)/RT]$ and by solving the kinetic master equation for scheme (1) with rate coefficients $k_{ij} = k_m \exp[-\Delta G_{ij}^\dagger / RT]$ (2, 4). The modeled thermodynamic populations of the four states, and a representative set of simulated kinetic population decays are shown in Fig. 3.2.

Fig. 3.3 plots a minimalist free energy landscape obtained from the sequential model as a function of two reaction coordinates $S_j = -\frac{\partial G(T)}{\partial T}$ (state entropy) and $I_j$ (state fluorescence intensity). In principle, any smooth experiment-derived state function such as $S_j$, $I_j$, or $\chi_j$ could be used as a reaction coordinate. We used $S$ because it correlates with the overall disorder of the RNA chain, and $I$ because it monitors base stacking (lower $I$ = greater base stacking). The thick arrows in Fig. 3.3 show the sequential path, and the thin arrow shows an additional path between $N$ and $U$ that appears in the full model.

**MD Simulations**

Equilibrium MD simulations were run near the melting temperature of the hairpin to sample different conformations during RNA unfolding. 50 ns long trajectories were obtained at 32 °C for the stem and loop mutants (Fig. 3.4). Starting configurations for the two mutants were sampled from a native ensemble obtained by relaxing an initial folded structure at 0 °C for 5 ns.

The trajectories were analyzed for native base stacking in the stem, native hydrogen bonding in the stem, and base stacking in the loop. A plot of base stacking in the stem and loop vs. hydrogen bonding in the stem is shown in Fig. 3.4. Regions where the trajectory was caught in a metastable state are shown within circles. Four or five such metastable states are seen for each trajectory, and they provide structural models for the states observed by experiment.
Representative structures of \( N \), \( E \), \( U \), and \( S \) from each run are shown in Fig. 3.5. For the stem mutant in Fig. 3.5A, the stacked stem bases are shown in red while the 2AP is shown in purple. All the hydrogen bonds and stacking are intact in native structure \( N \). In the frayed state \( E \) the single unstacked base is shown in yellow. The unfolded state \( U \) shows both stacked and unstacked bases in the stem. Finally in the unstacked state \( S \) all the stacking is lost. Similar representative structures for the loop mutant are shown in Fig. 3.5B. The loop does not have any stacked bases in the native state. The unstacked bases in the loop are shown in blue while the 2AP is shown in purple. Base stacking in the loop increases remains constant or increases slightly at higher temperature, as shown in the unfolded structure with the stacked base in silver. After running for 50 ns at the highest temperature, the loop mutant also populates a completely unstacked state.

**Discussion**

2AP turns out to be a conservative probe for RNA stability and folding dynamics when substituted for adenine in stem or loop positions. Both RNA mutants we studied can be fitted by the same global thermodynamic/kinetic model (Figs. 3.1-3.3). In addition, the absorption melt curves at 260 nm of the stem and loop mutant are shifted from each other by less than 2 °C, within the experimental error (Fig. B.2). Finally, molecular dynamics trajectories of the two mutants (Fig. 3.4) go though similar structural stages. Hence 2AP substitution does not extensively alter the overall stability of a small RNA hairpin, nor the activation barriers connecting different states. Since 2AP does not cause a significant perturbation in a10-mer RNA, it should be even more applicable as a minimally invasive probe for larger RNAs.

Previous experiments have already shown that 2AP exhibits site-specific fluorescence responses (13). Our results here are in agreement with these observations (Fig. 3.1). The 2AP fluorescence intensity decreases upon base stacking, as does the lifetime of its fastest fluorescence decay phase. To reduce temperature-dependent intrinsic baselines in the 2AP fluorescence intensity, the use of trinucleotides as controls is recommended (Fig. B.1). To use 2AP in fluorescence lifetime studies, the <4 ns phase must be resolved: the slow decay phase of 2AP switches instantaneously (< 10 ns) upon temperature jump, and appears to be related with an intrinsic conformational response of the nucleotide, not a probe of its local environment.
The global experimental fit of the thermal denaturation and kinetic relaxation data is consistent with the gacUACGguc hairpin making a steady progression from more folded states with high stem stacking and low loop stacking, to less folded states with low stem stacking and high loop stacking. The transition is not a simple two-state process, but samples at least 4 states, which are all significantly populated during relaxation kinetics (Fig. 3.2). Molecular dynamics trajectories monitoring the unfolding of the hairpin mutants likewise show 4-5 metastable states where the trajectory temporarily resides before moving on to a more unfolded state (circles in Fig. 3.4). Representative long-lived structures are shown in Fig. 3.5.

Fig. 3.3 shows that it is possible to extract a low-dimensional free energy landscape from the thermodynamic and kinetic data. The black arrows show the main population flows, and the letters \(N\), \(E\), \(U\) and \(S\) can be tentatively identified with structural states as shown in Fig. 3.5. Such an identification must be tentative: the reaction coordinates \(I\) and \(S\) in Fig. 3.3 are not the same as the computational reaction coordinates of base stacking and hydrogen bonding between bases in Fig. 3.4. Perhaps the hardest tasks for molecular dynamics simulators nowadays is not the simulations themselves, nor even the inaccuracies of the force fields, but computing experimental observables instead of purely theoretical coordinates such as “number of native contacts.”

The structures in Fig. 3.5 do match up with the experimental observations qualitatively as follows. The \(N\) structure corresponds to the native state, with maximal base stacking in the stem, and minimal base stacking in the loop. \(E\) is a frayed structure where the stem has lost some hydrogen bonding and some base stacking. \(U\) is a largely unfolded state, but some single stranded base stacking and non-native loop base stacking remains. \(S\) is a highly unstacked state, with a low probability of interconverting directly to the native state even in the full global model.

The thicknesses of the arrows in Fig. 3.3 indicate which interconversions among states are most likely based on the global model whose parameters are summarized in Table 3.1. It is clear that the sequential model \(N \Leftrightarrow E \Leftrightarrow U \Leftrightarrow S\) is not a bad approximation to the full dynamics on the two-dimensional free energy landscape. This is in agreement with our global fitting: a sequential model (with the parallel rate coefficients in scheme 1 deleted) is nearly as good as the full model, whereas a parallel model produced a noticeably lower quality fit. Multiple parallel and sequential pathways are involved in the folding of this small RNA hairpin, but the sequential scheme provides an adequate, if oversimplified, description of the folding-unfolding process.
Tables

Table 3.1: Equilibrium thermodynamic parameters and activation energy parameters of the sequential four state model. Errors shown are two standard deviations.

<table>
<thead>
<tr>
<th>State</th>
<th>Tm (ºC)</th>
<th>ΔG(1) (kJ.mol⁻¹.ºC⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native</td>
<td>48.43 (10)</td>
<td>0.4926 (12)</td>
</tr>
<tr>
<td>FrayEd</td>
<td>50.90 (10)</td>
<td>0.2525 (12)</td>
</tr>
<tr>
<td>Unfolded</td>
<td>53.85</td>
<td>0.0000</td>
</tr>
<tr>
<td>UnStacked</td>
<td>59.26 (22)</td>
<td>-0.3501 (40)</td>
</tr>
</tbody>
</table>

Table 3.1a.

<table>
<thead>
<tr>
<th>Transition state</th>
<th>ΔG(0)† (kJ mol⁻¹)</th>
<th>ΔG(1)† (kJ mol⁻¹.ºC⁻¹)</th>
<th>kₘ (µs⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native → FrayEd</td>
<td>8.56 (30)</td>
<td>0.0478 (22)</td>
<td></td>
</tr>
<tr>
<td>FrayEd → Unfolded</td>
<td>12.77 (22)</td>
<td>-0.036 (16)</td>
<td>1.000</td>
</tr>
<tr>
<td>Unfolded → UnStacked</td>
<td>19.42 (26)</td>
<td>0.049 (24)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1b.
Table 3.2: Equilibrium thermodynamic parameters and activation energies of the four state model summarized in eq. (1). Errors shown are two standard deviations.

<table>
<thead>
<tr>
<th>State</th>
<th>$T_m$ (°C)</th>
<th>$\Delta G^{(1)}$ (kJ.mol$^{-1}$.°C$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native</td>
<td>48.25 (02)</td>
<td>0.5743 (10)</td>
</tr>
<tr>
<td>FrayEd</td>
<td>55.48 (04)</td>
<td>0.3521 (08)</td>
</tr>
<tr>
<td>Unfolded</td>
<td>53.85</td>
<td>0.0000</td>
</tr>
<tr>
<td>UnStacked</td>
<td>71.54 (24)</td>
<td>-0.2400 (50)</td>
</tr>
</tbody>
</table>

Table 3.2a.

<table>
<thead>
<tr>
<th>Transition state</th>
<th>$\Delta G^{(0)\dagger}$ (kJ mol$^{-1}$)</th>
<th>$\Delta G^{(1)\dagger}$ (kJ mol$^{-1}$.°C$^{-1}$)</th>
<th>$k_m$ (µs$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native $\rightarrow$ FrayEd</td>
<td>13.82 (14)</td>
<td>0.207 (18)</td>
<td>1.000</td>
</tr>
<tr>
<td>FrayEd $\rightarrow$ Unfolded</td>
<td>9.40 (11)</td>
<td>0.246 (10)</td>
<td></td>
</tr>
<tr>
<td>Unfolded $\rightarrow$ UnStacked</td>
<td>20.28 (24)</td>
<td>0.033 (22)</td>
<td></td>
</tr>
<tr>
<td>Native $\rightarrow$ Unfolded</td>
<td>25.73 (48)</td>
<td>-0.652 (22)</td>
<td></td>
</tr>
<tr>
<td>FrayEd $\rightarrow$ UnStacked</td>
<td>30.25 (26)</td>
<td>-0.919 (15)</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2b.
Fig. 3.1: Experimental thermal denaturation and temperature jump data fitted by a four state model. (A) Thermodynamic data and model. (B and C) Relative fluorescence lifetime change and model fit for the loop mutant at two of the six temperatures measured (see also Appendix B). (D and E) Relative fluorescence lifetime change and model fit for the stem mutant at two of the five temperatures measured.
Fig 3.2: Populations thermodynamics and kinetics from the sequential fit (A). Thermodynamic populations from the 4 state sequential model. (B-E) Population decays resulting from the sequential fit.
Fig 3.3: Free energy landscape at 50 °C (near $T_m$) derived from the sequential fit ($\chi^2 = 0.41$) in Fig. 3.1. Entropy (S) and Intensity (I) have been used as the reaction coordinates. The values for the coordinates and the relative free energies are obtained from the sequential model. Bold arrows indicate the sequential path. A thin arrow between N and U has been used to indicate the only other contribution from the full model that yields a slightly better fit ($\chi^2 = 0.40$).
Fig 3.4: Molecular dynamics trajectories of RNA hairpin unfolding at 32 °C, starting with two native-like structures equilibrated at 0 °C: loop mutant (blue) and stem mutant (red).
Fig 3.5: Representative structures from the regions of the stacking-hydrogen bonding plot where individual trajectories spend most of their time. The configurations that were sampled in figure are labeled there. States are labeled with bold letters as in Fig. 3.2, although the correspondence between simulated and measured states is only approximate.
References


6. Hyeon, C., and D. Thirumalai. 2008. Multiple probes are required to explore and control the rugged energy landscape of RNA hairpins. Journal of the American Chemical Society 130:1538-1539.


Chapter 4

Pushing β-sheet proteins towards the folding speed limit with predictions from very long MD simulations*

Introduction

The protein folding problem has been a major challenge for the last fifty years (1, 2). A key question in this regard is whether we can predict the folding pathway(s) and the time taken by proteins to reach their folded state. The problem has motivated efforts to obtain a better description of the energy landscape involved in the folding process. It has been approached both by experimental and computational methods. Experiments using various probes and techniques including fluorescence, IR, CD and NMR have been employed to characterize the folding energy landscape of proteins. The use of multiple probes provides a more detailed view as they may access different reaction coordinates within the energy landscape.

Temperature jump experiments have contributed significantly to the understanding of fast folding proteins (3, 4). They have enabled us in designing even faster folders and helped us to push protein folding towards its speed limit. Fast folders allow folding timescales to be accessed by MD simulations. The speed limit for alpha helical proteins has already been experimentally demonstrated (3). Beta sheet proteins however, need to be designed better in order to reach such time scales. Experimental surveys on the WW domain, one of the smallest beta sheet structures, have yielded fast folding candidates that have an observed relaxation time as short as 10 µs (4). However, can even faster time scales for the folding of beta sheet proteins be achieved through better design?

While experimentalists were pondering on these questions all atom molecular dynamics simulations of proteins were reaching the fast folding time scales set by experimentalists (5-7). The advent of faster processors, parallelization (8), specifically designed computer architecture (9), and efficient algorithms (10) have only recently made long folding trajectories possible. This has allowed better parameterization of the molecular dynamics force fields for proteins in

* Parts of this chapter have been submitted for publication by S. Piana, K. Sarkar, K. Lindorff-Larsen, M. Guo, M. Gruebele and D.E. Shaw.
retrospect (11). Here we describe how insights gained from all-atom MD simulation of the WW domain protein Fip35, with explicit solvent, have allowed us to design a faster folding ultra-stable mutant. The predicted mutations were experimentally performed and the folding kinetics and thermodynamics tested. The results are in excellent agreement with the MD simulations performed before the experiments.

This chapter is the result of a collaboration with the D. E. Shaw research group. The computational part of the work was performed jointly by Kresten Lindorff-Larsen and Stefano Piana while the experimental part was performed by me and Minghao Guo.

Background

A survey (4) performed on the folding timescales of different variants of the WW domain protein showed us that a chimera of the human Pin WW domain and the Formin binding protein is one of the fastest folders having an observed folding time of ~10 µs. This protein is called the Fip35. The structural differences between the human Pin WW domain and the chimeric protein Fip35 are shown in Fig. 4.1. Two serines in loop 1 are replaced by an aspartic acid while the second tryptophan near the C-terminal is mutated to a phenylalanine. The protein consists of a three stranded beta sheet and a small hydrophobic core.

Following the discovery of this fast folding variant of the WW domain (4), Freddolino et al. (7) looked at the folding dynamics of this protein using all atom molecular dynamics simulation. They performed MD simulation of the protein in explicit solvent for 10 µs, near the melting temperature of the protein, starting from an extended conformation. The number of native contacts, secondary structure and exposed hydrophobic surface was monitored throughout the trajectory. The protein collapses to a molten globule state within 500 ns. This hydrophobic collapse is concurrent with the formation of alpha helices within residues 8-20 and 27-34. This is followed by a series of interconversions between multiple states having primarily helical conformations. The persistence of alpha-helical conformations over the entire length of the trajectory was anomalously long. This inability to fold the beta sheet protein to its correct folded state motivated further analysis. In a follow up work by the same author (12) it was shown that the alpha helical state observed in the simulation was the most stable state based on the CHARMM22 force field with CMAP corrections used in this simulation. The experimentally observed native state had a higher energy in this force field. This work clearly demonstrated the
need for improving the force field parameters to obtain results closer to reality. Ensign et al. (13) simulated the folding dynamics of Fip35 using implicit solvent and the AMBER96 force field. A total length of 2.73 ms of simulation was generated using “Folding@home” (8) and analyzed. Only a small number of folding events were observed and two distinct beta sheet conformations were obtained essentially indicating heterogeneity in the folding landscape. The lack of sufficient folding and unfolding events prevented detailed kinetic analysis.

In 2008, Shaw et al. developed a special purpose super computer, Anton (9), for performing MD simulations two orders of magnitude faster than currently possible. This dramatically changed the accessibility to very long time scale MD simulations. Equilibrium MD simulations near the melting temperature of Fip35 were performed on Anton. Multiple folding and unfolding transitions were observed (14). This allowed direct comparison to already available folding rates. Furthermore, based on these simulations accurate predictions were made possible.

Materials and Methods

Molecular Dynamics

Molecular dynamics simulations were performed in Anton (9), a computer specifically built for the purpose. The simulation of Fip35 along with the simulation conditions are described in detail in ref. (14). GTT was built in an extended conformation and solvated in a cubic box of ~4000 TIP3P water molecules having a length of ~50 Å. The systems were simulated using the Amber ff99SB-ILDN force field that includes recently improved side-chain torsion parameters (11). Four MD simulations of reversible folding of GTT (of lengths 83 µs, 118 µs, 124 µs and 272 µs) were performed. A series of shorter simulations of both GTT and FiP35, were run until the first folding event occurred. These additional simulations totalling 110 µs and 70 µs for FiP35 and GTT respectively were used to increase the accuracy of the calculated folding rates. A reaction coordinate was determined based on the method developed by Best and Hummer (15).

Sample preparation

The Fip35 sequence incorporated in a pGEX-2T vector was used for protein expression using the Rosetta DE3 pLysS cell line (EMD Chemicals). The cells were grown in LB media and
induced using IPTG. They were then centrifuged and lysed using a sonicator. Purification of the protein was performed by running it through a Glutathione Sepharose column. It was dialyzed extensively with phosphate buffer at pH = 7.0. Thrombin digestion was performed to cleave the GST tag from the protein. In order to remove the GST the digested protein solution was passed through a 10 kD Amicon membrane (Millipore, Billerica, MA). The purified protein was filtered out. Single point mutations to the peptide were performed using the QuickChange Site-Directed Mutagenesis Kit (Agilent Technologies, Santa Clara, CA). Two mutants were expressed. The first is a single mutant N26G and the second is a triple mutant N26G, A27T and S28T. The expressed proteins were tested for their purity by running SDS page and performing ESI mass spectrometry.

**Thermal denaturation**

Thermal denaturation measurements for Fip 35 and the mutants were performed using a spectropolarimeter (Jasco Inc, Easton, MD). The spectropolarimeter is used to measure both circular dichroism and fluorescence. The temperature of the sample is controlled by a peltier controller supported by a water bath. The same optical filter Hoya B370 (Hoya Optical, Fremont, CA) is used for the fluorescence denaturation measurements as is used for the temperature jump setup. As the samples were all found to have high melting points, 2M and 4M guanidinium chloride solutions were used to access the unfolded baselines.

The thermal denaturation curves were fit to the free energy expression as in eqn(1) along with linear signal baselines as in eqn (2).

\[ \Delta G = \Delta G^{(1)}(T-T_m) + \Delta G_{Ga}[GuHCl] \]  

\[ S = S^{(0)} + S^{(1)}(T-T_m) \]  

\[ T_m \] is the folding mid-point. A global fit was performed for each mutant using the Marquardt-Levenberg non-linear least squares fitting algorithm.

**Temperature jump measurement**
Temperature jump measurements were performed in a custom built setup (16). A 10 ns Raman shifted pulse from a YAG laser at a wavelength of 1.9 μm is used to jump the temperature of the sample solution in a 200 μm path length fused silica cuvette. The sample for temperature jump measurements was dissolved in a phosphate buffer having 30% heavy water so as to be able to compare with earlier measurements (17). The heavy water allows deeper penetration of the T-Jump pulse and hence produces a more uniform heating profile within the cuvette. A frequency tripled Titanium Sapphire laser is used to monitor the time resolved tryptophan fluorescence in the protein sample. The ~100 fs, 280 ± 1 nm pulses interrogate the sample every 14 ns. The fluorescence is collected by a photomultiplier and then digitized using a Tektronix oscilloscope and finally downloaded to a computer for analysis. The oscilloscope digitizes the signal every 500 ps, resulting in 28 data points per decay.

The time (t = 0) position for the data was determined from Raman scatter at the position of the T-jump. Three decays right after the jump were not analyzed. Thereafter the data was binned over 20 decays. This binned data was further analyzed by fitting a linear combination of the fluorescence decays right after the jump (f1) and 450 μs after the jump (f2). Thus any intermediate decay is fit to \( f = \chi f_1 + (1-\chi)f_2 \). \( \chi \) is then plotted against time to obtain the relaxation. Single exponential fits to the relaxation curves were then obtained using the Igor Pro software package.

Results and Discussion

Molecular Dynamics Simulations

MD simulations could provide non-intuitive strategies beyond the introduction of hydrophobic residues or modification of active sites (e.g. shortening loop 1) currently used in experimental redesign (18). In our simulations of FiP35, we found that the loop in the first hairpin formed easily and quickly. Interestingly, we also observed in our simulations that in the TSE, the third β-strand in FiP35, which is not fully formed at this point during folding, can make both native and non-native contacts with the second strand in the first hairpin. In particular, we found that hydrogen bonds between strands two and three have a non-negligible fraction of their native state energy present already in the TSE.

These observations suggested a viable strategy to further speed up folding of the already fast FiP35 WW domain. The TSE can be stabilized even more by increasing the propensity of
the third strand to form extended conformations, thereby increasing the interactions between the second and third strand in the TSE. To this end, we introduced three mutations in the second hairpin of FiP35. First, we noted that asparagine residue N26 in the loop between the second and third strand has a positive value for the backbone $\phi$-dihedral. While asparagine is highly favored for this position in the Ramachandran plot, glycine is even more favored (19) among natural amino acids. We therefore decided to mutate N26 to glycine, thus enhancing the formation of the turn between the second and third $\beta$-strands (20). Two additional mutations were motivated computationally by the fact that amino acids branched at the $\beta$-carbon stabilize extended conformations. Analysis of the amino acid sequences in WW domains showed that threonine, a $\beta$-branched amino acid, is tolerated in positions 27 and 28. We thus substituted A27 and S28 with threonine. We will refer to the triple mutant resulting from these three mutations (N26G, A27T, S28T; Fig. 4.1) as GTT.

C$_\alpha$ RMSD for residues 4 to 32 with respect to the crystal structure (pdb id: 2F21) was calculated for the entire length of the simulation run. Clear folding and unfolding transitions were observed as shown in Fig. 4.2. An overlay of the crystal structure of a variant of human Pin 1 (pdb id: 2F21) along with the folded structures obtained from the simulation of GTT and Fip35 is shown in Fig. 4.3. The reaction coordinate was obtained from the simulation using the method developed by Best and Hummer (15). Fig. 4.4 shows the difference in the 2D free energy profile for FiP35 and the GTT mutant. The unfolded state for GTT is destabilized with respect to FiP35.

We calculated the folding rate and stability of the GTT variant in silico by performing very long MD simulations at 395 K, a temperature that we previously (14) found was close to the melting temperature for FiP35 with the force field used (see Methods). We observed reversible folding and unfolding of GTT multiple times in the long MD trajectories at 395K (Fig. 4.2). GTT folded to a structure similar to Fip35 (Fig. 4.3). Based on a total of 35 folding events observed for GTT, we calculated a folding time of $7\pm1$ $\mu$s, twice as fast as FiP35. We also carried out additional simulations of FiP35 to increase the accuracy of our calculations of its folding time, which we found to be $13\pm3$ $\mu$s based on 16 observed folding events. It thus appeared that in silico the GTT mutation indeed had increased the folding rate and lowered the barrier to folding (Fig. 4.4). We also found the stability of GTT to be higher than that of FiP35: in the simulations the population of the folded state was 61% for FiP35 and 74% for GTT, corresponding to an increased stability of 0.5 kcal mol$^{-1}$ at 395K and an estimated increase in the
melting temperature of 6 K. The increased stability of GTT *in silico* is entropic in origin; while the folding of FiP35 is accompanied by a drop in entropy of 78.2 cal mol\(^{-1}\) K\(^{-1}\), the corresponding value for GTT is only 71.3 cal mol\(^{-1}\) K\(^{-1}\) (\(\Delta H\) for folding is -31.5 kcal mol\(^{-1}\) for FiP35 and -29.0 kcal mol\(^{-1}\) for GTT).

**Experimental Measurements**

To test the computational predictions, thermal unfolding measurements were performed on Fip35, and on its N26G and GTT mutants, to determine their stability. Laser temperature jump experiments were performed to determine the change in folding rate.

**Thermal denaturation**

Thermal denaturation measurements for the three proteins are shown in Fig. 4.5. The melting mid points for Fip35, N26G and GTT are 360 K, 369 K and 371 K respectively. The high stability of the proteins made it impossible to determine the unfolded baselines. Guanidinium chloride was used to destabilize the proteins and access the unfolded baselines.

**T-jump kinetics**

The observed relaxation rate measured at 70 °C for Fip35 provides a single exponential fit with a time constant of 13 ± 4 µs whereas the same for N26G at 80 °C yields 5.6 ± 0.4 µs and that for GTT at 80 °C yields 3.7 ± 0.4 µs as shown in Fig. 4.6. The different temperatures are chosen so as to measure the rate at the points of similar thermodynamic stability for the two proteins.

As predicted by molecular dynamics, GTT was found to be more stable than Fip35. Stability was determined by monitoring the fluorescence of the single tryptophan residue during thermal titration. The fluorescence intensity response is shown in Fig. 4.5. To determine reliable free energy parameters and melting temperatures, both temperature and denaturant (guanidine hydrochloride) were scanned to access the midpoints of the melting transitions. The folding midpoints were 360±2 K for Fip35, 369±2 K for FiP35 N26G, and 371±2 K for GTT. GTT melts 10 degrees higher than FiP35, slightly higher even than predicted by molecular dynamics.
As predicted by molecular dynamics, GTT folded considerably faster than FiP35. Temperature jump measurements were performed on Fip35 and its GTT mutant at final temperatures chosen to yield equal stabilities for both proteins, so rates could be compared directly (Fig. 4.4b). The equilibrium constants for both proteins were ≈2, in favor of the native state. The observed relaxation time of 3.7 μs for the GTT mutant was three times faster than for Fip35 itself. If we make the two-state assumption (whose validity may be limited for such fast rates (21), the mean refolding time of FiP35 at 353 K is 20 μs⁻¹, and the mean refolding time of GTT at 363 K is 5.7 μs⁻¹.

The experimental results validate the computational protein engineering. The FiP35 refolding rate is slightly faster computationally than experimentally, not unexpected for the 40 K higher temperature of the simulations and the resulting smaller solvent friction. The measured stability and relaxation rate of the GTT mutant both increase as predicted by molecular dynamics. It is remarkable that molecular dynamics is now capable of making rational design predictions when the free energy changes involved are on the order of 1 k_BT. The benchmark for force fields is no longer just whether they can predict a correct small protein fold (see Fig. 4.3), but whether they can determine ΔΔG values, changes in the free energy upon perturbation of a protein’s amino acid sequence or solvent environment.

With the simulation in hand, we can analyse the reasons for the stability shift from FiP35 to GTT in more detail. Our simulations show that the order of events during folding is the same for GTT and FiP35, and that the structure of the native states and the TSEs are very similar. On the other hand, we observe a substantial difference between the unfolded states of FiP35 and GTT. In particular, we found an increased amount of residual structure in the unfolded state of GTT, with residues 27-31, which form the third β-strand in the native state, having an increased population of β-strand structure. The increased propensity to form a turn at residue 26 and extended structure at residues 27 and 28 intended in the GTT design thus did increase the intrinsic stability of the third β-strand. We speculate that such “pre-organization” of the unfolded state to increased native-like structure may be a general property of the fastest-folding proteins. For example, the unfolded state of both the villin headpiece (22) and engrailed homeodomain (23) are highly helical, and it appears that unfolded state helicity of homeodomains correlates with folding rate (24). Too much of a good thing (e.g. helix propensity
in a helical protein) can however slow down folding and create intermediates when the protein previously was a fast two state folder (25).
Fig. 4.1: Structure of the Pin1 WW domain showing the required mutations for Fip35: Two serines in loop 1 are mutated to aspartic acid and the tryptophan in strand 3 is mutated to phenyl alanine.
Fig. 4.2†: Reversible folding simulations of Fip35 and GTT. Root mean square deviation (RMSD) in two representative 100 $\mu$s MD simulations of Fip35 and the GTT variant. The C$_\alpha$-RMSD was calculated for residues 4 to 32 with respect to the crystal structure pdb id 2F21.

† Figure courtesy S. Piana, K. Lindorff-Larsen and D.E. Shaw
Fig. 4.3†: WW domain structure. Superimposition of the crystal structure of a variant of the hPin1 WW domain with loop 1 shortened (PDB entry 2F21, green) and the structures obtained from the simulations of Fip35 (blue) and GTT (red); for clarity only backbone atoms are shown. The inset shows the backbone of the turn in the second hairpin; the side-chains of the residues that were mutated in GTT are also shown and labelled (Fip35 sequence in blue and GTT sequence in red).

† Figure courtesy S. Piana, K. Lindorff-Larsen and D.E. Shaw
Fig. 4.4†: Characterization of the free energy landscape and unfolded states of Fip35 and GTT. Free energy as a function of the optimal reaction coordinate. Higher values of the reaction coordinate correspond to more unfolded structures.

† Figure courtesy S. Piana, K. Lindorff-Larsen and D.E. Shaw
Fig. 4.5: Thermodynamic melting curves for Fip35, N26G and GTT. The melts were performed in 0 M, 2 M and 4 M Guanidinium hydrochloride in order to access the unfolded baselines. The dots are experimental data and the lines represent fits to the thermodynamic model described in the text. The melting midpoint temperatures based on the model fit are shown on the top right.
Fig. 4.6: Relaxation rates for the three proteins. The relaxation rates for Fip35, N26G and GTT are shown from top to bottom based on a single exponential fit. The choice of temperatures for each T-jump was made so as to access the same thermodynamic regions of the folding curves for the different proteins.
References


Chapter 5

Binding of ribosomal signatures S4 and h16*

Introduction

This chapter will discuss the nature of the interaction between the ribosomal signatures S4 and h16. All the computational work in this chapter was performed by Ke Chen and John Eargle from Zan Luthey-Schulten’s group while I performed the experiments.

Ribosomal signatures as the name suggests are characteristic of the three domains of life: Bacteria, Archaea, and Eucarya. It was first identified by Carl Woese more than two decades ago (1-3). The positions of the signatures in the rRNA of the ribosomal small subunit (SSU), were instrumental in constructing the universal phylogenetic tree which is still used today in the classification of biological organisms (4). As such, these molecular fossils offer insight into the early evolution of the translational apparatus. With the rapid growth of genomic and structural data, the definition of signatures has been extended to include structure motifs in the rRNA and r-proteins and arrangements of genomic content that are unique to one domain of life. In a study that includes over 90,000 16S and 23S sequences, Roberts et al. (5) demonstrated that the sequence and structure signatures of rRNA account for 50% of the phylogenetic separation between Bacteria and Archaea. Correlations between the rRNA and r-protein signatures show that the rRNA signatures coevolved with both domain-specific r-proteins and inserts in universal r-proteins.

The largest bacterial rRNA structure signature with such a coevolutionary protein partner is found in the five-way junction of the 16S rRNA 5’ domain, which is held together by the universal r-protein S4. The N-terminal bacterial signature of S4 (S4N) interacts specifically with helix 16 (h16), and a bioinformatic study showed that cysteine composition in this region varies within bacterial phylogeny in terms of its zinc binding ability (6). The five-way junction contains additional sequence and structure signatures in helix 17 (h17) and the pseudo-knot of helix 18 (h18).

* Parts of this chapter have been submitted for publication by K. Chen, J. Eargle, K. Sarkar, M. Gruebele and Z. Luthey-Schulten.
The Nomura map developed in the 1970’s (7, 8) showed that the assembly of the small subunit in bacteria was dependent on the presence of S4 and five other primary binding proteins that are required to be in place before the remaining fourteen r-proteins can be incorporated. S4, S17 and S20 bind directly to the 5’ domain of the 16S rRNA and the other primary binders S15/S8 and S7 bind to the central and 3’ domains respectively. Two of them, S4 and S7, were later identified as the only assembly-initiator proteins based on their noncooperative binding during the onset of assembly (9). More recently, the Nomura dependency map has been extended to include information about the kinetics and folding pathways for the assembly of the 30S ribosome. Using pulse-chase experiment monitored by quantitative mass spectrometry, Williamson and coworkers determined the binding rates of each ribosomal protein in the small subunit and found that the 5’ domain proteins, especially S4, bind more quickly than the proteins in the central or 3’ domain (10, 11), indicating a 5’ to 3’ directionality in the assembly process.

The folding of the 5’ domain of 16S rRNA was studied with time-resolved hydroxyl radical foot printing which established the time dependence for the formation of tertiary contacts (12). In the absence of proteins, the structure signature h16, folds earliest under a wide range of ion concentrations. The “minimal” rRNA binding site for the complete S4 (13) was established from deletion and mutation studies on the RNA. The measured binding free energies show that S4 binds tightly to the five-way junction formed by h3, h4, h16, h17 and h18, and that truncations in the first three helices give the largest variations (see Fig. 5.1).

The critical role of S4 in the early assembly of the small subunit and RAM mutations on the five way junction suggests that the interactions between the signatures in this region are functionally important for ribosomal assembly and fidelity of protein synthesis in bacteria, but further experiments and simulations are required to characterize these interactions. This chapter presents a detailed study of the dynamics of the ribosomal signatures h16 and S4N using MD simulations. Circular dichroism (CD) and fluorescence spectroscopy experiments were performed in conjunction with the computational work, and qualitative agreement between computational and experimental results give insight into the intrinsic disorder of the signature on S4 and the flexible nature of the interactions between the RNA:protein signatures. Our results are consistent with a fly-casting mechanism in which folding of S4N is induced by binding to h16 and suggest that this signature region on the ribosome was a domain specific invention in
evolution aimed at speeding up the molecular recognition between the rRNA and the early binding ribosomal protein S4.

**Materials and Methods**

**Protein and RNA Design and Sample Preparation**

Based on NMR (14), X-ray (15) and a bioinformatics study (6) of S4, the flexible N-terminus extends from residue 1 to approximately 45 in *Escherichia coli*. However, in order to study the interaction with h16, the S4N fragment in this study has been chosen such that the last residue is within 5 Å of the h16 in the crystal structure. The fragment of the wild type *E. coli* ribosomal protein S4 from positions 1 to 39 (PDB ID 2I2P (16)) was ordered from Genscript (Piscataway, New Jersey). In order to have a fluorescence probe within the protein for bulk measurements, the isoleucine at position 33 was replaced by tryptophan. The cysteine at position 31 was replaced by serine while the alanine at position 1 was replaced by cysteine. These two cysteine mutations were introduced to facilitate the labeling of the N-terminus with an Alexa-488 fluorophore for future single molecule experiments. The mutated residues were carefully chosen according to the sequence alignment of S4 provided in (6), ensuring minimum perturbation from native behavior. The N-terminus of the fragment was acetylated and the C-terminus was amidated.

The h16 RNA fragment from position 406 to 436 (PDB ID 2I2P (16)) was ordered from Integrated DNA Technologies (Coralville, Iowa) with two modifications: the adenine at position 435 was changed to 2-aminopurine (2AP), a fluorescent analog of adenine shown to not perturb RNA folding (17), and a nonfluorescent quencher Iowa Black FQ was attached to the 5’ end of the molecule. The original and modified sequences are shown in Fig. 5.1.

A standard PE buffer, 10 mM sodium phosphate and 0.1 mM ethylenediaminetetraacetic acid disodium salt (EDTA-Na₂) with pH 7.1, was used for all experiments performed.

**Experimental Measurements**

**Thermal absorbance melts**
Thermal absorbance melts of the 1μM h16 solution was performed in a 1 cm path length cell using spectropolarimeter (Jasco Inc., Easton, MD). Absorbance intensities were measured at 260 nm, and the resulting curve was fitted to a two-state thermodynamic model. Assuming that the free energy difference (ΔG) between the two states has a linear temperature dependence, the unfolding transition temperature can be estimated.

**Dilution measurements**

A dilution measurement was performed using the Cary Eclipse Fluorescence spectrophotometer (Varian Inc, Palo Alto, CA). A fluorescence spectrum was measured with 600 μL of 7.5 μM sample of 1:1 S4:h16. Thereafter, 100 μL of sample solution was removed from the cuvette and replaced with 100 μL of water, stirred and spectrum measured again. This was repeated until a concentration of 0.34 μM was achieved. The spectra were baseline subtracted. The integrated spectrum was used for calculating the binding constant. The dilution data was fitted to the following thermodynamic model to obtain the binding constant. For a two component system described by

\[ A + B \rightleftharpoons AB \]  

the binding constant K at equilibrium can be written as:

\[
K = \frac{[AB]}{[A][B]} = \frac{x}{(x_0 - x)^2} 
\]  

(2)

where \( x_0 \) is the starting concentration of each component A and B; x is the equilibrium concentration for the bound complex AB. Thus,

\[
x = \frac{(2x_0K + 1) \pm \sqrt{(2x_0K + 1)^2 - 4K^2x_0^2}}{2K} 
\]  

(3)

The overall fluorescence intensity (Flu) in solution can be expressed as follows:

\[
\frac{Flu}{x_0} = (I_A^B + I_B^A) \frac{(x_0 - x)}{x_0} + I_{AB}^A \frac{x}{x_0} 
\]  

(4)
where $I_A^β$, $I_B^β$, and $I_{AB}^β$ are the fluorescence intensities per unit concentration for A, B, and AB respectively. The curve in Figure 5.6A is fit using Eq. (3) and (4) to obtain the binding constant $K$. The dissociation constant is $K^{-1}$. The fluorescence contribution from the RNA is insignificant and hence is assumed to be zero.

**Circular dichroism** (CD) spectra were measured for 20 μM solution of the protein and RNA separately as well as a 1:1 mixture of the two molecules in a 1 mm path length cuvette using a spectropolarimeter (Jasco Inc., Easton, MD). The CD spectrum was scanned from 190-300 nm.

**Molecular Modelling and Simulation**

Coordinates for S4 and the five-way junction in 16S rRNA were taken from the crystal structure of E. coli ribosomal SSU at 3.22 Å resolution (PDB ID 2I2P (16)). The Ile33Trp, Cys31Ser, and Ala1Cys mutations on the protein and A435 (2AP) mutation on the RNA sequence were made according to experimental setup (see section above and Fig. 5.1). Parameters for 2-aminopurine were developed by analogy with the separate adenine and lysine parameters already present in the CHARMM27 force field. The five-way junction system, including h3, h4, h16, h18, a truncated h17, a five-membered loop CUCAA that caps h4 and a seven membered loop UUUUGCU that caps the truncated h17, was assembled according to the minimal S4 binding model suggested in Woodson’s lab (13). The two additional loops were taken from the *E. coli* SSU (PDB ID 2I2P, residue 618 – 622) and glutamine tRNA (PDB ID 2RD2, residue 932 – 938) with two mutations (C934U and A937C), respectively.

All systems studied in this paper were neutralized with Na⁺ or Cl⁻, and prepared using the protein: RNA protocol in (18) and software VMD (19). 10 mM NaCl was added to the final solution according to experimental conditions, and equilibration simulations were run using NAMD2 (20) with periodic boundary conditions and the NPT ensemble with pressure set to 1 atmosphere and temperature set to 298 K. Electrostatics were calculated with the particle mesh ewald method (21). The van der Waals interactions were calculated using a switching distance of 12 Å and a cutoff of 14 Å. Variable time steps for updates of bonded, van der Waals, and electrostatic calculations were 1 fs, 2 fs, and 4 fs, respectively.
Ten, 50-ns MD simulations were run for each of S4N, h16, and the S4N:h16 complex, and they are referred to as unbound S4N runs, unbound h16 runs, and complex runs, respectively. Four additional runs were performed to establish reference points: two 50-ns runs for the full-length, unbound S4 (~205 residues); one 50-ns run for the h16-containing five-way junction in the 5’ domain of 16S rRNA; and one 50-ns run for the complex of the five-way junction and full-length S4.

For comparison to the experimental melting data, unfolding of the S4N was also simulated. After 20 ns of equilibration, the temperature was raised linearly from 298K to 358K through one hundred 5-ps steps. Production runs at 358K were carried out for another 99.5 ns to achieve a total of 100 ns of unfolding simulations.

**Base Pairing and Base Stacking Interaction Determination**

Base pairing interactions were determined using 3DNA (22), which reports both canonical and non-canonical base pairs. Hydrogen bond patterns were also recorded for each base pair, including those between atoms on one base with atoms from the other sugar ring or backbone. As the base stacking information given in 3DNA is limited to the overlapping areas of successive base pairs, a base stacking detection program was developed to include the occurrence of stacking interactions between bases not involved in base pairing. The criteria were based on the geometric measures established in (23), with slightly relaxed cutoff values obtained from (24) to incorporate the intrinsic RNA structure fluctuations. The three criteria used were: 1) distance between the geometric centers of the two base rings should be smaller than 5.5 Å, 2) the angle between the base normal vectors of the two base rings should be smaller than 30°, and 3) the angle between one of the two base normal vectors and the vector connecting the two ring geometric centers should be less than 40°. Both rings in the purine bases were calculated individually, and if one of the rings met all the above criteria, the base was considered stacked. For more detail, please refer to (23).

To validate the algorithm, pairs of bases were placed in various conformations, and each conformation was determined to be stacked or not according to the algorithm. The non-bonded energies were then calculated in each conformation to show that the identified stacking interactions were energetically favored. The algorithm was further tested using three tRNA
simulations with high, medium and low magnesium concentrations from a previous study on EF-Tu:tRNA which demonstrated dramatic changes in both base pairing and stacking with ion concentrations. Trends in the amount of base stacking in the anticodon arm were in agreement with those obtained through both visual inspection and 3DNA calculations.

**Results and Discussion**

**Characterization of protein:RNA interactions in the crystal structure**

The interactions between S4N and h16 in *E. coli* (PDB ID 2I2P(16)) will serve as a reference point for the simulations and experiments on the dynamics of this signature complex. The 39 residue S4N consists of a short α-helix (Lys7–Glu14) followed by an unstructured loop with two coil turns. In the structure of *T. thermophilus* (PDB ID 2J00 (25)) and sequences of some other bacteria, this region of S4 contains a zinc-finger motif in which a zinc ion stabilizes the short helix and the two coils. While this motif is partially missing in *E. coli*, the eight-residue helix is stable without the presence of metal ions, but the two coils become quite mobile in the MD simulations. Similar to the other r-proteins, S4 is highly charged. The *E. coli* version of the S4N has nine positively-charged amino-acids (arginine/lysine) and four negatively-charged ones (aspartate/glutamate), comprising one-third of the total number of residues in this segment. In the crystal structure, these charged residues sit on opposite sides of the S4N — with the positively (negatively) charged ones oriented toward (away from) the S4N:h16 contact interface. In the crystal structure, no saltbridges are present within the S4N, but a salt bridge does exist between Glu14 and Arg55 on the first helix of the C-terminal domain of S4.

In the crystal structure, the 31-nucleotide RNA helix h16 (nucleotide 406 to 436, *E. coli* numbering) contains thirteen native base pairs (with a total of 29 hydrogen bonds) and 22 native base stacking pairs (See Methods). The complexity of its secondary structure is seen in the diversity of the base pairs, including five Watson-Crick base pairs (three GC bp, two AU bp), three GU wobble pairs, one sheared GA pair, and one AU Hoogsteen pair. Two stems are established through these base pairs and stabilized by the base stacking interactions. The two stems of h16 are connected by an internal loop which itself is stabilized by interstrand stacking of three adjacent base pairs, G410·A432, U429·A431 and A430·A411 (26). Other intrastrand base-stacking interactions, such as G428/A430, help keep the backbone continuity despite the
large curvature created by the intrastrand base pair U429·A431 and the nearby interstand stacking.

Positions of the protein:RNA contacts in the crystal structure are presented in Fig. 5.1 and the contact map and 3D structure in Fig. 5.2. h16 is located on the periphery of the SSU, and it has very limited interaction with the other parts of the RNA except the stem on h18 and the linkage into h17. However, it interacts with the protein S4 extensively. The backbone of h16 fits into the grooves defined on the protein:RNA interaction surface of S4N outlined by positively charged amino acids (shown in blue). The two main binding sites for S4N are on either side of h16 from residue U407 to G413 and from G425 to A430 with slightly more interactions existing with the 30 side of the internal loop. Phosphate atoms in these two binding sites make hydrogen bonds with either the sidechain amines of lysine/arginine or with the backbone amines from the nearby amino acids. G413, one of the two bases in h16 that has neither base-pairing nor base-stacking interaction within the RNA molecule, has the interaction between the O6 atom on the base with the backbone nitrogen atoms of Lys30 and Ser31. From the S4 side, binding to the RNA is driven mainly by non-specific electrostatic interactions. Each positively charged residue may interact with phosphate atoms from several nucleotides, while all four negatively charged residues (shown in red) are positioned away from the RNA. Contacts between the RNA and protein residues 13 to 19 are prevented by the rigidity of the helix, which is oriented at an angle of ~ 46° with respect to the helical axis of h16.

**The N-terminus of S4 is dynamically disordered**

The MD simulations reveal distinct behaviors in the two domains of unbound S4: the stability of its C-terminal domain and the disorder of its N-terminal fragment. The RMSD per residue shows large fluctuations in the S4N region which are compared in (Fig. 5.4) when it is bound to h16.

This flexibility of S4N seen in the unbound system and in the complex, makes the structural alignment among different conformations along the trajectories difficult to interpret when the alignment is based on the backbone of the full sequence. In order to maintain an understandable basis for comparison, the alignment is based on the relatively stable eight-residue α-helix (Fig. 5.4 A). This procedure results in a backbone RMSD of less than 1.5 Å for the α-
helix in all the simulated trajectories, and it is then clear that the overall RMSD value, though exceptionally large, is a measure of how much the coils and loops have moved away from the original positions. Another measure of the disorder is reported by examining the fraction of native contacts, $Q$, which does not depend on structure alignment.

Time averaged RMSD and $Q$ were plotted for the last 45 ns of each trajectory (Fig. 5.4 B). Unbound S4N is much more flexible than S4N bound to h16. All simulations of the S4N:h16 complex (blue squares) have S4N RMSD values smaller than 8 Å while all simulations of the unbound protein (red dots) have RMSD larger than 8 Å. Similar results are observed for the mean $Q$ values with the complex formed from either the full system or just the signature regions all having values greater than 0.4. From calculations carried out for protein folding and structural phylogenetic studies, values of $Q$ below 0.35 represent unfolded states or structurally unrelated proteins (27). Furthermore, the disordered structure of unbound S4N is consistent with an NMR study where the S4 N-terminal fragment was not resolved together with the globular C-terminal domain (14).

The backbone motions arise mostly from the coil and loop region between residues 23 and 30, as seen in the RMSD per residue plots in Fig. 5.4 A as well as the inset representative conformations colored by RMSD per residue in Fig. 5.4 B. When S4N is bound to h16, positively charged residues Lys21, Arg25, Lys30 and Lys32 interact with the RNA backbone strongly so that the coils vary little from their binding positions. Though the range of movements is limited, the sidechains of these lysines and arginines can interact with the phosphates of nucleotides neighboring the contact sites in the crystal structure as discussed below in “RNA:protein interactions through simulations.”

**Temperature unfolding studies**

During the temperature unfolding simulation (see Methods), the stability of the $\alpha$-helix seen in both the unbound and complex simulations of S4N was maintained The helix persisted across the 100-ns simulation with two smaller helices of length 4-5 transiently appearing over 40 ns after the temperature jump at 20 ns (Fig. 5.4 C). CD measurements taken during temperature melt experiments on unbound S4N also showed increasing $\alpha$-helical content. The CD signal at 222 nm, the wavelength characteristic of $\alpha$-helices, decreased linearly with increasing temperatures (Fig. 5.4 C). Furthermore, the CD temperature melt also supports the disordered
structure of S4N at room temperature, since no cooperative structural transition of S4N was observed. A fluorescence melt of S4N performed by exciting the tryptophan in the peptide showed no sigmoidal transition, another indication of the natively unfolded nature of S4N (Fig. 5.3).

**Stability of h16**

**Backbone movements**

In general, h16 shows greater backbone stability than S4N at room temperature, and although unbound h16 might have more conformational flexibility than the bound h16, the difference between the two is much more subtle than with S4. The thermo-stability of h16 was measured by a temperature melt experiment, in which the absorbance intensity at 260 nm of all the four RNA bases was recorded at increasing temperatures so that a global transition of the RNA molecule can be monitored. The result shows that h16 is stable at room temperature, and the gradual unfolding transition occurs at a melting midpoint of 58 ºC (Fig. 5.5 B). The unbound h16 runs have larger RMSD values (~3.87 Å, ensemble average) than the complex runs (~3.34 Å), however, the shift between unbound and bound forms of h16 is much smaller than for S4N. In Fig. 5.5 A, data points for the unbound and bound forms of h16 cannot be readily separated as in the same plot for S4N (Fig. 5.4 B), and average RMSD values for the most stable replicas in both unbound and complex runs are all around 3 Å . The slight increase in conformational flexibility for the unbound h16 is more clearly seen in individual plots.

In all twenty MD runs involving h16, time-averaged Q values are above 0.45, while only two complex runs of S4N have a Q value larger than 0.45. Q values for unbound and bound h16, which span from 0.45 to 0.55, are more inter-mixed than those for unbound and bound S4N, which span across a difference of 0.2. Only when included in the 5-way junction together with the full-length S4 bound is h16 significantly stabilized, with an average RMSD near 2.3 Å and a Q value over 0.6.

**Experimental and computational measure of flexibility in protein:RNA interactions**

How strongly the two molecules bind can be determined by measuring the fluorescence of the S4N:h16 complex during progressive dilution Fig. 5.6 A. Tryptophan fluorescence from
S4N is quenched in the complex due to the proximity of Iowa black which is attached to the 5’ end of h16, so higher fluorescence intensity indicates a larger fraction of unbound S4N in solution. The general trend of the increasing normalized fluorescence intensity clearly shows that as the solution is diluted, the binding equilibrium shifts to more unbound S4N and h16. The two-state thermodynamic model allows us to obtain a binding constant of 0.91 μM which is not surprising given the fact that only fragments of the rRNA and the protein S4 are used in this experiment.

CD measurements were taken to investigate any possible conformational changes for either S4N or h16 upon binding. At 20 μM concentration, the sum of the individual spectra of h16 and S4N were subtracted from that of the 1:1 S4N:h16 mixture to obtain the difference spectrum (Fig. 5.6 B red curve). This spectrum, together with the individual curves, indicates how the CD signal changes upon binding. At shorter wavelength (190 – 215 nm), CD signal changes are mainly due to conformational changes in the protein. Increase in this region is generally indicative of a decrease in random coil characteristics in S4N upon binding. However, the small decrease observed near 222 nm is inconclusive as to the change in helical content, because the amplitude of the signal is almost comparable to the noise. Most interesting is the drop in CD signal at longer wavelengths (260-280 nm). There are two possible explanations for this drop: either h16 changes conformation upon binding S4 or the difference is caused by transient interactions between h16 and S4N, as nothing is expected from the protein in this region.

To determine which scenario was occurring, CD spectra were calculated based on MD trajectory data. Snapshots were taken from every replica and coordinates therein were used to calculate the CD signals using the online software DichroCalc (Fig. 5.6 B insert, values have been scaled by 1200 times to enable direct comparison with experimental data). In the long wavelength region (260-280 nm) where unexpected decrease of CD signal was seen upon binding, the averaged calculated CD signals peaked at the same wavelength and spread out in the same order as in experimental measurements. Furthermore, the mean CD spectra calculated using h16 and S4N coordinates taken from both complex and unbound runs showed that neither h16 nor S4N has a conformational change that causes the CD signal to change at the peak wavelength ~268 nm. Therefore, the decrease of CD signal at the long wavelength range upon binding of h16 and S4N is a consequence of interactions between the two molecules.
In order to characterize the possible interactions that caused the CD signal to change around the 270 nm region, we looked into the trajectories and searched for specific interactions between protein and RNA molecules. Since the transition dipole moments of the bases are included for the RNA molecule in the calculation, interactions between protein and RNA bases are our main focus. As a result, three regions are identified to be rich in cation-π interactions during simulations. First, G413 and A412, the two bases which are neither base-paired or stacked to any of the other bases in the crystal structure. Second, U409 and G410, where the interacting protein residues can reach into the major groove of the RNA helix. Third, G406 and U407, which is the opening end of the molecule. All three places are on the 5’ strand of h16 and they form cation-π interactions with Lys32/Lys30/Arg25, Lys21/Lys7 and Lys21/Arg2 respectively. The first two regions show such interactions in every complex replicate, and the interactions are stabilized by the nearby salt bridges or hydrogen bonding to the functional group on the bases. On the contrary, cation-π interaction with G406 and U407 depends on the flexibility of the two bases and are mostly transient in the simulations. Once the end base pair opens, other interactions, such as the stacking between RNA base G406 and aromatic protein side chain from Phe19 or Tyr3, may also occur.

RNA:protein interactions through simulations

A detailed analysis of the salt bridges formed between protein residues arginine/lysine and phosphate oxygens of the RNA was performed to illustrate the variety of possible amino acid:nucleotide charge charge interactions in S4N:h16 complex. Among the fifty salt bridges identified in simulations, only three (Arg12-U429, Lys9-G428 and Lys32-U426) appeared in every replica run. Even for these three salt bridges, the largest occupancy is only ~68% and the longest average duration ~7.2 ns. Considering the other salt bridges present in the crystal structure, the average occupancy can be as low as ~7.5% and average duration shorter than 1 ns. Most of the salt bridges appearing during simulation had an occupancy under 20% and mean duration under 3 ns.

Four of the nine positively charged protein residues interact exclusively with the 3’ strand of h16, and two interact with the 5’ strand for the majority of time. Generally, the 3’ strand of h16 interacts much more strongly with the S4N than the 50 strand does. However, Arg25, Lys21
and Lys30 interact with both strands and especially the internal loop. The number of salt bridges formed for these three residues was large, but not one of them dominated in terms of occupancy or duration. These three residues either sit in or are close to the coil region on S4N where RMSD per residue is as high as the opening ends even in complex runs. These contacts explain why this particular region on the S4N is so dynamic.

The fly-casting mechanism

The “fly-casting” mechanism of molecular recognition was first illustrated by Shoemaker and colleagues by a simplified model of arc repressor dimer binding to the target DNA (28). They argued that target binding is sped up by the weak longer-range nonspecific interactions due to a larger capture radius of the partially unfolded protein monomer. In a subsequent study, it was shown that electrostatic forces are also coupled to the fly-casting mechanism by destabilizing the protein and lowering the folding barrier (29). In addition to the increase in capture radius, Huang et al. further attributed the speeding of the target recognition to the fewer encounter times required for disordered proteins before formation of the complex (30). All the above studies use coarse-grained model and the target molecules are kept fixed.

Our results suggest the possibility of a fly-casting mechanism. Guided by the electrostatic potential, the disordered N-terminal fragment on S4 searches for its binding site while changing its conformation constantly. The unstructured coil gives it a larger searching volume centered on the stable C-terminus, and the ever-changing conformation exposes different charged residues to the target RNA. As there are so many acceptable interactions between h16 and the disordered S4N, the initial contact can be established with very few tries. Although these interactions are very weak, with a binding constant 3 orders of magnitude larger than that of the binding of the full-length S4 to 16S rRNA, it can start the coupled folding and binding, and so the binding process completes faster on the whole.

It is!reasonable to suggest that ribosomal signatures occurred later during cellular evolution, promoting faster assembly of the ribosome using a “fly-casting” strategy with the help of cooperatively designed disordered r-proteins. Further validation is needed for the functional role of signatures other than h16 and S4N, since those proteins depend on the prior binding of primary binding proteins.
In conclusion, it is clear that instead of trying to search for some predefined rigid interaction pattern, the two molecules’ binding depends on an ensemble of hydrogen bonding and salt bridge networks. Exploring the flexibility of the interaction between the intrinsically disordered S4N and the stable h16, a fly-casting strategy is applied by the cell in a random search of partners inside the cell volume. The mechanism chosen by evolution ensures that r-protein S4 binds to the 16S rRNA quickly and nucleates the assembly of the SSU.
Fig. 5.1†: Three-dimensional visualization of the five-way junction and r-protein S4 on the ribosome. The first and second figures on top show the cartoon representation and the crystal structure of the five-way junction (dark blue) and r-protein S4 (yellow) on the ribosome, respectively. The third one shows the five-way junction and S4 as well as the signature regions S4N (red) and h16 (green) on the ribosomal small subunit (SSU). This figure has been rotated by 90° around the horizontal axis of the first two figures. Blow-up of the 3D structure for the system studied in this paper is showed below with the same coloring. In the secondary structure diagram that followed, bases colored red and yellow are in contacts with the S4N and S4C, respectively. S4N and h16 sequences in the wild type E. coli ribosome and the modified versions are listed at the bottom. Residues that have been mutated in this study are highlighted.

† Figure courtesy K. Chen, J. Eargle and Z. Luthey-Schulten.
Fig. 5.2†: Contacts between h16 in the rRNA and N-terminal fragment of ribosomal protein S4. A) Contact distances (Å, shortest distance between any pair of the heavy atoms for each pair of the residues): within the RNA molecule (top-left); within the protein molecule (bottom-right); and between protein and RNA molecules (bottom-left). Residue numbers for the arginines/lysines are colored blue and for the aspartic acids/glutamic acids colored red. Helix position in the protein is indicated by a purple bar beside the residue numbers. The scale bar applies to all three contact maps. B) Secondary structure diagram of h16 adapted from (31) (top) and generated by software jViz (32) (bottom). The bottom one was modified as follows: bonds in orange indicate canonical base pairing interactions, base pairs indicated by dots are noncanonical pairs that have two hydrogen bonds, base pairs indicated by open circles are noncanonical pairs that have only one hydrogen bond, and bonds in yellow indicate base stacking interactions. Nucleotides that are within 4.5 Å of S4 N-terminus are shown in red. C) Three dimensional visualization of the h16 backbone sitting on S4N interface (top) and S4N backbone sitting on h16 interface (bottom). Protein is colored by residue type, i.e. positively charged residues in blue, negatively charged residues in red.

† Figure courtesy K. Chen, J. Eargle and Z. Luthey-Schulten.
Fig. 5.3: Fluorescence melt of S4N.
A temperature dependence very similar to tryptophan itself in solution is seen for S4N indicating the absence of clear structural transition.
Fig. 5.4†: Backbone and secondary structure movement of S4 N-terminus.

A) RMSD Vs. Q (see Methods) for the backbone of S4 N-terminus averaged over time. Each marker represents one run, with red dots for unbound runs and blue squares for the complex runs and green triangles for the full-length S4 runs. Orange triangle stands for S4N in the full system, which is the 5-way junction with full-length S4. B) RMSD per residue from selected representative replicates. C) CD signal at 222 nm measured for S4N alone when temperature is increased gradually (left). The right plot shows how secondary structure changes over time for an unfolding simulation of S4N at 85 °C, with residues on the y-axis. Secondary structures are color-coded such that pink represents α-helix, green represents turn and blue represents 3_10-helix.

† Figure courtesy K. Chen, J. Eargle and Z. Luthey-Schulten.
Fig. 5.5†: Backbone and secondary structure fluctuations of h16 on the RNA. A) Time averaged backbone Q plotted against RMSD for all runs involving h16. Each marker represents one run, with red dots for unbound runs, blue squares for the complex runs, green triangles (4) for the five-way junctions and orange triangle (5) for the five-way junction with full-length S4. B) The absorption measured at 260 nm (absorption wavelength for 2-aminopurine) as temperature is gradually increased. C) Examples of fluctuations in the base pairing interactions, taken from both unbound h16 and complex runs. D) Each pair of base pairing interaction and base stacking interaction is colored by the fraction of time it stayed during the last 45 ns of all ten replicates. Interaction with black box in the plot indicates native interaction present in the crystal structure.

† Figure courtesy K. Chen, J. Eargle and Z. Luthey-Schulten.
Fig. 5.6: Binding of S4 N-terminus and h16 measured experimentally.

A) The dilution measurement showing increasing signal as the concentration of S4N:h16 is lowered. Insert plot shows the fluorescence quenching spectrum of S4N, h16 and 1:1 mixture of S4N:h16 at the highest concentration of 7.5 μM. The fluorescence was excited at 280 nm. B) Experimentally obtained CD signals for unbound h16 (cyan), unbound S4N (purple), and the S4N:h16 complex (green), as well as signals for the sum of unbound h16 and unbound S4N (blue) and difference signal obtained by subtract the sum from the S4N:h16 complex signals. Insert plot is the theoretically calculated CD signals in the region from 250 nm to 290 nm, color-coded accordingly.
References


Chapter 6

Design and fabrication of a high throughput fluorescence detector

Introduction

Single molecule spectroscopy (SMS) has become a very valuable technique for looking at heterogeneities within a given population (1). Bulk experiments show an ensemble average that may not be representative of any member of the ensemble. SMS allows us to look at the entire distribution, one molecule at a time and hence provides a more detailed description of the ensemble. The technique of SMS has in recent times been applied to the field of folding of biological macromolecules providing a molecule by molecule description of the process [(2-5)]. However, the current generation single molecule folding studies have certain limitations. One of the key limitations is the lack of throughput resulting in few observations. This greatly reduces the chance of observing states at higher energies in the energy landscape which are of great interest in the folding of biopolymers like DNA, RNA and protein. Furthermore, ultrafast time resolution is needed for experimentally determining the transition path time for the folding of biopolymers at the single molecule level. Until now single molecule experiments have only been able to put an upper bound to this quantity (6). In this chapter I describe the development of an instrument that intends to address these problems. The design goal of this instrument is to generate droplets containing a single molecule of interest, at a predetermined frequency which are optically trapped, probed by a pulsed femtosecond laser beam and the fluorescence photons collected in all $4\pi$ steradians using a time correlated single photon counting setup.

An overview of the instrument

The overall design can be decomposed into five separate components: the cube assembly, the generation of the femtosecond probe beam, the droplet generator, the optical trapping assembly and the detection system. Dilute solutions of sample are loaded into the droplet generator which generates 5-20 μm diameter droplets at regular intervals. The concentration can
be controlled so as to have only one molecule on average in each droplet. As of now, our detection limit has reached ~ 100 molecules of cerulean, a green fluorescent protein mutant, in every drop. The drops fall through the body diagonal of a fused silica cube made up of six lenses. As a drop reaches the center of the cube it is trapped by two counter propagating trapping diode lasers. They hold the drop long enough to be interrogated by a femtosecond probe beam which excites a suitable fluorophore in the biomolecule. Single photons from the fluorescence emission are then collected in all $4\pi$ steradians using a custom built time correlated single photon counting setup. The experiment is completely automated, starting from drop generation, trapping to data acquisition. In the following sections I will describe in detail all the individual components of the instrument. A schematic of the experiment is shown in Fig. 6.1.

**Cube Assembly**

The primary components of the cube assembly are the six photon collection tubes. Each tube is made up of three pieces. The front piece of a tube holds two of the focusing lenses. The first lens (in green) is a standard, off-the-shelf component (Lambda Research Optics, Inc., Costa Mesa, California) while the second one (in blue) is a custom-made aspheric lens whose parameters were optimized using Beam 4 (Stellar Software, Berkeley, CA) a commercially available optics simulation package. The two lenses together focus the photons coming from the droplet onto the PMT as shown in Fig. 6.1. These front pieces come together to form a “lens cube” (Fig. 6.3) in the center of which the droplet is levitated.

The second piece is capable of holding two filters, while the third piece has a slot for one (Fig. 6.5). The third piece also holds a baffle to limit the amount of stray light passing through and a PMT holder sits at its end. The first filter is tilted at an angle of 30 ° while the second and third filters are tilted at an angle of 20 ° with respect to an axis perpendicular to the cylindrical axis of the tube. This arrangement of the filters reflects the probe beam scattered by the droplet out through exit apertures in the tube. The three filters together help remove the laser scatter in stages. The scattered beam reflected from the filters ends up in a pyrex horn coated on the inside with colloidal graphite. The graphite along with the horn geometry absorbs the entire laser scatter. The insides of the tubes are also coated with the same colloidal graphite suspension in order to prevent stray reflections. The sensitivity of the setup requires very clean components.
All six tubes come together and are then held by scaffolding. Three aluminum legs hold the scaffolding down to the optical table. The PMTs sit at the end of the tube in a 3D adjustable mount designed specially for the purpose. The adjustability is required to position the PMT to the point where the fluorescence photons are focused down.

As shown in Fig. 6.4, the drops are introduced into the cube from the top (white arrow). The counter propagating IR trapping lasers come from the two corners of the cube while the probe beam is introduced through a small aperture along the bottom edge center.

**Droplet generator**

Droplet generators are useful in many different industrial applications that require the accurate delivery of a small amount of liquid or solution to a specific region (7). Inkjet printers are one of the most common appliances that use advanced droplet generation technology. Here we describe the construction of our own droplet generator for generating 5-20 μm diameter droplets at regular intervals.

The nozzles for the droplet generator were pulled using a custom built “pipette puller”. The fabricated pipette puller is shown in Fig. 6.7. It consists of a DC motor, a flexible coupling and a long axle supported on ball bearings. At the end of the axle an adapter is used to attach a borosilicate capillary along a ridge coaxial with the axle. The capillary is attached to a cylindrical weight at the end maintaining the axial symmetry. An oxy-acetylene flame is used to melt the capillary at the middle as it rotates so as to draw it out into a taper. The weight at the bottom rotates within a slightly oversized cylindrical well preventing stray off-axis motions and finally allowing a vertical drop after the capillary is melted. The angular velocity of the rotation and the intensity of the flame can be controlled to obtain nozzles of various shapes. The fused tapered end is then polished with abrasive paper of progressively lower mesh sizes to open up an orifice of predefined diameter. As a final polish a 5 μm silica powder is used. A low powered microscope is used to check the size of the orifice. Hydrogen fluoride etching of nozzles was also attempted to open up orifices of predefined diameter. The method is more time consuming and creates surface islands at the orifice that affect droplet stream stability and often results in multiple streams.

A piezo cylinder is glued on top of a fused borosilicate capillary using conductive glue to maintain the connection to the inner surface electrode for sending voltage pulses to obtain compression in the piezo. This piece of capillary is then coupled to the nozzle with a teflon
adapter. The two part assembly allows reuse of the part containing the piezo even if the nozzle itself breaks or gets irreversibly clogged. The capillary is supplied with solution from a syringe using teflon tubing. The droplet generator assembly is mounted on two piezo driven translation stages that are screwed together at 90°. This allows the droplet generator to be translated in two perpendicular directions which in turn makes it possible to compensate for any deviation in the droplet ejection direction. Finally the entire assembly is mounted onto a kinematic mount so as to have reproducibility in its positioning. The droplet generator assembly is shown in Fig. 6.9.

The characterization of the droplet generation and optical assessment of droplet trapping was performed by stroboscopic viewing (7). A CMOS camera was used to view the droplets. The 60 Hz frame rate trigger from the camera was used to synchronize the stroboscopic viewing. A Wavetech signal generator was triggered by the camera externally. The Wavetech in turn triggered the droplet generator and an LED with variable delay. The width of the LED pulse could also be controlled. Smaller pulses yielded sharper but dim images and vice versa. The overall triggering scheme is shown in Fig. 6.15. Since the camera provided a constant 60 Hz trigger a frequency divider box was built that allowed droplets to be run at 30, 20, 10, 4, 2, 1 Hz. The droplet video is directly transmitted from the camera to a TV. It is also sent to a video converter box for recording on a computer. A TV grab of a stroboscopically imaged droplet stream is shown in Fig. 6.10.

**Laser system**

The instrument is currently capable of exciting molecules in the ultra-violet and blue wavelengths. A Ti:Sapphire laser, pumped by a 5 W Millenia solid state laser at 532 nm, produces 700-800 mW of CW power in the wavelength range of 820-890 nm. On mode-locking, the Ti:Sapphire produces ~100 fs pulses at 92 MHz with an average output power of 400-550 mW. The center frequency of the emission wavelength under mode-locked condition can be tuned from 830-900 nm with a FWHM of 30-45 nm. This infrared beam is then fed into a custom built frequency tripler. An LBO crystal doubles the frequency of the incoming femtosecond beam and then recombines this doubled blue beam with the remaining infrared beam on a BBO crystal generating UV light. The overall scheme is shown in Fig. 6.11. Thus either the doubled (420-445 nm) or tripled (275-295 nm) light may be used for performing the experiment.
depending on the choice of fluorophore. The probe beam is spatially filtered through a 25 µm pinhole to obtain a TEM00 focal spot at its point of intersection with the droplet stream.

**Optical trapping assembly**

The technique of optical trapping was pioneered by Arthur Ashkin at the Bell laboratories almost 50 years ago. It was known for a long time that light exerts radiation pressure. However its successful use to trap light objects had to wait until 1970 (8). An object that has a larger difference in refractive index between itself and the surrounding medium is easier to trap. Three different geometries have been shown to exhibit optical trapping (9). The single beam gradient force optical trap is by far the most commonly used trapping geometry, as it provides a very stable trap. However because of the high numerical aperture requirement, the distance of the trapped particle from the microscope objective is limited to a sub millimeter range. Hence this technique could not be implemented in our experiment as our design criterion requires the focusing lens to be at least 12 mm away from the trap position. A levitation trap is another feasible geometry where the trapped particle is held up by just the forward scattering force balancing its weight. This method is more prone to stability problems and is also not geometrically feasible in our design of the experimental setup. Finally, a counter propagating geometry in which two beams coming from opposite directions are focused down with a short distance of separation of between the two foci also leads to trapping. This geometry is implemented in our experiment.

In order to test the feasibility of optically trapping water droplets with the counter propagating laser beam geometry we performed simulations based on Generalized Lorentz-Mie theory (GLMT) while considering Brownian motion of droplets in air. Evaporation of the droplet from laser beam irradiation and humidity of the surrounding air was also taken into account. The simulations were run with different input parameters including trapping beam power, ejection velocity of the droplet from the nozzle, distance of the nozzle from the trap position, relative humidity of the environment around the droplet, diameter of the ejected droplet and focal spot size of the trapping laser beam. A plot of the distance travelled by the droplet in the vertical direction against the time of travel, for different ejection positions is shown in Fig. 6.12. The distance travelled is measured from its ejection position. The results of the simulation show that it is possible to trap the water droplet at the desired position using realistic input parameters.
The optical trapping assembly consists of two counter propagating diode lasers that are focused close to the center of the cube. The separation of the foci is controlled to obtain maximum trapping efficiency. Optical trapping was first tested in a “mock chamber” shown in Fig. 6.13 before implementing it in the actual setup. Two collimated beams with identical beam profiles come in from the left and right arm and are focused in to a plexi-glass containment cube. This containment cube prevents air current disturbances from affecting the stability of the droplet. The overlap of the focal spot is controlled by the moving the translation stages onto which the focusing lenses are mounted. Ideally the beams from the two sides focus just before the center position. The droplet stream can be viewed either by stroboscopic illumination using the LED in which case the IR laser scatter is filtered from the CMOS camera or by direct recording of movies using the IR laser scatter as the droplet passes through the trapping region.

A frame by frame image of a droplet being trapped as it passes through the trapping beam is shown in Fig. 6.14. Each frame is 33.33 ms long. The droplet stays within $\pm 3 \mu m$ of the center position for longer than 100 ms while it evaporates. A overlaid image of the time progression of the droplet scatter is shown in Fig. 6.14.

The trapping lasers are modulated as shown in Fig. 6.15 in order to synchronously turn them on and off so as to allow drops to be retarded for observation and removed in an automated fashion. For this purpose a modulating voltage was generated by programming the analog outputs of the National Instruments board 6229. A software delay allows for synchronization with the droplet stream.

**Detection system**

The detection system consists of 6 photomultipliers. The probe beam is polarized in a direction parallel to four PMTs in plane as shown in Fig. 6.16 while it is perpendicular to the two PMTs on the top and bottom. The PMTs on the top and bottom will be referred to as out-of-plane PMTs while the ones on the side will be referred to as in-plane PMTs. All the PMTs were pre-calibrated for sensitivity using a constant power white light source. The calibration was performed by adjusting the supplied voltage to the individual PMTs so that all of them detect on average the same photon flux from the constant white light source. Six of the PMT channels are fed to a home built data acquisition box that is synchronized to a mode locked Ti:Sapphire laser, serving as a clock for the entire experiment.
The overall sequence of the electronic data acquisition process is shown in Fig. 6.17. Each single photon counting PMT sends its signal first to an amplifier. The amplifier unit on detecting the photon sends out a square wave pulse to a voter board. The voter board receives input from six of the PMT amplifiers. As a photon is detected, the voter board simultaneously sends out a 6 bit binary number and two square wave pulses. One of the square wave pulses triggers a time-to-amplitude converter while the other one triggers a 16 bit binary counter. The 6 bit binary number sent out from the voter board specifies which PMT the photon was detected by. Each of the 6 bits represents a PMT. For a single detection event all of them are set to zero expect for the one in which the photon was detected. The example in Fig. 6.17 shows that PMT 2 detected a photon. Thus these 6 bits provide us with polarization information for the emitted photon. The pulse sent out to the time-to-amplitude converter from the voter board starts a capacitor ramp which is terminated by the next laser pulse detected by the photodiode. The capacitor voltage is read at the point of discharge and is converted to a 10-bit binary number. This number is proportional to the delay time between the laser pulse and the detected photon and will be referred to as the “micro time”. Meanwhile, the 92 MHz laser pulses detected by the photodiode trigger a 16 bit counter. The counter sends out a 16 bit number as soon it receives the second square pulse from the voter board. This provides the time of emission of this photon in laser pulse units since the beginning of data acquisition. This time will be referred to as the “macro time”. Thus in total 32 bits of binary data are generated for each photon detected by the PMTs. A sample 32 bit data and its breakdown into the different components is shown in Table 6.1. In case of simultaneous photon detection by multiple PMTs within two consecutive laser pulses the delay time of only the first photon is recorded however both the PMTs are registered in the 6 bit PMT information. These are very few in number and are not used in the final analysis of the data. The electronics currently has a 100 ns dead time preventing the detection of photons that are separated by less than 100 ns.

**Calibration of time to amplitude converter**

The time-to-amplitude converter (TAC) does not have a perfect linear response. However the response itself is reproducible. Thus a calibration of this response is required to obtain accurate micro time information from the instrument. This is implemented by exposing the detection system to an incoherent white light source and then calibrating the distribution to a
uniform one. This process generates a look up table which is used for calibrating the raw micro
time output. A temporal distribution profile, before (red) and after (blue) calibration, obtained
from an incoherent white light source is shown in Fig. 6.18.

The slight differences in cable lengths for each PMT slightly offsets their time zero
positions with respect to each other. This is corrected in the data analysis using temporal offsets
for each PMT so that the scatter spike from the droplet is perfectly overlaid. Any asymmetry in
the positioning of the droplet within the probe beam is corrected by looking at the intensity of the
scatter signal from the 6 PMTs in real time. Ideally based on the symmetry of the overall
detection system the four in-plane PMTs are expected to have similar intensity levels between
themselves whereas the two out-of-plane ones are expected to have much lower intensity levels
in comparison to the in-plane PMTs. In between themselves the out-of-plane PMTs are expected
to show similar intensity levels.

Data Acquisition and Analysis

The entire experiment is run in an automated manner through a program written in
LabWindows (National Instruments). The experiment itself is capable of acquiring data at an
average photon flux of a 1,000,000 photons/sec. At this rate with 32 bits required for each
photon, the average data rate is 4 MB/s. Thus both the hardware and the software needed to be
capable of acquiring data at this rate. We assembled a computer that would be robust enough to
handle such needs. In total 4 TB of hard disk space was made available to acquire the data.
Gigabit ethernet adapters connected to Gigabit data transfer cables allowed direct transfer of the
data to a UNIX computer cluster for analysis.

A description of the data acquisition and analysis program along with the hardware
components involved follows. A droplet generator module helps assign the waveform parameters
to the voltage pulse sent out to the droplet generator driver for compression and relaxation of the
piezo, resulting in expulsion of the droplet. A module for trapping laser modulation allows the
user to assign the parameters for the modulating waveform sent out to the trapping laser. Two
National Instrument cards are used in triggering data acquisition, modulating the trapping lasers
and acquiring the data from the DAB described earlier. The NI PCI-6229 Multifunction Data
Acquisition device generates analog outputs to modulate the trapping lasers and to trigger the
droplet generator. The NI PCI-6534 High-Speed Digital I/O device obtains the raw 32 bit binary
data from the DAB. The NI cards are synchronized via a RTSI (Real-Time System Integration) bus cable. A “reset” signal is used to trigger data download and storage. In its high state the DAB stops functioning momentarily allowing data reading and storage. The automatically triggered reset is set to go “high” in between drops to prevent any loss of data from the droplets. Multiple threading is used to acquire and protect the integrity of the data. A data reading thread flushes all the raw data into a “thread safety queue”. This protects the raw data while being shared by multiple threads. Furthermore, multiple threading allows the user to operate the programs front panel for other tasks increasing efficiency of the overall process. A pattern I/O method is used for data acquisition with the NI PCI-6534 card.

The data is analyzed to look at the lifetime of the fluorophore obtained from all six PMTs independently and in average by creating a histogram of the micro time stamps. An overall intensity trace against time is also obtained using the macro time information and binning the photons in time bins. The 16 bit macro time information allows us to track time from 0 to ~700 µs after which it resets to zero. However the minimal photon flux obtained in our instrument allows us to track the change and correct for it to get the full length time trace. A concatenation analysis allows us to extract and combine only those portions of the data where the droplet is present. This is then be used for any further analysis.

Sensitivity

A 92 MHz, 13 mW probe beam at 425 nm generates a flux of 80,000 photon/sec when passing through the “cube” in the absence of droplets. This signal drops down to 9,500 photons/sec on purging with argon or helium gas as shown in Fig. 6.19. The difference in the extent of attenuation is not significantly different between the two gases. The reduction in the background signal shows that it is crucial to keep the Raman scatter from reaching the PMT. When a droplet is present at the center of the cube, the Raman scatter primarily comes from the aqueous solution. A significant portion of this scatter signal was removed by careful design of the optical filter sets. However, the Raman signal from air directly falls within the transmission window for the optical filters and can only be removed by purging the chamber with argon. On the other hand the Raman signal can be used as a calibration for the size of the droplet and removed from the analysis by time gating at the shortest times. This comes with the drawback that the photons with the shortest time delay are lost.
Choice of chromophore

The experimental setup has been tested with two sets of chromophores, each set requiring a different probe wavelength and a different set of rejection filters. Tryptophan, the intrinsic fluorophore in proteins was chosen for the first test. It was excited with a 280 nm, 5 mW probe beam at the 92 MHz repetition rate. The scatter from the drop itself was too high, making detection impossible below micro-molar concentrations at the currently generated drop sizes of 12-18 μm. Furthermore the high scatter caused the first optical filter to fluoresce. This fluorescence being red shifted was easily transmitted through the remaining filters causing a large background. The use of the high energy UV excitation makes the problem of intrinsic fluorescence from the filters even more formidable. Moreover tryptophan has very low excitation cross section (~6500 M⁻¹cm⁻¹) and quantum yield (~10%) compared to other fluorophores making this task more difficult.

As a second chromophore, a mutant of the green fluorescent protein (GFP) was tested next. This chromophore has an absorption cross section of 43,000 M⁻¹cm⁻¹ and fluorescence quantum yield of 62 %[]. At this point the apparatus reliably detects ~100 cerulean molecules per 12 ± 2 μm drops.

Results

A solution of Cerulean prepared in phosphate buffer in pH = 7.2. It was diluted to a concentration of 36 pM such that a 12 ± 2 μm drop would contain on average 20 molecules. A portion of the trajectory is shown in the Fig. 6.20. The intensity time trace on the bottom shows the drops passing through the probe beam at 20 ms intervals. The time axis for this plot is obtained from the 16 bit macro time information while the photons are binned in 1 ms time bins. On the top a portion of the time trace is used to show the delay time distribution along with polarization sensitive detection. An integrated fluorescence profile as a projection of the distribution is shown on the top right. The out-of-plane PMTs (green) provide a better signal-to-background (yellow) than the in-plane one.

Addressing current problems
In order to lower the detection limit of this experimental setup and make real use of the high throughput a number of improvements need to be made. Firstly, the scatter background needs to be lowered by reducing the size of the droplet. This was initially attempted by trying to make the nozzle orifice diameter smaller. However, smaller orifices in the range of 5-10 μm diameter cause severe clogging preventing droplet generation for long durations. Higher piezo voltages are required to generate drops with small orifice nozzles, which in turn induce forced vibrations in the droplet generator. This reduces the stability of the droplet trajectory. Smaller drops reach the terminal velocity at a shorter distance and hence the nozzle needs to come very close to the path of the probe beam. Secondary scatter results from the closeness of the tip surface that generates unnecessary background. Thus direct reduction in droplet size by reducing the orifice diameter of the nozzle was not feasible.

An alternative approach is being attempted. This involves holding the droplet stably in the optical trap for a longer duration such that it evaporates on its own and reduces in size while maintaining its position. This would result in increasing signal-to-background as the droplet size reduces. Furthermore, a mild heating laser may be used to evaporate the droplet faster to maintain the high throughput.

In another direction, a set of new brighter dyes, Alexa 488 and Alexa 594 and their corresponding filters are being implemented. They are almost twice as bright as Cerulean. A combination of our approach towards using smaller droplets and brighter dyes should provide us with the single molecule resolution in the near future.
### Table 6.1: The 32 bits of data received for each photon detected.

<table>
<thead>
<tr>
<th>Raw 32 bit binary data</th>
<th>0000 0011 1111 0001 01000 0010 0010 0101</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information content</td>
<td>High 16 bits, UV pulse tick count</td>
</tr>
<tr>
<td>Binary</td>
<td>0000 0011 1111 0001</td>
</tr>
<tr>
<td>Decimal</td>
<td>1009</td>
</tr>
</tbody>
</table>
Automated data collection with pulsed lasers, which can provide **Photon Delay** and **Photon Spacing** information for large numbers of single molecules, allowing a meaningful statistical analysis.

**Experimental Setup**

Fig. 6.1: An overview of the experiment.
Fig. 6.2: A scale drawing of the focusing optics, with beams originating from a drop on the left and ending at a PMT on the right. Only the optics and rays on one side of the principal axis are shown here along with distances in cm.

Fig. 6.3: On the left is the cross section of a single front piece. On the right is a cross section through four front pieces. The region shaded in green is the first lens and the one in blue is the second lens. A front piece would come from the front and back to complete the “lens cube”.

106
Fig. 6.4: A photograph showing the assembly of the front pieces. The two lower front pieces have been moved back slightly for ease of visualization. A razor blade placed at the center is used for beam profile measurements. The focusing lens for the trapping laser is also visible on the right.
Fig. 6.5: Schematic of the complete tube made up of the front piece, the middle piece and the end piece. The horns are shown in black. The beam is shown being initially reflected by the first filter that removes most of the scattered probe beam and lets the fluorescence photons through. Thereafter the second filter reflects almost all of the rest of the scattered probe beam, and the fluorescence photons are let through. A third filter may also be used if needed.
Fig. 6.6: An overview of the experimental setup showing the droplet generator, the trapping lasers and the “Cube”.
Fig. 6.7: The custom-built pipette puller showing the DC motor, the flexible coupling, the shaft and the capillary with the cylindrical weight inside the drop well. An oxy-acetylene flame is used is heating and pulling process as shown.
Fig. 6.8: Fabrication of droplet generator nozzles. (A) A side view of the front end of a pulled capillary showing the taper. (B) A top view of the front end of a nozzle after final polish. (C) An SEM image of the top of a nozzle with a 10 µm orifice diameter. (D) A zoom in of the same nozzle as in (C) showing the 10 µm orifice and the surrounding area in detail.
Fig. 6.9: The droplet generator. (A) A complete view of the droplet generator assembly showing the piezo driven translation stages, the adjustable shaft, the syringe, the capillary with the cylindrical piezo, the adapter and the nozzle. (B) A close of the nozzle showing the piezo cylinder, the adapter and o-ring in detail.
Fig. 6.10: (A) A schematic of the stroboscopic illumination setup. (B) A TV grab of a stroboscopic movie being recorded showing an ejected droplet.
Fig. 6.11: A schematic of the femtosecond laser system showing the generation of the UV and blue pulses.

Fig. 6.12: Simulation of droplet trapping after ejection from the nozzle. Each curve corresponds to the droplet being ejected at different distances from the nozzle orifice to the trapping center. In two of the extreme cases the trapping laser was turned off.
Fig. 6.13: The “mock” optical trapping setup. This setup has a geometry identical to that in the actual experiment and is used for testing optical trapping of droplets while imaging it with a CMOS video camera.

Fig. 6.14: Trapping of droplets. (A) Consecutive frames showing trapping of a droplet. (B) An overlay of the outer rings of the bright scatter spots from the frames in (A). The drop is forced back up by the trapping beam.
Fig. 6.15: The synchronization scheme. The droplet generator pulse and the trapping pulse are delayed with respect to each other. The trapping pulse has a higher voltage to start with to cause greater retardation initially.

Fig. 6.16: Polarization sensitivity of the “cube”. Four of the sides are parallel to the polarization of the incoming laser pulse while the top and bottom are perpendicular.
117

Fig. 6.17: A flowchart of the data generation process inside the data acquisition box.

Fig. 6.18: Calibration of the TAC. The uncalibrated response against the raw time delay (bottom axis) is shown in red. After calibration with the red data, a new data set from the same white light source is shown in blue against the calibrated time delay in ns (top axis).
Fig. 6.19: Raman scatter reduction. Change in photon flux detected by PMTs while the chamber is purged with Argon or Helium gas.
Fig. 6.20: Intensity vs time trajectory of droplets and the decay time distribution.
References


Chapter 7

Conclusion

Our fundamental understanding of the folding landscapes of biomolecules has come a long way since Anfinsen’s pioneering discovery. Experimentally, the use of various probes such as fluorescence, IR, CD, Raman, X-Ray and NMR has provided valuable complementary insight into the folding problem. These have been augmented from the theoretical and computational side by the proposition of the energy landscape theory for the folding of biomolecules and the use of better force fields with faster computers for running long MD simulations that are capable of reproducing experimental results. In chapters 2 and 3, we have shown how fluorescence can be used as an efficient probe for studying ultrafast RNA folding. Experiments along with MD simulations provide us with a better understanding of the folding process.

However we are not yet able to generally and accurately predict the three dimensional structures of biomolecules (1) or enumerate the various folding pathways possible based on the sequence and solvent conditions alone. We are surely headed in the right direction. In chapter 4, we have shown the implications of molecular dynamics time scales reaching those of experiments. In the days to come, computational and experimental “microscopes” will take the understanding of the protein and RNA folding problem to a level of unprecedented detail and accuracy. New experimental techniques such as femtosecond X-rays hold great promise (2). As computers and algorithms improve, explicit quantum corrections within classical simulations of large systems will become more feasible (3). With a better understanding of the folding energy landscapes of unimolecular systems we would then be able to apply those to many body problems. In chapter 5, we describe a computational and experimental effort to that end.

The advent of single molecule spectroscopy has brought in a new dimension to the understanding of the folding problem. The ability of experiments to look at the entire distribution instead of the average provides a lot more detail and clarity. Current generation single molecule experiments are limited by time resolution and throughput while attempting to address the folding problem. In chapter 6, we have described the development of an instrument that holds great promise in removing those shortcomings.
References


Appendix A

Supplementary Information for Chapter 2

Fig. A.1: The complete fits for the sequential model discussed in the main text and in figure 1. Two additional temperatures are shown here.
Fig. A.2: The complete fits for the trapped state model. The fit does not quite match the thermodynamic data in the 335 K to 350 K range, and also provides a less accurate kinetic fit. This model is more similar to the interpretation given by Ma et al., J. Am. Chem. Soc. (see references in main text). We cannot rule this model out entirely, but the basic result remains: that the hairpin landscape is rugged with at least four local minima detected by absorption, fluorescence or infrared spectroscopy.
Fig. A.3: The complete fits for the parallel state model.
Fig. A.4: The individual MD run pairs at each of the three temperature jumps simulated. The equilibrium run at the top is the same as in the main text.
Fig. A.5: (a) Fluorescence melt at an excitation of 308 nm. (b) CD melt at 280 nm. (c) Absorbance melt at 280 nm. The melting temperatures obtained by fitting to a two-state model vary by 5 °C, two standard deviations outside the combined measurement and fitting uncertainty. This could be an indication of multistate behavior in the thermodynamics, even though each individual probe fits to a two-state sigmoid.
Fig. A.6: Fluorescence decays of the 2AP labeled RNA at different temperatures. Note that the decay levels off to near zero amplitude close to 14 ns.
### Table. A.1(a): Free Energy Parameters for the trapped state model

<table>
<thead>
<tr>
<th>State</th>
<th>$T_0$ (K)</th>
<th>$\Delta G^{(0)}$ (kJ mol$^{-1}$)</th>
<th>$\Delta G^{(1)}$ (kJ mol$^{-1}$ K$^{-1}$)</th>
<th>$\Delta G^{(2)}$ (kJ mol$^{-1}$ K$^{-2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native</td>
<td>330.22</td>
<td>0.00</td>
<td>0.4089</td>
<td>0.0021</td>
</tr>
<tr>
<td>FrayEd</td>
<td>312.53</td>
<td>0.00</td>
<td>0.1818</td>
<td>-0.0001</td>
</tr>
<tr>
<td>Unfolded</td>
<td>328.00</td>
<td>0.00</td>
<td>0.0000</td>
<td>-0.0001</td>
</tr>
<tr>
<td>Trapped</td>
<td>348.47</td>
<td>0.00</td>
<td>0.0836</td>
<td>0.0384</td>
</tr>
</tbody>
</table>

### Table. A.1(b): Activation free energy parameters and Kramers prefactor for the trapped state model in Fig. A.2. Note that a large $\Delta G^{(2)}$ had to be fitted for the S state.

<table>
<thead>
<tr>
<th>Connecting States</th>
<th>$\Delta G^{(0)}$ (kJ mol$^{-1}$)</th>
<th>$\Delta G^{(1)}$ (kJ mol$^{-1}$ K$^{-1}$)</th>
<th>$k_m$ (µs$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native $\rightarrow$ FrayEd</td>
<td>6.5107</td>
<td>0.0411</td>
<td>1.2114</td>
</tr>
<tr>
<td>FrayEd $\rightarrow$ Unfolded</td>
<td>10.1897</td>
<td>0.0699</td>
<td></td>
</tr>
<tr>
<td>Unfolded $\rightarrow$ Trapped</td>
<td>20.5917</td>
<td>0.1371</td>
<td></td>
</tr>
</tbody>
</table>

### Table. A.2(a): Free Energy Parameters for the parallel state model

<table>
<thead>
<tr>
<th>State</th>
<th>$T_0$ (K)</th>
<th>$\Delta G^{(0)}$ (kJ mol$^{-1}$)</th>
<th>$\Delta G^{(1)}$ (kJ mol$^{-1}$ K$^{-1}$)</th>
<th>$\Delta G^{(2)}$ (kJ mol$^{-1}$ K$^{-2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native</td>
<td>328.00</td>
<td>0.00</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>FrayEd</td>
<td>344.74</td>
<td>0.00</td>
<td>-0.0743</td>
<td>0.0000</td>
</tr>
<tr>
<td>Unfolded</td>
<td>342.32</td>
<td>0.00</td>
<td>-0.2059</td>
<td>0.0000</td>
</tr>
<tr>
<td>UnStacked</td>
<td>335.58</td>
<td>0.00</td>
<td>-0.5339</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

### Table. A.2(b): Activation free energy parameters and Kramers prefactor for the parallel state model in Fig. A.3.

<table>
<thead>
<tr>
<th>Connecting States</th>
<th>$\Delta G^{(0)}$ (kJ mol$^{-1}$)</th>
<th>$\Delta G^{(1)}$ (kJ mol$^{-1}$ K$^{-1}$)</th>
<th>$k_m$ (µs$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native $\rightarrow$ FrayEd</td>
<td>9.198</td>
<td>-0.0784</td>
<td></td>
</tr>
<tr>
<td>Native $\rightarrow$ Unfolded</td>
<td>10.7952</td>
<td>0.3012</td>
<td>1.0000</td>
</tr>
<tr>
<td>FrayEd $\rightarrow$ UnStacked</td>
<td>16.6644</td>
<td>0.2722</td>
<td></td>
</tr>
<tr>
<td>Unfolded $\rightarrow$ UnStacked</td>
<td>15.9690</td>
<td>-0.4044</td>
<td></td>
</tr>
</tbody>
</table>
Appendix B

Supplementary Information for Chapter 3

Figure B.1: Thermodynamics and kinetics of the loop and stem mutants. (A) Relative fluorescence intensity with increasing temperature for the stem and loop mutants. The stem and loop controls and 2AP itself are also plotted on a log-normal scale. Normalization of the hairpin data by the short controls eliminates the intrinsic temperature dependence of 2AP in its local environment. (B and C) T-Jumps to 60 ºC and 80 ºC along with sequential model fits for the loop mutant. (D – F) T-Jumps to 39 ºC, 59 ºC and 79 ºC along with sequential model fits for the stem mutant.
Fig. B.2: Normalized absorbance melt measured at 260 nm of the stem and loop mutant.
Fig. B.3: Lifetime profiles of the stem (A) and loop (B) mutant at different temperatures.
Fig. B.4: The instantaneous response of the loop (A) and stem (B) controls.
Author’s Biography

Krishnarjun Sarkar obtained his B.Sc. degree in 2002 from St. Xavier’s College, Kolkata, majoring in Chemistry. He then obtained a M.Sc. degree in Chemistry in 2004 from the Indian Institute of Technology, Bombay, specializing in Physical Chemistry. In that same year, he enrolled for graduate studies at the Department of Chemistry in the University of Illinois at Urbana-Champaign, where he works under the guidance of Prof. Martin Gruebele. His graduate research focuses on studying the energy landscapes of RNA and protein folding, from bulk towards high throughput single molecule experiments.