UNIVERSITY OF CALIFORNIA, SAN DIEGO

Energy Landscapes of Biomolecular Function

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

in

Physics(Biophysics)

by

Paul Charles Whitford

Committee in charge:

Professor José N. Onuchic, Chair
Professor Patricia Jennings
Professor Herbert Levine
Professor Andrew McCammon
Professor Peter Wolynes

2009

The dissertation of Paul Charles Whitford is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2009

DEDICATION

To the carbon footprint this work has produced.

EPIGRAPH

Even the smartest person
can learn something
from the dumbest.
—Peking Noodle Company

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor José Onuchic for excellent professional and scientific guidance.

I would also like to thanks the co-authors of the papers in this thesis: John Saunders, Kevin Sanbonmatsu, Alex Schug, Shachi Gosavi, Scott Hennelly, Koby Levy, Osamu Miyashita and Jeff Noel. I would also like to thank the Onuchic and Wolynes groups for many stimulating discussions, especially Peter Wolynes for teaching me how to sharpen my arguments and be more critical of my own work.

Chapter 2, in full, appears in Journal of Molecular Biology, 2007, Whitford, Miyashita, Levy, Onuchic. The dissertation author is the primary investigator and author of the paper.

Chapter 3, in full, appears in Journal of Biological Chemistry, 2008, Whitford, Gosavi, Onuchic. The dissertation author is the primary investigator and author of the paper.

Chapter 4, in full, appears in Proteins: Structure, Funtion, Bioinformatics, 2009, Whitford, Noel, Gosavi, Schug, Sanbonmatsu, Onuchic. The dissertation author is the primary investigator and author of the paper.

Chapter 5 and Appendix A, in full, appear in Biophysical Journal, 2009, Whitford, Schug, Saunders, Hennelly, Onuchic, Sanbonmatsu. The dissertation author is the primary investigator and author of the paper.

VITA

2001-2003	Research Assistant in Physics with Professor George Phillies, Department of Physics, Worcester Polytechnic Institute
2001-2003	Undergraduate Teaching Assistant, Department of Mathematics, Worcester Polytechnic Institute
2003	B. S. in Physics and minor in Mathematics with *High Distinction*, Worcester Polytechnic Institute
2005	Master in Physics, University of California, San Diego
2007	Candidate of Philosophy in Physics, University of California, San Diego
2009	Doctorate of Philosophy in Physics(Biophysics), University of California, San Diego

PUBLICATIONS

Whitford PC, Schug A, Saunders J, Hennelly SP, Onuchic JN, Sanbonmatsu KY "Nonlocal helix formation is key to understanding S-adenosylmethionine riboswitch function. ", *Biophys. J.*, 96(L7-9), 2009.

Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, Onuchic JN, "An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields", *Proteins: Struct. Func. Bioinfo.*, 75(430-441), 2009.

Gosavi A, Whitford PC, Jenning PA, Onuchic JN, "Extracting function from a β-trefoil folding motif.", *Proc. Nat. Acad. Sci. USA*, 105(10384-10389), 2008.

Whitford PC, Onuchic JN, Wolynes PG, "Energy landscape along an enzymatic reaction trajectory: hinges or cracks?", *HFSP J.*, 2(61-64), 2008.

Whitford PC, Gosavi S, Onuchic JN, "Conformational transitions in adenylate kinase - Allosteric communication reduces misligation.", *J. Biol. Chem.*, 283(2042-2048), 2008.

Schug A, Whitford PC, Levy Y, Onuchic JN, "Mutations as trapdoors to two competing native conformations of the rop-dimer.", *Proc. Nat. Acad. Sci. USA*, 104(17674-17679), 2007.

Whitford PC, Miyashita O, Levy Y, Onuchic JN, "Conformational transitions of adenylate kinase: Switching by cracking.", *J. Mol. Biol.*, 366(1661-1671), 2007.

Mills JE, Whitford PC, Shaffer J, Onuchic JN, Adams JA, Jennings PA, "A novel disulfide bond in the SH2 domain of the C-terminal Src kinase controls catalytic activity", *J. Mol. Biol.*, 365(1460-1468), 2007.

Whitford PC, Phillies GDJ, "Enhanced septahedral ordering in cold Lennard-Jones fluids.", *Phys. Rev. E*, 72, 2005.

Whitford PC, Phillies GDJ, "Extended-range order, diverging static length scales, and local structure formation in cold Lennard-Jones fluids.", *J. Chem. Phys.*, 122, 2005.

Phillies GDJ, O'Connell R, Whitford P, Streletzky KA, "Mode structure of diffusive transport in hydroxypropylcellulose : water.", *J. Chem. Phys.*, 119(9903-9913), 2003.

ABSTRACT OF THE DISSERTATION

Energy Landscapes of Biomolecular Function

by

Paul Charles Whitford
Doctor of Philosophy in Physics(Biophysics)
University of California San Diego, 2009
Professor José N. Onuchic, Chair

In the context of the protein folding funnel and the energy landscape theory of protein folding, this work seeks to explain the origins of functionally related conformational transitions in proteins and RNA. Several avenues of investigation were employed. First, we extend the the well know C_α structure-based model for protein folding, such that the energy landscape has multiple competing basins. Through this extension of the model the energetics of conformational rearrangement in Adenylate Kinase were characterized. Specifically, we found several types of motion are required to explain its functionally related conformational rearrangements. The most exciting type of motion was partial unfolding, or cracking. The next approach used was a normal mode-based model where the conformational transitions are represented as motion along the lowest frequency normal modes. While the method itself was largely borrowed from earlier work, this application provided evidence of a catalytic cycle in AKE that can serve to reduce misligation. The third line of investigation explored folding of small proteins using an all-atom structure-based model. The aim of this study was to fully understand the limits of the model, such that functional transitions may now be studied with model related artifacts removed. The final line of investigation employed an all-atom model to study the folding and function of the SAM-1 riboswitch. We hypothesize that the rate limiting steps in riboswitch folding may be related to the decision point in riboswitch function. This work explicitly included the associated ligand and

identified a likely mechanism for riboswitch function.

Chapter 1

Introduction

1.1 Protein folding funnel

Through the application of energy landscape theory[1, 2] and the principle of minimal frustration,[3, 4] it has become clear that, for many proteins, the landscape associated with folding is smooth.[6, 7, 8, 9] That is, the energy gap between the unfolded and folded ensembles is much larger than the energetic traps encountered during folding. Non-native interactions are only transiently formed and native interactions dominate the folding process. This naturally led to the protein folding funnel paradigm[1] (Figure 1.1). At the top of the funnel is a large ensemble of rapidly interconverting unfolded structures. As energy decreases, the ensemble condenses to a smaller set of structures, which results in a loss of entropy. This decrease in energy is also strongly correlated with the formation of native structure. As the protein becomes more native-like, the imperfect cancelation of energy and entropy gives rise to free energy barriers.

There are many ways to schematically represent the protein folding funnel. While each representation has its own strength, here I focus on a discrete representation of the funnel (Figure 1.1). While protein folding is not composed of a small number of discrete states, this representation can still be illustrative. As described, this representation treats folding as the process of moving through a small number of levels of nativeness (vertical axis). As the protein becomes more folded, it moves vertically from one subset of states to a smaller subset of states

and eventually finds the native ensemble.

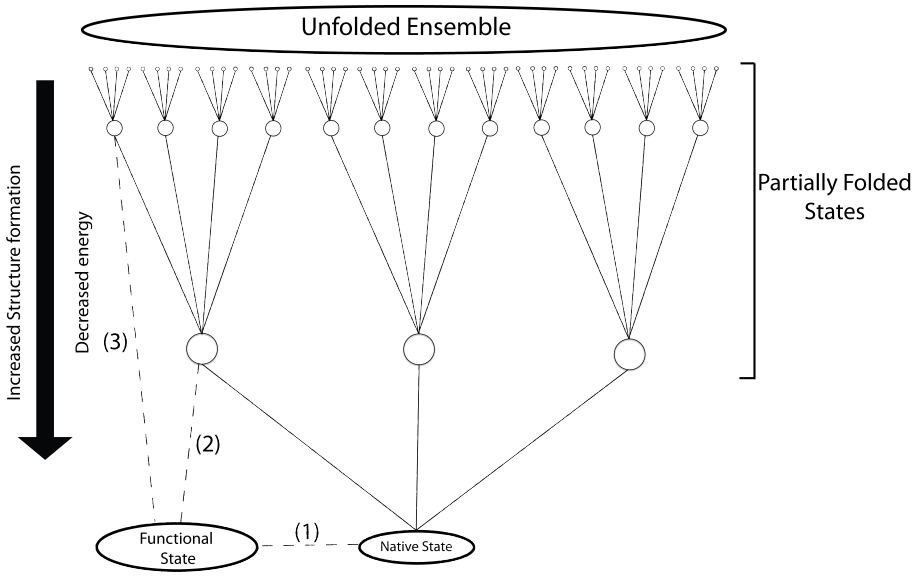

Figure 1.1: Discrete representation of the folding funnel. As the molecule becomes more native there is a decrease in energy and the number of available conformations is reduced. "functional" transition may be connected to the "native" ensemble in three ways: 1) simple hinge motions result in near-barrierless transitions, 2) partial unfolding, or cracking, occurs during the transition, or 3) complete, or nearly complete, denaturation is required to reach the functional state.

1.2 Including function in the funnel

While the ability to fold is a necessary property of any functional protein, it is clear that the structural dynamics must also be considered. Functionally related structural motions can range from simple side chain reorientation to large-scale domain rearrangements[10] and may even include order-disorder transitions[11, 12]. To account for these motions in the context of protein folding, we revisit the protein folding funnel. Schematically, each of these types of motion will modify the funnel differently. Let's consider a hypothetical "functional" conformation (Figure 1.1). From a static structure, it is not clear how this functional state will be energetically connected to the native basin. The most simple picture is that the functional state is connected directly to the native basin by simple, low-energy, "hinge" motions

(1 in 1.1). While this picture may be appealing, since it is consistent with how macroscopic machines function (i.e. levers and hinges), there is no reason why microscopic machines (where the structural integrity is maintained largely by many weak interactions) should be governed by the same principles. A more interesting mechanism for making a transition is that the protein partially unfolds,[13, 14] effectively moves up the funnel, and then refolds into the functional conformation (2 in 1.1). Along this same line of reasoning, there is no physical mandate saying that a native ensemble and functional ensemble be connected anywhere near the bottom of the funnel. In this case, there may be complete denaturation(3 in 1.1), and an orthogonal funnel is centered around the functional state. The aim of this thesis is to explore the possible ways in which biopolymer functional motions modify the folding funnel picture.

1.3 Progress towards function in the funnel

In folding funnels the energetic bias to fold is large relative to the energetic traps. In the most extreme case, the funnel is completely smooth, there are no energetic traps and only native interactions are stabilizing. To simulate an unfrustrated landscape one uses a structure-based Hamiltonian. With the completely smooth representaiton of the landscape one can simulate the entire folding process, with atomic detail, and determine the role of entropy during folding and function.

The present chapter provides a general outline of four projects and describes how each one has advanced our understanding of biomoleculr motions and how each one will lead to a more complete understanding of biopolymer function.

1.3.1 Adenylate Kinase: Multi-basin Molecular Dynamics model

Chapter 2 describes a study in which a C_α model was used to study the three-domain protein Adenylate Kinase (AKE).[15] AKE is know to undergo large structural rearrangements of two domains. The motion of these domains is strongly coupled to the rate of enzymatic catalysis. To study this system, we employed a

structure-based model and then added specific "frustration" to induce the conformational transitions. In this model, the functional motions in AKE are the result of competing stabilizing interactions. While this exact representations is obviously limited (namely the coarse-graining of residues), we were able to characterize differential dynamics of the two domains of AKE. In one domain, we found that it's conformational transitions likely occur easily, where thermal fluctuations are enough to induce domain rearrangement. In the other domain, we found strong competing energetic terms lead to the transitions. From these results we were able to make a variety of predictions, including which residues will shift the population from one ensemble to the other. Relating this to the folding funnel: this study showed that the motions in AKE are not simple hinge motions. Rather, partial unfolding, or cracking, occurs during the conformational transitions.

1.3.2 AKE: Normal modes and functional cycles

Chapter 3 describes an investigation where we approximate the energetics of each functional state of AKE by a network of harmonic interactions.[16] This description is very similar to a previous investigation that modeled the conformational transitions as "jumps" between 2 energy surfaces. The primary deviation in this study was the inclusion of 2 addition functional conformations. By describing the motion as jumping between these 4 energy surfaces, we were able to suggest a mechanism that can reconcile the high efficiency of AKE with its structural motions. These results led to the prediction of a catalytic cycle, where phosphoryl transfer and ligand binding redirect the fluctuations in the domains which lead to the jumps. The discussion of rates of domain motion, specifically the relationship to vibrational frequencies, are not intended to be conclusive. The primary advance of this work is that it suggests that the functional rearrangement observed in AKE should not be considered simple 2-state process. Additionally, this work suggests which specific sub-rearrangements will be more likely to include partial unfolding, or cracking.

1.3.3 Developing a structure-based all-atom model

Chapter 4 is focused on extending the well established C_α structure-based model to an all-atom representation.[17] Since protein folding is a well understood process, the aim of this study was to understand the robustness of the results obtained with this model. While the focus was on folding, there were observed processes that will likely resurface when using this representation for functional studies. Specifically, we showed that backbone collapse and sidechain ordering can occur separately. Even in the folded ensemble (i.e. ordered backbone), many ($\sim 40\%$) of the native atom-atom interactions are not formed. As described above, a key question in biomolecular function is to what extend does unfolding occur. Chapter 2 describes possible backbone cracking, though it may occur in a more subtle fashion where the backbone is relatively native, but the side chains undergo unfolding/refolding transitions.

1.3.4 SAM-1: Riboswitch

Chapter 5 discusses the application of the folding funnel to RNA.[18] Specifically, we studied the S-Adenosylmethionine-1 (SAM-1) riboswitch. Riboswitches are structured RNA fragments that can take different tertiary structures upon ligand binding. In the case of SAM, only the ligand-bound structure is known. The rate limiting step in folding provides a pause in structure formation. Since RNA folds as it is transcribed by the RNA polymerase, we hypothesized that this pause could be an ideal time for the riboswitch to "decide" which final structure it will take. In our simulations, we determined the rate limiting step and showed that the ligand decreases the associated free energy barrier. From a purely technical perspective, this work demonstrated the viability of a new method for studying ligand binding. Since many conformational transitions in biomolecules involve ligand binding, this methodology will be a strong tool for determine binding mechanisms and the ligand's influence on the functional motion. In the funnel framework, this study suggested that a near-native conformation may be the decision point for this system.

Bibliography

[1] Leopold PE, Montal M & Onuchic, J.N. *Proc. Nat. Acad. Sci. USA* **1992**, 18, 8721-8725.

[2] Onuchic JN & Wolynes PG *Curr. Opin. Struct. Biol.* **2004**, 14, 70-75.

[3] Bryngelson JD, Onuchic JN, Socci ND & Wolynes, P.G. *Proteins* **1995**, 21, 167-195.

[4] Bryngelson JD & Wolynes PG *Proc. Nat. Acad. Sci. USA.* **1987**, 84, 7524-7528.

[5] Nymeyer H, Garcia AE & Onuchic JN *Proc. Nat. Acad. Sci. USA.* **1998**, 95, 5921-5928.

[6] Chavez LL, Onuchic JN & Clementi C *J. Am. Chem. Soc.* **2004**, 126, 8426-8432.

[7] Gosavi S, Chavez LL, Jennings PA & Onuchic, J.N. *J. Mol. Biol.* **2006**, 357, 986-996.

[8] Clementi C, Jennings PA & Onuchic JN *J. Mol. Biol.* **2001**, 311, 879-890.

[9] Clementi C, Nymeyer H & Onuchic JN *J. Mol. Biol.* **2000**, 298, 937-953.

[10] Xu W, Harrison SC, Eck MJ. Three-dimensional structure of the tyrosine kinase c-Src. Nature 1997;385:595-602.

[11] Hyeon C & Onuchic JN. *Proc. Nat. Acad. Sci. USA* **2007**,104,2175-2180.

[12] Hyeon C & Onuchic JN. *Proc. Nat. Acad. Sci. USA* **2007**,104,17382-17387.

[13] Miyashita O, Onuchic JN & Wolynes, P.G. *Proc. Nat. Acad. Sci. USA* **2003**, 100, 12570-12575.

[14] Miyashita O, Wolynes PG & Onuchic, J.N. *J. Phys. Chem. B* **2005**, 109, 1959-1969.

[15] Whitford PC, Miyashita O, Levy Y & Onuchic JN. *J. Mol. Biol.* **2007**,366,1661-1671.

[16] Whitford PC, Gosavi S & Onuchic JN. *J. Biol. Chem.* **2008**,283,2042-2048.

[17] Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, & Onuchic JN. *Prot. Struct. Func. Bioinfo.* **2009** DOI: 10.1002/prot.22253.

[18] Whitford PC, Schug A, Saunders J, Hennelly SP, Onuchic JN & Sanbonmatsu KY *Biophys. J.* **2009**, 96, L7-9.

Chapter 2

Conformational transitions of Adenylate Kinase: switching by cracking

2.1 Abstract

Conformational heterogeneity in proteins is known to often be the key to their function. We present a coarse grained model to explore the interplay between protein structure, folding and function which is applicable to allosteric or non-allosteric proteins. We employ the model to study the detailed mechanism of the reversible conformational transition of Adenylate Kinase (AKE) between the open to the closed conformation, a reaction that is crucial to the protein's catalytic function. We directly observe high strain energy which appears to be correlated with localized unfolding during the functional transition. This work also demonstrates that competing native interactions from the open and closed form can account for the large conformational transitions in AKE. We further characterize the conformational transitions with a new measure Φ_{Func}, and demonstrate that local unfolding may be due, in part, to competing intra-protein interactions.

2.2 Introduction

Flexibility and conformational changes are well acknowledged to be indispensable properties of proteins. New experiments using ultrafast laser technology and detailed computer simulations have begun to reveal the motions of these proteins, which encompass a rich repertoire of movements on various length and time scales. These motions, which complement the static three-dimensional structures provided by X-ray crystallography and NMR measurements, are essential to understand protein functions[1].

Protein flexibility and plasticity allow proteins to bind ligands, form oligomers, aggregate, and perform mechanical work. Therefore, the ability to alter protein dynamics may enable quantitative control of protein functionality. While this form of functional control is very important to biology, it is not well understood from either a theoretical or experimental basis. Thus, the question arises: How can we quantitatively connect conformational dynamics with biomolecular recognition and function? To address this question, we propose a new structure-based model to study the dynamical properties of proteins, specifically, conformational rearrangement.

Large conformational changes in proteins are important in many cellular signaling pathways, which can be generally described by the following steps. First, a signaling protein becomes activated, which then activates, or deactivates, signal transducing proteins, such as kinases. Signal transducing proteins are mobile and communicate with receptor proteins, which then produce specific reactions. The activity of many signal transducing proteins is associated with large conformational changes. For example, C-terminal Src Kinase protein[2], the Cyclin Dependent Kinase family [3], the Protein Kinase C family [4] and Adenylate Kinase (AKE)[5] have stable inactive conformations, in addition to active forms. Since the balance between conformations regulates protein activity, conformational transitions play important roles in the machinery of the cell[6].

Functional conformational transitions require a biomolecule to have at least a pair of conformational states of nearly equal free energy. The energy landscapes of these proteins have several basins of attraction and the transitions between

basins dictates the conformational dynamics[7]. Despite the biological significance of these processes, the details of these processes are not fully understood. With a complete understanding of conformational changes we hope to predict which proteins have multiple conformations, predict these alternate conformations, determine the properties of the conformational transition ensemble, explain how proteins have evolved to have these properties and eventually design novel macromolecular machines which can execute any given biological function. To work towards these objectives, we explore the relationship between the structure, folding and function of AKE.

While many studies have investigated the relationship between protein structure and folding, fewer have focused on the relationship between structure and function, and even fewer have explored the interplay between protein structure, folding mechanism and function. Current experimental methods, including NMR, X-ray crystallography and fluorescence spectroscopy have been successful in describing the structural properties of individual states. These methods sometimes also manage to capture the chain flexibility[8, 9]. Nonetheless, experimental techniques have not been able to provide the molecular details necessary to fully understand the mechanism of conformational changes. Due to these limitations, there has been significant effort to develop a theoretical framework for describing functional transitions in proteins[10, 11, 12, 13, 14, 15]. With a developed framework, one may study the energetic barriers associated with conformational transitions, their coupling to folding/unfolding (cracking), the role of ligands, and the role of energetic heterogeneity and frustration in conformational transitions[10, 11]. In this work we propose a structure-based model that has a clear physical interpretation. Our model demonstrates that intraprotein contacts formed in the ligand bound structure of AKE can be responsible for the observed functional conformational changes. There has been success in applying simplified models to conformational changes, but our model provides a new physical interpretation that has not been proposed elsewhere.

The simplest model to describe functional transitions is based on landscape hopping and cracking between elastic networks[10, 11]. To lowest order approxima-

tion, all interactions about a minimum are harmonic. Thus, this approach uses the most simplified approximation to the landscape about two energetic basins. From this model, the energetics of transitions are determined. This approach has been successful in demonstrating the physical relationship between protein fluctuations (low frequency normal modes) and protein function (conformational transition), and thus serves as a benchmark for further work.

To elucidate the relationship between protein structure, folding and function, functional transitions have been modeled as a result of "hopping" [12, 14] between structure-based energy surfaces. These structure-based potentials, which were inspired by the work of Gō[16], have had great success in explaining the interplay between protein structure and protein folding[17, 18]. A limitation of these models is that the two structure-based energy surfaces have many nearly redundant contributions, since the conformations of interest have structural overlap. When applying these models to entire proteins, these near-redundancies may, or may not, contribute to the conformational changes. These redundancies add a degree of uncertainty to the physical interpretation of the system. Therefore, redundant interactions have been removed and replaced by single contacts for both structures.

Inspired by the successes of minimalist structure-based models in advancing our understanding of protein folding and molecular recognition[17, 18, 19, 20, 21, 22, 23, 24, 25, 26], our approach begins with the established theoretical framework of protein folding. As described below, we extend this framework to account for large conformational changes.

It is well established that protein folding is the result of a globally funneled, minimally frustrated energy landscape[1, 2, 4]. The application of the principle of minimal frustration via structure-based potentials with single native basins has had considerable success in explaining the physics of protein folding. To now explain large functional transitions, there is a need for multiple basins. Thus, we generalize the minimally frustrated energy landscape of protein folding studies to incorporate biologically functional motions. We propose protein structure dominates functional behavior, as well as protein folding. Thus, we begin with

a structure-based potential and add gradual perturbations, based on an alternate structure, to produce multiple minima. Our model implies, as does our previous model[10], that the transition ensemble can be determined from information of the conformations of interest. This is an implication of structure-based models in general (folding transition states can be determined by information of the native state). Using our model, we also show that multiple stable conformations may be due to amplified roughness in the global energetic landscape upon ligand binding.

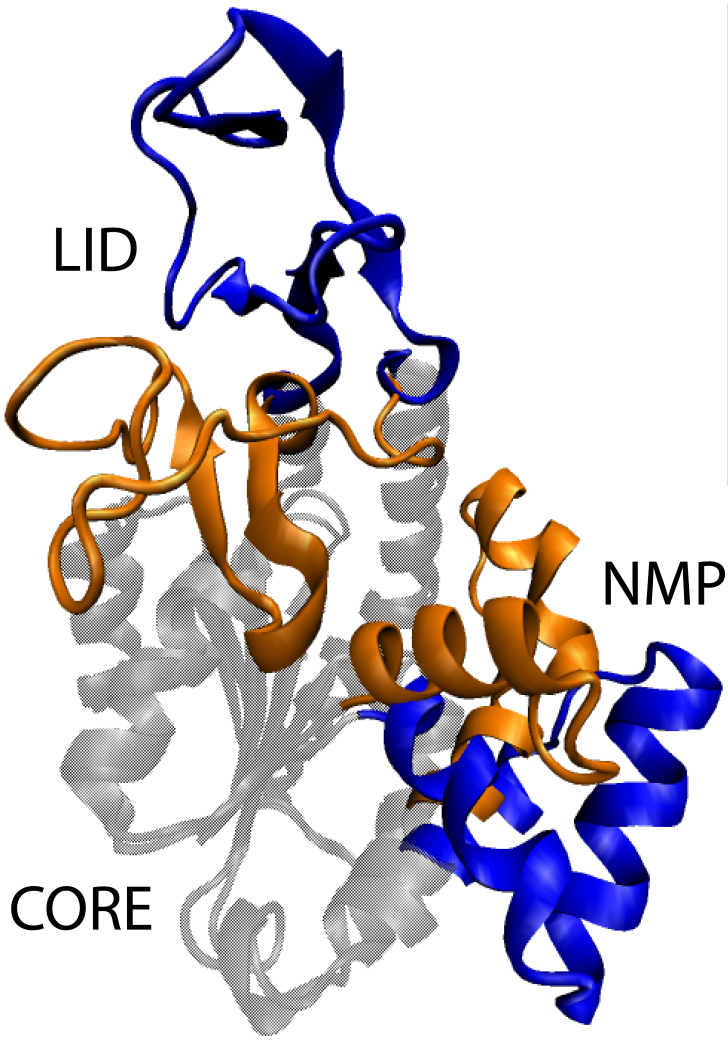

Figure 2.1: Functionally Relevant Conformations of AKE. Structure of the open (blue)[42] and closed (orange)[5] forms of AKE, with the CORE domain spatially aligned (grey). ATP binds in the pocket formed by the LID and CORE domains. AMP binds in the pocket formed by the NMP and CORE domains. Figure prepared with VMD[43].

Some proteins undergo large conformational changes without the aid of a co-factor. In allosteric proteins, however, such as Calmodulin and AKE, large conformational changes are associated with a co-factor, often an ion or a small biological molecule. Our model is general enough to be applied to both allosteric and non-allosteric conformational changes.

The model protein used in this study is E. Coli Adenylate Kinase. AKE is a 214 residue 3 domain protein (Figure 4.2) that catalyzes the reaction

$$ATP + AMP \overset{AKE}{\rightleftarrows} 2ADP \tag{2.1}$$

while undergoing large conformational changes which are believed to be the rate limiting steps of the reaction[30]. This protein was chosen mainly because it is well established that its multiple structures are catalytically relevant and because there is evidence that the conformational changes are rate limiting. Moreover, AKE is a good protein system to study the physics of conformational switching because there is a large amount experimental and theoretical data available on this process.

2.3 Results

2.3.1 Hamiltonian determination and implications

First, we developed several potentials and determined which reproduces the structural properties of the open and closed forms of AKE. Second, we employed the superior potential to study the conformational transitions of AKE. Analogous to protein folding models where information of the native state is used to model the folding properties, this work uses information about two stable forms of AKE to infer conformational transition properties. To determine which potential most accurately accounts for the structural properties of AKE's conformations, we compared conformational preference (i.e., open or closed), interresidue distance distributions and B-factors to experimental results. Four proposed Hamiltonians were compared: H^{open-D}_{open-C} (open structure potential), $H^{open-D}_{closed-C}$ (open/closed mixed structure potential), $H^{closed-D}_{open-C}$ (closed/open structure potential) and $H^{closed-D}_{closed-C}$ (closed structure-based potential, see *Models and Methods*). Each potential stabilizes the

contacts native to the open or closed form (denoted by C-open and C-closed) and the dihedral angles found in the open or closed form (D-open and D-closed). According to the above criteria, H_{open-C}^{open-D} reproduces experimental results most accurately (explained below).

The first experimentally known property of AKE that our potential must reproduce is that the unligated protein must be predominantly in the open form. Since we later propose ligand binding can be represented by introducing contacts unique to the closed form (which are scaled by ϵ_2; see *Models and Methods*), the simulated AKE without the contacts unique to the closed form (i.e. $\epsilon_2 = 0.0$) must also be in the open form. H_{open-C}^{open-D} and $H_{open-C}^{closed-D}$ have this property. $H_{closed-C}^{open-D}$ and $H_{closed-C}^{closed-D}$ do not exhibit this property under any conditions (data not shown). Since the open state is not an energetic minimum (global or local) for $H_{closed-C}^{open-D}$ nor $H_{closed-C}^{closed-D}$ these are not appropriate potentials for our investigation, and were not further considered. This result suggests that the open state is not purely a consequence of entropy, but energetic contributions are important as well.

Distance distributions $P(r)$ of residues A55 and V169 (located in the NMP and LID domains, respectively) have been determined experimentally for unligated and ligated AKE[31] and were compared to the values obtained for the remaining two Hamiltonians: H_{open-C}^{open-D} and $H_{open-C}^{closed-D}$. R_{max} is the value of r at which $P(r)$ is a maximum. In simulations, R_{max} does not vary significantly for $T < T_f$. The resulting R_{max} values from simulations and experiments are summarized in Table 2.1. R_{max} for H_{open-C}^{open-D} with $\epsilon_2 = 0$ (unligated) agrees very well with experiments. For $H_{open-C}^{closed-D}$ residues A55 and V169 are closer than in experiments, indicating the relative locations of the LID and NMP domains do not reflect in-solution dynamics. R_{max} for the closed form ($\epsilon_2 = 1.6$ maintains a closed conformation) for both potentials agrees equally well with the crystal structure distances but not as well with the fluorescence results. These results support H_{open-C}^{open-D} as the more appropriate potential for our studies.

B-factors for each conformation, simulated under the different potentials, were compared to crystal structure B-factors. Correlation coefficients between simulated B-factors and experimental B-factors were computed. The correlation

Table 2.1: Summary of $C_\alpha - C_\alpha$ distance distributions for A55 and V169 from experiments[31] and simulations. R_{max} is value for which the $C_\alpha - C_\alpha$ distance distribution is peaked. This data suggests H^{open-D}_{open-C} is the more appropriate potential to represent the dynamical properties of AKE. [a] from pdb file 4AKE. [b] from pdb file 1AKE.

System	R_{max} (Å)	R_{cryst} (Å)
unligated AKE (experimental)	29.2-31.4	29.5[a]
H^{open-D}_{open-C} $\epsilon_2 = 0$	27.0-32.7	29.5
$H^{closed-D}_{open-C}$ $\epsilon_2 = 0$	22.0-30.0	29.5
$AKE - Ap_5A$ (experimental)	10.0-10.7	12.3[b]
H^{open-D}_{open-C} $\epsilon_2 = 1.6$	11.6-13.8	12.3
$H^{closed-D}_{open-C}$ $\epsilon_2 = 1.6$	11.8-13.4	12.3

coefficient for the B-factors from the open crystal structure and H^{open-D}_{open-C} is 0.68. To correlation coefficient between B-factors from the crystal structure and $H^{closed-D}_{open-C}$ is 0.56. Both potentials' B-factors were poorly correlated for the closed conformation (r<0.5). B-factors for all simulations were significantly larger than crystal structure B-factors, which is expected when comparing simulated to crystallographic B-factors[32]. Large B-factors have also been observed in all-atom, explicit solvent, simulations and were attributed to not including crystal contacts[33]. To validate this claim, we simulated the dimeric form (PDB entry 4AKE:chains A and B) of the open conformation using a structure-based potential. The native interactions were determined with CSU[68]. The average B-factor was reduced from the monomer value of 273 Å^2 to 78 Å^2 for the dimer, and the correlation with experimental B-factors decreased to 0.44. The experimental average B-factor is around 38 Å^2. Since including a fraction of the crystal contacts (via a dimer) significantly alters the dynamics, crystallographic B-factors may not accurately describe in-solution multi-domain protein dynamics. Though, the discrepancy may also imply that our model is too flexible. Since the simulated dimeric B-factors are two times larger than the crystal B-factors, it is possible that the energetic barriers in our model are too small by up to a factor of two.

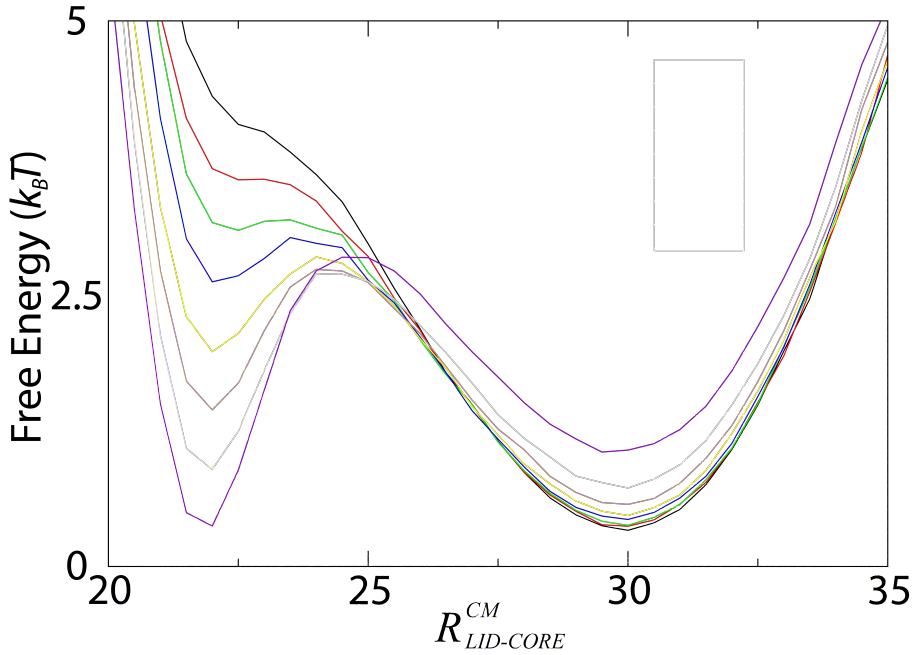

Figure 2.2: Contacts Native to the Closed Conformation Can Account for Large Conformational Changes. Free energy as a function of the distance between center of mass of the LID domain and CORE domain ($R^{CM}_{LID-CORE}$) for ϵ_2 = 0.5 – 1.2 (incremented by 0.1, colored black to purple). ϵ_2 is the interaction strength of closed conformation contacts, which represent ligand binding. For $\epsilon_2 > 0.6$ there are multiple minima indicating ϵ_2 can represent ligand binding accurately. For $\epsilon_2 < 0.7$ there is only one minimum corresponding to the open form, indicating non-open interactions can exist without distorting the open form.

The analysis of conformational preference for the four designed potentials shows that H^{open-D}_{open-C} most accurately represents the open and closed conformations of AKE and is thus used to study the conformational transitions. This decision has several implications. First, it suggests that the energy landscape has a single minimum corresponding to the open form, and additional minima are the result of perturbations from ligand binding. Second, the open conformation is the only stable state for some non-zero values of ϵ_2 (Figure 2.2). This demonstrates that contacts that are not native to the open form can exist and aid in conformational changes while not destabilizing the open form. Finally, ligand binding is well represented by increasing ϵ_2, therefore, to first approximation the effects of ligand binding are manifested in contacts found in the closed form.

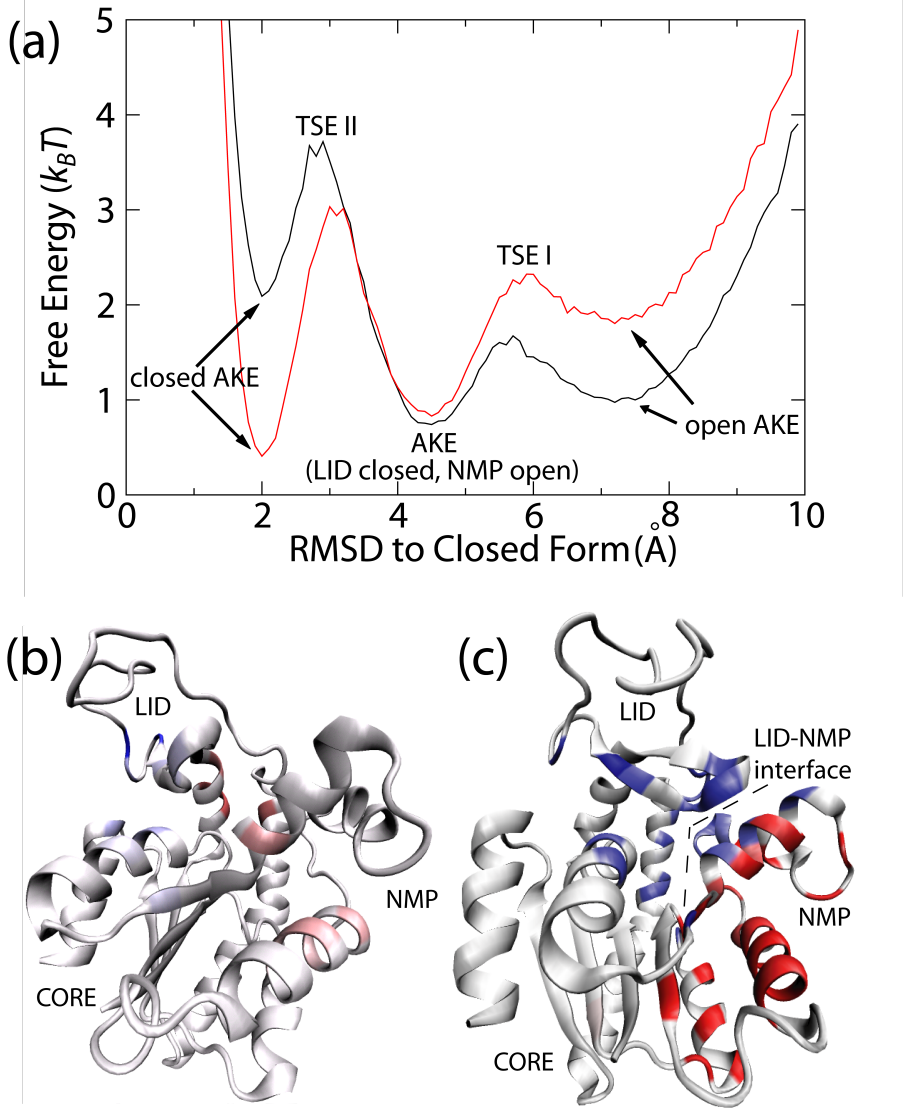

Figure 2.3: Multiple Transitions Seen in Conformational Rearrangement of AKE. (a) Free energy versus RMSD from the closed conformation for ϵ_2 = 1.2 (black) and ϵ_2 = 1.3 (red) shows the free energy barriers to close the LID domain (TSE I) and the NMP domain (TSE II). This result suggests NMP domain closure is rate limiting. Φ_{Func}-values mapped onto the closed structure for LID closure (b, rotated for clarity) and NMP closure (c). For residues with $\Delta\Delta G_{Y-X} < 0$ (residues that resist closing), Φ_{Func}-values are colored white (= 0) to red ($\geq$ 1). For $\Delta\Delta G_{Y-X} >$ 0 (residues that contribute to closing), Φ_{Func}-values are colored white (= 0) to blue ($\geq$ 1). The dotted line represents the LID-NMP interface, which contributes strongly to NMP domain closure. Figures (b) and (c) prepared with VMD[43].

2.3.2 Energetic barriers of conformational transitions

Free energy as a function of RMSD from the closed form is shown in Figure 2.3(a). TSE I and TSE II correspond predominantly to LID domain closure, and NMP domain closure, respectively (see Figure 2.4(d)). Figures 2.3(b) and 2.3(c) illustrate the energetic properties of TSE I and TSE II (see *Functional* Φ*-values* below). This finding is surprising, since there are more interactions between the NMP domain and CORE domain than there are between the LID domain and CORE domain. The barrier for LID closure is 0.9 k_BT and the barrier for NMP closure is 3.0 k_BT. Thus, if structural contributions dominate functional kinetics, then NMP domain closure should be rate limiting in AKE catalysis. While this prediction has not been made previously, it agrees with previous results from elastic network models where the lowest frequency normal mode corresponds to LID domain motion and the second lowest mode corresponds to NMP domain motion[35]. Thus, it is reasonable to expect the curvature of the NMP closure barrier will be greater than that of LID closure. While elastic network models have not explored the possibility of the intermediate we observe here (LID closed, NMP open), they have predicted a steeper strain energy barrier when opening the NMP domain than closing the LID domain[10], also in agreement with our results.

2.3.3 Localized strain energy and unfolding govern conformational changes

We previously proposed a cracking mechanism for allosteric proteins, based on normal mode analysis of AKE[10]. Upon translation along the lowest frequency modes, insurmountably high strain energy accumulated in very localized regions of the protein. This strain was enough to unfold the entire protein (>20kcal/mol), thus we predicted that localized regions of the protein unfold during conformational transitions, as a mechanism to reduce strain and enhance catalytic efficiency. The simulations reported here support the high strain energy and unfolding hypothesis (Figure 2.5).

Figure 2.5 (top left) shows the average strain energy (defined as the total

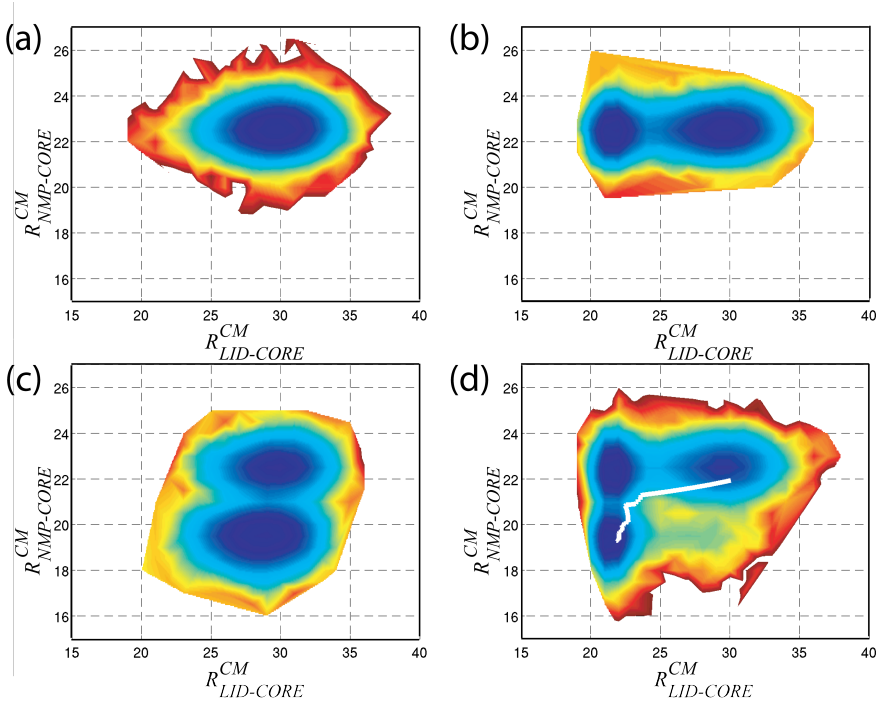

Figure 2.4: Proposed Hamiltonian Captures Dynamics of AMP, ATP and Ap_5A Binding. Free Energy surfaces for H^{open-D}_{open-C} with four subsets of Q_{Ligand} and varied ligand binding parameter ϵ_2. (a) ϵ_2 = 0.0 represents the unligated AKE. (b) $Q^{LID-CORE}_{Ligand}$ with ϵ_2 = 1.5 represents ATP binding. (b) $Q^{NMP-CORE}_{Ligand}$ with ϵ_2 = 1.9 represents AMP binding. (d) All Q_{Ligand} contacts, ϵ_2 = 1.3, represents Ap_5A binding, or simultaneous AMP and ATP binding. A predicted pathway generated via normal mode analysis[10] (white line in (d)) shows excellent agreement with our results. 10 k_BT energy scale (dark blue to dark red).

potential energy) by residue. There are clear peaks near residues 60-70, 120-125 and to a lesser extent residues 10-20, 30-35, 80-90 and 170-180. These finding are in excellent agreement with normal mode predictions of high strain in residues 10, 110-125, 150 and 160-170[10]. In normal mode studies, residues 30, 60 and 80 have only weak peaks at the later stage of the conformational transition. This is likely due to the previous data mainly reporting the strain associated with LID closure, and not NMP closure [10].

We believe that the high strain energy in AKE is the result of competing energetic contributions. Since competing energetic terms can not be satisfied simultaneously, internal strain must result. Some regions of strain drive the pro-

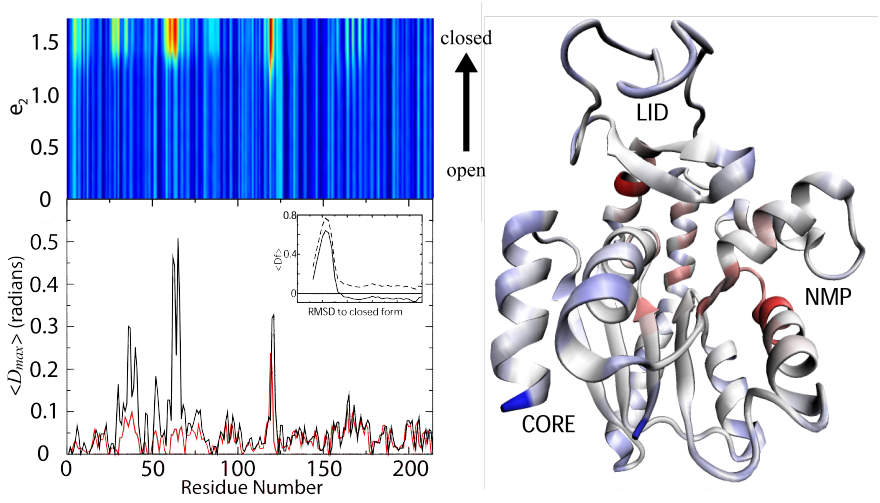

Figure 2.5: High Strain Energy Gives Rise to Local Unfolding. Strain energy as a function of binding parameter, ϵ_2, and residue number (top left), colored blue (low strain) to red (high strain). Unfolding measure, $< D_{max} >$, by residue number (bottom left). Red and black lines correspond to $< D_{max} >$ for LID and NMP transition. Average deviation from PDB dihedral values for dihedral angle 63 as a function of RMSD from closed form (inset). Strain energy for ϵ_2 = 1.7 mapped onto closed form of AKE. Red indicates high strain energy, blue indicates low strain energy and white indicated intermediate strain energy. The correlation between high strain energy and protein unfolding suggests unfolding is a mechanism by which strain energy is released during conformational changes. Analysis of individual domain motion (as seen in Figures 2.4(b) and (c)) shows that each peak in strain energy and $< D_{max} >$ is due to NMP or LID domain motion (not shown). Figure of structure prepared with VMD[43].

tein's opening transition and other regions of strain drive the closing transition (see *Functional* Φ*-values* below). Thus, the balance between competing strains energies is very important for the function of AKE.

Local unfolding was measured by calculating the average deviation in dihedral angle i from the closed and open conformations $< \Delta\phi_i^{open/closed} >$ as a function of RMSD from the closed conformation. The inset in Figure 2.5 shows $< \Delta\phi_{63}^{open} >$ and $< \Delta\phi_{63}^{closed} >$ (residues 63-66). The fact that $< \Delta\phi_{63}^{open} >$ and $< \Delta\phi_{63}^{closed} >$ both exceed zero indicates an unfolding event is occuring. To quantify the unfolding we calculated the D-value, defined as

$$D * 2 = | < \Delta\phi_i^{open} > | + | < \Delta\phi_i^{closed} > | - | < \Delta\phi_i^{open} > - < \Delta\phi_i^{closed} > | \qquad (2.2)$$

as a function of RMSD from the closed form. By definition, D is greater than zero unless $\phi_{closed} < \phi < \phi_{open}$, for which values D is zero (if $0 < \phi_{open} - \phi_{closed} < \pi$). In other words, if a given dihedral angle is between the corresponding angles of the open and closed forms, D equals zero, otherwise $D > 0$ as it deviate from the boundary. The maximum value of D for each TSE, D_{max}, was calculated for each dihedral angle and averaged by residue[1]. $< D_{max} >$ for the two energetic barriers is shown in Figure 2.5.

There is excellent agreement between regions of high strain energy and local unfolding (Figure 2.5). To no surprise the hinge region of the LID domain is under significant strain. More surprisingly, and in excellent agreement with normal mode predictions, is the strain experienced by residues 60-63. The different regions of strain correspond to strain that drives the protein opening and strain that drives the protein closing. Thus, large conformational changes in AKE are, in part, the result of competing intra-protein interactions. These interaction give rise to large strain energies, which are reduced through local unfolding.

2.3.4 Functional Φ-values

To determine which interactions are responsible for conformational changes, high strain energy and local unfolding, we calculated functional Φ-values for each residue[10, 11, 12], defined as

$$\Phi_{Func} = \frac{\Delta\Delta G_{\ddagger - X}}{\Delta\Delta G_{Y-X}}, \tag{2.3}$$

where $\Delta\Delta G_{\ddagger - X} = \Delta G_{\ddagger} - \Delta G_X$, $\Delta\Delta G_{Y-X} = \Delta G_Y - \Delta G_X$ and ‡ is the transition state between states X and Y[2]. The Φ_{Func} values for LID closure (TSE I) and NMP closure (TSE II) are mapped onto the structure of the closed form (Figures 2.3(b) and (c)). Before discussing the results, it is important to clarify the similarities and differences in the definition and interpretation of Φ_{Func}-values and protein folding

[1]For each residue, there are two D_{max} values for which that residue is one of the middle two residues constituting the angle. We average over these two D_{max} values for each residue.

[2]$\Delta\Delta G$ is the calculated change in free energy upon removal of all non-local interactions via perturbation theory. Thus, positive $\Delta\Delta G_{Y-X}$ values indicate that a residue stabilizes state Y more than state X.

Φ-values.

Protein folding Φ-values measure the amount of native content present in a residue at the folding transition state ensemble[36]. Owing to the funneled global landscape[2] and the principle of minimal frustration[3] $\Delta\Delta G_{Y-X} \geq \Delta\Delta G_{\ddagger-X} \geq 0$, where Y = native state and X = denatured state (i.e. non-local interactions stabilize the transition state more than the denatured state, and the native state more than the transition state). Thus, for protein folding, negative Φ-values and Φ-values greater than 1 are not easily interpreted.

As with folding Φ-values, Φ_{Func} measures the amount of state Y content present in the transition state ensemble. Though, due to the complexity of conformational changes, the same restrictions on $\Delta\Delta G$ and Φ_{Func} are not applicable. Since conformational changes arise from perturbations to the global landscape, $|\Delta\Delta G_{Y-X}|$ can be less than $|\Delta\Delta G_{\ddagger-X}|$ and $\Delta\Delta G_{\ddagger-X}$ can be negative. In order to obtain a complete picture of the dynamics, we must study Φ_{func} in conjunction with $\Delta\Delta G$.

A positive Φ_{Func}-value indicates $\Delta\Delta G_{Y-X} \geq \Delta\Delta G_{\ddagger-X} \geq 0$ or $\Delta\Delta G_{Y-X} \leq \Delta\Delta G_{\ddagger-X} \leq 0$. i.e., the residue stabilizes both the transition state and state Y (relative to state X), or destabilizes the transition state and state Y. Likewise, negative Φ_{Func}-values ($\Delta\Delta G_{Y-X} \geq 0 \geq \Delta\Delta G_{\ddagger-X}$ or $\Delta\Delta G_{Y-X} \leq 0 \leq \Delta\Delta G_{\ddagger-X}$) indicate the residue's energetic effect on the transition state is opposite of that on state Y. The ambiguity is resolved by observing $\Delta\Delta G_{Y-X}$.

An additional deviation from Φ-values is Φ_{Func} is not calculated if $|\Delta\Delta G_{Y-X}| < 0.75\ k_BT$ AND $|\Delta\Delta G_{\ddagger-X}| < 0.75\ k_BT$, rather than just using a cut-off for $|\Delta\Delta G_{Y-X}|$. Since we can't assume $|\Delta\Delta G_{Y-X}| > |\Delta\Delta G_{\ddagger-X}|$, residues may strongly affect the transition state and not state X nor Y. We do not want to filter this feature out accidentally.

Our model indicates, as shown in Figure 2.3(b), that residues 131, 135 and 143 (blue residues in LID domain) stabilize the closed LID domain ($\Delta\Delta G_{Y-X} > 0$) and contribute to closure of the LID domain ($\Phi_{Func} > 0$). Our model also predicts that residues 118 and 121 (red residues at hinge region between the LID and CORE domains) resist LID domain closure ($\Delta\Delta G_{Y-X} < 0$) and this resistance is accumu-

lated during the closing transition state ($\Phi_{Func} > 0$), in agreement with high strain energies observed during the closing transition in normal mode calculations[10]. Additionally, the large Φ_{Func}-values for $\Delta\Delta G_{Y-X} > 0$ that span the LID-NMP interface in Figure 2.3(c) and the lack of Φ_{Func}-values at the interface in Figure 2.3(b) indicate that NMP closure is stabilized substantially by the closed LID domain while LID closure is not highly influenced by the NMP domain. To further illustrate the NMP domain's dependence on the LID domain and the LID domain's independence of the NMP domain, we excluded the LID-NMP contacts and observed strong inhibition of NMP closure, with little effect on LID closure (not shown). Finally, our model shows that NMP intra-domain interactions resist conformational changes. Therefore, we predict that mutations in the core of the NMP domain will disrupt the interhelical interactions, reduce the energetic barrier to change conformation and/or stabilize the closed NMP domain.

These results support the claim that strain energy can be characterized as strain due to opening ($\Delta\Delta G_{Y-X} > 0$) or strain due to closing ($\Delta\Delta G_{Y-X} < 0$).

2.3.5 High strain and cracking are robust features of conformational changes

Using subsets of our ligand binding interactions, Q_{Ligand}, we were able to simulate the unligated (Figure 2.4(a)), ATP bound (2.4(b)), AMP bound (2.4(c)) and Ap_5A bound (2.4(d)) states. Since Ap_5A is a bisubstrate analogue for AMP and ATP, which are biologically relevant, we have focused our attention there, thus far.

Strain energy and $< D_{max} >$ values were also computed for ATP binding and AMP binding (not shown). We observe that each peak in strain energy and $< D_{max} >$ in Figure 2.5 corresponds exclusively to either LID closure or NMP closure (not shown). This suggests the coupling between high strain and local unfolding (cracking) is a property common to many multi-domain proteins, where the number of domains is inconsequential.

2.4 Conclusions

Using a coarse-grained model, we have shown that large conformational changes in AKE can be accounted for by the intra-protein contacts that are formed upon ligand binding. This work demonstrates that the energy landscape of AKE follows the principle of minimal frustration, with the addition of contacts of two competing native states. Analysis of the structural properties of AKE has been performed via this model and has yielded several novel findings. One finding is that energetically competing native interactions can exist in AKE and contribute to its functional dynamics. This model provides the first direct measurements of cracking (which was proposed based on the results from normal mode analysis of AKE[10]). We have further demonstrated that this local unfolding is the result of competing strain energies in the protein and that this phenomenon applies to the motion of individual domains, suggesting it is not limited to three-domain proteins.

While this energetically heterogenous structure-based model has had considerable success, it is yet to be established whether or not it is common for intra-protein interactions to produce competing strain energies that give rise to local unfolding during functional transitions, or if AKE is somehow unique. Additionally, it will be interesting to see to what extent non-local interactions contribute to conformational changes. These answers, in addition to a more detailed understanding of conformational changes, will hopefully become clear as theoretical models become more widely applied and refined.

2.5 Models and methods

2.5.1 Construction of the energy function

We study the conformational transitions of AKE by employing a structure-based Hamiltonian[21] with a modified contact map (described below). Structure-based potentials account for native interactions which are usually given the same energy weighting and produce a single funneled energy landscape. In a C_α-model each residue is represented by a single bead centered at the C_α position. Inter-

action energies of adjacent beads are harmonic in bond length and angle, with the geometry of the native state included through a dihedral term and non-local bead-bead interaction terms. Non-local contacts are included via

$$E_{contacts} = \epsilon_n(5(\frac{\sigma_{ij}}{r})^{12} - 6(\frac{\sigma_{ij}}{r})^{10}) \tag{2.4}$$

which has a minimum of depth ϵ_n at $r = \sigma_{ij}$, with σ_{ij} being the native distance between the $C_\alpha(i)$ and $C_\alpha(j)$ atoms in the crystal structure. In homogenous structure-based models, there is a single value for ϵ_n. To model conformational changes, different sets of contacts n are given different values of ϵ_n (see below). Temperature and energy values reported in this work are in units of ϵ_1 (interaction strength of contacts in Q^{open} or Q^{closed}, see below), and distances are reported in Angstroms. A detailed description of a structure-based potential can be found elsewhere[21].

A simple structure-based potential has a single minimum. Making such a model for each form of AKE is straightforward. In this work, however, we construct a single potential that has multiple minima. Thus, to extend a structure-based model to systems with multiple minima, we modified the contact map and the dihedral angles. To determine whether dihedral angles of non-local interactions govern conformational change, two sets of non-local interactions were determined, one based on the open conformation and the other based on the closed conformation. In addition, two sets of dihedral values were considered, one from the open conformation and the other based on the closed conformation. All four combinations of contact map and dihedral values were simulated and compared to experimental data.

Native contact maps Q^{open} and Q^{closed} were generated using the CSU software package[68], and are assigned interaction distances from their respective structures. Contacts that are unique to the closed form, and are over 50% further apart when in the open form (Q_{Ligand} in Figure 2.6) are deleted from Q^{closed}. Both contact sets, Q^{open} and Q^{closed}, were used in different simulations to model the open (unligated) conformation and were given an energetic weighting of $\epsilon_1 = 1.0$. These two contact maps were simulated to determine if non-local interactions govern the open conformational dynamics (see Results). To simulate the closed (ligand

bound) conformation, Q_{Ligand} contacts are added to each simulation with an energetic weight of ϵ_2, which is varied from 0.0 to 2.0. Q_{Ligand} is defined as the contacts when in the closed conformation with $C_\alpha - C_\alpha$ distances that are over 50% further apart in the open conformation. In summary, Q^{open} and Q^{closed} are contact maps based on the open and closed structures where additional contacts (Q_{Ligand}) are added to represent ligand binding.

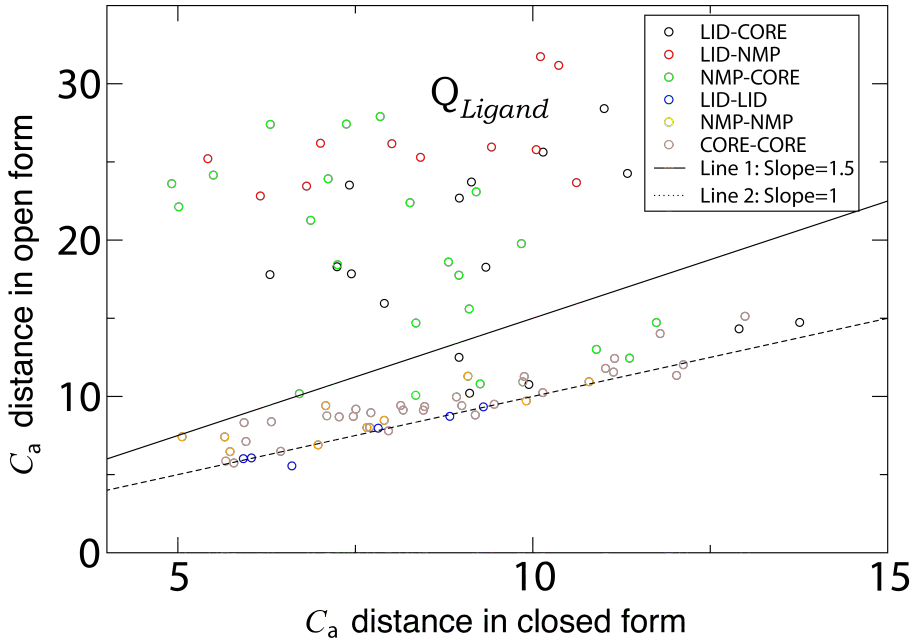

Figure 2.6: Contacts Unique to Closed Form. Each point represents a contact between residue i and residue j that is unique to the closed form. The Y-axis is the distance between the C_α atoms of residues i and j in the open form and the X-axis is the distance in the closed form. The locations of the residue pairs are indicated by color. i.e., black circles indicate the contact is between a residue in the LID domain and a residue in the CORE domain. Contacts above the line of slope 1.5 (solid line) constitute the set Q_{Ligand}.

Modeling ligand binding by including Q_{Ligand} interactions is a warranted approximation. Q_{Ligand} represents interactions that, inarguably, stabilize the ligand bound form. When $\epsilon_2 > 0$ there is an energetic bias to the closed form. Inspection of the ligand-protein interactions in AKE shows that many ligand mediated interactions are included in Q_{Ligand} through direct residue-residue interactions. Additionally, some missing ligand mediated interactions are accounted for by lo-

cal (in sequence) contacts. For example, the ligand mediated interaction between residues L58 and R167 is not included in Q_{Ligand}, though several contacts between L58 and residues 170-175 are included.

The mixing of multiple structure-based potentials has been successful in representing conformational tranistions[12, 14]. Previous studies mixed energetic terms at the global level (conformational switching is the result of all energetic contributions switching simultaneously between the two minima), whereas our potential is local (one interaction can stabilize the open conformation while a different interaction stabilizes the closed conformation). Thus, we must ask: Are dihedral angles or non-local contacts responsible for conformational transitions? To explore this question, Q^{open} and Q^{closed} were permuted with the open and closed values for dihedral angles, resulting in 4 possible Hamiltonians $H(Q^{open}, \phi^{open})$, $H(Q^{closed}, \phi^{open})$, $H(Q^{open}, \phi^{closed})$ and $H(Q^{closed}, \phi^{closed})$ [3] where C and D signify contacts and dihedral angles). The properties of AKE were examined with each potential. Several properties (conformational preference, interresidue distance distributions and B-factors) were compared to experimental values in order to determine the potential that most accurately captures AKE's dynamics about the two known structures. As discussed in the Results section, we find the potential H^{open-D}_{open-C} (single structure-based potential, based on the open structure, plus Q_{Ligand} contacts) is the most appropriate to study the conformational transitions of AKE.

2.5.2 Calculating thermodynamic properties

We used Molecular Dynamics (MD) to simulate the conformational changes of AKE. We developed our own software to simulate constant temperature runs. Temperature was maintained using the Berendsen algorithm to couple the system to a thermal bath[69]. Our code was tested extensively to ensure correct calculations of energy and force through systematic debugging and simulated folding of CI2 and α-spectrin SH3 domain proteins. The folding mechanisms and thermodynamics of these proteins are in excellent agreement with previously reported

[3] denoted by H^{open-D}_{open-C}, $H^{open-D}_{closed-C}$, $H^{closed-D}_{open-C}$ and $H^{closed-D}_{closed-C}$

simulations using AMBER[20].

Free energy profiles were obtained by simulating several constant temperature runs near room temperature and combining them by using the WHAM algorithm[39, 40].

The folding temperature, T_f, was approximated, via kinetic unfolding simulations to be between, 1.15 and 1.25 (in reduced units). All results in this paper are at T=0.9≈ $0.8T_f$, which corresponds approximately to room temperature.

2.5.3 R^{CM}_{X-CORE} and RMSD calculations

Allosteric conformational changes involve major domain motion. We measure this motion via the spatial distance between the centers of mass of the domains and the RMSD from the closed conformation. $R^{CM}_{LID-CORE}$ and $R^{CM}_{NMP-CORE}$ are the distances between the centers of mass of the LID and the CORE domains, and the distance between the centers of mass of the NMP and the CORE domains, respectively. $R^{CM}_{LID-CORE}$ is 30.1 Å and 21.0 Å in the open and closed form. $R^{CM}_{NMP-CORE}$ is 22.0 Å and 18.4 Å in the open and closed form. Domain definitions are given in Table 2.2. RMSD was calculated using the McLachlan algorithm[41] in the PROFIT software package.

2.5.4 Definition of open and closed states

The open and closed structures were obtained from PDB entry 4AKE, chain A[42] and 1AKE, chain A[5]. The system is "open" when $F(R^{CM}_{LID-CORE})$ and $F(R^{CM}_{NMP-CORE})$ have only one minimum each located near 30.1 Å and 22.0 Å. Figure 2.2 shows $F(R^{CM}_{LID-CORE})$ for several values of ϵ_2.

2.6 Acknowledgements

We are are particularly grateful to Peter Wolynes for many discussions, in particular on the concept of Φ_{Func}. We also thank S. Takada and K. Okazaki for insightful comments on the manuscript. This work was funded by the NSF-

Table 2.2: Summary of symbols and names.

Name/Symbol	*Description*
CORE domain	Residues 1-29, 68-117 and 161-214
NMP domain	Residues 30-67
LID domain	Residues 118-160
H^{X-D}_{Y-C}	Hamiltonian with dihedral angles (D) from conformation X(=open/closed) and contact set (C) Q^Y (Y=open/closed). There are 4 possibilities: H^{open-D}_{open-C}, $H^{open-D}_{closed-C}$, $H^{closed-D}_{open-C}$ and $H^{closed-D}_{closed-C}$
Q^Y and σ^Y	Contact set and contact distances for conformation Y (=open/closed)
Q_{Ligand}	Contacts above line 1 in Figure 6: N=39
$Q^{LID-CORE}_{Ligand}$	Subset of Q_{Ligand}, contacts between LID domain and CORE domain: N=11
$Q^{NMP-CORE}_{Ligand}$	Subset of Q_{Ligand}, contacts between NMP domain and CORE domain: N=17
$Q^{LID-NMP}_{Ligand}$	Subset of Q_{Ligand}, contacts between LID domain and NMP domain: N=11
ϵ_1	Interaction strength of contacts in Q^Y
ϵ_2	Interaction strength of contact set Q_{Ligand}
$R^{CM}_{NMP-CORE}$	Distance between center of masses of NMP and CORE domains
$R^{CM}_{LID-CORE}$	Distance between center of masses of LID and CORE domains

sponsored Center for Theoretical Biological Physics (Grants PHY-0216576 and 0225630) and the NSF Grant 0543906. PW is supported by the NIH Molecular Biophysics Training Grant at UCSD (Grant T32 GM08326).

Chapter 2, in full, appears in Journal of Molecular Biology, 2007, Whitford, Miyashita, Levy, Onuchic. The dissertation author is the primary investigator and author of the paper.

Bibliography

[1] Swain, J. & Gierasch, L. *Curr. Opin. Struct. Biol.* **2006**. 16, 102-108.

[2] Ogawa, A., Takayama, Y., Sakai, H., Chong, K.T., Takeuchi, S., Nakagawa, A., Nada, S., Okada, M. & Tsukihara, T. *J. Biol. Chem.* **2002**, 277, 14351-14354.

[3] DeBondt, H.L., Rosenblatt, J., Jancarik, J., Jones, H.D., Morgant, D.O. & Kim, S. *Nature*, **1993**, 363, 595-602.

[4] Xu, Z., Chaudhary, D., Olland, S., Wolfrom, S., Czerwinski, R., Malakian, K., Lin, L., Stahl, M., Joseph-McCarthy, D., Benander, C., Fitz, L., Greco, R., Somers, W. & Mosyak, L. *J. Biol. Chem.* **2004**, 279, 50401-50409.

[5] Muller, C.W. & Schulz, G.E. *J. Mol. Biol.*, **1992**,224, 159-177.

[6] Gunasekaran, K., Ma, B., & and Nussinov, R. **2004**,57, 433-443.

[7] Tsai, C.J., Kumar, S., Ma, B. & Nussinov, R. *Prot. Sci.* **1999**, 8, 1181-1190.

[8] Eisenmesser, E.Z., Bosco, D.A., Akke, M. & Kern, D. *Science* **2002**, 295, 1520-1523.

[9] Kern, D. & Zuiderweg, E.R.P. *Curr. Opin. Struct. Biol.* **2003**, 13, 748-757.

[10] Miyashita, O., Onuchic, J.N. & Wolynes, P.G. *Proc. Nat. Acad. Sci. USA* **2003**, 100, 12570-12575.

[11] Miyashita, O., Wolynes, P.G. & Onuchic, J.N. *J. Phys. Chem. B* **2005**, 109, 1959-1969.

[12] Okazaki, K., Koga, N., Takada, S., Onuchic, J.N. & Wolynes, P.G. *Proc. Nat. Acad. Sci. USA* **2006**, 103, 11844-11849.

[13] Zuckerman, D.M. *J. Phys. Chem. B* **2004**, 108, 5127-5137.

[14] Best, R.B., Chen, Y. & Hummer, G. *Structure* **2005**, 13, 1755-1763.

[15] Maragakis, P. & Karplus, M. *J. Mol. Biol.* **2005**, 352, 807-822.

[16] Ueda, Y., Taketomi, H. & Gō, N. *Int. J. Pept. Res.* **1975**, 7, 445-459.

[17] Shoemaker, B.A., Wang, J. & Wolynes, P.G. *Proc. Nat. Acad. Sci. USA.* **1997**, 94, 777-782.

[18] Nymeyer, H., Garcia, A.E. & Onuchic, J.N. *Proc. Nat. Acad. Sci. USA.* **1998**, 95, 5921-5928.

[19] Chavez, L.L., Onuchic, J.N. & Clementi, C. *J. Am. Chem. Soc.* **2004**, 126, 8426-8432.

[20] Clementi, C., Jennings, P.A. & Onuchic, J.N. *J. Mol. Biol.* **2001**, 311, 879-890.

[21] Clementi, C., Nymeyer, H. & Onuchic, J.N. *J. Mol. Biol.* **2000**, 298, 937-953.

[22] Gosavi, S., Chavez, L.L., Jennings, P.A. & Onuchic, J.N. *J. Mol. Biol.* **2006**, 357, 986-996.

[23] Levy, Y. & Onuchic, J.N. *Acc. Chem. Res.* **2006**, 39, 135-142.

[24] Levy, Y., Cho, S.S., Shen, T., Onuchic, J.N. & Wolynes, P.G. *Proc. Nat. Acad. Sci. USA.* **2005**, 102, 2373-2378.

[25] Levy, Y., Cho, S.S., Onuchic, J.N. & Wolynes, P.G. *J. Mol. Biol* **2005**, 346, 1121-1145.

[26] Yang, S.C., Cho, S.S., Levy, Y., Cheung, M.S., Levine, H., Wolynes, P.G. & Onuchic, J.N. *Proc. Nat. Acad. Sci. USA* **2004**, 101, 13786-13791.

[27] Leopold, P.E., Montal, M. & Onuchic, J.N. *Proc. Nat. Acad. Sci. USA* **1992**, 18, 8721-8725.

[28] Bryngelson, J.D., Onuchic, J.N., Socci, N.D. & Wolynes, P.G. *Proteins* **1995**, 21, 167-195.

[29] Onuchic, J.N. & Wolynes, P.G. *Curr. Opin. Struct. Biol.* **2004**, 14, 70-75.

[30] Wolf-Watz, M., Thai, V., Henzler-Wildman, K., Hadjipavou, G., Eisenmesser, E.Z. & Kern, D. *Nat. Struct. Biol.* **2004**, 11, 945-949.

[31] Sinev, M.A., Sineva, E.V., Ittah, V. & Haas, E. *Biochemistry* **1996**, 35, 6425-6437.

[32] Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O. & Bahar, I. *Biophys. J.* **2001**, 80, 505-515.

[33] Krishnamurthy, H., Lou, H., Kimple, A., Vieille, C. & Cukier, R. *Proteins: Struct. Funct. Bioinfo.* **2005**, 58, 88-100.

[34] Sobolev, V., Wade, R., Vried, G. & Edelman, M. *Proteins: Struct. Funct. Genet.* **1996**, 25, 120-129.

[35] Temiz, N.A., Meirovitch, E. & Bahar, I. *Proteins: Struct. Funct. Bioinfo.* **2004**, 57, 468-480.

[36] Fersht, A., *Structure and mechanism in protein science: A guide to enzyme catalysis and protein folding*, W.H. Freeman and Company, (1999).

[37] Bryngelson, J.D. & Wolynes, P.G. *Proc. Nat. Acad. Sci. USA.* **1987**, 84, 7524-7528.

[38] Berendsen, H.J.C., Postma, J.P.M., VanGunsteren, W.F., Dinola, A. & Haak, J.R. *J. Chem. Phys.* **1984**, 81, 3684-3690.

[39] Ferrenberg, A.M. & Swendsen, R.H. *Phys. Rev. Letters.* **1988**, 61, 2635-2638.

[40] Ferrenberg, A.M. & Swendsen, R.H. *Phys. Rev. Letters.* **1989**, 63, 1195-1198.

[41] McLachlan A.D. *Acta. Cryst. A.* **1982**, 38, 871-873.

[42] Mueller, C.W., Schlauderer, G.J., Reinstein, J. & Schulz, G.E. *Structure* **1996**, 4, 147-156.

[43] Humphrey, W., Dalke, A. & Schulten, K. *J. Molec. Graphics* **1996**, 14, 33-38.

Chapter 3

Conformational Transitions in Adenylate Kinase: Allosteric Communication Reduces Misligation

3.1 Abstract

Large conformational changes in the LID and NMP domains of Adenylate Kinase (AKE) are known to be key to ligand binding and catalysis, yet the order of binding events and domain motion is not well understood. Combining the multiple available structures for AKE with the energy landscape theory for protein folding, a theoretical model is developed for allostery, order of binding events and efficient catalysis. Coarse-grained models and non-linear normal mode analysis are used to infer that intrinsic structural fluctuations dominate LID motion, while ligand-protein interactions and cracking (local unfolding) are more important during NMP motion. In addition, LID-NMP domain interactions are indispensable for efficient catalysis. LID domain motion precedes NMP domain motion, both during opening and closing. These findings provide a mechanistic explanation for the observed 1:1:1 correspondence between LID domain closure, NMP domain closure and sub-

strate turnover. This catalytic cycle has likely evolved to reduce misligation, and thus inhibition, of AKE. The separation of allosteric motion into intrinsic structural fluctuations and ligand-induced contributions can be generalized to further our understanding of allosteric transitions in other proteins.

3.2 Introduction

Kinase mediated phosphoryl transfer is a key component of many signalling pathways. Tight control of signalling requires regulation of kinase activity, which is influenced by allosteric transitions[1, 2]. The classical description of allostery involves ligand induced conformational rearrangements between static protein structures. It is now acknowledged that protein dynamics is statistical in nature, and that allostery is often due to a change in the balance of pre-existing conformational substates upon ligand binding[3]. This manuscript explores the subject of how protein structure determines allostery, order of ligand binding events and ultimately efficient catalysis, in the context of adenylate kinase (AKE).[1] AKE is a three domain (LID, NMP and CORE) protein (Figure 3.1) that undergoes large conformational changes as it catalyzes the reaction

$$ATP + AMP \overset{AKE}{\rightleftarrows} 2ADP \tag{3.1}$$

Large motions of the LID and NMP domains are associated with nucleotide binding. It has been shown that substrate turnover and domain rearrangements ("open" state to "closed" state) occur at the same frequency[2]. Despite several theoretical[6, 7, 8, 9, 10, 11, 12, 13] and experimental studies[2, 14, 15, 16, 17] on AKE, a catalytic mechanism that explains this high efficiency has not been proposed.

Since it is known that conformational transitions can be well described as a superposition of normal modes,[6, 7, 13, 18, 19, 20] we use a simplified non-linear elastic network model and a structure-based model withimplicit ligand interactions to demonstrate a likely catalytic mechanism in AKE. We show that intrinsic

[1]The abbreviations used are: AKE, adenylate kinase; NMA, normal mode analysis; MD, molecular dynamics; PDB, Protein Data Bank; CM, center of mass.

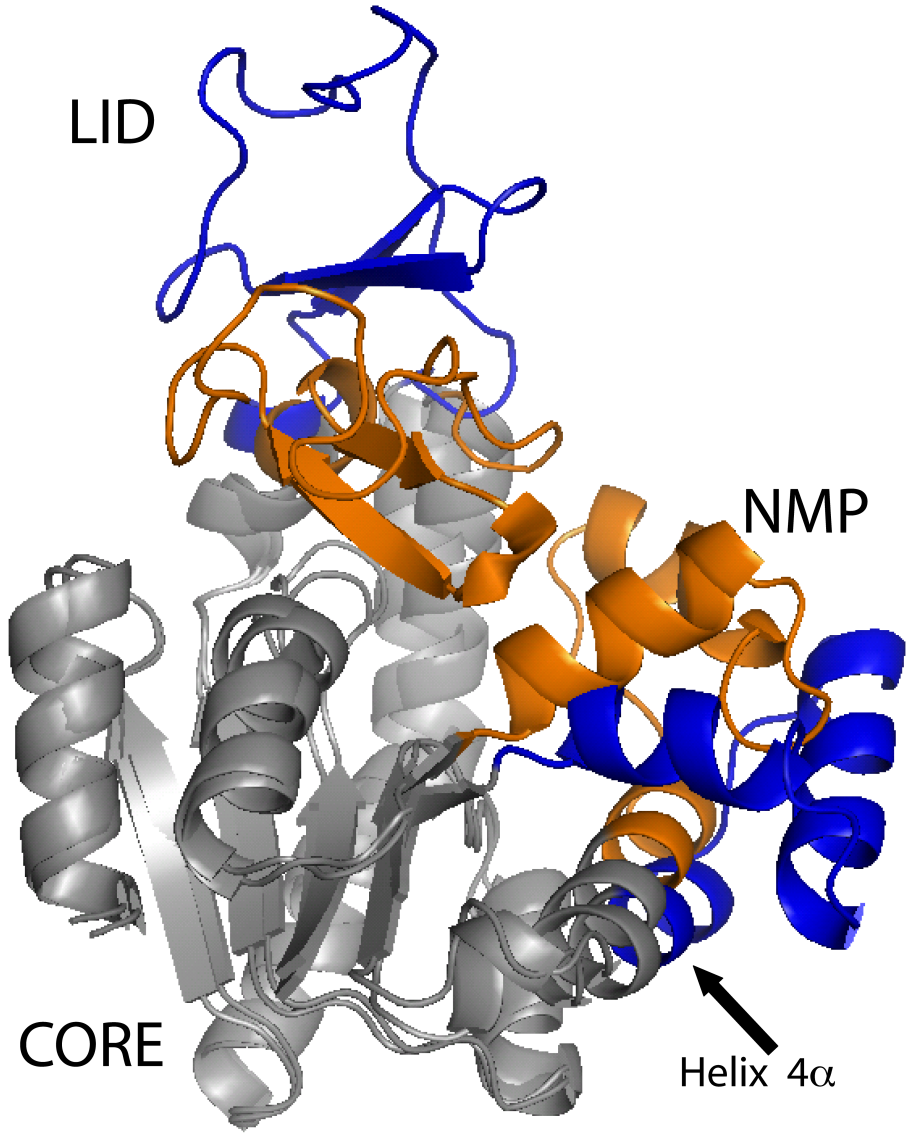

Figure 3.1: Crystal Structures of AKE. Structure of the open (blue, PDB entry 4AKE4) and closed (orange, PDB entry 5) LID and NMP domains of AKE, with the CORE domain (grey) spatially aligned. AKE can accommodate two ligands, one in the pocket between the LID and CORE domains and one in the NMP-CORE pocket. ATP, ADP, and AMP are able to bind the LID-CORE pocket. ADP and AMP are able to bind to the NMP-CORE pocket.

structural fluctuations[21, 22] in the LID domain account for the majority of the domain's motion, and that ATP binding likely serves to lock the domain closed. The LID motion is an example of allosteric regulation that results from a shift in the populations of available substates. The intrinsic motion of the NMP domain does not correlate with domain closure. Thus the NMP domain is an example of a ligand-induced conformational change. Additionally, the high strain in the NMP domain supports the claim that the NMP domain likely partially unfolds (cracks) during conformational transitions. Upon closure of the LID domain, the LID-NMP interface provides enthalpic interactions that stabilize the closed NMP domain. This provides a driving force for NMP closure and leads to phosphoryl transfer. Intrinsic structural fluctuations then drive the LID domain open, which destabilizes the LID-NMP interface and drives the NMP domain open. This mechanism which involves correlated motion of the domains reduces mis-ligation and may have evolved to increase the efficiency of AKE.

3.3 Methods

3.3.1 Multiple minima Molecular Dynamics simulations

Structure-based models allow one to study the relationship between protein folding[24, 25, 26, 27, 28, 29] and protein function[8, 30, 31, 32]. To characterize the energetics of the conformational change transition states of AKE, we simulated it using a coarse-grained structure-based potential that utilizes information from the open and closed forms. Each residue is represented by a bead located at the C_α position. Backbone geometry is maintained through harmonic potentials for adjacent bond distances and bond angles, and cosine functions for dihedrals, with minimum values corresponding to the open conformation (PDB entry 4AKE[4]). Non-local contacts are determined using CSU[33] analysis of the open conformation and are included via a 10-12 potential. Non-contacts are given repulsive terms. In addition to the open contacts, contacts unique to the closed structure (PDB entry 1AKE[5]) were included to induce domain closure. The closed contacts provide an implicit representation of the ligand. A complete description of the potential

can be found elsewhere[8]. For this study, we reduce the interaction strength of all contacts formed with helix 4α by 30%. Reducing these interaction strengths noticeably alters the population of the NMP-closed-LID-open conformation, which illustrates the significant effect helix 4α has on the catalytic dynamics.

3.3.2 Non-Linear normal mode analysis

Normal modes are known to capture the dynamics of large conformational changes in many proteins[6, 7, 23]. In this study, we use a non-linear normal mode analysis to describe the intrinsic contributions a given structure provides to each conformational rearrangement. This approach, first introduced by Miyashita et al., provides a better representation of global motion far from equilibrium by minimizing steric and energetic contradictions observed by standard linear normal mode analysis.

Initially, normal modes $\overrightarrow{N}$ for the open form of AKE (pdb entry 4AKE) are determined using a Tirion potential[34], where all atom pairs within 5 Åare connected by harmonic springs. AKE is translated along each mode a distance $dR_i = \Lambda \overrightarrow{d} \cdot \overrightarrow{N}_i B_i/B_t$, where B_i/B_t is the fraction of the total B-factor contributed by mode i and Λ is a chosen length scale (=0.05 Å). $\overrightarrow{d}_i$ is the spatial displacement of atoms between the current structure and the final structure (closed structure, pdb entry 1AKE) after rms fitting using the McLachlan algorithm[35] in the PROFIT software package (Since $\overrightarrow{d}_i$ and rms fitting require 2 structures with an identical number of atoms and the open structure does not have a ligand, ligands were not explicitly represented). Thus, the intrinsic overlap with a conformational change is $\frac{1}{\Lambda} \sum_{i}^{modes} dR_i$. At every step, a new Tirion potential is determined, based on the translated structure, and the non-linear normal mode analysis is repeated. This process is repeated until the closed conformation is reached. The same process is used for the opening of AKE and for individual domain rearrangements. This method is similar to a previously applied method,[6, 7] except that the B_i/B_t term in dR_i was not included in previous studies. While the B_i/B_t term does not qualitatively alter the dynamics, it modifies overlap to be the percent of intrinsic motion a structure has in the direction of a given conformational transition. The

strain energy is determined by calculating the total potential energy as defined by the Tirion potential for the initial structure. As discussed by Miyashita et al.,[6, 7] if too much local strain is accumulated then cracking (local unfolding) may reduce the strain during conformational transitions.

3.3.3 Reaction coordinates

Allosteric conformational changes involve major domain motion. We measure this motion using the spatial distance between the center of masses of the domains. $R_{CM}^{LID-CORE}$ and $R_{CM}^{NMP-CORE}$ are the distances between the center of mass of the LID and the CORE domains, and the distance between the center of mass of the NMP and the CORE domain, respectively. $R_{CM}^{LID-CORE}$ is 30.1 Åand 21.0 Åin the open and closed forms while $R_{CM}^{NMP-CORE}$ is 22.0 Åand 18.4 Å. The LID domain is defined as residues 118-160, the NMP domain as residues 30-67 and the CORE domain as residues 129, 68-117 and 161-214.

3.4 Results

3.4.1 Intrinsic motion in the LID domain of AKE contributes significantly to allosteric motion

The LID and NMP domains of AKE rearrange relative to the CORE domain during catalysis (Figure 3.1). To determine the order of domain motion, closing and opening trajectories were constructed using recursive non-linear normal mode analysis (see methods). At every step of this recursive procedure, the domains were translated along each mode an amount proportional to the mode's overlap with the conformational transition. This process filters for high-B-factor contributing and high-conformational overlap modes. Since B-factors are a measure of atom mobility within a protein, this method not only determines a likely pathway but provides a quantitative measure for the intrinsic (ligand-free) propensity of the protein to undergo a given conformational rearrangement.

Figure 3.2 shows the trajectories as functions of $R_{CM}^{LID-CORE}$ and $R_{CM}^{NMP-CORE}$,

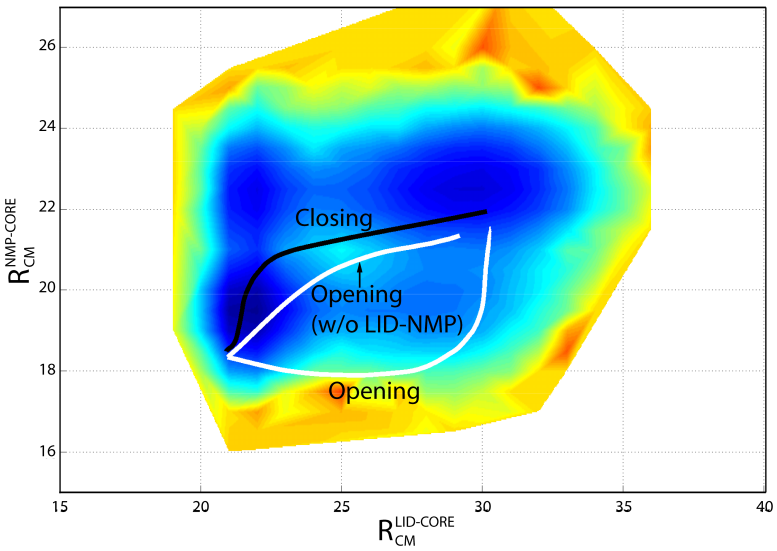

Figure 3.2: Intrinsic fluctuations direct conformational dynamics. Non-linear normal mode trajectories are superimposed on the free energy landscape obtained via MD simulations. Axes are the distance between LID domain and CORE domain centers of mass, $R_{CM}^{LID-CORE}$, and the distance between NMP domain and CORE domain centers of mass, $R_{CM}^{NMP-CORE}$. NMA suggests sequential closure and opening of the LID and NMP domains, in agreement with the free energy landscape. Intrinsic fluctuations promote LID opening prior to NMP opening. Removal of LID-NMP interactions eliminates the NMP domain's dependence on the LID domain during opening. The NM trajectories begin and end at the crystal structures of AKE. The free energy minima for the open and closed forms are at slightly larger values of $R_{CM}^{LID-CORE}$ and $R_{CM}^{NMP-CORE}$, due to higher entropy in more extended structures. When NM trajectories are shifted to slightly larger R_{CM} values, there is excellent agreement between the two methods.

the distances between the center of masses of the LID and CORE domains and the NMP and CORE domains. These trajectories indicate that LID domain motion precedes NMP domain motion, both during the opening and closing processes. There are two possible explanations for this observed sequential motion: 1) The open (closed) protein has the intrinsic ability to close (open) the LID domain and not the NMP domain. Once the LID domain closes (opens), the dynamics of the NMP domain are altered such that the intrinsic fluctuations about the LID-closed (open) state favor NMP closure (opening). 2) There is an intrinsic ability to close (open) the LID domain, and not the NMP domain. After LID closure (opening), the method employed merely forces the NMP domain to close (open).

3.4.2 The LID-NMP interface facilitates communication between LID and NMP motion

To determine the reason for the observed sequential motion, trajectories were constructed for individual domain rearrangements (Figure 3.3). LID domain motion has a higher intrinsic overlap and lower strain associated with its motion than does NMP motion. When both domains are open or closed, there are lower barriers associated with LID motion than NMP motion. Also, a significant increase (Figure 3.3: starred versus double-starred) in the slope of the strain associated with NMP opening is observed upon the LID-NMP interface formation.

The combination of a reduced intrinsic overlap and the higher energetic barriers associated with NMP motion suggests that NMP motion is more dependent on enthalpic contributions, such as ligand binding and LID-NMP interface formation. Additionally, the low intrinsic overlap suggests the NMP domain may enter a disordered (locally unfolded) state during conformational transitions. This agrees with structure-based simulations of the open form of AKE that have shown the closed LID domain is more easily accessible than the closed NMP state (Figure 4a of Ref. [8]), and that local unfolding is observed in the NMP domain (Figure 5 of Ref. [8]). Additionally, since the LID domain has a propensity to close, and the NMP domain does not, stabilizing interactions must be available to close the NMP domain.

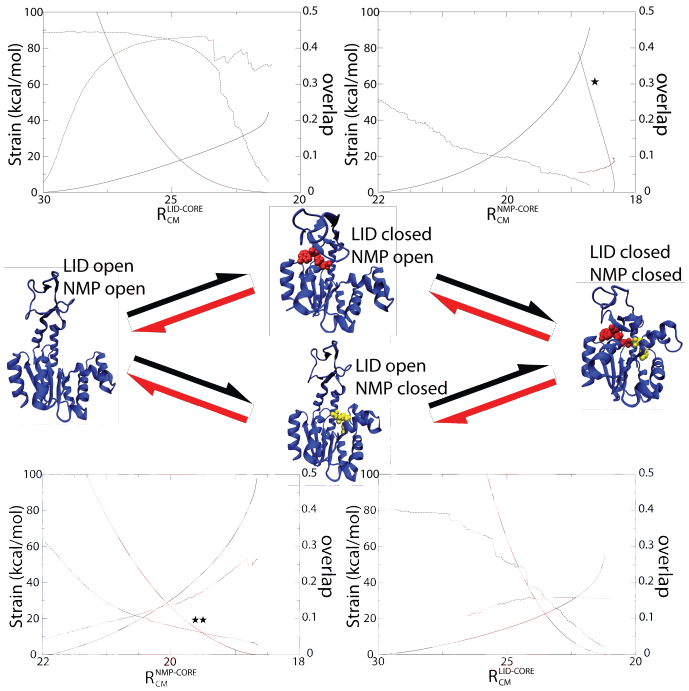

Figure 3.3: Intrinsic motion of LID domain overlaps with allosteric conformational transition. Four possible conformations during catalysis are shown with the associated overlap and energetic barriers: both domains open, the LID domain closed with the NMP domain open, both domains closed, and the LID domain open with the NMP domain closed. Solid lines correspond to the strain associated with opening (black) and closing (red) each domain. Dotted lines represent the intrinsic overlap of opening (black) and closing (red). LID domain closure has a higher intrinsic overlap (0.45) than does NMP domain closure (0.3), and the NMP overlap drops more quickly as the domain closes (decreasing R_{CM}). Thus, it is likely that the interactions that stabilize the closed state are more important for NMP closure than for LID closure. The largest energetic barrier is associated with NMP opening prior to LID opening. The barrier associated with NMP motion when the LID domain is closed (starred) is greater than when the LID domain is open (double starred). This larger barrier height is due to the steep strain profile associated with opening of the NMP domain. Because the most significant structural difference involving the NMP domain is the degree of LID-NMP interface formation, this interface probably plays a role in regulating domain motion, and ultimately activity. The third important feature is the higher intrinsic overlap with LID opening (0.15) than with NMP opening (0.1). The fourth feature is that the overlap of NMP opening increases by 300% (from 0.1 to 0.3) when the LID domain is already open. These last two features further illustrate the significant effect the LID-NMP interface has on catalytic dynamics.

3.4.3 LID domain binding of ATP assists NMP closure

Using a simplified structure-based potential (see methods) and MD simulations, the free energy surface of the conformational rearrangements in AKE was calculated (Figure 3.2)[2] . This model suggests strong coupling between NMP and LID motion in agreement with the non-linear normal model trajectories. Using the transition state ensembles for each domain's motion, we can calculate functional Phi values Φ_{Func} which are analogous to protein folding Φ-values. Protein folding Φ-values measure the free energy contribution to the folding transition state relative to the native state[36] for a given residue. Since proteins tend to be minimally frustrated, folding results from the balance between the entropy of the unfolded state and native enthalpic interactions. Thus, Φ-values measure the amount a given residue drives the folding process. Since the entropy of the various conformations of an allosteric protein may be comparable, enthalpic contributions can play a large role in both the forward and reverse conformational transitions. Thus, high Φ_{Func}-values show a given transition is dominated by enthalpic interactions and low Φ_{Func}-values show intrinsic structural fluctuations dominate the transition. Therefore, one can determine whether a given residue with a high Φ_{Func} stabilizes the forward or the backward reaction.

Larger Φ_{Func}-values for the NMP domain suggest that energetics are more important for NMP motion than for LID motion. Φ_{Func}-values for residues at the LID-NMP interface are large for the NMP transition state, but nearly zero for the LID transition state. Therefore, not only is NMP motion highly dependent on enthalpic interactions but a significant number of these interactions are made accessible though closure of the LID domain. This suggests that NMP motion is strongly influenced by the state of the LID domain (open/closed) as well as AMP binding to its binding site. Thus, both non-linear normal mode analysis and MD simulations suggest that intrinsic fluctuations are not the main contributors to the NMP transition.

[2] A discrepancy exists between the NM trajectories and the locations of the free energy minima because the MD landscape includes entropic contributions.

3.4.4 Domain dynamics can be controlled by mutating the LID-NMP interface and helix 4α

To test the importance of the LID-NMP interface on NMP domain dynamics, Non-Linear Normal Mode Analysis (NMA) was performed with all LID-NMP interface interactions removed (Figure 3.2). As expected, the LID domain no longer opens prior to NMP opening[3]. Additionally, if the LID-NMP interactions are not included in our MD simulations, the population of the NMP closed state is largely reduced. Since the interface plays a large role in stabilizing the closed NMP domain, and the entropy of the open and closed states is approximately the same (data not shown), enthalpic contributions must stabilize the open form. The 4α helix in the NMP domain fulfils this role. We repeated our previous simulations[8] using varying interaction strengths for native contacts made with the helix 4α (Figure 3.2). By reducing the strength of these interactions in simulations, we are able to alter the domain dynamics such that NMP closure followed by LID closure were as probable as LID closure followed by NMP closure. Additionally, by removing the LID-NMP interactions in our NMA we change the opening dynamics such that NMP opening precedes LID opening (Figure 3.2). These results illustrate the dramatic effects the LID-NMP interface and the helix 4α have on the catalytic dynamics of AKE.

3.5 Discussion

3.5.1 Catalytic cycle explains the 1:1:1 relationship between NMP motion, LID motion and substrate turnover.

Although the NMP and the LID domains are allowed to move independently[37, 38], their conformational transition rates are equal[2]. Despite evidence that substrate turnover and domain motion are correlated, the details of the mechanism are still unknown. The data presented in this manuscript suggests a catalytic cycle

[3]The closing trajectories were not repeated without LID-NMP interactions as these interactions are not formed in the open conformation.

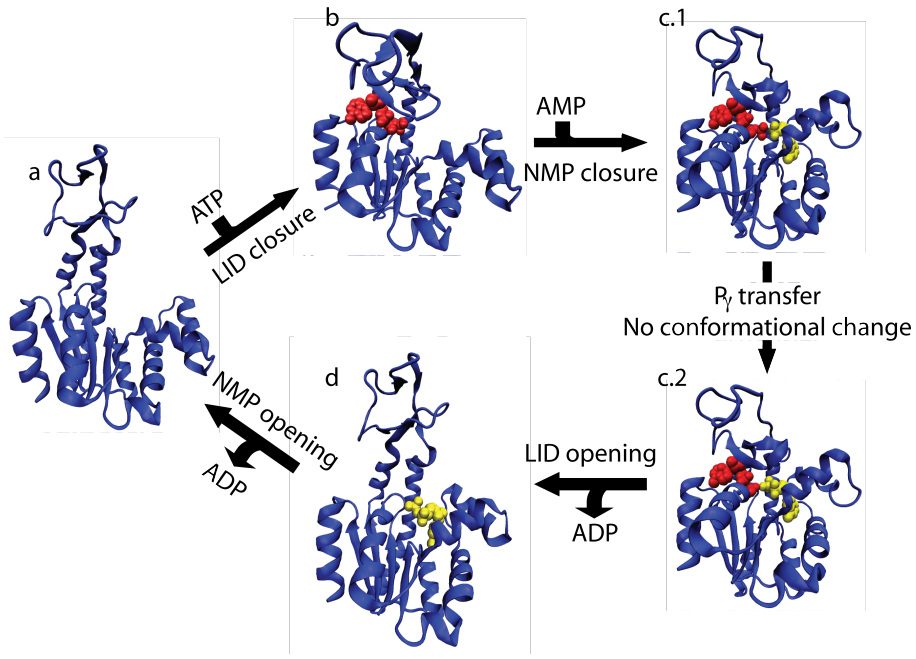

Figure 3.4: Proposed mechanism for AKE catalysis. All results presented suggest the following catalytic mechanism for AKE: Open, unligated AKE (a). ATP binds while the LID domain closes (b) followed by AMP binding/ NMP domain closure (c.1). Phosphoryl transfer occurs, resulting in 2 bound ADPs (c.2). Thermal fluctuations open LID domain and 1 ADP is released (d). Loss of LID-CORE interactions induces opening of NMP domain, loss of second ADP and a return to the open conformation (a). States c.1 and c.2 are modeled by deletion of one phosphoryl group from Ap5A in pdb structure 1AKE.

in which there is time ordered and sequential domain motion (Figure 3.4), i.e., LID domain motion precedes NMP motion. Assuming fast phosphoryl transfer, independent domain motion, correct ligand binding and conformational exchange rates of k_{LID} and k_{NMP}, substrate turnover rate would be $k_p = \frac{2}{\frac{1}{k_{LID}}+\frac{1}{k_{NMP}}}$. This rate is less than k_{LID} and k_{NMP} unless $k_{LID} = k_{NMP}$. Having independent domain motion that coincidentally had the exact same exchange rate is unlikely. Thus, an alternate explanation for $k_p = k_{LID} = k_{NMP}$ is that domain rearrangement is correlated, where the closure (opening) of one domain signals the rapid closure (opening) of the second domain. Energetic and structural considerations from simplified simulations and non-linear normal mode analysis suggest LID motion signals NMP motion in the following catalytic mechanism:

1. AKE is in the unligated open conformation (Fig. 3.4a)
2. LID domain closes and ATP binds (Fig. 3.4b)
3. LID closure enables AMP to bind concomitantly with NMP domain closure (Fig. 3.4c.1)
4. Phosphoryl transfer occurs, resulting in two bound ADPs (Fig. 3.4c.2)
5. ADP is released from the ATP site and the LID domain opens (Fig. 3.4d)
6. Opening of the LID domain signals the NMP domain to open and release the second ADP (Fig. 3.4a)

3.5.2 Catalytic cycle prevents mis-ligation

The proposed closing mechanism ensures that each conformational rearrangement contributes to the turnover of a substrate by preventing non-productive substrate binding. Several structural features of AKE are consistent with the proposed mechanism 1) the LID^{C}_{ATP} – NMP^{C}_{AMP} complex is the only product forming state,[4] 2) the ATP binding site (LID-CORE pocket) can accommodate ATP, ADP and AMP (in order of decreasing affinity), 3) the AMP binding site (NMP-CORE pocket) is only known to accommodate ADP and AMP[5] and 4) when in the LID^{C}_{ATP} – NMP^{O}_{AMP} state, the AMP site can only accommodate AMP.

During step 1 the binding affinities at the ATP site discriminate ATP over AMP by 3kcal/mol[17]. Since the LID^{C}_{ATP} – NMP^{O}_{AMP} state can only accommodate an AMP[6] , proper ligation in AKE would be limited by the ability of the ATP site to discriminate between AMP and ATP. If the alternate closing mechanism was common (followed by), AKE would have more opportunities to form

[4]The following notations were used: LID^{Y}_{X} = LID domain in state X (O=open,C=closed) with Y (ATP, ADP, AMP, or 0=no ligand) bound to the LID-CORE pocket; and NMP^{Y}_{X}=NMP domain in state X (O=open, C=closed) with Y (ATP, ADP, AMP) bound to the NMP-CORE pocket.

[5]Due to communication between domains and AMP binding of the ATP site, binding affinities for AMP may be difficult to determine exactly.

[6]Since reaction 1 is reversible LID^{C}_{ADP} – NMP^{C}_{ADP} can result in a product, though in this manuscript we assume steady state conditions for the forward reaction.

non-product forming (inhibited) complexes. In the $LID_o^O - NMP_{AMP}^C$ state the NMP domain is closed. NMP domain closure grants the LID domain access to the LID-NMP interface interactions. These interactions could reduce discrimination between the $LID_{ATP}^C - NMP_{AMP}^C$ and $LID_{AMP}^C - NMP_{AMP}^C$ states by perturbing the closing dynamics of the LID domain. In addition, the LID-NMP interface may provide enough stability to the closed LID domain that the LID may temporarily close without a ligand. If an ATP-free LID-closed state is formed, the LID domain would have to open prior to substrate binding and turnover. Thus, these non-functional states would increase domain motion frequency without yielding product.

Similar structural/functional arguments support the proposed opening process (steps 5 and 6). In Figure 3.4 the LID bound ADP is released first (LID opening), resulting in the $LID_0^O - NMP_{ADP}^C$ state. Upon LID opening the interface between the NMP domain and the LID domain is lost, which destabilizes the closed NMP state. In addition to this destabilization, the intrinsic fluctuations of NMP increase in the direction of domain opening (Figure 3.3). The combination of NMP domain destabilization, increased motion in the opening direction and the chemical potential driving ADP out of the NMP-CORE pocket likely signal the NMP domain to open. These contributions ensure rapid release of the second ADP, allowing NMP to open without the $LID_{AMP}^C - NMP_{ADP}^C$ state forming.

Strain energy considerations from molecular dynamics simulations[8] and NMA suggest the alternate opening mechanism, namely NMP opening followed by LID opening, would also be more error prone. To see why, let's consider the two closed domains to be loaded springs that have been pulled from their equilibrium (open) states. MD simulations and NMA suggest there is a larger amount of strain build-up during NMP closure than during LID closure (approximately three times more; Figure 3.3 and Ref.[8]). Assuming one of the domains is open and the other is closed, the NMP domain's smaller mass and larger effective spring constant (? strain) make it open approximately twice as fast as the LID domain. Thus, when the NMP domain opens first there is twice as much time available for AKE to misligate (forming a $LID_{ATP}^C - NMP_{ADP}^C$ complex) than if the LID opens first

(forming a $LID^C_{ADP} - NMP^C_{AMP}$ complex).

3.5.3 Predictions for the AKE catalytic cycle

As described above, helix 4α and LID-NMP interface appear to play large roles in the catalytic activity of AKE. To probe these effects in experiments, activity assays of different mutants can be performed. Mutations that provide strong interactions between helix 4α and nearby residues may either maintain wild type efficiency (substrate turnover/domain rearrangement ≈ 1) or result in NMP domain motion becoming extremely rate limiting. In the latter case, the NMP domain will undergo a single transition per substrate turnover, even though the LID domain undergoes many closing/opening transitions.

On the other hand, mutations that destabilize helix 4α and the LID-NMP interface would allow NMP to close more easily, diminish the role of the LID-NMP interface and decouple the motion of the two domains. This decoupling would effectively poison the protein and substrate turnover would drop to $\frac{2}{\frac{1}{k_{LID}}+\frac{1}{k_{NMP}}}$, assuming no mis-ligation events.

3.5.4 Allosteric motion may be decomposed into intrinsic and ligand-gated contributions

Allosteric conformational transitions are becoming a prevalent theme in cellular signalling, occurring in a wide range of systems including kinases, G-proteins and ion channels. In the classical view of allostery conformational changes occur between static structures and are induced by ligand binding. Allosteric changes of this type can be considered "ligand gated" or "extrinsic" motion. In contrast, the current understanding of allostery assumes that protein structures are not static three dimensional structures, but are ensembles of conformations where ligands change the balance of pre-existing substates. In this perspective, the protein is constantly interconverting between multiple conformations and the conformational change could be considered "intrinsic". In this manuscript, we include the intrinsic and extrinsic descriptions in a single statistical picture and provide two structure-

based methods that quantitatively distinguish between them. Using AKE as a test protein, we have demonstrated that the LID domains conformational rearrangement can be captured by intrinsic motion and that the NMP domains motion is likely more dependent on ligands. As we target allosteric changes in therapeutics, this type of classification will allow for more sophisticated drug design.

3.6 Acknowledgements

We would like to thank Professor Peter Wolynes and Jeff Noel for valuable discussions, especially regarding non-linear NMA. We would additionally like to thank Pat Jennings for discussion about enzymatic kinetics. This work was funded by the NSF-sponsored Center for Theoretical Biological Physics (Grants PHY-0216576 and 0225630) and the NSF Grant 0543906. PCW is supported by the NIH Molecular Biophysics Training Grant at UCSD (Grant T32 GM08326). SG is supported by a Burroughs Wellcome Fund La Jolla Interfaces in Science fellowship.

Chapter 3, in full, appears in Journal of Biological Chemistry, 2008, Whitford, Gosavi, Onuchic. The dissertation author is the primary investigator and author of the paper.

Bibliography

[1] Lieser, S. A., Shaffer, J. & Adams J. A. *J. Biol. Chem.* **2007**, 281, 38004-38012.

[2] Wolf-Watz, M., Thai, V., Henzler-Wildman, K., Hadjipavou, G., Eisenmesser, E. Z. & Kern, D. *Nat. Struct. Biol.* **2004**, 11, 945-949.

[3] Eisenmesser, E., Millet, O., Labeikovsky, W., Korzhnev, D. M., Wolf-Watz, M., Bosco, D. A., Skalicky, J. J., Kay, L. E. & Kern, D. *Nature* **2005**, 438, 117-121.

[4] Mueller, C. W., Schlauderer, G. J., Reinstein, J. & Schulz, G. E. *Structure* **1996**, 4, 147-156.

[5] Mueller, C. W. & Schulz, G. E. *J. Mol. Biol.* **1992**, 224, 159-177.

[6] Miyashita, O., Onuchic, J. N. & Wolynes, P. G. *Proc. Nat. Acad. Sci. USA* **2003** 100, 12570-12575.

[7] Miyashita, O., Wolynes, P. G. & Onuchic, J. N. *J. Phys. Chem. B* **2005**, 109, 1959-1969.

[8] Whitford, P. C., Miyashita, O., Levy, Y. & Onuchic, J. N. *J. Mol. Biol.* **2007**, 366, 1661-1671.

[9] Krishnamurthy, H., Lou, H., Kimple, A., Vieille, C. & Cukier, R. *Proteins: Struct. Funct. Bioinfo.* **2005**, 58, 88-100.

[10] Lou, H. & Cukier, R. I. *J. Phys. Chem. B* **2006**, 110, 24121-24137.

[11] Snow, C., Qi, G. & Hayward, S. *Proteins. Struct. Func. Bioinfo.* **2007**, 67, 325-337.

[12] Temiz, N. A., Meirovitch, E. & Bahar, I. *Proteins: Struct. Funct. Bioinfo.* **2004**, 57, 468-480.

[13] Maragakis, P. & Karplus, M. *J. Mol. Biol.* **2005**, 352, 807-822.

[14] Bae, E. & Phillips, G. N. *J. Biol. Chem. 279* **2004**, 2820228208.

[15] Bae, E. & Phillips, G. N. *Proc. Nat. Acad. Sci. USA* **2006**, 103, 2132-2137.

[16] Sinev, M. A., Sineva, E.V., Ittah, V. & Haas, E. *Biochemistry* **1996**, 35, 6425-6437.

[17] Sanders, C. R., Tian, G. & Tsai, M. D. *Biochemisty* **1989**, 28, 9028-9043.

[18] Moritsugu, K. & Kidera, A. *J. Phys. Chem. B* **2004**, 108, 3890-3898.

[19] Horiuchi, T. & G?, N. *Proteins:Struct. Funct. Gen.* **1991**, 10, 106-116.

[20] Kong, Y., Ma, J., Karplus, M. & Lipscomb, W. *J. Mol. Biol.* **2006**, 356, 237-247.

[21] Gerstein, M., Lesk, A. M. & Chothia, C. *Biochemistry* **1994**, 33, 6739-6749.

[22] Bahar, I., Atilgan, A.R. & Erman, B. *Folding and Design* **1997**, 2, 173-181.

[23] Tama, F. & Sanejouand, Y. H. *Prot. Eng.* **2001**, 14, 1-6.

[24] Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. *Proteins* **1995**, 21, 167-195.

[25] Bryngelson, J. D. & Wolynes, P. G. *Proc. Nat. Acad. Sci. USA.* **1987**, 84, 7524-7528.

[26] Leopold, P. E., Montal, M. & Onuchic, J. N. *Proc. Nat. Acad. Sci. USA* **1992**, 18, 8721-8725.

[27] Onuchic, J. N. & Wolynes, P. G. *Curr. Opin. Struct. Biol.* **2004**, 14, 70-75.

[28] Clementi, C., Nymeyer, H. & Onuchic, J. N. *J. Mol. Biol.* **2000**, 298, 937-953.

[29] Nymeyer, H., Garcia, A.E., & Onuchic, J. N. *Proc. Nat. Acad. Sci. USA* **1998** 95, 5921-5928.

[30] Levy, Y., Cho, S. S., Shen, T., Onuchic, J. N. & Wolynes, P. G. *Proc. Nat. Acad. Sci. USA.* **2005**, 102, 2373-2378.

[31] Levy, Y., Cho, S. S., Onuchic, J. N. & Wolynes, P. G. *J. Mol. Biol* **2005**, 346, 1121-1145.

[32] Yang, S. C., Cho, S. S., Levy, Y., Cheung, M. S., Levine, H., Wolynes, P.G. & Onuchic, J.N. *Proc. Nat. Acad. Sci. USA* **2004**, 101, 13786-13791.

[33] Sobolev, V., Wade, R., Vried, G. & Edelman, M. *Proteins: Struct. Funct. Genet.* **1996**, 25, 120-129.

[34] Tirion, M. M. *Phys. Rev. Lett.* **1996**, 77, 1905-1908.

[35] McLachlan A. D. *Acta. Cryst. A.* **1982**, 38, 871-873.

[36] Fersht, A. *Curr. Opin. Struct. Biol.* **1994**, 4, 79-84.

[37] Schlauderer, G. J., Proba, K. & Schulz, G. E. *J. Mol. Biol.* **1996**, 256, 223-227.

[38] Diederichs, K. & Schulz, G. E. **J. Mol. Biol. 1991**, 217, 541-549.

Chapter 4

An All-atom Structure-Based Potential for Proteins: Bridging Minimal Models with All-atom Empirical Forcefields

4.1 Abstract

Protein dynamics take place on many time and length scales. Coarse-grained structure-based (Gō) models utilize the funneled energy landscape theory of protein folding to provide an understanding of both long time and long length scale dynamics. All-atom empirical forcefields with explicit solvent can elucidate our understanding of short time dynamics with high energetic and structural resolution. Thus, structure-based models with atomic details included can be used to bridge our understanding between these two approaches. We report on the robustness of folding mechanisms in one such all-atom model. Results for the B domain of Protein A, the SH3 domain of C-Src Kinase and Chymotrypsin Inhibitor 2 are reported. The interplay between side chain packing and backbone folding is explored. We also compare this model to a C_α structure-based model and an all-atom empirical forcefield. Key findings include 1) backbone collapse is accom-

panied by partial side chain packing in a cooperative transition and residual side chain packing occurs gradually with decreasing temperature 2) folding mechanisms are robust to variations of the energetic parameters 3) protein folding free energy barriers can be manipulated through parametric modifications 4) the global folding mechanisms in a C_α model and the all-atom model agree, although differences can be attributed to energetic heterogeneity in the all-atom model 5) proline residues have significant effects on folding mechanisms, independent of isomerization effects. Since this structure-based model has atomic resolution, this work lays the foundation for future studies to probe the contributions of specific energetic factors on protein folding and function.

4.2 Introduction

In recent years the energy landscape theory of protein folding [1, 2, 3, 4, 5] has been validated through its application to protein folding [6, 7, 8, 9, 10], oligomerization [11, 12, 13, 14], functional transitions [15, 16, 17, 18, 19, 20] and structure prediction [21, 22]. The theory states that proteins are minimally frustrated, that their energy landscape is funnel shaped and that the folded state of the protein is at the bottom of the funnel. Because of the shape of the landscape there is a strong energetic bias towards the folded state of the protein with relatively infrequent trapping caused by non-native interactions. The resulting heterogeneity observed during folding is due to the geometric constraints of the native structure. Thus, models of proteins that have only the native structure encoded have had great success in determining folding mechanisms. Until recently, most models tended to be coarse-grained, which are very useful in understanding global folding dynamics. In commonly used structure-based (Gō) potentials [9], each residue is represented by a bead centered at the location of the C_α atom (Figure 4.1b) and only native interactions are stabilizing.

On the other end of the spectrum of structural and energetic details are the computationally intensive all-atom empirical forcefields [25, 26, 27, 28, 29, 30]. These forcefields include an atomistic representation of a protein either with an

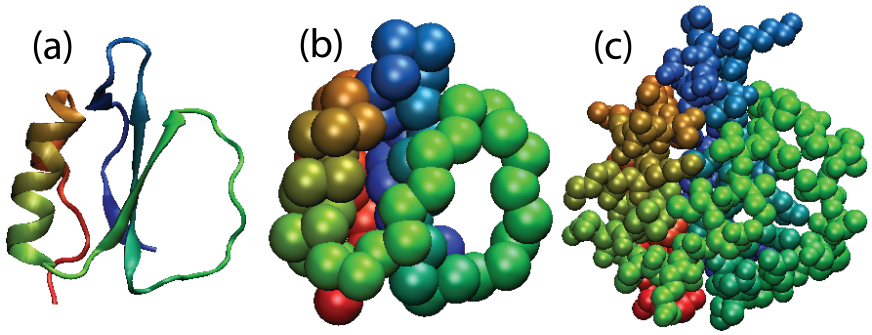

Figure 4.1: CI2 (Protein Data Base Entry 1YPA [23]) shown in (a) cartoon representation, (b) C_α representation and (c) all-atom (AA) representation. Structures are colored Red (C-terminus) to Blue (N-terminus). The size of the atoms in the C_α and AA representations correspond to the excluded volume radii used in the C_α [9] and AA models studied in this paper. Structures visualized using VMD [24].

implicit or an explicit solvent. In these potentials, the parameters which determine the interaction between atoms, such as partial charges and van der Waals radii, are fit to experimental measurements and quantum mechanical calculations. With accurate calibration, a single parameter set may be applied to any protein and with sufficient computing resources, the dynamics of a protein can be calculated on a computer. The physics-based representation of atom-atom interactions automatically includes electrostatic interactions as well as any non-native interactions that may be present. In principle, these models render knowledge of a native structure unnecessary. A major limitation of these potentials is that they are often too expensive to fold all but small proteins [31, 32, 33, 34, 35, 36, 37, 38, 39]. The timescales that can currently be calculated vary from hundreds of nanoseconds to microseconds, depending on the size of the protein. Biological timescales are usually several orders of magnitude larger and these dynamics cannot be accessed using all-atom empirical forcefields. In addition, sensitivity analysis of the dynamics to the parameters is not possible with these all-atom empirical forcefields.

In all-atom empirical forcefields an observed specificity of (i.e. preference for) native interactions is seen as a consequence of many energetic contributions. Due to the complex formulation of these potentials, it is impossible to partition geometric effects from energetic ones. There is a similar restriction in coarse-

grained models due to their simplicity. Partitioning these effects is often impossible since geometry is included implicitly through energetic interactions. By studying all-atom models with structure-based potentials [40, 41, 42, 43, 44], since atomic geometry is explicitly included, we can ask to what extent energetics contribute to the apparent native specificity in protein structure, folding and function. In contrast to enzyme catalysis where specific atomic interactions directly control the chemical reactions, in most cases the energetic specificity required in protein folding is less stringent.

Providing a complete picture of specificity in protein folding and function will require the study of many proteins and many parametric variations. In this manuscript, we lay the foundation for this line of investigation through systematic characterization of a completely specific (only, and all, native interactions are stabilizing) AA structure-based model. We study the effect of varying the parameters of the model on folding barriers, mechanisms, contact formation and side chain dynamics. The test proteins, B domain of Protein A, SH3 domain of C-Src Kinase and Chymotrypsin Inhibitor 2 (CI2) (Figure 4.2) have been experimentally [47, 48, 49] and computationally [8, 50, 51, 52] well characterized. Additionally, they possess two-state folding dynamics and represent different secondary and tertiary structures. The present model is energetically unfrustrated, with an explicit representation of all non-hydrogen atoms and homogeneous interaction strengths. We find that folding in the model is robust to parameter changes and dynamics agrees well with both the C_α model and an all-atom empirical forcefield with explicit solvent. Further, side chain ordering can be probed explicitly and the effect of prolines can be calculated. This study and model will serve as a basis for future AA models which incorporate non-specific contributions of energetic frustration, electrostatics and hydration.

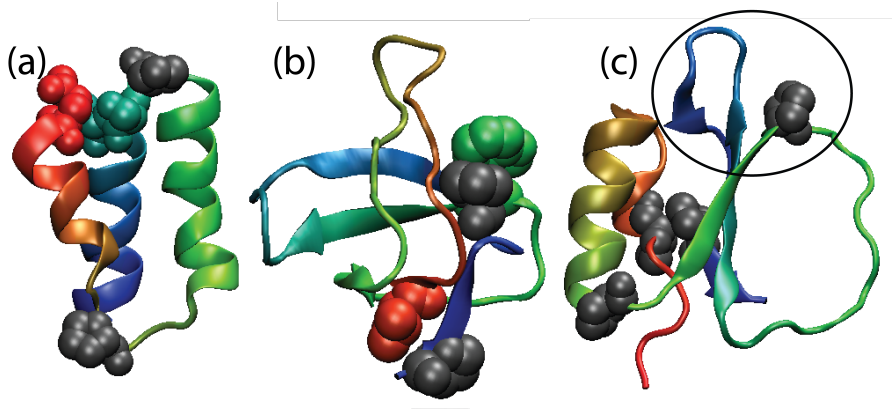

Figure 4.2: Structures of (a) Protein A, (b) SH3 and (c) CI2 (PDB entries 1BDD [45], 1FMK [46] and 1YPA [23]) colored Red (C-terminus) to Blue (N-terminus). These three proteins represent differing structural content and topological complexity. Protein A is a three-helix bundle, SH3 is composed of multiple β strands and in CI2 an alpha helix flanks a β sheet. Proline residues are shown as grey spheres. In Protein A, Gln1 and Ser31 are shown as colored spheres. In SH3, Val4 and Trp35 are shown as spheres. The mini-core of CI2 is circled.

4.3 Results

4.3.1 Folding mechanisms are robust to parameter changes

We employ a model where the potential energy function is defined by the native state and all heavy (non-hydrogen) atoms are explicitly represented. Any two atoms that are close in the native structure are said to form a native contact. We describe the folding process by using the fraction of native residue pairs in contact Q_{AA} (*see methods*). Figure 4.3a shows Q_{AA}, Q_{CA} (fraction of C_α contacts, *see methods*) and radius of gyration R_g as functions of time for an AA simulation of CI2, near folding temperature. Since Q_{AA} captures the same collapse events as R_g and Q_{CA} (Figure 4.3b), Q_{AA} is a useful measure of backbone folding in addition to side chain packing.

It is crucial to understand the parameter dependence of a model before it can be used to make reliable predictions of folding mechanisms. The robustness of the folding mechanism is probed here by characterizing Protein A, SH3 and CI2 for variants of the AA structure-based energy function. Due to the debate

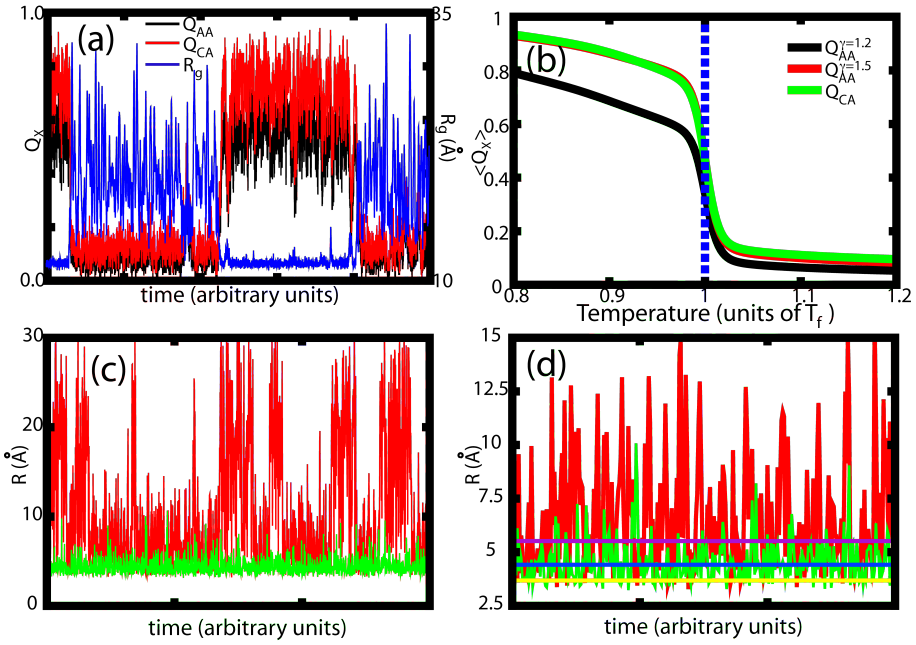

Figure 4.3: (a) Fraction of C_α contacts $Q_{CA}(t)$, AA contacts $Q_{AA}(t)$ and Radius of Gyration $R_g(t)$ as functions of time for a representative trajectory of CI2 with the AA model. (b) Average structure formation for several reaction coordinates. A contact between residues is formed when a single atom-atom contact between them is formed. An atom-atom contact is considered formed when the pair is at a distance $r < \gamma\sigma$ where σ is the native pair distance. The fraction of native residue contacts formed Q_{AA}^X is shown for γ = 1.2 (black) and γ = 1.5 (red). A C_α contact is formed when the C_α atoms are within 1.2 times their native distance (green). All three coordinates capture the same folding events. (c) Atom-atom distance for a contact in the active loop of CI2 versus time at T_f (red) and $T < T_f$ (green). Large changes in distance ($> 20\mathring{A}$) coincide with folding transitions. Side chain rearrangements in the folded state ($R < 10\mathring{A}$) occur on much faster time scales than folding of the entire protein. (d) Same as Figure (c) with time scale decreased by a factor of 100. Horizontal lines correspond to σ (yellow), 1.2σ (blue) and 1.5σ (purple). As temperature is decreased, distance fluctuations and average distances decrease.

about the balance between secondary and tertiary interactions, we vary the ratio of non-local contact energy to dihedral angles $R_{C/D}$ and the relative strength of backbone dihedral angles to side chain dihedral angles $R_{BB/SC}$ (see *methods*).

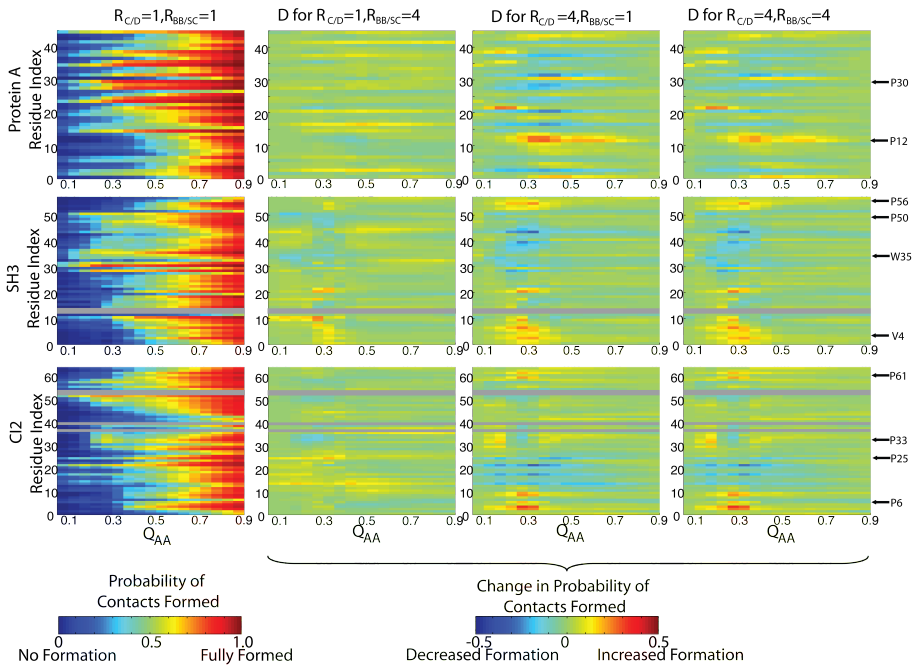

Figure 4.4: The left column shows the probability of contacts being formed for each residue $P(Q_i, Q_{AA})$ as a function of Q_{AA} for $R_{C/D}$ = 1.0 and $R_{BB/SC}$ = 1.0. The three right columns show $P(Q_i, Q_{AA})$ for different Hamiltonians relative to $R_{C/D}$ = 1.0 and $R_{BB/SC}$ = 1.0. Blue indicates a decrease in formation, relative to $R_{C/D}$ = 1.0 and $R_{BB/SC}$ = 1.0, and red an increase. Proline containing regions are often sensitive to contact energy. In Protein A, both P12 and P30 fold earlier with increased contact strength. In SH3, the increase in formation of Val4 may be attributed to interactions with Pro56, though Pro50 and Trp35 do not exhibit increased formation. In CI2, both Pro6 and Pro61 exhibit increased formation with increased contact strength. Residues that lack native contacts are shown in grey.

To characterize the folding mechanism for different parameter sets we computed the probability of contacts formed as a function of the folding process $P(Q_i, Q_{AA})$. $P(Q_i, Q_{AA})$ is the probability that the set of contacts involving residue i, Q_i, are formed as a function of Q_{AA}. $P(Q_i, Q_{AA})$ was calculated for the three proteins for 16 different parameter sets (all combinations of $R_{C/D}$ = $1.0, 2.0, 3.0, 4.0$ and $R_{BB/SC}$ = $1.0, 2.0, 3.0, 4.0$). Figure 4.4 shows the folding mech-

anisms for four parameter sets. The difference in folding mechanism between parameter sets i and j can be quantified by the root mean squared deviation in $P(Q_i, Q_{AA})$ over all Q_{AA} and Q_i, ($P_{rms} = \sqrt{\langle (P_i(Q_i, Q_{AA}) - P_j(Q_i, Q_{AA}))^2 \rangle}$). The largest values of P_{rms} for Protein A, SH3 and CI2 were 0.057, 0.097 0.077. SH3 is a complicated fold, Protein A a simple fold and CI2 an intermediate fold [53]. Thus, it is not surprising that energetic modifications have the largest effects on Protein A and the smallest effects on SH3.

Figure 4.4 shows proline containing regions are less stable to parametric modifications. Regions with prolines, and regions interacting with prolines, form structure earlier (at lower Q) with increased contact strength. This is because contact strength is increased at the expense of dihedral strength. Prolines possess a covalent $C_\delta - N$ bond, which limits the mobility of the ϕ dihedral. Removing energy from the dihedrals does not increase flexibility in prolines. However, adding energy to contacts increases structure formation around prolines. For this reason, increasing $R_{C/D}$ stabilizes and promotes earlier formation of proline containing regions.

4.3.2 Fully folded backbone allows for disordered side chains

While Q_{AA} and Q_{CA} capture the same cooperative folding events, at folding temperature, Q_{CA} is higher than Q_{AA} for the folded ensemble. This suggests that while the backbone structure is native ($Q_{CA} \approx 0.8$), many of the native residue interactions form as temperature is decreased (Figure 4.3c and d). To account for this structurally and quantitatively, we calculated the difference between the probability of C_α contacts being formed $P(Q^i_{C_\alpha}, Q_{C_\alpha})$ and AA contacts being formed $P(Q^i_{AA}, Q_{C_\alpha})$ (Figure 4.5). A value of 0 indicates that, on average, the C_α atoms of a residue pair are near their native distance when the side chains are in contact. Positive values are seen when extended side chains are interacting, resulting in the C_α atoms being far from their native distance. Negative values indicate backbone folding precedes side chain ordering[1]. Side chains in Protein A appear

[1] Q_{AA} is a generous definition of side chain packing, since a side chain is "packed" when one or more atom-atom contacts are formed. Thus, "underpacked" residues clearly have very little native structure.

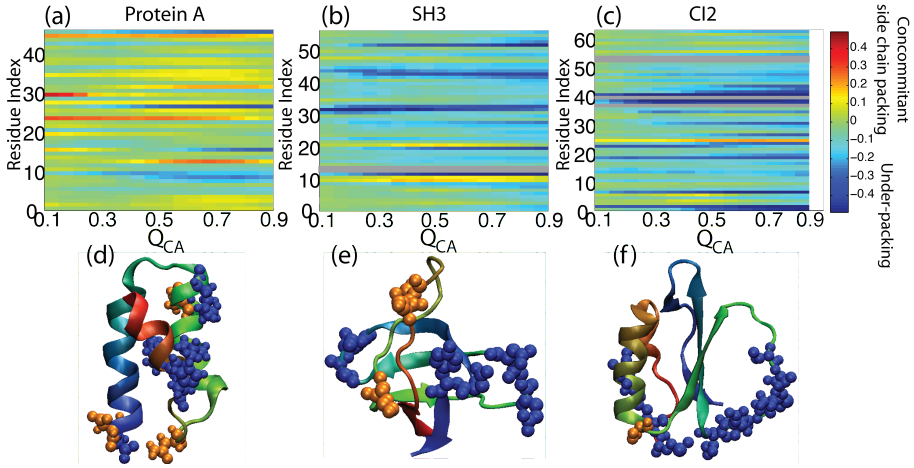

Figure 4.5: Difference in AA contact formation and C_α contact formation $P(Q^i_{AA}, Q_{C_\alpha}) - P(Q^i_{C_\alpha}, Q_{C_\alpha})$ for (a) Protein A , (b) SH3 and (c) CI2. Positive values (red) indicate that residues are interacting without the C_α atoms being near. Negative values (blue) indicate the residues are "underpacked": the C_α atoms are near each other without the side chains interacting. Residues that lack native contacts are shown in grey. (d-f) Underpacked (blue spheres) and well packed (orange spheres) residues are shown on the native structures. In Protein A, to order the backbone of a helix the side chains must be packed around it. Beta sheets are stabilized by non-local interactions. Thus, a small number of contacts can maintain the tertiary structure of SH3 without the side chains in the turn regions interacting, hence the underpacking. In CI2, the active site loop is significantly underpacked.

to be well-packed, in that there is concomitant side chain and backbone folding. In SH3, the turns have negative values, and are thus underpacked. In CI2, underpacking is primarily found in the active site loop and the C-terminal tail. These results reveal a signature of complicated folds [52, 53]: a small subset of native contacts is sufficient to constrain the backbone to its native orientation, resulting in significantly underpacked regions in the native state. This occurs in complicated folds because an individual contact can impose a high level of order on the system. In order to form contacts that are distant in sequence a large number of residues must also order. In Protein A, many contacts are local and only constrain single helical turns. In SH3 and CI2, fewer contacts are required to constrain the entire backbone (including the turns and loops).

Figures 4.3c and 4.3d show the dynamics of a typical underpacked contact.

As T is lowered below T_f the underpacked contact's average distance and distance fluctuations smoothly decrease. This results in a gradual increase in Q without a noticeable free energy barrier (See Figure 4.6e). We hope that these subtle dynamics will be experimentally probed and tested in the future.

4.3.3 Understanding free energy profiles through parametric variation: Free energy profiles can be altered through parametric changes

While the folding mechanisms are stable, the free energy barriers associated with folding and the locations of the folded basins vary systematically with parameters. Figure 4.6 shows free energy profiles for SH3, CI2 and Protein A for several values of $R_{C/D}$ with $R_{BB/SC}$ = 2.0. There are four distinct, interrelated, trends shared by all three proteins. First, there are two folding processes: backbone collapse and side chain packing. Second, the free energy minimum for the folded state moves to lower Q with increasing $R_{C/D}$. Third, the free energy barrier decreases with increasing $R_{C/D}$. Finally, increasing $R_{BB/SC}$ has similar effects as increasing $R_{C/D}$ (not shown).

The free energy basins for the folded states are located at $Q_{CA} \approx 0.8$ and $Q_{AA} \approx 0.5$ (Figure 4.6d), indicating that the backbone orders while many native atom-atom interactions remain extended. Thus, the entropy loss during the cooperative folding transition is likely dominated by backbone ordering. Side chain packing occurs both concomitantly with, and after, backbone ordering.

There are likely two major factors that lead to the observed trends. First, increasing $R_{C/D}$ increases contact strength. As seen in other simplified models[54], when each contact is stronger, a smaller number of contacts is required (lower Q) to provide an equal amount of stabilizing energy. The second contributing factor is the change in side chain entropy. While entropy loss in the backbone dominates the collapse transition, the gradual side chain packing can also lead to shifting basins. Increasing $R_{BB/SC}$ or $R_{C/D}$ reduces the strength of side chain dihedrals, resulting in more mobile unfolded side chains. Therefore, there is an increased entropy loss

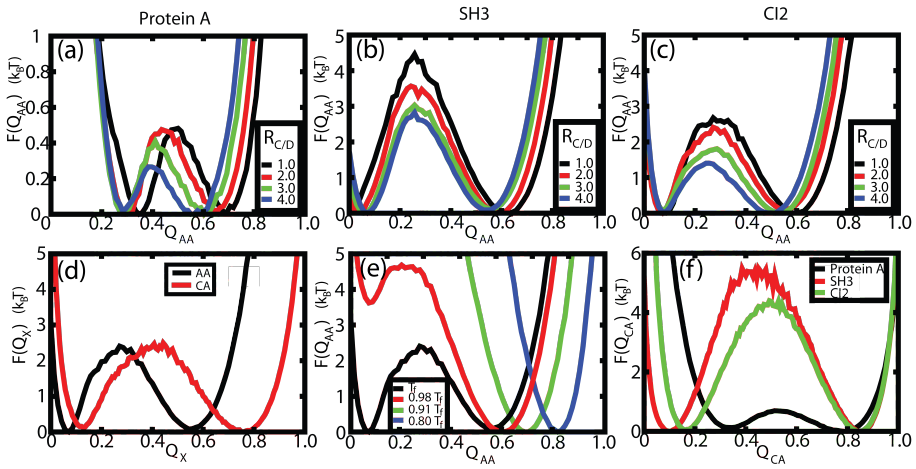

Figure 4.6: Free energy barriers in the AA model for (a) Protein A , (b) SH3 and (c) CI2. Profiles in (a-c) are for $R_{BB/SC}$ = 2.0 with $R_{C/D}$ = 1.0 (black), $R_{C/D}$ = 2.0 (red), $R_{C/D}$ = 3.0 (green) and $R_{C/D}$ = 4.0 (blue). In SH3 and CI2, barrier height decreases and the folded basins move to lower Q with increasing $R_{C/D}$ and increasing $R_{BB/SC}$. (d) $F(Q_{CA}(t))$ and $F(Q_{AA}(T))$ for a typical parameter set demonstrate that the folded basins in (a-c) correspond to collapsed states. (e) Two distinct folding processes observed in our model: backbone collapse and side chain packing. (f) Free energy barriers obtained from C_α structure-based simulations for Protein A, SH3 and CI2. Barrier heights in the C_α simulations are greater than in AA simulations. Both models predict the largest barriers for SH3 and smallest for Protein A.

per side chain upon folding ΔS_{sc} when $R_{C/D}$ or $R_{BB/SC}$ is increased. Since side chains can pack independently of the collapse transition, when ΔS_{sc} increases, a fraction of the side chain interactions extend, while leaving the overall fold intact. Since the folded basin shifts to lower Q, the overall structure required to form a stable fold is reduced. A reduced barrier height naturally results when the folded basin is less ordered.

Free energy barriers, in conjunction with diffusion constants, provide a direct connection to experimental folding rates [55, 56, 57]. We find that the relative barrier heights calculated using our AA model are similar to those from a C_α model (Fig. 4.6f). The relative barrier heights calculated from this model are known to correlate well with experimental rates [57]. We note in passing, that the absolute free energy barriers in the AA model can be parametrically changed by up to a factor of two for a given protein and that the relative barrier heights between proteins remain constant. Thus, while the magnitude of the rates will be determined by the diffusion constant, the correlation between experimental folding rates and theoretical barriers is independent of the choice of parameters.

4.3.4 All-atom structure-based simulations capture C_α folding mechanism

Next we compare the backbone folding mechanisms of our AA model and a commonly used C_α model [9]. The C_α representation has been successful at capturing experimentally determined protein folding mechanisms [8, 9, 11]. The first column in Figure 4.7 shows the differences in folding mechanisms between the AA model and an energetically homogeneous C_α model. Every contact and dihedral in the homogeneous C_α model has the same interaction strength. Since the AA model distributes contact energy inhomogeneously between residue pairs, it is not surprising that the mechanisms differ.

To remove differences arising from energetic homogeneity in the C_α model, we modified it such that each contact is weighted by the number of contacts between each residue pair in the AA model (Figure 4.7, second column). For Protein A this modification improves agreement. The remaining difference is in a single

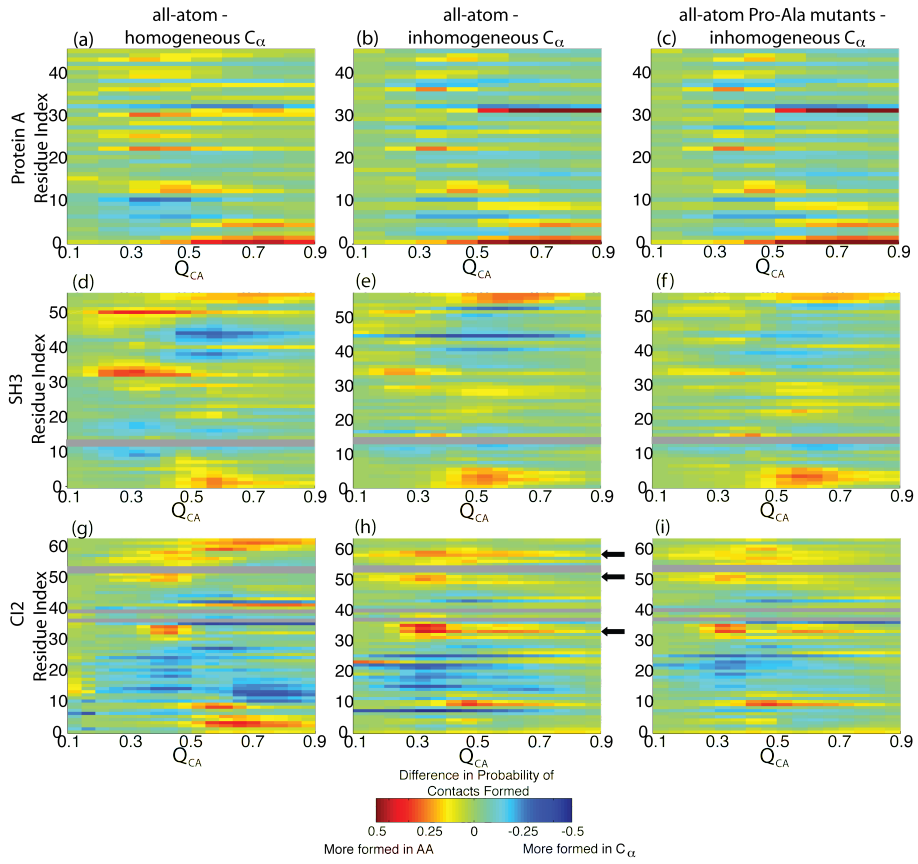

Figure 4.7: Comparison of backbone folding in C_α and AA structure-based models. The probability of contacts being formed in a C_α model, minus the probability of C_α contacts being formed in an AA model, is shown for (a-c) Protein A, (d-f) SH3 and (g-i) CI2. (a, d, g) Comparison of AA simulation to a C_α model with homogenous contact strength. (b, e, h) Comparison between AA results to an energetically inhomogeneous C_α model. Regions of increased formation in the AA representation correspond largely to proline containing regions, or regions that interact with proline, such as the minicore in CI2 (black arrows indicate mini-core residues), the tails of SH3 and turn 2 of Protein A. Increased formation in the tails of CI2 can largely be accounted for by the large number of contacts between GLU4 and ARG62. (c, f, i) The inhomogeneous C_α model compared to the AA model with all prolines mutated to alanines. Mutating proline to alanine improved agreement between models. Residues that lack native contacts are shown in grey.

turn-to-tail contact (Gln1 with Ser31, Fig. 4.2a) that rarely forms in C_α simulations. In SH3, agreement improves around residues Asp34 and Asn52, while differences persist in Gln45 and the tails. The overall effect is increased formation around Gln45 at the expense of the tails. In CI2, there is significant agreement in the tails, though the mini-core still forms earlier (in the AA model), at the expense of the helix. For all three proteins, several regions of disagreement possess proline residues, whose $C_\delta - N$ bond is not included in the C_α model.

To eliminate effects specific to proline, we repeated the AA simulations with all prolines mutated to alanines. The third column of Figure 4.7 shows the Pro-Ala mutants compared to the inhomogeneous C_α model. Improved agreement is observed in Pro-Ala mutants of SH3 and CI2. In both proteins Pro-Ala mutations delay folding of proline regions, in agreement with proline effects on model stability. In SH3 the tails still form slightly earlier in the AA model, at the expense of residues 35-55. In CI2, the balance between minicore and helix formation is clearly improved, highlighting the importance of prolines in the folding process. Pro-Ala mutations have almost no effect on the folding mechanism of P12 and P30 in Protein A and P25 in CI2. This is likely because these prolines are located in turn regions. In our model, turns are highly constrained by short range contacts, and the reduced dihedral constraint (imposed by a proline) acts as a small perturbation. The remaining differences between the Pro-Ala AA mutants and the inhomogeneous C_α model demonstrate, to no surprise, that the inclusion of side chains alters the relative entropy of residues.

4.3.5 Native basin dynamics of AA structure-based model correlate with the dynamics of an all-atom empirical forcefield with explicit solvent

Two common measures of native state dynamics are native contact formation and root mean squared deviations in structure *rmsd*. Figure 4.8 shows the average contact formation in the native ensemble for the structure-based model and an all-atom empirical forcefield with and explicit solvent. While the average

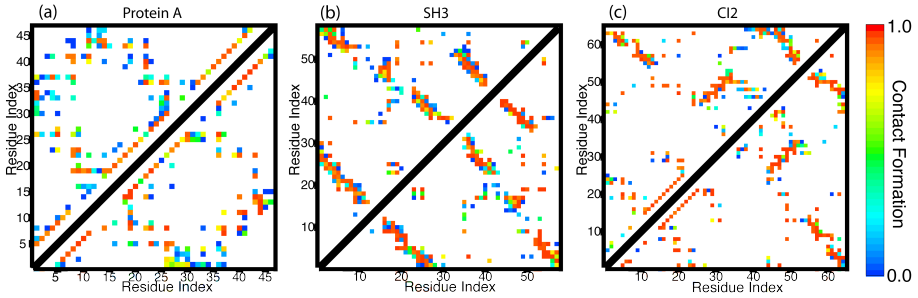

Figure 4.8: Probability of contacts being formed $P(i,j)$ at $T \approx 0.8T_f$ for the AA structure-based potential (top left) and an all-atom empirical forcefield (bottom right) for (a) Protein A, (b) SH3 and (c) CI2. Dark red indicates that residue i (x axis) and residue j (Y axis) are always in contact under native conditions. Dark blue indicates the contact is formed rarely (less than 10% of the time). White indicates $P(i,j) < 0.025$. In all three proteins, contacts are more broadly distributed (higher number of low probability contacts) in the structure-based simulations than in all-atom empirical forcefield simulations (fewer contacts, but with higher probabilities). There are approximately four times as many contacts with $P(i,j) < 0.01$ for the structure-based simulations than are seen in all-atom empirical simulations, indicating more mobile dynamics.

contacts are not identical, no major differences in contact formation are observed. The overlaps between the AA maps and the all-atom empirical forcefield maps of Protein A, SH3 and CI2 are 0.85, 0.97 and 0.84. An overlap of 1 indicates identical maps, and 0 indicates the two maps have no contacts in common.

In a uniquely defined native state, the probability of each contact being formed is 1. Since we sample the native ensemble at finite temperatures, atom mobility leads to additional contacts being formed. In the structure-based model, these additional interactions are strictly repulsive. In an all-atom empirical forcefield these interactions can be attractive, yet they are observed more frequently in the structure-based model[2]. These contacts are likely due to increased mobility in the structure-based simulations. In all-atom empirical forcefields, hydration shells can result in less mobile side chains, and hence a narrower distribution of contacts.

The increased mobility is quantified by the structural *rmsd*. The magnitude of fluctuations in all-atom empirical simulations is much lower than in structure-based simulations (not shown). For the all-atom empirical forcefield at

[2]In Figure 4.8 only interactions present more than 2.5% of the time are shown.

300 K, the average *rmsd* for Protein A, SH3 and CI2 are 1.53, 1.00 and 0.97 Å. The *rmsd* of the C_α atoms are 1.23, 0.66 and 0.74. The same values are obtained in structure-based simulations at around $T = 0.55T_f$. In real temperature units, $0.55T_f$ corresponds to temperatures significantly less than 300 K. A likely cause for the increased structural fluctuations is hydration effects of explicit solvent molecules in the all-atom empirical forcefield. To compare the distribution of *rmsd* fluctuations between models, correlation coefficients (r) were computed for the *rmsd* by atom in the all-atom empirical forcefield and the structure-based potential. For all parameter sets of the structure-based potential, the $r \approx 0.7$ for CI2 and SH3 and $r \approx 0.8$ for Protein A[3].

4.4 Discussion

In this manuscript, we describe a systematic analysis of an AA structure-based model which bridges the gap between coarse-grained models and all-atom empirical forcefields. We show that in our C_α and AA structure-based models the global folding mechanisms agree and the main differences are largely due to energetic heterogeneity and the explicit representation of prolines in the AA model. Also, the native basin dynamics are similar in the AA structure-based model and an all-atom empirical forcefield with explicit solvent. In agreement with previous studies, the folding mechanisms in complicated folds are stable to parametric variation. On the other hand, the free energy barriers associated with folding vary systematically with parameters. Since free energy barriers are not a robust feature of this model, understanding the interplay between barrier heights and diffusion will be important before attempting to predict folding rates [55, 58, 59].

Using this model we characterized two folding processes: one associated with backbone collapse and the other with side chain packing. We observed that backbone collapse is accompanied by partial side chain packing in a cooperative transition and residual side chain packing occurs as temperature is reduced below the global folding temperature. One explanation for the partial separation of

[3]Comparison of *rmsd* of the C_α atoms yields similar values of r.

backbone folding and side chain ordering may be that mobility in specific residues is necessary for the functional properties of proteins. Proteins are selected for their function. Orthogonal networks of residues responsible for stability and function have been proposed [60, 61]. The observation in our model that some residues are not necessary to maintain the backbone structure is consistent with this proposal. In CI2, the backbone of the active site loop is in the native orientation, yet the side chains are not packed. In SH3, several turns are also disordered. Since binding sites are often found in loops, flexible loops may be more easily adapted to new sequences and functions.

Gradual side chain packing can also allow for proteins to functionally respond to cellular stress by affecting side chain orientations, without denaturing the entire protein. This is consistent with the prediction that localized unfolding, or cracking, is important for biological function of kinases and motor proteins [15, 18, 62, 63, 64, 65, 66, 67].

The current model explicitly includes the effects of topological contributions to protein folding, and the role of energetic contributions may now be elucidated. Our results are a significant step forward in understanding protein dynamics from the C_α to the all-atom level. In the coming years, it will be interesting to probe the effects of electrostatics, non-native interactions, water and explicit mutations in this model.

4.5 Models and methods

4.5.1 Energy function

In our AA model of the protein, only heavy (non-hydrogen) atoms are included. Each atom is represented as a single bead of unit mass. Bond lengths, bond angles, improper dihedrals and planar dihedrals are maintained by harmonic potentials. Non-bonded atom pairs that are in contact in the native state between residues i and j, where $i > j + 3$, are given a Lennard-Jones potential, while all other non-local interactions are repulsive. All contacts identified by the Contact of Structural Units software package (CSU)[68] were included. The functional form

of the potential is,

$$V = \sum_{bonds} \epsilon_r (r - r_o)^2 + \sum_{angles} \epsilon_\theta (\theta - \theta_o)^2 + \sum_{impropers/planar} \epsilon_\chi (\chi - \chi_o)^2 + \sum_{backbone} \epsilon_{BB} F_D(\phi) + \sum_{sidechains} \epsilon_{SC} F_D(\phi) + \sum_{contacts} \epsilon_C [(\frac{\sigma_{ij}}{r})^{12} - 2(\frac{\sigma_{ij}}{r})^6] + \sum_{non-contacts} \epsilon_{NC} (\frac{\sigma_{NC}}{r})^{12} \quad (4.1)$$

where,

$$F_D(\phi) = [1 - cos(\phi - \phi_o)] + \frac{1}{2}[1 - cos(3(\phi - \phi_o))] \quad (4.2)$$

and $\epsilon_r = 100$, $\epsilon_\theta = 20$, $\epsilon_\chi = 10$ and $\epsilon_{NC} = 0.01$. r_o, θ_o, χ_o, ϕ_o and σ_{ij} are given the values found in the native state and $\sigma_{NC} = 2.5\mathring{A}$. When assigning dihedral strengths, we first group dihedral angles that share the middle two atoms. For example, in a protein backbone, one can define up to four dihedral angles that possess the same $C - C_\alpha$ covalent bond as the central bond. Each dihedral is given the interaction strength of $1/N_D$, where N_D is the number of dihedral angles in the group. ϵ_{BB} and ϵ_{SC} are then scaled so that $R_{BB/SC} = \frac{\epsilon_{BB}}{\epsilon_{SC}}$. Next, dihedral strengths and contact strengths are scaled such that our other system parameter, the ratio of total contact energy to total dihedral energy $R_{C/D} = \frac{\sum \epsilon_C}{\sum \epsilon_{BB} + \sum \epsilon_{SC}}$, is satisfied. The total stabilizing energy is equal for all parameter sets (i. e. $\sum \epsilon_C + \sum \epsilon_{BB} + \sum \epsilon_{SC}$=Constant).

As a reaction coordinate we use Q_{AA} and Q_{CA}. Q_{AA} is the fraction of natively interacting residues that are in contact. Two residues are considered in contact if any native atom-atom interactions between the residues are within 1.2 times the native distance σ_{ij}. At $1.2\sigma_{ij}$ the potential energy of a native pair is approximately half of the minimum. Similarly, Q_{CA} is the fraction of natively interacting residue pairs whose C_α atoms are within 1.2 times their native distance.

4.5.2 Proline to Alanine mutations

To investigate the role of proline residues in the AA model, proline to alanine mutants were constructed. This was achieved by removing the C_γ and C_δ atoms of each proline. Native contacts formed with the C_γ and C_δ of a proline

were included as contacts with the C_β of the corresponding alanine. This ensured the energetics of the system were unperturbed, and only topology was modified.

4.5.3 Simulation details

All-atom structure-based simulations were performed using the GROMACS software package [26]. No modifications to the source code were necessary. Reduced units were used. The timestep τ was 0.0005. The Berendsen algorithm [69] was used[4] with the coupling constant of 1. For all folding results in this paper, several constant temperature runs were performed, with temperatures that corresponded to the protein being always folded to always unfolded. The Weighted Histogram Analysis Method [70, 71] was used to combine data from multiple temperatures into single free energy profiles.

4.5.4 All-atom empirical forcefield simulations

All-atom empirical forcefield simulations were performed using GROMACS [26, 72], with the OPLS-AA forcefield [73] with TIP3P water molecules [74]. Each protein was simulated for 10 ns at T=300K and a pressure of 1 atm. A timestep of 2 fs was used in conjunction with the LINCS [75, 76] algorithm for constraining covalent bonds with hydrogen. Protein A, SH3 and CI2 were simulated with 2810, 3617 and 4644 water molecules in cubic boxes of initial dimensions 45.15 Å, 48.98 Åand 53.07 Å. Temperature was maintained using the Berendsen algorithm [69]. 1 ns was allowed for equilibration. For the remaining 9 ns, structures were saved at 1 ps intervals.

[4]When using the Berendsen thermostat, numerical instabilities can arise when the bath-molecule coupling timescale is shorter than the timescale for internal energy diffusion. In our experience, these problems tend to surface when you simulate weakly interacting domains with implicit solvation. Since the present study investigates folding of single domain proteins under weak temperature coupling, these features are not likely a source of significant errors. Nonetheless, future work will also employ Langevin or Nose-hoover temperature coupling.

4.5.5 Comparison of contacts

In the all-atom empirical forcefield simulations contacts were determined for each saved structure using CSU [68]. The average number of contacts $< Q >$ was calculated for each protein. The probability of individual contacts being formed was averaged over all structures with $Q =< Q >$. With the all-atom empirical potential $< Q >$ was 80, 135 and 146 for Protein A, SH3 and CI2. This analysis was repeated for folded simulations with our AA structure-based simulations. For the structure-based simulations $< Q >$ was 80, 138 and 144. To compare contact maps, the dot product of the two maps was taken.

4.6 Acknowledgements

We would like to thank Angel Garcia and Peter G. Wolynes for useful discussions regarding all-atom modeling. PCW and JKN were supported in part by the National Institutes of Health Molecular Biophysics Training Program at University of California at San Diego Grant T32 GM08326. This work was supported in part by Grants PHY-0216576 and 0225630 from the National Science Foundation (NSF)-sponsored Center for Theoretical Biological Physics, NSF Grant 0543906, the LANL LDRD program and NIH Grant R01-GM072686.

Chapter 4, in full, appears in Proteins: Structure, Funtion, Bioinformatics, 2009, Whitford, Noel, Gosavi, Schug, Sanbonmatsu, Onuchic. The dissertation author is the primary investigator and author of the paper.

Bibliography

[1] Leopold PE, Montal M & Onuchic JN. *Proc. Nat. Acad. Sci. USA* **1992**,18, 8721-8725.

[2] Bryngelson JD, Onuchic JN, Socci ND & Wolynes PG. *Proteins* **1995**, 21, 167-195.

[3] Bryngelson JD & Wolynes PG. *Proc. Nat. Acad. Sci. USA.* **1987**,84,7524-7528.

[4] Onuchic JN & Wolynes PG. *Curr. Opin. Struct. Biol.* **2004**,14,70-75.

[5] Ueda Y, Taketomi H & Gō N. *Int. J. Pept. Res.* **1975**,7,445-459.

[6] Shoemaker BA, Wang J & Wolynes PG. *Proc. Nat. Acad. Sci. USA* **1997**,94,777-782.

[7] Nymeyer H, Garcia AE & Onuchic JN. *Proc. Nat. Acad. Sci. USA* **1998**,95,5921-5928.

[8] Clementi C., Jennings PA & Onuchic JN. *J. Mol. Biol.* **2001**,311, 879-890.

[9] Clementi C, Nymeyer H & Onuchic JN. *J. Mol. Biol.* **2000**,298,937-953.

[10] Gosavi S, Chavez LL, Jennings PA & Onuchic JN. *J. Mol. Biol.* **2006**,357,986-996.

[11] Levy Y & Onuchic JN. *Acc. Chem. Res.* **2006**,39,135-142.

[12] Levy Y, Cho SS, Shen T, Onuchic JN & Wolynes PG. *Proc. Nat. Acad. Sci. USA.* **2005**,102,2373-2378.

[13] Levy Y, Cho SS, Onuchic JN & Wolynes PG. *J. Mol. Biol* **2005**,346,1121-1145.

[14] Yang SC, Cho SS, Levy Y, Cheung MS, Levine H, Wolynes PG & Onuchic JN. *Proc. Nat. Acad. Sci. USA* **2004**,101,13786-13791.

[15] Whitford PC, Miyashita O, Levy Y & Onuchic JN. *J. Mol. Biol.* **2007**,366,1661-1671.

[16] Whitford PC, Gosavi S & Onuchic JN. *J. Biol. Chem.* **2008**,283,2042-2048.

[17] Schug A, Whitford PC, Levy Y & Onuchic JN. *Proc. Nat. Acad. Sci. USA* **2007**,104,17674-17679.

[18] Okazaki K, Koga N, Takada S, Onuchic JN & Wolynes PG. *Proc. Nat. Acad. Sci. USA* **2006**,103,11844-11849.

[19] Best RB, Chen Y & Hummer G. *Structure* **2005**,13,1755-1763.

[20] Zuckerman, D.M. *J. Phys. Chem. B* **2004**,108,5127-5137.

[21] Eastwood MP, Hardin C, Luthey-Schulten Z & Wolynes PG. *IBM J. Res. Dev.* **2001**,5,475-497.

[22] Rohl CA, Strauss CEM, Misura KMS & Baker D. *Methods Enzymol.* **2004**,383,66-93.

[23] Harpaz Y, Elmasry N, Fersht AR & Henrick K. *Proc. Nat. Acad. Sci. USA* **1994**,91,311-315.

[24] Humphrey W, Dalke A & Schulten K. *J. Molec. Graphics* **1996**,14,33-38.

[25] Adcock SA & McCammon JA. *Chem. Revs.* **2006**,106,1589-1615.

[26] Lindahl E, Hess B & van der Spoel D. *J. Mol. Mod.* **2001**,7,306-317.

[27] Ponder JW & Case DA. *Adv. Prot. Chem.* **2003**,66,27-85.

[28] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L & Schulten K. *J. Comp. Chem.* **2005**,26,1781-1802.

[29] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S & Karplus M. *J. Comp. Chem.* **1983**,4,187-217.

[30] Eisenberg D & McLachlan AD. *Nature* **1986**,319,199-203.

[31] Zhou R. *Proc. Nat. Acad. Sci. USA* **2003**,100,13280-13285.

[32] Paschek D, Nymeyer H & Garcia A. *J. Struct. Biol* **2007**,157,524-533.

[33] Duan Y & Kollman PA. *Science* **1998**,282,740-744.

[34] Garcia A & Onuchic JN. *Proc. Nat. Acad. Sci. USA* **2003**,100,13898-13903.

[35] Schug A, Herges T & Wenzel W. *Phys. Rev. Letters* **2003**,91,158102.

[36] Schug A, Verma A, Herges T, Lee KH & Wenzel W. *ChemPhysChem* **2005**,6,2640-2646.

[37] Schug A & Wenzel W. *Biophys. J.* **2006**,90,4273-4280.

[38] Jayachandran G, Vishal V & Pande VS. *J. Chem. Phys.* **2006**,124,164902.

[39] Freddolino PL, Liu F, Gruebele M & Schulten K. *Biophys. J.* **2008**,94,L75-L77.

[40] Clementi C, Garcia AE & Onuchic JN *J. Mol. Biol.* **2003**,326,933-954.

[41] Shimada J, Kussell EL & Shakhnovich EI. *J. Mol. Biol.* **2001**,308,79-95.

[42] Linhananta A & Zhou Y. *J. Chem. Phys.* **2002**,,8983-8995.

[43] Zhou Y, Zhang C, Stell G & Wang J. *J. Amer. Chem. Soc.* **2003**,125,6300-6305.

[44] Linhananta A, Boer J & Mackay I. *J. Chem. Phys.* **2005**,122,114901.

[45] Gouda H, Torigoe H, Saito A, Sato M, Arata Y & Shimada I. *Biochemistry* **1992**,31,9665-9672.

[46] Xu W, Harrison SC & Eck MJ. *Nature* **1997**,385,595-602.

[47] Sato S, Religa TL, Daggett V & Fersht AR. *Proc. Nat. Acad. Sci. USA* **2004**,101,6952-6956.

[48] Viguera AR, Martinez JC, Filimonov VV, Mateo PL & Serrano L *Biochemistry* **1994**,33,2142-2150.

[49] Jackson SE & Fersht AR. *Biochemistry* **1991**,30,10428-10435.

[50] Shea JE, Onuchic JN & Brooks CL III. *Proc. Nat. Acad. Sci. USA* **2002**,99,16064-16068.

[51] Hoang TX. & Cieplek M. *J. Chem. Phys.* **2000**,113,8319-8328.

[52] Shea JE, Onuchic JN & Brooks CL III. *Proc. Nat. Acad. Sci. USA* **1999**,96,12512-12517.

[53] Plaxco KW, Simonsa KT & Baker D. *J. Mol. Biol.* **1998**,277,985-994.

[54] Prieto L & Rey A. *J. Chem. Phys.* **2007**,126,165103.

[55] Bryngelson JD & Wolynes PG. *J. Phys. Chem.* **1989**,93,6902-6912.

[56] Kramers HA. *Physica* **1940**,7,284-304.

[57] Chavez LL, Onuchic JN & Clementi C. *J. Amer. Chem. Soc.* **2004**,126,8426-8432.

[58] Socci ND, Onuchic JN & Wolynes PG. *J. Chem. Phys.* **1996**,104,5860-5868.

[59] Chahine J, Oliveira RJ, Leite VBP & Wang J. *Proc. Nat. Acad. Sci.* **2007**,104,14646-14651.

[60] Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH & Ranganathan R. *Nature* **2005**,437,512-518.

[61] Gosavi S, Whitford PC, Jennings PA & Onuchic JN. *Accepted PNAS.*

[62] Miyashita O, Onuchic JN & Wolynes PG. *Proc. Nat. Acad. Sci. USA* **2003**,100,12570-12575.

[63] Miyashita O, Wolynes PG & Onuchic JN. *J. Phys. Chem. B* **2005**,109,1959-1969.

[64] Whitford PC, Onuchic JN & Wolynes PG. *HFSP J.* **2008**,2,61-64.

[65] Hyeon C & Onuchic JN. *Proc. Nat. Acad. Sci. USA* **2007**,104,2175-2180.

[66] Hyeon C & Onuchic JN. *Proc. Nat. Acad. Sci. USA* **2007**,104,17382-17387.

[67] Wonmuk H, Lang MJ & Karplus M. *Structure* **2008**,16,62-71.

[68] Sobolev V, Wade R, Vried G & Edelman M. *Proteins, Struct. Funct. Genet.* **1996**,25,120-129.

[69] Berendsen HJC, Postma JPM, VanGunsteren WF, Dinola A & Haak JR. *J. Chem. Phys.* **1984**,81,3684-3690.

[70] Ferrenberg AM & Swendsen RH. *Phys. Rev. Letters.* **1988**,61,2635-2638.

[71] Ferrenberg AM & Swendsen RH. *Phys. Rev. Letters.* **1989**,63,1195-1198.

[72] Berendsen HJC, van der Spoel D & van Drunen R. *Comp. Phys. Comm.* **1995**,91,43-56.

[73] Jorgensen WL & Tirado-Rives J. *J. Am. Chem. Soc.* **1988**,110,1657-1666.

[74] Jorgensen WL, Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW & Klein ML. *J. Chem. Phys.* **1983**,79,926935.

[75] Miyamoto S & Kollman PA. *J. Comp. Chem.* **1992**,13,952-962.

[76] Hess B, Bekker H, Berendsen HJC & Fraaije JGEM. *J. Comp. Chem.* **1997**,18,1463-1472.

Chapter 5

Non-local helix formation is key to understanding SAM-1 riboswitch function

5.1 Abstract

Riboswitches are non-coding RNAs that regulate gene expression in response to changing concentrations of specific metabolites. Switching activity is affected by the interplay between the aptamer domain and expression platform of the riboswitch. The aptamer domain binds the metabolite, locking the riboswitch in a ligand-bound conformation. In absence of the metabolite, the expression platform forms an alternative secondary structure by sequestering the 3 end of a non-local helix called P1. We use all-atom structure-based simulations to characterize the folding, unfolding and metabolite binding of the aptamer domain of the S-adenosylmethionine-1 (SAM-1) riboswitch. Our results suggest that folding of the non-local helix (P1) is rate limiting in aptamer domain formation. Interestingly, SAM assists folding of the P1 helix by reducing the associated free energy barrier. Because the 3 end of the P1 helix is sequestered by an alternative helix in the absence of metabolites, this observed ligand-control of P1 formation provides a mechanistic explanation of expression platform regulation.

5.2 Results

Structure formation in mRNA often regulates genetic expression. Multiple compact conformations may be accessed while kinetic and thermodynamic competition of these structures determines the functional state of the mRNA [1]. In these systems the folding dynamics can play a critical role in biological function. Riboswitches are one class of functional mRNA units that are often found in specific 5-untranslated regions of mRNA [2]. They regulate transcription and translation in response to changing concentrations of metabolites via communication between an aptamer (metabolite binding) domain and the expression platform (Figure 5.1a). Conformational changes in the aptamer domain are essential for this functional response. Little is known about riboswitch function from a theoretical perspective, as computational efforts have largely been focused on smaller RNA fragments [3]. One question of interest is: How does ligand binding influence the formation of secondary and tertiary structure? Recent single molecule force spectroscopy experiments [4] have suggested the helix formed by the 3 and 5 ends of a pbuE adenine riboswitch is the least thermodynamically stable helix and is the helix most sensitive to metabolite concentrations. In contrast, fluorescence experiments suggest native 5-3 helix formation occurs prior to metabolite binding in a thiM riboswitch [5].

In this letter we describe the role of the 5-3 helix (P1) folding and S-Adenosylmethionine (SAM) binding in the activity of the SAM-I riboswitch [6] (Fig. 5.1). We adopt the energy landscape theory of protein folding [7] and apply it to RNA via an all-atom structure-based model ([8, 9]; see Supplementary Information). We compare aptamer domain folding with and without its associated metabolite, SAM. The functional state of the riboswitch is regulated by the balance of aptamer domain folding and formation of an alternate conformation involving a terminator sequence binding the 3 tail of the riboswitch [10]. It has been suggested that breaking of the 3 tail (in the non-local helix) is needed to regulate the expression platform. While the terminator sequence has been identified, the structure of the full riboswitch has not been solved and the precise details of the decision process need to be determined. However, the folding of both confor-

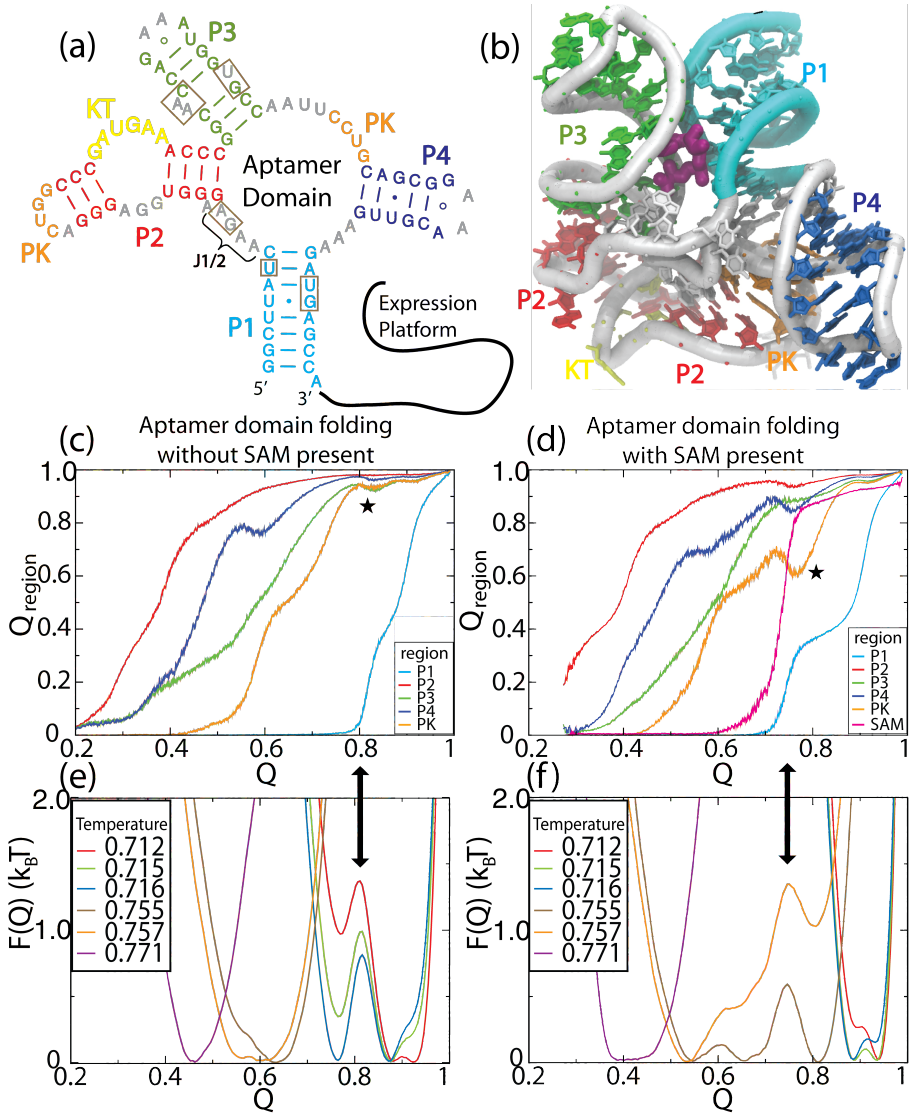

Figure 5.1: (a) Secondary and (b) tertiary structure (PDB entry: 2GIS) of the SAM-I Riboswitch. Average secondary structure formation as a function of the fraction of native contacts formed (Q; see Data S2) for the (c) SAM-free and (d) SAM-present simulations. Figures a-d use the same color scheme; P1=cyan, P2=red, P3=green, P4=blue, PK=orange, SAM=purple in (b) and (d). In (a) SAM contacting residues are highlighted by brown boxes. The most notable difference in folding mechanism is earlier initial folding of P1 (black arrows) at the expense of the PK (starred) when SAM is present. The folding free-energy profiles for the (e) SAM-free and (f) SAM-present simulations are shown for several temperatures (with temperature indicated by color). The most significant free-energy barrier in both systems is associated with initial P1 folding. When SAM is present, the free-energy barrier is reduced and encountered earlier in the folding process.

mations must occur on the same energy landscape. Thus, rate limiting steps in aptamer formation may provide opportunities for the alternate structure to form and the functional decision to be made. We perform simulations using the recently solved x-ray structure of the SAM-1 riboswitch aptamer domain [6], allowing us to isolate the role of P1 formation in aptamer folding. Our results suggest the rate limiting step in aptamer domain folding is the initiation of P1 helix formation. SAM reduces the associated free-energy barrier by binding to the pre-formed P3 helix and then attracting the unstructured strands of the P1 helix.

Energy landscape theory states that nature has selected for protein sequences that maximize the energetic bias for the native state and minimize trapping of non-native structures. Namely, they have been selected to be minimally frustrated. The principle of minimal frustration has been validated through comparison of structure-based models and experimental results, which has led to the funnel paradigm of protein folding [7]. For structured RNA, one can envision a frustrated landscape where there is a margninal bias to reach the native state. The RNA would then randomly search all possible base pairs and the folded state would only be reached by chance. This would result in a "Levinthals paradox," where searching takes the age of the universe, whereas, in reality, folding of functional RNAs can be fast (~ ms). Therefore, evolutionary pressure to reduce frustration must exist. While RNA is likely frustrated to some degree, by understanding energetically unfrustrated models one can partition the structural and energetic effects in folding and function.

The principle of minimal frustration is applied via structure-based simulations in which all heavy atoms are explicitly represented. The model is energetically unfrustrated since only native interactions are attractive and all other interactions are repulsive. Kinetic (temperature jump) and thermodynamic (constant temperature) simulations of the aptamer domain were performed, both with and without SAM present. Thermodynamic simulations ranged in temperature such that the full folding/binding landscape could be characterized (Fig. 5.1c-f and 5.2). For SAM-present simulations, one copy of the aptamer domain and 100 copies of the SAM molecule are placed in a box with periodic boundary conditions.

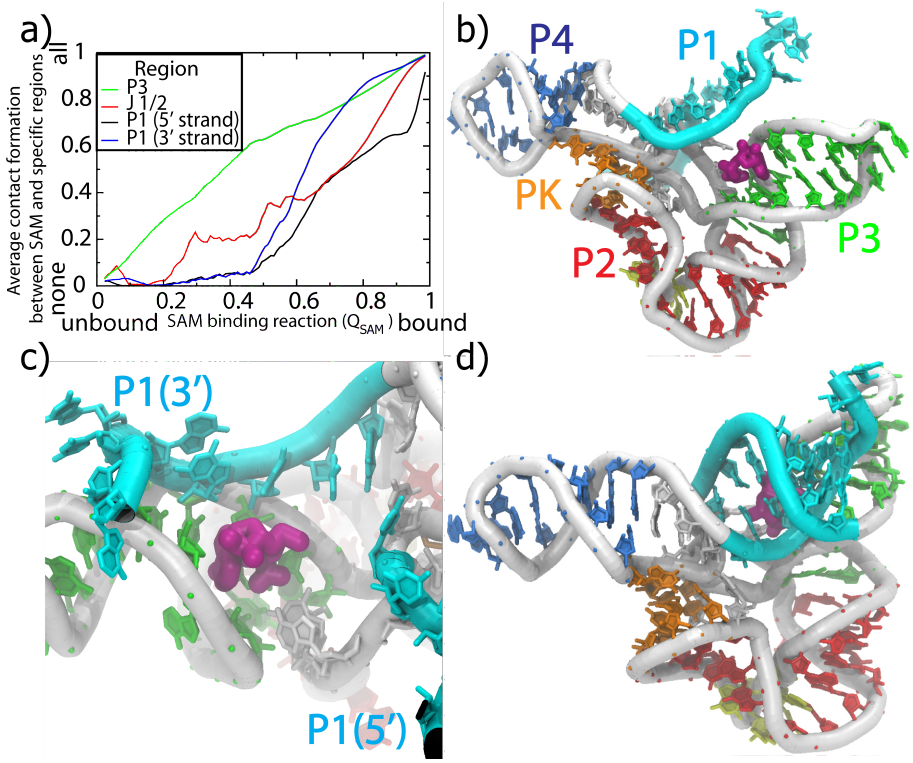

Figure 5.2: (a) Average percent of SAM-aptamer domain interactions formed by region as a function of the fraction of native SAM-aptamer domain contacts formed QSAM. Simulation images illustrating SAM binding mechanism: (b) SAM binds a preformed P3 helix, (c) SAM recruits 3 strand of P1, (d) SAM binds 5 strand of P1 and P1 helix formation proceeds.

SAM molecules are free to associate and only native SAM-aptamer interactions are attractive. Since SAM-SAM interactions are strictly repulsive, metabolite aggregation and non-specific binding are not possible. To our knowledge, this is the first simulation in which a bath of ligands (with atomic resolution) is able to freely associate and dissociate with a RNA molecule during folding.

In thermodynamic simulations of the apo aptamer domain, the largest free-energy barrier is associated with initial formation of the P1 helix (Fig. 5.1c and e, black arrow). In the presence of SAM, the initiation of P1 helix formation and the free-energy barrier are encountered earlier in the folding process (Fig. 5.1d and f, black arrows) and the free-energy barrier is reduced. P1 forms after all other secondary structure (and some tertiary structure) is formed and SAM primarily affects P1 folding in both thermodynamic and kinetic simulations (see Fig. A.1

in SI). In the SAM riboswitch, the SAM molecule stabilizes the rate limiting step (largest free energy barrier; see SI) in folding, which leads to a kinetically accessible and thermodynamically more stable folded aptamer domain.

Since the P3 domain is formed prior to SAM binding (Fig. 5.1c, green curve), P3 can serve as a platform for SAM binding. Figure 5.2 shows that upon binding to P3, SAM stabilizes the P1 domain by predominantly interacting with the 3 strand and then the 5 strand of P1.

Another notable feature in Figures 5.1c and d is the apparent interplay between P1 and the pseudo-knot (PK, starred). In kinetic simulations (see SI) this partial unfolding of the PK is more pronounced, suggesting that a dynamic balance between PK and P1 formation exists.

The current picture of RNA folding is hierarchical [11]. In this view, it is important to distinguish between local helices (formed by simple stem-loops) and non-local helices (formed by two strands distant in sequence) [12]. Relative to a stem loop, a non-local helix has a larger loss of entropy associated with its formation. This unfavorable driving force is often accounted for in secondary structure prediction algorithms, where scoring penalties are imposed on large loops [13]. Thus, it may not be surprising to find a non-local helix (P1) that is less stable than the local helices. As we have shown, the entropic barrier due to bringing together distant (in sequence) bases also gives rise to the rate limiting step, initiation of P1 folding.

Since P1 folding is rate limiting, it is an ideal stage for SAM to bind and the on/off decision to be made. Our results provide a detailed mechanism for both this switching decision and SAM binding. Our results also suggest the structural mechanism of control is the same, regardless of whether the process is thermodynamically, or kinetically, regulated [14]. The less stable, and slower forming, P1 helix results in a compact state where some tertiary structure (the PK) can be formed. In this partially structured state SAM may bind to a pre-formed P3 helix. After SAM binds to P3, it localizes the 3 and 5 strands that compose the P1 helix. SAM binding to P1 initiates P1 helix formation (Fig. 5.1d), after which P1 continues to form without any significant free-energy barriers.

Since the P1 helix is a fragile structure (relative to P2, P3 and P4), it is likely more sensitive to the cellular environment. Force spectroscopy experiments have shown a coupling between non-local helix formation and ligand binding in an adenine riboswitch [4]. In Azoarcus ribozyme [15], a near-native, compact state with partial tertiary structure has been experimentally observed. This is also consistent with non-local helix formation being the final folding step. While non-local helix formation is important in some RNA-ligand systems, loop ordering [16, 17] and tertiary structure formation [5] may also be important in the decision processes of other riboswitches.

Several recent results have shown that molecular recognition, control and signaling do not necessarily occur by surface matching between biomolecules. Rather, a more interesting process occurs where folding of the biomolecular parts is signaled through binding. Our results suggest that initial P1 formation is a central step for further recognition and function in the SAM aptamer.

5.3 Acknowledgement

PCW and AS thank Ulrich Mller at UCSD for useful discussions. This work was supported in part by Grants PHY-0216576 and 0225630 from the National Science Foundation (NSF) sponsored Center for Theoretical Biological Physics, NSF Grant 0543906, the LANL LDRD program and NIH Grant R01-GM072686. We would also like to thank the Tera-grid computing facilities at Virginia Tech.

Chapter 5, in full, appears in Biophysical Journal, 2009, Whitford, Schug, Saunders, Hennelly, Onuchic, Sanbonmatsu. The dissertation author is the primary investigator and author of the paper.

Bibliography

[1] Breaker RR.*Science*, **2008** 319,1795-1797.

[2] Montange RK, & Batey RT.*Annu. Rev. Biophys.* **2008**, 37,117-133.

[3] Garcia AE, & Paschek D. *J. Amer. Chem. Soc.* **2008** 130,815-817.

[4] Greenleaf W, Frieda JKL, Foster DAN, Woodside MT, & Block SM. *Science* **2008**, 319,630-633.

[5] Lang K, Rieder R & Micura R.*Nuc. Acid. Res.* **2007**, 35,5370-5378.

[6] Montange RK & Batey RT. *Nature* **2006**, 441,1172-1175.

[7] Onuchic JN, Luthey-Schulten Z & Wolynes PG. *Ann. Rev. Phys. Chem.* **1997**, 48,545-600.

[8] Our model is based on the ideas presented in Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY & Onuchic JN. *Prot. Struct. Func. Bioinfo.* **2009** DOI, 10.1002/prot.22253.

[9] Structure-based models have been used for RNA folding in, Sorin EJ, Nakatani BJ, Rhee YM, Jayachandran G, Vishal V & Pande VS. *J. Mol. Biol.* **2004**, 337,789-797; Hyeon C & Thirumalai D.*Biophys. J.* **2007**, 92,731-743; Hyeon C, Dima DI, & Thirumalai D.*Structure* **2006**, 14,1633-1645.

[10] Winkler WC, Nahvi A, Sudarsan N, Barrick JE & Breaker RR. *Nat. Struct. Biol.* **2003** 10,701-707.

[11] Brion P & Westhof E. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, 26,113-117.

[12] Woodson SA *Biochem . Soc. Trans.* **2002**, 30,1166-1169.

[13] Chen S *Annu. Rev. Biophys.* **2008**, 37,197-214.

[14] Coppins RL, Hall KB, & Groisman EA *Curr. Opin. Struct. Bio.* **2007**, 10,176-181.

[15] Chauhan S & Woodson SA. *J. Amer. Chem. Soc.* **2008**, 130,1296-1303.

[16] Gilbert SD, Stoddard CD, Wise SJ, & Batey RT. *J. Mol. Biol.* **2006**, 359,754-768.

[17] Ottink MO, Rampersad SM, Tessari M, Zaman GJR, Heus HA, & Wijmenga SS. *RNA* **2007**, 13,2202-2212.

Appendix A

Supporting Information for Chapter 5

A.1 Formulation of the Hamiltonian

Coarse-grained structure-based models have been applied with varying degrees of resolution to simulate RNA and protein molecules[1, 2, 3, 4, 5, 6]. To simulate the SAM riboswitch, we employ an all-atom structure based potential, similar to that used previously for proteinsvii and RNA[8]. All heavy atoms are explicitly included and are given unit mass. Since thermodynamic quantities do not depend on mass. In kinetic simulations using unit mass for all atoms could have an effect. Though, rescaling each phosphate atom to be three times the mass of a carbon atom would only increase the total mass by 10%. Harmonic potentials restrain the bond lengths and angles, as well as planar dihedrals. Flexible dihedral angles are included via cosine terms. Non-local contacts formed in the native structure are given attractive 6-12 interactions. All other non-local interactions are repulsive. The functional form of the potential is

$$
\begin{aligned}
V = & \sum_{bonds} \epsilon_r (r - r_o)^2 + \sum_{angles} \epsilon_\theta (\theta - \theta_o)^2 + \sum_{planar} \epsilon_\chi (\chi - \chi_o)^2 \\
& + \sum_{backbone} K_{\phi,BB}[[1 - cos(\phi - \phi_o)] + 0.5[1 - cos(3(\phi - \phi_o))]] \\
& + \sum_{sidechains} K_{\phi,SC}[[1 - cos(\phi - \phi_o)] + 0.5[1 - cos(3(\phi - \phi_o))]] \\
& + \sum_{contacts} \epsilon(i,j)[(\frac{\sigma_{ij}}{r})^{12} - 2(\frac{\sigma_{ij}}{r})^6] + \sum_{non-contacts} \epsilon_2(i,j)(\frac{\sigma_{NC}}{r})^{12}
\end{aligned}
\tag{A.1}
$$

where r_0, θ_0, ϕ_0, & σ_0 are determined by the native structure and σ_r = 2.5Å. K_r = 100Å^2, K_θ = 20/rad^2, K_χ = 10/rad^2 & ϵ_2 = 0.01. Native contact distances σ_{ij} are assigned to all atom-atom pairs that are within 4 Åin the native structure. According to this definition of a contact, stacked bases have about 5 times the number of contacts that hydrogen bonded base pairs have. To account for this, the contacts between stacked bases are rescaled by 1/3[1]. Contact interactions and dihedral interactions are weighted such that

$$
R_{C/D} = \frac{\sum \epsilon(i,j)}{\sum K_\phi} = \frac{\text{total contact energy}}{\text{total dihedral energy}} = 2.0 \tag{A.2}
$$

and

$$
R_{BB/SC} = \frac{K_{\phi,BB}}{K_{\phi,SC}} = \frac{\text{Strength of each Backbone dihedral}}{\text{Strength of each Side Chain dihedral}} = 1.0 \tag{A.3}
$$

are satisfied. Reduced units are used for all calculations.

A.2 Features of the Hamiltonian

A.2.1 Distribution of energy

Energies were assigned such that $R_{C/D}$ = 2.0 and $R_{BB/SC}$ = 1.0. In protein folding, these parameters are easily interpreted, since dihedrals are local and contact are usually between non-local (in sequence) pairs of amino acids. For RNA,

[1]Stacking interactions are chemically different from hydrogen bonding. When using a cut-off distance criterion for contacts, there is no reason that each stacking contact should be the same strength as a base pairing contact. The rescaling is only intended to give a more reasonable distribution of energy (Table A.1).

Table A.1: Summary of Distribution of energy by type of interaction. When setting $R_{C/D}$ = 2.0 the dihedral angles in the sugar ring are included. Though, since sugar ring dihedrals are far less flexible due to bonds and bond angles, it is more appropriate to exclude them in this summary. This results in an effective $R_{C/D}$ of 4.0 (20% of the energy in dihedrals).

Types of Interaction	**Energy (% of total)**
Base Pairing contacts	20.7
contacts between adjacent (in sequence) bases	39.9
Tertiary contacts	16.8
pseudo-knot base pair contacts	2.6
backbone dihedrals	16.0
sidechain dihedrals	4.0
Total secondary energy	80.6
Total tertiary energy	19.4
Total SAM-Aptamer energy	2.3 (relative to total aptamer energy)

it is instructive to further partition the energy. Table A.1 indicates that this parameterization of the potential corresponds to 80.6% of the energy in secondary structure and 19.4% in the tertiary interactions. While these ratios might suggest the parameterization corresponds to a low [Mg^{2+}], a direct comparison is not appropriate here[2]. To substantiate claims about ionic effects, it would be necessary to include (or approximate) tertiary electrostatic interactions, with an appropriate measure for [Mg^{2+}]. By varying these parameters, the electrostatic effects may be probed[8]. Low [Mg^{2+}] results in denaturation of RNA through backbone-backbone repulsion. Increasing [Mg^{2+}] provides additional stability to the native structure by mediating these tertiary interactions[9, 10, 11, 12, 13, 14]. In the present study, native backbone-backbone interactions are attractive, but they are the same strength as other types of contacts. Our parameterization roughly corresponds to an intermediate [Mg^{2+}], as tertiary interactions are not repulsive, but are weakly attractive.

To study the effects of SAM, we explicitly include the ligand. In doing so,

[2]Since folding mechanisms in structure-based models are robust to parametric changes (see below), it is only necessary that our parameters are reasonable.

Table A.2: Summary of contact energy distribution in each helix.

Region	Total Contact E	# of base pairs	$< E >$ per pair
P1	187.7	8	23.5
P2	170.0	7	24.3
P3	150.1	7	21.4
P4	151.7	6	25.3
PK	84.4	4	21.1

we perturb the system by increasing the stabilizing energy and adding structural constraints. Table A.1 indicates that the SAM-Aptamer contacts only increase the total stabilizing energy by 2.3%. The observed increase of 5% in the folding temperature of P1 upon addition of SAM (from T=0.716 to 0.755; Figure 1) further emphasizes the focused effect SAM has on P1.

A.2.2 G-C vs A-U base pairing

It is well established that G-C base pairs interact more strongly A-U base pairsxvi. While we do not give a bias to G-C contacts over A-U contacts, our simple distance criterion results in more contacts, and hence more stabilizing interactions, between G-C pairs. Figure A.1 shows representative G-C and A-U base pairs with all native contacts indicated by dotted purple lines.

A.2.3 Stacking interactions

An additional feature that is implicit to our model is the distribution of stacking energy. Since our model is structure-based, bases that are well stacked (as seen in helices) will have the highest number of contacts, and bases that are poorly stacked (as seen in turns and loops) will have fewer contacts (Figure A.2). Our model does not treat stacking in local and non-local helices differently. The main factors in determining the stabilizing energy assigned to a helix are 1) base pairing and 2) how well the bases are stacked. The stabilizing energy per base pair is given for each helix. Since P1 does not have significantly less enthalpy per base pair than P2-P4, the observed differential stability of P1 is a result of entropy.

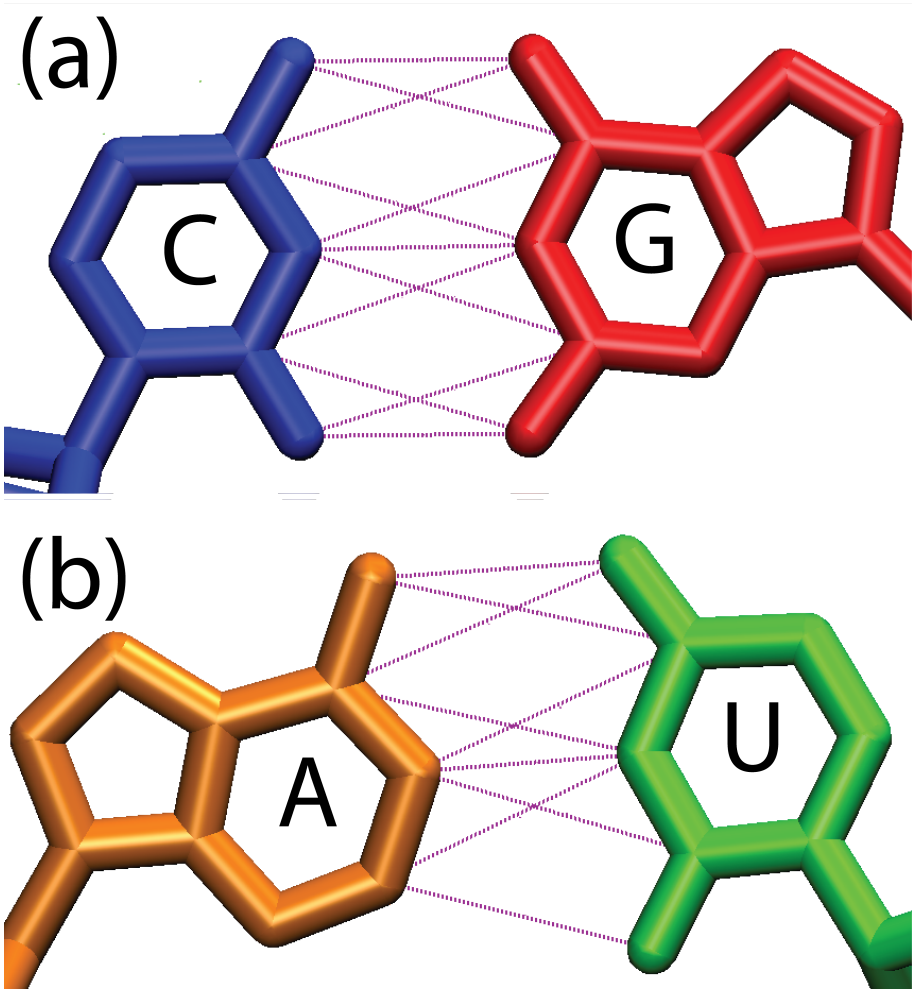

Figure A.1: A representative (a) G-C base pair and (b) A-U base pair with native contacts represented by purple lines. The G-C pair has 11 contacts while the A-U pair only has 9. Image prepared with VMD.

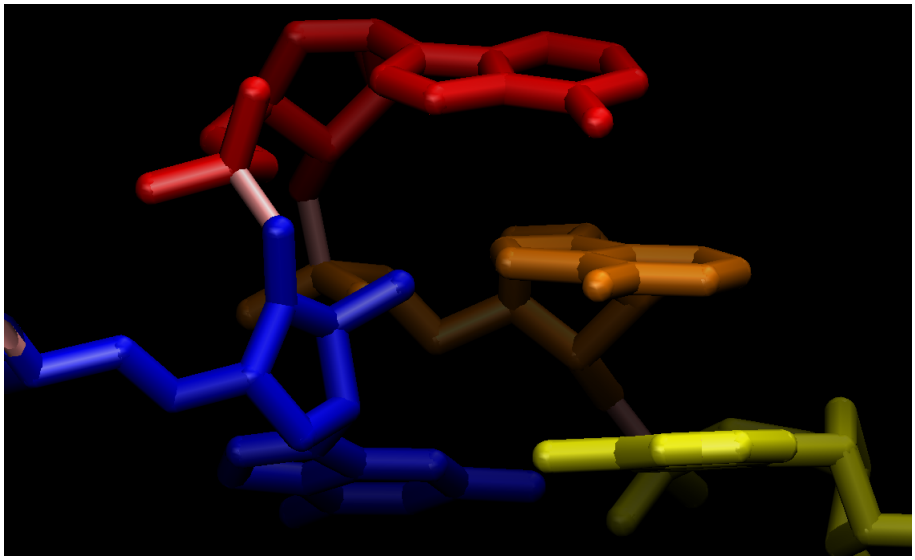

Figure A.2: Stem loop of helix P3 (G50 to A53: colored blue, red, orange and yellow). Bases 51 to 53 are well stacked. Bases 50 and 51 are adjacent in sequence, but are not stacked.

A.2.4 Potential hamiltonian effects on Pseudo-knot, P1 and kink-turn stability

In the formulation of our model, pseudo-knot base pairs are not treated differently from helical base-pairs. That is, the PK contacts are given the same energetic weight as helical base pair contacts. There are several reasons that PKs should be less stable than helices. First, PKs tend to have a small number of base pairs (¡5). Second, PKs are tertiary interactions and are therefore entropically unfavorable. Third, PKs can require bending of a helix in order to form.

The number of base pairs in a PK does not affect the formulation of our model. Though, as described above, well stacked bases have more stabilizing energy than poorly stacked bases. This can lead to boundary effects at the ends of helices. The terminal base in a helix is not always stacked against the first base of a loop (Figure A.2). Table A.2 indicates that this boundary effect is negligible when comparing the helices of this particular riboswitch and each helix has 6-8 base pairs. In the PK, there are only 4 base pairs and the stabilizing energy per base pair is less than in helices, but only marginally less than P3.

The second contributing factor to PK stability is the large loss in entropy upon structure formation. Despite both the entropic penalty and the smaller stabilizing energy, the PK is still more stable than P1 for this parameterization of the model.The formation of a PK can also introduce structural constraints that result in a "bent" helix. That is, the helix will have a persistence length that is shorter when the PK is formed than when it is not. This bending of a helix can introduce an energetic penalty to PK formation. Since our model is structure-based, there is not a penalty for "bending" the P2 helix. By adding an energetic term that favors a longer persistence length of P2, the PK could be destabilized. It is important to realize the persistence length would be increased if the Kink-turn was replaced by canonical base pairs, to form a continuous P2 helix. In such a system, the PK could potentially be less stable than P1, which is consistent with the interplay between the PK and P1 seen in kinetic folding simulations. Less drastic mutations to the KT region may also perturb the PK, though at this point the exact effects would be speculation.

A.3 Robustness of the results to changes in the distribution of energy

Whitford et al.[7] have shown that folding mechanisms in proteins are not sensitive to the choice of parameters in an all-atom structure based model. Since we employ a similar model here, the mechanisms in this paper should not be sensitive to the exact choice of parameters. They have also shown that folding free energy barriers can be varied with parameters, but the relative heights of different barriers remain similar for each parameterization. The folding barriers between the Apo and SAM-present simulations should be considered relative barriers, in that the largest barriers are likely rate limiting, yet the overall magnitudes are not meant to be exact.

A.4 Simulation details

All simulations were performed using the Gromacs[17, 18] software package. No changes to the source code were necessary. Reduced units were used in all calculations and figures. The time step τ was 0.0004. Stochastic dynamics were used with a drag coefficient γ = 1.0.

Kinetic folding simulations were obtained by first thermally denaturing the RNA. An ensemble of unfolded structures was obtained from several high-temperature simulations. The unfolded structures were then thermally quenched to around $0.8T_f$. Approximately 100 kinetic simulations were performed each for the Apo and SAM-present simulations.

To obtain thermodynamic sampling, over 50 simulations were performed for the Apo and SAM-present systems, each at a different temperature. Temperatures were chosen such that the aptamer domain was always folded in the lowest temperature simulations and always unfolded in the highest temperature simulations. The Weighted Histogram Analysis Method[19, 20] was used to calculate thermodynamic quantities.

In the SAM present simulations, 100 copies of SAM and 1 copy of the aptamer domain are present in a cubic box of dimension 250Å. In experiments, a concentration of this magnitude could result in non-specific binding of SAM and large non-physiologically relevant electrostatic contributions. In our simulations, since these effects are not present we can focus specifically on the binding event.

A.5 Reaction coordinates

For all results presented here, we use the fraction of native contacts Q as a reaction coordinate. The primary requirement of any good reaction coordinates is that is can distinguish between folded and unfolded structures. Q has been extensively studied in the context of protein folding[24]. Since RNA is also a polymer, there is no reason a priori to believe Q is inappropriate for the characterization of RNA folding. A native contact is defined in our model as any two atoms that are separated by at least 3 bonds, are in different residues and are within 4Åof

each other in the native structure. Q is defined as the fraction of natively contacting pairs that are within 1.5 times their native distance, at a given time. In this manuscript, secondary structure contacts Qregion are defined as contacts for which both atoms are in the same structural unit (region=P1, P2, P3, P4 or PK, though the PK is not a secondary structure unit) and are not in adjacent (in sequence) bases.

A.6 Folding mechanisms from kinetic simulations

Figure A.3 shows the folding mechanisms of the aptamer domain in kinetic simulations. As seen in thermodynamic simulations, P1 is the last helix to form. The apparent partial unfolding of the Pseudo-knot (PK) upon P1 formation is more pronounced in kinetic simulations than in thermodynamic simulations. Decreases in Q_{region} with increasing total Q has been observed in protein folding and is referred to as "backtracking"[25, 26]. Backtracking occurs in unfrustrated models when structure formation of one region impedes the folding of another region. It was recently shown that the folding of functional regions of proteins can lead to backtracking[27]. In the SAM-1 riboswitch, backtracking suggests P1 and the PK are functionally relevant. Perturbations to the PK (mutation, ligand binding, ionic concentration, etc.) may effect on the functional state of the riboswitch.

It is worth noting that P2 folds later in kinetic simulations than in thermodynamic simulations. This is likely due to P2 being a significantly longer (spatially) helix than P3 or P4. The longer length leads to a kinetic delay, though it is thermodynamically the most stable helix. Despite the longer length scale, it still forms before P1.

Figure A.4 shows the binding mechanism for SAM in the kinetic simulations. SAM clearly binds P3 prior to P1. This feature is observed in thermodynamics simulations, but it is more pronounced here. The sequential docking of the 3 and 5 strands is also accentuated.

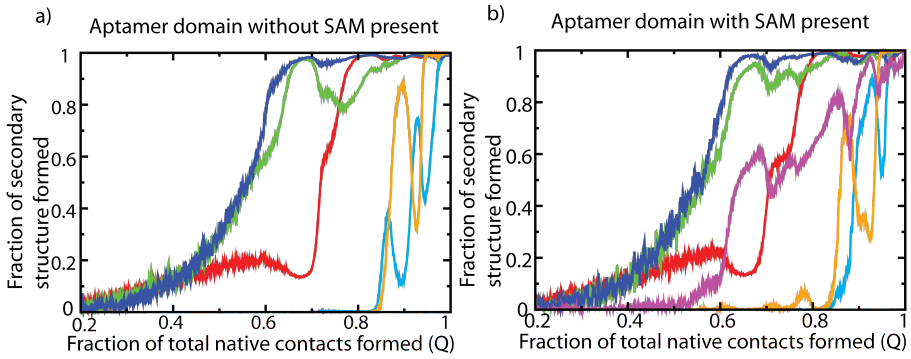

Figure A.3: Folding mechanisms in kinetic simulations. Fraction of native contacts formed by region as a function of the fraction of total contacts formed Q (P1=cyan, P2=red, P3=green, P4=blue, PK=orange, purple=SAM-aptamer contacts). SAM appears to influence the folding of P2, P3 and P4 less than P1. P1 and PK formation are late in both Apo and SAM-present simulations. Folding of the PK appears to be correlated with partial unfolding of P1. SAM reduces the folding of P1 from 3 to 2 steps.

A.7 Relating rates and free energy barriers

The rate of a chemical reaction, such as polymer folding, can be related to the diffusion and free energy barrier in Q via, $R \sim De^{-\beta \Delta F}$, where ΔF and D are the free energy barrier associated with the reaction and the diffusion coefficien[28, 29, 30]. When relating our free energy barriers to rates, higher free energy barriers correspond to slower folding. In protein folding, free energy barriers are often larger for small fast folding proteins than the ones obtained in this study. This does not indicate riboswitch folding is faster than folding of a small protein. To make quantitative arguments about the magnitude of our rates, one would need approximations of D for protein and RNA folding. Since D is the diffusion in reaction coordinate space, the projection of the diffusion in Cartesian coordinates to diffusion in Q space will be different for different molecular structures. The length scale is larger for the SAM riboswitch than for small proteins. The larger the system, the farther apart native pairs can be when in the unfolded state. Consequently D can be smaller, and rates will be slower, for the riboswitch than for small proteins. Thus, our results are limited to relative rates of similar systems. In the present work, we compare the free energy barriers of the Apo and SAM-

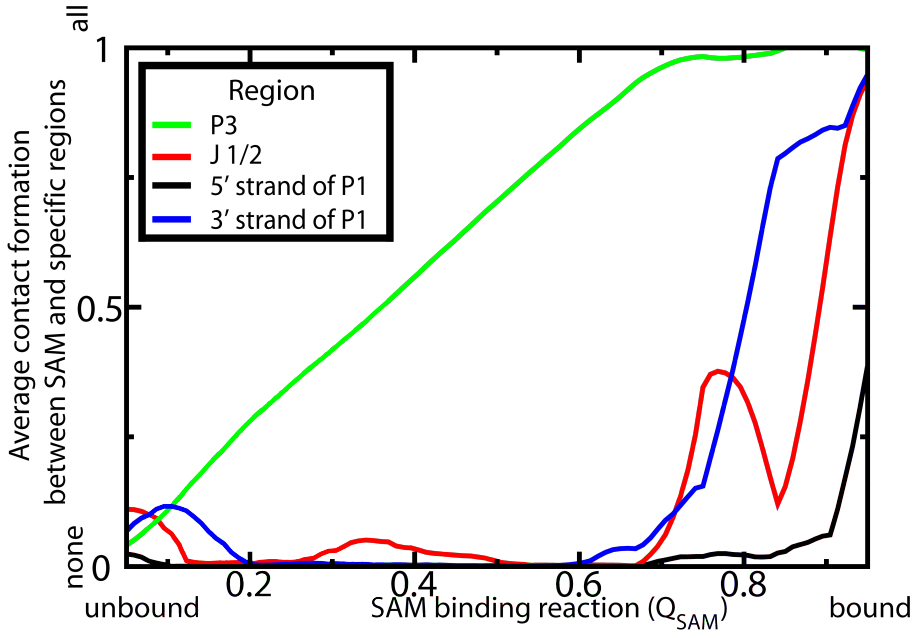

Figure A.4: Mechanism of SAM binding in kinetic simulations. The sequential binding of SAM, where SAM first binds P3, then the 3 strand of P1 and finally the 5 strand of P1, is accentuated in the kinetic simulations.

present riboswitch. Since these two systems are identical, except for a perturbation from the ligand, it is reasonable to assume that differences in D will be negligible, and the differences in rates are due solely to the free energy barriers.

Bibliography

[1] Hyeon C, Thirumalai D. *Biophys. J.* **2007**, 92,731-743.

[2] Hyeon C, Dima RI, Thirumalai D *Structure* **2006**, 14,1633-1645.

[3] Clementi C, Nymeyer H, Onuchic JN *J. Mol. Biol.* **2000**, 298,937-953.

[4] Oliveira LC, Schug A, Onuchic JN *J. Phys. Chem. B.* **2008**, 112,6131-6136.

[5] Cheung M S, Garcia AE, Onuchic JN *Proc. Nat. Acad. Sci. USA* **2002**, 685-690.

[6] Murarka R K, Liwo A, Scheraga HA *J. Chem. Phys.* **2007**, 127,155103.

[7] Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, Onuchic JN.*Prot. Struct. Func. Bioinfo.*, **2009**DOI, 10.1002/prot.22253.

[8] Sorin EJ, Nakatani BJ, Rhee YM, Jayachandran G, Vishal V, Pande VS *J. Mol. Biol.* **2004**, 337,789-797.

[9] Zuker M. *Nuc. Acid. Res.* **2003**, 31,3406-3415.

[10] Matthews DH, Sabina J, Zuker M, Turner DH *J Mol. Biol.* **1999**, 288,911-940.

[11] Laing LG, Draper DE *J. Mol. Biol.* **1994**, 237,560-576.

[12] Fang X, Pan T, Sosnick TR *Biochem.* **1999**, 38,16840-16846.

[13] Gonzalez RL, Tonico I Jr *J. Mol. Biol.* **1999**, 289,1267-1282.

[14] Misra VK, Draper DE. *J. Mol. Biol.* **2000**, 299,813-825.

[15] Cate JH, Hanna RL., Doudna JA *Nat. Struct. Biol.* **1997**, 4,553-558.

[16] Shiman R, Draper DE *J. Mol. Biol.* **2000**, 302,79-91.

[17] Humphrey W, Dalke A and Schulten K, *J. Molec. Graphics*, **1996**, 14,33-38.

[18] Walker AE, Turner DH *Biochem.* **1994**, 33,12715-12719.

[19] Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruther MH, Neilson T, Turner DH *Proc. Nat. Acad. Sci. USA* **1986**, 83,9373-9377.

[20] Lindahl E, Hess B, van der Spoel, D. *J. Mol. Mod.* **2001**, 7,306-317.

[21] Berendsen HJC, van der Spoel, D, van Drunen, R. *Comp. Phys. Comm.* **1995**, 91,43-56.

[22] Ferrenberg AM, Swendsen RH *Phys. Rev. Lett.* **1988**, 61,2635-2638.

[23] Ferrenberg AM, Swendsen RH *Phys. Rev. Lett.* **1989**, 63,1195-1198.

[24] Cho SS, Levy Y and Wolynes PG. *Proc. Nat. Acad. Sci. USA* **2006**, 103,586-591.

[25] Gosavi S, Chavez LL, Jennings PA. , Onuchic J. N. *J. Mol. Biol.* **2006**, 357,986-996.

[26] Chavez LL, Gosavi S, Jennings PA , Onuchic JN *Proc. Nat. Acad. Sci. USA*, **2006**, 103,10254-10258.

[27] Gosavi S, Whitford PC, Jennings PA , Onuchic JN *Proc. Nat. Acad. Sci. USA*, **2008**, 105, 10384-10389.

[28] Kramers HA *Physica* **1940**, 7,284-304.

[29] Bryngelson JD and Wolynes PG *J. Phys. Chem.* **1989**, 93,6902-6915.

[30] Zwanzig R *Proc. Nat. Acad. Sci. USA*. **1988**, 85,2029-2030.

A.8 Acknowledgement

Appendix A, in full, appear in Biophysical Journal, 2009, Whitford, Schug, Saunders, Hennelly, Onuchic, Sanbonmatsu. The dissertation author is the primary investigator and author of the paper.