Advances in Coarse-grained Models for Protein Folding and Protein-protein Interactions

by

Yanming Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry and Scientific Computing)
in the University of Michigan
2020

Doctoral Committee:

Professor Charles L. Brooks III, Chair
Assistant Professor Aaron T. Frank
Professor Kevin J. Kubarych
Professor Anna K. Mapp

Yanming Wang

ymwang@umich.edu

ORCID iD: 0000-0002-2383-2470

Acknowledgements

I would first and foremost like to thank Prof. Charlie Brooks for his guidance and help in my graduate school study. I feel very grateful to have had the opportunity to work in his group over the past four years. Much of the work presented in the dissertation would not have happened without his insightful advice. I also want to thank my thesis committee members Prof. Aaron Frank, Prof. Anna Mapp, and Prof. Kevin Kubarych for their valuable suggestions on my research projects. I thank my experimental collaborators Kevin Halloran, Prof. Bob Matthews and Prof. Osman Bilsel for helping me to see the broader implications of my work. I would next like to thank past and present members of the Brooks lab including Karunesh Arora, Xinqiang Ding, Ryan Hayes, Hedieh Torabifard for their kind help. I also want to thank David Braun, Kathleen Dyki, and Nicholas O'Hair for their administrative support. Finally, I would like to thank my parents for their understanding and support.

Table of Contents

List of Tables

List of Figures

Abstract

Almost all biological functions rely on the dynamics of proteins. Protein folding and protein-protein interactions are the two most fundamental problems of protein dynamics. Molecular dynamics (MD) simulation is a powerful tool to elucidate the mechanism of protein folding and protein-protein interactions by providing atomic-level resolution. However, the timescales of protein folding and protein association-dissociation are often not accessible by all-atom MD simulations due to the high computational cost. Coarse-grained models effectively address this issue by reducing the degrees of freedom of the system to only a few that are essential for the properties to be studied. In this dissertation, I describe the developments and applications of coarse-grained modeling of protein folding and protein-protein interactions through several case studies.

I first present a computational study of the folding mechanism of a triosephosphate isomerase (TIM) barrel protein using a coarse-grained model. This is the first time this model was used to study a large protein with more than 200 amino acid residues. From the simulations, we proposed a 3-channel folding mechanism with one major and two minor folding pathways. The simulations show overall good agreements with the experiments in capturing the regions that are first to fold and capturing a rate-limiting intermediate state found in the major folding channel. The simulations advance our understanding of the folding mechanism of this TIM barrel proteins by directly providing structural details of the protein folding intermediates as suggested by experiments.

In the realm of protein-protein interactions, I developed a new sampling method based on Hamiltonian replica exchange (HREX) that allows efficient calibration of the coarse-grained model and fast calculation of the dissociation constant. This HREX method was used to study the protein-protein interaction in the context of the allosteric regulations in the KIX and TAZ1 domain of the CBP/P300 transcription coactivator. The simulations captured both the positive/cooperative allosteric effect in the KIX domain and the negative allosteric effect in the TAZ1 protein switch with two vastly different allosteric mechanisms. The simulations suggest the positive allosteric regulation in KIX, in which a prebound ligand favors the binding of the second ligand, is due to a favorable entropic change. Whereas the negative allosteric regulation in TAZ1, in which two ligands compete for binding the same target, is mainly driven by long-range electrostatic forces. The simulations also suggest the importance of electrostatics for the coarse-grained model to be generally successful in modeling the allosteric effect involving intrinsically disordered proteins. These studies advance our understanding of the allosteric effect by generating testable hypotheses that help quantify the problem.

Chapter 1 Introduction

Proteins are the building blocks of life. Almost all biological functions rely on the dynamics of proteins. Understanding protein dynamics in the cell is a central pillar of biology. Protein folding and protein-protein interactions are the two most fundamental problems of protein dynamics. Molecular dynamics (MD) simulation has proven to be a robust method to help elucidate the mechanism of protein dynamics (1, 2). Nowadays, the capability of MD simulations is still largely limited by the high computational cost to reach timescales long enough for biologically relevant events such as protein folding and protein association-dissociation to occur. To address this issue, there are typically two approaches. First, by increasing the sampling speed using specially designed hardware such as Anton (3) or using a massively distributed computing framework such as Folding@home (4). However, these resources are only available to a few researchers. The second approach exploits speedup by reducing the degrees of freedom while keeping the essential features of the system (5, 6). The reduced complexity allows problems with larger system size and longer timescales to be studied. The application of coarse-grained models in studying protein dynamics problems dates back to 1975 when Levitt and Warshel carried out the first computer simulation of protein (7). They developed a coarse-grained model that uses two beads to represent each amino acid residue to study the folding of bovine pancreatic trypsin inhibitor (BPTI). In their paper, they pointed out the two fundamental assumptions of their model (7). First, that much of the protein's fine structure can be eliminated by averaging. Second, that the overall chain folding can be obtained by considering only the most effective variables (those that vary most slowly yet

cause the greatest changes in conformation). These two assumptions are still generally true for most coarse-grained models used today.

This thesis focuses on the development and application of the coarse-grained model developed by Karanicolas and Brooks (KB model) (8). This model was originally designed to study protein folding mechanisms (9), and was recently adopted to study protein-protein interactions with some modifications (10, 11). In this model, each amino acid residue is represented as a single bead with mass equal to the corresponding amino acid, centered at the Cα position, and connected to neighboring residues via virtual bonds. The KB model mainly considers two types of interactions, bonded and non-bonded interactions (equations 1.1-1.2). The bonded interactions include bond, angle, and dihedral terms. The bond and angle terms are harmonic potentials with equilibrium values set to those calculated from the native structure. The dihedral term is a statistical potential based on probability distributions obtained from the Ramachandran plot and provides additional sequence-specific information to avoid overfitting to the native structure (8). The non-bonded interactions consider native contacts, and the non-native excluded volume interactions. A variant of the original KB model with explicit electrostatics named electrostatic inclusive KB model (EIKB) model was developed in this dissertation and was used to study protein-protein interactions in transcription regulation. In the EIKB model, the electrostatic interactions were represented by a simple Debye-Hückel potential (equation 1.3) with two parameters: the dielectric constant D and the screening length κ. This simple representation of electrostatics has been shown to be compatible with the coarse-grained model (12). The model only considers the electrostatic interactions among charged residues: ASP (-1), GLU (-1), HIS (+0.5), LYS (+1), and ARG (+1). The native contacts formed between residue pairs are modeled using a modified 12-10-6 Lennard-Jones potential with a desolvation penalty and the interaction strength is proportional to the

statistical potential (i, j) of the Miyazawa-Jernigan matrix (13). The KB model and its variant EIKB model were used throughout this dissertation.

$$U_{bonded} = \sum_{bonds}^{N-1} K_b(r_i - r_0)^2 + \sum_{angles}^{N-2} K_\theta(\theta_i - \theta_0)^2 + \sum_{dihedrals}^{N-3} \sum_{n}^{4} K_\phi(1 + \cos(n\theta - \theta_0)) \quad (1.1)$$

$$U_{unbonded} = \sum_{\substack{native \\ |i-j|>3}} \varepsilon \left\{ 13\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 18\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} + 4\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} \right\} + \sum_{\substack{non-native \\ |i-j|>3}} \varepsilon \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} \quad (1.2)$$

$$U_{elec} = \sum_{i>j} \frac{332 q_1 q_2}{D r_{ij}} e^{-\frac{r_{ij}}{\kappa}} \quad (1.3)$$

Chapter 2 focuses on the application of the KB model in the folding mechanism of a TIM barrel protein. Triosephosphate isomerase (TIM) barrel proteins have not only a conserved architecture that supports a myriad of enzymatic functions, but also a conserved folding mechanism that involves on- and off-pathway intermediates. Although experiments have proven to be invaluable in defining the folding free energy surface, they only provide a limited understanding of the structures of the partially folded states that appear during folding. Coarse-grained simulations employing native centric models are capable of sampling the entire energy landscape of TIM barrels and offer the possibility of a molecular-level understanding of the readout from sequence to structure. In this chapter, sequence-sensitive native centric simulations with small angle X-ray scattering and time-resolved FRET were used to monitor the formation of structure in an intermediate in the sulfollobus solfataricus indole-3-glycerol phosphate synthase TIM barrel that appears within 50 μs and must at least partially unfold to achieve productive folding. Simulations reveal the presence of a major and two minor folding channels not detected in experiments. Frustration in folding, i.e., backtracking in native contacts, is observed in the major channel at the initial stage of folding, as well as late in folding in a minor channel prior to the appearance of the native conformation. Similarities in global and pairwise dimensions of the early intermediate, the formation of structure in the central region that spreads progressively towards each terminus, and

a similar rate-limiting step in the closing of the β-barrel underscore the value of combining simulation and experiment to unravel complex folding mechanisms at the molecular level.

Chapter 3 focuses on the development of an enhanced sampling method to study protein-protein interactions in the context of transcription regulation. Allosteric regulation by intrinsically disordered proteins (IDPs) is an important class of cellular processes, including transcription. Molecular dynamics (MD) simulation is a promising approach to unravel the complex molecular interactions involved in the allosteric regulation by IDPs. While allosteric regulation is often characterized by the effect of a ligand on the binding affinity of a distal ligand, the binding affinity is often challenging to calculate by MD simulations due to insufficient sampling of the rare events in this binding/unbinding process. In this chapter, I present a new sampling approach based on Hamiltonian replica exchange (HREX) that allows accurate and efficient calculation of binding affinities using a native-centric coarse-grained model. I also demonstrate the utility of the new method by studying the positive allostery associated with the kinase-inducible domain interacting (KIX) domain of the CREB binding protein (CBP), in which a pre-bound ligand enhances the binding of the second ligand. The simulations reaffirm the reduced-entropy mechanism of the cooperative allosteric effect in KIX in which the prebound ligand reduces the entropic cost for the second ligand to bind.

Chapter 4 focuses on the application of the HREX method developed in Chapter 3 to the negative allosteric regulation in a disordered protein switch. The transcriptional adaptor zinc-binding 1 (TAZ1) domain of the transcriptional coactivator CBP/P300 and two disordered peptides HIF-1α and CITED2 form a delicate protein switch that regulates cellular hypoxic response. In hypoxia, HIF-1α binds TAZ1 to control the transcription of adaptive genes critical for the recovery from hypoxic stress. CITED2 acts as the negative feedback regulator to rapidly displace HIF-1α and

efficiently attenuate the hypoxic response. Though CITED2 and HIF-1α have the same dissociation constant (K_d = 10nM) in their binary complexes with TAZ1, CITED2 is much more competitive than HIF-1α upon binding the same target TAZ1 in ternary (14). In this chapter, I demonstrate that a simple coarse-grained model can recapitulate this negative allosteric effect and provides detailed physical insights into the displacement mechanism. The long-range electrostatic forces were found to be essential for the efficient displacement of HIF-1α by CITED2. The strong electrostatic interactions between CITED2 and TAZ1, along with the unique binding mode, make CITED2 more competitive than HIF-1α in binding TAZ1.

Chapter 2 Frustration and Folding of a TIM Barrel Protein

This chapter has been published in the following paper:

Kevin T. Halloran[*]; Yanming Wang[*]; Karunesh Arora; Srinivas Chakravarthy; Thomas C. Irving; Osman Bilsel; Charles L. Brooks III; C. Robert Matthews; Frustration and Folding of a TIM Barrel Protein. *Proc. Natl. Acad. Sci.* **2019**, *116* (33), 16378–16383. ([*]these authors contributed equally to this work)

This chapter was a collaborative effort with the lab of Prof. C. Robert Matthews. The experiments were performed by Dr. Kevin T. Halloran, Dr. Srinivas Chakravarthy, and Prof. Osman Bilsel. The simulations were performed by myself and Dr. Karunesh Arora. All authors took part in the analysis of the data. Dr. Kevin T. Halloran, Prof. Osman Bilsel, Prof. Charles L. Brooks III, Prof. C. Robert Matthews and myself wrote the paper.

2.1 Introduction

The folding of globular proteins involves the formation of numerous noncovalent interactions as the polypeptide chain samples the folding free energy surface on its journey to the global free energy minimum. Given the complexity of the conformational transition, it is surprising that proteins execute their folding reactions within a few 10's of seconds on a relatively smooth energy surface (15). The folding reactions of small proteins and domains, <100 amino acids, usually follow a 2-state process with a single barrier between unfolded and native states that controls a simple exponential response. Larger proteins, however, often have more complex responses involving multiple exponential phases whose rate constants progressively decrease as they approach the native state (16). These phases have been attributed to the appearance of partially-

folded states whose role in folding may be to avoid aggregation, accelerate folding or simply be a consequence of the interplay between the sequence and topology of the protein (17). A complementary view of such intermediates posits that their presence reflects frustration in folding that precludes the direct formation of the native state by the topological incompatibility of preformed elements of structure (18, 19). Experiments and simulations have revealed that these intermediates may be native-like in secondary structure but contain non-native interactions (20) or contain structural elements not found in the native structure (21).

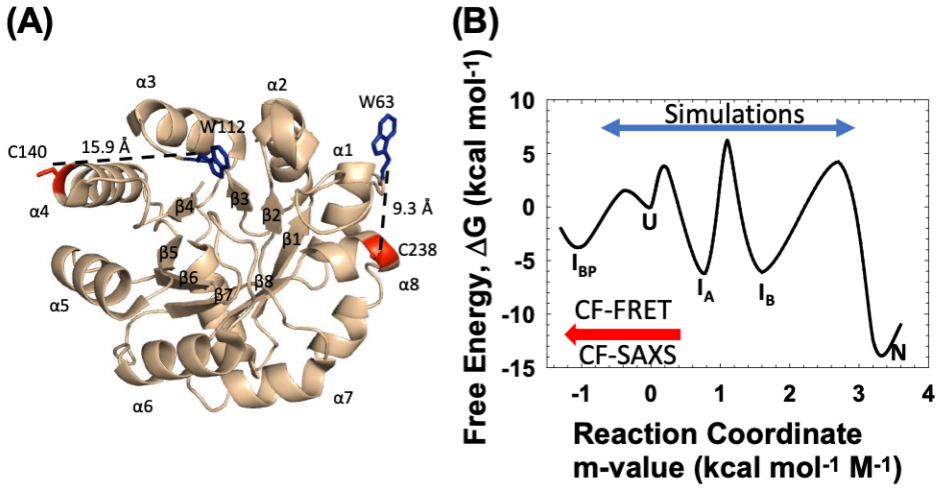

Figure 2.1 (A) A ribbon representation of the structure of SsIGPS (PDB code 2C3Z). The FRET pairs employed to study the central α_3-α_4 segment, W112-C140, (containing the strongly protected $(\beta\alpha)_{3\text{-}4}$ module) and the N- and C-termini, W63-C238, are highlighted with W112 and W63 residues in blue and the C140 and C238 residues in red. (B) The reaction diagram of SsIGPS (22). The barrier heights were estimated using the Kramer's formalism with a prefactor of 1 μs. The blue arrows indicate aspects of the free energy landscape probed in this study by simulations and the red arrow indicates the focus of the present experiments.

The triosephosphate isomerase (TIM) barrel family (Figure 2.1A) provides a rich example of complex folding reactions whose kinetic responses are largely conserved while the underlying sequences vary widely (23). Previous folding studies on several family members have consistently

found that folding comprised a sub-millisecond, burst phase followed by a phase whose relaxation time decreases with increasing final urea concentrations, a characteristic of an unfolding reaction but under refolding conditions. The escape from a misfolded, off-pathway intermediate is followed by the further acquisition of secondary structure and stability in an on-pathway intermediate(s) and the rate-limiting formation of the native state (23, 24). The kinetically trapped species could be frustration in folding whereby the burst phase species must unfold to enable productive folding to the native state. Mutational analysis and hydrogen exchange (HDX) experiments on several TIM barrel proteins have revealed a relationship between sequence and topology in the structures of the kinetically-trapped and on-pathway intermediates (23, 25, 26). Clusters of branched aliphatic side chains, isoleucine, leucine and valine, local in sequence and local in space, form water-resistant cores that stabilize these partially-folded forms (27). Although the kinetic species are conserved, the evolution of the sequences over time has resulted in alternative locations for the cores of stability in the TIM barrel architecture.

The mutational and HDX experiments are valuable in pinpointing the regions where structure appears in intermediates detected after the first few milliseconds of folding but leave unanswered questions about the crucial initial folding reaction. Recent advances in microfluidic mixers now enable access to folding events in the microsecond time range and have allowed us to examine the earliest events in the folding of a candidate TIM barrel protein (28). In the present study, pair-wise distance measurements from time-resolved FRET (trFRET) and global size and shape measurements from small angle X-ray scattering (SAXS), combined with coarse-grained computer simulations, are employed to probe the earliest events in the folding of the *S. solfataricus* indole-3-glycerol phosphate synthase (SsIGPS) TIM barrel. Surprisingly, the global collapse of the unfolded chain to a misfolded, off-pathway intermediate occurs within 50 μs. The simulations

reveal the potential complexities of this exceedingly rapid reaction and show that the rate-limiting step in the folding of SsIGPS is the frustration encountered by the competition between the N- and C-terminal β-strands to close the eight-stranded β-barrel.

2.2 Results

The SsIGPS TIM barrel is a representative member of the most common architecture for enzymes in biology. The 8 alternating β and α elements are arranged sequentially as a central parallel-stranded β-barrel encompassed by an α-helical shell (Figure 2.1A). Its folding mechanism has previously been shown to begin with the sub-millisecond formation of an off-pathway intermediate, I_{BP}, followed by two on pathway intermediates, I_A and I_B, before reaching the native state (Figure 2.1B). The I_{BP} intermediate has an apparent stability of 3.5 kcal·mol^{-1}, is rich in secondary structure and displays strong protection against exchange of amide hydrogens with solvent in the central $(\beta\alpha)_4$ module within 75 ms (22). As the folding reaction proceeds, the protection expands to encompass $(\beta\alpha)_{2-6}$ in I_A and $(\beta\alpha)_{1-8}$ in I_B. The fully folded TIM barrel appears in the final step of folding.

Measuring Global Dimensions by Small Angle X-ray Scattering (SAXS)

To obtain global insights into the structures of the intermediates, SAXS profiles were obtained under equilibrium and kinetic refolding conditions. At equilibrium, the native state of the protein has a radius of gyration (Rg) of approximately 18 Å, and the unfolded state has an estimated Rg of 46 Å, by linear extrapolation from high denaturant conditions (Figure 2.2A). The I_A intermediate, highly populated at 4 M urea (22), self-associates at the 80 μM protein concentration required for reliable SAXS measurements, precluding an estimate of its Rg (Figure 2.3).

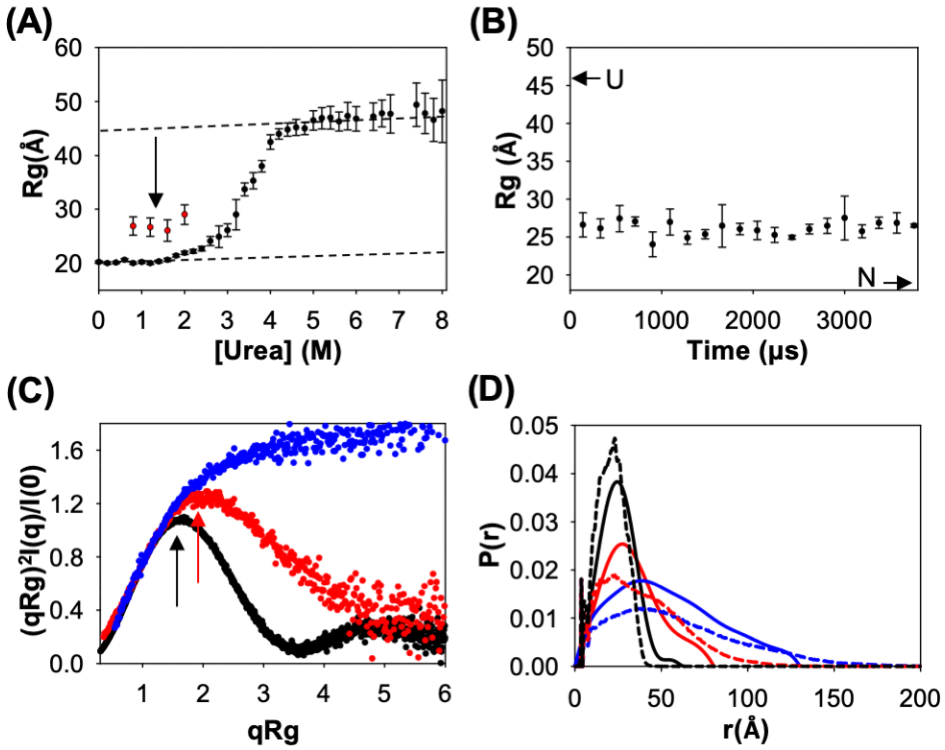

Figure 2.2 (A) Rg as a function of urea concentration for the unfolding of SsIGPS (black circles). The estimated Rg's of the native and unfolded states in water are indicated by linear extrapolation of the native and unfolded baselines. The estimated Rg after 150 µs of refolding at several final urea concentrations in the native baseline region (red circles). (B) Rg as a function of folding time at 0.8 M urea. The Rg's of the unfolded and native states in water are indicated. (C) Dimensionless Kratky plots of the unfolded (blue), I_{BP} intermediate (red) and native state (black). The arrows indicate the maxima in the plots for the N and I_{BP} species. (D) The P(r) of the unfolded (blue), I_{BP} intermediate (red) and native states (black). The dashed lines represent the P(r)'s for these states calculated from the simulations.

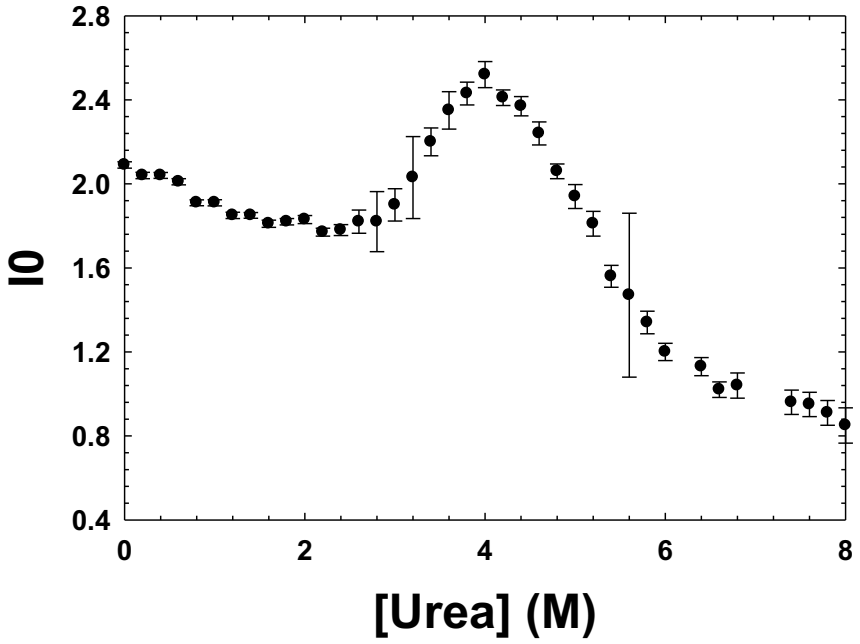

Figure 2.3 The zero-angle scattering intensity as a function of denaturant concentration. As the protein unfolds, the I0 is expected to decrease linearly, however, the deviation from 2 M through 4.5 M indicates that the intermediate is dimerizing.

The Rg of the sub-ms burst phase intermediate I_{BP} was determined by a ten-fold dilution from 8 M urea, using a custom, single piece microfluidic mixer. Within 150 μs, the dead time of the mixer, the Rg of SsIGPS is 26 ± 1.5 Å (Figure 2.2B). The absence of change in Rg out to 4 ms, where I_{BP} can be detected by stop-flow CD, shows that the I_{BP} intermediate appears within 150 μs. The conclusion that the SAXS detected burst-phase species is a discrete thermodynamic state, and not a collapsed form of the unfolded state, is supported by the observation of a Rg that is insensitive to the urea concentration up to 2 M urea (Figure 2.2A). A collapsed form of the unfolded state would have been expected to swell with increasing urea concentration (29).

Transformation of the scattering curve from native state of SsIGPS in 0.8 M urea to a dimensionless Kratky plot (30) shows the parabolic shape typical of globular structure. The maximum in the plot occurs at $(\sqrt{3}, 1.1)$ as expected for the Guinier approximation (Figure 2.2C)

(30). At 8 M urea, SsIGPS has an extended random coil-like structure with the expected hyperbolic plateau shape at high qRg. By contrast, the dimensionless Kratky plot from the continuous flow refolding jump from 8 to 0.8 M urea shows that the I_{BP} state has a peak shift on the qRg-axis to a approximately (2 ,1.25). This behavior deviates from the Guinier approximation and shows that the protein has regions that are not yet fully globular. The pair distribution function, P(r), for I_{BP} confirms a large collapse of the chain from U to I_{BP} within 150 μs (Figure 2.2D). The maximum distance between any two atoms, D_{max}, concomitantly decreases from 130 Å to 80 Å, and the significant shoulder at ~70 Å shows that I_{BP} is not fully globular.

Pair-wise Dimensional Analysis by Time Resolved FRET

To complement the global dimensional data obtained by SAXS, 2 sets of pair-wise distances were measured by time-resolved tryptophan-AEDANS Förster resonance energy transfer (trFRET) experiments on SsIGPS. One FRET pair was positioned to monitor barrel closure by measuring the distances between α_1 and α_8, W63-C/AEDANS238. The second pair was positioned to monitor the formation of the strongly protected $(\beta\alpha)_4$ module (23), by measuring the distance between α_3-α_4, W112 and C/AEDANS140.

The average Trp lifetimes for donor-only (DO) and donor-acceptor (DA) samples for both the α_1-α_8 and α_3-α_4 pairs at 8 M urea were identical, consistent with the absence of FRET in the unfolded state (Figures. 2.4 A, B). The continuous flow trFRET (CF-trFRET) data for the refolding of SsIGPS containing the α_1-α_8 pair shows a decrease to a non-native-like lifetime of 4.5 ns for the DO sample and 3.3 ns for the DA sample, within the dead time of the mixer (50 μs) (Figures 2.4 A, B). As was the case for Rg, there are no significant changes in lifetimes for both the DO and DA samples from ~50 μs out to ~1 ms. Similar behavior was observed for the α_3-α_4 FRET pair

during refolding jumps to 0.8 M urea, demonstrating a global collapse of unfolded SsIGPS within 50 μs.

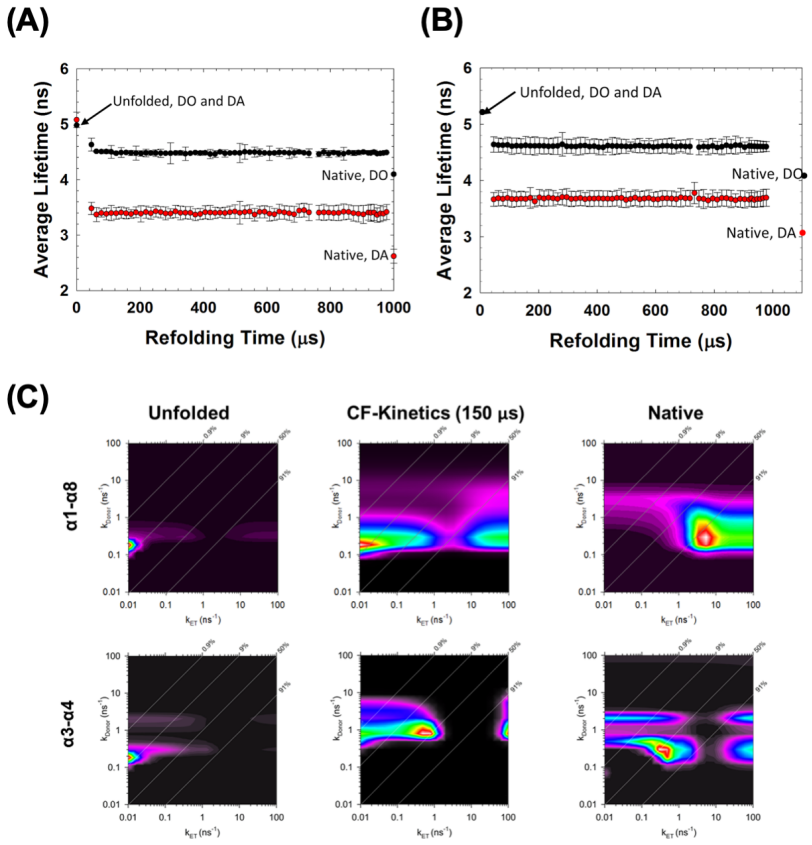

Figure 2.4 (A) The average Trp lifetimes for donor-only (black) and donor-acceptor (red) samples of the α_1-α_8 FRET pair and (B) the α_3-α_4 FRET pair after refolding 8.0 to 0.8 M urea. The average lifetimes for both DO and DA samples in their unfolded states at 8.0 M urea and in water are indicated. (C) Maximum Entropy Modeling (MEM) results for the unfolded states, continuous-flow kinetics (CF-Kinetics) at 150 μs, and the native states for both the α_1-α_8 and the α_3-α_4 FRET pairs. The grey lines represent 0.9%, 9%, 50%, and 91% FRET efficiency. The unfolded state for both pairs shows no significant FRET taking place. The CF-Kinetics for the α_1-α_8 pair shows both a low and high FRET state, while the α_3-α_4 pair has the major peak at ~50% FRET efficiency. The native state for the α_1-α_8 pair has a strong, >90%, FRET signal. For the α_3-α_4 pair, the native state has the peak at ~50% efficiency.

Maximum Entropy Modeling (MEM)

The trFRET data for both the α_1-α_8 and α_3-α_4 pairs were analyzed by two-dimensional maximum entropy modeling (2D-MEM) (31, 32) for the unfolded state (8 M urea), I_{BP} (0.8 M urea, 100 μs),

and the native state (0 M urea) (Figures 2.5 A, B and Figure 2.4C). The unfolded state for both FRET pairs show very small normalized amplitudes from 12 to 35 Å, the distances most sensitive to FRET for the Trp-AEDANS pair (R_0 = 22 Å). The maximum normalized amplitude in the native state for α_1-α_8 pair (Figure 2.5A) is ~13 Å, in good agreement with the calculated distance between residues 63 and 238 in the crystal structure, 9.3 Å. The maximum normalized amplitude in the native state for the α_3-α_4 FRET pair has a peak at ~19 Å, also in good agreement with the distance between Cα's of residues 112 and 140 in the crystal structure, 15.9 Å. Surprisingly, the normalized amplitude after 100 μs for the α_1-α_8 FRET pair revealed the presence of two distinct distributions of distances. One distribution of distances is more compact than native, and the other is more extended than native but more compact than the unfolded state. By contrast, the α_3-α_4 pair after 100 μs shows a single peak around 20 Å (Figure 2.5B) that is similar to the native protein but has a broader distribution.

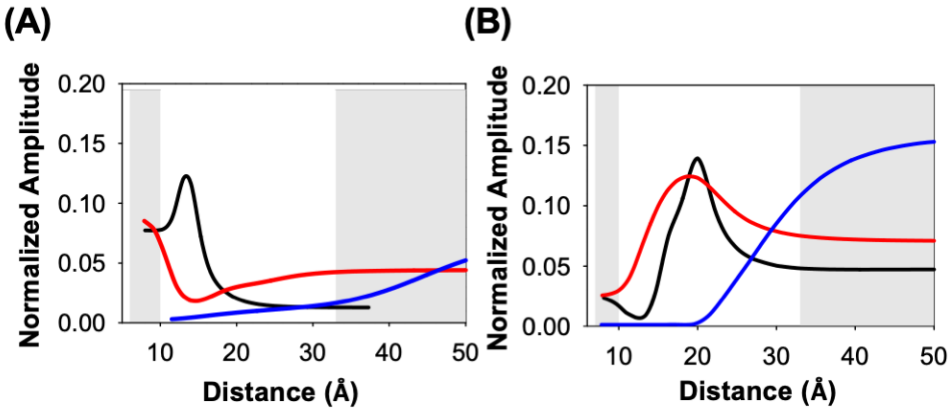

Figure 2.5 (A) The 1D conversion of the MEM analysis of the trFRET data for the α_1-α_8 pair to generate distance distributions for the unfolded state (blue), the native state (black), and the apparent pair of states appearing after 50 μs (red). (B) The 1D conversion of the MEM analysis of the trFRET data for the α_3-α_4 pair to generate distance distributions for the unfolded state (blue), the native state (black), and the two states appearing after 50 μs (red).

Ensemble Averaged Folding Properties from Simulations

To gain deeper insight into the development of structure during the folding of SsIGPS, we complemented the present and previous experimental studies (23, 33) with a native centric simulation to sample the entire folding landscape.

Native-centric coarse-grained Gō model (8) refolding simulations were initiated from an unfolded ensemble of structures sampled from simulations at high temperature and 100 independent 2,000 time-units (1 time-unit = 10,000 dynamic steps) folding trajectories were sampled in the analysis. Because the underlying model is coarse-grained, the landscape is smoother and the folding timescales are compressed, and thereby do not directly correspond to the times observed in experiments. However, we anticipate that the time ordering, as well as the relative lag times between folding phases should reflect what is observed in kinetic experiments (34). The progress of the folding reaction for each trajectory was monitored by the radius of gyration (Rg) and the fraction of total native contacts (Q_t). Over the time courses of the 100 trajectories, persistent values were observed for Q_t of 0.3, ~0.5, ~0.6, ~0.8, and 0.9 (Figure 2.6A). The initial (0.3) and final (0.9) values correspond to the unfolded and native forms of the protein, with the intermediate plateaus (~0.5, ~0.6, ~0.8) suggesting the presence of partially folded states. Examination of the entire set of trajectories revealed that only a small fraction of the simulations reached the native state within 2,000 time-units (Figure 2.7). The majority of the simulations reached a Q_t of 0.8. Plateaus at similar time steps were observed for Rg (Figure 2.6B), beginning with the unfolded state at ~45 Å and progressively decreasing to 18 Å for the native state.

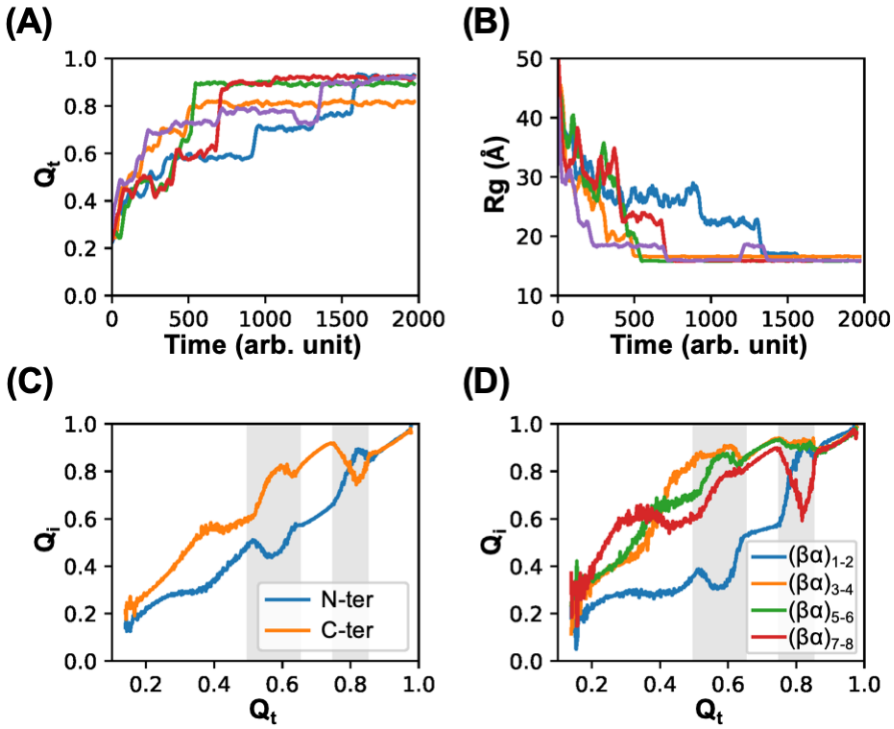

Figure 2.6 (A) Representative trajectories of the fraction of total native contacts Q_t. (B) Representative trajectories of Rg. For clarity, kinetic traces are shown as moving averages of 30 successive snapshots. The leveling off at various Rg and Q_t values indicates multiple intermediates are formed during the simulations. (C) Fraction of native contacts Q_i of the of the N- and C-terminal halves of the protein and (D) Q_i of the four βαβα modules as a function of Q_t. Decreases in the fraction of native contacts in panels C and D represents the backtracking that occurs during the simulations. There are two main backtracking events that involve the N and C-termini (Q_t = 0.50 to 0.65 and Q_t = 0.75 to 0.85).

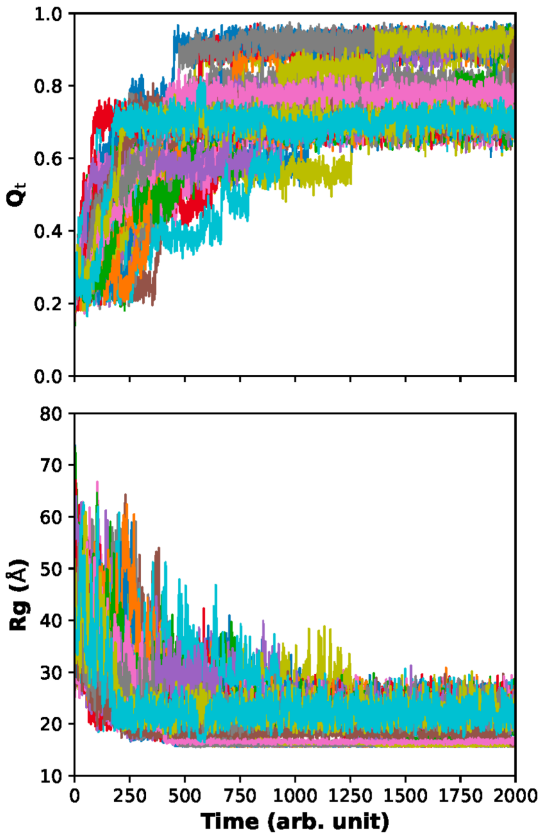

Figure 2.7 Protein folding trajectories of the fraction of total native contacts (Q_t) and radius of gyration (Rg). For clarity, only 50 out of the 100 trajectories were shown in the figure.

Decomposition of Q_t into contributions (Q_i) from the N- and C-terminal halves, $\alpha_0(\beta\alpha)_{1\text{-}4}$ and $(\beta\alpha)_{5\text{-}8}$, reveals frustration in folding (Figure 2.6C). The striking anti-correlated gain/loss in native contacts in N- and C-terminal halves was observed at $Q_t = 0.50$ to 0.65 and $Q_t = 0.75$ to 0.85, where the intermediate states persist, suggest the frustration in these two regions might be related to those intermediate states. Further decomposition of Q_t into the four $(\beta\alpha)_{i\text{-}(i+1)}$ modules of stability (35) pinpoint the major sources of frustration (Figure 2.6D). At $Q_t = 0.5$ the source of frustration derives from $(\beta\alpha)_{1\text{-}2}$ competing with $(\beta\alpha)_{7\text{-}8}$ and to a lesser extent $(\beta\alpha)_{5\text{-}6}$. The frustration event at $Q_t = 0.6$ is $(\beta\alpha)_{1\text{-}2}$ driving folding while $(\beta\alpha)_{3\text{-}4}$ and $(\beta\alpha)_{5\text{-}6}$ lose native contacts. The final frustration event

at Q_t = 0.8, is mainly the competition between the $(\beta\alpha)_{1\text{-}2}$ and $(\beta\alpha)_{7\text{-}8}$ modules; however, the competition also effects $(\beta\alpha)_{3\text{-}4}$ and $(\beta\alpha)_{5\text{-}6}$.

An examination of the contact probability maps at different times gives further insight into the folding mechanism (Figure 2.8). The central region (residue ~90 to residue ~180) formed most of its contacts within 400 time-units compared to the total simulation time of 2,000 time-units. Then, more contacts were formed in the C-terminal region within 1000 time-units. At the end of the simulation, most of the low-probability contacts were those formed between $\alpha_0\beta_1$ (residues ~1 to ~40) and other regions, suggesting many trajectories ended up with a structure with unfolded $\alpha_0\beta_1$.

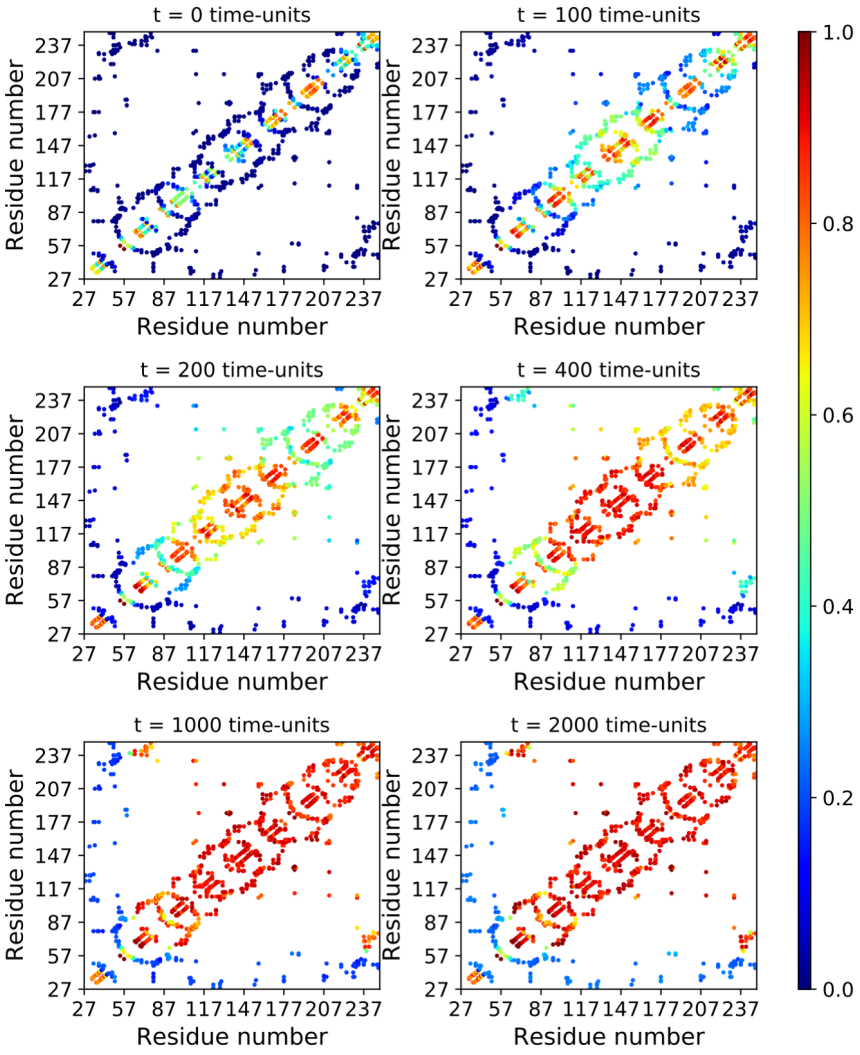

Figure 2.8 Contact probability maps at different times with red colors indicating high probabilities of forming contacts while blue colors indicating low probabilities. The contacts are calculated from the native structure and are the same as those used in the Gō model. Each contact dot represents the probability of forming a contact averaged over 100 trajectories at the given time.

Multiple Folding Pathways Revealed from Simulations

To obtain further insights into the molecular events that occur during the folding of SsIGPS, we examined each trajectory in detail. Three significant folding pathways were found, based on the assembly order of secondary structural units (Figure 2.9). The classification of different states

found in all trajectories was based upon both their distinct Q_t and Rg values (Table 2.1) and their visual differences in structure (Figure 2.9 and Figure 2.10).

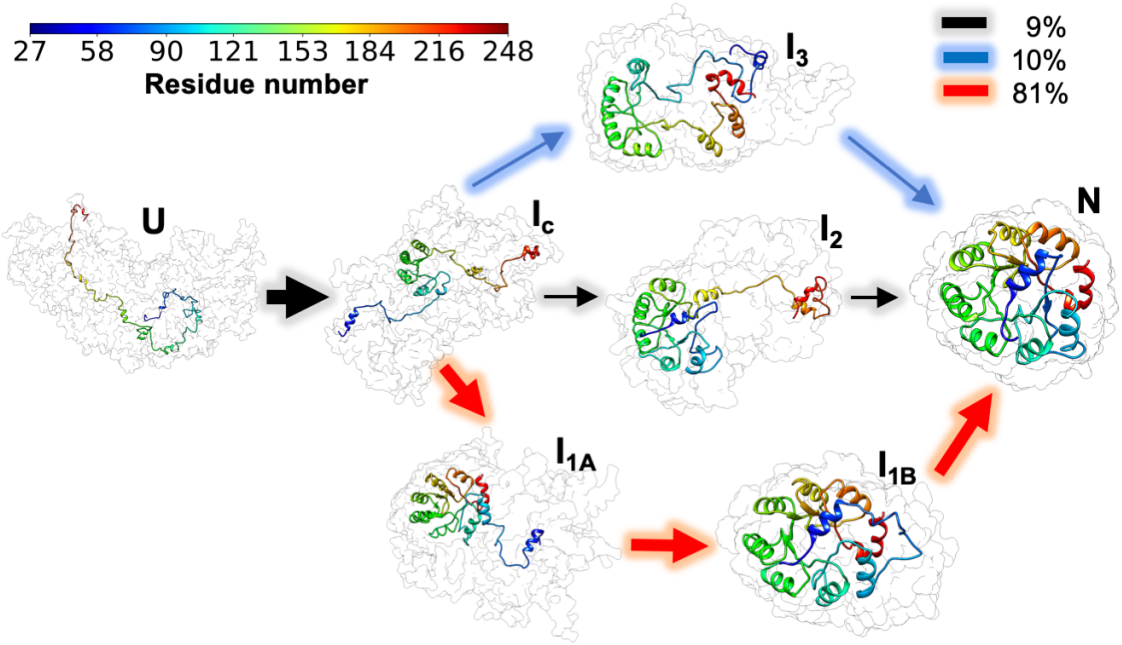

Figure 2.9 Multiple folding pathways discovered by simulations, the upper right legend shows the transition probabilities from I_c to I_{1A}, I_2, and I_3. Additional structural details for I_{1A}, I_{1B}, I_2, and I_3 are shown in Figure 2.10. The gray contours show the overlay of approximately 50 protein conformations, sampled from the corresponding states.

Table 2.1 Fraction of total native contacts Q_t, radius of gyration Rg and their associated standard deviations of all states in simulations. All states are listed sequentially form left to right with increasing Q_t.

State	U	I_c	I_2	I_1	I_{1A}	I_{1B}	N
Q_t	0.313	0.536	0.698	0.714	0.788	0.847	0.996
Std (Q_t)	0.084	0.039	0.046	0.040	0.011	0.026	0.007
Rg (Å)	43.22	31.84	22.07	24.21	22.18	16.96	15.87
Std (Rg)	9.55	4.72	1.96	3.58	2.07	0.33	0.12

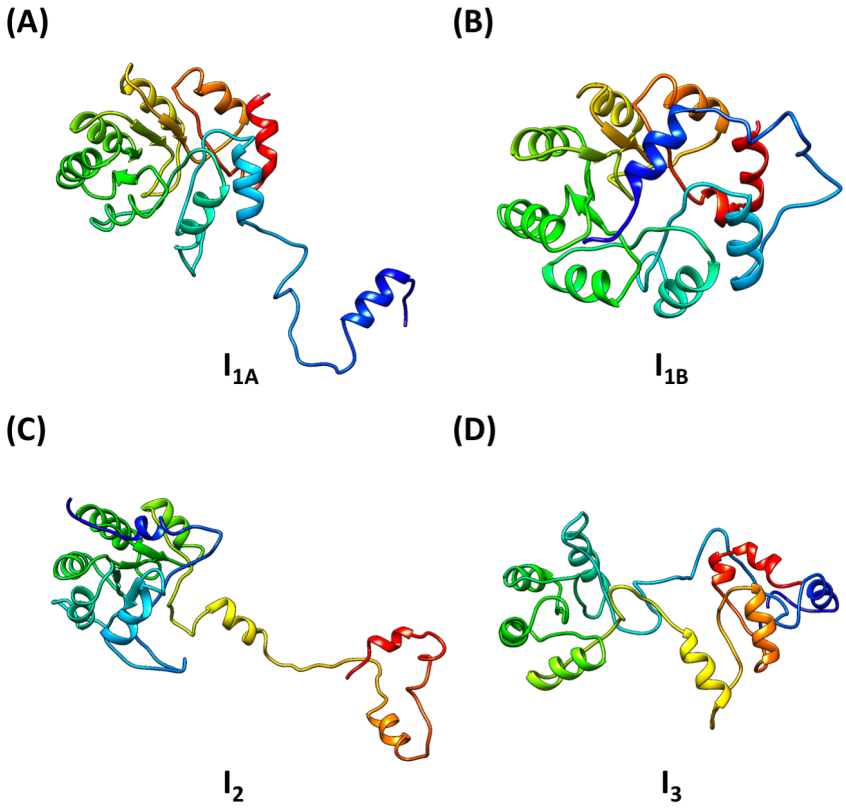

Figure 2.10 Intermediate states I_{1A}, I_{1B}, I_2 and I_3 found in the three folding pathways.

The unfolded state (U) initially collapses to a single intermediate, I_c, with a well-folded central region (residues 75-175, $(\beta\alpha)_{2\text{-}5}$). The I_c state then partitions into the I_{1A} state, the I_2 state or the I_3

state with transition probabilities of 81%, 9%, and 10% respectively, entering three separate folding channels.

The I_1 pathway is characterized by the formation of an extremely stable I_{1A} state after I_c. The I_c state transited to I_{1A} by spreading its folded structure from $(\beta\alpha)_{2\text{-}5}$ to $(\beta\alpha)_{2\text{-}8}$, leaving an unfolded α_0 tethered by β_1. The I_{1A} state has a 7-stranded barrel, with native-like contacts between helices α_1 and α_8 that prevent the incorporation of α_0 and β_1 into the barrel architecture. I_{1A} persists in the great majority of the trajectories, with only a small fraction escaping to dock α_0 across the bottom of the barrel to form the I_{1B} state. The I_{1B} state has an unfolded β_1 with two of its ends fixed on the folded β-barrel. I_{1B} then rapidly folds to the native state by the insertion of β_1 into the β-barrel through the channel between α_1 and α_8. To test the stability of the I_{1A} state, we further sampled another set of 100 trajectories beginning from the I_{1A} state for 8000 time-units. Even with a quadrupled simulation time, only 17% of the trajectories reached the native state, confirming the extremely long lifetime of the I_{1A} state.

In comparison to the I_1 pathway, in which $\alpha_0\beta_1$ is the last to fold, both the I_2 and I_3 pathways require the $\alpha_0\beta_1$ element to fold before the closure of the β-barrel. In the I_2 pathway, I_c incorporates the $\alpha_0\beta_1$ element but excludes the C-terminal $(\alpha\beta)_{7\text{-}8}$ elements to form the I_2 state. I_2 then readily folds to the native state by incorporating the C-terminal $(\alpha\beta)_{7\text{-}8}$. In the I_3 pathway, I_c first forms two partially folded $\alpha_2(\alpha\beta)_{3\text{-}5}$ and $\alpha_6(\alpha\beta)_{7\text{-}8}$ before the docking of $\alpha_0\beta_1$ on $\alpha_6(\alpha\beta)_{7\text{-}8}$ to form the I_3 state. The I_3 state has two partially folded halves of the β-barrel, the $\alpha_2(\alpha\beta)_{3\text{-}5}$ subdomain and the $(\alpha\beta)_0+\alpha_6(\alpha\beta)_{7\text{-}8}$ subdomain, linked by unfolded $\beta_1\alpha_1\beta_2$ and β_6 strands. Interestingly, the I_3 state, fully connected by contacts from head to tail, then folds to native by the cooperatively merging the two partially folded halves. More structural details of the I_{1A}, I_{1B}, I_2, and I_3 intermediates are shown in Figure 2.10.

To examine the relationship between the two regions of frustration (Q_t = 0.50 to 0.65 and Q_t = 0.75 to 0.85) found in the ensemble averaged analysis (Figure 2.6) and the three folding pathways, the Q_i vs. Q_t data for the N- and C-terminal halves of SsIGPS of the three folding pathways were compared with the data of all trajectories, as shown in Figure 2.11. The Q_t = 0.50 to 0.65 and Q_t = 0.75 to 0.85 regions, representing the early and final folding stages respectively, differ in their sources of frustration. The great similarity of the Q_i vs. Q_t plots at Q_t = 0.50 to 0.65 between all trajectories and the I_1 pathway (Figures 2.11 A and B) indicates the global frustration at Q_t = 0.50 to 0.65 is primarily contributed by the frustration in the major I_1 channel.

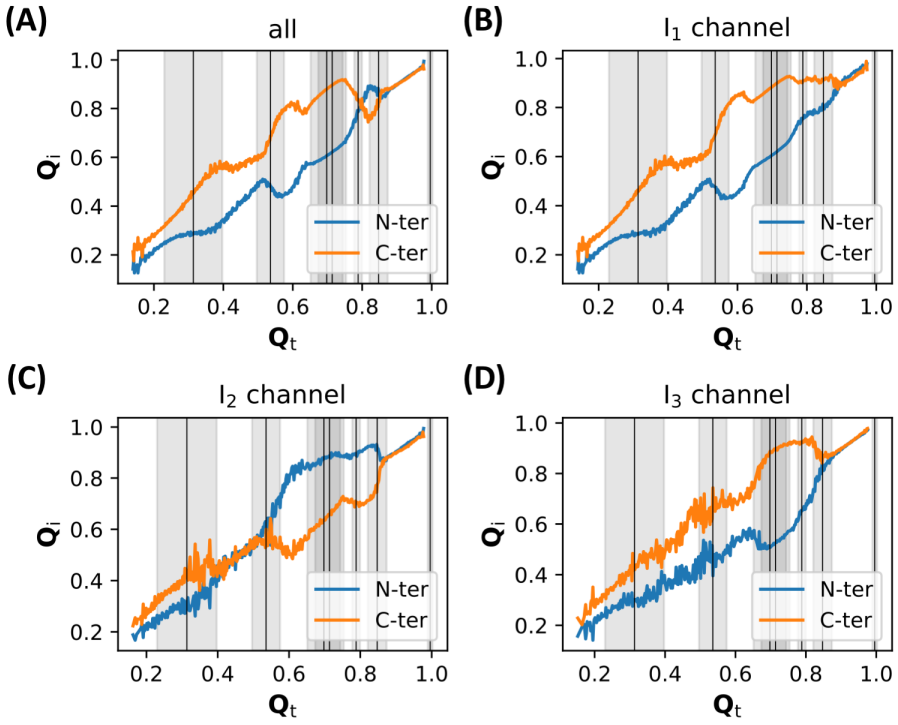

Figure 2.11 Fraction of native contacts Q_i of N- and C-terminal halves as a function of Q_t of all trajectories (A), trajectories of the I_1 pathway (B), trajectories of the I_2 pathway (C), and trajectories of the I_3 pathway (D). The average Q_t of all states U, I_c, I_3, I_2, I_{1A}, I_{1B}, and N are shown sequentially from left to right as vertical lines with shaded area indicating plus/minus one standard deviation.

However, the global frustration at Q_t = 0.75 to 0.85 is different, since no significant backtracking events were observed in the same region of all three folding channels (Figures 2.11 B, C, and D). The three folding channels, differing in their assembly order of the protein, have very different values of $Q_{C\text{-}ter}$ and $Q_{N\text{-}ter}$ at different times. In the I_1 and I_2 channels, the C-terminal $(\alpha\beta)_{7\text{-}8}$ and the N-terminal $\alpha_0\beta_1$ are the last to fold, respectively. Before reaching the final stage of folding at Q_t = 0.75 to 0.85, the $Q_{C\text{-}ter}$ of the I_1 channel reaches ~0.9, which is significantly larger than that of the I_2 channel (~0.7). During most of the 2000 time-units the I_1 pathway trajectories (74 out of 81) were trapped in the extremely stable I_1 state ($Q_t \approx 0.788$), which precludes sufficient sampling at $Q_t > 0.788$ in the I_1 channel and makes the I_2 and I_3 channels dominant in this region. Therefore, the backtracking of $Q_{C\text{-}ter}$ at Q_t = 0.75 to 0.85 is a result of the significantly smaller $Q_{C\text{-}ter}$ of the I_2 pathway and the sparse sampling of the I_1 pathway in this region.

2.3 Discussion

A combined experimental and computational study of the folding reaction of the SsIGPS TIM barrel has revealed insights into the structures of partially folded states and the potential role of frustration that occurs in simulations of the folding reaction.

Mechanistic Analysis

Previous experimental studies of the folding kinetics generated a 5 species model, $I_{BP} \rightleftharpoons U \rightleftharpoons I_A \rightleftharpoons I_B \rightleftharpoons N$ (Figure 12) (26). Current native-centric simulations predict a more complex model involving partitioning between 3 different pathways to reach the native conformation (Figure 9). The data from both models can be used to generate “kinetic species” plots to compare the flow of material from the U state to the N state during a folding reaction (Figure 12).

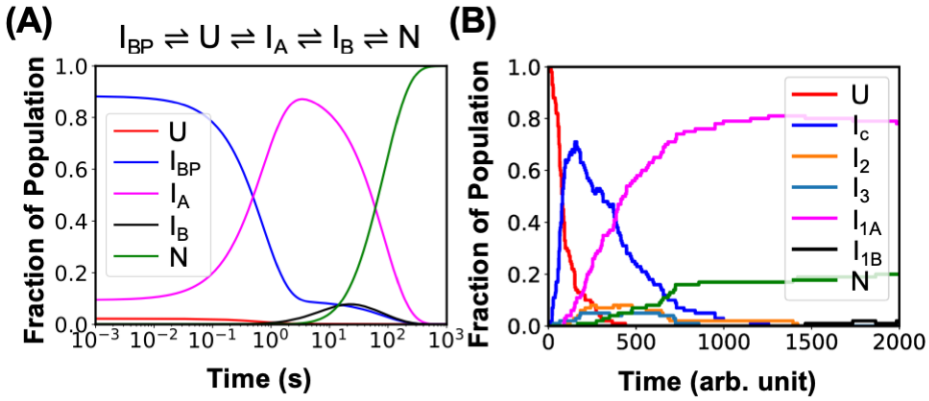

Figure 2.12 Experimental (A) and simulated (B) populations of states vs. time.

The two kinetic species plots are remarkably similar in several respects, but obviously differ with respect to the number of folding channels. The CF-FL and CF-SAXS measurements report the collapse of the unfolded chains within 50 μs, however, the distance distribution of the α_1-α_8 FRET pair (Figure 2.5A) indicates the presence of two states. One state is more compact than the native state, implying a non-native structure, and the other more expanded than native but more compact than the unfolded state. Because the experimental kinetic species plot (Figure 2.12A) predicts the simultaneous presence of the I_{BP} and I_A states after a few ms, with the I_{BP} state predominant, we presume that the overly compact state corresponds to I_A and the expanded state to I_{BP}. Examination of the predictions of the simulations after ~200 time-units (Figure 2.12B) supports this interpretation when comparing the populations of the I_c and I_{1A} states. The subsequent increase in the population of the I_A state in experiments is also mimicked by the I_{1A} state in the simulations. Particularly striking is the correspondence of the very long lifetimes of both the I_A and I_{1A} states, consistent with their rate-limiting roles in folding by both experiment and simulations. Both experiments and the major refolding channel in the simulations then reveal a final intermediate, I_B and I_{1B}, respectively, before proceeding to the native state. The partitioning of I_c into 3 channels

in the simulations is not evident in the experimental data. However, Channels 2 and 3 each carry only ~10% of the population and would be difficult to detect experimentally. Both experiments and the major refolding channel in the simulations then reveal a final intermediate, I_B and I_{1B}, respectively, before proceeding to the N state.

Structural Analysis

The pairwise and global dimensional analysis provided by CF-trFRET and CF-SAXS enables a direct comparison with the results of the simulations on the structures of the unfolded state, U, and the I_{BP}/I_c intermediates that appear in microseconds.

U State

SAXS measurements of the unfolded state in high concentrations of urea (Figure 2.2A), when extrapolated to the absence of denaturant, yield an estimated Rg in water, 46 ± 5 Å, that is consistent with a random-coil ensemble for a chain of 226 amino acids (36). Remarkably, native-centric simulations of the U state (Figure 2.6B) obtained the same estimate of Rg, ~45 Å, and both approaches revealed the breadth of the unfolded manifold of conformers (Figures 2.2D and 2.6B).

I_{BP}/I_c State

The trFRET data for the α_1-α_8 pair show that the microsecond folding reaction partitions into two distinct distributions, with different degrees of contraction (Figure 2.5A). One ensemble is more compact than that for the N state and the other much more expanded. Unfortunately, the limitations of FRET measurements of distance outside efficiencies of 0.2 to 0.8 preclude estimates of the relative populations of these distributions. As described above, these results are consistent with the previous global analysis of the folding of SsIGPS that found U partitioning into the I_{BP} and I_A states (Figure 2.12A). Although the α_1-α_8 pair distances of the I_c state from simulations do not capture the overly compact conformation that appear after 50 μs (Figure 2.5A), native centric

simulations are incapable of detecting non-native structures. As the time steps increase, more compact states appear at ~20 and ~10 Å, reflecting the progression of the folding reaction towards the I_{1A}, I_{1B} and N states. The α_3-α_4 FRET pair show a single distribution centered near that for the N state (Figure 2.5B), but whose greater breadth indicates a larger, dynamic ensemble. The SAXS data reveal a denaturant-independent Rg of 26 Å below 2 M urea after 150 μs (Figure 2.2A), demonstrating that this specie(s) is not a collapsed form of the unfolded state (29). The Kratky plot after 150 μs (Figure 2.2C) is not consistent with a globular structure, and the associated P(r) (Figure 2.2D) shows a peak at ~30 Å and a tail at longer distances.

The simulations show a remarkable degree of correspondence with experimental results for the structures that appear at the initial stage of folding for SsIGPS. The I_c intermediate has native-like structure in the $(\beta\alpha)_{2\text{-}5}$ segment, consistent with the formation of secondary structure detected by previous HDX-MS experiments (23, 26) and the distance measured for the α_3-α_4 FRET pair. The N- and C-terminal segments are unstructured and give rise to the tail at high values of the P(r) (Figure 2.2D), very similar to that seen in the SAXS data.

I_A/(I_{1A} and I_2)

Although the present experimental study does not address the I_A intermediate, a previous HDX-MS study (23) found strong protection against exchange in the $(\beta\alpha)_{2\text{-}6}\beta_7$ region and a lack of protection in β_1 and β_8. As described above, the simulations show partitioning after the formation of I_c (Figure 2.9). The dominant I_1 pathway involves the formation of a 7-stranded β-barrel, $\alpha_1(\beta\alpha)_{2\text{-}8}$, by locking out β_1 and α_0. The minor I_2 pathway formed a 6-stranded β-barrel by excluding $(\beta\alpha)_{7\text{-}8}$. In both cases, the nascent barrels appear to be stabilized by native-like interactions between α_1 and α_8. The exclusion of the N- and C-terminal β-strands would expose them to solvent and explain in the lack of protection against HDX exchange.

I_B

The final intermediate in the experimental folding mechanism, I_B, has a fully-formed β-barrel, providing protection against HDX in all 8 β-strands (23). This species appears after the rate-limiting step in folding and is only transiently populated as a minor species before the appearance of the N state (Figure 2.12A). The simulations differ in that the I_{1B} state in Channel 3 excludes β_1.

Frustration in folding

Two major regions of topological frustration were found in the ensemble averaged analysis of the simulation, shown in the Q_i vs. Q_t plots (Figure 2.6). Multiple asynchronous folding pathways complicate the descriptions of the frustration in folding.

Frustration at Q_t = 0.50 to 0.65

The frustration at Q_t = 0.50 to 0.65, where the I_{BP}/I_c state persists, is mainly contributed by the backtracking events in the dominate I_1 channel. The backtracking event of the $Q_{N\text{-}ter}$ at Q_t = 0.50 (Figure 2.6C), corresponds primarily to the unfolding of the $(\beta\alpha)_{1\text{-}2}$ element (Figure 2.6D). This conclusion is consistent with experimental results in which some premature structures in the I_{BP}/I_c state are required to unfold before reaching the productive folding pathway. The two minor pathways I_2 and I_3 show different outcomes of the I_{BP}/I_c state in this region. The I_2 pathway shows backtracking in the C-terminus while the I_3 pathway shows no obvious frustration. Interestingly, the different sources of frustration in the I_1 and I_2 pathways reflect alternative forms of an incomplete TIM barrel. The major I_1 pathway excludes the N-terminus while the minor I_2 pathway excludes the C-terminus. Both contain the central $(\beta\alpha)_{2\text{-}6}$ region that is protected against HDX (26) and, evidently, is capable of propagating structure in either direction.

Frustration at Q_t = 0.75 to 0.85

The frustration at Q_t = 0.75 to 0.85 is a combined result of the three folding pathways that differ in their assembly order of the protein, and therefore does not represent actual loss of structures. The global backtracking of $Q_{C\text{-}ter}$ at Q_t = 0.75 to 0.85 is mainly caused by the I_2 channel, in which the C-terminal $(\alpha\beta)_{7\text{-}8}$ elements are the last to fold that lowers the global $Q_{C\text{-}ter}$ value. Although no evidence of backtracking of the I_{1A} state ($Q_t \approx 0.79$) was found, it remains possible that I_{1A} may first partly unfolds, so that $\alpha_0\beta_1$ folds before the barrel closure to reduce the tremendous entropic cost required to make the transition from I_{1A} to I_{1B}.

2.4 Conclusions

A combined experimental and computational study of the folding reaction for a TIM barrel protein has yielded remarkable agreement between their complementary views of a complex process. The mechanism defined by the major refolding pathway in simulations agrees closely with the mechanism determined by a variety of experiments. Striking similarities include the formation of stable structure in the central region of the sequence early in folding and a rate-limiting step prior to the formation of an 8-stranded β-barrel. Global and pairwise distance measurements of the early intermediate find a very similar degree of compactness, likely with disordered tails at both termini. In contrast to the experiments, the simulations reveal the presence of two minor channels that delineate alternative pathways to the native conformation. The frustration in folding detected by the simulations result in the formation of a pair of nascent TIM barrels that differ in the exclusion of either the N- or C-terminal segments of the protein, consistent with the presence of intermediates observed in experiments. Although it is likely that these incomplete barrels contain non-native structures inaccessible to native centric simulations, the remarkable similarities in the minima on the experimental and computational folding free energy surfaces argue that they are dominated by native-like structures.

2.5 Methods

Molecular Dynamics Protocol and Data Analysis

The Cα-based native centric coarse-grained model (see chapter 1 for more details) of SsIGPS was generated by an in-house script using Protein Data Bank (PDB) structure 2C3Z. Molecular dynamics simulations were performed using the CHARMM package (37). Langevin dynamics was used to propagate the equation of motion with a friction coefficient of 1.36 ps^{-1} and a time-step of 22 fs. Each snapshot was recorded every 100,000 time-steps (1 time-unit). The refolding simulations sampled one hundred 2,000 time-units trajectories at 230 K initiated from unfolded conformations generated from 2 time-units short simulations at 510 K. Two order parameters, the fraction of native contacts (Q) and the radius of gyration (Rg) were used as order parameters for the ensemble averaged analysis. Each native contact was considered formed if the residue pair was within a cutoff distance chosen such that the given contact was satisfied 85% of the time in native-state simulations at 300 K. The three folding pathways and intermediates were discovered through a combined analysis of the Q/Rg vs. time plots and the visualization of the trajectory using VMD (38). The trajectories were first projected onto the Q and Rg order parameters to help identify transitions between different states (e.x. some transitions are clearly visible in Figures 2.6 A and B). The timestamps when transitions occur on the Q/Rg vs. time plots were recorded and then visually examined in each trajectory. The states that each frame corresponds to were labeled based on those transitions.

Protein Production

Protein was expressed in DE3 cells and purified using metal affinity, ion exchange, and sizing chromatography before labeling with IAEDANS. See SI Appendix for full details.

Fluorescence

Details of the TCSPC apparatus equipped with a microsecond continuous-flow mixer (Translume) have been described previously (32) and are discussed more completely in the SI Appendix.

MEM

The 2D-MEM is used to obtain a distance distribution without any assumptions as to the number of lifetime components in the donor-only excited state decay, the shape of the distance distribution or the number of sub-populations. Details of the mathematical framework used in this analysis is given in the SI Appendix.

Small angle X-ray scattering

Small-angle X-ray scattering measurements were performed at the BioCAT beamline at the Advanced Photon Source, Argonne, IL. Equilibrium SAXS measurements were performed by interfacing an autosampler running custom software to the standard quartz sample capillary (39). Kinetic experiments were performed as previously described (39) with the exception that flow to the quartz mixer for the kinetic experiments was controlled by syringe pumps (Harvard Apparatus) at a total flow rate of 4 to 5 $ml \cdot min^{-1}$. Scattering images were reduced using scripts provided by BioCAT and analyzed as previously described (32, 39).

Chapter 3 Enhanced Sampling Applied to Modeling Allosteric Regulation in Transcription

This chapter has been published in the following paper:

Yanming Wang; Charles L. Brooks III; Enhanced Sampling Applied to Modeling Allosteric Regulation in Transcription. *J. Phys. Chem. Lett.* **2019**, *10* (19), 5963–5968.

3.1 Introduction

Allosteric regulation is an important cellular process, in which the binding of the first ligand on the target protein affects the second ligand at a distal site (40). Allosteric regulation by intrinsically disordered proteins (IDPs) is fundamental for cellular signaling and regulation, such as gene transcription, and is often involved in many human diseases (41). IDPs lack tertiary structure in the unbound state and fold upon binding a protein (42). The coupled folding and binding property of IDPs facilitates high binding specificity with low binding affinity, enabling IDPs to bind multiple targets (43). Understanding the allosteric mechanism by IDPs is essential for the rational design of drugs targeting these processes (44, 45). However, the allosteric mechanism, including the thermodynamics, remains poorly understood.

Molecular dynamics (MD) simulation is a powerful tool to help elucidate molecular-level insights into the mechanism of allosteric regulation by IDPs. Recently, MD simulation using a coarse-grained model successfully recapitulated the coupled folding and binding (10) and the cooperative allosteric effect (11) of IDPs. In this study, we focus on the cooperative allosteric effect of two IDP peptides binding to the same target protein, in which the binding of the first ligand enhances

the binding affinity of the second ligand. Therefore, it is essential to accurately calculate the dissociation constant (K_d) to characterize the allosteric effect quantitatively. One challenge is that the calculation of K_d is computationally expensive for MD simulations even using coarse-grained models, due to the extensive sampling needed to observe multiple long-timescale binding/unbinding events. In this work, we first present a new enhanced sampling method based on Hamiltonian replica exchange (HREX) that effectively addresses this issue. The new method allows accurate and efficient calculation of K_d using an electrostatics inclusive (EI) coarse-grained model (10) related to that originally developed by Karanicolas and Brooks (8) (EIKB model). We then demonstrate the utility of the new method by a case study of the cooperative allostery in the kinase-inducible domain interacting (KIX) domain.

KIX is a globular domain of the transcriptional coactivator CREB binding protein (CBP) and its paralogue P300 (46). Previous studies have shown KIX is vital for transcriptional activity in cells and is a potential target for drug design (47). KIX has a three-helix bundle structure that simultaneously binds two peptides on its two opposite binding surfaces, of which one can bind c-Myb/pKID and the other binds MLL (Figure 3.1). The binding of c-Myb/pKID on KIX increases the binding affinity of MLL and vice versa, which is a cooperative allosteric effect (48). The allosteric effect in KIX has been extensively studied both experimentally and theoretically (11, 49–51), and therefore is ideal for testing and validating the new method.

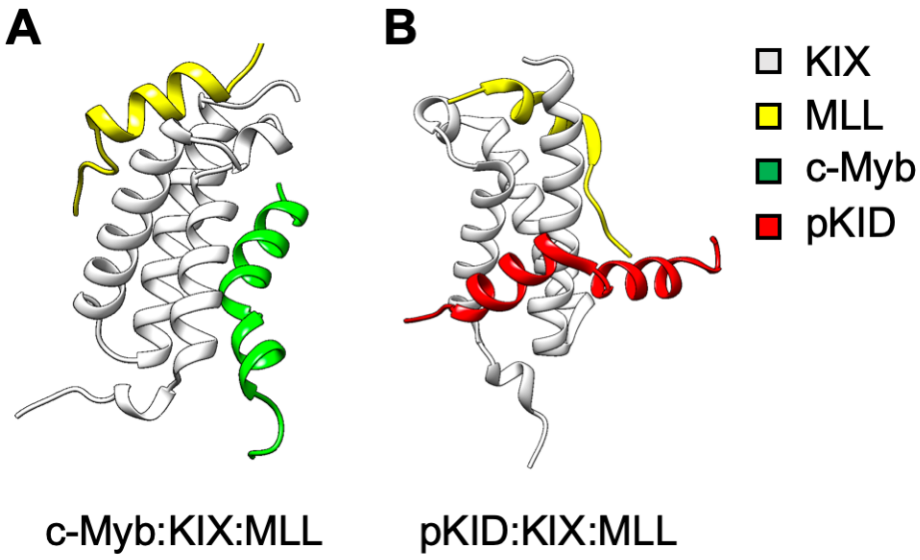

Figure 3.1 Ribbon diagram of the c-Myb:KIX:MLL ternary complex (panel A, PDB code: 2AGH) and the pKID:KIX:MLL ternary complex (panel B, PDB code: 2LXT).

3.2 Methods

Model Setup

Two systems c-Myb:KIX_c:MLL_c and pKID:KIX_p:MLL_p were first built using the EIKB model. This coarse-grained model uses a single bead to represent an amino acid residue. In this model, intermolecular interactions mainly consider short-range native contacts, and long-range electrostatic interactions are represented by a simple Debye-Hückel potential, similar to earlier work (12). The default dielectric constant (D) that controls the strength of electrostatic interactions is chosen to be 40 unless otherwise specified, and our data suggest that by setting D = 40 the EIKB model gives the closest agreement with experiment when electrostatics plays a role in binding, e.g., in pKID binding to KIX:MLL in the present study (see below). For the c-Myb:KIX_c:MLL_c and pKID:KIX_p:MLL_p systems, we use subscripts ‘c’ and ‘p’ on “KIX” and “MLL” to distinguish

models built from the Protein Data Bank (PDB) structures c-Myb:KIX:MLL (2AGH) or pKID:KIX:MLL (2LXT), respectively. Though ideally the models of KIX_c:MLL_c and KIX_p:MLL_p should be the same, they were treated as different systems due to the native-centric nature of the EIKB model, although this difference does not affect the validity of our conclusions (see below).

A total of four binary systems: KIX_c:c-Myb, KIX_c:MLL_c, KIX_p:pKID, and KIX_p:MLL_p; and two ternary systems: c-Myb:KIX_c:MLL_c and pKID:KIX_p:MLL_p were simulated and examined in this study. First, the force fields of ternary complexes were built from the PDB structures. Then, the force fields of binary complexes were directly extracted from their ternary models. One assumption of building the binary model based on the ternary model is that the structure of the binary complex is identical to that in the corresponding ternary complex. This assumption is supported by NMR studies (51) in which the pairwise RMSD between backbone atoms of the well-structured parts of the KIX domain in the KIX_p:MLL_p binary complex and pKID:KIX_p:MLL_p ternary complex is only 1.07 Å and the KIX protein backbone is not significantly affected by binding pKID. Ideally, the EIKB models of KIX_c:MLL_c (built from c-Myb:KIX:MLL) and KIX_p:MLL_p (built from pKID:KIX:MLL) should be the same. However, they are treated differently due to the native-centric nature of the EIKB model. The native contacts used by the EIKB models of KIX_c:MLL_c and KIX_p:MLL_p are listed in ***Appendix*** Tables 3.3 and 3.4, respectively. A total of 10 native contacts are found in common (marked as red) out of the 28 intermolecular native contacts of KIX_c:MLL_c and the 27 intermolecular native contacts of KIX_p:MLL_p. The RMSD between the EIKB models of KIX_c:MLL_c and KIX_p:MLL_p is 3.72 Å. Given the facts that KIX_c:MLL_c and KIX_p:MLL_p are structurally similar, their EIKB models show reasonable number of identical native contacts (>30%), and the sums of force constants of the intermolecular native contacts for both KIX_c:MLL_c (sum = -25.50 kcal/mol/Å^2) and KIX_p:MLL_p

(sum = -27.24 kcal/mol/Å^2) are very close, it is reasonable to assume the EIKB models of KIX_c:MLL_c and KIX_p:MLL_p are approximately the same. In fact, our simulations reproduced the positive allosteric effects in both c-Myb:KIX:MLL and pKID:KIX:MLL. Therefore, it is reasonable to assume that this approximation does not affect the validity of our conclusions.

Model Calibration

After setting up the system, two scaling factors α and β are introduced to calibrate the model so that the model reproduces some essential experimental findings. The scaling factor α is used to scale the intramolecular contact strengths to compensate the often-overestimated helicities of IDPs by the EIKB model (11, 52). The values of the scaling factor α were adopted from previous studies with $\alpha_{c\text{-}Myb} = 0.45$ (11), $\alpha_{MLL} = 0.05$ (11), $\alpha_{pKID\text{-}\alpha1} = 0.75$ (10), and $\alpha_{pKID\text{-}\alpha2} = 0.15$ (10). Those values were obtained by fitting the model with helicities from either experiment or bioinformatics predictions. For pKID, the phosphoserine SER133 was modeled as glutamic acid and the salt-bridge native contact interactions were scaled down by 40% to avoid double-counting similar to previous work (10). The scaling factor β is used to scale the strength of the intermolecular contacts between IDP ligands and KIX so that the modeled IDPs recapitulate the binary experimental K_ds upon binding free KIX. For each peptide, the optimal value of β was determined by interpolation on the K_d vs. β plot of binary systems: KIX_c:c-Myb, KIX_c:MLL_c, KIX_p:pKID, and KIX_p:MLL_p (see Figure 3.2) by matching experimental K_ds. Importantly, our model does not directly encode the allosteric effect by tuning the peptides to match the ternary K_ds when binding to KIX that is prebound by a ligand.

The HREX Method

The K_d vs. β data are essential for model calibration. In previous studies, the K_d vs. β calibration curve was obtained by running a series of brute-force unbiased simulations (11) or by running

temperature replica exchange simulations (10) at different values of β. However, the high computational cost of calculating K_d allows only a few (K_d, β) data points to be evaluated, leading to large errors due to discretization. The disadvantage of these methods is significant, especially for tightly bound ligands for which the bound/unbound transitions are rare during the simulation. The HREX method addresses this issue by allowing systems with high binding affinities to exchange coordinates with systems with low binding affinities and therefore facilitates the sampling of the unbound states for tightly bound ligands. The HREX method uses two exchange coordinates: the temperature (T) and the scaling factor β. The nearest two windows of trajectories in either dimension, (β_i, T_i) and (β_{i+1}, T_i) or (β_i, T_i) and (β_i, T_{i+1}) exchange their coordinates periodically using the Metropolis algorithm in the simulation (53). Trajectories running at high temperatures or small βs facilitate the sampling of the unbound states for trajectories at low temperatures or large βs. Notably, the HREX method only requires a single simulation with multiple trajectories running in parallel on the (β, T) grid to obtain the K_d vs. β data.

The MD Simulation Protocol

All simulations were performed using the OpenMM library (54), which allows simulations with highly customizable force fields and GPU acceleration. The HREX method shows some resemblance with the Replica Exchange with Solute Tempering (REST) method (55) in which scaling factors are used to control interactions between different groups of atoms. The HREX method was implemented using in-house C++ codes based on the OpenMM C++ API and Message Passing Interface (MPI). Each MPI process runs a single simulation. Two Hamiltonian variables, temperature T and the scaling factor β, were chosen as exchange coordinates. The HREX method allows replicas to exchange coordinates with each other and therefore enhances the sampling. The system cartesian coordinates of the two nearest windows (T_i, T_{i+1}) or (β_i, β_{i+1}) were exchanged

every 10,000 steps using the Metropolis algorithm (53). The average exchange rate (number of successful exchanges / total number of exchange trials) is 24.86% for all HREX simulations in Figure 3.2.

Each HREX simulation uses 11 β windows covering a range of systems from low binding affinity to high binding affinity. The β windows are shown in Figures 3.2 and 3.3. Three temperature windows at 300 K, 320 K, and 340 K were used for each HREX simulation. Therefore, a total of 33 replicas were used for a single HREX simulation. Only the replicas at 300 K were used for data analysis. The force field of the EIKB model was implemented using different force objects of the OpenMM library and is already described in the Model section. The Langevin integrator was used to propagate the equations of motion with a friction coefficient of 0.1 ps^{-1}. For each HREX simulation, 400 million steps were simulated for each replica with a time step of 22 fs. The first 50 million steps were discarded in the data analysis. For the unbiased simulation, 3 billion steps were simulated for each system with the β^{opt} obtained from the HREX simulation while other parameters were the same as HREX. For all trajectories, snapshots were collected every 10,000 steps for data analysis. Periodic boundary condition with a cubic box of 150 Å was used for all simulations. All HREX simulations were repeated 3 times, and the unbiased simulations were repeated 10 times for each system of interest.

3.3 Results

The HREX Simulation Results

This HREX method was used to obtain the K_d vs. β data (Figure 3.2) for the four IDPs binding to both free-KIX (solid blue curves) and KIX prebound by a ligand (solid orange lines). Both the binary (binding to free KIX) and ternary (binding to KIX prebound by a ligand) K_d vs. β curves monotonically decrease with β, since β scales the interactions between KIX and the peptides. The

HREX method also resembles the isothermal titration calorimetry (ITC) experiment (56), in which the binding enthalpy (ΔH) and entropy (ΔS) can be calculated by curve fitting. The K_d vs. β plot can be used to calculate ΔH and ΔS by fitting Equation 3.4, which is derived based on Equations 3.1-3.3. Equation 3.3 expresses K_d as a function of the fraction of unbound ligands (P_u) in simulation while Equation 3.2 converts P_u to the binding free energy ΔG, which is also a function of ΔH and ΔS as shown in Equation 3.1. Therefore, the HREX method can not only be used as a model calibration protocol, but it can also be used to study the binding thermodynamics.

$$\Delta G = \Delta \mathrm{H} - \mathrm{T}\Delta \mathrm{S} \tag{3.1}$$

$$\Delta G = -RTln\left(\frac{1-P_u}{P_u}\right) \tag{3.2}$$

$$K_d = \frac{1660}{V} \times \frac{P_u^2}{1-P_u} \tag{3.3}$$

$$\ln(K_d) = \ln\left(\frac{1660}{V}\right) - \frac{T\Delta \mathrm{S}}{RT} + \frac{\Delta \mathrm{H}}{RT}\beta - \ln\left\{exp\left(\frac{T\Delta \mathrm{S}}{RT} - \frac{\Delta \mathrm{H}}{RT}\beta\right) + 1\right\} \tag{3.4}$$

We first calculated the optimal values of β (Figure 3.2, blue dashed vertical lines) for the four ligands from the binary K_d vs. β calibration curves (Figure 3.2, blue solid curves) using equation 3.4 by locating the β that matches the experimental K_d (Figure 3.2, black dashed horizontal lines). The optimal values of β (β^{opt}, Figure 3.2, blue dashed vertical lines) were then used to obtain the ternary K_d vs. β data (Figure 3.2, orange solid curves) when KIX is prebound by a ligand. The β of the first bound ligand is set to be β^{opt} in simulations for the calculation of the ternary K_d for the second ligand. The allosteric effect can be examined by comparing the binary K_d with ternary K_d at β^{opt}. We observe that all ternary K_d vs. β curves (Figure 3.2, orange solid curves) are systematically lower than their corresponding binary curves (Figure 3.2, blue solid curves),

suggesting all the four peptides bind more favorably with KIX that is prebound by a ligand, compared to free KIX. The ternary K_ds predicted by our simulations (Figure 3.2, red dashed horizontal lines) are calculated by interpolating the ternary K_d vs. β data at β^{opt} and are listed in Table 3.1. Overall, the predicted ternary K_ds (Table 3.1) of the four peptides show excellent agreement with the experimental results. Therefore, we conclude that our simulations captured the cooperative allosteric effect in KIX for both the c-Myb:KIX_c:MLL_c and pKID:KIX_p:MLL_p systems.

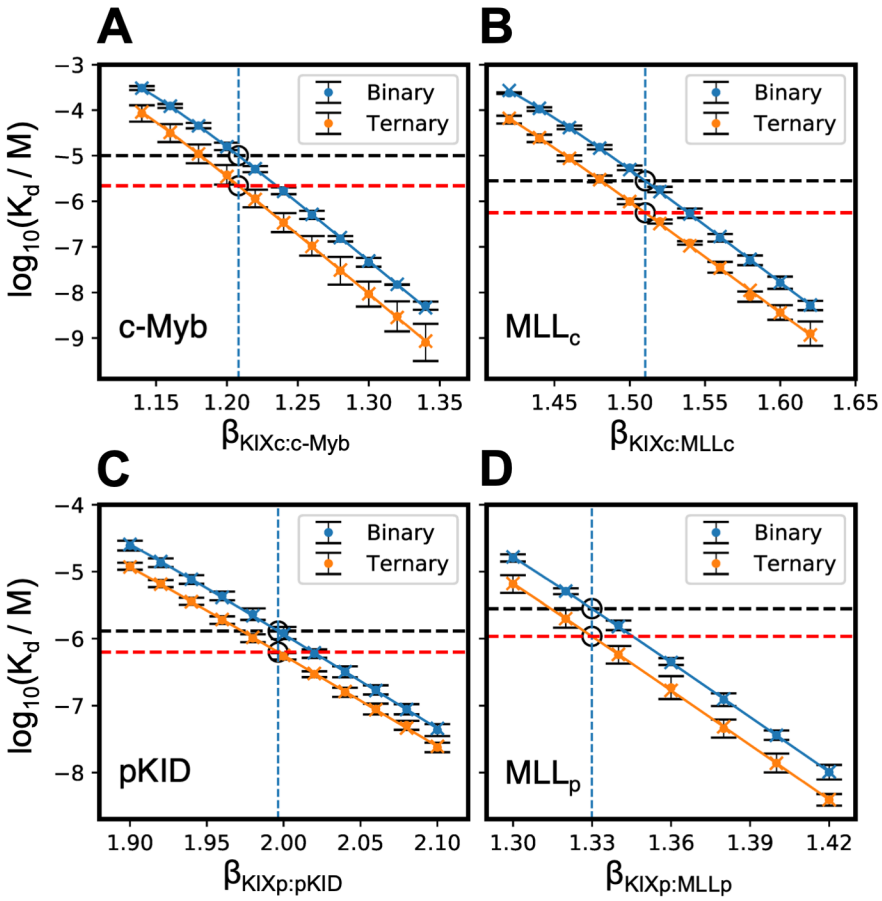

Figure 3.2 Binary (blue) and ternary (orange) K_d vs. β calibration curves of c-Myb (A), MLL_c (B), pKID (C), and MLL_p (D) with a dielectric constant D = 40. Each data point is the average of three independent HREX simulations. The blue dashed vertical lines denote β^{opt} for each peptide. The black dashed horizontal lines denote the corresponding experimental values of K_d while the red dashed horizontal lines denote the simulated K_d in ternary when $\beta = \beta^{opt}$ for each peptide.

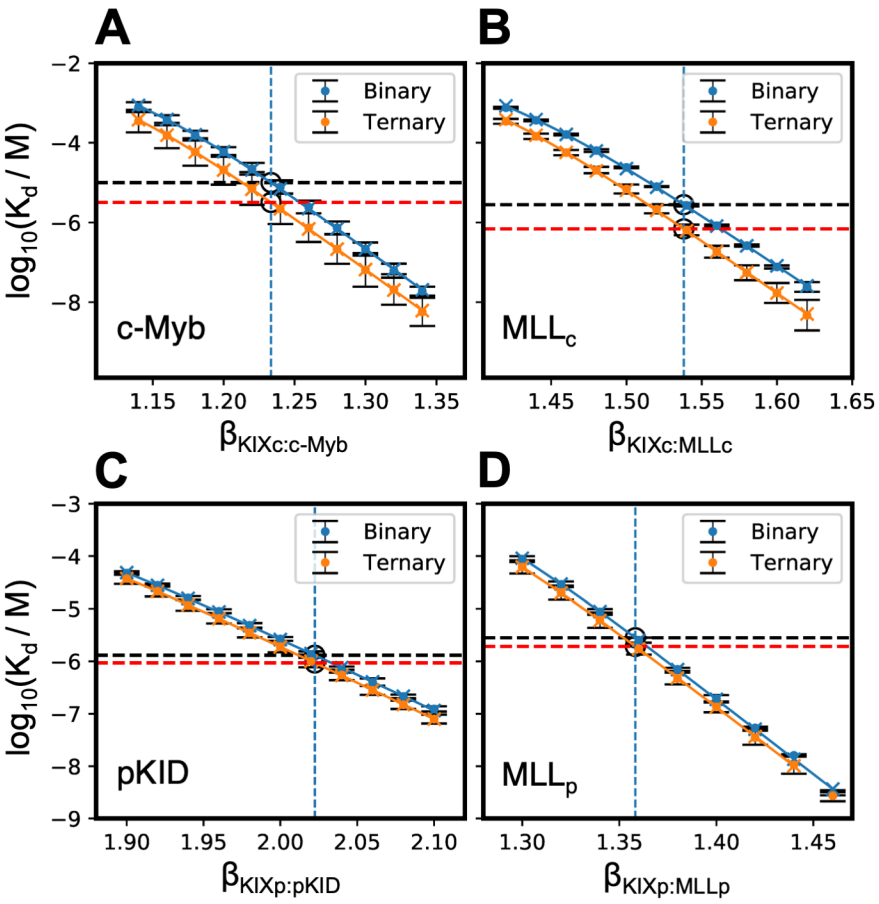

Figure 3.3 Calibration curves of the binary and ternary systems at D = 80. Each panel represents IDP calibration curves in the binary or ternary complexes.

Table 3.1 Thermodynamic data of IDPs binding to free and bound KIX. All simulation data were calculated at 300 K and D = 40. The units for $-T\Delta S_{OP}$, $T\Delta S_{tot}$, and ΔH_{tot} are kcal/mol.

Ligand	Binding to	Exp. K_d, μM[a]	Sim. K_d, μM[b]	$-T\Delta S_{OP}$[c]	$-T\Delta S_{tot}$[d]	ΔH_{tot}[d]
c-Myb	KIX_c	10 ± 2	10	1.06	20.24	-21.35
c-Myb	KIX_c:MLL_c	4 ± 1	2.19	0.49	20.02	-21.62
MLL_c	KIX_c	2.8 ± 0.4	2.8	1.30	24.50	-26.02
MLL_c	KIX_c:MLL_c	1.7 ± 0.1	0.56	0.73	23.46	-25.48
pKID	KIX_p	1.30 ± 0.02	1.30	0.27	17.73	-19.48
pKID	KIX_p:MLL_p	0.65 ± 0.05	0.63	0.12	16.79	-18.76
MLL_p	KIX_p	2.8 ± 0.4	2.8	0.64	23.48	-24.99
MLL_p	KIX_p:pKID	1.5 ± 0.2	1.08	0.49	23.11	-24.92

[a.] Experimental data from (48).
[b.] Simulation results from HREX.
[c.] Conformational entropy contributions calculated from order parameters using method described elsewhere from (11).
[d.] Total conformational entropic and enthalpic contributions calculated from HREX.

The Unbiased Simulation Results

With the optimized force fields from HREX, we further carried out a set of long unbiased simulations for all binary and ternary systems. The unbiased simulations give very close ternary K_ds (Table 3.2) compared to ternary K_ds calculated by the HREX method, suggesting the β^{opt} calculated by HREX simulations are accurate. Therefore, the HREX method is also an effective and efficient way to optimize the force fields of models similar to the EIKB model for unbiased MD simulations, as the unbiased MD simulations often contain richer information such as binding kinetics.

Table 3.2 K_ds of different systems calculated by HREX and unbiased simulations at D = 40.

Ligand	Binding to	Exp. K_d, μM[a]	HREX K_d, μM[b]	Unbiased K_d, μM[c]
c-Myb	KIX_c	10 ± 2	10	27.7 ± 26.1
c-Myb	KIX_c:MLL_c	4 ± 1	2.19	7.28 ± 12.7
MLL_c	KIX_c	2.8 ± 0.4	2.8	3.69 ± 1.22
MLL_c	KIX_c:MLL_c	1.7 ± 0.1	0.56	0.76 ± 0.36
pKID	KIX_p	1.30 ± 0.02	1.30	1.73 ± 0.98
pKID	KIX_p:MLL_p	0.65 ± 0.05	0.63	0.66 ± 0.52
MLL_p	KIX_p	2.8 ± 0.4	2.8	4.08 ± 1.46
MLL_p	KIX_p:pKID	1.5 ± 0.2	1.08	1.15 ± 0.69

[a.] Experimental data from (48).
[b.] K_d calculated by HREX.
[c.] K_d calculated by long unbiased simulations.

Thermodynamics

To examine the thermodynamics of the allosteric mechanism in KIX, the binding enthalpy (ΔH) and entropy (ΔS) of binary (binding to free KIX) and ternary (binding to KIX prebound by a ligand) complexes were calculated for both c-Myb:KIX_c:MLL_c and pKID:KIX_p:MLL_p by fitting the K_d vs. β data with equation 3.4 and are listed in Table 3.1. A previous study using the EIKB model found the cooperative allostery in c-Myb:KIX_c:MLL_c is due to a favorable conformational entropic change of KIX, in which the first bound ligand pays the entropic costs for the second ligand to bind (11). Our data from the HREX simulations also support this reduced entropy mechanism, as the entropic cost ($-T\Delta S_{tot}$) for a ligand to bind KIX prebound by a ligand is always lower than the cost to bind free KIX (Table 3.1). Notably, the ΔH/ΔS from the HREX method is considered to be the total binding enthalpy/entropy change while the ΔS from the previous study

(11) is calculated as the Gibbs entropy from the state probability distributions of three order parameters that reflect the major conformational fluctuations of KIX.

The entropic cost based on order parameters ($-T\Delta S_{OP}$) was re-calculated here using our unbiased trajectories and is shown to be in line with the total entropic cost ($-T\Delta S_{tot}$) and the total enthalpic change (ΔH_{tot}) from HREX in Table 3.1. For the four IDPs, both $-T\Delta S_{tot}$ and $-T\Delta S_{OP}$ are qualitatively consistent in values where the cost for the second ligand to bind KIX prebound by the first ligand is always lower than the cost to bind free KIX. Therefore, our data support the reduced entropy mechanism for both c-Myb:KIX_c:MLL_c and pKID:KIX_p:MLL_p and is consistent with the previous study on c-Myb:KIX_c:MLL_c. The fact that these quantities agree, i.e., the entropy from the temperature derivative and the configurational entropy arising from specific fluctuation changes, is not surprising given the fact that the EIKB model really only has peptide and protein conformational degrees of freedom. The configurational entropy change in the disorder to order transition of the peptide upon binding to the protein represents the ~20 kcal/mol offset observed in the differences between the $-T\Delta S_{OP}$ and $T\Delta S_{tot}$ values and is assumed to remain constant for both free KIX and KIX prebound by a ligand.

3.4 Discussion

The idea of the reduced entropy mechanism is that the first bound ligand reduces the entropic cost for the second ligand to bind. The first ligand binds and reduces the dynamical fluctuations of KIX, indicated by the more narrowly distributed states on the three order parameters for the bound KIX (Figure 3.4 and 3.5). We also examined how the binding affinity of the first ligand changes the binding affinity of the second ligand in Figure 3.6, which shows the K_d of the second ligand as a function the first bound ligand's scaling factor β. The K_d (ligand 2) vs. β (ligand 1) data were obtained at two different dielectric constants, D = 40 and 80. For the pKID:KIX_p:MLL_p system,

the K_d of neither pKID nor MLL_p show significant responses with the increase of the first ligand's β values (or K_d). However, for the c-Myb:KIX_c:MLL_c system, the K_ds of both c-Myb and MLL_c are gradually decreasing when increasing the first ligands' β values (or K_d). Therefore, there are no simple relationships between the K_d (ligand 2) and K_d (ligand 1) for the two different KIX systems.

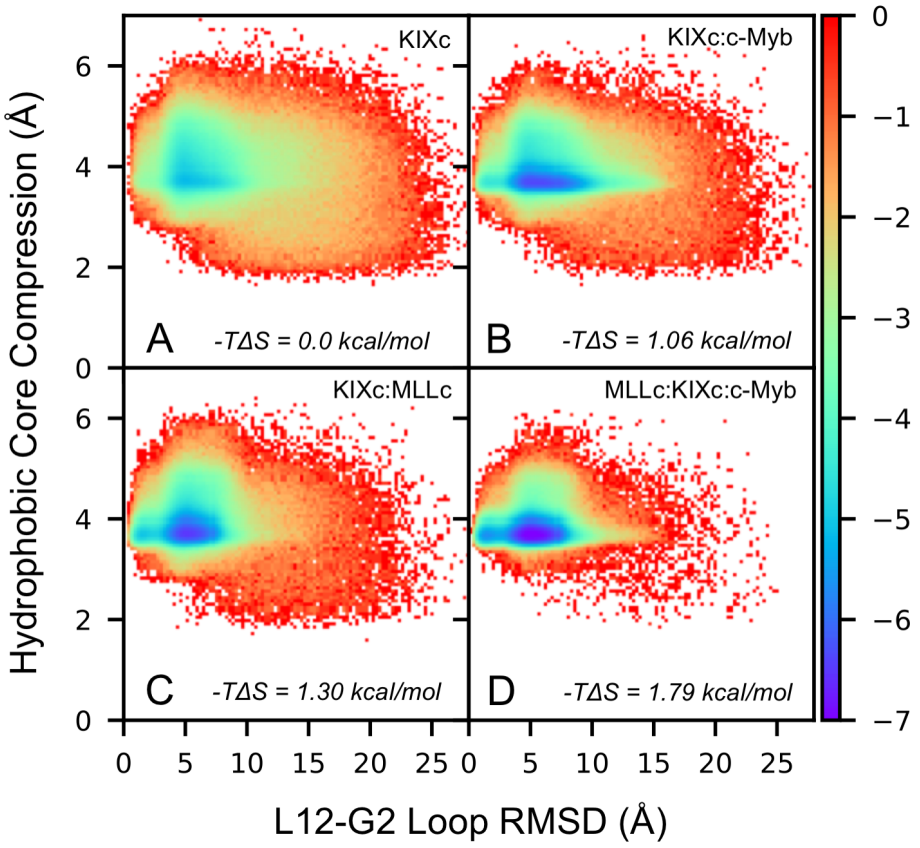

Figure 3.4 Free energy plots of the hydrophobic core compression and L_{12}-G_2 loop RMSD of KIX_c in the free state (A), bound with c-Myb (B), bound with MLL_c (C), and bound with both c-Myb and MLL_c (D). Definitions of these order parameters are described in a previous paper (11).

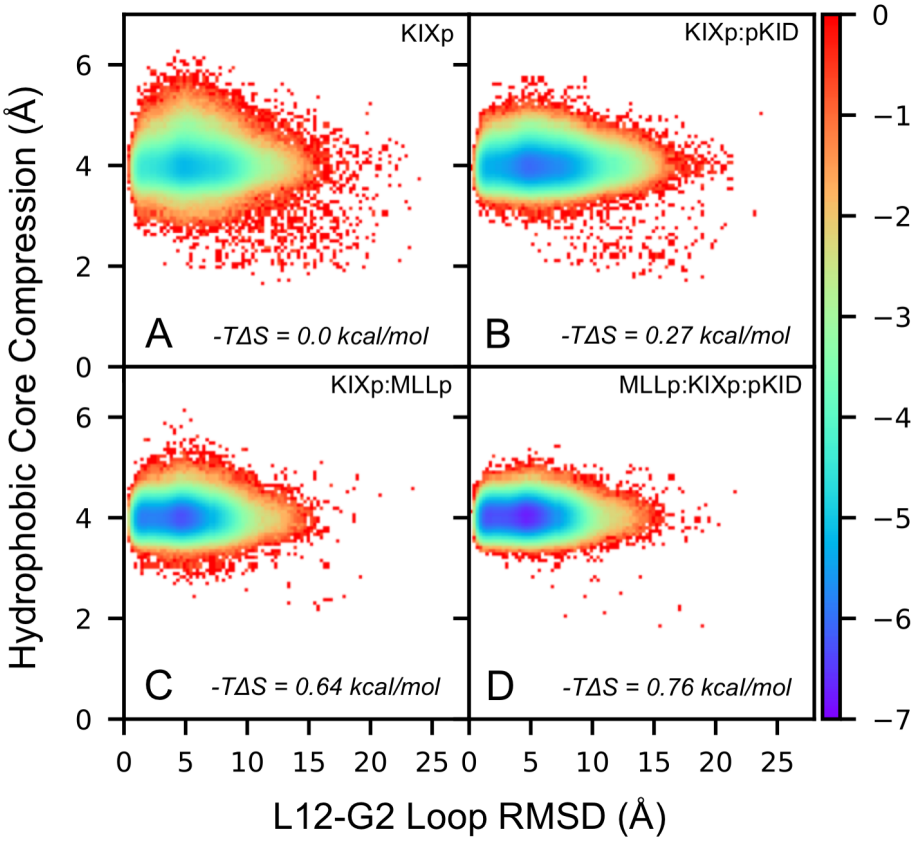

Figure 3.5 Free energy plots of the hydrophobic core compression and L_{12}-G_2 loop RMSD of KIX_p in the free state (A), bound with pKID (B), bound with MLL_c (C), and bound with both c-Myb and MLL_c (D). Definitions of these order parameters are described in a previous paper (11).

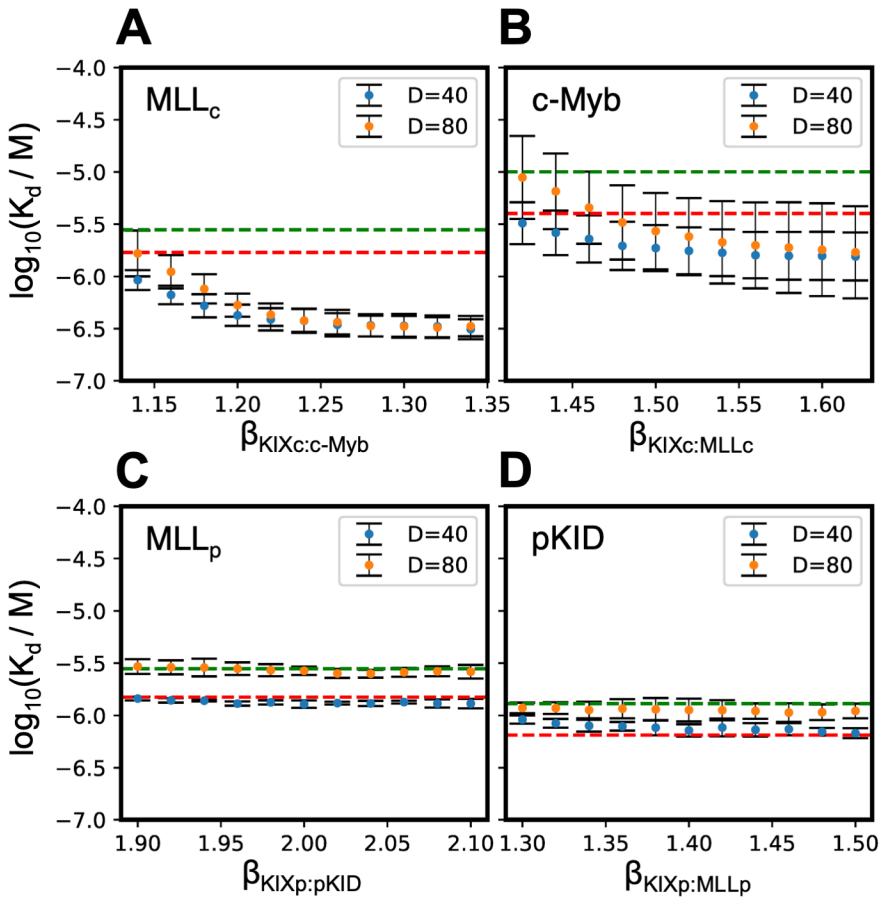

Figure 3.6 The K_d (ligand 2) vs. β (ligand 1) plots of MLL_c (A), c-Myb (B), MLL_p (C), and pKID (D), respectively. The green and red dashed horizontal lines correspond to the binary and ternary experimental K_ds of their corresponding IDPs in each panel.

The original KB model in the previous study on c-Myb:KIX_c:MLL_c does not encode explicit electrostatic interactions, although it successfully recapitulated the cooperative allostery (11). In this study, we found that explicit electrostatic forces with a dielectric constant D = 40 is essential for the pKID:KIX_p:MLL_p system to recapitulate the cooperative allostery, whereas electrostatics is unimportant for c-Myb:KIX_c:MLL_c. We compared the K_d vs. β data at D = 40 (Figure 3.2) with the data at D = 80 (Figure 3.3) and we found no significant decreases of K_ds from binary to ternary at D = 80 for both pKID and MLL_p. Similar results were also found in the K_d (ligand 2) vs. β (ligand 1) plots (Figures 3.6 C and D), in which the ternary K_ds (Figure 3.6, orange dots) for the

two peptides at D = 80 are almost unchanged from their binary K_ds (Figure 3.6 green lines). However, the electrostatic interactions are not essential for c-Myb:KIX_c:MLL_c to reproduce the positive allosteric effect, consistent with the previous study. We further calculated the residue-residue pairwise electrostatic interaction energy (E_{elec}) between c-Myb and MLL (***Appendix*** Table 3.5) as well as the E_{elec} between pKID and MLL (***Appendix*** Table 3.6) at D = 40 based on the topology of the native PDB conformation. The total E_{elec} between c-Myb and MLL is 0.0178 kcal/mol while the total E_{elec} between pKID and MLL is -0.236 kcal/mol, which suggests the interactions between pKID and MLL are electrostatically more favorable. Therefore, explicit electrostatics with D = 40 is more compatible with the EIKB model and should be used for future studies.

A recent review by Wright et al. (57) pointed out a seeming discrepancy of the allostery mechanism between the previous study (11) and their ITC experiment (48). Their data suggested that the overall thermodynamic driving force in c-Myb:KIX_c:MLL_c is due to a favorable enthalpic change whereas this process in pKID:KIX_p:MLL_p is governed by an overall favorable change in the total entropy. Though their results seem to disagree with the reduced entropy mechanism proposed by the previous study (11), which is also supported our current simulations, we wish to emphasize two points. First, the entropy calculated by coarse-grained MD simulations, either from order parameters or HREX, corresponds to the protein conformational entropy. The total thermodynamic function changes (entropy and enthalpy) calculated from the ITC experiments include significant, and non-deconvolvable, contributions from solvation and desolvation of the interacting proteins as well as the configurational entropy change of the solvent degrees of freedom. Our model explicitly looks at only the protein conformational entropy. Thus, while we cannot directly suggest what the overall driving force for the association is, our conclusion that “pre-paying” the

configurational entropic cost of binding provides a mechanism for allostery, is not necessarily inconsistent with the ITC measurements. Moreover, by focusing on the changes in conformational flexibility, we can form direct and testable hypotheses regarding protein flexibility and allosteric regulation, which cannot emerge from calculations of overall thermodynamic function changes from ITC. For example, the conformational entropy has an established mapping to protein structural dynamics and can be measured by NMR relaxation experiments (58) through a dynamic proxy as well as by molecular dynamics simulations (11, 59). Therefore, the reduced conformational entropy model may be useful to guide such experiments in the future.

3.5 Conclusions

In conclusion, we developed a new sampling method that can accurately and efficiently calculate K_d and optimize force fields for coarse-grained models similar to the EIKB model. The new method can be readily used to study allosteric regulation in systems similar to KIX. Our simulations recapitulate the cooperative allosteric effects in both c-Myb:KIX_c:MLL_c and pKID:KIX_p:MLL_p and our data support the reduced entropy mechanism in which the first bound ligand reduces the entropic cost for the second ligand to bind. As a whole, our work provides new tools to study the allosteric regulation by IDPs and also provides new insights into the allosteric mechanism.

3.6 Appendix

Table 3.3 Native contacts interactions between KIX_c and MLL_c in PDB structure 2AGH. Contacts that are the same as those found in KIX_p:MLL_p are marked as red.

Residues of KIX	Residues of MLL	Force constant (kcal/mol/Å^2)	Distance (Å)
PHE612	PHE852	-1.444	9.15
PHE612	VAL853	-1.251	6.39
PHE612	ASN856	-0.746	8.82
PHE612	THR857	-0.851	8.20
PRO615	ASN856	-0.304	6.95
ARG624	ASP848	-0.456	10.14
ARG624	ASP851	-0.456	10.89
ARG624	PHE852	-0.792	7.74
MET625	PHE852	-1.305	8.48
ASN627	ASP848	-0.334	8.11
ASN627	ILE849	-0.644	6.39
LEU628	ILE849	-1.400	5.89
LEU628	PHE852	-1.448	8.23
LEU628	VAL853	-1.289	8.57
TYR631	ILE844	-1.044	10.08
TYR631	LEU845	-1.128	8.53
TYR631	PRO846	-0.635	8.43
TYR631	ILE849	-1.044	7.47
MET639	ILE844	-1.197	11.51
LYS656	ILE844	-0.599	7.53
LYS659	ILE844	-0.599	8.76
ILE660	ILE844	-1.301	8.06
LEU664	LEU845	-1.466	11.32
LEU664	VAL853	-1.289	8.70
LEU664	THR857	-0.863	7.66
ARG668	LEU854	-0.802	7.83
ARG668	THR857	-0.378	6.43
ARG668	PRO858	-0.338	5.10
Total		-25.40	

Table 3.4 Native contacts interactions between KIX_p and MLL_p in PDB structure 2LXT. Contacts that are the same as those found in KIX_c:MLL_c are marked as red.

Residues of KIX_p	Residues of MLL_p	Force constant (kcal/mol/Å^2)	Distance (Å)
ILE611	LEU845	-1.610	10.45
ILE611	MET850	-1.376	8.59
ILE611	VAL853	-1.383	7.25
PHE612	PHE852	-1.660	10.05
PHE612	VAL853	-1.438	6.78
PHE612	ASN856	-0.857	8.95
THR614	PRO858	-0.434	7.49
ARG624	PHE852	-0.910	8.27
ASN627	ILE849	-0.741	8.46
LEU628	ILE849	-1.610	7.16
TYR631	LEU845	-1.296	8.62
ASP638	ALA841	-0.389	7.94
MET639	ASN843	-0.675	10.49
GLU655	ASN843	-0.345	8.15
LYS656	ASN843	-0.277	5.15
LYS659	ILE844	-0.688	6.53
ILE660	LEU845	-1.610	7.31
ILE660	MET850	-1.376	8.69
GLU663	MET850	-0.661	7.98
LEU664	MET850	-1.466	6.84
LEU664	VAL853	-1.482	8.27
LEU664	LEU854	-1.685	6.80
LYS667	LEU854	-0.771	6.92
ARG668	LEU854	-0.921	6.55
ARG671	ASP851	-0.524	10.36
ARG671	LEU854	-0.921	7.42
ARG671	LYS855	-0.135	7.56
Total		-27.24	

Table 3.5 Electrostatic interactions between c-Myb and MLL in PDB structure 2AGH at D = 40.

Residues of c-Myb	Residues of MLL	Distance (Å)	Energy (kcal/mol)[a]
LYS291	ASP841	22.83	-0.0371
LYS296	ASP841	23.82	-0.0322
ARG294	ASP841	24.27	-0.0302
ARG294	ASP851	24.87	-0.0277
LYS291	ASP840	24.89	-0.0277
LYS296	ASP840	25.10	-0.0269
GLU297	LYS855	25.51	-0.0254
GLU299	LYS855	25.75	-0.0246
LYS291	ASP851	26.07	-0.0235
ARG294	ASP840	26.32	-0.0227
LYS293	ASP841	26.46	-0.0223
GLU292	LYS855	26.74	-0.0214
ARG294	ASP848	27.68	-0.0188
LYS296	ASP851	27.82	-0.0185
LYS293	ASP840	28.20	-0.0175
LYS293	ASP851	28.56	-0.0167
LYS291	ASP848	29.46	-0.0148
LYS296	ASP848	29.76	-0.0142
LYS293	ASP848	31.41	-0.0114
GLU306	LYS855	33.49	-0.0087
GLU308	LYS855	33.77	-0.0084
LYS310	ASP848	36.33	-0.0060
LYS310	ASP841	36.47	-0.0059
LYS310	ASP840	36.75	-0.0057
LYS310	ASP851	37.27	-0.0054
LYS310	LYS855	38.09	0.0048
GLU308	ASP840	35.13	0.0070
GLU308	ASP841	34.31	0.0078
GLU308	ASP851	33.73	0.0084
GLU308	ASP848	33.63	0.0086
GLU306	ASP851	32.52	0.0099
GLU306	ASP848	31.80	0.0109
GLU306	ASP840	31.25	0.0117
GLU292	ASP848	30.98	0.0121
GLU306	ASP841	30.86	0.0123
GLU297	ASP848	28.52	0.0168
GLU292	ASP851	28.09	0.0178

LYS296	LYS855	27.19	0.0201
GLU297	ASP840	26.92	0.0209
LYS293	LYS855	26.82	0.0212
GLU299	ASP848	26.58	0.0219
GLU297	ASP851	26.53	0.0220
GLU292	ASP840	25.55	0.0252
GLU299	ASP851	25.47	0.0255
GLU297	ASP841	25.31	0.0261
LYS291	LYS855	24.38	0.0297
GLU292	ASP841	23.91	0.0318
ARG294	LYS855	23.24	0.0350
GLU299	ASP840	22.56	0.0386
GLU299	ASP841	21.43	0.0454
Total			0.0178

[a.] Electrostatic energy at the pair distance calculated from the native structure. Attractive interactions are negative.

Table 3.6 Electrostatic interactions between pKID and MLL in PDB structure 2LXT at D = 40.

Residues of c-Myb	Residues of MLL	Distance (Å)	Energy (kcal/mol)[a]
ASP840	ARG124	10.26	-0.290
ASP840	ARG125	11.27	-0.239
ASP840	LYS123	12.35	-0.195
ASP840	ARG131	18.09	-0.075
ASP840	ARG135	19.22	-0.063
ASP840	ARG130	19.31	-0.062
ASP840	LYS136	21.87	-0.043
LYS855	GLU133	24.50	-0.029
ASP848	ARG124	25.57	-0.025
ASP851	ARG124	25.78	-0.024
ASP851	ARG131	26.22	-0.023
ASP851	LYS136	27.27	-0.020
ASP851	ARG135	27.62	-0.019
ASP848	ARG131	27.71	-0.019
ASP848	ARG125	28.03	-0.018
ASP851	ARG125	28.27	-0.017
ASP848	LYS123	28.43	-0.017
ASP851	LYS123	28.79	-0.016
ASP848	ARG135	28.97	-0.016
LYS855	ASP140	28.99	-0.016
ASP848	LYS136	29.08	-0.016
ASP851	ARG130	29.60	-0.015
ASP848	ARG130	30.90	-0.012
LYS855	ASP120	31.19	-0.012
LYS855	ASP144	31.91	-0.011
LYS855	GLU126	33.37	-0.009
LYS855	ASP116	33.81	-0.008
LYS855	LYS123	32.82	0.010
ASP848	ASP144	32.53	0.010
LYS855	ARG125	31.82	0.011
ASP851	ASP144	31.80	0.011
LYS855	ARG130	31.48	0.011
ASP848	ASP140	30.43	0.013
ASP848	GLU126	30.37	0.013
ASP851	GLU126	30.16	0.013
LYS855	ARG124	29.61	0.015
ASP851	ASP116	29.07	0.016

ASP851	ASP140	28.97	0.016
LYS855	ARG135	28.78	0.016
ASP848	ASP116	28.23	0.017
LYS855	ARG131	27.87	0.018
LYS855	LYS136	27.61	0.019
ASP851	ASP120	26.43	0.022
ASP848	GLU133	25.99	0.024
ASP840	ASP144	25.78	0.024
ASP848	ASP120	25.34	0.026
ASP851	GLU133	23.92	0.032
ASP840	ASP140	23.63	0.033
ASP840	GLU133	20.17	0.055
ASP840	GLU126	14.78	0.128
ASP840	ASP116	14.27	0.140
ASP840	ASP120	8.93	0.380
Total			-0.236

[a.] Electrostatic energy with the distance calculated from the native structure. Attractive interactions are negative.

Chapter 4 The Negative Allosteric Regulation in a Disordered Protein Switch

This chapter is adapted from the following manuscript:

Yanming Wang; Charles L. Brooks III; Electrostatic Forces Control the Negative Allosteric Regulation in a Disordered Protein Switch. (submitted)

4.1 Introduction

The allosteric effect, by which a ligand affects another ligand at a distal binding site of the same target (40, 60), is essential for many fundamental biological processes and is manifest in many human diseases (41, 42). One important type of allosteric regulation utilizes intrinsically disordered proteins (IDPs) (61). The intrinsic structural flexibility of IDPs allows rapid but highly specific interactions with multiple cellular targets, which is essential for their versatile roles in cellular signaling and regulation (62). Understanding the mechanism of allosteric regulation by IDPs forms the basis for other related higher-level regulatory processes in the cell cycle. However, the underlying physical principles of allosteric regulation involving IDPs are still largely elusive due to the complicated protein-protein interactions involved in these processes.

One such example is the negative allosteric regulation in the TAZ1 protein switch (Figure 4.1). The TAZ1 domain of transcriptional coactivators CBP/P300 is critical for cellular hypoxic response (46, 57). In hypoxia, TAZ1 binds the disordered α-subunit of the transcription factor HIF-1 (HIF-1α) to trigger the transcription of adaptive genes to respond to the hypoxic stress (63). At the same time, the disordered CITED2 peptide acts as the negative feedback regulator to rapidly

displace HIF-1α and efficiently attenuates the hypoxic response (64). Though both HIF-1α and CITED2 have the same dissociation constants (10 nM) in their binary complexes with TAZ1, recent experiments by Wright et al. discovered that CITED2 is extremely efficient in displacing HIF-1α under equimolar conditions upon binding the same target TAZ1 (14). This discovery is surprising and it overturns the originally proposed naive mass-action displacement mechanism (65). The experiments suggest the formation of a ternary intermediate complex (also captured by our simulations and is shown in Figure 4.1C) that facilitates the dissociation of the partially bound HIF-1α is essential for the displacement mechanism (14, 66).

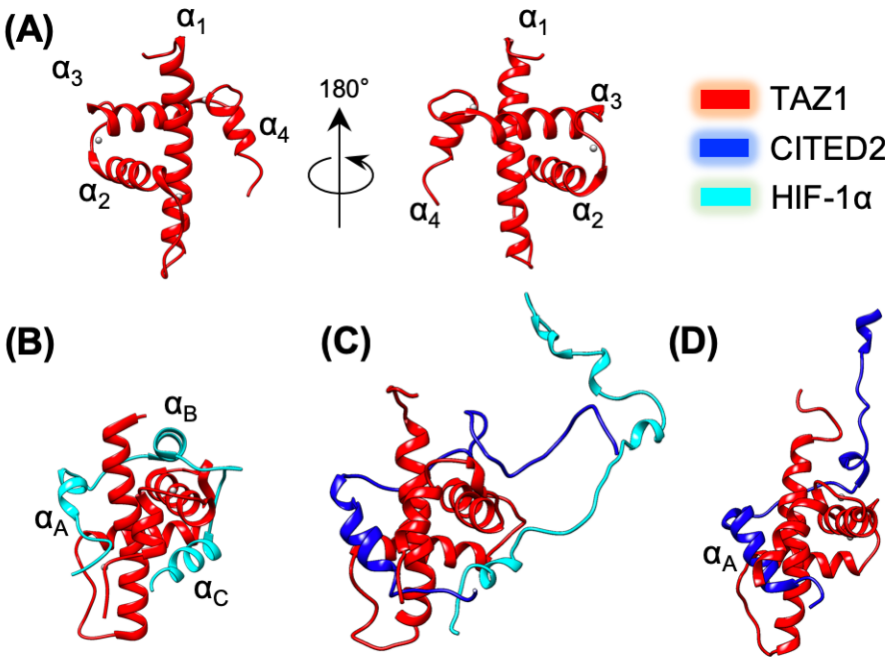

Figure 4.1 Structures of free TAZ1 (panel A, PDB code 1U2N), TAZ1:HIF-1α binary complex (panel B, PDB code 1L8C), CITED2:TAZ1:HIF-1α ternary intermediate complex (panel C, structure captured by simulations), and TAZ1:CITED2 binary complex (panel D, PDB code 1R8U).

Even though significant insights were provided into the allosteric mechanism by recent experimental (14, 66) and computational (67) studies, the underlying physical principles are still

not fully elucidated. To understand the essential physics governing the negative allosteric effect in the TAZ1 protein switch, we carried out molecular dynamics (MD) simulations using a coarse-grained model (8, 11) that has shown success in modeling the positive/cooperative allosteric effect in the KIX domain of the same parent CBP/P300 transcription coactivator. Our simulations show excellent agreement with experiment in reproducing the overall allosteric effect and capturing the ternary intermediate complex critical for the displacement mechanism. Notably, our simulation data also pinpoint the decisive role of electrostatics in the TAZ1 protein switch for the first time to our knowledge.

The protein data bank (PDB) structures of free TAZ1 (1U2N), the TAZ1:CITED2 binary complex (1R8U), and the TAZ1:HIF-1α binary complex (1L8C) were used to build the coarse-grained models for MD simulations. The coarse-grained model only utilizes the experimental binding data from the binary systems (TAZ1:HIF-1α and TAZ1:CITED2) and does not explicitly encode the allosteric effect. This model mainly considers short-range native contacts and long-range electrostatic forces as intermolecular interactions (8). To balance the short-range native contacts and the long-range electrostatic interactions between TAZ1 and the two peptides, all TAZ1-peptide native contact interaction strengths were scaled by a factor β so that the modeled binary complex reproduces the experimental K_d at the optimal value of β (β_{opt}). A dielectric constant (D) was used to modulate the strengths of electrostatic interactions and was examined at 80, 60, 50, and 40 to mimic increasing strengths of electrostatic interactions. The values of β_{opt} for each peptide at different dielectric constants were obtained by the Hamiltonian replica exchange (HREX) method developed previously (68). The scaling factor β plays the role of balancing the short-range and long-range forces of the model so that the K_ds of the two simulated binary complexes are always

kept at the experimental target of 10 nM. Further details of the simulation setup can be found in the ***Method*** section below.

4.2 Results and Discussion

As shown in Figures 4.2 A and B, the K_ds of the two ligands binding to free TAZ1 in the binary complex monotonically decrease as the scaling factor β increases. When the dielectric constant (D) decreases from 80 to 40, the β^{opt} (Figure 4.2, vertical lines) of HIF-1α that matches the binary experimental K_d = 10 nM (-8 in the log scale, black dashed horizontal lines) decreases from 1.529 to 1.441 (decreased by -5.8%) while the β^{opt} of CITED2 decreases from 1.206 to 1.114 (decreased by -7.6%), respectively. As the dielectric constant is decreasing or the strengths of electrostatic interactions are increasing, the strengths of native contacts (controlled by β) needed to keep the same K_d are also decreasing for both peptides, suggesting significant electrostatic interactions are contributing to the binding of the two peptides. Notably, the percentage decrease of β^{opt} for CITED2 (-7.6%) is larger than that for HIF-1α (-5.8%), indicating the percentage contribution by electrostatic forces to the K_d in the TAZ1:CITED2 complex is larger than that of the K_d for the TAZ1:HIF-1α complex, consistent with the apparent net charges of the two peptides in the model (-7 for CITED2 and -5 for HIF-1α). Recent mutational experiments also confirmed the hot-spots residues for HIF-1α binding energetics are all hydrophobic residues rather than charged residues (69).

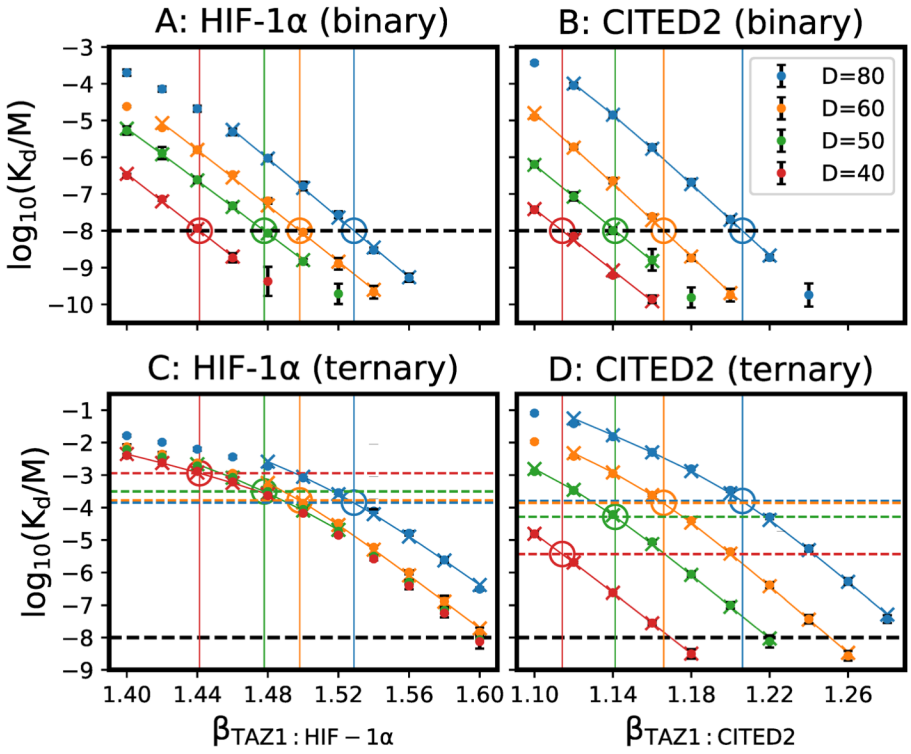

Figure 4.2 K_d as a function of the scaling factor β for HIF-1α and CITED2 in binary complex (A, B) and ternary complex (C, D). All simulations were carried out at 4 different dielectric constants D = 40 (red), 50 (green), 60 (orange), and 80 (blue). The experimental K_d (10 nM, or -8 in log scale) of the binary complexes of the two peptides are denoted by black horizontal dashed lines. The β^{opt} at different dielectric constants are denoted by vertical lines with different colors.

To study the allosteric effect, the competing ligand was added into the binary system with the competitor ligand's β kept at its predetermined value of β^{opt}. The same HREX method was used again to compute the K_d of the given ligand in the ternary complex with its competing ligand (Figures 4.2 C and D). The predicted K_d of the given ligand in the ternary complex is calculated by interpolation on the K_d vs. β plot at $\beta = \beta^{opt}$. As shown in Figures 4.2 C and D, the predicted K_d of CITED2 in the ternary complex decreases from ~100 μM to 3.8 μM whereas the K_d of HIF-1α in ternary increases from ~100 μM to 1.26 mM when the dielectric constant decreases from 80 to 40. The striking asymmetrical changes of the K_ds for the two peptides at different dielectric constants indicate HIF-1α can only be efficiently displaced when electrostatic interactions are

strong enough. Notably, a slight decrease of the dielectric constant from 80 to 60 has little effect on shifting the overall result of the competition between the two peptides. The displacement of HIF-1α by CITED2 is only significant when the strengths of electrostatic interactions are doubled or the dielectric constant changes from 80 to 40. Overall, our data suggest electrostatic forces are essential for the negative allosteric effect in the TAZ1 protein switch.

Table 4.1 Experimental and simulated K_ds (nM) of different systems studied in the TAZ1 protein switch. Only the K_ds simulated at the dielectric constant of 40 are shown.

Peptide	Binding to	K_d (Exp.)[a]	K_d (HREX)[b]	K_d (unbiased)[c]
CITED2	TAZ1	10 ± 1	10	12.4 ± 4.3
CITED2	TAZ1:HIF-1α	0.2 ± 0.1	3.8×10^3	$8.20\times10^3 \pm 11.3\times10^3$
HIF-1α	TAZ1	10 ± 1	10	13.9 ± 12.4
HIF-1α	TAZ1:CITED2	900 ± 100	1.26×10^6	$1.24\times10^6 \pm 0.44\times10^6$

[a] Data from reference (14).
[b] Data from HREX simulations.
[c] Data from unbiased simulations.

To further check and complement with the HREX simulations, we carried out long unbiased simulations modeling electrostatics with a dielectric constant of 40. We simulated each system for an aggregate time of 1.32 ms for the ternary complex and 0.66 ms for the two binary complexes. The unbiased simulations adopted the optimal values of $\beta = \beta^{opt}$ for the two peptides directly from the HREX simulations. The K_ds of the two peptides in the binary and ternary complex calculated from unbiased simulations, along with K_ds calculated from HREX simulations and K_ds obtained from the experiments are shown in Table 4.1. Overall, the K_ds from unbiased simulations show good agreement with K_ds from the HREX simulations. Notably, our simulated K_ds in the ternary complex deviate significantly from the experimental values for both CITED2 (3.8×10^3 nM by simulation vs. 0.2 nM from the experimental analysis) and HIF-1α (1.26×10^6 nM by simulations vs. 900 nM by experiment). However, we need to emphasize the K_ds of the two peptides in the

ternary complex from experiment are apparent dissociations constants (14) and only qualitatively reflect the relative binding affinities of the two peptides in the ternary complex. They do not quantitatively mean CITED2 has a higher binding affinity in the ternary complex compared to that in the binary complex (K_d decreases from 10 nM to 0.2 nM). Considering the simplicity of our model, it is remarkable that our model captured the asymmetric responses of the two peptides' K_ds in the ternary complex with respect to the changes of the dielectric constant. Therefore, we conclude that our simulations captured the negative allosteric effect in the TAZ1 protein switch.

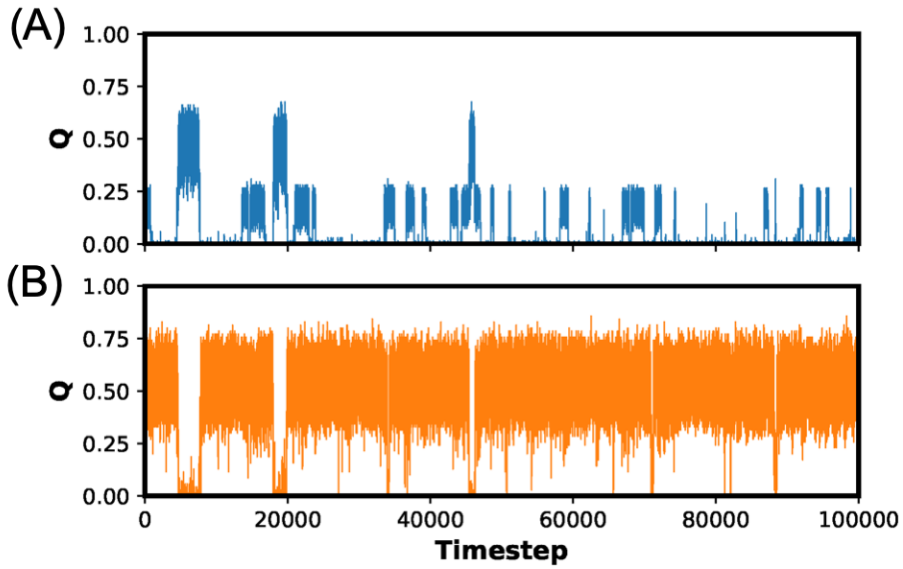

Figure 4.3 Representative trajectories of fraction of native contacts (Q) of HIF-1α (A) and CITED2 (B). HIF-1α shows two bound states: the bound state (0.3 < Q) and the partially bound state (0.1 < Q < 0.3). CITED2 shows a single bound state (Q > 0.1).

The experiments proposed a displacement mechanism with a CITED2:TAZ1:HIF-1α ternary state as an important intermediate (14, 66). Our simulations demonstrate this ternary intermediate complex (Figure 4.1 C) with a bound CITED2 and a partially bound HIF-1α, primarily due to the weak interactions between the highly flexible α_1-α_2 regions of HIF-1α and TAZ1, consistent with the NMR experiments (14, 66). The competition between the two peptides in the ternary complex

is evident in the trajectories of the fraction of native contacts (Q) formed between TAZ1 and the two peptides from the unbiased simulations (Figure 4.3), in which $Q_{HIF\text{-}1\alpha}$ = 0.1 to 0.3 for the partially bound HIF-1α, $Q_{HIF\text{-}1\alpha}$ > 0.3 for the fully bound HIF-1α, and Q_{CITED2} > 0.1 for the bound CITED2. To better understand the displacement mechanism, especially the role of the ternary intermediate complex, we constructed a Markov state model (70, 71) with five states involved in the allosteric regulation of the TAZ1 protein switch: TAZ1 (free-TAZ1 state), TAZ1 with bound CITED2 (TAZ1:CITED2 state), TAZ1 with partially bound HIF-1α (TAZ1:p-HIF-1α state), TAZ1 with bound HIF-1α (TAZ1:HIF-1α state), and TAZ1 with bound CITED2 and partially bound HIF-1α (CITED2:TAZ1:p-HIF-1α state), from the unbiased simulations. The five different states are classified based on their characteristic values of Q_{CITED2} and $Q_{HIF\text{-}1\alpha}$. More details of the Markov state model setup can be found in the ***Method*** section.

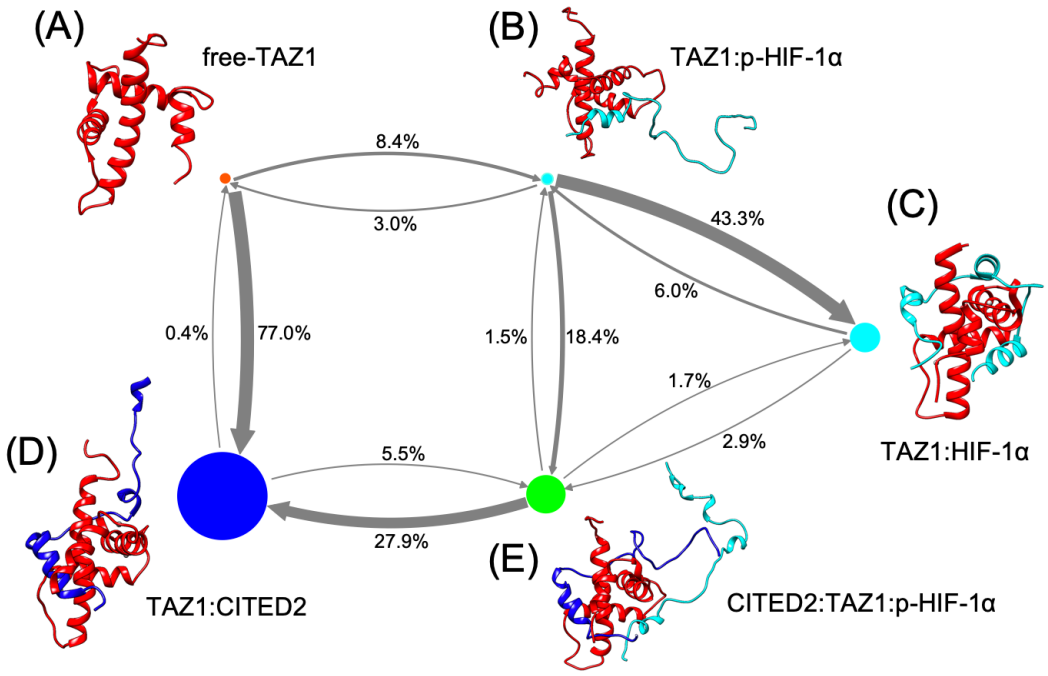

Figure 4.4 Schematic diagram of the five-state Markov state model of the allosteric mechanism of the TAZ1 protein switch. Only transition probabilities with a flux over 10^{-4} are shown. The circle size is proportional to the equilibrium population of the corresponding state. The transition probability matrix is shown in Table 4.2.

Table 4.2 Transition matrix of the 5-state Markov state model using the unbiased trajectories when D = 40.

From \ To	free-TAZ1	TAZ1:CITED2	TAZ1:p-HIF-1α	CITED2:TAZ1:p-HIF-1α	TAZ1:HIF-1α
free-TAZ1	5.54%	76.96%	8.38%	4.90%	4.23%
TAZ1:CITED2	0.44%	93.78%	0.18%	5.53%	0.08%
TAZ1:p-HIF-1α	3.03%	11.29%	24.00%	18.43%	43.25%
CITED2:TAZ1:p-HIF-1α	0.14%	27.92%	1.48%	68.77%	1.69%
TAZ1:HIF-1α	0.21%	0.66%	6.00%	2.93%	90.21%

The kinetics network built from the Markov state model shows that the CITED2:TAZ1:p-HIF-1α ternary intermediate complex (Figure 4.4E) acts as the state connecting the two HIF-1α-bound states (Figures 4.4 B, C) and the CITED2-bound state (Figure 4.4 D). Notably, the CITED2:TAZ1:p-HIF-1α ternary intermediate has a significantly higher transition probability (27.9%) to the CITED2-bound state TAZ1:CITED2 than that to the HIF-1α-bound states: TAZ1:p-HIF-1α (1.5%) or TAZ1:HIF-1α (1.7%), indicating the CITED2:TAZ1:p-HIF-1α ternary intermediate facilitates the dissociation of HIF-1α, which is also consistent with the experiments (14). The partially bound HIF-1α in the CITED2:TAZ1:p-HIF-1α ternary intermediate is more prone to dissociate than the fully bound CITED2 due to the partially lost contacts with TAZ1. The free-TAZ1 state has transition probabilities of 77% and 8.4% to TAZ1:CITED2 and TAZ1:p-HIF-1α, respectively, which suggests CITED2 has much higher association rate on TAZ1 than HIF-1α, primarily due to the strong long-range electrostatic interactions. Our coarse-grained model encodes +9.5 net charges for TAZ1, -7 for CITED2, and -5 for HIF-1α. The columbic potential maps of free TAZ1, TAZ1:CITED2 binary complex, and TAZ1:HIF-1α binary complex are shown in Figure 4.5 and we observe TAZ1:CITED2 largely shifts the sign of the electrostatic potential on the surface, whereas no significant shifts are observed on the TAZ1:HIF-1α surface compared to free TAZ1. Therefore, both the existence of the CITED2:TAZ1:p-HIF-1α ternary intermediate

and the long-range electrostatic forces favor the formation of the CITED2-bound state TAZ1:CITED2.

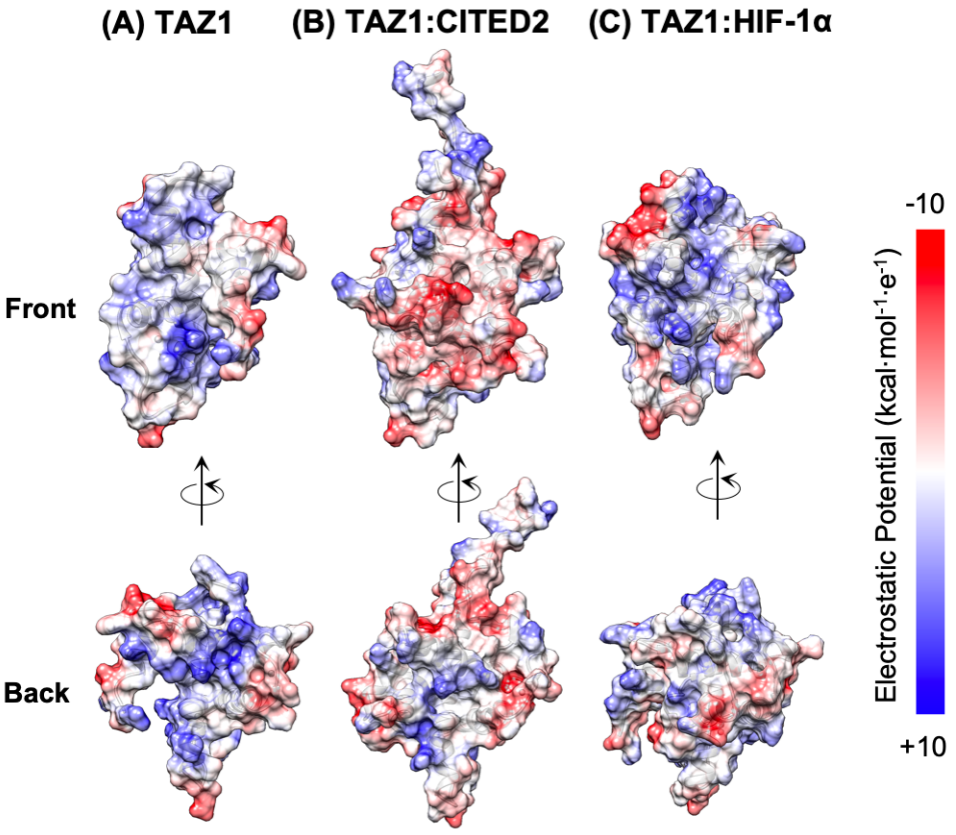

Figure 4.5 Columbic potential maps of free-TAZ1 (A), TAZ1:CITED2 binary complex (B), and TAZ1:HIF-1α binary complex (C).

In the cell, the ionic strength influences the electrostatic interactions through salt screening effects (72). We performed a series of unbiased simulations of the ternary complex at different screening lengths to mimic the effect of different salt concentrations. As shown in Tables 4.3 and 4.4, at very short screening length (1Å and 5Å) or high salt concentration, HIF-1α shows higher binding affinity with TAZ1 than CITED2, which is not surprising since HIF-1α has a higher percentage of non-electrostatic interactions contributing to its binding to TAZ1. At larger screening length (>=10Å) that corresponds to low or moderate salt concentration, CITED2 begin to dominate the

binding of TAZ1. Further increase of the screening length to 15Å, 20Å, and infinity leads to higher population of the CITED2:TAZ1:p-HIF-1α ternary intermediate complex (Table 4.3). Therefore, our simulations suggest ionic strength strongly influences the allosteric effect of the TAZ1 protein switch and low salt concentration facilitates the formation of the CITED2:TAZ1:p-HIF-1α ternary intermediate complex.

Table 4.3 Equilibrium population of the 5 states in the CITED2:TAZ1:HIF-1α ternary complex at different screening lengths when D = 40.

Screening length	free-TAZ1	TAZ1:CITED2	TAZ1:p-HIF-1α	CITED2:TAZ1:p-HIF-1α	TAZ1:HIF-1α
1 Å	81.40%	0.88%	13.47%	0.06%	4.18%
5 Å	34.80%	4.41%	28.69%	0.98%	31.12%
10 Å	0.45%	74.93%	1.20%	14.81%	8.62%
15 Å	0.11%	64.66%	0.40%	34.25%	0.58%
20 Å	0.05%	48.83%	0.51%	50.29%	0.32%
Inf Å[a]	0.03%	14.57%	0.69%	84.41%	0.29%

[a] Infinite screening length.

Table 4.4 K_ds (M) of CITED2 and HIF-1α in the CITED2:TAZ1:HIF-1α ternary complex at different screening lengths when D = 40.

Screening length	K_d (CITED2) / M	K_d (HIF-1α) / M
1 Å	$5.25\times10^{-2} \pm 1.40\times10^{-2}$	$1.94\times10^{-3} \pm 0.44\times10^{-3}$
5 Å	$9.11\times10^{-3} \pm 3.73\times10^{-3}$	$1.28\times10^{-4} \pm 0.36\times10^{-4}$
10 Å	$8.20\times10^{-6} \pm 11.3\times10^{-6}$	$1.24\times10^{-3} \pm 0.44\times10^{-3}$
15 Å	$7.79\times10^{-8} \pm 13.8\times10^{-8}$	$5.93\times10^{-4} \pm 0.90\times10^{-4}$
20 Å	$4.25\times10^{-8} \pm 6.36\times10^{-8}$	$2.31\times10^{-4} \pm 0.26\times10^{-4}$
Inf Å[a]	$6.07\times10^{-8} \pm 10.2\times10^{-8}$	$1.23\times10^{-5} \pm 0.17\times10^{-5}$

[a] Infinite screening length.

4.3 Conclusions

In conclusion, we calculated the K_d values of CITED2 and HIF-1α in the CITED2:TAZ1:HIF-1α ternary complex at different dielectric constants using MD simulations and we found that CITED2 only displaces HIF-1α at low dielectric constant when the electrostatic interactions are sufficiently strong. Previous experimental studies mainly focus on the role of the CITED2:TAZ1:HIF-1α ternary complex (14) and backbone dynamics (66) in explaining the mechanism of the allosteric effect. Our simulations provide an alternative explanation of the negative allosteric regulation in the TAZ1 protein switch with more physical insights, in addition to the mechanism suggested by experiments. The kinetics network built from the Markov state model reveals that the CITED2:TAZ1:HIF-1α ternary intermediate complex is an important state connecting the CITED2-bound and HIF-1α-bound states. Two factors make CITED2 outcompete HIF-1α in the competition of binding to TAZ1. First, the long-range electrostatic interactions allow the fast association of CITED2 to form the TAZ1:CITED2 binary complex, which further discourages the binding of HIF-1α by neutralizing the positive charges on TAZ1. Second, the shared α_A binding site of HIF-1α in the TAZ1:HIF-1α binary complex is prone to be attacked by CITED2, facilitating the formation of the CITED2:TAZ1:HIF-1α ternary intermediate complex, which then favors the dissociation of the partially bound HIF-1α due to the impaired interactions between HIF-1α and TAZ1.

4.4 Methods

Model Setup

The protein data bank (PDB) structures of free TAZ1 (1U2N), TAZ1:CITED2 binary complex (1R8U), and TAZ1:HIF-1α binary complex (1L8C) were used to build the coarse-grained models for MD simulations. This model mainly considers short-range native contacts and long-range

electrostatic forces as intermolecular interactions as described in Chapter 1 of this dissertation. The PDB structures of binary complexes TAZ1:HIF-1α and TAZ1:CITED2 were used to construct the intra-molecular force-field terms of the two peptides HIF-1α and CITED2 and the inter-molecular force-field terms between the two peptides and TAZ1. The PDB structure of the free TAZ1 (1U2N) was used to construct the intra-molecular terms of TAZ1 to avoid potential biases.

Modeling the Zinc Finger

Notably, TAZ1 has three zinc fingers with the zinc atom binding to three adjacent cysteine residues and a histidine residue (Zn-CCCH). Since no zinc finger force-field has ever been built for the EIKB coarse-grained model, we chose the Zinc AMBER Force Field (73) (ZAFF) as a reference and use a single bead to represent each zinc finger. Each coarse-grained zinc finger bead has a net charge of -1 as calculated from the ZAFF charge distributions. A harmonic bond potential with an empirical force constant of 50 kcal/mol is used for the four Zn-residue bonds of each zinc finger bead. A harmonic angle potential with an empirical force constant of 30 kcal/mol is used for the residue1-Zn-residue2 angle force field of each zinc finger bead.

Calibration of the Model

Previous studies have shown that the EIKB Gō model tends to overestimate the intrinsic helicities of intrinsically disordered proteins (IDPs) while underestimating the strengths of IDP-protein interactions. To address these two issues, we adopted an approach similar to previous studies (10, 11). Two scaling factors α and β were used to scale the intra- and inter-molecular interactions so that the model recapitulates some fundamental experimental results. The scaling factor α tunes the intrinsic helicities of HIF-1α and CITED2 and was added to scale all intra-molecular native contacts of the two peptides. Since experiments show the unbound HIF-1α and unbound CITED2 are highly disordered (63, 64), we empirically set the values of α to be 0.05 for both peptides. To

balance the short-range native contacts and the long-range electrostatic interactions between TAZ1 and the two peptides, all TAZ1-peptide native contacts interactions were scaled by a factor β so that the modeled binary complex reproduces the experimental K_d = 10 nM at the optimal value of β (β_{opt}). The dielectric constant (D) was used to modulate the strengths of electrostatic interactions and was examined at 80, 60, 50, and 40 to mimic decreasing strengths of electrostatic interactions. The values of β_{opt} of the two peptides at different dielectric constants were obtained by the Hamiltonian replica exchange (HREX) method developed previously (68). The scaling factor β plays the role of balancing the short-range and long-range forces of the model so that the K_ds of the two simulated binary complexes are always kept at the experimental value 10 nM.

Molecular Dynamics Simulation Protocol

We used an enhanced sampling method based on Hamiltonian replica exchange (HREX) that has shown success in modeling the positive allosteric effect in the KIX domain of CBP/P300 (68). This method uses two variables β and the temperature (T) in the Hamiltonian as the two exchange variables. The values of β for HIF-1α ($\beta_{\mathrm{TAZ1:HIF-1\alpha}}$) were chosen to be span the range from 1.40 to 1.60 with an increment of 0.02. The values of β for CITED2 ($\beta_{\mathrm{TAZ1:CITED2}}$) were chosen to be span the range from 1.10 to 1.30 with an increment of 0.02. The values of T were set to be 300 K, 320 K and 340 K. Therefore, each HREX simulation has 11 β windows and 3 temperature windows with a total of 33 combined simulation windows. Only the trajectories at 300 K were used for the data analysis. The cartesian coordinates of the two nearest windows (T_i, β_i) and (T_{i+1}, β_{i+1}) were exchanged every 10,000 steps using the Metropolis algorithm (53). All molecular dynamics (MD) simulations were carried out using the OpenMM library. The HREX method was implemented using the OpenMM C++ API and Message Passing Interface (MPI). All MD simulations used the Langevin integrator with a friction coefficient of 0.1 ps^{-1} and a timestep of 22 fs to propagate the

equation of motion. Periodic boundary conditions with a 150 Å cubic box was applied for all simulations. For the HREX simulations, 400 million steps were simulated for each replica. The HREX simulation yields the K_d vs. β data (Figure 4.2). For each HREX simulation, the first 50 million steps were discarded prior to data analysis. The optimal values of β for the two peptides ($\beta^{opt}_{TAZ1:HIF-1\alpha}$ and $\beta^{opt}_{TAZ1:CITED2}$) that reproduce the experimental K_d of the binary complexes were calculated from the K_d vs. β plot through curve interpolation using equation 4.1, which describes the relationship between K_d and β with the entropy change ΔS, enthalpy change ΔH, and temperature T as three parameters. The derivation of this equation is shown in a previous paper (68). For each unbiased simulation, we simulated 30 billion steps (0.66 ms) for binary systems and 60 billion steps (1.32 ms) for ternary systems with β set to β^{opt} calculated from HREX. Snapshots were collected every 10,000 steps for data analysis.

$$\ln(K_d) = \ln\left(\frac{1660}{V}\right) - \frac{T\Delta S}{RT} + \frac{\Delta H}{RT}\beta - \ln\left\{exp\left(\frac{T\Delta S}{RT} - \frac{\Delta H}{RT}\beta\right) + 1\right\} \quad (4.1)$$

Data Analysis.

The dissociation constant K_d is calculated based on the fraction of native contacts (Q) formed between TAZ1 and the ligand. Native contact was considered formed if the pair distance is within 1 Å of the native distance. The ligand is considered to be bound if Q > 0.1 and the associated K_d is calculated from the fraction of unbound states P_u (number of unbound snapshots / total number of snapshots) using equation 4.2.

$$K_d = \frac{1660}{V} \times \frac{P_u^2}{1-P_u} \quad (4.2)$$

Markov State Model Analysis.

The unbiased trajectories provide valuable kinetics information of the TAZ1 protein switch and can be better analyzed by a Markov state model (70, 71). As discussed in the main text, we consider CITED2 to be either bound (Q > 0.1) or unbound (Q < 0.1); and HIF-1α to be bound (Q > 0.3),

partially bound ($0.1 < Q < 0.3$), or unbound ($Q < 0.1$). We build a Markov state model using the unbiased trajectory of the CITED2:TAZ1:HIF-1α ternary complex with 5 states based on this classification: TAZ1 (free-TAZ1), TAZ1 with bound CITED2 (TAZ1:CITED2 complex), TAZ1 with partially bound HIF-1α (TAZ1:p-HIF-1α complex), TAZ1 with bound HIF-1α (TAZ1:HIF-1α complex), and TAZ1 with bound CITED2 and partially bound HIF-1α (CITED2:TAZ1:p-HIF-1α complex), as shown in Figure 4.3. We chose 200 snapshots (2,000,000 dynamic steps) as the lag time for the model. The Markov state model was built based on the trajectories with discretized states using the MSMBuilder package (70).

Chapter 5 Conclusions

This dissertation describes new advances of coarse-grained models in protein folding and protein-protein interactions. By studying the folding mechanism of SsIGPS, a TIM barrel protein, the KB Gō coarse-grained model was used to study a large protein (>200 residues) for the first time. The simulations show good overall agreement with the experiments in capturing the regions that fold first and in capturing a rate-determining state. The simulations not only enhance our understanding of the folding mechanism of SsIGPS by providing atomic-level resolution but also consolidate the robustness of the KB Gō model in modeling large proteins. By studying the allosteric regulations in the KIX and the TAZ1 domain of the CPB/P300 transcription co-activator, the EIKB Gō coarse-grained model shows success in modeling both positive/negative allosteric effects. The HREX enhanced sampling method developed in this dissertation provides a framework to rapidly calibrate this type of model and to efficiently calculate the dissociation constant, which is essential for studies of the allosteric effect. The EIKB model is a variant of the original KB model with explicit electrostatics. The MD simulations using the EIKB model found two vastly different allosteric mechanisms in KIX and TAZ1. The cooperative/positive allosteric effect in KIX is through a reduced entropy mechanism in which a prebound ligand reduces the entropy cost for the second ligand to bind, whereas the negative allosteric effect of the TAZ1 protein switch is due to electrostatic forces in which the strong electrostatic interactions between CITED2 and TAZ1 allow CITED2 to efficiently displace HIF-1α. From these case studies as described in this dissertation,

we can conclude that the coarse-grained model is a powerful tool to investigate and quantify the mechanism of protein folding and protein-protein interactions.

Though coarse-grained modeling provides significant insights into some problems of protein folding and protein-protein interactions that cannot be easily answered by experiments, it is still far from being perfect. Here I wish to share some possible future directions. One possible direction is in the field of trajectory analysis. MD simulations generate trajectory data which are high-dimensional time series. In order to make sense of these data and get human-comprehensible knowledge, we typically need to perform proper dimension reductions. However, the choice of the right dimension reduction approach is still very subjective and heavily depends on the researcher's experience and the knowledge of the simulated system. In chapter 2, two conventional physics-based metrics, the radius of gyration and the fraction of native contacts, were chosen as the reaction coordinate to study the protein folding mechanism. However, these physics-based metrics are not perfect and some important states may not be distinguishable using these reaction coordinates. To develop more objective and automated dimension reduction methods, new techniques in the field of machine learning and signal processing may be helpful. For example, one recent methodological progress is called the time-structure based independent component analysis (tICA), which has been shown to be a good dimension reduction method for protein folding (74). The tICA method considers the time correlation of the snapshots and selects the reaction coordinates that correspond to slow timescales. A potential method that can systematically improve the data analysis for coarse-grained modeling is the Markov state model (MSM) (70, 71). MSM provides an automated framework for the MD trajectory analysis that provides a human-comprehensible picture of the underlying processes. Nowadays, most of the applications of MSM are in the trajectory analysis of full-atomic simulations. The major challenge of using MSM for coarse-grained modeling is the

feature engineering in which many commonly used features for full-atomic models (e.g. dihedral angles) may no longer exist nor remain suitable for coarse-grained models.

Another possible future direction is the force field improvement for the coarse-grained model. The general framework of the KB Gō model has been proved to be robust in the past two decades. No obvious deficiency was found in modeling the protein folding processes as compared to the experimental results (75–77). However, some issues were found in the KB Gō model in modeling protein-protein interactions such as the overestimated intrinsic helicities of IDPs (10), the underestimated binding affinities of IDPs (10), and the lack of electrostatics (68). These issues were still not adequately addressed so far. Current method to compensate the overestimated intrinsic helicities of IDPs is to use a scaling factor to scale down the intra-molecular native contacts of the peptide. Though the scaling factor has huge impact on the binding entropy, the method to compute the scaling factor is still qualitative and empirical. Therefore, more rigorous method should be developed in the future. In chapter 3 and 4, electrostatics was found to be important for modeling IDPs in some systems. Therefore, the balance between the short-range native contact interactions and the long-range electrostatic interactions is very important for studying protein-protein interactions. However, the rigorous method to properly balance the native contacts forces and the electrostatic forces remains to be explored in the future. Another interesting problem that has not been fully addressed in this dissertation is to model ions in the KB model. In chapter 5, a zinc-finger force field for the KB model was empirically developed. It would be interesting to develop a general framework to parameterize ions such as zinc for the KB model in the future.

Bibliography

1. Karplus M, Petsko GA (1990) Molecular dynamics simulations in biology. *Nature* 347(6294):631–639.

2. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. *Nat Struct Biol* 9(9):646–652.

3. Shaw DE, et al. (2008) Anton, a special-purpose machine for molecular dynamics simulation. *Commun ACM* 51(7):91–97.

4. Shirts M, Pande VS (2000) Screen savers of the world unite. *Science (80-)* 290(5498):1903–1904.

5. Tozzini V (2005) Coarse-grained models for proteins. *Curr Opin Struct Biol* 15(2):144–150.

6. Clementi C (2008) Coarse-grained models of protein folding: toy models or predictive tools? *Curr Opin Struct Biol* 18(1):10–15.

7. Levitt M, Warshel A (1975) Computer simulation of protein folding. *Nature* 253(5494):694–698.

8. Karanicolas J, Brooks CL III (2009) The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci* 11(10):2351–2361.

9. Hills RD, Brooks CL III (2009) Insights from coarse-grained go models for protein folding and dynamics. *Int J Mol Sci* 10(3):889–905.

10. Ganguly D, Chen J (2011) Topology-based modeling of intrinsically disordered proteins: Balancing intrinsic folding and intermolecular interactions. *Proteins Struct Funct Bioinforma* 79(4):1251–1266.

11. Law SM, Gagnon JK, Mapp AK, Brooks CL III (2014) Prepaying the entropic cost for allosteric regulation in KIX. *Proc Natl Acad Sci U S A* 111(33):12067–12072.

12. Borgia A, et al. (2018) Extreme disorder in an ultrahigh-affinity protein complex. *Nature* 555(7694):61–66.

13. Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair

term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 256(3):623–644.

14. Berlow RB, Dyson HJ, Wright PE (2017) Hypersensitive termination of the hypoxic response by a disordered protein switch. *Nature* 543(7645):447–451.

15. Onuchic JN, Wolynes PG (2004) Theory of protein folding. *Curr Opin Struct Biol* 14(1):70–75.

16. Hills RD, Brooks CL III (2008) Subdomain Competition, Cooperativity, and Topological Frustration in the Folding of CheY. *J Mol Biol* 382(2):485–495.

17. Bartlett AI, Radford SE (2009) An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. *Nat Struct Mol Biol* 16(6):582–588.

18. Kathuria S V., Day IJ, Wallace LA, Matthews CR (2008) Kinetic Traps in the Folding of βα-Repeat Proteins: CheY Initially Misfolds before Accessing the Native Conformation. *J Mol Biol* 382(2):467–484.

19. Baldwin RL (1996) On-pathway versus off-pathway folding intermediates. *Fold Des* 1(1):R1–R8.

20. Nishimura C, Dyson HJ, Wright PE (2006) Identification of native and non-native structure in kinetic folding intermediates of apomyoglobin. *J Mol Biol* 355(1):139–156.

21. Matsumura Y, et al. (2013) Transient helical structure during PI3K and Fyn SH3 domain folding. *J Phys Chem B* 117(17):4836–4843.

22. Forsyth WR, Matthews CR (2002) Folding mechanism of indole-3-glycerol phosphate synthase from Sulfolobus solfataricus: A test of the conservation of folding mechanisms hypothesis in (βα)8 barrels. *J Mol Biol* 320(5):1119–1133.

23. Gu Z, Rao MK, Forsyth WR, Finke JM, Matthews CR (2007) Structural Analysis of Kinetic Folding Intermediates for a TIM Barrel Protein, Indole-3-glycerol Phosphate Synthase, by Hydrogen Exchange Mass Spectrometry and Gō Model Simulation. *J Mol Biol* 374(2):528–546.

24. Gangadhara BN, Laine JM, Kathuria S V., Massi F, Matthews CR (2013) Clusters of branched aliphatic side chains serve as cores of stability in the native state of the HisF TIM barrel protein. *J Mol Biol* 425(6):1065–1081.

25. Wu Y, Vadrevu R, Kathuria S, Yang X, Matthews CR (2007) A Tightly Packed Hydrophobic Cluster Directs the Formation of an Off-pathway Sub-millisecond Folding Intermediate in the α Subunit of Tryptophan Synthase, a TIM Barrel Protein. *J Mol Biol* 366(5):1624–1638.

26. Gu Z, Zitzewitz JA, Matthews CR (2007) Mapping the Structure of Folding Cores in TIM Barrel Proteins by Hydrogen Exchange Mass Spectrometry: The Roles of Motif and

Sequence for the Indole-3-glycerol Phosphate Synthase from Sulfolobus solfataricus. *J Mol Biol* 368(2):582–594.

27. Kathuria S V., Chan YH, Nobrega RP, Özen A, Matthews CR (2016) Clusters of isoleucine, leucine, and valine side chains define cores of stability in high-energy states of globular proteins: Sequence determinants of structure and stability. *Protein Sci* 25(3):662–675.

28. Kathuria S V., et al. (2013) Advances in turbulent mixing techniques to study microsecond protein folding reactions. *Biopolymers* 99(11):888–896.

29. Borgia A, et al. (2016) Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J Am Chem Soc* 138(36):11714–11726.

30. Rambo RP, Tainer JA (2011) Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers* 95(8):559–571.

31. Kumar ATN, Zhu L, Christian JF, Demidov AA, Champion PM (2001) On the rate distribution analysis of kinetic data using the maximum entropy method: Applications to myoglobin relaxation on the nanosecond and femtosecond timescales. *J Phys Chem B* 105(32):7847–7856.

32. Wu Y, Kondrashkina E, Kayatekin C, Matthews CR, Bilsel O (2008) Microsecond acquisition of heterogeneous structure in the folding of a TIM barrel protein. *Proc Natl Acad Sci U S A* 105(36):13367–13372.

33. Forsyth WR, Bilsel O, Gu Z, Matthews CR (2007) Topology and Sequence in the Folding of a TIM Barrel Protein: Global Analysis Highlights Partitioning between Transient Off-pathway and Stable On-pathway Folding Intermediates in the Complex Folding Mechanism of a (βα)8 Barrel of Unknown Function from B. . *J Mol Biol* 372(1):236–253.

34. Karanicolas J, Brooks CL III (2003) The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: Lessons for protein design? *Proc Natl Acad Sci U S A* 100(7):3954–3959.

35. Yang X, Kathuria S V., Vadrevu R, Matthews CR (2009) βα-Hairpin clamps brace βαβ modules and can make substantive contributions to the stability of TIM barrel proteins. *PLoS One* 4(9):e7179.

36. Kohn JE, et al. (2004) Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc Natl Acad Sci U S A* 101(34):12491–12496.

37. Brooks BR, et al. (2009) CHARMM: The biomolecular simulation program. *J Comput Chem* 30(10):1545–1614.

38. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. *J Mol Graph* 14(1):33–38.

39. Kathuria S V., et al. (2014) Microsecond barrier-limited chain collapse observed by time-resolved FRET and SAXS. *J Mol Biol* 426(9):1980–1994.

40. Motlagh HN, Wrabl JO, Li J, Hilser VJ (2014) The ensemble nature of allostery. *Nature* 508(7496):331–339.

41. Uversky VN, Oldfield CJ, Dunker AK (2008) Intrinsically Disordered Proteins in Human Diseases: Introducing the D 2 Concept . *Annu Rev Biophys* 37(1):215–246.

42. Wright PE, Dyson HJ (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* 16(1):18–29.

43. Wright PE, Dyson HJ (2009) Linking folding and binding. *Curr Opin Struct Biol* 19(1):31–38.

44. Mapp AK, Pricer R, Sturlis S (2015) Targeting transcription is no longer a quixotic quest. *Nat Chem Biol* 11(12):891–894.

45. Krishnan N, et al. (2014) Targeting the disordered C terminus of PTP1B with an allosteric inhibitor. *Nat Chem Biol* 10(7):558–566.

46. Dyson HJ, Wright PE (2016) Role of intrinsic protein disorder in the function and interactions of the transcriptional coactivators CREB-binding Protein (CBP) and p300. *J Biol Chem* 291(13):6714–6722.

47. Thakur JK, Yadav A, Yadav G (2014) Molecular recognition by the KIX domain and its role in gene regulation. *Nucleic Acids Res* 42(4):2112–2125.

48. Goto NK, Zor T, Martinez-Yamout M, Dyson HJ, Wright PE (2002) Cooperativity in transcription factor binding to the coactivator CREB-binding protein (CBP): The mixed lineage leukemia protein (MLL) activation domain binds to an allosteric site on the KIX domain. *J Biol Chem* 277(45):43168–43174.

49. De Guzman RN, Goto NK, Dyson HJ, Wright PE (2006) Structural basis for cooperative transcription factor binding to the CBP coactivator. *J Mol Biol* 355(5):1005–1013.

50. Palazzesi F, Barducci A, Tollinger M, Parrinello M (2013) The allosteric communication pathways in KIX domain of CBP. *Proc Natl Acad Sci U S A* 110(35):14237–14242.

51. Brüschweiler S, Konrat R, Tollinger M (2013) Allosteric communication in the KIX domain proceeds through dynamic repacking of the hydrophobic core. *ACS Chem Biol* 8(7):1600–1610.

52. Ganguly D, Zhang W, Chen J (2013) Electrostatically Accelerated Encounter and Folding for Facile Recognition of Intrinsically Disordered Proteins. *PLoS Comput Biol* 9(11):e1003363.

53. Bussi G (2014) Hamiltonian replica exchange in GROMACS: A flexible implementation.

Mol Phys 112(3–4):379–384.

54. Eastman P, et al. (2017) OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* 13(7):e1005659.

55. Liu P, Kim B, Friesner RA, Berne BJ (2005) *Replica exchange with solute tempering: A method for sampling biological systems in explicit water* doi:10.1073/pnas.0506346102.

56. Fox JM, Zhao M, Fink MJ, Kang K, Whitesides GM (2018) The Molecular Origin of Enthalpy/Entropy Compensation in Biomolecular Recognition. *Annu Rev Biophys* 47(1):223–250.

57. Berlow RB, Dyson HJ, Wright PE (2018) Expanding the Paradigm: Intrinsically Disordered Proteins and Allosteric Regulation. *J Mol Biol* 430(16):2309–2320.

58. Boonstra S, et al. (2018) Hemagglutinin-Mediated Membrane Fusion: A Biophysical Perspective Influenza hemagglutinin (HA): a class I homotrimeric glycoprotein of the influenza virus, responsible for target membrane binding and fusion. *Annu Rev Biophys* 47721(1):1–7.

59. Kasinath V, Sharp KA, Wand AJ (2013) Microscopic insights into the NMR relaxation-based protein conformational entropy meter. *J Am Chem Soc* 135(40):15092–15100.

60. Dragovic RA, et al. (2011) Sizing and phenotyping of cellular vesicles using Nanoparticle Tracking Analysis. *Nanomedicine Nanotechnology, Biol Med* 7(6):780–788.

61. Tompa P (2012) Intrinsically disordered proteins: A 10-year recap. *Trends Biochem Sci* 37(12):509–516.

62. Madan Babu M, et al. (2011) Intrinsically disordered proteins: regulation and disease This review comes from a themed issue on Sequences and topology Edited. *Curr Opin Struct Biol* 21(3):1–9.

63. Dames SA, Martinez-Yamout M, De Guzman RN, Jane Dyson H, Wright PE (2002) Structural basis for Hif-1α/CBP recognition in the cellular hypoxic response. *Proc Natl Acad Sci U S A* 99(8):5271–5276.

64. De Guzman RN, Martinez-Yamout MA, Dyson HJ, Wright PE (2004) Interaction of the TAZ1 domain of the CREB-binding protein with the activation domain of CITED2: Regulation by competition between intrinsically unstructured ligands for non-identical binding sites. *J Biol Chem* 279(4):3042–3049.

65. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6(3):197–208.

66. Berlow RB, Martinez-Yamout MA, Dyson HJ, Wright PE (2019) Role of Backbone Dynamics in Modulating the Interactions of Disordered Ligands with the TAZ1 Domain of the CREB-Binding Protein. *Biochemistry* 58(10):1354–1362.

67. Gao M, Yang J, Liu S, Su Z, Huang Y (2019) Intrinsically Disordered Transactivation Domains Bind to TAZ1 Domain of CBP via Diverse Mechanisms. *Biophys J* 117(7):1301–1310.

68. Wang Y, Brooks CL III (2019) Enhanced Sampling Applied to Modeling Allosteric Regulation in Transcription. *J Phys Chem Lett* 10(19):5963–5968.

69. Lindström I, Andersson E, Dogan J (2018) The transition state structure for binding between TAZ1 of CBP and the disordered Hif-1α CAD. *Sci Rep* 8(1). doi:10.1038/s41598-018-26213-x.

70. Harrigan MP, et al. (2017) MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys J* 112(1):10–15.

71. Scherer MK, et al. (2015) PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J Chem Theory Comput* 11(11):5525–5542.

72. Shammas SL, Travis AJ, Clarke J (2014) Allostery within a transcription coactivator is predominantly mediated through dissociation rate constants. *Proc Natl Acad Sci U S A* 111(33):12055–12060.

73. Peters MB, et al. (2010) Structural survey of zinc-containing proteins and development of the zinc AMBER force field (ZAFF). *J Chem Theory Comput* 6(9):2935–2947.

74. Schwantes CR, Pande VS (2013) Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. doi:10.1021/ct300878a.

75. Hills RD, et al. (2010) Topological Frustration in βα-Repeat Proteins: Sequence Diversity Modulates the Conserved Folding Mechanisms of α/β/α Sandwich Proteins. *J Mol Biol* 398(2):332–350.

76. Nobrega RP, et al. (2014) Modulation of frustration in folding by sequence permutation. *Proc Natl Acad Sci U S A* 111(29):10562–10567.

77. Halloran KT, et al. (2019) Frustration and folding of a TIM barrel protein. *Proc Natl Acad Sci* 116(33):16378–16383.