2019

Flexible Intermediates During SH3 Binding

Gabriella Gerlach
Skidmore College, ggerlach@skidmore.edu

Flexible Intermediates During SH3 Binding

By Gaby Gerlach
December 22, 2018
Skidmore College

SH3 domains are the most common protein interaction domains and are found across all forms of life with at least 400 in humans alone. Theses domains often bind to flexible proteins known as intrinsically disordered proteins (IDPs). However, little is known about the binding mechanism between SH3 domains and their IDP binding partners which tend to be proline rich. One SH3 domain found in yeast, AbpSH3, has a binding site for the IDP ArkA. Molecular dynamics simulations were used to model the binding mechanism of AbpSH3 with ArkA. AbpSH3 is hypothesized to undergo a multi-step binding process with ArkA, beginning with the formation of an encounter complex where an ensemble of ArkA conformations are populated in an equilibrium exchange. The two halves of the ArkA sequence, segments 1 (N-terminal) and 2 (C-terminal), are also believed to bind independently. Segment 1, which contains the PxxP motif, is more structured than segment 2. We characterized the structural ensemble of ArkA alone. Then, we performed simulations of initial binding interactions between the SH3 domain and ArkA. The peptide was initially placed at least 10 Å away from the SH3 domain in explicit water. Upon binding, ArkA sampled a wider range of contacts with the domain, compared to simulations started from the bound structure. This suggests that ArkA is forming a flexible encounter complex with the SH3 domain as a binding intermediate. We also observe that the PxxP motif in segment 1 can bind to the AbpSH3 in both the forward and reverse orientation in the encounter ensemble. We saw agreement, within an order of magnitude, between the ArkA binding rate in our simulations and that determined from experimental data. In the future, we will explore the role of electrostatics in this binding interaction.

Flexible Intermediates During SH3 Binding

By Gaby Gerlach

December 22, 2018

Skidmore College

Table of Contents

Chapter 1: Introduction 3-8

Chapter 2: Methods 9-13

MD Simulations 9-13

Experimental 13

Chapter 3: ArkA Ensemble 14-22

Chapter 4: ArkA binding to AbpSH3 23-32

ArkA12 and s1 binding to AbpSH3 23-29

Markov State Models 29-32

Chapter 5: Discussion 33-37

References 38-44

Supplemental 45-46

Chapter 1: Introduction

Intrinsically disordered proteins (IDPs) play an important role in many cellular functions. IDPs do not have a folded structure like globular proteins, rather they sample a wide range of conformations. There are several factors that make IDPs different from globular proteins including their sequence composition, and their flat energy landscapes without a clear global energy minimum (1, 2). Disorder has a cellular function, but in many cases IDPs must bind to globular proteins to perform these functions. Regions of disorder are now known to be present in between 25% and 41% of eukaryotic proteins, and their flexibility allows functional diversity by having multiple interaction partners (3). Binding of IDPs and disordered regions to globular proteins is prevalent throughout the proteome, but there is little known about what drives these interactions. There are several advantages to using disorder as a method of protein interaction. IDPs can bind with specificity and promiscuity, allowing them to interact with multiple partners (4), and IDPs can better regulate processes which require rapid responses, such as signaling, since they typically have fast turnover rates within cells (5).

When IDPs bind to globular proteins the binding event is often coupled with IDP folding and takes place in at least two steps (6, 7, 8, 9). Folding upon binding leads to one of the most compelling rationales for the value of IDPs in cellular processes. At the thermodynamic level, folding and binding are very similar, both processes involving burial of hydrophobic residues and the formation of hydrogen bonds and salt bridges to minimize free energy. For proteins to bind or fold, their partners must be near enough, so interactions can start. IDPs have a larger “capture radius” than folded proteins of the same length because they generally adopt a more extended conformational ensemble (10). Binding begins with the creation of an encounter complex ensemble when the IDP “dances” on top of the domain it is binding to (11). This initial interaction

is generally driven by one segment of the peptide which has specific residue composition (12). Electrostatic interactions have been shown to drive the formation of encounter complex ensembles which can accelerate association up to 4-fold (12). The specific binding pathway of many IDPs is still unknown including the intermediate steps and the timescale of the encounter complex.

One common IDP binding domain is the SH3 domain. It is conserved through more than one billion years of evolution from yeast to humans, and frequently occurs in protein-protein interaction modules (13). Additionally, most components of the yeast cytoskeleton have mammalian homologues where they play similar roles (14). Their prevalence across all three domains of life and wide range of functions makes understanding their functionality important. We focus on the binding mechanism between an SH3 domain of yeast, Actin Binding Protein 1 (AbpSH3) and ArkA, a portion of the yeast actin patch kinase, Arka1p. AbpSH3 is significant to the actin cytoskeleton through localization of cortical actin patches, actin organization, and endocytosis (14, 15). The structures of AbpSH3 alone and bound to ArkA have been solved by x-ray crystallography and NMR, respectively (13, 16). AbpSH3 has the typical SH3 fold with a five-stranded β-sandwich and long irregularly structured RT-loop (13).

Outside of ArkA, there are other peptides which bind to AbpSH3, but ArkA has the strongest binding, so we focused on this interaction (13). Binding between AbpSH3 and ArkA has great sequence conservation across fungal species and higher eukaryotes making it a relevant system of study (17). There are other biologically relevant interactions over a 20-fold range of affinity. It has also been shown that the 12-residue version of ArkA with truncated C and N terminal ends has a 6-fold reduction in binding affinity compared to the 17-residue peptide, but versions longer than 17 residues do not increase the binding affinity (13). This range of binding

partners makes examining the interaction between AbpSH3 and ArkA particularly interesting because there are several components which are involved.

Based on the solved bound structure, ArkA is divided into two segments (s1 and s2) where s1 is the N-terminal proline rich end and s2 is the C-terminal segment (Fig. 1c). This division is based on its interaction with AbpSH3. The proline rich s1 interacts with AbpSH3 in the typical manner for SH3 domains with a PxxP region, where P is proline and x is any residue, with each Px well packed into a groove (13). The region which binds to PxxP is referred to as surface I (SI) (Fig. 1a). The residues which s2 makes contacts to are distinct from those for s1 and the region is referred to as surface II (SII) (Fig. 1a).

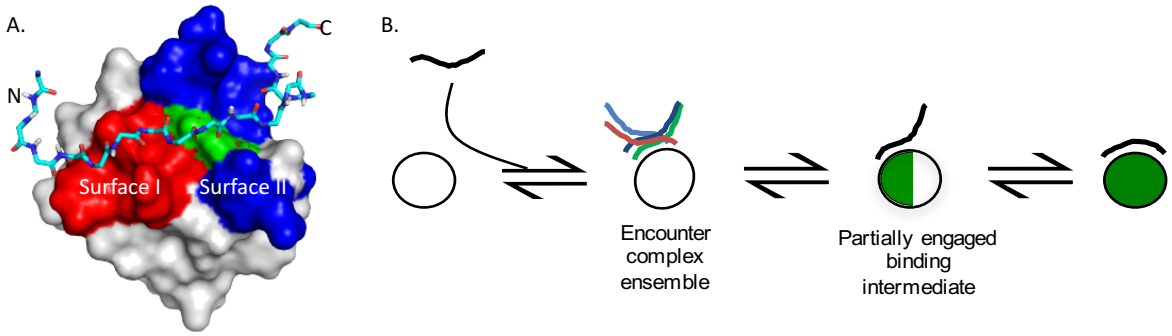

Fig. 1. Description of system studied. A. Surface view of AbpSH3 bound to ArkA showing the two binding pockets (Surface 1 (red) and Surface 2(blue)) and their interface (green) with bound ArkA in sticks the C and N-termini are labeled. B. Proposed binding mechanism with ArkA shown as a line and AbpSH3 as the circle. C. Sequence of ArkA12 used in all simulations with Segment 1 (s1) shown in green and segment 2 (s2) in bold. The protecting groups on the C and N-terminal ends are also shown.

Through binding experiments, it was determined that s1 is required for significant binding and s2 is not. This along with the literature examples of IDP complexes sampling encounter complexes before becoming tightly bound led to the creation of the proposed model of binding for

ArkA with AbpSH3 (8, 18, 19, 20). Generally, the proposed model involves the formation of a loosely bound encounter complex followed by the tight binding of s1 then s2 (Fig. 1b). There is a lot about the binding of ArkA to AbpSH3 that is still not well understood including the intermediates that the complex goes through as it begins to bind. The initial interaction involves an extended IDP. We and others (21, 22) have hypothesized a multistep binding process where the distinct segments contribute to binding through different methods. Similar stepwise binding processes has been seen in other SH3 domains (21) and other extended peptides (18).

IDPs going through multistep binding and gaining structure is a common theme in the field with several specific mechanisms being studied including the phosphorylated kinase inducible activation domain binding to the KIX domain of the CREB binding protein (18), the C-terminal domain of the measles virus nucleo-protein and the X domain of the viral phosphoprotein (9), and the tumor suppressor p53 binding to MDM2 (23), to name a few. These systems have shown there is multistep recognition in the process of the IDPs gaining structure (36). Several IDP complexes, p53-TAD1/TAZ2, HIF-1α/TAZ1, and NCBD/ACTR, have been shown to have an electrostatic driven encounter complex in their binding path (24).

Binding, whether involving an encounter complex driven by electrostatics or not, has been described by a series of models. There are several models of protein binding proposed including conformational selection and induced fit. The major distinction between these models is whether the peptide, in this case ArkA, folds and then binds, or binds and then folds. Though the distinction between these two models was initially treated as a dichotomy, recently several studies have shown a combination of these two mechanisms used in binding (25). In many cases the binding mechanism is not a dichotomy, but a scale with proteins binding both through conformational selection and induced fit (26, 27). Where ArkA and other IDPs lie on this scale is still unknown.

The prevalence of SH3 domains and relevance of the observed PxxP motif makes this an excellent model for understanding binding properties of extended IDPs.

The N-terminal s1 of ArkA contains the PxxP sequence which is common to peptides that bind SH3 domains and previous binding experiments have shown its importance in the binding interaction. There are also examples of SH3 domains binding the same peptide rotated 180 degrees, this shows how important the two xP sites are to the binding process (28). At the interface of SI and SII, the SH3 domain contains a 'specificity pocket,' which is negatively charged and electrostatically interacts with a positively charged residue outside the PxxP motif, K(-3) (28). As the name implies, the specificity in SH3 domain ligand interaction stems from this pocket. Both ArkA and AbpSH3 contain charged residues and electrostatics have been shown to have an important role in IDP binding, so long range electrostatics likely play a significant role in the formation of the initial complex (21, 24).

The effects of changes to ArkA on the rate of binding to AbpSH3 have not been published previously, so this specific analysis will add significant understanding to the binding process. Though there has been work on SH3 domains and their fast recognition of proline rich motifs, this will give a clearer understanding of roles of the C and N terminal residues of ArkA and how tightly s1 can bind alone (29).

Here, AbpSH3 binding to ArkA was examined using Molecular Dynamics (MD) simulations and ^{15}N relaxation dispersion Carr-Purcell-Meiboom-Gill (CPMG) which is an NMR method to study protein dynamics on the microsecond to millisecond scale. Looking at protein folding and binding through MD simulations was until very recently outside the scope of computational power; however, with recent advancements in both hardware and the accuracy of force fields it has become possible to simulate proteins on the microsecond timescale and observe

folding and binding. The binding rates obtained from CPMG are compared to those obtained from MD simulations as a way to examine the accuracy of the simulations. (30)

We are focused on atomic level detail and timescales that are not possible to capture using experimental techniques. MD has recently been extensively used for the study of binding pathways (31, 32, 33). Studying binding with MD simulations allows the use of established analysis methods such as Markov State Models (MSMs) a method to overcome the challenge of analyzing the extremely large data sets that are created by MD simulations carried out over physiologically relevant timescales (34). This method has been used in the study of many simulated binding studies and is established as effective it will be applied to ArkA binding to AbpSH3 (35, 36, 37). Overall, we have gained understanding of the binding mechanism of ArkA and AbpSH3 and begun to construct an overall model of IDPs binding to SH3 domains.

Chapter 2: Methods

MD Simulations

MD simulations were run on five constructs: AbpSH3 bound to ArkA12, AbpSH3 binding to ArkA12, AbpSH3 binding to s1, ArkA12 unbound, and ArkA17 unbound. The simulations of AbpSH3 bound to ArkA were started from the first NMR structure in the bound NMR because all structures were similar, these are referred to as the bound simulations (2RPN). Two different starting structures were used to initiate simulations of ArkA12 alone. One starting structure was obtained from the NMR structure (2RPN) of ArkA bound to AbpSH3 and the other as a fully extended peptide. In all cases, ArkA12 or s1 were edited to have a chloroethyl carbamate protecting group on the C-terminus and an amine-terminal protecting group on the N-terminus for all simulations except the bound simulations which only has the chloroethyl carbamate group. For the binding simulations, the ArkA construct was placed at least 10 Å from AbpSH3, which is further than the cutoff distance, 9 Å, for calculating long-range interactions in these simulations. For the binding simulations, the starting structure of both ArkA constructs came from the ArkA unbound simulations and AbpSH3 from the bound NMR (2RPN).

All simulations were run on Amber 16 using the AmberFF14SB forcefield with frcmod.ff99SB_w_dih modifications (38, 39, 40). The CUDA version of pmemd in Amber 16 was used to run the simulations on GPUs (40). The binding simulations were solvated with TIP3P water in an octahedral box (41), all other simulations were solvated with TIP3P-FB, the same water with a modification (41). The modification makes the model more accurate in terms of dielectric constant and transport properties. The bound structure was solvated such that the edge of the box was at least 9 Å from any peptide or protein. Binding simulations were solvated with the edge at least 12 Å from any peptide or protein. The simulations of ArkA unbound were solvated with water

15 Å from the edge of the peptide. All systems were neutralized with sodium or chloride ions. Simulations were run with 9 Å as the electrostatic cut off distance.

All systems were subject to energy minimization (1000 steps using harmonic restraints with a force constant of 10 kcal/mol, 1000 steps without restraints) where the first 500 steps were steepest descent and the second 500 steps conjugate gradient in both cases. The systems were then subject to heating from 100 to 300 K (harmonic restraints with a force constant of 10 kcal/mol), and equilibration (50 ps with harmonic restraints with a force constant of 10 kcal/mol). All structures, except ArkA alone, were equilibrated again for 200 ps without restraints. The independent simulations were started with each atom given a random velocity based on the time on the wall clock.

ArkA unbound was simulated using Replica Exchange MD (5). 48 replicas were simulated from 290.00 - 425.00 K with geometric spacing to allow for equal exchange probabilities for all replicas (supplemental material) (42). Each replica was equilibrated without restraints for 500 ps. The simulations were run with an integration step every 2 fs and coordinates stored every 5 ps. Three independent simulations of ArkA12 and ArkA17 from an extended peptide and the conformation in the NMR structure were run for 300 ns giving a total of 1.8 μs of simulation for each ArkA construct.

The bound and binding simulations were run with Monte Carlo barostat with new system volumes attempted every 100 steps, an integration step every 2 fs, and coordinates stored every 10 ps. The number and length of all simulations are summarized below (table 1). The replica exchange simulations were run on the XSEDE resource Xstream (43), and all other simulations were run on a local cluster.

Table 1. Summary of simulations run

Construct	# of simulations	Length (ns)	Total Length for construct (μs)
Unbound ArkA12	1	350	1
	1	300	
	1	200	
	2	75	
Unbound ArkA17	1	300	0.575
	1	275	
Bound ArkA12	5	2100	10.5
Binding ArkA12	5	1600	15
	5	800	
	10	300	
Binding s1	5	4500	40.8
	5	3000	
	11	300	

Analysis

To analyze the trajectories, the water was stripped from the simulations using the cpptraj module in the AmberTools16 package (40). The AmberTools 16 package was also used in most analysis including the calculation of dihedral angles, end to end distance, and secondary structure.

Complete Sampling

The running average of secondary structure per residue was used as a measure of completeness of sampling for the unbound simulations. The autocorrelation between replicas was also calculated to ensure the replicas were exchanging as expected (44).

Structural analysis of ensemble

Dihedral RMSD from the NMR structure of ArkA bound to AbpSH3 was calculated as described by Kreiger et al. (45). The distance between the center of masses of s1 and AbpSH3 was also calculated using AmberTools16 and compared to the bound simulations and NMR structure. The dihedral angles were also used to calculate the polyproline II helix (PPII) content as described

by Masiaux et. al. (46). Residue distances were calculated based on the Cα for each residue in ArkA and AbpSH3, and 8 Å was used as the cut off distance to define a contact. Contact maps were created based on the percentage of the simulation during which residue contacts were made.

Calculation of k_{on}

The rate constant for binding of ArkA12 and s1 to AbpSH3 was calculated from the 20 independent binding simulations,

$$rate = k_{on}[Abp1SH3][ArkA] \quad (1)$$

the rate and concentrations were calculated from the simulations. Both the rate and concentration were based on having one AbpSH3, ArkA, and thus one binding event. The concentrations are one divided by the volume of the box which is then converted to Molar. The rate is one binding event per time. The volume of the boxes varied slightly over the simulation time and between the two ArkA constructs, so the concentrations and rates were slightly different (table 2).

Table 2. Summary of volume, concentration, and rate for the two binding simulations.

	Volume (Å^3)	Concentration (mM)	Rate (M s^{-1})
Segment 1	253000 ± 29000	6.86	1.24 x 10^5
ArkA12	230000 ± 700	7.22	4.03 x 10^5

The time is the first frame where the distance between the center mass of s1 and AbpSH3 is below 13 Å. This distance is based on the distance that the bound simulations stay under for 96% of simulation time. Equation 1 was rearranged,

$$k_{on} = \frac{rate}{[Abp1SH3][ArkA]} \quad (2)$$

and using the values calculated above k_{on} was determined for each simulation and averaged. The multiplicative factor that separates the two ArkA constructs was calculated by dividing the two k_{on} values and used to compare simulated k_{on} to the one determined through NMR.

Markov State Models

The MSMs of both ArkA12 and s1 encountering AbpSH3 were built using PyEMMA, a python module (47). The backbone torsion angles and Cα distances were the features used to construct the models. Time independent component analysis was run to lower the degrees of freedom present in the simulations. The first and second time independent components (TIC) were used to construct microstates. K-means clustering was performed with 500 clusters. The lag time for K-means, which is the autocorrelation time, was determined to be 500. Perron Cluster Cluster Analysis (PCCA) was used to group microstates into three macrostates. The MSM was coarse grained into a Hidden Markov Model (HMM) and the transition rates and stationary distributions were determined. The MSMs were plotted over the energy graph of the first two TICs.

Experimental

To measure the binding rates for ArkA and s1 experimentally, ^{15}N relaxation dispersion CPMG was used. The experiments used 0.5 mL sample of 1 mM ^{15}N labeled domain with 5% bound to unlabeled ArkA. The data was collected at two different static magnetic field strengths and comes as a series of 2D NMR spectra. The spectra were processed using standard software developed by our collaborator (48) which has been applied to a few other domains (49, 50). The binding rate constant, k_{on} determined from this analysis was compared to that calculated from the simulations.

Chapter 3: ArkA ensemble

Before examining the ArkA-AbpSH3 interaction, it is important to examine the free ArkA ensemble to understand how prevalent the bound state is when ArkA is free. As it is an IDP, ArkA has a broader conformational ensemble than free globular proteins, so the NMR structures of it bound to AbpSH3 does not show the conformations of ArkA that are less populated or not possible when bound to AbpSH3. To examine the less populated and not possible parts of the ensemble, simulations were run, but because it is disordered, traditional MD simulations were unlikely to capture the missing pieces. Advanced sampling techniques allow proteins to cross energy barriers that are not possible to overcome in normal MD. There are several that have been developed including Umbrella Sampling (51), Infrequent Metadynamics (52), Gaussian Accelerated Molecular Dynamics (GaMD) (53), and Replica Exchange Molecular Dynamics (REMD) (54).

Here, REMD was used to obtain more complete sampling of ArkA. REMD has been used to study IDPs previously including showing the transient secondary structure that makes amyloid-β42 much more neurotoxic than amyloid-β40 both of which are found in the brains of Alzheimer's patients (55). The convergence of REMD simulations has also been explicitly studied using the number of unique states visited, sampling errors, and relative RMSD showing that for a disordered peptide solvated in water, like ArkA, REMD gives acceptable sampling (56). REMD has also shown that binding can stabilize IDPs, in the case of the NCBD protein the helix distribution seen was in agreement with experimental data and pointed toward a binding mechanism that utilized both conformational selection and induced fit (57). The roles of induced fit and conformational selection in ArkA binding to AbpSH3 is of interest. There are several other examples of REMD characterizing states that are not seen experimentally or with straight MD making it an established method for examining IDP states (8, 9, 10).

REMD works by running parallel simulations at a number of different temperatures some above and some below the temperature of interest. The higher temperatures enhance the probability of sampling high energy conformations. The neighboring simulations can exchange based on a Boltxmann-weighted probability. The exchange attempts are based on a Monte Carlo criterion; when it is met, the conformations are swapped, and the velocities are rescaled to the new replica. This process is repeated at rapid intervals, so each final simulation is made from a wide range of temperatures which enhances conformational sampling (Fig. 2).

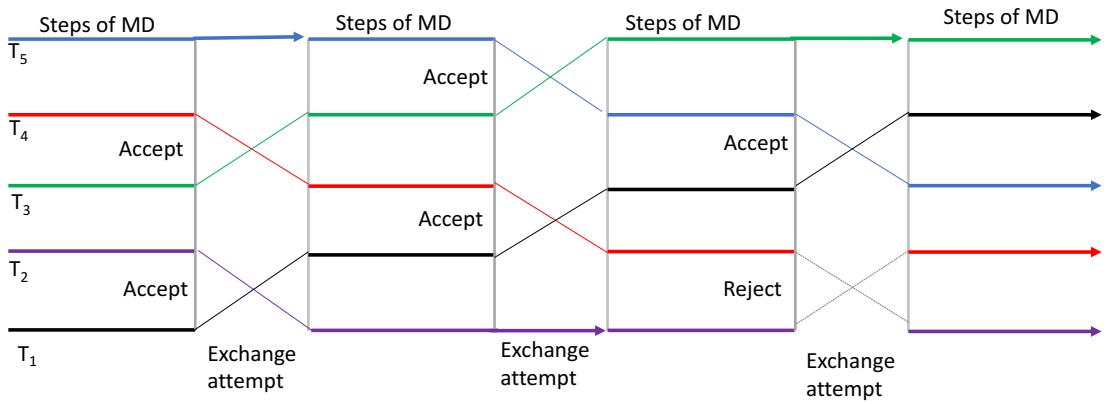

Figure 2. Visual representation of REMD simulations with five temperatures.

When running simulations there is no way to definitively know if you have sampled all possible states. However, it is possible to improve the sampling and have an idea of how complete it was. Here, we used two different starting structures, one an extended chain of amino acids and the other from the NMR structure of ArkA bound to AbpSH3 (2RPN). For these simulations, AbpSH3 was removed from the NMR structure and ArkA was simulated. Six independent simulations were run for ArkA12 and two for ArkA17. Half of the simulations were started from the extended structure, and half from the NMR structure. The number and length of simulations is summarized in the methods (Table 1).

Convergence of these two starting structures for both ArkA12 and ArkA17 was measured using running averages of percent of the peptide in 3-10 helix, bend, turn, and the end to end distance, the shortest distance between the C and N-terminal ends of the peptide at each frame (Fig. 3 & 4). Both sets of simulations appear to be converging as the difference between each set of simulations decreases as time goes on. The ArkA17 results are less robust because there are fewer independent simulations but do show the beginnings of convergence.

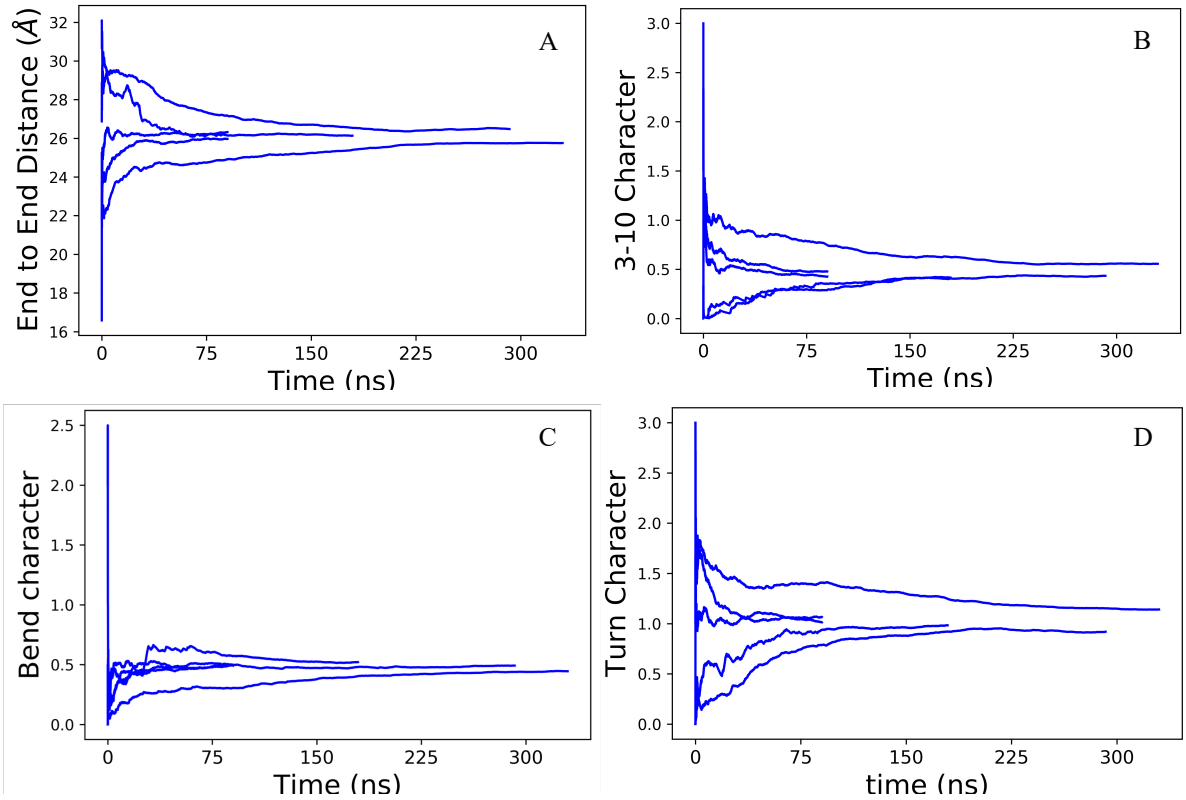

Figure 3. Running average of A. end to end distance B. 3-10 helix character C. Bend Character D. Turn Character of all independent ArkA12 REMD simulations showing the simulations beginning to converge.

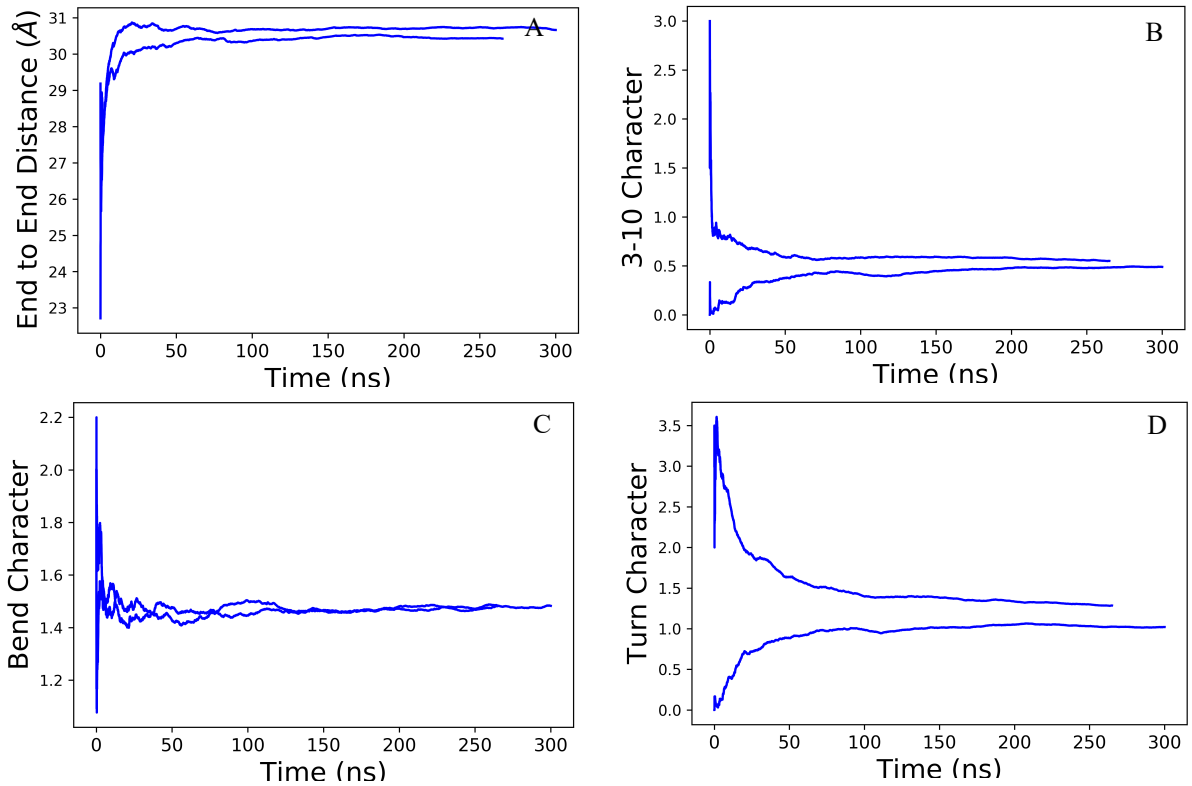

Figure 4. Running average of A. end to end distance B. 3-10 helix character C. Bend Character D. Turn Character of both independent ArkA17 REMD simulations showing the simulations beginning to converge.

The running averages show that the sampling of states is beginning to converge, and to ensure that REMD was exchanging properly the autocorrelation of the replica state index was measured. This shows how long it took for each replica to not be correlated with the temperature it started at. For replica exchange to be functioning correctly, the replicas should be moving across the temperatures, so the states sampled at those higher temperatures are present in the temperature of interest at the end (Fig. 5) (58). After around 12 ns, the replicas are not correlated, so useful exchange is taking place and the number of replicas is sufficient.

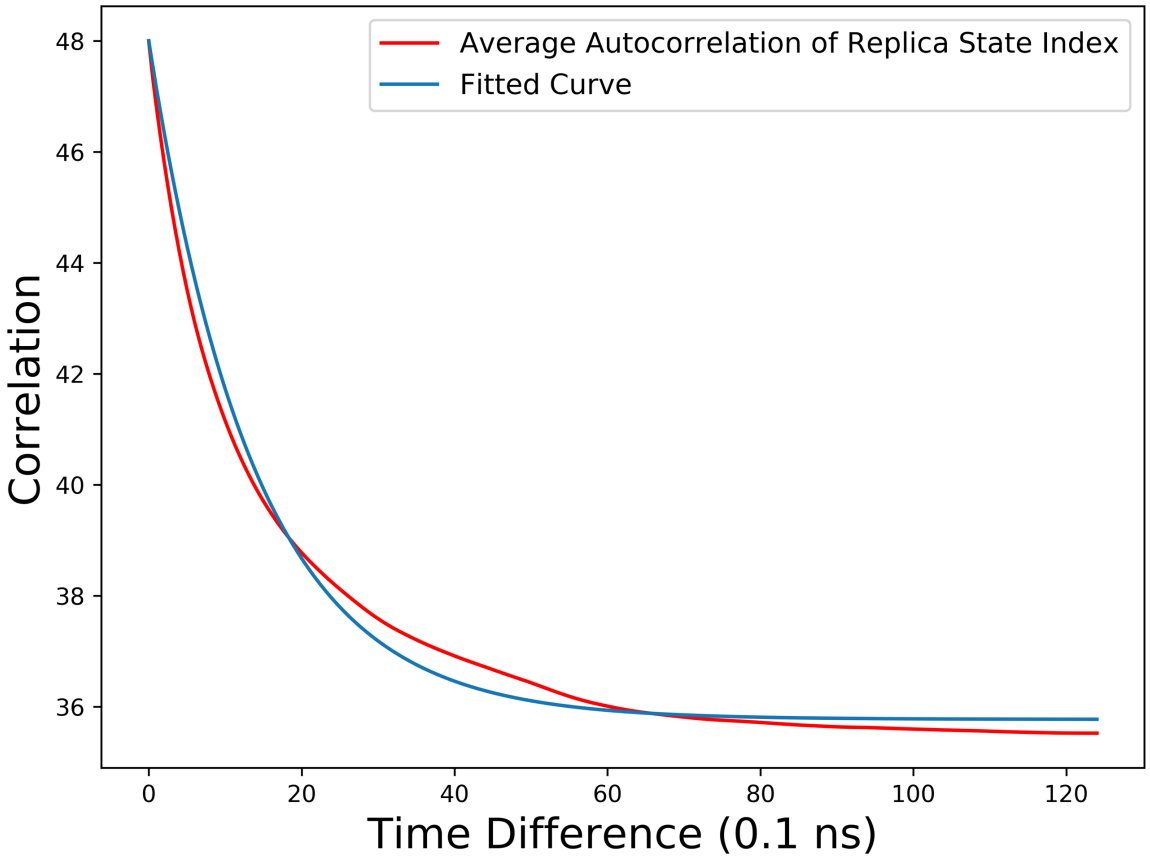

Figure 5. Representative autocorrelation of replica state index graph, for one simulation, showing that the replica exchange was exchanging as expected and the number of replicas was sufficient.

From the REMD data, a structural ensemble was created by using two independent coordinates (45). The first was end to end distance and second Dihedral Root Mean Square Deviation (DRMSD) which is a measure of how similar the dihedral angles of the simulation are to the angles in the NMR structure. This was first done with ArkA12 and six clusters were seen (Fig. 6A). Each cluster represents a conformation of the peptide, based on the two coordinates that are significantly sampled. The clustering was done visually with each of the six clusters having a highly populated center. These clusters were then examined for their different 3-10 helix and Polyproline II helix (PPII) content, which was calculated based on the method used by Mansiaux *et al.* (46).

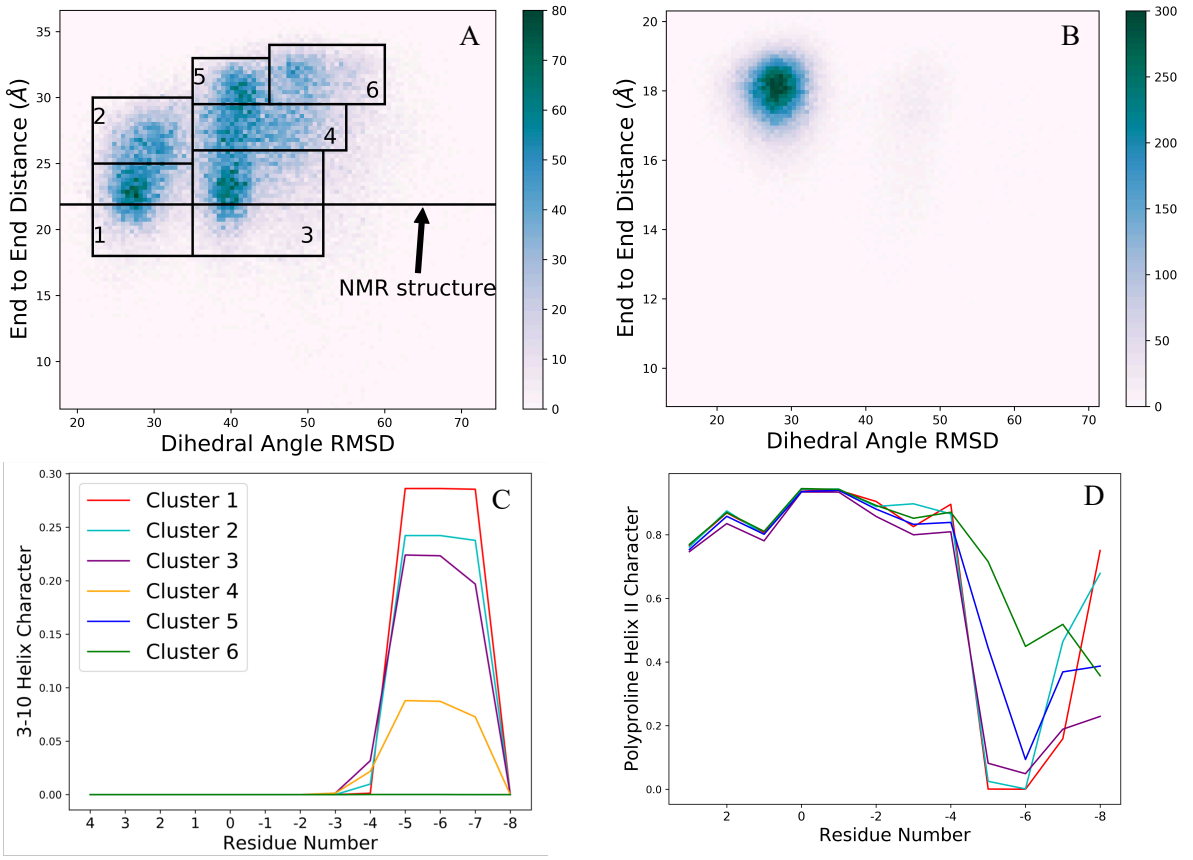

Figure 6. Structrual ensemble of ArkA12. A) ArkA12 graphed by Dihedral RMSD and end to end distance with 6 clusters and the NMR end to end distance labeled. B) Segment 1 strucutral ensemble. C) 3-10 helix character of the 6 Clusters. D) Polyproline II Helix character of the six clusters with same legend as C.

These clusters were analyzed to determine what made them distinct. From examining their PPII and 3-10 Helix content, it was clear that most of the distinction came from s2 of ArkA12 (Fig. 6C & D). When structural ensemble of just the s1 region of ArkA12 was plotted, only one major and one minor conformation were seen (Fig. 6B). This shows that most of the motion in the peptide comes from s2. In contrast, the proline rich C-terminal end is mostly stuck in PPII helix. Segment 2 only samples PPII to an extent comparable to s1 in Cluster 6, the other clusters either do not have a defined secondary structure or sample 3-10 helix. Examining the 3-10 helix graph shows that only at most 30% of the simulation time is spent in this conformation. There was no appreciable sampling of other secondary structures in s2, so for the unaccounted part of the simulation, s2 has

no secondary structure. In general, the clusters with a smaller end to end distance sample more 3-10 helix, and the more extended clusters are more PPII or unstructured.

The preliminary data of ArkA17 was also examined using the two independent coordinates, end to end distance and DRMSD (Fig. 7A). The ArkA17 structural ensemble showed much more concentrated sampling than ArkA12 with what appears to only be 2 clusters at nearly the same end to end distance. The PPII content of ArkA17 was examined as well and shows the same higher content in s1 than s2, but because of the extended N terminus there is an additional spike (Fig. 7B). Additionally, the overall content is closer to 50% than the 80% seen in ArkA12. The amount of the other secondary structures present was also examined and there is more turn and bend content than in ArkA12 and less 3-10 helix (Fig. 7C). Meaning that the extended ends are changing the amount of time spend in helices, but it could also be a result of the limited data. However, more REMD data on ArkA17 is needed to confirm that the ensemble has been sampled completely.

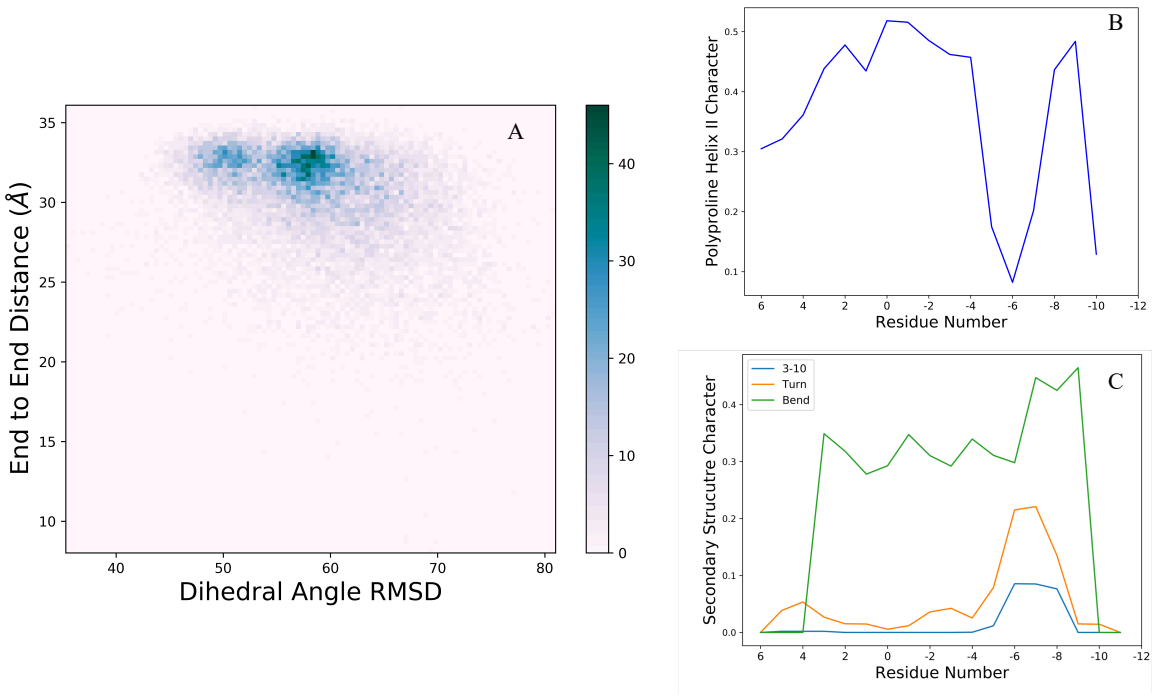

Figure 7. Preliminary analysis of ArkA17 structural ensemble. A. ArkA17 graphed by Dihedral RMSD and end to end distance showing concentrated sampling. B. PPII contenct per residue. C. 3-10 helix content, turn content, and bend content per residue.

A limitation of these simulations is that they do not sample the *cis* state of proline. Proline is unique among the 20 biological amino acids in that it spends up too 20% of the time in the *cis* conformation this number can vary dramatically based on the sequence, length, and temperature of the system (59, 60). This *cis, trans* isomerization has been shown to be a rate determining step in protein folding and binding (59, 61, 62). The specific timescale of the isomerization depends on the system but has been estimated to be as high as four minutes (63). This is far outside the time scale that is sampled with the REMD simulations used here. This could mean the amount of PPII calculated is an overestimate.

States involving *cis* proline are relevant to ArkA-AbpSH3 binding because this interconversion between the *cis* and *trans* states could be a rate limiting step in binding. However,

it is known that *trans* is the dominate conformation and the conformation when ArkA is bound to AbpSH3. It is a reasonable assumption that the states with proline in *trans* would be the ones to interact with and successfully bind to AbpSH3. Therefore, I used conformations from the ArkA12 conformational ensemble to initiate simulations of binding to AbpSH3.

Chapter 4: ArkA binding to AbpSH3

ArkA12 and s1 binding to AbpSH3

The binding of both ArkA12 and s1 to AbpSH3 were simulated using MD with 20 independent simulations of at least 300 ns, and 10 simulations of at least 1.6 μs for each ArkA construct, these are referred to as "binding simulations" (supplemental table 1). In all 40 binding simulations, spontaneous binding was observed. Additionally, five independent "bound simulations" starting from the NMR structure of ArkA12 bound to AbpSH3 were run for 2100 ns to compare with the binding simulations. Until recently, seeing spontaneous binding this consistently in unbiased MD, that is MD simulations without any enhancement, was considered out of the range of current computational power (64). With improvement of hardware and the advent of MD specific supper computers it is possible to simulate the multiple microseconds of data that is needed to observe many biological processes (65). Here, we observed spontaneous binding on the scale of nanoseconds and it occurred in every simulation run.

The simulations were initiated with every atom of either s1 or ArkA12 further than the minimum distance for electrostatic interactions, 10 Å, from every atom of AbpSH3. A bound structure was defined as less than 13 Å between the center of mass of s1 and AbpSH3. This distance comes from the bound simulations of ArkA12 which start from the NMR structure. Those simulations spend 96% time with ArkA12 at or below 13 Å. We do not observe ArkA dissociation from AbpSH3 in these simulations, so they were used to determine an appropriate distance. The binding rate constants calculated from simulation data was compared to relaxation dispersion experiments of the same two binding partners (table 2). The binding rate constant was calculated using all the independent simulations.

Table 2. Comparison of k_{on} from simulations to NMR experiments. Including the ratio of the k_{on} for ArkA12 compared to s1 (multiplicative factor). The error was determined using the standard deviation of the k_{on} values and propagating the error through the division.

	k_{on} (M^{-1} s^{-1})	
Construct	**Simulation**	**NMR**
segment 1	3.28×10^9	3.00×10^7
ArkA12	8.23×10^9	1.60×10^8
ArkA12/segment 1 ratio	3 ± 4	5.33

The on-rate observed in the simulations is faster than what is seen from relaxation dispersion. However, the rates are within 2 orders of magnitude making them similar and a good measure of the accuracy of the simulations. Seeing the same trend with the addition of s2 shows that the simulations are capturing the effect s2 has on binding. In relaxation dispersion, the presence of s2 increases the rate by a factor of 5, but in the simulations, the increase is only by a factor of 3 with uncertainty that makes it possible for s1 to bind faster than ArkA12. Despite this, none of the individual simulations run have on-rates which are faster. This implies that the effects of the addition of s2 are less pronounced in the simulations than they are in the relaxation dispersion experiment. During the binding simulations, the ArkA constructs begin far from AbpSH3, spend some time in a further interacting state we refer to as encounter20, and then reach a closer interacting state referred to as encounter13 (Fig. 8). ArkA is in encounter13 when the center of mass of the ArkA construct is less than 13 Å from AbpSH3, and in encounter20 when the center of mass of the ArkA construct is between 13 and 20 Å. The cut off for encounter13 comes from the bound simulations which spend 96% of simulation time at or below this distance. Encounter20 comes from the furthest distance that the ArkA construct could be while still interacting with AbpSH3. The amount of time spent in encounter20 varies between simulations with some returning to encounter20 from encounter13 for an extended period during the middle of the simulation run (Fig. 9). Each independent simulation spends a different percentage of time in these

partially bound states (supplementary table 1). Encounter20 is not seen in the simulations which start bound, and because the bound simulations do not ever come unbound, the state they are in below 13 Å is likely not the same state as seen in the binding simulations (Fig. 9a).

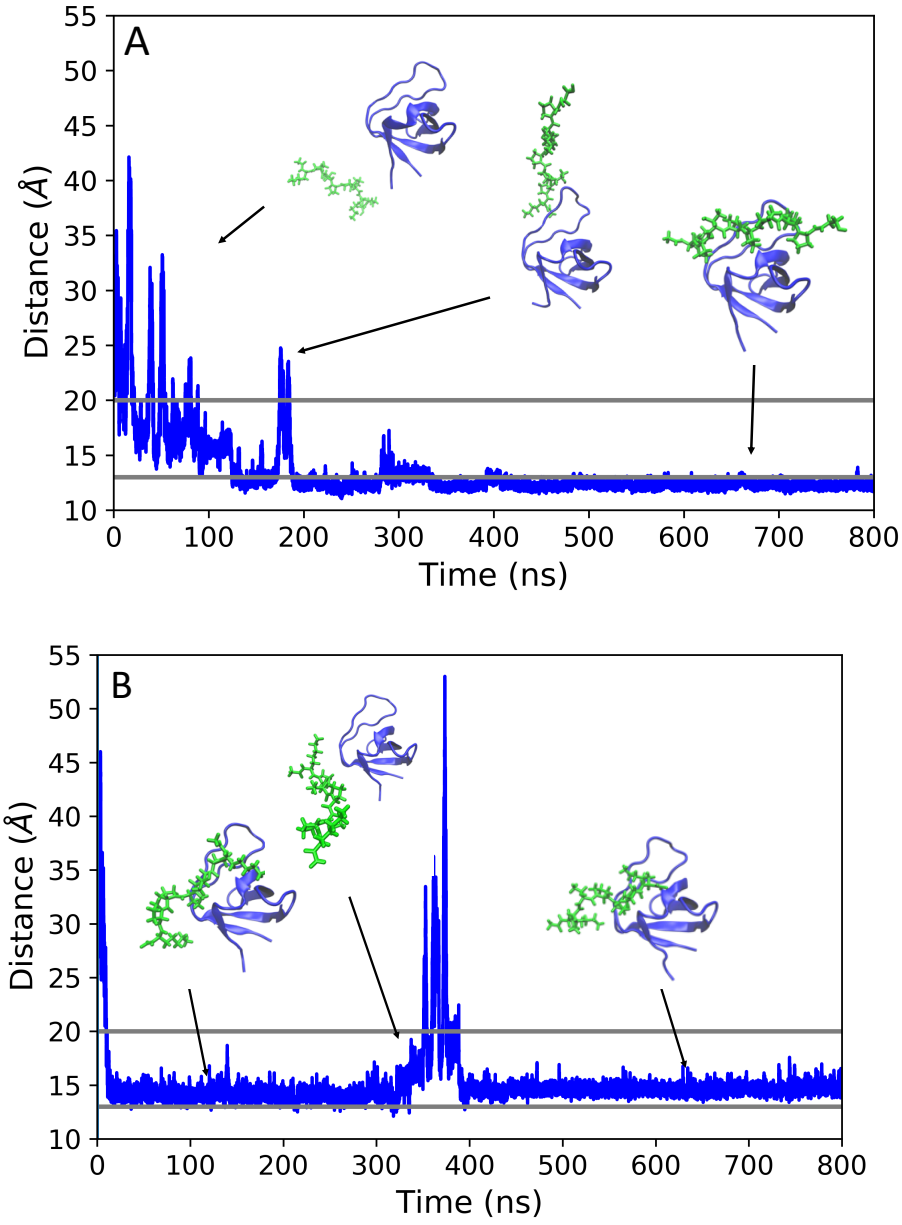

Figure 8. Representative structures of (A) ArkA12 or (A) s1 (green) binding to AbpSH3 (blue) shown with the center of mass distance between s1 and AbpSH3. The grey lines correspond to encounter13, below the first grey line and encounter20 between the two lines.

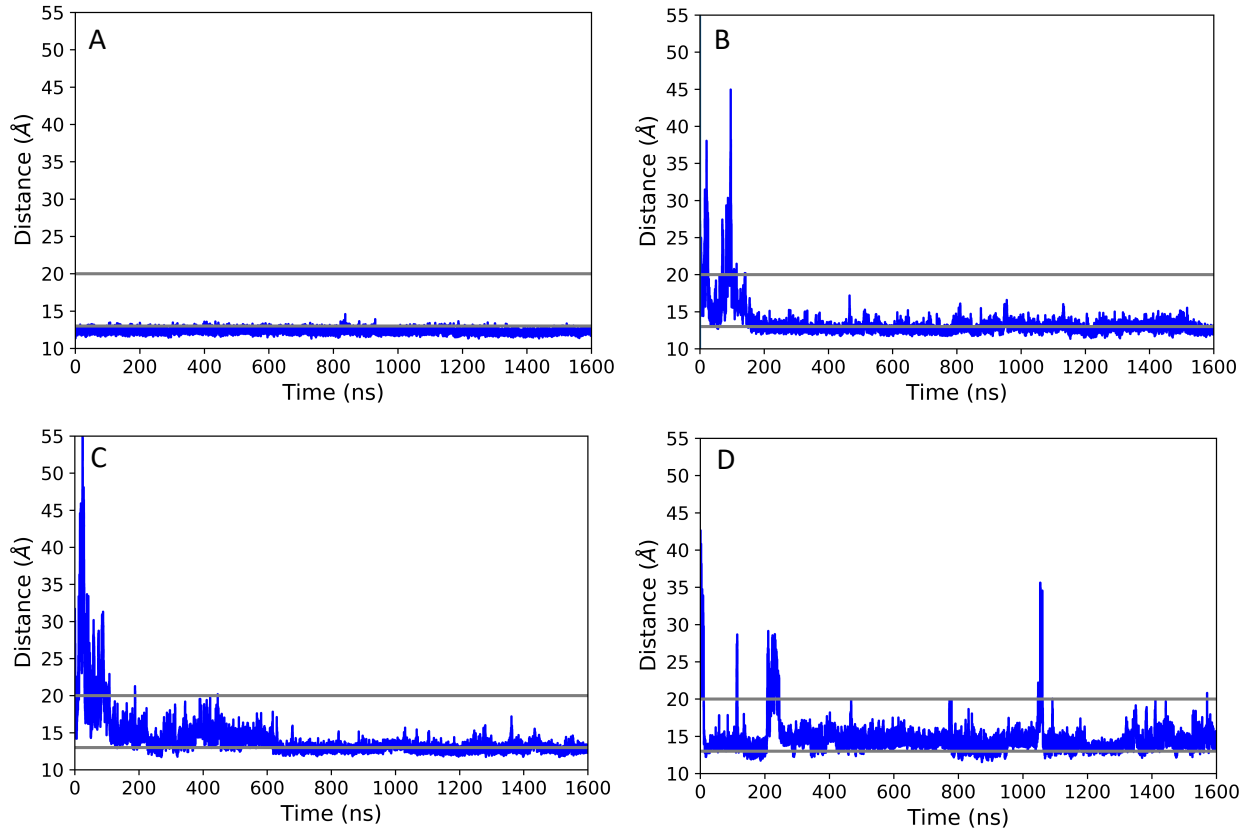

Figure 9. Center of mass distance time course for simulations of ArkA12 starting bound (A), s1 binding (B) and ArkA12 binding to AbpSH3 (C and D), showing the partial unbinding which occurs in the simulations which start separate, but not those that start from the bound structure. The grey lines correspond to encounter13, below the first grey line and encounter20 between the two lines.

In the ArkA12 binding simulations, the percent of time spend in encounter13 is 28%, only two-thirds of the 42% seen in the s1 binding simulations. The addition of s2 allows ArkA to exist in encounter20 more often than is seen with only s1, so s2 stabilizes encounter20. The bound simulations do not sample encounter20 and all the binding simulations initially bind in encounter20 before transitioning to encounter13. Since the bound simulations do not sample encounter20 and the binding simulations interchange between encounter13 and encounter20, we conclude that our binding simulations do not reach the fully bound state, and that therefore encounter13 and encounter20 both represent intermediate states or the 'encounter' complex.

The DRMSD of ArkA compared to the conformation seen in the NMR structure (2RPN) was examined for the binding and bound simulations. The simulations of ArkA bound only sample one confirmation, while in the s1 binding simulations it samples a second conformation with a larger DRMSD (Fig. 10). In the ArkA12 binding simulations, the larger DRMD conformation is not always sampled. Without s2, ArkA is capable of getting close to the NMR dihedral angles, but with s2 it becomes more stable in this conformation. We find no correlation between whether the complex is in encounter13 or encounter20 and the ArkA DRMSD. This means that the distinct states described by the DRMSD are not the same as what is seen with measuring center of mas distance.

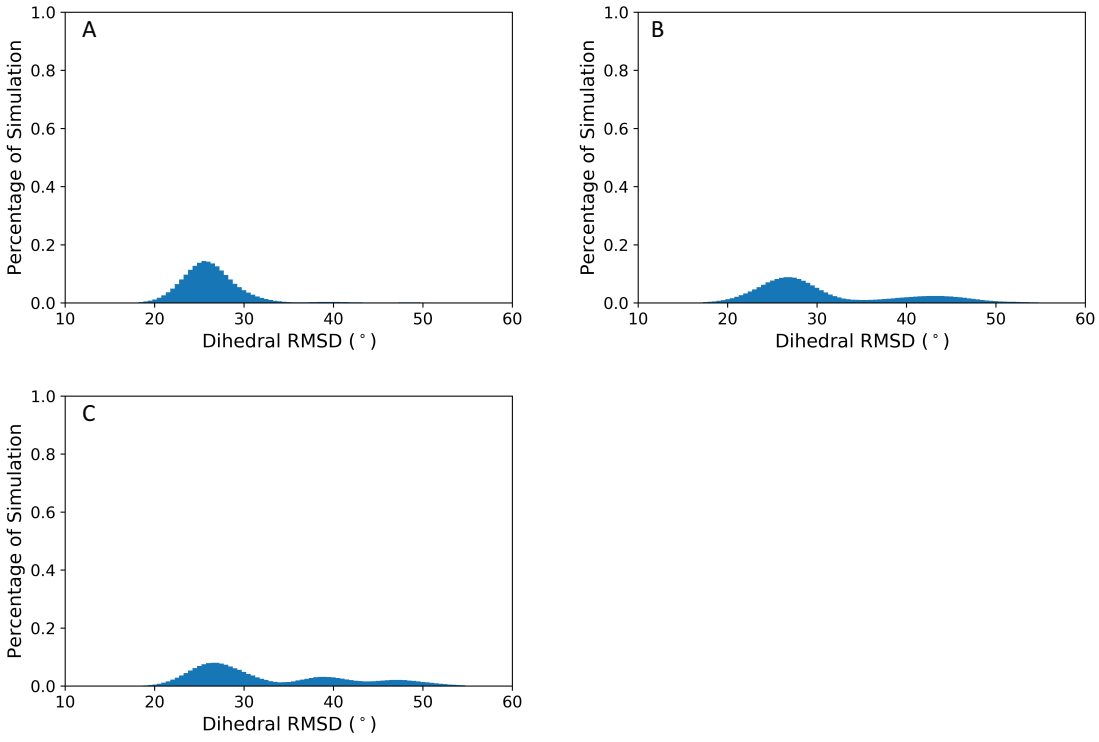

Figure 10. DRMSD of the ArkA peptide measured in degrees, for A) of ArkA12 bound to AbpSH3, B) s1 binding to AbpSH3, and C) ArkA12 binding to AbpSH3.

The contacts made between each ArkA construct and AbpSH3 were compared using contact maps. These measure the normalized percentage of time the Cα of an ArkA residue is within 8 Å of the Cα of an AbpSH3 residue. The binding simulations have broader contacts across ArkA compared to the bound simulations, and consequently have lower individual percentages (Fig. 11). There are not contacts made by the binding simulations to residues outside SI and SII of AbpSH3. However, there are more nonnative contacts in the ArkA12 binding simulations than the s1 binding simulations. The contact maps are based on simulations after initial binding, so they do not capture contacts made before either encounter13 or encounter20 has been sampled.

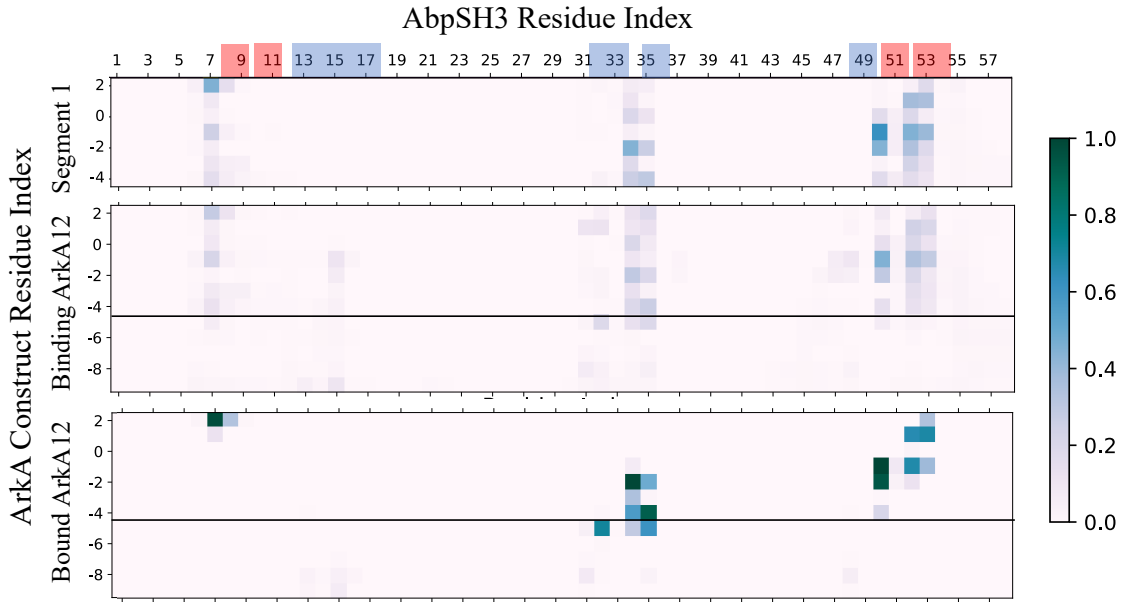

Figure 11. The contacts (distance less than 8 Å) between the residues of AbpSH3 and the ArkA constructs with, when present, a black line separates s2 from s1. Shown is s1 binding (top), ArkA12 binding (middle), and ArkA12 bound (bottom). SI and SII of AbpSH3 are labeled in red and blue respectively.

The contact maps also show that it is possible for s1of ArkA to bind backward. Binding backward was defined as the C and N-terminal lysines making one of the three contacts made by

the opposite end in the bound ArkA12 graph. During the s1 simulations, the ArkA construct bound backward 22% of the time. This is also reflected in the contact map which shows the wider range of contacts across ArkA.

Markov State Models

To further characterize the binding pathway of ArkA constructs to AbpSH3, including intermediate states and transition rates, MSMs were made for each set of binding simulations (Fig. 12 & 13). Neither of the MSMs have a completely unbound state because both ArkA constructs enter either encounter20 or encounter13 quickly. The distinct states seen in the MSMs do not correlate to encounter20 or 13 or the different dihedral RMSD states. However, states 2 and 3 of the ArkA12 model have more different dihedral RMSD than the other states which get closer to the bound state. This means that the MSMs are capturing distinctions that are not seen in the other methods we are using. Additionally, looking through the representative structures for each cluster there are large visual variations within each cluster. Multiple clusters contain states that appear unbound for both ArkA12 and s1 binding to AbpSH3. For these models to be more descriptive the features used to construct the model should be optimized and the lag time adjusted.

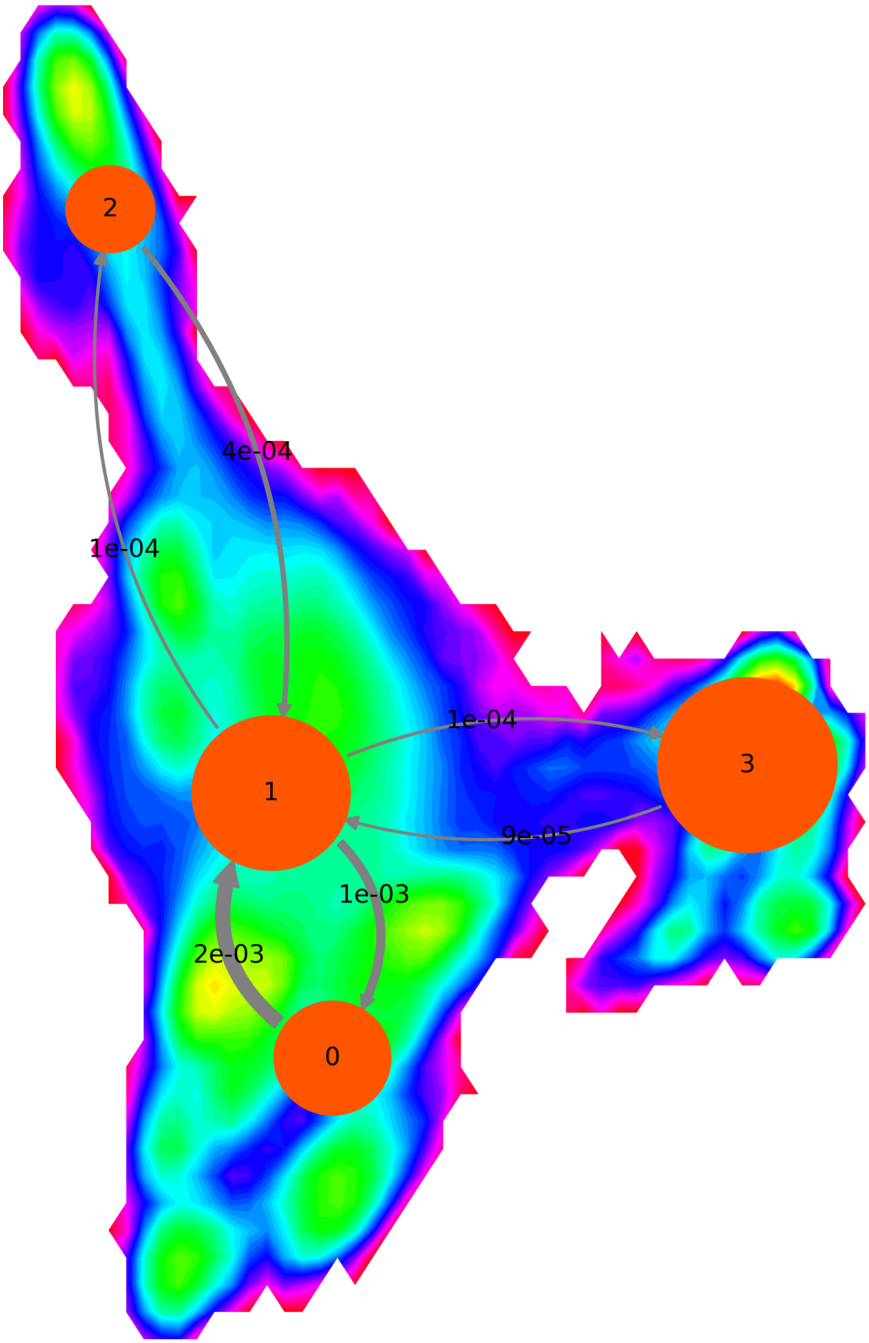

Figure 12. MSM of s1 binding to AbpSH3. Four distinct states are seen with the transition probabilities shown in black over the arrows.

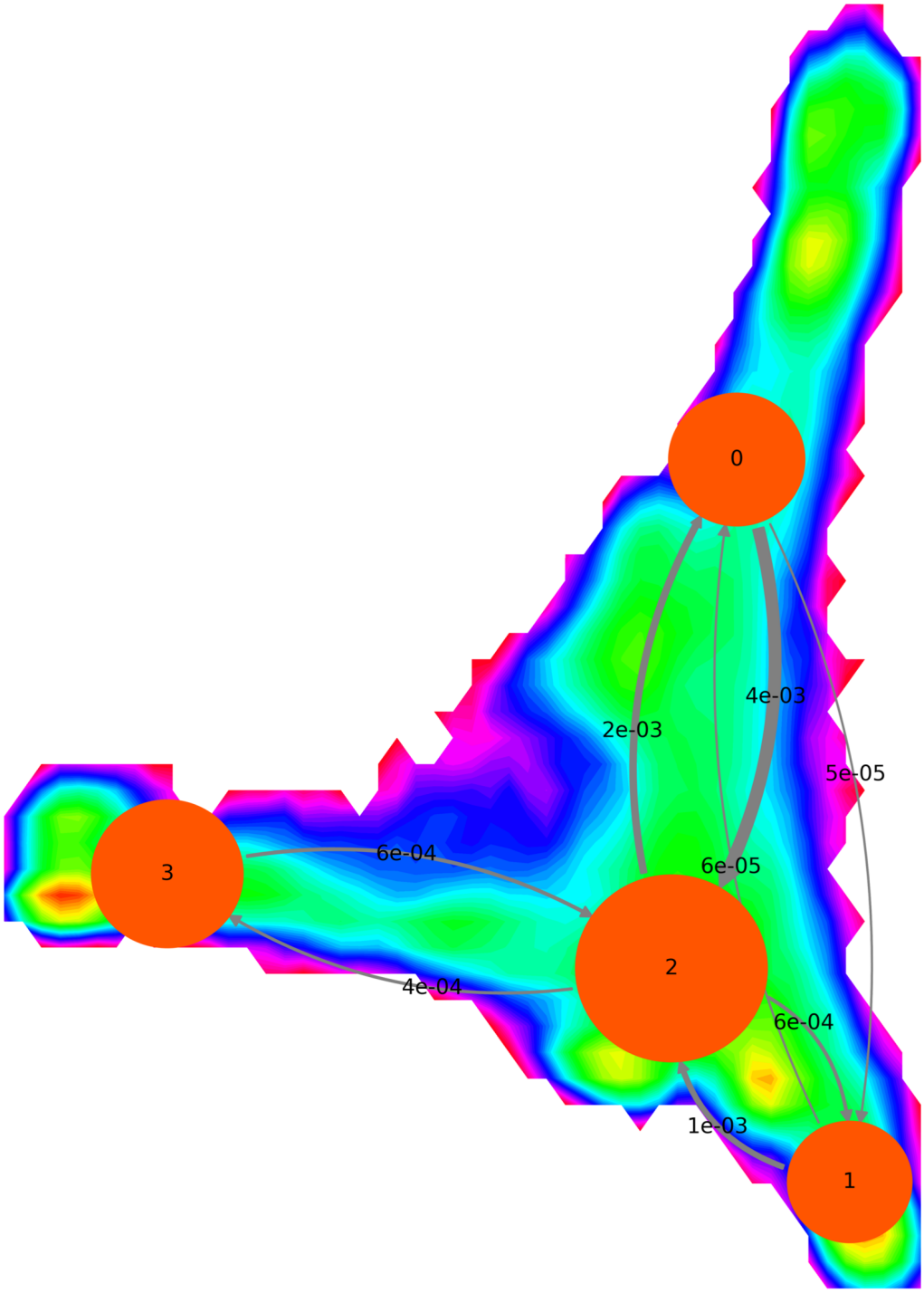

Figure 13. MSM of ArkA12 binding to AbpSH3. The four distinct states are shown with probability of transition.

The binding rates of both ArkA constructs were calculated and are within 2 orders of magnitude of the rates determined from NMR experiments. Neither ArkA construct interacts with portions of AbpSH3 outside of S1 and SII which contributes to the fast binding rate. The binding simulations sample two initial binding states, encounter13 and encounter20, and they switch between the two states. ArkA12 spends more time in encounter20 than when only s1 is binding. The addition of s2 also stabilizes the state with a lower DRMSD and leads to more nonnative contacts. Because the bound simulations do not sample encounter20, it seems that ArkA12 does not reach the fully bound state in the binding simulations. Though the MSMs do not correlate to encounter13 and 20 they do reinforce the conclusion that there are multiple intermediates during the binding process.

Chapter 5: Discussion

In our simulations of ArkA binding to AbpSH3, we have captured the encounter ensemble that is in our proposed binding model (Fig. 1b). From the large differences between the conformational ensembles of the bound and binding simulations we conclude that the binding simulations have not reached a completely bound state. The simulations which start from the bound structure do not sample the further encounter state, encounter20. Since the simulations which start bound do not significantly sample encounter20, it is implied that the binding simulations, which all sample encounter20, have not sampled the bound state.

The encounter ensemble is made up of distinct states which are examined using several reaction coordinates. One coordinate is the distance between the center of mass of ArkA and AbpSH3. The distance ranges from below encounter13 to above encounter20. These different distances show that ArkA goes from close to far and can spend time at different places. The DRMSD was also used to characterize the encounter ensemble showing population from 20° to 60° with the largest peak at 27° for the bound and binding simulations. The s1 binding simulations show one other peak at 42°, and the ArkA12 binding simulations have peaks at 36° and 48°. The encounter ensemble has a wide range of DRMSD values. The range of angles that in the encounter ensemble is slightly smaller than what is seen for the unbound simulations of ArkA12 which shows binding has concentrated the ensemble. The encounter ensemble of segment 1 binding also has both forward and reverse binding giving more distinct states. All of these reaction coordinates measure the heterogeneity in the encounter ensemble, but none of these reaction coordinates are correlated in the binding simulations, so there is no one reaction coordinate which captures all the heterogeneity seen in the encounter complex. Understanding these distinct states shows that the encounter complex of the binding model is made of several different states.

The initial binding between ArkA and AbpSH3 takes place between residues in s1 and SI. The motif which is common across many SH3 domains, PxxP, is in s1 of ArkA, so s1 making more contacts is consistent with the current understanding of SH3 domains (13, 28). Even during the bound simulations, s2 does not make a large number of contacts, demonstrating the importance of the s1 motif for the bound state. When unbound ArkA12 was examined, all six clusters were stable in a PPII helix; in contrast, the six clusters of s2 sample some amount of 3-10 helix or have no secondary structure. The flexibility of s2 is maintained as the interaction with AbpSH3 begins, shown by the lack of contacts it makes in the encounter ensemble. The stability of s1 leaves it in a conformation that is able to take advantage of the PxxP binding motif and results in the quick binding that has been observed.

The quick binding is quantified by the binding rate which, like the experimental results, show s1 alone rapidly binds. The contacts maps show that there is no appreciable nonnative interaction between either ArkA12 or s1 and AbpSH3 during binding. When s1 is alone, it can bind in reverse, as s1 is nearly palindromic, meaning even in reverse the PxxP motif can be utilized (66, 67, 68) . The rapid s1 binding rate shows the importance of the interaction between s1 and AbpSH3. There were differences between the simulation and experimental k_{on}, but some disagreement (at least an order of magnitude) was expected. Similar to the NMR experiments, the simulations capture that ArkA12 binds faster than s1 this is a useful metric that the simulations are agreeing with the NMR experiments. In previous work, simulations have been slower, faster, and nearly aligned with experimental work (69, 70, 71). There are several factors that could account for the difference between the simulation and NMR experiments including not sample the *cis* conformation of proline, the salt concentration, the specifics of the water model. The water model we used, TIP3P, moves closer to the speed of water at 70 °C, and likely accounts for part of the

variation between experiments and simulations (72). The water solvating the system moving faster makes the proteins move faster. Often in studies which report a slower binding rate an implicit water model was used (70, 73). Implicit water is when there are no physical molecules of water, rather the force that would be present if there was water is simply calculated. Generally, this method of simulation is less accurate.

The salt concentration of the NMR experiments is expected to affect the binding rate because of the high content of charged residues in both ArkA and AbpSH3. ArkA is positively charged and AbpSH3 is negatively charged, with many of the charged residues concentrated in SI and SII. The large number of charged residues makes it probable that electrostatics play a large role in this binding interaction. One was that electrostatics can be involved is electrostatic steering, which is the role of nonspecific long-range charge interactions to drive a binding interaction. Their role in allowing IDPs to bind has been seen in several systems (74, 75, 76). Electrostatics also are likely involved with the increase in k_{on} with the addition of s2. The third lysine in ArkA12 is in s2 so extending s1 adds another charged residue to interact the negatively charged residues of AbpSH3 which increases the strength of the interaction. The relaxation dispersion experiments were done in 100mM of NaCl, and electrostatic forces are important in the binding mechanisms of SH3 domains (12). Differences in electrostatic interactions between experiments and simulation has been seen as a source of error in k_{on} calculations (6). The salt in the NMR experiments interacts with the charged residues, so they do not as strongly interact with each other, this slows the binding rate. Simulations run with salt could align better with the NMR experiments.

The fast binding could also come from the proline residues which are unique residues because they are capable of sampling the *cis* conformation unlike most residues which only sample the *trans* state. When in the *trans* state, proline forms PPII helices and when in the *cis* conformation

there are Polyproline I helices. Depending on what residues surround the proline and the structure of the protein, the *cis* is sampled between 5-60% of the time (77, 78). When ArkA is free it should be able to sample *cis,* but there is a high-energy barrier that must be crossed which was not reached during the straight or REMD simulations (Chapter 3). Only confirmations with the *trans* conformations are likely to bind, so sampling *cis* would likely show down the binding rate from what is seen currently in the simulations. ArkA binds with all proline residues in the *trans* conformation, so the simulation data still give extensive insight into the binding mechanism but is likely leading to a faster binding rate.

The contribution of induced fit and conformational selection to ArkA binding is unknown, but from the results presented here we propose a role of both models. The conformational ensemble of ArkA changes upon binding, which implies a role of induced fit, but because of the prominent role of the very stable s1 in binding conformational selection is also likely involved. To confirm this, more analysis of the secondary structure of ArkA after binding is required to determine whether s2 becomes more structured after binding. If s2, which has very little structure when unbound, becomes structured after binding that would mean induced fit is important for tight binding.

To continue characterizing this interaction, simulations with unbinding events and salt concentrations comparable to those of the NMR experiments should be run. It is possible that encounter20 was not sampled by the bound simulations because the water box used was more than 2.5 times smaller than that of the binding simulations. When unbinding has been previously simulated, the protein complex was simulated in a larger box to allow for the larger intermolecular distance needed to see unbinding (79). Having this larger area could make it possible for ArkA12 to come unbound, but it is also possible that for unbinding to be seen advanced sampling would

have to be employed. In previous studies of IDP unbinding mechanisms, systems were simulated bound at physiologically relevant temperature and then the temperature was increased to nearly 500 K to observe unbinding (80, 81). Running simulations with salt concentrations comparable to those of the NMR experiments and comparing the k_{on} could explain the portion of the difference that is a result of the screened electrostatic interactions in the NMR experiment.

For further information on the different intermediates in the interaction, the reaction coordinates used to build the MSMs need to be optimized. The MSMs show that there are distinct intermediates sampled over the simulation time, but the states do not correlate with any other measure we are using. Currently, the Cα distances and backbone torsion angles are being used, but because of the large role of electrostatics in the binding interaction using distances between charged residues could give more valuable models. Additionally, more independent simulations with more binding events would improve the model.

The simulations here showed fast and specific binding that only requires s1 of the disordered peptide. The heterogeneity of the encounter ensemble is not characterized by any one reaction coordinate, but there are several flexible binding intermediates. The intermediates which have been characterized do not reach the tightly bound state. The simulation binding rate shows fairly good agreement with the NMR experiments. These results also suggest a role of both induced fit and conformational selection in the binding event, but further examination of the changes in secondary structure after binding is required to confirm this. We have gained understanding of the loosely bound flexible intermediate whose binding effected by changing salt concentration. This has also further shown the importance of the PxxP binding motif when proline rich peptides bind to SH3 domains. This binding event of one IDP to an SH3 domain gives insight to the SH3 binding which takes place throughout all three domains of life.

References

1. Uversky,V.N., Gillespie,J.R. and Fink,A.L. (2000) Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Genetics,* **41,** 415-427.

2. Fisher,C.K. and Stultz,C.M. (2011) Constructing ensembles for intrinsically disordered proteins. *Current Opinion in Structural Biology,* **21,** 426-431.

3. Liu,J., Faeder,J.R. and Camacho,C.J. (2009) Toward a quantitative theory of intrinsically disordered proteins and their function. *Proceedings of the National Academy of Sciences of the United States of America. USA,* **106,** 19819.

4. Dyson,H.J. and Wright,P.E. (2002) Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology,* **12,** 54-60.

5. Wright,P.E. and Dyson,H.J. (1999) Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology,* **293,** 321-331.

6. Gianni,S., Dogan,J. and Jemth,P. (2016) Coupled binding and folding of intrinsically disordered proteins: What can we learn from kinetics? *Current Opinion in Structural Biology,* **36,** 18-24.

7. Kiefhaber,T., Bachmann,A. and Jensen,K.S. (2012) Dynamics and mechanisms of coupled protein folding and binding reactions. *Current Opinion in Structural Biology,* **22,** 21-29.

8. Uversky,V.N. and Dunker,A.K. (2010) Understanding protein non-folding.*Biochimica et Biophysica Acta,* **1804,** 1231-1264.

9. Bonetti,D., Troilo,F., Brunori,M., Longhi,S. and Gianni,S. (2018) How robust is the mechanism of folding-upon-binding for an intrinsically disordered protein? *Biophysical Journal,* **114,** 1889-1894.

10. Shoemaker,B.A., Portman,J.J. and Wolynes,P.G. (2000) Speeding molecular recognition by using the folding funnel: The fly-casting mechanism *Proceedings of the National Academy of Sciences of the United States of America. Sci. USA,* **97,** 8868-8873.

11. Dyson,J.H. and Wright,P.E. (2012) Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology,* **12,** 54-60.

12. Meneses,E. and Mittermaier,A. (2014) Electrostatic interactions in the binding pathway of a transient protein complex studied by NMR and isothermal titration calorimetry. *The Journal of Biological Chemistry,* **289,** 27911-27923.

13. Stollar,E.J., Garcia,B., Chong,P.A., Rath,A., Lin,H., Forman-Kay,J. and Davidson,A.R. (2009) Structural, functional, and bioinformatic studies demonstrate the crucial role of an

extended peptide binding site for the SH3 domain of yeast Abp1p. *The Journal of Biological Chemistry,* **284,** 26918-26927.

14. Barbara Fazi, M Jamie T V Cope, Alice Douangamath, Silvia Ferracuti, Katja Schirwitz, Adriana Zucconi, David G Drubin, Matthias Wilmanns, Gianni Cesareni and Luisa Castagnoli. (2002) Unusual binding properties of the SH3 domain of the yeast actin-binding protein Abp1: Structural and functional analysis. *The Journal of Biological Chemistry,* **277,** 5290-5298.

15. Garcia,B., Stollar,E.J. and Davidson,A.R. (2012) The importance of conserved features of yeast actin-binding protein 1 (Abp1p): The conditional nature of essentiality. *Genetics,* **191,** 1199-1211.

16. Barbara Fazi, M Jamie T V Cope, Alice Douangamath, Silvia Ferracuti, Katja Schirwitz, Adriana Zucconi, David G Drubin, Matthias Wilmanns, Gianni Cesareni and Luisa Castagnoli. (2002) Unusual binding properties of the SH3 domain of the yeast actin-binding protein Abp1: Structural and functional analysis. *The Journal of Biological Chemistry,* **277,** 5290-5298.

17. Haynes,J., Garcia,B., Stollar,E.J., Rath,A., Andrews,B.J. and Davidson,A.R. (2007) The biologically relevant targets and binding affinity requirements for the function of the yeast actin-binding protein 1 src-homology 3 domain vary with genetic context. *Genetics,* **176,** 193-208.

18. Sugase,K., Dyson,H.J. and Wright,P.E. (2007) Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature,* **447,** 1021.

19. Rajamani,D., Thiel,S., Vajda,S. and Camacho,C.J. (2004) Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America.* **101,** 11287-11292.

20. Xue,Y., Yuwen,T., Zhu,F. and Skrynnikov,N.R. (2014) Role of electrostatic interactions in binding of peptides and intrinsically disordered proteins to their folded targets. 1. NMR and MD characterization of the complex between the c-crk N-SH3 domain and the peptide sos. *Biochemistry,* **53,** 6473.

21. Xue,Y., Yuwen,T., Zhu,F. and Skrynnikov,N.R. (2014) Role of electrostatic interactions in binding of peptides and intrinsically disordered proteins to their folded targets. 1. NMR and MD characterization of the complex between the c-crk N-SH3 domain and the peptide sos. *Biochemistry (N. Y.),* **53,** 6473-6495.

22. Uversky,V.N. (2011) Multitude of binding modes attainable by intrinsically disordered proteins: A portrait gallery of disorder-based complexes. *Chemical Society Reviews,* **40,** .

23. Schon,O., Friedler,A., Bycroft,M., Freund,S.M.V. and Fersht,A.R. (2002) Molecular mechanism of the interaction between MDM2 and p53. *Journal of Molecular Biology,* **323,** 491-501.

24. Ganguly,D., Zhang,W. and Chen,J. (2013) Electrostatically accelerated encounter and folding for facile recognition of intrinsically disordered proteins. *PLOS Computational Biology,* **9,** e1003363.

25. Csermely,P., Palotai,R. and Nussinov,R. (2010) Induced fit, conformational selection and independent dynamic segments: An extended view of binding events. *Trends in Biochemical Sciences.,* **35,** 539-546.

26. Wlodarski,T. and Zagrovic,B. (2009) Conformational selection and induced fit mechanism underlie specificity in noncovalent interactions with ubiquitin. *Proceedings of the National Academy of Sciences of the United States of America.,* **106,** 19346-19351.

27. Grünberg,R., Leckner,J. and Nilges,M. (2004) Complementarity of structure ensembles in protein-protein binding. *Structure,* **12,** 2125-2136.

28. Feng,S., Chen,J.K., Yu,H., Simon,J.A. and Schreiber,S.L. (1994) Two binding orientations for peptides to the src SH3 domain: Development of a general model for SH3-ligand interactions. *Science,* **266,** 1241-1247.

29. Ahmad,M., Gu,W. and Helms,V. (2008) Mechanism of fast peptide recognition by SH3 domains. *Angewandte Chemie International Edition,* **47,** 7626-7630.

30. Wang,Y., Martins,J.M. and Lindorff-Larsen,K. (2017) Biomolecular conformational changes and ligand binding: From kinetics to thermodynamics. *Chemical Science,* **8,** 6466-6473.

31. Martínez,L., Sonoda,M.T., Webb,P., Baxter,J.D., Skaf,M.S. and Polikarpov,I. (2005) Molecular dynamics simulations reveal multiple pathways of ligand dissociation from thyroid hormone receptors. *Biophysical Journal,* **89,** 2011-2023.

32. Pabon,N.A. and Camacho,C.J. (2017) Probing protein flexibility reveals a mechanism for selective promiscuity. *eLife,* **6,** e22889.

33. Miao,Y., Huang,Y.M., Walker,R.C., McCammon,J.A. and Chang,C.A. (2018) Ligand binding pathways and conformational transitions of the HIV protease. *Biochemistry (N. Y.),* **57,** 1533-1541.

34. Pande,V.S., Beauchamp,K. and Bowman,G.R. (2010) Everything you wanted to know about markov state models but were afraid to ask. *Methods,* **52,** 99-105.

35. Gu,S., Silva,D., Meng,L., Yue,A. and Huang,X. (2014) Quantitatively characterizing the ligand binding mechanisms of choline binding protein using markov state model analysis. *PLOS Computational Biology,* **10,** e1003767.

36. Thayer,K.M., Lakhani,B. and Beveridge,D.L. (2017) Molecular Dynamics–Markov state model of protein ligand binding and allostery in CRIB-PDZ: Conformational selection and induced fit. *Journal of Physical Chemistry B,* **121,** 5509-5514.

37. Plattner,N. and Noé,F. (2015) Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and markov models. *Nature Communications,* **6,** 7653.

38. Maier,J.A., Martinez,C., Kasavajhala,K., Wickstrom,L., Hauser,K.E. and Simmerling,C. (2015) ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation.,* **11,** 3696-3713.

39. Doshi,U. and Hamelberg,D. (2009) Reoptimization of the AMBER force field parameters for peptide bond (omega) torsions using accelerated molecular dynamics. *Journal of Physical Chemistry B,* **113,** 16590-16595.

40. D.A. Case, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman. (2017) Amber 2017. *University of California, San Fransisco, .*

41. Jorgensen,W.L., Chandrasekhar,J., Madura,J.D., Impey,R.W. and Klein,M.L. (1983) Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics,* **79,** 926-935.

42. Sugita,Y. and Okamoto,Y. (1999) Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters,* **314,** 141-151.

43. Towns,J., Cockerill,T., Dahan,M., Foster,I., Gaither,K., Grimshaw,A., Hazelwood,V., Lathrop,S., Lifka,D., Peterson,G., et al. (2014) XSEDE: Accelerating scientific discovery. **16,** 62-74.

44. Nguyen,T.H. and Minh,D.D.L. (2016) Intermediate thermodynamic states contribute equally to free energy convergence: A demonstration with replica exchange. *Journal of Chemical Theory and Computation,* **12,** 2154.

45. Krieger,J., Fusco,G., Lewitzky,M., Simister,P., Marchant,J., Camilloni,C., Feller,S. and De Simone,A. (2014) Conformational recognition of an intrinsically disordered protein. *Biophysical Journal,* **106,** 1771-1779.

46. Mansiaux,Y., Joseph,A.P., Gelly,J. and de Brevern,A.G. (2011) Assignment of PolyProline II conformation and analysis of sequence – structure relationship. *Plos One,* **6,** e18401.

47. Scherer,M.K., Trendelkamp-Schroer,B., Paul,F., Pérez-Hernández,G., Hoffmann,M., Plattner,N., Wehmeyer,C., Prinz,J. and Noé,F. (2015) PyEMMA 2: A software package for estimation, validation, and analysis of markov models. *Journal of Chemical Theory and Computation,* **11,** 5525-5542.

48. Vallurupalli,P., Hansen,D.F., Stollar,E., Meirovitch,E. and Kay,L.E. (2007) Measurement of bond vector orientations in invisible excited states of proteins *Proceedings of the National Academy of Sciences of the United States of America,* **104,** 18473-18477.

49. Demers,J. and Mittermaier,A. (2009) Binding mechanism of an SH3 domain studied by NMR and ITC. *Journal of the American Chemical Society,* **131,** 4355-4367.

50. Meneses,E. and Mittermaier,A. (2014) Electrostatic interactions in the binding pathway of a transient protein complex studied by NMR and isothermal titration calorimetry. *The Journal of Biological Chemistry,* **289,** 27911-27923.

51. Torrie,G.M. and Valleau,J.P. (1977) Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics,* **23,** 187-199.

52. Laio,A. and Parrinello,M. (2002) Escaping free-energy minima. *Proceedings of the National Academy of Sciences of the United States of America,* **99,** 12562-12566.

53. Miao,Y., Feher,V.A. and McCammon,J.A. (2015) Gaussian accelerated molecular dynamics: Unconstrained enhanced sampling and free energy calculation. *Journal of Chemical Theory and Computation,* **11,** 3584-3595.

54. Sugita,Y. and Okamoto,Y. (1999) Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters,* **314,** 141-151.

55. Sgourakis,N.G., Yan,Y., McCallum,S., Wang,C. and Garcia,A.E. (2007) The alzheimer's peptides Aβ40 and 42 adopt distinct conformations in water: A combined MD / NMR study. *Journal of Molecular Biology,* **368,** 1448-1457.

56. Smith,A.K., Lockhart,C. and Klimov,D.K. (2016) Does replica exchange with solute tempering efficiently sample aβ peptide conformational ensembles? *Journal of Chemical Theory and Computation,* **12,** 5201-5214.

57. Knott,M. and Best,R.B. (2012) A preformed binding interface in the unbound ensemble of an intrinsically disordered protein: Evidence from molecular simulations. *PLoS Computational Biology,* **8,** e1002605.

58. Nguyen,T.H. and Minh,D.D.L. (2016) Intermediate thermodynamic states contribute equally to free energy convergence: A demonstration with replica exchange. *Journal of Chemical Theory and Computation,* **12,** 2154-2161.

59. Wedemeyer,W.J., Welker,E. and Scheraga,H.A. (2002) Proline Cis−Trans isomerization and protein folding. *Biochemistry,* **41,** 14637-14644.

60. Xiong,Y., Juminaga,D., Swapna,G.V., Wedemeyer,W.J., Scheraga,H.A. and Montelione,G.T. (2000) Solution NMR evidence for a cis tyr-ala peptide group in the structure

of Pro93Ala] bovine pancreatic ribonuclease A. *Protein Science: A Publication of the Protein Society,* **9,** 421-426.

61. Sarkar,S.K., Young,P.E., Sullivan,C.E. and Torchia,D.A. (1984) Detection of cis and trans X-pro peptide bonds in proteins by 13C NMR: Application to collagen. *Proceedings of the National Academy of Sciences of the United States of America,* **81,** 4800-4803.

62. Brandts,J.F., Brennan,M. and Lung-Nan Lin. (1977) Unfolding and refolding occur much faster for a proline-free proteins than for most proline-containing proteins. *Proceedings of the National Academy of Sciences of the United States of America,* **74,** 4178-4181.

63. Fassolari,M., Chemes,L.B., Gallo,M., Smal,C., Sánchez,I.,E. and de Prat-Gay,G. (2013) Minute time scale prolyl isomerization governs antibody recognition of an intrinsically disordered immunodominant epitope. *The Journal of Biological Chemistry,* **288,** 13110-13123.

64. De Vivo,M., Masetti,M., Bottegoni,G. and Cavalli,A. (2016) Role of molecular dynamics and related methods in drug discovery. *Journal of Medical Chemistry,* **59,** 4035-4061.

65. De Vivo,M., Masetti,M., Bottegoni,G. and Cavalli,A. (2016) Role of molecular dynamics and related methods in drug discovery. *Journal of Medical Chemistry,* **59,** 4035-4061.

66. Lim,W.A., Richards,F.M. and Fox,R.O. (1994) Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains. *Nature,* **372,** 375.

67. Feng,S., Chen,J.K., Yu,H., Simon,J.A. and Schreiber,S.L. (1994) Two binding orientations for peptides to the src SH3 domain: Development of a general model for SH3-ligand interactions. *Science,* **266,** 1241-1247.

68. Yu,H., Chen,J.K., Feng,S., Dalgarno,D.C., Brauer,A.W. and Schrelber,S.L. (1994) Structural basis for the binding of proline-rich peptides to SH3 domains. *Cell,* **76,** 933-945.

69. Saglam,A.S. and Chong,L.T. (2016) Highly efficient computation of the basal kon using direct simulation of Protein–Protein association with flexible molecular models. *Journal of Physical Chemistry B,* **120,** 117-122.

70. Wang,Y., Martins,J.M. and Lindorff-Larsen,K. (2017) Biomolecular conformational changes and ligand binding: From kinetics to thermodynamics. *Chemical Science,* **8,** 6466-6473.

71. Decherchi,S., Berteotti,A., Bottegoni,G., Rocchia,W. and Cavalli,A. (2015) The ligand binding mechanism to purine nucleoside phosphorylase elucidated via molecular dynamics and machine learning. *Nature Communications,* **6,** 6155.

72. Mark,P. and Nilsson,L. (2001) Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *Journal of Physical Chemistry A,* **105,** 9954-9960.

73. Onufriev,A., Bashford,D. and Case,D.A. (2004) Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins,* **55,** 383-394.

74. Wicky,B.I.M., Shammas,S.L. and Clarke,J. (2017) Affinity of IDPs to their targets is modulated by ion-specific changes in kinetics and residual structure. *Proceedings of the National Academy of Sciences of the United States of America,* **114,** 9882-9887.

75. Sun,B., Cook,E.C., Creamer,T.P. and Kekenes-Huskey,P.M. (2018) Dual roles of electrostatic-steering and conformational dynamics in the binding of calcineurin's intrinsically-disordered recognition domain to calmodulin. *bioRxiv,* .

76. Meneses,E. and Mittermaier,A. (2014) Electrostatic interactions in the binding pathway of a transient protein complex studied by NMR and isothermal titration calorimetry. *The Journal of Biological Chemistry,* **289,** 27911-27923.

77. Brown,A.M. and Zondlo,N.J. (2012) A propensity scale for type II polyproline helices (PPII): Aromatic amino acids in proline-rich sequences strongly disfavor PPII due to Proline–Aromatic interactions. *Biochemistry (N. Y.),* **51,** 5041-5051.

78. Williamson,M.P. (1994) The structure and function of proline-rich regions in proteins. *Biochemical Journal,* **297 (Pt 2),** 249-260.

79. Perthold,J.W. and Oostenbrink,C. (2017) Simulation of reversible Protein–Protein binding and calculation of binding free energies using perturbed distance restraints. *Journal of Chemical Theory and Computation.,* **13,** 5697-5708.

80. Chen,H. and Luo,R. (2007) Binding induced folding in p53−MDM2 complex. *Journal of American Chemical Society,* **129,** 2930-2937.

81. Chen,H. (2008) Mechanism of coupled folding and binding in the siRNA-PAZ complex. *Journal of Chemical Theory and Computation,* **4,** 1360-1368.

Supplemental Material

The 48 temperatures used during REMD are recorded below in Kelvin. The replica used for analysis is bolded.

290.00, 292.37, 294.76, 297.16, **299.59**, 302.03, 304.50, 306.99, 309.49, 312.02, 314.57, 317.14, 319.73, 322.34, 324.97, 327.62, 330.30, 332.99, 335.71, 338.45, 341.22, 344.00, 346.81, 349.65, 352.50, 355.38, 358.28, 361.21, 364.15, 367.13, 370.13, 373.15, 376.20, 379.27, 382.36, 385.49, 388.63, 391.81, 395.01, 398.23, 401.48, 404.76, 408.07, 411.40, 414.76, 418.14, 421.56, 425.00

Supplemental Table 2. Binding simulations with length, time in encounter 13, in encounter 20, and other

	Simulation	Length (ns)	Encounter 13 (ns)	Encounter 20 (ns)	Other distance (ns)
ArkA12	1	1600	1392	160	48
	2	1600	1520	60.8	19.2
	3	1600	704	816	80
	4	1600	592	960	48
	5	1600	100.8	1440	59.2
	6	800	1.92	776	22.08
	7	800	30.4	712	57.6
	8	800	30.4	760	9.6
	9	800	10.4	768	21.6
	10	800	120	608	72
	11	300	0	189	111
	12	300	0	189	111
	13	300	14.1	252	33.9
	14	300	12.9	270	17.1
	15	300	45	240	15
	16	300	4.2	285	10.8
	17	300	0.15	240	59.85
	18	300	25.8	171	103.2
	19	300	0	159	141
	20	300	6.9	270	23.1
Segment 1	1	4500	2655	1710	135
	2	4500	85.5	4365	49.5
	3	4500	1305	3105	90
	4	4500	2565	1845	90
	5	4500	2745	1665	90
	6	300	105	120	75

	7	300	0.66	243	56.34
	8	300	0.93	261	38.07
	9	300	11.1	267	21.9
	10	300	10.5	270	19.5
	11	300	177	108	15
	12	300	189	66	45
	13	300	195	78	27
	14	300	90	162	48
	15	300	3.3	261	35.7
	16	300	30	240	30
	17	3000	1260	1560	180
	18	3000	1050	1830	120
	19	3000	2520	420	60
	20	3000	297	2550	153
	21	3000	1860	960	180