PROTEIN CONFORMATIONAL TRANSITIONS USING COMPUTATIONAL METHODS

by

Heng Wu

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Medicinal Chemistry and Molecular Pharmacology

West Lafayette, Indiana

December 2018

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Carol B. Post, Chair

Department of Medicinal Chemistry and Molecular Pharmacology

Dr. Daisuke Kihara

Department of Biological Sciences/Computer Science

Dr. Markus A. Lill

Department of Medicinal Chemistry and Molecular Pharmacology

Dr. Casey J. Krusemark

Department of Medicinal Chemistry and Molecular Pharmacology

Approved by:

Dr. Andy Hudmon

Head of the Graduate Program

To whoever cares to read

ACKNOWLEDGMENTS

I would like to thank my advisor Carol Beth Post for her mentorship over the years. Her insightful comments and scientific assessment have helped to shape my projects into this thesis work. The problem-solving skills and all-around way of thinking I have developed under her mentorship would be an invaluable asset to my future career.

I would like to acknowledge my committee members, Prof. Daisuke Kihara, Markus Lill, and Casey Krusemark. They provide suggestions that help to expand my projects into more dimensions. My presentation skills have been greatly polished over the conversations with my committee members during the past years.

I would like to express my gratitude to the previous and current Post lab members for both scientific insights and companionship. Duy Hua, who guided me on technical skills when I first joined the lab and for scientific discussions over the time. Amit Roy, He Huang and Bradley Dickson, who gave their impartial opinions on my project, even though I have never had the opportunity to meet the latter two fomer lab members. Chao Feng and Mehul Joshi, who I consulted with for my preliminary exam and class projects.

My parents have showed their unconditional support for my pursuit of a graduate degree. Thank you for providing me the best education you can, and for giving me the freedom to follow my interest.

I do not include a full list of the names of the friends that I met at Purdue over the years. It is their companionship and support that give me the strength to go through the mental hardship that never becomes too strong. Special thanks go to a special friend who has been a friend for a while, and who has accompanied me the most.

The professional technical support at Purdue has made my journey smoother. Steve Wilson in Hockmeyer building has always responded timely to technical issues related to our local workstations. We have also built our first GPU workstation with his help. Lev Gorenstein, a former Post group member and now ITAP staff, among others has helped with solving issues on the community clusters.

And I would thank the local stores that provide me places to wander during the evenings, and my headset that has accompanied me through countless research hours.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABPO	adaptively biased path optimization
ADK	adenosine kinase
ATP	adenosine triphosphate
CML	chronic myelogenous leukemia
CSP	chemical shift perturbation
FES	free energy surface
GPCR	G protein coupled receptor
HTVS	high through-put virtual screening
IGF1R	insulin-like growth factor 1 receptor
NOE	nuclear overhauser effect
NtrC	nitrogen regulatory protein C
PKA	cAMP-dependent protein kinase
PMF	potential of mean force
PRE	paramagnetic relaxation enhancement
RMSD	root mean squared deviation
RV	reduced variable
RXR	retinoid X receptor
SFK	Src family kinase
SP	standard precision
XP	extra precision

LIST OF SYMBOLS

k_B	Boltzmann constant
ns	nanosecond
μs	microsecond

ABSTRACT

Author: Wu, Heng PhD
Institution: Purdue University
Degree Received: December 2018
Title: Protein Conformational Transitions Using Computational Methods
Major Professor: Carol B. Post

Protein conformational transitions are fundamental to the functions of many proteins, and computational methods are valuable for elucidating the transitions that are not readily accessible by experimental techniques. Here we developed accelerated sampling methods to calculate optimized all-atom protein conformational transition paths. Adaptively biased path optimization (ABPO) is a computational simulation method to optimize the conformational transition path between two states. We first examined the transition paths of three systems with relatively simple transitions. The ways to define reduced variables were explored and transition paths were built at convergence of the optimizations. We constructed the all-atom conformational transition path between the active and the inactive states of the Src kinase domain. The αC helix rotation was identified as the main free energy barrier in the all-atom system, and the intermediate conformations and key interactions along the transition path were analyzed. This is the first demonstration of the robustness of a computational method for calculating protein conformational transitions without restraints to a specified path. We also evaluated protein-peptide interactions using both molecular dynamics simulations and peptide docking. Long unbiased simulations were used to evaluate Src-SSP interactions and complex stability in both implicit and explicit solvent. Molecular docking was used to build possible protein-peptide interaction models, using both Src regulatory domain SH2 and the kinase domain. Possible Src-SSP complexes were built as the first Src-substrate complex structure models.

CHAPTER 1. INTRODUCTION

1.1 Biological Functions of Src Family Kinases

Protein kinases regulate intracellular and intercellular signal transduction pathways in a series of important functions, including cell migration, cell cycle and survival[1-2]. Dysregulation of the kinase activity often leads to aberrant signaling pathways, cell malignancy and diseases including cancer, diabetes and inflammation, making them good targets for drug design purposes[3-6].

Src family kinases (SFKs) are a group of non-receptor tyrosine kinases. Src protein is the product of the first proto-oncogene discovered[7]. SFKs consists of 11 members in human[8], namely Src, Lyn, Lck, Hck, Fyn, Yes, Blk, Fgr, Brk, Frk and Srm. While the expression of Src, Yes, Yrk and Fyn are ubiquitous, the expression of the other family members are usually tissue or cell specific[9]. SFKs are involved in various signal transduction pathways and cellular functions[9-12]. The members interact with a broad set of transmembrane receptors and signal to downstream DNA synthesis, MAPK activation, cytoskeletal rearrangements and cell migration[13]. A few examples of the receptors that interact with SFKs are GPCRs[14], integrins[15], steroid receptors[16], immunoreceptors[17] as well as receptor tyrosine kinases such as growth factor receptors[18]. The activity of Src is regulated via phosphorylation of tyrosine residues. The first regulation site is Tyr416, whose phosphorylation is associated with the upregulation of the kinase activity[19-21]. Another site is Tyr527, the phosphorylation of which will downregulate the kinase activity[22-23].

SFKs are involved in multiple types of cancer and autoimmune diseases as a result of its important roles described above. The overexpression and/or mutation of the Src family tyrosine kinases (SFKs) has been observed in various types of cancers, including carcinomas of the lung, ovary, gastrointestinal tract, pancreas[24], and breast[4][25-27]. Besides, SFKs play roles in autoimmune diseases due to their involvement in the immune cell signaling[28-29]. While targeting Src for autoimmune diseases might be difficult due to the complex and sometimes conflicting signaling pathways and tissue-specific expressions[29], the development of Src inhibitors for several types of cancers has seen

encouraging progress. Two selective kinase inhibitors that target BCR-ABL are imatinib (Gleevec)[30][31] for chronic myelogenous leukemia (CML), and dasatinib for imitanib resistant CML. Those two inhibitors were approved by FDA. Allosteric inhibitors targeting BCR-ABL such as asciminib[32-33] that bind to the myristoyl-binding site has also reached phase III trial[34]. Several Src inhibitors have reached phase II or phase III testing, such as saracatinib that inhibits SRC but failed in several phase II trials [35-38], bosutinib that inhibits SRC/ABL that is in ongoing, promising phase III trial[39-41]. Given that the published inhibitors in early stages from Pharma companies are not usually the most promising ones, it is likely that there are more SFKs inhibitors under development. A list of SFKs inhibitors, the target and binding site, the development stages, and PDB code if available, including these inhibitors described here, is in Table 1.1.

Table 1.1 A list of BCR-ABL and SFKs inhibitors and the development stage.

Inhibitor	target	Binding site	PDB code	stage
imatinib	BCR-ABL	catalytic		FDA approved
dasatinib	SFKs, BCR-ABL	catalytic	3GSD, 3QLG	FDA approved
saracatinib	SRC	catalytic	2H8H	Phase II
bosutinib	SRC-ABL	catalytic		Phase III
asciminib	BCR-ABL	allosteric		Phase III

1.2 Structures of Src Kinase and the Conformational Activation

Non-receptor tyrosine kinase structures consist of regulatory domains and kinase domain. SFKs are about 60kD in molecular weight and have conserved domain organizations. The domain structures consist of a short SH4 domain, a unique region, SH3 domain[42], SH2 domain[43], a catalytic domain (kinase) and a short C-terminal tail[44]. Each large region has about 40-70 (unique region), 50(SH3), 100(SH2) and 250 (kinase) residues respectively. A schematic representation is in Figure 1.1A. The regions and the residue numbering in c-Src are listed in

Table 1.2. The two tyrosine regulatory sites Y416 and Y527 mentioned before are in the kinase domain and C-terminal tail respectively. SFKs associate with the cytoplasmic side of the plasma membrane via myristylation and palmitylation of the residues in N-

terminal SH4 domain[45-46]. Gly 2 is a myristoylation site in SH4, and residue 3 is the palmitoylation site in Lck, Hck but does not exist in Src or Blk[44]. SH3 domain and SH2 domain are regulatory units that recognize proline-rich regions[47] and phosphorylated tyrosine[48] respectively.

The crystal structures of the down-regulated and active conformations of Src are available for both full-length protein (excluding SH4) and the kinase domain. The full-length Src structure is shown in Figure 1.1B. In the down-regulated form, the kinase is in an assembled form with the phosphorylated C-terminal tail forming an intra-molecular interaction with the SH2 phosphotyrosine binding pocket. The SH3 domain interacts with the proline-rich linker between SH2 and the kinase domain. The linker adopts a left-handed polyproline II helix conformation in the down-regulated form. Both SH3 and SH2 domains are at the back side of the kinase domain. In the active conformation where Y527 is unphosphorylated, the SH2 domain is released from the C-terminal tail, and the SH3 and SH2 domains have no interactions with the kinase domain. The relative orientation of the three domains also changes during the activation process, and the SH3-SH2 linker partially unfolds and extends to accommodate the domain movements.

There are several functionally important conformational changes happening within the kinase domain during the activation indicated from the two forms of crystal structures. The downregulated and active kinase domain structures are detailed in Figure 1.1C. The two structures differ in three aspects, the αC-helix orientation, the A-loop conformation, and the relative orientation of N-lobe and C-lobe, with the former two being the main differences. In the down-regulated form, the A-loop residues form a two-helical structure and reside between the N-lobe and the C-lobe, preventing binding in the catalytic site. In the active form, the A-loop unfolds and extends to a free loop structure, exposing the catalytic site for substrate phosphorylation. The αC-helix orientation also alters between the two states. In the down-regulated form, the αC-helix is out to be away from the cleft, exposing the catalytically important Glu310 in solution. In this conformation, E310 interacting with R409 on the A-loop. In the active conformation, the αC-helix rotates ~25º to bring the E310 to the cleft to interact with K295 and form a catalytic important salt bridge[49]. The details of the residue-residue interactions difference of the two forms can

be described in terms of a switched electrostatic network[50][51]. Finally, the lobe-lobe distance increases in the active state, resulting in a more open catalytic cleft.

Table 1.2 The residue numbering of the domains in c-Src kinase.

Region	SH3	SH2	linker	N lobe	C lobe	C-terminus
Residue NO.	84-153	154-245	246-259	260-341	342-521	522-533

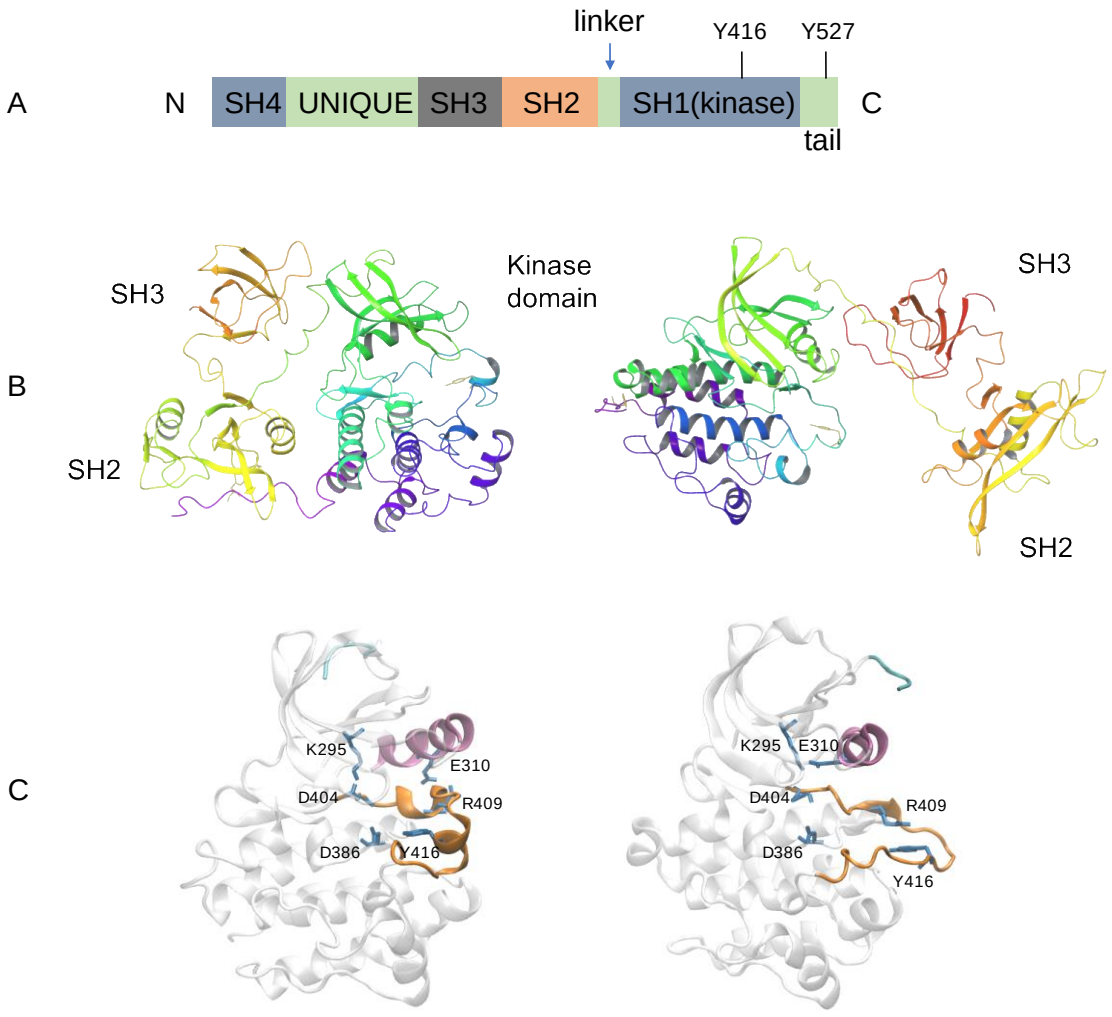


Figure 1.1 The SRC protein schematic and crystal structures. A: a schematic representation of SFKs structure, showing the domains and the C-terminal tail. B: the down-regulated and active full-length Src kinase. Left: the down-regulated form, PDB ID 2SRC. Right: the active form, PDB ID 1Y57. C: the Src kinase domain with the down-regulated form on the left and the active form on the right to show the structural difference between the two forms. Pink: the αC helix. Orange: the A-loop. Cyan: the N-terminal linker. The residues in the electrostatic network are shown in stick representation.

1.3 Computational Methods for Conformational Transitions, and Computational Studies on Src

Conformational transitions are fundamental to functions of many proteins[52-54], such as signaling proteins that convert between an enzymatically active and down-regulated forms, membrane proteins that transport molecules via open/closed forms, and molecular machines that couple chemical energy to molecular motion. Many conformational transitions are between states with disparate functionality and often tightly regulated for proper control of cellular processes, which highlights the importance of studying the transition processes. Computational methods are valuable for elucidating such transitions in atomistic detail not achievable by experimental observation. Mechanistic insights, and an understanding of molecular recognition or regulation of enzymatic activity can be gained from knowledge of a free-energy surface or free-energy profile along a pathway.

Most functionally interesting protein conformational transitions are activated processes, and the timescales are typically longer than can be adequately sampled with current unbiased molecular dynamics (MD) simulations. Enhanced sampling methods are therefore required to overcome the free-energy barriers that separate different protein conformational states[55-57].

Several of the most notable methods that utilize features of the transition to efficiently explore relevant regions of the conformational space are reviewed here. Metadynamics[58-60] explores multidimensional free energy surface (FES) by defining a few collective coordinates. A sampling history-dependent potential energy term, or, adaptive biasing potential, fills minima on the FES to allow the efficient exploration of the FES. Adaptive biasing force[61-62] evaluates the derivatives of the free energy with respect to both Cartesian coordinates and time for free energy calculations based on thermodynamic integration. It works that when the force acting on the coordinate of interest (defined as instantaneous force) is subtracted from the equation of motion, the acceleration along the coordinate becomes zero. The potential of mean force is calculated once the simulation is completed by integrating the derivatives. It is simple to use for complicated

order parameters and straightforward to implement under the available simulation framework. Milestoning[63] defines a sequence along the reaction coordinate as milestones and launches short trajectories at each of the milestone hypersurface. The trajectories are terminated when for the first time they reach one of the neighboring milestones. Then the probability densities of the short trajectories along the reaction coordinates are calculated. Accelerated MD[64] adds a continuous non-negative bias boost potential to the true potential when the true potential is below a threshold value. This results in an enhanced escape rate for the modified potential and accelerate sampling on the free energy surface rather than stay at a local minimum.

Path-directed approaches seek to specify the transition pathway between two known states, A and B, using a series of images to define the pathway through a space of a reduced set of variables. Evolution to the optimal pathway in most cases involves restrained sampling near the images. The finite temperature string method[65], implemented with restraints in the hyperplanes, or with swarms of trajectories[66], or with umbrella sampling underlies many of these techniques to find the minimum free-energy path[67] or maximum flux transition path[68].

Src catalytic domain activation features αC helix displacement and A-loop extension. The atomic details are not readily accessible by experimental techniques. Also, the transition is not easily accessed by conventional computational methods due to the high free energy barrier between the two states. Several studies indicate that calculating the Src activation transition using long simulations very likely exceeds the current simulation computation power[69-70]. Unbiased long simulations suggested an allosteric network underlying kinase regulation, involving the key structural elements in Src kinase domain, including αC helix, the regulatory spine (R spine), the catalytic spine (C spine), the HRD motif and several loops[70].

Accelerated sampling methods have been applied to study the conformational transition. A simulation using string method with swarms of trajectories identified the activation as a two-step process with A-loop opening followed by αC helix rotation[71]. The DFG-flip conformational transition was characterized using the same method combined with umbrella sampling[72]. A Markov state model was built to understand the kinetics of the transition[73], while the main features of the MSM transition path was studied by a

transition path theory framework[74]. Umbrella sampling was used to explore the free energy landscape of the active form, and the structural features like the regulatory spine(R-spine), HRD motif, and the electrostatic switch were analyzed[75]. A connectivity map was built from simulations to show the Src family member (Hck and Lck) conformational activation and identify intermediate states[76]. While the above summarized simulations mostly focus on the kinase domain, a simulation using mutations emphasized the importance of the linker connecting the SH2 and kinase domain[77]. Besides, full length Src activation was studied by a combination of string methods and umbrella sampling[78]. While these simulations provide valuable insights into the activation process, an initial path is often required for the calculation. Also, statistical methods were utilized to build the path from discontinuous simulations in several cases, while a continuous activation process is not observed.

Several questions remain unclear in the Src conformational transition process. The first is the order of events that happen along the transition path. It is not clear that during the inactivation process, if the A-loop folding happens before or after the αC helix rotation. It is explicitly stated in several studies that the αC helix rotation happens before the A-loop folding[71][74][76] during the inactivation. For other studies, this order of events is not explicitly discussed. The SFK Lyn kinase transition path built by MFTP suggested that the αC helix rotation happens after the A-loop folding and locks the two-helical A-loop conformation[79]. Further examinations would be needed to solve this inconsistency. Second is the structural information along the transition path. Previous results either have the structures from trajectories launched from a pre-determined path[71], or have the path built from short trajectories using statistical methods[74], or a combination of both[80], while a continuous and unrestrained path is not observed. Third is that the function of the Src regulatory domains and how they affect the kinase domain transition are not clear. Most of the reviewed works use the Src kinase domain starting from residue 260[71-72] [74-76][80]. The roles of the linker in Src activation regulation will be inspected in our work.

1.4 A brief introduction on adaptively biased path optimization (ABPO)

The ABPO methodology was developed and described in detail in reference[81] and is summarized here. ABPO uses an adaptive biasing potential in an iterative scheme to evolve an initial path to the optimal principal path connecting two pre-determined stable states by following the formula of the string method[65][82]. Unlike the string method where an initial path with structures along the path is required and the sampling is restrained to the initial path, ABPO does not start from an initial path and no restraints are required. The path is optimized through a reduced variable (RV) space and parameterized by $\phi(\lambda)$ with λ varying from 0 to the total length of the curve. λ indicates the position along the path, and when normalized, λ ranges from 0 to 1 from one state to the other. Multiple trajectories are launched from each end state and visits along a path defined with initial RV values are accumulated with a histogram. The histogram records the number of samplings on the positions along the path and is a function of both λ and the simulation time t, and is denoted as $h(\lambda, t)$.

Path optimization proceeds in a set of cycles to update the position of the path and re-parameterize $\phi(\lambda)$. In each cycle of the computation, the path is evolved according to the mean RV position of the trajectories associated with the hyperplane perpendicular to the path at each λ[67]. At the end of each cycle, the optimized path is updated to the mean position of the sampling on each hyperplane for each λ along the path.

An adaptive biasing potential[83] is constructed on the path to accelerate sampling along the path through the reduced-dimensional space. There are no restraints to localize trajectories to the path. The sampling is further facilitated by computing multiple independent trajectories in parallel (replicas) to determine the mean position for a cycle. The replicas have identical initial coordinates and simulation parameters, and they only differ in the initial velocity for the atoms. The replicas are used to enhance the sampling along the path.

ABPO accelerates the sampling in the region surrounding the path by adding the bias potential V_b[83] at point λ on the path. The bias potential, up to an arbitrary constant, is

$$V_b(\lambda, t) = k_B T \frac{b}{1-b} \ln[c(1-b)h(\lambda, t) + 1] \quad (1.1)$$

The histogram $h(\lambda, t)$ counts visits to the region around λ over time *t*, *b* is the fraction of the free energy that is flattened by the bias, *c* controls how the bias couples to the dynamics and has inverse time units, k_B is the Boltzmann constant, and *T* is temperature. The bias potential at λ increases with visits so the region is 'flooded' on the potential energy surface. The gradient of V_b is an ensemble average and well converged. For details on the gradient of V_b and a discussion of its convergence in the integration to propagate trajectories, the reader is referred to references[81][83] , where it was also shown that in the limit $b \rightarrow 1$, equation 1 reduces to the potential for standard metadynamics[83].

The only positional restriction in ABPO is a one-sided harmonic tube-wall potential to limit the sampling within a tube-shaped space around the path. The tube potential constructs a conformational space that has a distance *R* from the current path. The sampling is only allowed within the distance *R* from at least one point on the path[81] to discourage sampling far away from the current path that would not help with updating the current path. This tube potential is proved to be necessary and would improve the path optimization efficiency. A schematic representation of the construction of the tube is shown in Figure 1.2. The two ends of the tubes are modeled as hemispheres to avoid sudden change in tube potential.

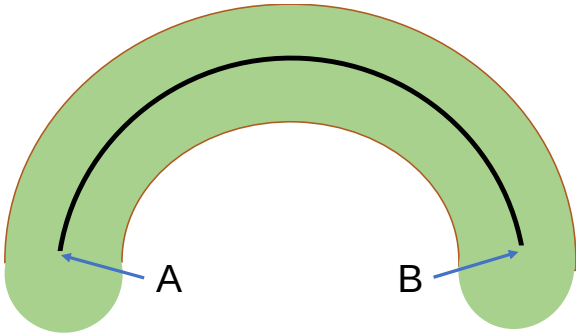


Figure 1.2 A schematic representation of the tube defined around the path. Point A and B indicate two state on the free energy surface, and the black line shows a path connecting the two states. The green area shows the tube space around the path.

Within each cycle, the trajectories proceed in blocks. At the end of each block, the histograms from all replicas are pooled together to check if the combined sampling at each slice has reached a pre-set minimum threshold. The cycle is terminated when the threshold

is reached, and the path is updated to the mean RV position of the replica ensemble for each hyperplane. Another cycle is started with reinitialized histograms, and cycles continued until the path is converged based on the distance between the current path and previous paths. At convergence, the path is the principal curve through RV space connecting the two end states. A flow chart of ABPO simulations in execution is in Figure 1.3.

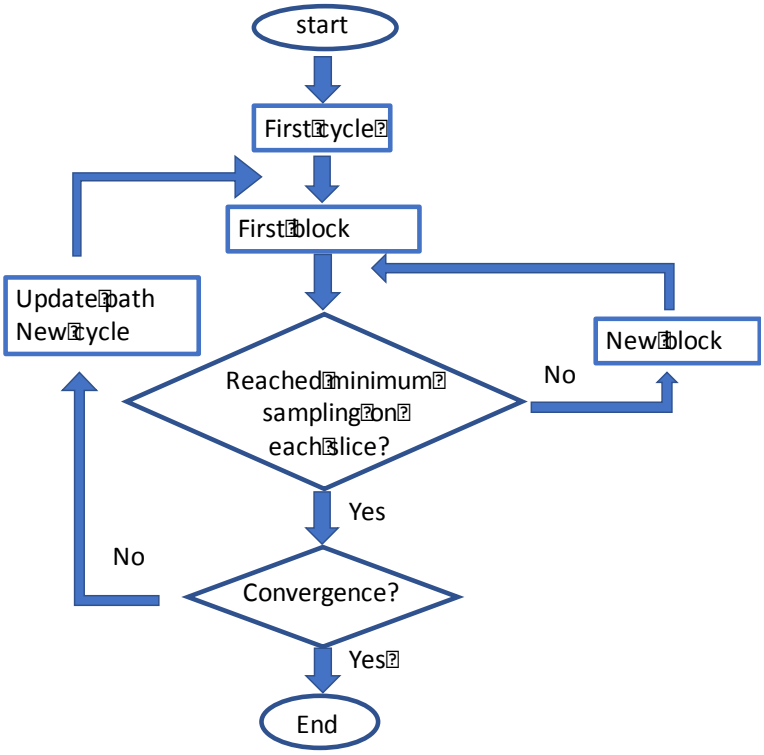


Figure 1.3 A flow chart for ABPO simulations. The path is updated through cycles, and within each cycle, the simulation progresses with blocks. The convergence of the simulation is checked to determine when to end.

The free energy for each slice, up to an additive constant, is obtained from the converged path and the histograms that contain only information for this final path. The

PMF $A(\lambda, t)$ [83] at the principal curve is computed from the combined histogram according to the following equation:

$$A(\lambda, t) = -k_B T \frac{1}{1-b} \ln[h(\lambda, t)] \quad (1.2)$$

With this equation, the PMF is calculated from the sampling history $h(\lambda, t)$ along the path. This approach to estimate free energy from the dynamics accumulated with the bias potential (equation 1.2) was introduced previously[83][84], where was also shown that *b* specifies the percentage of the free energy canceled by the bias potential ($A(\lambda,t) = -b^{-1}V_b(\lambda,t)$). Equation 2 limits b to values less than 1.0; when b$\rightarrow$1, A(λ,t) diverges. Similar divergence behavior exists with metadynamics computation[85-86]. ABPO implementations have used b values of 0.8 to 0.9.

Compared with the aforementioned computational methods that accelerate sampling, ABPO has several advantages. First, an initial path that might be unphysical between the two states is not required. The transition path can be built from trajectories initiated from the two end states. Second, the sampling is unrestrained on the free energy surface, except the tube wall potential to restrict the sampling within a certain distance. This tube wall potential has been proven to be necessary; it reduces sampling on the plane perpendicular to the path, and encourages sampling in the direction of diffusion along the path. The lack of such potential might lead to computational time wasted on sampling far from the path and failure in identifying the optimal path. Third, the sampling is enhanced by launching multiple replicas in addition to the adaptive biasing potential, which further improves the optimization speed.

1.5 Outline of the Thesis

The aims of this thesis are to 1) develop a methodology to build protein conformational transition pathways, 2) elucidate protein-ligand interactions using simulations and molecular docking. The ultimate goal is to facilitate structure-based drug design efforts targeting protein kinases. To achieve the aims, in Chapter 2, we report applying ABPO to three all-atom systems with relatively small transitions in response to ligand binding. Two systems have a local loop flipping, and the third system experiences a helix displacement. As introduced in Chapter 1.4, ABPO works in reduced variable space. Therefore, we

explore the methods to define proper reduced variables for each type of transitions. We analyze the transition paths for the three systems and recover the potential of mean force (PMF) profiles along the paths from the sampling. In Chapter 3, we describe the Src all-atom conformational transition path from ABPO calculations. We analyze the reduced variables and the structures along the transition path. Also, we identify the main free energy barrier along the path and determine the functions of the key residue. In Chapter 4, we look at how Src kinase domain interacts with its substrate. We use long, unbiased molecular dynamic simulations to study the protein-peptide complexes behavior in both explicit and implicit solvent. The protein and ligand backbone dynamics behavior are reported, and the protein-ligand interaction energy profiles are analyzed. In Chapter 5, we use molecular docking to model Src regulatory SH2 domain-flexible molecule complexes, and Src kinase domain-substrate peptide complexes. Glide is used for docking, and several types of techniques are reported to improve the performance of modeling the complex. The main issues in flexible molecule docking are also discussed. Finally, in Chapter 6, we summarize our work and provide future directions for this thesis.

CHAPTER 2. SMALL-SCALE PROTEIN CONFORMATIONAL TRANSITIONS FROM ALL-ATOM ABPO

2.1 Introduction

Adaptively biased path optimization (ABPO) is an approach[81] to optimize conformational transition pathways by constructing the adaptive biasing potential introduced in reference[83] in terms of a one-dimensional path in a reduced-variable space. The path is evolved with trajectories determined from the gradient of the adaptively biased potential and without restraints that localize the trajectories to the path. The path is optimized according to the description of the finite temperature string[65][82], and proceeds iteratively by updating the path variables according to the mean position of trajectories in cross sections, or hyperplanes orthogonal to the path at the images. The ABPO methodology differs from other path methods by allowing unrestrained sampling along the path rather than performing a linearly restrained path search. An adaptively biased potential is constructed on-the-fly to enhance the sampling along the path, without employing restraints, which are thought to potentially impede convergence or be difficult to sample[62][87]. Further, the approach does not require the generation of initial structures at specified positions along the path to initiate the computation as is needed by approaches that utilize restrained images in discretizing the path. Such initial structures could be unphysical and lead to instability or poor convergence when trying to move the system to the optimal pathway.

A first step of transition path optimization is to define the reduced variables that adequately capture the structural changes necessary and sufficient for the transition. We use the term "reduced variable" to reflect the low dimensionality of the space in which the pathway is determined rather than the previously used term "collective variable"[81] [88-89] to avoid inference of a collective motion being involved in optimizing the transition. The simplification afforded by using a reduced dimensionality for defining a path in a

complex conformational space has been appreciated for some time[90]. Four types of geometric reduced variables have been typically used: internal distances[58] or a linear combination[79] thereof, angles[91], and torsion angles[58][92]. The choice of reduced variables is a critical step for achieving a converged pathway but remains an ill-defined step in practice.

ABPO was introduced[81] using a Gō-model[93-94] and applied[81] to define the pathway for conformational activation of Lyn kinase, a Src-family protein tyrosine kinase. The Gō potential models a protein at the residue level with a single Cα position representing each residue, compared with an all-atom model, which includes greater than ten times more particles for the protein molecule. An important result of this earlier study was that the transition between down-regulated and activated conformations of Lyn obtained by ABPO and the maximum flux transition path (MFTP) method[79] were mechanistically similar. This direct comparison of the two computed pathways determined independently, using different computational approaches, gives confidence in the pathway. In addition, convergence of the pathway was achieved using ABPO with a 4.5 times smaller computational cost, demonstrating the efficiency of the ABPO method. Examination of the pathway found that the conformational changes contributing to the highest free-energy barrier were associated with the rotation of helix C, and thus provided a physical rationale for a large number of structurally diverse, kinase regulatory complexes for which the mechanism of the regulation was not always apparent from the crystal structure alone[79]. Here, the application of ABPO is extended to an all-atom description of the protein systems. How well ABPO can sample with the increased resolution and ruggedness of an all-atom force field has not yet been reported. Exploration of a higher resolution energy surface requires an appropriate choice of reduced variables that define the transition pathway keeping in mind that an atomistic model offers many additional descriptors of the transition beyond those specified in terms of only Cα positions. We examine conformational transitions with biological relevance in three protein systems (Figure 2.1): triose phosphate isomerase (TIM) has a flexible loop that closes when the protein is bound to a ligand[95]; dihydrofolate reductase (DHFR) has a loop that can adopt open and occluded forms in two states[96];and estrogen receptor has a helix that adopts two distinct positions when the protein is bound to an agonist or an antagonist[59][97]. For each system, distance-based or

torsion-angle-based reduced variables were identified from equilibrium trajectories. We discuss the choice of reduced variables, which is a critical step for path sampling methods. ABPO was launched to obtain the transition path between the two states. Our results suggest that ABPO works efficiently to converge an optimized transition pathway using an all-atom description of the protein systems. The all-atom transition pathways for the three systems identified from the simulations are described.

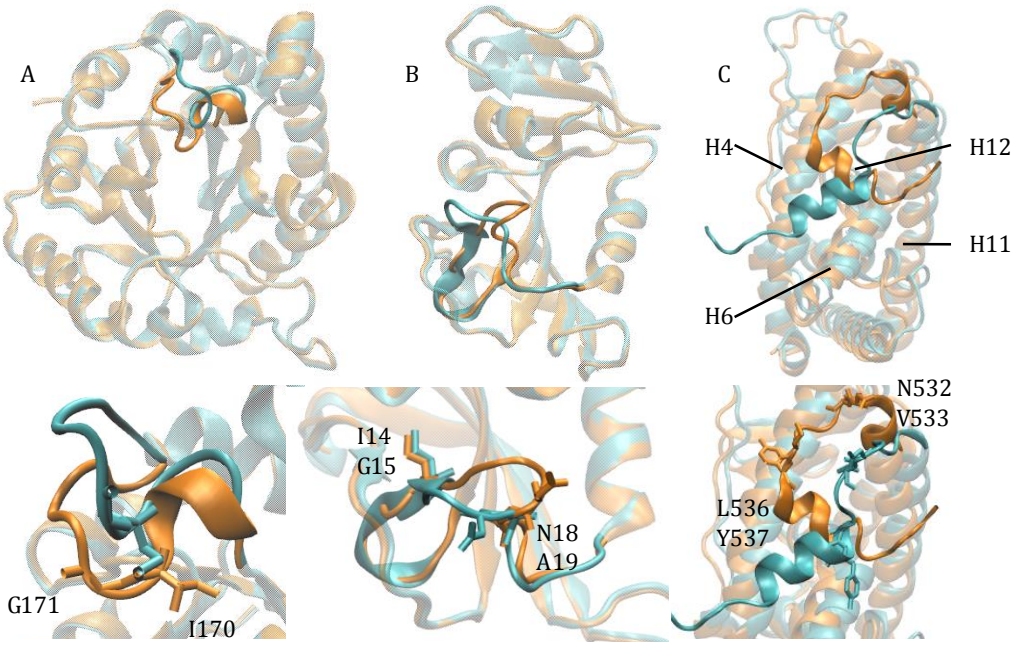


Figure 2.1 The conformational transitions of the three systems, illustrated in ribbon representation, are shown with opaque ribbon for the region of the transition. Zoomed-in views of the transition regions are at the bottom. The residues in stick representation in the zoomed-in views have disparate ϕ-ψ distributions in the two end states, as detailed in the results for each system. A: triose phosphate isomerase (TIM) transition of residues 166-176, including loop6. Cyan: open form, PDB ID 8TIM; orange: closed form, PDB ID 1TPH. B: dihydrofolate reductase (DHFR) transition of residues 9-24, including Met20 loop. Cyan: open form, PDB ID 1RA2; orange: occluded form, PDB ID 1RX7. C: estrogen receptor α ligand binding domain (ERα LBD) transition of residues 528-550, including helix12. Helix 4, 6, and 11 that interact with H12 are labeled. Cyan: antagonist-bound form, PDB ID 3ERT; orange: agonist-bound form, PDB ID 1QKU.

2.2 Methods

2.2.1 Simulation systems

TIM is an enzyme in glycolysis that catalyzes the reversible conversion of dihydroxyacetone phosphate to glyceraldehyde 3-phosphate. Loop 6, a flexible loop that contacts the active site (Figure 2.1A), is in a closed state when TIM is bound with a ligand, and in an open state on average in the absence of ligand. The transition between the two states is a key feature of the catalytic function of the enzyme by allowing substrate access to the active site in the open state while excluding water in the closed state[98-100]. Here we modeled the transition pathway between the two states using ABPO.

DHFR is an enzyme that reduces dihydrofolate to tetrahydrofolate. The Met20 loop adjacent to the active site is highly flexible (Figure 2.1B)[101], and immobilizes NADPH to promote the transfer of hydride from NADPH to dihydrofolate. Three Met20 loop conformations have been observed in various crystal structures, distinguished depending on if the active site is open, closed or occluded by the loop[96]. Met20 loop conformational flexibility is closely linked to the function of the enzyme given its alternating positions that either occlude or stabilize NADPH in the active site[101]. Here we study the transition between the open and occluded states, which is the largest conformational transition among the three states. The DHFR loop motion is more complex by involving more than a single peptide-bond motion.

ERα is a member of the nuclear receptor (NR) superfamily. Dysregulation of NR signaling often results in diseases such as cancer, diabetes, infertility and obesity. Specifically, ERα overexpression is often identified in breast cancer, and various studies have established ERα as one of the therapeutic targets in breast cancer[102-103].

The NR superfamily structure comprises three domains, and the ligand-binding domain (LBD) is the focus of our study. The NR LBD structure is highly conserved with eleven α-helices packing into a three-layer sandwich motif (Figure 2.1C). In ERα LBD, only H12 on the C-terminus is highly dynamic.

H12 is an essential element in ERα function by serving as a gate to regulate the binding of coactivators[104]. When LBD is bound to an agonist, the H12 gate is positioned to form the coregulatory surface that binds coactivators to activate downstream signaling for gene transcription. When LBD is bound to an antagonist, H12 adopts a new conformation, and

is positioned in the coactivator binding site, prohibiting the activation of the receptor[104]. NMR studies[105-107] show that the unligated forms of NR LBDs are conformationally dynamic, and the motions of LDB occur on a ms timescale.

2.2.2 Simulation details

Three systems, each in two states, are from the PDB entries 8TIM, 1TPH, 1RA2, 1RX7, 1QKU and 3ERT. Missing atoms, including all hydrogens atoms, were added to the set of coordinates retrieved from the PDB using the IC BUILD facility of CHARMM[108]. All crystal water molecules and ligands were removed. Simulations of the proteins were carried out using the CHARMM22 all-atom force field with CMAP dihedral angle corrections[109-110]. Unless stated otherwise, the solvent was modeled by the implicit solvent model FACTS[111] . It has been established that the structure and dynamics of single-domain, globular proteins are accurately reproduced with FACTS by comparison with explicit water TIP3P[112]. As the preparation steps for ABPO, we first performed energy minimization on the two structures for the end states of the transition. The energy was minimized using the steepest descent and Powell algorithms to a gradient less than 1.0 in the following stages: 1) with the position of protein heavy atoms fixed, 2) with harmonic restraints on protein heavy atoms, 3) with harmonic restraints on protein backbone (N, C, C_α) atoms, and 4) without restraints.

The energy-minimized structures were heated from 100 K to 300 K and equilibrated at 300 K over a total period of 500 ps. The initial velocities were generated from Gaussian distributions at the specified temperature. The leapfrog integrator was used to calculate the trajectories with a 2 fs time step.

A 10-ns simulation was initiated using coordinates from the equilibration run and Langevin dynamics with a temperature of 300 K, with long-range interactions cutoff distances set to 10, 12 and 14 Å. Coordinates were saved every 2 picosecond. The time series of temperature, potential energy, and heavy atom RMSD with respect to the energy-minimized structure were monitored to assess the simulations were stable.

A “closest-to-average structure” is a frame taken from the trajectory in place of a structure generated from the statistically averaged coordinates, which are often unphysical even after energy minimization. The closest-to-average structure was generated from the

last 4ns of the trajectories and used to define the distance RVs, to set values for the RVs at the end states and initiate the ABPO simulations. The coordinates averaged over the last 4 ns of the unbiased MD were compared to coordinates of each frame. The frame with the minimum heavy-atom RMSD with respect to the average structure was extracted from the trajectory as the closest-to-average structure.

2.2.3 ABPO parameters

The transition pathways were computed using the ABPO module in CHARMM. ABPO is an implementation of the path optimization and calculation of path free energy based on the bias potential in equation 1 and its gradient[81]. The number of replicas, the tube radius (R), the number of blocks per cycle, time steps per block, number of cycles and total simulation time for each system are summarized in Table 2.1. The effect of R on sampling efficiency and the free energy along the pathway are discussed in 2.2.3.

In all simulations, a time step of 2 fs was used. Langevin dynamics was used with a temperature of 300K. From equation 1, the fraction of the free energy cancelled by the bias potential, b, was 0.8, and the coupling of the bias to the dynamics, c, was 2.5 t^{-1}. The histogram for visits to path slices are smoothed using a Gaussian mollification factor set to 0.05. The number of slices are indicated in the plots for each path. The parameter values for the radius and force constant in the tube-wall potential[81] were chosen to enable efficient sampling; transition paths with more complex RVs require a larger radius. For the paths specified by dihedral RVs, the tube radius was 0.2 and 0.4 rad, and the force constant was 15 and 5 kcal/mol for TIM and DHFR transitions, respectively. For the distance-based RVs of ERα LBD, the tube radius was 10 Å and force constant 5 kcal/mol/Å^2. Ref[81] provides guidance for setting ABPO parameters.

Here, the initial paths were discretized to a set of linearly interpolated points between the two end-state values of the RVs. The end-state values were set equal to the population average from the distributions obtained in an unbiased simulation of the two known forms of the protein. The number in the set, or number of slices, varied depending on the complexity of the transition path. Initial coordinates to launch ABPO for path optimization are needed only for the end states; no coordinate sets are required at intermediate points of the path. The closest-to-average structure from the unbiased simulations were used for

end-state initial coordinates to start multiple trajectories running in parallel to accumulate sampling information to adapt the bias potential (Equation 1.1).

The string method[67][82] followed here to describe the evolution of the path includes a metric tensor, **D**, with the dimension of a diffusion coefficient. **D** was evaluated with equation 2 in reference[81], with averages estimated from short unbiased simulations at each end state and the inverse of **D** stored for input to ABPO optimization and free energy evaluation. As such, **D** is assumed to be approximately constant, which was verified by comparison of the elements computed at each end state and finding that the elements are acceptably close in value. For TIM and DHFR, **D** was evaluated from unbiased simulations over 100 ps with a 2 fs timestep. For ERα LBD, **D** was evaluated from unbiased simulations over 2 ns with a 2 fs timestep.

Distance combination RVs

A combination of individual inter-atom distances[79] was calculated using equation (2.1). For a distance combination RV with *n* individual inter-atom distances, *Z* is the combined value, r_j^{state} is the inter-atom distance of residue pair *j* in the state.

$$Z = \sum_{j=1}^{n} \frac{r_j^{state1} - r_j^{state2}}{|r_j^{state1} - r_j^{state2}|} r_j \qquad (2.1)$$

For all simulations, a time step of 2fs was used. Langevin dynamics was used with a temperature of 300K. The simulation details for each system are summarized in Table 2.1.

Tube potential effect on sampling efficiency and free energy of the path

The tube potential is a harmonic restraint to restrict the sampling of trajectories within a distance, *R*, of the closest point on the path. The choice of *R* should be large enough to allow a degree of exploration that finds alternative channels in the energy surface. *R* affects not only the sampling efficiency for optimization, but can also impact the entropic contribution computed from an ABPO final optimized path if the tube is more narrow than the actual reaction channel for the transition[81].

For the three conformational transitions of this study, the choice of R was made based on the efficiency of the pathway moving away from the initial path and toward a well converged path. Too small a value for R can limit the possibility of exploring multiple channels on a rugged surface and the path does not evolve, while too large a value can slow the progress of converging the path position. R was varied to define a value such that

trajectories traveled along the path and the path moved in the reduced-variable space in an acceptable amount of simulation time. An example for varying R in the case of the DHFR loop transition, with main-chain dihedral-angle reduced variables, is shown in Figure 2.2. For the *R* values 0.8 and 0.6 the time progress of A_RMSD is deemed too slow, while the convergence rate with 0.4 is acceptable.

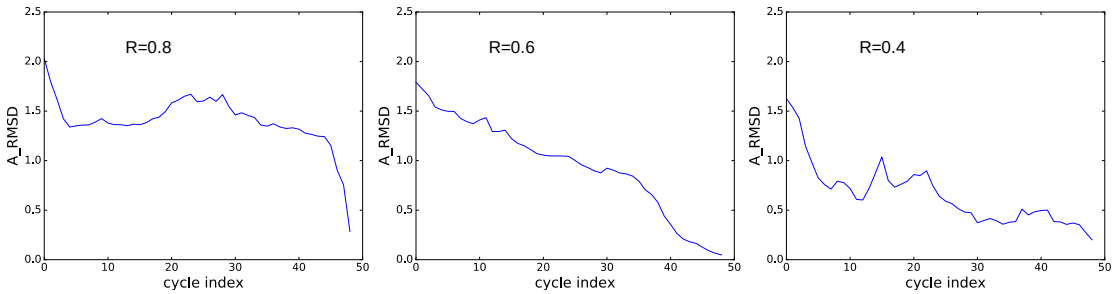


Figure 2.2 The tube radius, *R,* was varied to determine its effect on the rate of convergence of the ABPO calculation. When *R* is large, the A_RMSD against the last cycle is consistently large so that more cycles are required for the calculation to converge as a result of larger RV space surrounding the path.

Whether the tube is too narrow to accurately capture the entropic contribution to the free energy of the transition path can be explored by varying the value of *R* when accumulating histogram information from the final path (Equation 1.2). The possibility of *R* limiting an entropic contribution to the FE was tested for the transition of ERα LBD and the results are shown in Figure 2.3.

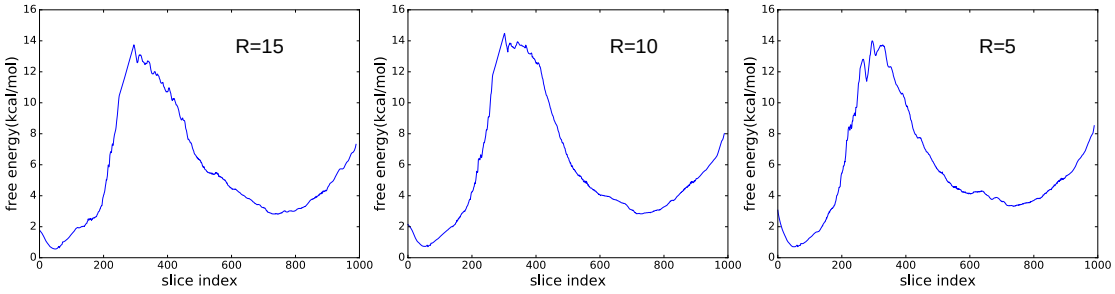


Figure 2.3 *R* was altered in the ABPO cycle to compute the PMF from the optimized path of ERα LDB for the purpose of detecting affects due to a possible entropic contribution; higher entropy manifests as a broad reaction channel. A reaction channel broader than the tube width would lead to an inaccurate free energy profile, and a larger *R* would be needed to accurately capture the entropic effect. We observe that varying *R* from 5 to 15 Å has no substantial effect on the PMF. The three PMF profiles have the same shape and a single peak at a similar slice index along the path. Thus, these tube widths adequately capture entropic contributions to the path free energy.

2.2.4 Data analysis

A_RMSD

To validate convergence in ABPO, we define A_RMSD as follows:

$$A\_RMSD = \sum_{i=1}^{N} \frac{\sqrt{\sum_{j=1}^{n}\left(Cycle_{ij}^{num} - Cycle_{ij}^{last}\right)^2}}{n} \quad (2.2)$$

In this equation, N is the number of RVs, n is the number of slices, $Cycle_{ij}^{num}$ is the RV value for the ith RV on the jth slice at the end of the cycle *num*. $Cycle_{ij}^{last}$ is the RV value for the ith RV on the jth slice from the last cycle. If there are S cycles in total, we compare the RV values of the last cycle (S) to the previous cycles (1 to S-1) and calculate A_RMSD for each point 1 to S-1 and plot the data. The curve should first decrease and go flat at convergence.

Normalized RV

We used equation (2.3) to calculate the normalized RV value for a RV for each slice along the path. For each RV, RV_{state1} and RV_{state2} are the RV values for the two end-states respectively, while RV_i is the value on the i^{th} slice.

$$RV_{normalized} = \frac{(RV_i - RV_{state1})}{(RV_{state2} - RV_{state1})} \quad (2.3)$$

Table 2.1 The ABPO simulation details for each system.

System	NO. of cycles	NO. of replicas	Time steps per block	Blocks[1] per cycle	Blocks[2] per cycle	R[3]	Path optimization time (ns)
TIM	50	4	20,000	2-3	1-2	0.2	15.84
DHFR	100	4	20,000	2-3	1-2	0.4	24.64
ERα LBD	70	16	30,000	25	7	10	989.76

[1] at the beginning of ABPO

[2] near convergence of ABPO

[3] tube radius

2.3 Results and Discussion

2.3.1 TIM

As for any path-transition computational method, a first step is to identify RVs that capture the motion and are effective for sampling the transition. A natural choice of reduced variables for transitioning between open and closed positions of a short loop is the main chain torsion angles, ϕ and ψ, that distinguish the two forms. We defined the torsion angles ϕ and ψ from simulating the open (PDB ID 8TIM) and closed (PDB ID 1TPH)[95] forms of TIM shown in Figure 2.1A for 10 ns to obtain the equilibrium distribution of ϕ-ψ torsion angle values for the loop residues 166-176 at each end state. Only two residues showed distinct ϕ-ψ distributions with less than 5% overlap in the populations from the two forms (Figure 2.4A and B). Based on these populations, we defined ψ of residue 170 and ϕ of residue 171 to be the reduced variables for the transition pathway, and set the end-state path values close to the population average of the equilibrium distribution. The other residues in the loop had overlapping dihedral angle distributions in the two states (Figure 2.5) and were therefore not selected to be a reduced variable.

To convey the structural nature of the transition, we note that residues 170 and 171 are in the middle of the loop (Figure 2.1A), and not at the end of the loop as expected for a hinge-like motion. Examining the ϕ-ψ plot, we find that the loop closes by the two central residues acting as a single switch involving rotations of the dihedrals flanking the intervening peptide bond.

The initial path was set up as follows. The dihedral angle values for ψ170 and ϕ171 at the two ends of the pathway were set equal to the population average values (Table 2.2) determined from the distributions generated with unbiased simulations of the open and closed forms of TIM (Figure 2.4). The initial transition path was a set of 200 linearly interpolated points between each of the two end-state RV values. The combined set with the two RVs defined the initial 200 hyperplanes perpendicular to the curve.

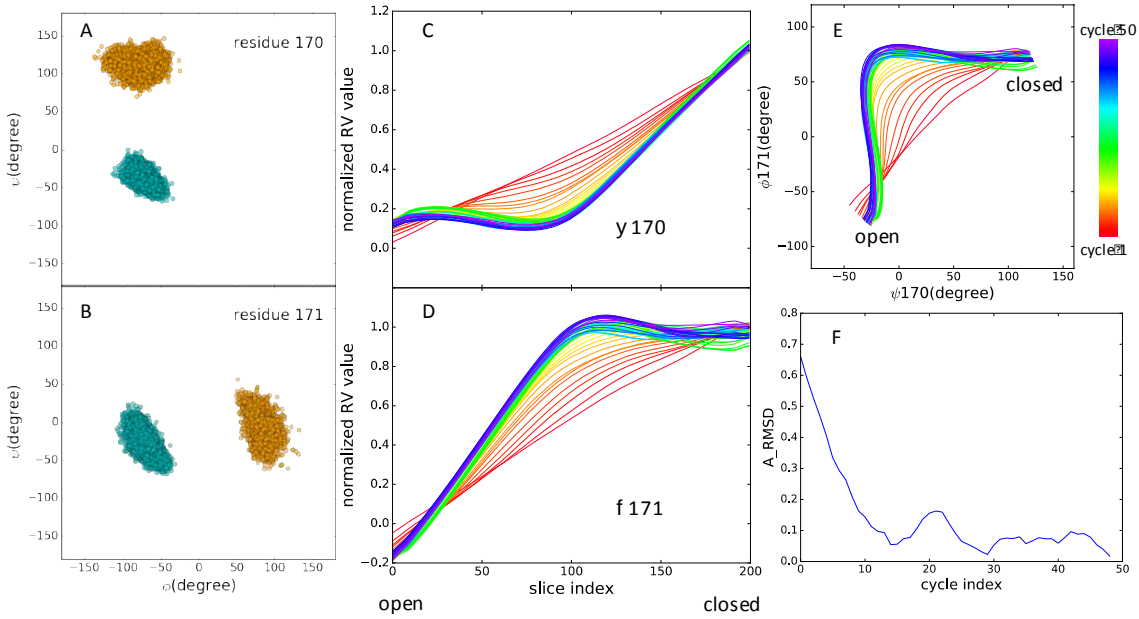


Figure 2.4 TIM transition path results from all-atom ABPO. A and B: Distributions of ϕ,ψ angles for residues 170 and 171 show distinct populations in the open and closed forms of TIM. Each dot represents a frame in the 10 ns trajectory. Cyan: open form; orange: closed form. C and D: normalized values for the two RVs at each slice (hyperplane) of the path evolving from cycle 1 to cycle 50. E: the two RVs of the ABPO calculation plotted together for each cycle to show progress and convergence of the path optimization. F: A_RMSD (see Methods) of the path at each cycle compared to the final path at cycle 50. The plateau near zero further demonstrates convergence of the simulation.

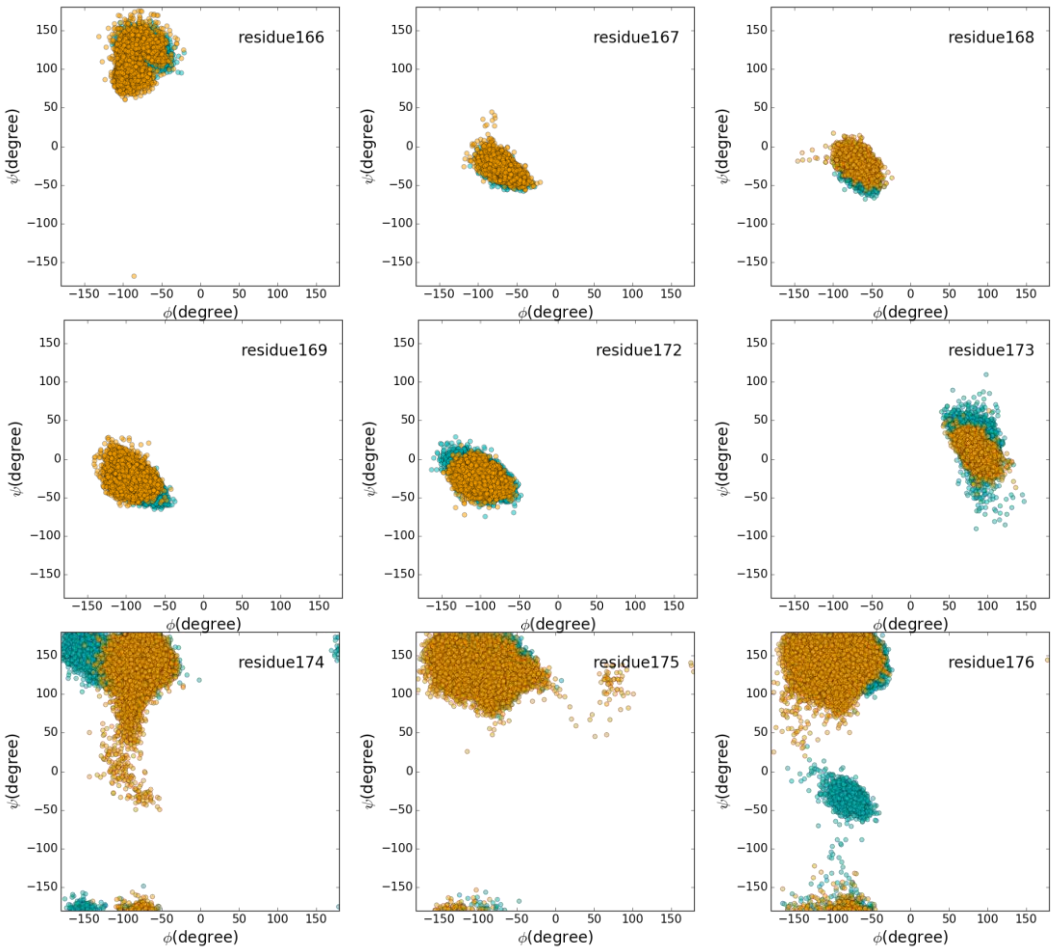


Figure 2.5 ϕ-ψ distributions for residues 166-176, except 170-171, for the open (cyan) and closed (orange) states of TIM from 10 ns equilibrium simulations. For the residues, the two states have highly overlapping dihedral angle populations sampled, and therefore are not good choices as the reduced variables for ABPO computations.

Table 2.2 The values in degree for torsion-angle RVs obtained from the close-to average structure of the MD simulation of the end states 1 and 2.

System	TIM		DHFR		
Torsion angle	ψ170	ϕ171	ψ14	ψ18	ψ19
State 1	-50.1	-56.0	-7.6	169.0	119.6
State 2	117.4	76.0	133.1	-13.9	-31.0

The ABPO computation was initiated with coordinates of the 'closest-to-average structure' of the unbiased simulations, defined by the minimum heavy-atom RMSD to the population-average structure (see Methods). The closest-to-average structure has ψ170 and ϕ171 values near the population average values. Multiple trajectories were computed with adaptive bias starting from the two end states as described above and in methods.

The ABPO approach generated a good transition path for the loop motion of TIM. The effectiveness of the bias potential (Equation 2.1) constructed along a path specified by the two RVs, ψ170 and ϕ171, is evident from the observation that the unrestricted trajectories freely sample along the path; each replica traverses nearly the full λ range (shown in Figure 2.6). The evolution of the two RVs along the transition path between the two states is shown in Figure 2.4C and D. From the plot, the transitions of the two dihedral angles do not occur simultaneously. Starting from the open state (slice 1) and moving to the closed state (slice 200), the transition of ϕ171 occurs first, followed by the ψ170 transition to complete the transition. How the two RVs of the transition path evolve together from the first cycle to the final cycle can be seen in the two-dimensional RV plot (Figure 2.4E). Based on the RV evolution along the pathway, it is concluded the optimization converged quickly after about 15 cycles. To further examine convergence of the ABPO results, we compared the path at each cycle to the final path from cycle 50 by evaluating A_RMSD (see Methods). The results are shown in Figure 2.4F as a function of cycle index. The plateau with values near zero after 15 cycles further establishes good convergence of the path. A PMF plot from the final cycle has two low free energy barriers between the two

states equal to ~2.5 and 1.5 kcal/mol respectively (Figure 2.7A), which is consistent with the short convergence time.

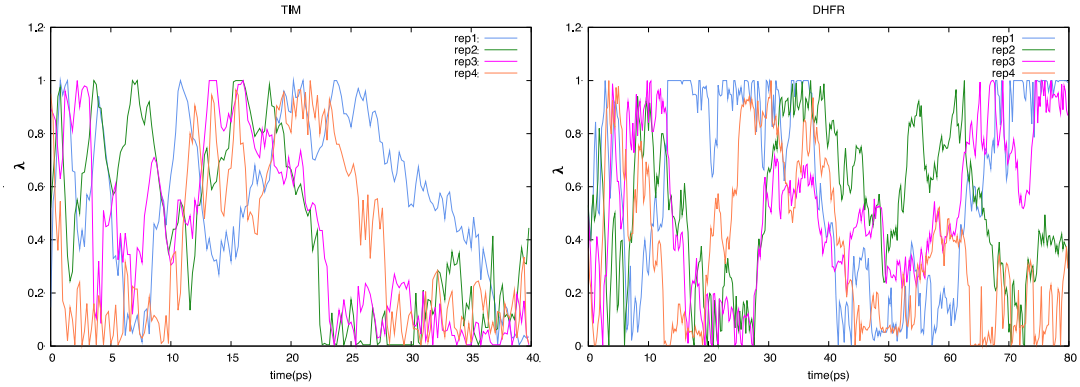


Figure 2.6 An illustration of the unrestricted sampling along the optimized path by ABPO trajectories. The visits of individual replicate trajectories to slices λ parameterizing the path length are plotted as a function of simulation time. The path sampled here is the optimized path for the loop motions of TIM (left) or DHFR (right). The bias potential is effective at enhancing sampling along the full path.

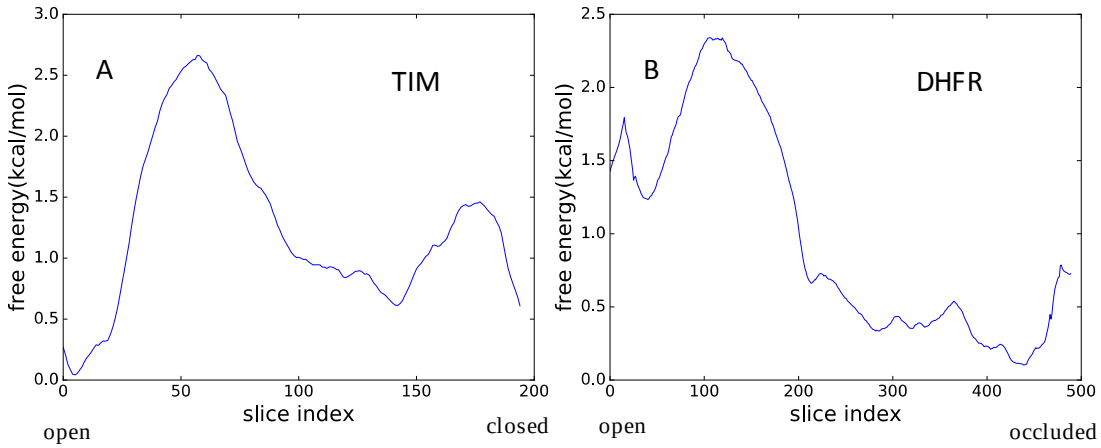


Figure 2.7 PMFs for the TIM and DHFR conformational transitions for the paths obtained using the defined RVs and ABPO. A: PMF from cycle 50 for TIM. Two barriers are shown in the PMF, corresponding to rotation of each of the two torsion angle RVs. The minimum number of visits to each slice was set to 200. B: PMF from cycle 100 for DHFR. One major barrier is observed. The smaller peaks are due to the noise of the PMF. The minimum number of visits to each slice was set to 500.

2.3.2 DHFR

As in the case of TIM, the structural difference between the two states of DHFR is the position of a loop (residues 14 to 19, Figure 2.1B), and therefore we looked to define dihedral angle RVs for the DHFR conformational transition. This reasoning that the backbone torsion angle RVs should be good descriptors of the localized loop transition is sound; however the selection of angles was not as straightforward as for the TIM loop.

ABPO reduced variables for Dihydrofolate Reductase (DHFR) loop transition

The ϕ-ψ distributions for the DHFR residues 9-24, which includes the loop residues 14 to 19 plus a few residues at each end, were determined from 10-ns simulations of the two end states starting from the crystallographic coordinates (PDB IDs 1RA2 and 1RX7)[96]. Four residues, 14, 15, 18 and 19, showed distinct ϕ-ψ distributions (Figure 2.8A) while all other distributions had overlapping populations (Figure 2.9). The main chain dihedral angles differ at the ends of the loops for residues 14,15 and 18,19, as anticipated for a hinge-like motion and in contrast to the TIM loop.

Based on the distinct populations observed in ϕ-ψ distributions (Figure 2.8A), we first defined four RVs to conduct the ABPO calculation: ψ14, ϕ15, ψ18, and ψ19. The alternate conformations of DHFR differ in both ϕ and ψ populations for residue 15; however, ϕ15 was chosen and ψ15 was not because of the overlap in the ψ15 sampling in the two states. The end-state path values were set to the population averages of the distributions shown in Figure 2.8A. Nonetheless, the four-RV bias potential did not lead to a transition path between the known end-states using the unrestrained ABPO computation. The ABPO simulations converged in 100 cycles, but to a path that lacked transition of ψ14 and ϕ15 as shown in Figure 2.10A. It is noted that the ABPO trajectories were launched from both end states and sampled the range of ψ14 and ϕ15 values in initial ABPO cycles, but converged to a path with values corresponding to only one of the end states along the path. That is, using unrestrained ABPO the normalized RV value of ψ14 and ϕ15 did not transition between 0 and 1 in the final cycle (Figure 2.10A).

We reasoned that additional features beyond the four RVs were needed for the transition and added ψ15 to the set of RVs, constituting a five-RV case. Still, the unrestrained ABPO with this set of RVs did not converge to a path with transition for all

dihedrals with distinct populations. In the five-RV path shown in Figure 2.10B, ψ15, ψ18, ψ19 transition between one state and the other, while in the converged path (dark blue curves) ψ14 showed an incomplete transition and ϕ15 remains in the vicinity of one state over the length of the optimized path. All major structural differences of the backbone ϕ-ψ angles between the two states of DHFR are included in the five-RV case, yet an acceptable transition pathway was not achieved even though the path optimization was efficient and converged in 25 ns of simulation time. We therefore conclude that missing components in the RV set is not the cause of the incomplete transition. Alternatively, we considered that the near overlap in the ϕ-ψ space sampled by residue 15 in the two end-states, coupled with the inclusion of ψ14 RV, which rotates the same peptide group as ϕ15, might be the cause of the problem.

The possible issue related with residue 15 was tested by launching ABPO with only three RVs: ψ14, ψ18, and ψ19. Unrestrained ABPO generated the expected path with the three RVs showing complete transitions along the path as described in the Figure 2.8C-E.

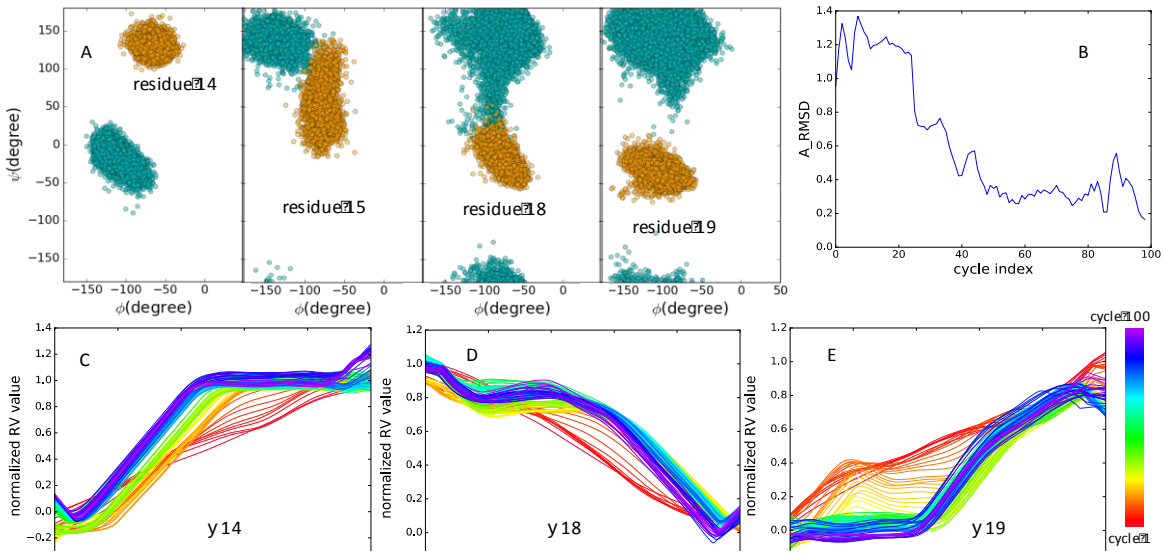


Figure 2.8 DHFR transition path results from all-atom ABPO. A: distributions of ϕ,ψ backbone angles for the four residues with largely distinct populations in the open (cyan) and occluded (orange) states. B. A_RMSD of the path at each cycle compared to the final path at cycle 100. C-E: evolution of the path during the ABPO computation showing the normalized value for the three RVs at each slice (hyperplane) of the path from cycle 1 to cycle 100. The tight overlap of the paths indicates convergence of the optimization.

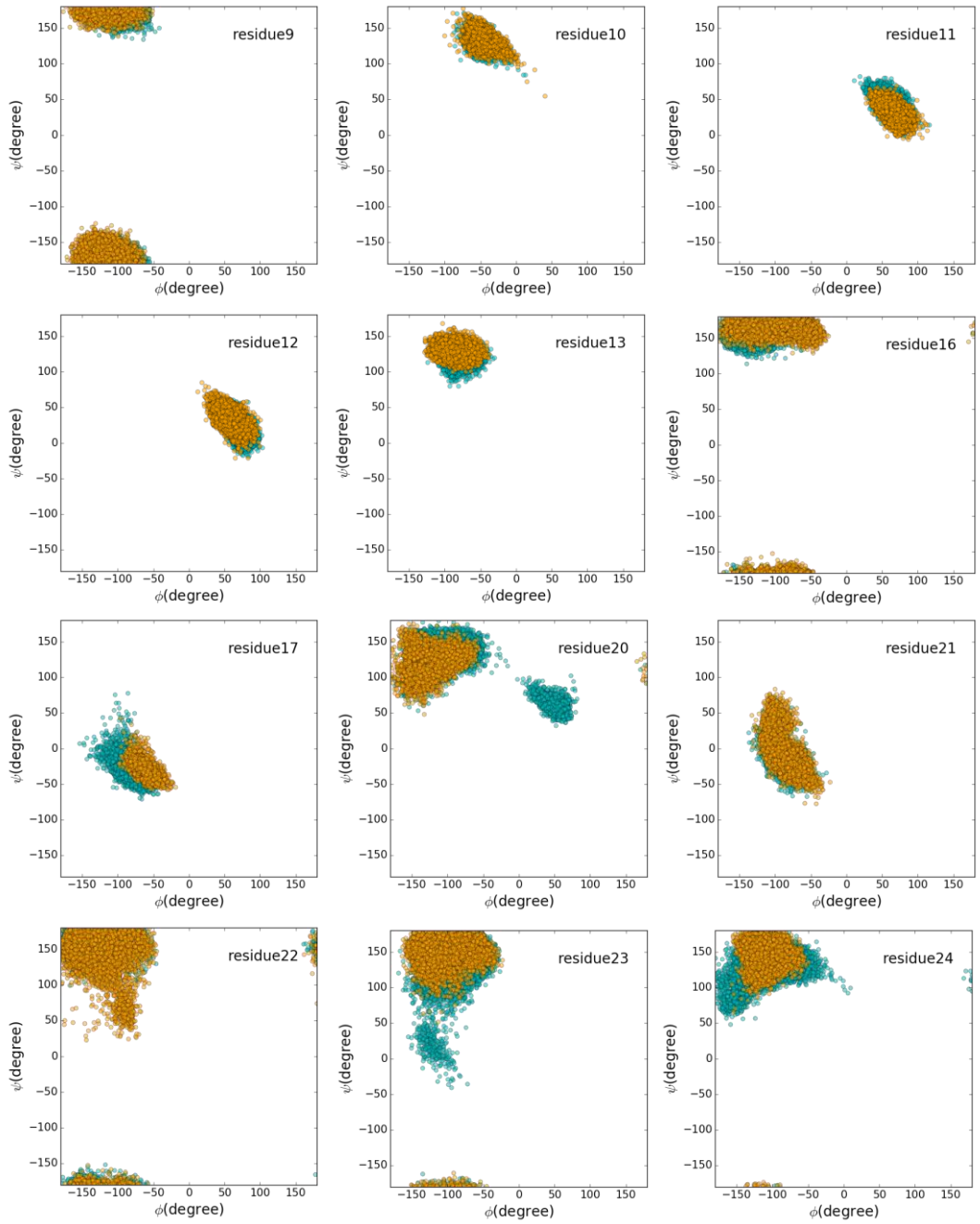


Figure 2.9 ϕ-ψ distributions for residues 9-24, except 14,15, 18 and19, for the open (cyan) and occluded (orange) states of DHFR from 10 ns equilibrium simulations. For these residues, the two states have highly overlapping dihedral angle populations and were not selected for RVs.

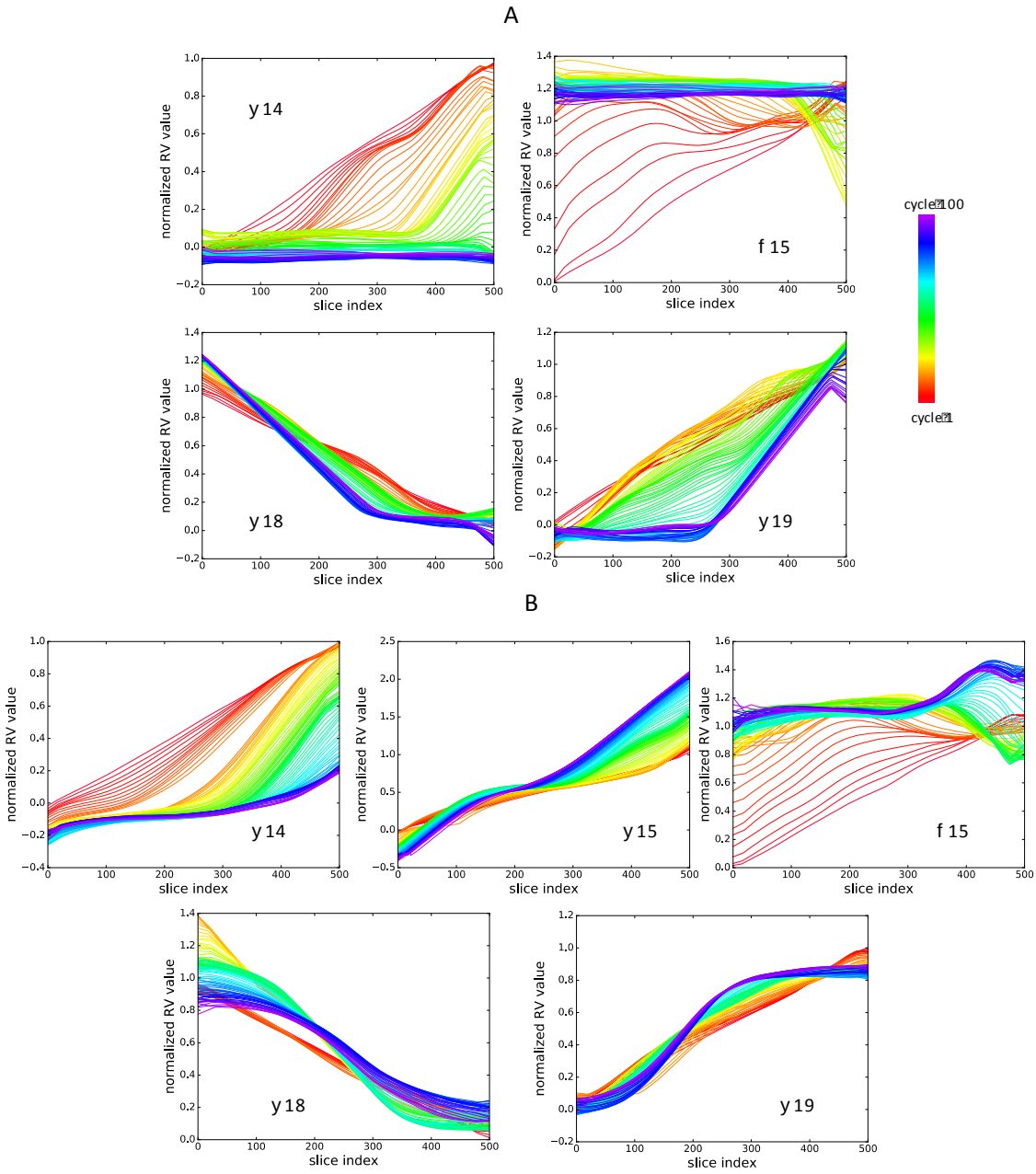


Figure 2.10 A: Evolution of the four normalized RV values for ABPO simulations of DHFR. In the plot, ψ14 and ϕ15 did not transition between the end states in the final cycle. ψ14 stayed in the open state, while ϕ15 stayed in the occluded state. B: Evolution of the five normalized RV values for ABPO simulations of DHFR. ψ14 showed an incomplete transition in the final cycle, and ϕ15 mostly stayed in one state. Normalized RV values are adjusted to remove dihedral periodicity.

The three-RVs yielded a good transition path suggests the nearly overlapping distributions in ϕ-ψ of residue 15, along with inclusion of the neighboring ψ14, could be the reason residue 15 dihedral angles are poor RVs.

To further examine the dependence of defining the pathway on the choice of RVs, we asked if fewer than three RVs are sufficient to formulate the transition, and investigated the use of two RVs: ψ 14 and ψ18, or ψ14 and ψ19. The two RV cases demonstrated that although the transition of the two residues developed, the transitions of other three dihedrals did not complete, showing that two RVs were not sufficient to characterize the loop movement. In Figure 2.11, ψ14 and ψ18 were defined as RVs, and from the scarcity of points in the range of values intermediate to the end states, the other three dihedral angles did not transition along the path.

As implemented here, the ABPO approach does not fix the path ends; the RVs over the length of the path, including the ends evolve during path optimization, which can assist in judging if a set of RVs is appropriate for capturing the features relevant to the conformational transition. Some steps that developed out of the ABPO implementations covered in this study include: 1) a directional and efficient evolution of the RV value from the initial path value is needed to converge the transition pathway in a tolerable number of cycles; 2) each RV should converge to its known value at the two end states in the optimized pathway. Because our implementation of ABPO does not fix the ends of the path, the expected end-state values are not guaranteed. (In the previous ABPO work, the ends of the path were fixed[81].); and 3) all parameters that characterize the difference between the two states should transition, even if some of the parameters are not defined as RVs. We note that guidelines 2 and 3 are simply requirements that the final path connects the expected end-states.

The main chain dihedral angle populations from unbiased MD simulations differ for residues 14,15 and 18,19 at the ends of the loop (Figure 2.8A), as anticipated for a hinge-like motion and in contrast to the TIM loop. On the criterion of having distinct populations in ϕ or ψ with essentially no overlap along the given dimension, we first selected four RVs to conduct the ABPO calculation: ψ14, ϕ15, ψ18, and ψ19. The ABPO pathway converged; however, the path that was generated did not include rotation of ψ14 and ϕ15. To find better RVs, we reasoned that including both ψ14 and ϕ15 could be problematic because

they both act on the same peptide group connecting residue 14 and 15. We therefore removed ϕ15 as an RV and used only three RVs: ψ14, ψ18 and ψ19. The bias potential constructed on these three RVs generated an optimal path with the expected complete transitions of ψ14, ψ18 and ψ19 over the course of the pathway (Figure 2.8C-E).

In the RV plots from the open to the occluded state, the ψ14 transition starts early, then the ψ19 transition initiates around 200 slice of the path, while the ψ18 transition is a continuous process that traverses the whole path. The good convergence of the path optimization is shown by the tight overlap of the final cycles (purple, Figure 2.8C-E), and further described by the A_RMSD against the final path (Figure 2.8B). From the plot, the optimization converged after about 50 cycles, taking almost twice the computation time compared with TIM (Table 2.1). Further, as with the TIM loop transition computation, the ABPO trajectories sample nearly the full length of the path given the bias potential and unrestrained sampling; plots of the position λ as a function of time for the replicate trajectories are given in Figure 2.6. The PMF from the last cycle has a single free-energy barrier between the open and occluded loop conformations of DHFR (Figure 2.7B).

To determine if in the three-RV case of ABPO all five dihedral angles identified in the ϕ-ψ plots (ψ14, ϕ15, ψ15, ψ18, and ψ19) actually transitioned between both end-state populations shown in Figure 2.8, we extracted the time series for the five dihedral angles from the final ABPO cycle. We found from the time series (Figure 2.12) that even though only three of the five torsion angles, ψ14, ψ18 and ψ19, were used to bias sampling, the other two dihedral angles, ϕ15 and ψ15, also transition given that the time series includes angle values intermediate to the two end-state values. This result testifies that adaptive biasing in the space of the three RVs can properly determine the complete transition process including features that are not part of the RVs.

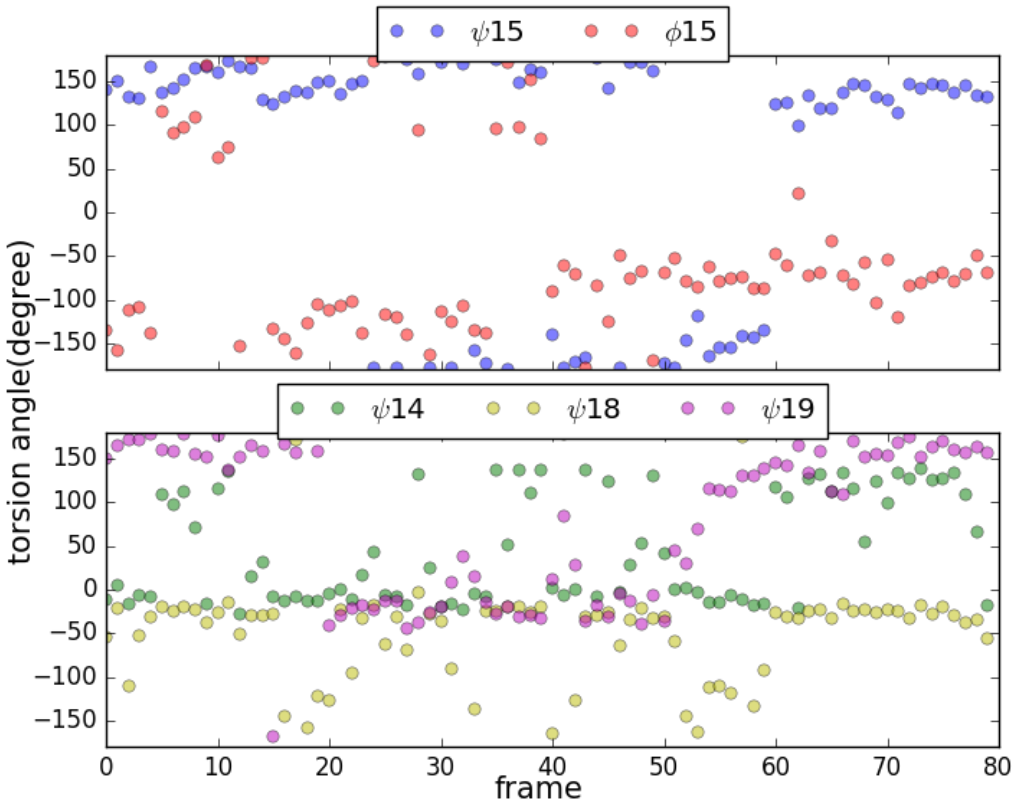


Figure 2.11 When ψ14 and ψ18 are defined as RVs, these two dihedrals sample intermediate angles along the path. The other three dihedrals that were not defined as RVs, ϕ15, ψ15 and ψ19, had significantly less sampling along the path as shown from few points intermediate to the end states. The ABPO computation included four replicas, with two trajectories initiated from each end state. The final cycle had one block of 20,000 steps with a 1,000-step saving frequency for each replica, totaling 80 frames.

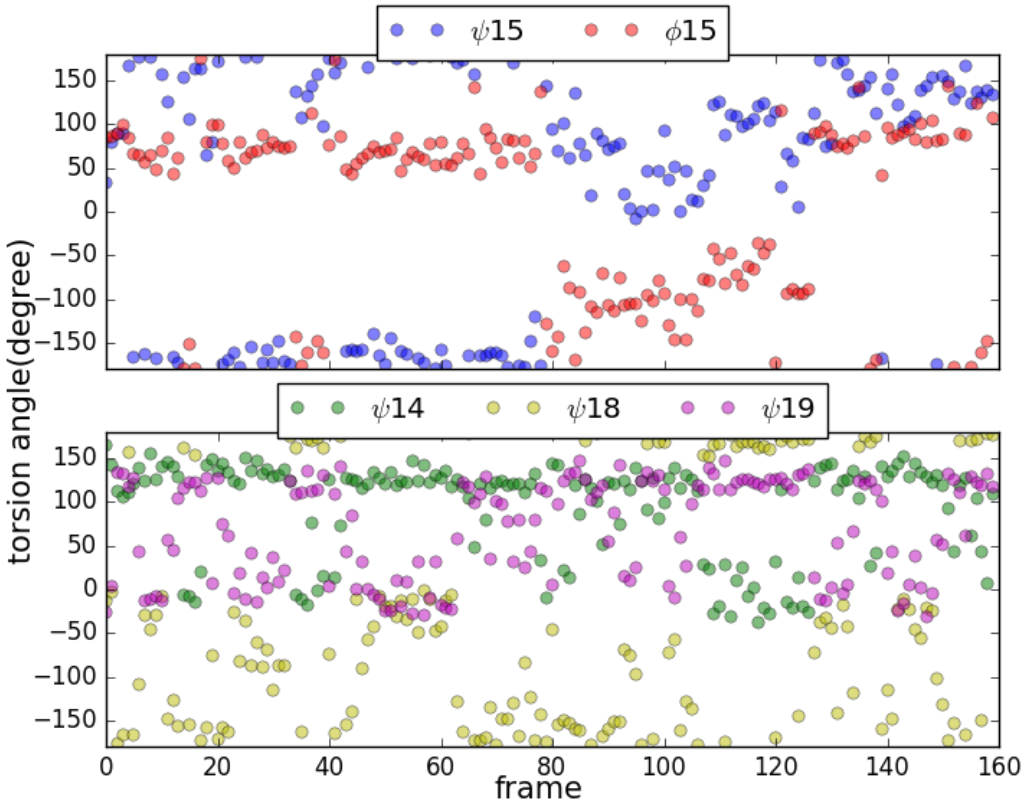


Figure 2.12 In the three-RV case, ψ14, ψ18 and ψ19 are used to compute the transition path of DHFR using ABPO. The four replicate trajectories from the last cycle, two initiated from each end state, were concatenated and torsion angle time series were extracted from the combined trajectory for the three RVs and for the two other torsion angles that differ between the two end states, ϕ15 and ψ15. The last cycle included two blocks, and each block had 20,000 steps with a 1000-step saving frequency, totaling 160 frames. The three RVs ψ14, ψ18 and ψ19 sample the full path as shown by visits to intermediate angle values. The other two dihedrals, ϕ15 and ψ15, also sample the two end states and intermediate angles in this final cycle, showing the computed path captures the conformational transition.

2.3.3 ERα LBD

We applied ABPO to examine the transition of H12 between the agonist and antagonist-bound forms. The actual timescale specific to the H12 transition is unknown, although the NMR studies suggest a longer timescale (ms) than associated with localized loop transitions. To better understand the dynamics of H12 in ERα LBD, we computed unbiased MD simulations for 6.2 μs of each of the two structures. The agonist and the antagonist that stabilize the two states were removed in the simulations. Even without the ligands stabilizing the two structure, no transition-like conformational changes were observed in the long simulations, and the position of the H12 on the surface of ERα LBD is stable over low μs timescales (

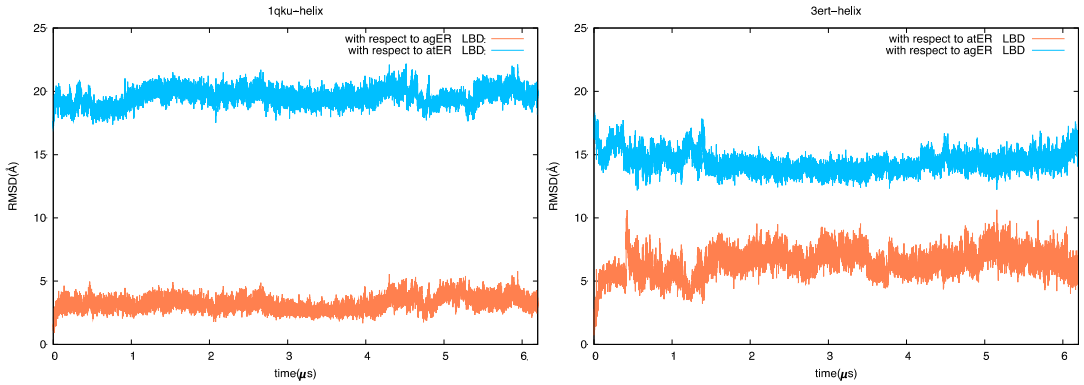


Figure 2.13). These MD results combined with NMR suggest that H12 displacement is unlikely to occur within the timeframe of simulations possible with currently available resources, so that to explore the transition details it is necessary to use enhanced methods, such as ABPO.

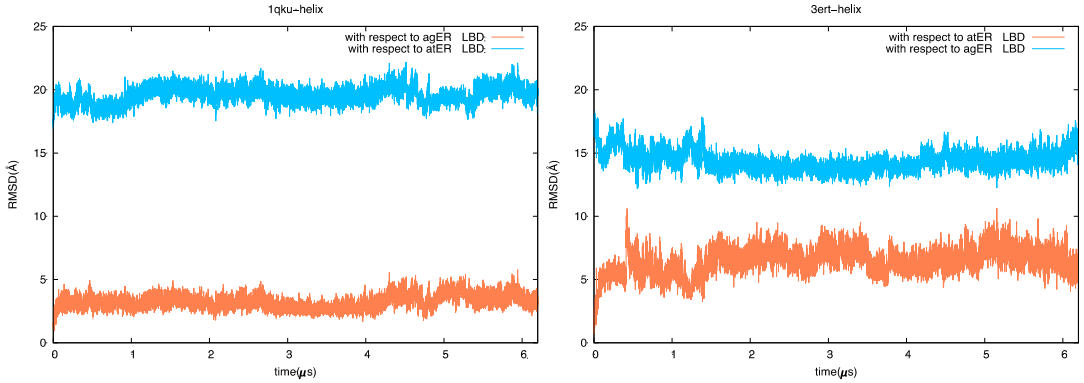


Figure 2.13 The two end structures from crystallography for ERα LBD were solvated with TIP3P and an unbiased simulation was computed for 6.2 μs. For frames from each trajectory, the RMSD for H12 was calculated with respect to the initial coordinates of H12 in either of the two forms of ERα LBD. The time series is to detect any potential movement of the helix. The RMSD values for agERα LBD show the position of H12 is stable after about 0.2 μs with a RMSD value around 3.5 Å. For atERα LBD, H12 RMSD fluctuates around 5 Å, but no transition between the two states was observed in the simulation. These results show enhanced sampling techniques are necessary to study the transition.

Alternative variables were explored to define the RVs for computing the transition between the two states of ERα LBD. First, we examined the backbone dihedral angles as RVs. In contrast to the loop transitions in TIM and DHFR, the transition in ERα LBD is a helix movement. In the agonist-bound form (called agER LBD hereafter), H12 (residues 537-543) interacts with H11 and the N-terminus of H4; in the antagonist bound form (called atER LBD hereafter), H12 interacts with H6 and the C-terminus of H4 (Figure 2.1C). We extracted the ϕ-ψ time series of the residues 526-545 from 10-ns simulations using the crystal structures to obtain the ϕ-ψ distributions for atER LBD (PDB ID 3ERT)[97] and agER LBD (PDB ID 1QKU)[113]. These residues include the coil region connecting H12 and H11. Only four residues, located at either end of the coil N-terminal to H12, had distinct ϕ-ψ distributions (

Figure 2.14). Based on the distributions, five dihedral-angle RVs were defined: ϕ532, ψ532, ψ533, ψ536, and ψ537. The ABPO simulation converged within 30 cycles based on the tight overlap of the evolution of the normalized RV values along the path. Nevertheless, examination of the structures at the computed end states of the path found that the dihedral rotation of the torsion angles RVs was achieved by residue movements localized to the coil without reposition of H12 on the surface of the protein. Based on this observation, we concluded that dihedral RVs were insufficient for transitions involving helix contacts and movements more complex than loop motions.

Distance-based RVs were therefore explored with the rationale that inclusion of the H12 contacts with other ERα LBD residues is needed to capture the essential structure features of the transition. The RVs chosen are a linear combination of multiple inter-residue distances[79]. The procedure to obtain a linear combination of multiple inter-residue distances to define RVs for the ERα LBD transition was as follows. We ran a 10-ns simulation of the two end structures, and calculated the closest-to-average structure for each system from the last 4 ns of the trajectories as described in methods. For each structure, we identified the residue pairs that had side chain heavy atoms within 4.5 Å of each other, and compared the residue pairs from the two structures to determine Cα-Cα distances that differed by more than 1 Å in the two structures. These residue pairs were grouped based

on their spatial proximity, and the individual Cα-Cα distances for each residue pair within a group were combined linearly using equation (3) in Methods to define an RV. In general, the pairs between one residue on H12 and other residues outside H12 are grouped into the same RV. The distance values for the RVs were extracted from the two closest-to-average structures, and the RVs comprise only Cα-Cα distances with no side chain distances involved. For the transition in ERα LBD, we first defined nine combined-distance RVs, but two of them were eliminated due to "high noise" or large fluctuations on the normalized RV path evolution plots, indicating the RVs are not effective for promoting the transition. The final ABPO calculation included seven combined-distance RVs. A list of the residue pairs is in Table 2.3. The location of the residues in the seven RVs is shown in Figure 2.15H-I where the residue pairs in one RV have the same color and each RV color is unique. The progression of the seven combined-distance RVs is shown in Figure 2.15. Although less cycles were used, the optimization in fact took much longer to converge compared with the previous two systems. For the previous two systems, only 3 blocks per cycle were needed; in ERα LBD, the simulation reached the specified maximum of 25 blocks per cycle at the beginning of the optimization (Table 2.1), which indicates the bias potential did not adapt sufficiently in the allotted simulation time for the initial cycles to promote complete sampling of the path. Nevertheless, the changes in the path variables from the initial guess were sufficient enough to achieve convergence to the principal curve in the subsequent cycles.

The PMF computed from the principal curve for the ERα LBD transition is shown in Figure 2.16A. The choice for the tube width, R, used to compute the trajectories that determine $h(\lambda,t)$ to estimate $A(\lambda,t)$ (Equation 2.2) must be larger than the reaction channel in order to accurately capture the entropic contribution to the free energy. The PMF is roughly invariant to the R values used (results shown in Figure 2.3) so the entropic component is reasonably accounted for in the PMF shown in Figure 2.16A. In addition, the results in Figure 2.3 include a second round of accumulating histograms with the same optimized path as Figure 2.16A and with R equal to 10Å. The similarity in the plots demonstrates reasonable certainty in the estimates for the PMF.

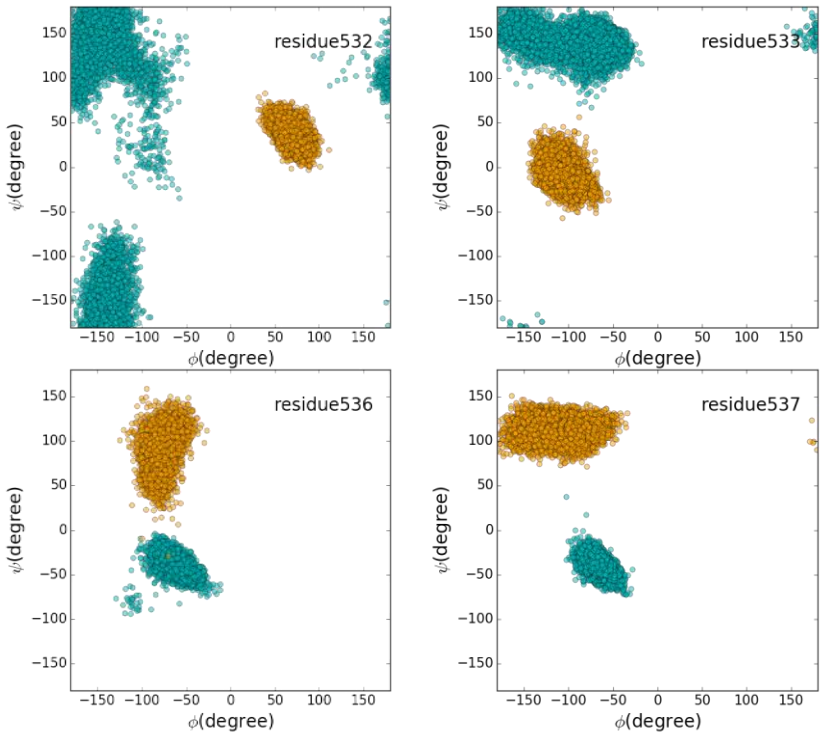


Figure 2.14 The ϕ-ψ distributions of the four residues in ERα LBD. Cyan: atERα LBD; orange: agERα LBD.

Table 2.3 The individual Cα-Cα distance pairs in each distance combination RV for ERα conformational transition.

RV number	Residue pairs
1	531-339, 533-339, 533-340, 533-343, 534-343, 535-340, 535-343, 535-344
2	536-350, 536-354, 536-383
3	537-347, 537-351, 537-355, 537-376, 537-380
4	540-346, 540-347, 540-358, 540-376, 541-372
5	543-355, 543-358, 543-359, 543-362, 543-376, 543-379, 543-383
6	544-362, 544-372, 544-383, 544-522, 544-525
7	546-380, 546-381

The transition path has a maximum free-energy near 0.4 of the path total length (Figure 2.16A), which likely is the reason for the more slowly converging optimization of this

transition path. Although the barrier height seems high given the timescale of the motion estimated from NMR[105-107], it is useful to examine the mechanism of the transition. By looking at the principal curve for each RV (Figure 2.15), RV3 and RV4 plateau near the same slice index as the free energy barrier, followed by a rapid change. RV3 and RV4 together describe the switching of the position of H12; RV3 and RV4 distances are the contacts between H12 and the two helices, H4 and H6, respectively. Based on this behavior of RV3 and RV4, the breaking and reforming of the interactions of H12 with H4 and H6 are suggested to be a key step in the transition.

A minimum in the PMF falls near slice 700, rather than near the end-state slice 1000, where the RV values are defined from the crystallographic structure (Figure 2.16A). The reduced coordinates that vary between slice 700 and 1000 are RV1, RV2 and RV3 (Figure 2.15A-C), which correspond to displacement of the flexible loop between H12 and H11. Further, the normalized value of RV1 extends to almost 2, *i.e.* RV1 values beyond those of the atER LBD end state. The shift of the free-energy minimum, and the extended values of RV1 could occur as a result of the flexibility of the loop. That is, if the conformational fluctuations of the loop differ in solution compared to the conformational space accessible in the crystal, a shift in the position of the free-energy minimum could occur. Alternatively, the implicit-solvent simulations may not properly account for the conformational equilibrium of this solvent-exposed loop[112], which could also give rise to the observed displacement of the minimum.

We extracted structures along the principal curve from the trajectories of the final cycle of the ABPO simulation to visualize the transition pathway. A frame for each slice was used to construct a structure series showing the transition from agER LBD to atER LBD. The structure series showed some clear features in this process: from agER LBD to atER LBD, the N-terminus of H12 breaks interaction with the N-terminus of H4, then the H12 C-terminus interaction with H11 is lost, so that H12 is more solvated. Gradually the C-terminus of H12 forms alternative interactions with H4 C-terminus and H12 N-terminus with H6. In structures extracted at the transition state of the PMF, H12 contacts neither H4, H6 nor H11. A representative structure near the transition state is shown in Figure 2.16B.

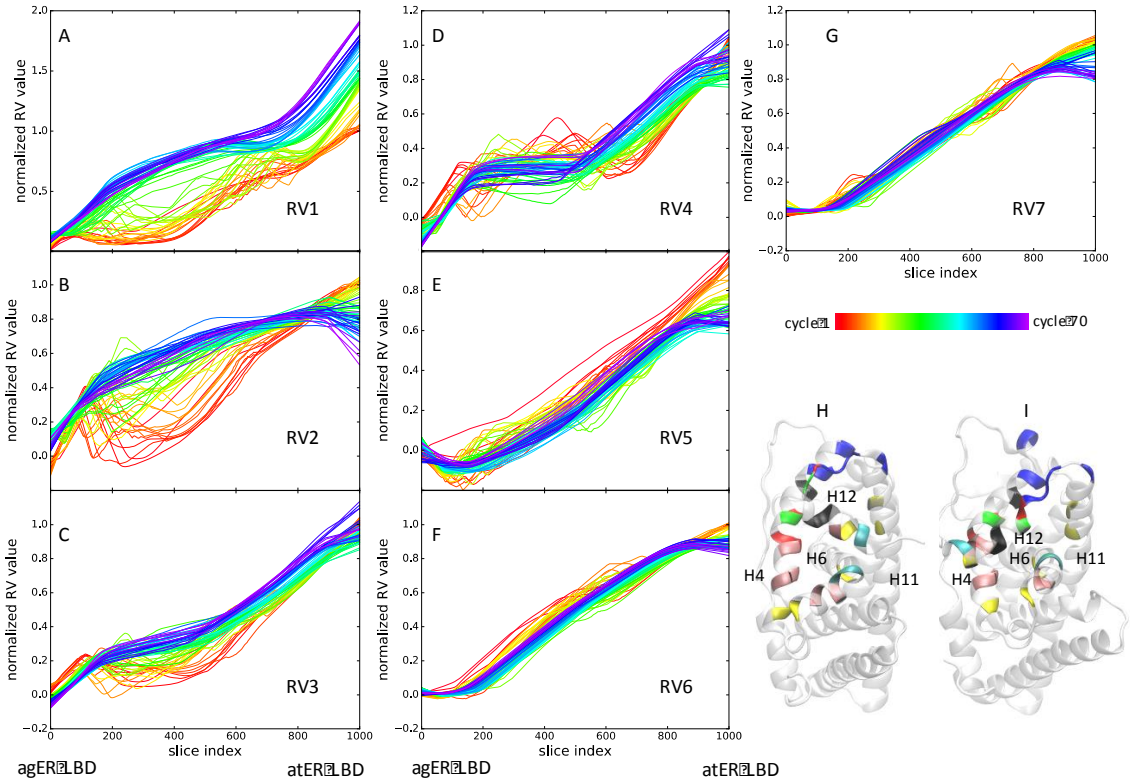


Figure 2.15 ERα LBD ABPO results. A to G: evolution of the normalized value for the seven Cα-Cα-distance RVs (see Supplementary Table 1 for a list of residues) show convergence. H agERα LBD and I atERα LBD closest-to-average structures from the equilibrium simulation: residues in each RV are colored differently to show their locations. From RV1 to RV7, each RV is colored 1) blue, 2) red, 3) green, 4) black, 5) pink, 6) yellow and 7) cyan.

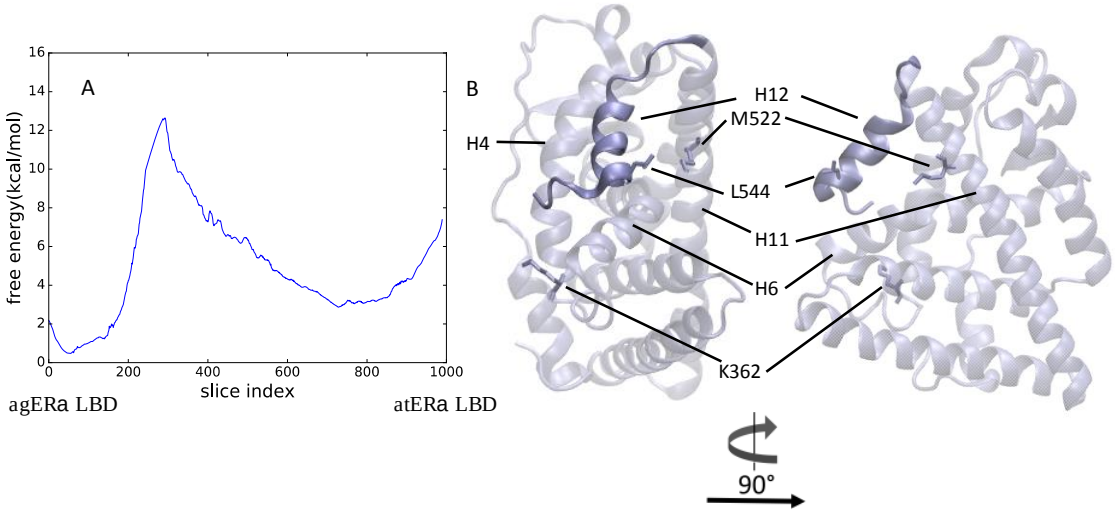


Figure 2.16 An intermediate state structure in ERα LBD transition path from agERα LBD to atERα LBD. A: PMF along the transition path from the final ABPO cycle shows one major free-energy barrier. The minimum number of visits to each slice was set to 100. B: a snapshot at the free-energy barrier, where, in the transition from agERα LBD to atERα LBD, the interactions of H12 with H11 are broken and the interactions of H12 with H4 have not formed. H12 is in opaque ribbon representation. In agERα LBD, L544 interacts with M522, while in atERα LBD, L544 interacts with K362.

2.4 Conclusions

In conclusion, ABPO is demonstrated here to be an effective method to optimize conformational transitions of three protein systems using an all-atom force field. The path evolves efficiently from a starting path to the final converged path under the forces of the adaptive bias potential. A highlight of ABPO is the free sampling in a tube around the path, which allows exploring multiple channels within the tube radius, and also the quality of the RVs to be readily assessed based on the ability of trajectories to traverse the path and describe the expected transition. In addition, the ABPO formalism provides a straight forward evaluation of the PMF.

We expanded the application of the unrestrained ABPO approach from the coarse-grained Gō model[81] to all-atom protein conformational transitions and found the bias potential to be effective at enhancing the sampling along pathways specified in reduced-variables in the higher resolution space and more rugged potential surface of an atomistic protein model. The trajectories, localized to the path by a tube potential but otherwise not restrained to the path, freely sample the path and showed good convergence to the final optimized curve. Compared to Gō-models with only Cα atoms represented, the all-atom model allows RVs to be in terms of main chain dihedral angles for motions localized to two sequential residues, which is convenient for exploring loop transitions.

The unrestricted sampling of ABPO was effective for moving the path from the initial guess, which is not only important for converging to an optimal path but also allows for efficient exploration in the choice of select RVs that capture the features of the motion. In the case of the switch in H12 position in ERα LBD, RVs derived from a linear combination of Cα-Cα distances that differed in the end states was found effective while dihedral angles were not. And, it was not necessary to include side-chain atom distances as the side chains moved with the main chain in the ABPO trajectories.

The transition-pathway computations from the three systems validate ABPO as an efficient method to calculate protein conformational transitions. The path computed with the ABPO approach for the DHFR loop included rotations of dihedral angles that were not specified as RVs, which speaks to the reliability of approach.

CHAPTER 3. SRC KINASE DOMAIN ALL-ATOM CONFORMATIONAL ACTIVATION USING ABPO

3.1 Introduction

The Src conformational activation process has been suggested from crystal structures[114-117]. NMR has been used to study the dynamics of Src SH3, SH2, unique and kinase domains[118-119][112][121-122]. The solution dynamics of Src-ligand complex reveals long-range communication between ligand binding and regulatory sites[123]. However, the atomic details of the activation process remain unrevealed. Besides, the study of the active form of Src is limited by the detection of the A loop in biophysical experiments. In the crystal structures, the active form of Src usually has the A-loop partially unsolved. Only two crystal structures, an unphosphorylated c-Src in complex with an inhibitor (PDB ID 1Y57 [117]) and c-Src kinase domain T338I mutant in complex with ATP (PDB ID 3DQW [124]), has the complete A-loop. In NMR, the peaks for the loops are usually missing due to the high flexibility of the regions.

Computational simulation methods are valuable in elucidating the atomic details of conformational transitions. Specific techniques are required to overcome the high free energy barrier in the transition process. Several types of accelerated sampling methods have been applied to study the Src conformational transition as introduced in Chapter 1. The transition kinetics, the order of some specific events during the transition, the important intermediate states and residue-residue interactions have been reported from these simulations. These techniques usually require an initial path which can be arbitrary, or statistical methods to build the path, while a continuous, unrestrained transition has not been established.

Here we apply ABPO to the all-atom Src kinase domain. Different from the transitions in Chapter 2, Src conformational transition is more complicated in both types of the transition and the number of residues involved. It involves rigid body movement/rotation like the displacement of αC helix; also, it embraces protein unfolding like the structural change of the A-loop. The protein folding/unfolding problem itself remains an ongoing research topic in computational simulations. How to define reduced variables for a combined set of transitions is a challenge. The details of the transitions like the sequence

of the events also need to be determined. We first define and optimize the reduced variables to get a converged transition path. Then we analyze the transition path and the structures generated during the optimization process. This is the first direct observation of an unrestrained, continuous path optimization of Src kinase domain using computational methods.

3.2 Methods

3.2.1 Simulation systems

Src kinase domain in active and inactive states are from the PDB entries 1Y57 [117] and 2SRC [116] respectively. Residues 255-521 were included in the simulations. All ions, ligand and crystal waters were removed from the structures. The coordinates of the missing atoms in the two structures and all hydrogen atoms were added using CHARMM IC BUILD facility. Simulations of the proteins were carried out using CHARMM 22 all-atom force field with CMAP dihedral angle corrections with implicit solvent model FACTS. To prepare the systems for ABPO, we first performed energy minimization on the two crystal structures. The energy was minimized using the steepest descent and Powell algorithms to a gradient less than 1.0. The initial velocities were generated from Gaussian distributions at 100K, then the energy-minimized structures were heated to and equilibrated at 298 K for a total of 500 ps. The leapfrog integrator was used to calculate the trajectories with a 2 fs time step.

A 10-ns simulation in Langevin dynamics was initiated using coordinates from the end of the equilibration run at 298 K. The long-range interactions cutoff distances were set to 10, 12 and 14 Å. The time series of temperature, potential energy, and heavy atom RMSD with respect to the energy-minimized structure were monitored to assess the stability of the simulation. The closest-to-average structures from the last 4ns of the trajectories were used to define the distance RVs, to set values for the RVs at the end states and initiate the ABPO simulations. The closest-to average structure is determined as described in Chapter 2.

3.2.2 ABPO parameters

The transition pathway was optimized using the ABPO module implemented in CHARMM. The ABPO was started from two end structures with 16 replicas, 8 replicas for

each end structure. Each block has 40,000 time steps. The maximum number of blocks per cycle was set to 30. The total simulation time was 1.225 μs. In all simulations, Langevin dynamics was used at 298K with a time step of 2 fs. The fraction of the free energy cancelled by the bias potential was 0.8 (b in equation 1.1, CHARMM BFCT = 0.8). The coupling of the bias to the dynamics was 2.5 t^{-1} (c in equation 1.1). The histogram for visits to path slices are smoothed using a Gaussian mollification factor set to 0.05. The number of slices was set to 2000 as indicated in the plots for each path to achieve desired sampling resolution along the path. For the tube radius and potential parameters, the radius was 20Å, and the force constant was 5 kcal/mol to enable efficient sampling of the system. The tube radius was chosen based on an estimation of the RMSD difference between the two states.

The metric tensor **D** was evaluated using equation 2 in reference [81]. The **D** was evaluated from short unbiased simulations at each end state and the inverse of **D** was stored for input to ABPO. For Src kinase domain, **D** was estimated from 2 ns unbiased simulations with a 2 fs time step. The **D** is viewed as a constant for further calculations.

Distance combination RVs for Src kinase domain

We divide the kinase domain into 4 regions: αC helix, Nlobe (excluding αC helix), Aloop, and Clobe (excluding A-loop). The residue numbering of the regions is listed in Table 3.1. We looked at residue pairs in the region pairs: αC helix-αC helix, αC helix-Nlobe, αC helix-Aloop, αC helix-Clobe, Aloop-Aloop, Aloop-Nlobe and Aloop-Clobe. The Cα-Cα distances were extracted from the two close-to-average structures from the 10ns equilibrium simulations, and the residue pairs that have distances differ by 2.5 Å in the two states were identified. Then the residue pairs in each region pair were grouped into distance combination reduced variables using equation 2.1. For the residues with long side chains, the farthest heavy atom from Cα atom was also used to calculate the residue-residue distances, including Cα-sidechain distance and sidechain-sidechain distance. The sidechain-related pairs were included in the calculation where the Cα-Cα distance difference between the two states is smaller than the sidechain-sidechain distance difference. Preliminary ABPO simulations for ~10 cycles were used to eliminate inappropriate reduced variables. Inappropriate RVs were determined that 1) there is a high flat region in the pmf indicating no sampling in the middle of the path, 2) the normalized RV does not evolve from 0 to 1, except for the N-linker residue based RVs, 3) the

normalized RV evolution profiles are not reproduced well in independent simulations with identical input. The N-linker residue based RVs are very flexible and might not transition from exact 0 to 1 as other RVs. The RV profiles that are not reproduced well indicate that the distance difference between the two states might be due to the structural flexibility of the protein in solution, not that the difference is essential to the transition. A complete list of the final reduced variables is in

Table 3.2.

Table 3.1 The residue numbering of the regions in c-Src kinase domain.

Region	N-linker	N lobe	C lobe	A-loop	αC helix
Residue NO.	255-259	260-341	342-521	404-424	304-316

Table 3.2 List of residue pairs in each distance combination RV for Src kinase domain activation conformational transition. Cα-Cα distances are used unless otherwise noted.

RV number	Residue pairs
1	311-260(NE1), 255-308, 255-311, 255-312, 256-311, 256-312, 257-311, 257-312
2	307-295, 307-296, 307-297, 307-334, 307-335, 307-336
3	310-295, 310(CD)-295(CE), 310(CD)-382, 310-382, 310(CD)-409(CZ), 310-410, 310(CD)-403, 310(CD)-404, 311-325
4	406-410, 407-410, 407-411, 408-411, 409-412
5	412-417, 413-416, 413-417
6	414-417, 414-418, 415-418, 415-419, 415-420, 416-419
7	413-423, 415-423, 416-424
8	410-302, 411-278, 411-301, 411-302, 412-278
9	410-380, 410-381, 410-382, 411-381
10	416-386, 416-388, 416-428, 417-385
11	422-437, 422-439, 422-433, 423-433, 423-429

3.2.3 Data analysis

The normalized RVs are calculated using equation 2.3.

3.3 Results and Discussion

3.3.1 Reduced variable selections and path evolution profiles

ABPO uses reduced variables that capture the structural changes necessary and sufficient for the transition, and the choice of reduced variables is a critical step for the pathway optimization. Four types of geometric reduced variables are implemented in ABPO, namely internal distances, or a linear combination of distances, angles and torsion angles. The scale of the conformational transition of Src kinase domain is larger than previous systems, and several types, or combinations of reduce variables have been attempted for the transition. From the active to the inactive conformation, the Aloop undergoes a structural change from an extended loop to a folded two-helical structure, and the values for torsion angles would have altered during this process. The difference in backbone torsion angle values is validated by visualizing the ϕ-ψ distributions of the residues in the loop region residue 404-424 (Figure 3.1). Some residues with largely overlapping distributions are not included in the figure. We ran ABPO with a combination of distance-based RVs for αC helix rotation and dihedral angle RVs for A-loop folding. Nevertheless, in the simulation, we observed that when the torsion angle RVs were used for the loop, the RVs showed progress along the cycles, while the structure did not change along the path. We concluded that the flexibility of the loop backbone atoms partially compensated for the difference in the dihedral angle values. Also, the lack of sampling along the path suggests that dihedral angle RVs are not good choices for protein folding conformational transition. This is also an example showing that only monitoring the progress of the reduced variables might not be sufficient for RV-based path sampling methods.

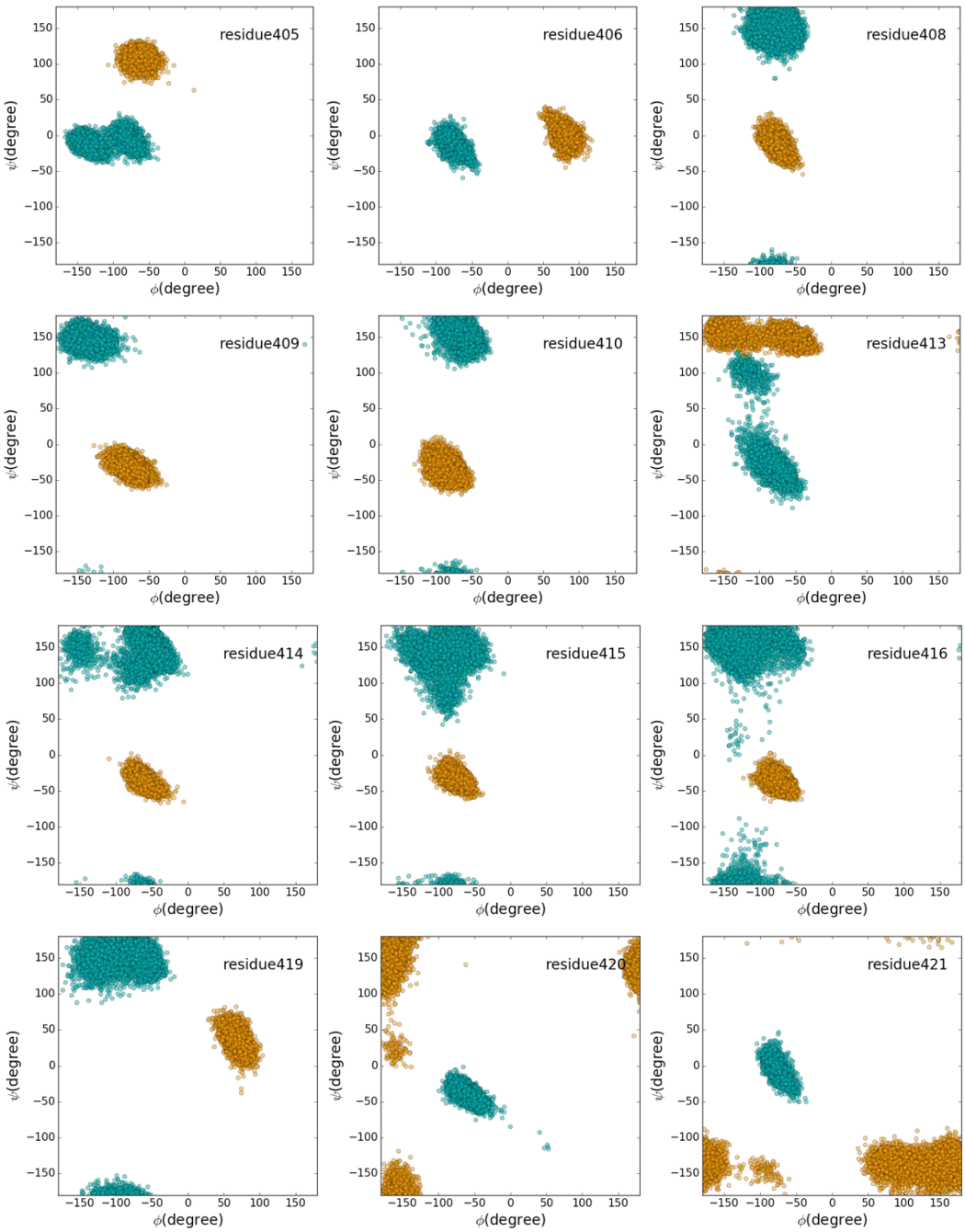


Figure 3.1 The ϕ-ψ distributions for Src 404-424 A loop region from 10 ns equilibrium simulations. Each dot represents a frame in the simulation trajectories. Most of the residues have non-overlapping distributions as shown in the figure. Cyan: the active form. Orange: the inactive form.

Based on these observations, we explored distance based RVs for the loop transition, along with the RVs for the αC helix rotation. A group of 11 RVs were chosen to calculate the transition path. The procedure and criteria of selecting the 11 RVs are detailed in methods. The evolvement of the 11 RVs from the first cycle to the last cycle is shown in Figure 3.2. The RVs converged within 70 cycles as shown in the figure.

RV1 did not traverse from 0 to 1 directly as shown in Figure 3.2A, which can be explained by the nature of the movement described by RV1. RV1 describes the distance change between the N-linker and the αC helix. In Figure 3.2A the normalized RV value first increases to ~1.5 then moves back to 0; it fluctuates in the 0-1range twice and finally reaches 1. This RV profile shows that the N-linker is very flexible. A further examination of the structure shows that the N-linker needs to move away to free αC helix and allow αC helix rotation. The two end states for the N-linker can be flexible, for the two ends are not fixed in the simulation.

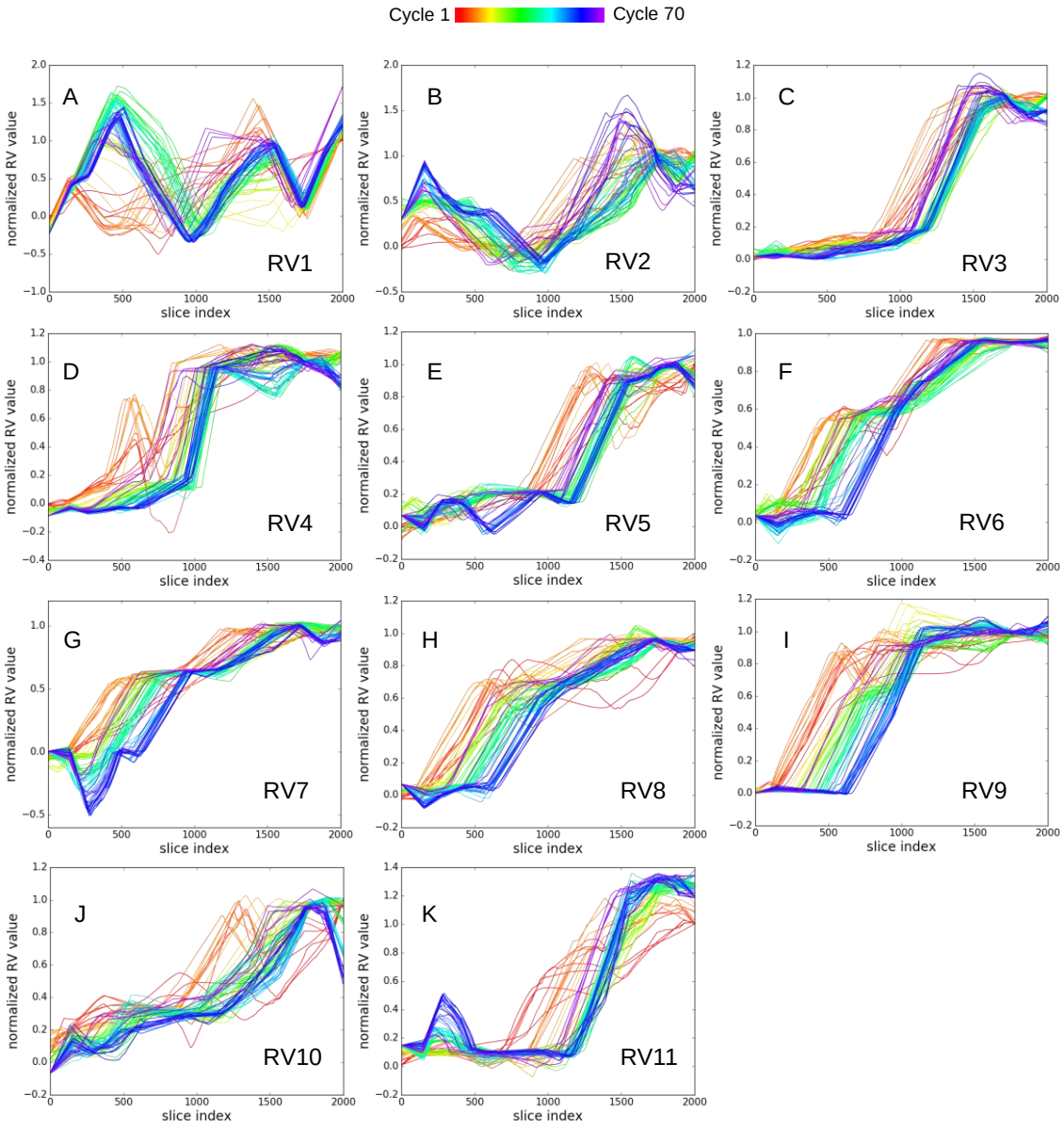


Figure 3.2 Evolution of the normalized value for the 11 RVs show convergence.

3.3.2 Transition path analysis

The transition trajectory

We built a transition path structure series using the trajectories from the last cycle. A frame for each slice was used to construct a structure series showing the transition from the active to the inactive form of Src kinase. The structure series shows atomistic details of the transition events along the transition path. The representative structures are shown in

Figure 3.3. From the active state to the down-regulated state, Tyr 416 first switches among several states with the movement of A-loop to break the interaction with Arg 409. Both Arg 409 and Tyr 416 moves towards the center of the kinase, with Arg 409 in the middle of the partially formed 1st helix and Tyr 416 with the following residues forming the 2nd helix on the A-loop. Arg 409 is freed with the interaction with Tyr 416 and moves up towards αC helix to complete the formation of the 1st helix. αC helix rotates out, and Glu 310 breaks interaction with Lys 295 and forms new interaction with Arg 409; this rotation of αC helix locks the conformation of the 1st helix on the A-loop. In the mean time, with the αC helix rotation, the freed Tyr416 moves further towards the center of the kinase to complete formation of the 2nd helix structures with other residues on the A-loop. The 2nd helix fluctuates and Tyr 416 moves closer to Asp 386 to form new interactions.

αC helix rotation is the main free energy barrier along the transition path

αC helix has been identified as a switch in the Src/CDK-like kinase domain conformational transitions in coarse-grained models[79][125]. The role of αC helix regulation in the human kinome is also suggested by structural analysis[79]. In our simulations, αC helix rotation has been constantly observed as the main free energy barrier along the path.

The potential of mean force (PMF) (Figure 3.4) is computed from the combined histogram of sampling of cycle 70 using equation 1.2. The PMF profile and visualizing structures along the transition path reveals that the αC helix rotation corresponds to the highest point on the PMF profile. Interestingly, we experimented different combinations and numbers of RVs, and the αC helix rotation constantly shows up as the highest energy barrier on the path, while the Aloop transition did not complete with those RVs. These results show that αC helix has an important role in regulating Src conformational activation.

The movement of αC helix rotation is presented as a switch completed in a short time in our simulation. This is concluded from the scarcity of structures obtained from the

simulation before and after the helix rotation. Also, the time profiles for Lys295-Glu310 and Glu310-Arg409 from long equilibrium simulations reported before also showed the switch behavior[126]. However, the conformational change of Aloop was not reported in that work.

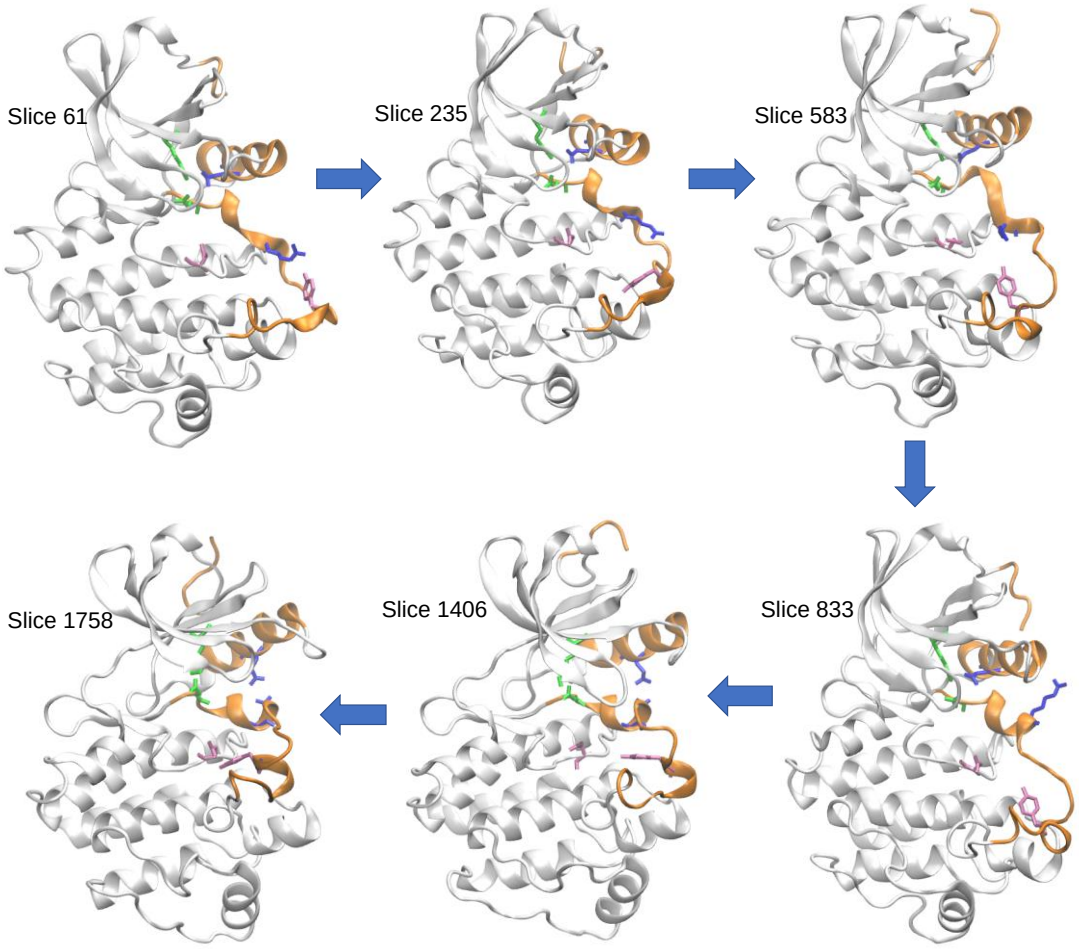


Figure 3.3 The representative structures along the transition path. The αC helix and the Aloop region are in orange. The three residue pairs, 404-295, 310-409 and 386-416 are in green, blue and mauve respectively with stick representation.

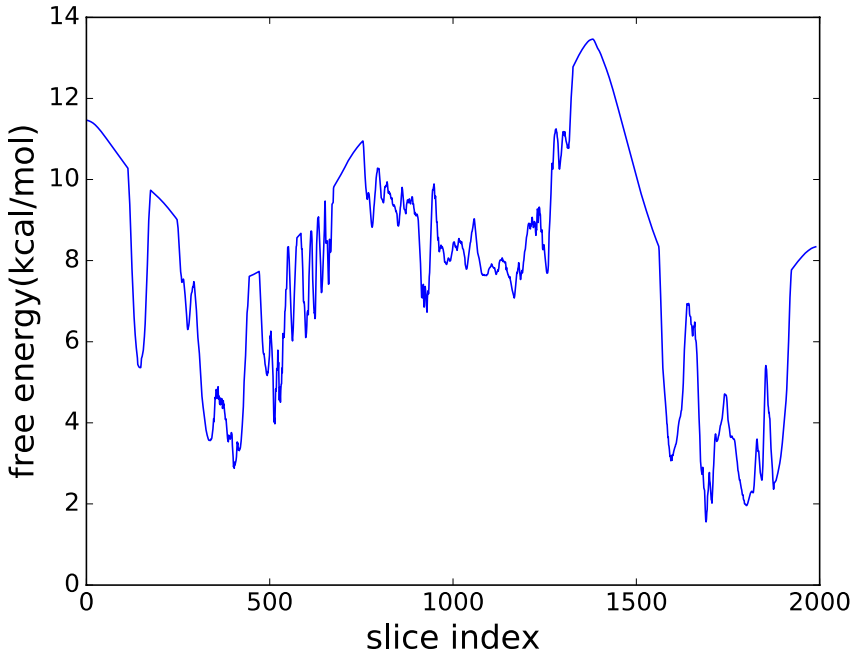


Figure 3.4 The PMF profile for Src conformational transition from active (left) to inactive (right) state. The highest point corresponds to the αC helix rotation.

The correlated motions of RV3 and RV12

The correlated motions of residue groups in Src kinase domain have been observed in previous computational simulations[70][73]. The αC helix movement was observed to be coupled with conformational changes in the C-lobe centered on Lys 427 and Trp428. The dynamic coupling between Aloop and the residues in the electrostatic network, R-spine and DFG motif was also observed.

In our ABPO simulation, the motion of two reduced variables are highly correlated as shown in the normalized RV profile. In the converged curve, both RV3 and RV12 stayed relatively invariant till ~1300 slices, then increased to the final value. RV3 describes the αC helix rotation with respect to Nlobe and Aloop, including the electrostatic network. R12 describes the relative movement of Aloop and Clobe. The correlated motion of the two reduced variables indicates that the two main events, αC helix rotation and Aloop

extension, in the Src conformational activation are highly concerted and not independent. Also, it shows the concerted motions of residues spanning both lobes in the Src kinase domain. The coupling of residues more than 20 Å apart reveals the existence of a long-range allosteric network in the Src kinase domain.

3.3.3 The function of the linker residues

Previous experimental[127-130] and computational studies[77-78] suggest that the SH2-kinase domain linker plays an important role in regulating kinase activation. Leu 255 has suggested to be a critical component of the SFKs intramolecular inhibition mechanism[129]. The W260A studies suggest that the conserved Trp 260 has a critical role in coupling the regulatory domains and the catalytic domain[128].

The preparation steps we used have little effect on the linker positions comparing the crystal structure and the close-to-average structure from the equilibrium simulations (Figure 3.5). In the active crystal structure, residue 255 is close to residues 304 and 305, and residue 260 is stacking against residue 315. In the active ABPO starting structure, residues 255, 256 are interacting with the middle part of αC helix while residue 260 stays in the same position as in the crystal structure. In the inactive crystal structure, residue 260 is in between the αC helix and a β sheet, while residues 255-259 do not interact with the helix. In the inactive ABPO starting structure, the residues stay in the same positions as they are in the crystal structure.

To test the importance of the linker residues, we compared the transition paths with and without the linker. In the first construct, the linker residues 255-259 were not included in the two end structures for simulation. We found that when the linker is not included, the αC helix rotation may occur independently and earlier in the inactivation process, without correlating with the Aloop transitions. This can be visualized as the position of the highest peak shift on the PMF profile.

To further evaluate the function of the linker, we deleted the first RV from the 11 RV groups and ran ABPO using 10 RVs. In this simulation, the conformations are not favored in the direction of moving the linker away from the αC helix with the bias potential. We observed that if the linker did not move in the free sampling, the transition would not happen featured by a high free energy peak along the PMF.

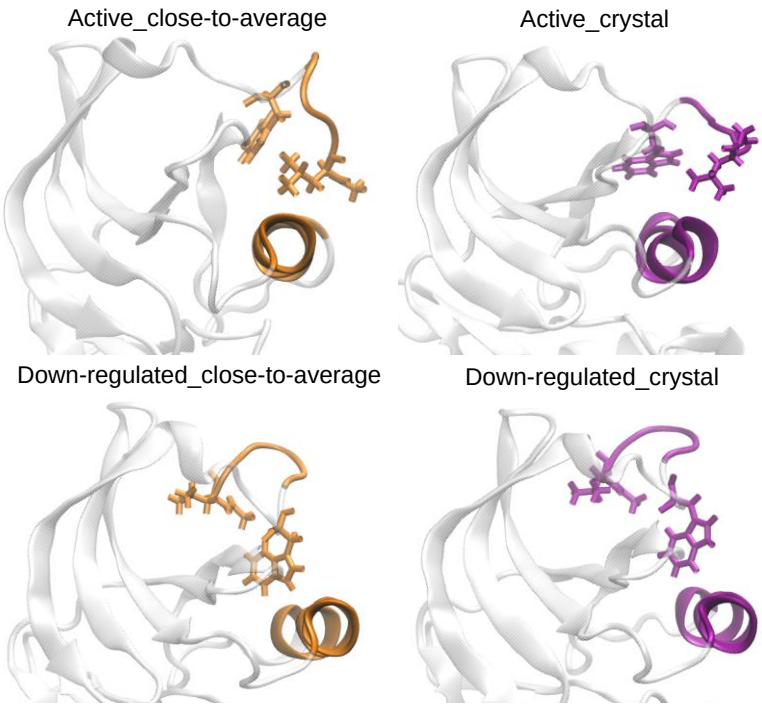


Figure 3.5 The linker position in the active form and the down-regulated form showing its interaction with the αC helix. Leu 255 and Trp 260 are in stick representation. The preprocess steps for ABPO do not alter the linker-αC helix interactions.

3.3.4 The conformational flexibility of the A-loop in the all-atom ABPO simulations

Large conformational variations of A-loop have been observed in our all-atom simulations. Before the αC helix rotation, the A-loop undergoes a series of conformational changes and gradually forms the 1st helix and partially the 2nd helix. After the αC helix rotation, the 2nd helix continues to form and moves to its position in the down-regulated conformation. Comparing the all-atom deactivation transition path we built using ABPO and the Gō potential model path[79], the αC helix remains as the switch for the conformational transition. However, there is a sharp decrease in the PMF profile in the Gō

potential model, while in our PMF profile, the highest peak corresponds to ~0.7 of the total path length, indicating there are structural changes after αC helix rotation. In the Gō potential model, the alpha-C helix rotation locks the two-helical structure, and no obvious structural changes was observed after αC helix rotation[79]. In our model, the 2nd helix on the A-loop continues to form and move to its final position. This observed difference is probably due to the lack of the 2nd helical structure in the Gō model. Gō model uses Cα atoms to present residues and are missing interactions between side chains and associated structural differences that are reflected on secondary structures. The 2nd half of A-loop in the Gō model showed high flexibility and were excluded when defining reduced variables for the transition. Also, we have a higher resolution of sampling along the path in ABPO, and more structures along the path were retrieved. The Gō-model only observes 40 path images. In our simulation, we divided the path into 2000 slices. While a large number of slices would cause some noise in the PMF, the higher resolution of sampling does help with visualizing structures along the path.

The distance combination reduced variables successfully promote the formation of two helices from coil. RVs 4-6 describe the loop formation, with residue pairs 3 or 4 residues apart (Table 3.2). The standard α-helix in proteins contains 3.6 residues per turn. In this sense, our reduced variables specify the largest distance changes in residue pairs that at least interact in one state of the protein. The formation of helix is promoted with adaptive biasing potential with this type of reduced variables.

3.3.5 Electrostatic network analysis

The atomic details of the conformational inactivation can be described in terms of an electrostatic network[50][51]. The residues that are involved in the electrostatic network are important for catalysis or regulation of the kinase. To analyze how the electrostatic network evolves during the inactivation process, the bins of distance distributions of electrostatic network residue pairs are visualized as described here. After the path optimization is completed, for each frame in all the trajectories in the last cycle, the inter-residue distances for the 5 residue pairs and the corresponding slice index for that frame was extracted. The whole path is divided into 10 bins by slice index, for example, bin 1 covers slices 1-200, bin 2 covers slices 201-400. The inter-residue distances for each frame

are assigned to the bins per the slice index associated with that frame. The inter-residue distances were then plotted for each bin for each pair.

The analysis of the electrostatic network identifies three residues pairs as important to the Src inactivation process. The residue pair distances and their distributions in each bin are shown in

Figure 3.6. The 310-409 and 295-310 pairs have the large distance change around the same time, further proving the αC helix rotation as the switch for the conformational transition. For the two residue pairs 386-416 and 409-416, the distance distributions showed some trend of distance change with the increase of bin numbers that is not as obvious as stable switches. The behavior of residue pairs 386-416 and 409-416 are partially due to the flexibility of the Aloop and can be explained in terms of the transition structures. Asp 386 is a residue in the HRD motif on the C-lobe of the kinase domain, and Tyr 416 is a residue on the A-loop. The two residues are not interacting with each other in the active form of the kinase. However, Tyr 416 might move closer to Asp 386 in space due to the flexibility of the A-loop, while the two-helical structure does not form completely with the loop movement. After the two helices on the A-loop form, Tyr 416 is part of the 2nd helix. The 386-416 distance might still change, for the position of the 2nd helix is flexible for the short linkers on both ends of the short helix, while the helical structure remain intact. Similarly, for residue pair 409-416, the distance between the two residues increases in bin 2 and 3 due to the temporary breaking of interaction, and the A-loop is still in extended loop conformation in this stage. Later, the 409-416 distances increases in bin 8 and remains stable till the end of the path, for the 2nd helical structure has formed, and it is not structurally viable for Tyr 416 to go back to interact with Arg 409 without breaking the helical structure. While the two helices are relatively stable after formation, the position of the 2nd helix, is observed as flexible after αC helix switch. The 404-295 distance is mostly stable as observed in our simulations. The observed distance change in a previous work might be due to the choice of different crystal structures.

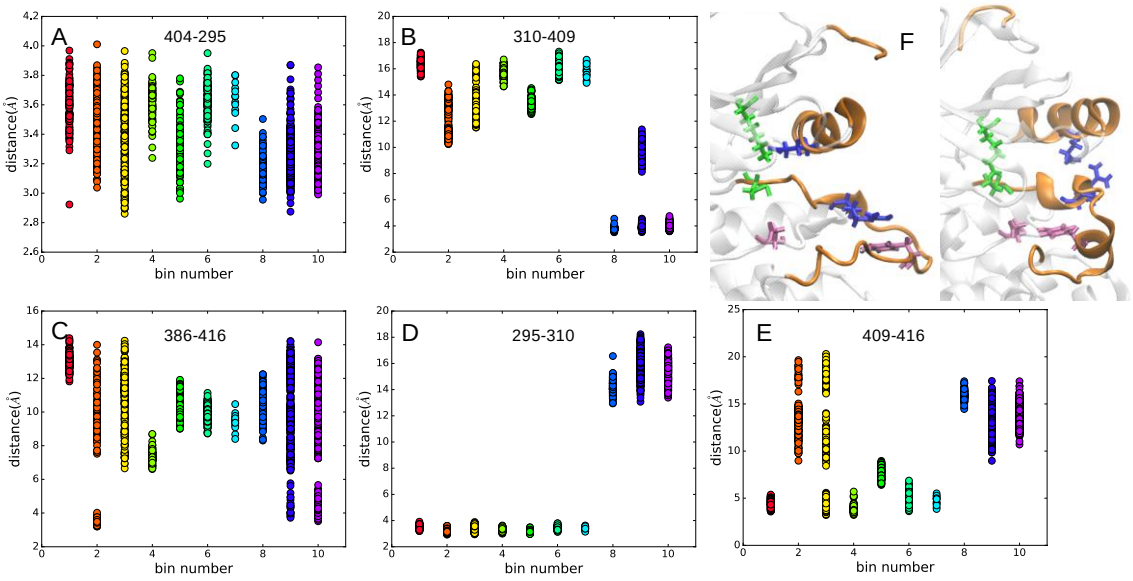


Figure 3.6 The distance distributions for the residue pairs in the electrostatic network. A-E are showing the distance distributions for the 5 pairs, F is showing three pairs in stick representation in both active and down-regulated structure. Green: 404-295. Blue: 310-409. Pink: 386-416.

3.4 Conclusions

3.4.1 Structural and methodological insights

Src transition pathway

We examined the Src conformational transition and analyzed the features during the progress. The αC helix rotation is established as the main free energy barrier along the transition path in the all-atom model. Also, the large flexibility of the Aloop structure was observed. Specifically, from the active to the inactive form, the Aloop first forms the two helical-like structure and moves towards the middle part of the protein, then the αC helix rotates to lock the Aloop in place, and the 2nd helix on the Aloop continues to form and fluctuates gradually to its final position. This process reveals a series of intermediate states which will assist drug design efforts targeting Src kinase domain.

The concerted motions of the reduced variables are observed as shown on the RV profiles. The concerted motions reveal key residues involved in the transition process across the kinase domain, which would help with molecule designs to regulate the kinase activity.

The all-atom transition path is compared with the previous simulations in our group. Specifically, the difference observed in the ABPO simulation and MFTP Gō model is discussed and explained in terms of structures and the reduced variables used. The electrostatic network was analyzed and the distance changes are inspected in details.

ABPO on large-scale conformational transitions

Methodology insights are acquired from building the Src conformational transition pathway. In our project, we experimented several types of reduced variables, including single inter-residue distances, linear combination of inter-residue distances, and dihedral angles. It is demonstrated that while the protein folding features dihedral angle value changes in the protein backbone atoms, the dihedral angles might not be good reduced variables to describe the transition, for the inter-residue interaction information is not included. Distance combination reduced variables would be good choices for large conformational changes. For Src kinase domain, we incorporated both protein structural information and preliminary ABPO calculation to get the final set of reduced variables. In addition, the choice of simulation parameters is gaining importance in complicated transitions.

3.4.2 Advancement in the field

This is the first example using all-atom ABPO to optimize a large-scale conformational transition involving partial protein folding. Also, this is one of the few cases where the continuous, unrestrained conformational transition in an all-atom protein system is observed.

The proper choice of reduced variables remains a main challenge for such calculations. Here we demonstrated that linear distance combination reduced variables defined from a combination of algorithmic approach and system structural information would work properly for large-scale all-atom conformational transitions.

Our results answer several questions in Src conformational transition. The order of the two main events during the transition, A-loop folding and αC helix rotation, was explained in details. Also, the linker residues are shown to be important in interacting with and regulating the kinase domain. Further, we relate our results with previous studies and provide a comprehensive explanation to the observed simulation results.

CHAPTER 4. SRC-SSP COMPLEX STABILITY IN IMPLICIT AND EXPLICIT SOLVENT

4.1 Introduction

How protein kinase interacts with its substrate is important to understanding the enzymatic activity and structure-based drug designs. Src phosphorylates its substrate by transferring a phosphate group from ATP to the substrate that binds to the Src kinase domain. Due to the disease relevance of SFKs misregulations, ATP analogs and some allosteric inhibitors have been designed to modulate Src activity. However, the substrate-binding site is relatively unexplored. In serine/threonine kinases, the substrate binds to the cleft between the two lobes on top of the activation loop[131-135], while in tyrosine kinases, the substrate binds to the C-lobe below the activation loop[136-139]. To our knowledge, there is no crystal structure of Src in complex with a peptide/protein substrate to directly show Src-substrate interactions.

Several mutation studies of Src identify some residues that are associated with substrate binding, while a specific site is not clearly identified. The R385A mutant of cSrc has low kinase activity suggests R385 is required to stabilize substrate binding[140]. The D404N mutation that mimics D404 protonation at the DFG motif promotes substrate-peptide binding[126]. Besides, NMR studies show allosteric communication across the whole kinase domain[123] upon substrate binding. The allosteric network within Src kinase domain complicates the interpretation of the mutagenesis studies, for the effect of the mutations might be due to long-range effect rather than direct change of the substrate binding site.

Previous experimental results reported in a manuscript from a former group member Mehul Joshi suggest a change in substrate recognition within the TK group, that Src substrate might bind to the cleft between Nlobe and Clobe. NMR chemical shift perturbation (CSP) experiment indicates large-scale conformational motions upon SSP binding, involving both N and C lobes, and the paramagnetic relaxation enhancement (PRE) showed that multiple SSP bound forms might exist. Finally, residue L407 in the cleft with a strong PRE was studied. The enzymatic activity of a mutant L407D was determined to be 33-fold lower than the wild-type Src, supporting a cleft binding mode. Also, an earlier

study in our group by Beverly Gaul et al of activated SFK Lyn kinase recognition of immunoreceptor tyrosine-based activation motif (ITAM) substrate using both NMR and molecular docking suggests that ITAM binding in an orientation of the cleft-binding mode is strongly favored[141].

The Src substrate peptide we use here is a 11-residue peptide with sequence AEEEIYGEFEA (named SSP here). The peptide has a K_M = 300 μM, and was identified from peptide library experiments [142].

The ensemble MD with PRE restraints obtained from NMR experiments was used to identify possible SSP-bound state orientations. The frames from the PRE-restrained simulations were clustered according to SSP backbone RMSD values, and the clusters have either cleft or clobe binding mode to satisfy the restraints. Two models were built from ensemble MD. Model 1 gives 9 clusters, with clusters 1-5 and 9 in cleft mode, and clusters 6-8 in C-lobe mode. The clusters in cleft mode is further divided into two groups cleft-A and cleft-B. For cleft-A mode, the N-lobe and cleft PREs are satisfied, while in cleft-B mode, the cleft and C-lobe PREs are satisfied. Among the 6 clusters in cleft mode, only cluster 2 is in cleft-B mode. For the C-lobe mode, only the C-lobe PRE is satisfied. An illustration showing the three binding modes using cluster 2, 4 and 6 is in Figure 4.1. The tyrosine residue on the peptide is positioned in between Asp 386 and Arg 388 for catalysis. Model 2 yields 8 clusters, with clusters 1, 2 and 5 in cleft binding mode and the other five clusters in C-lobe binding mode. It is worth noting that the more number of clusters are in C-lobe mode does not indicate the mode has a higher probability in simulations. The three cleft-mode clusters contain more total number of frames. Here we use unbiased MD simulations to understand the dynamics and stability of Src-SSP complexes, and further explore the possible binding modes.

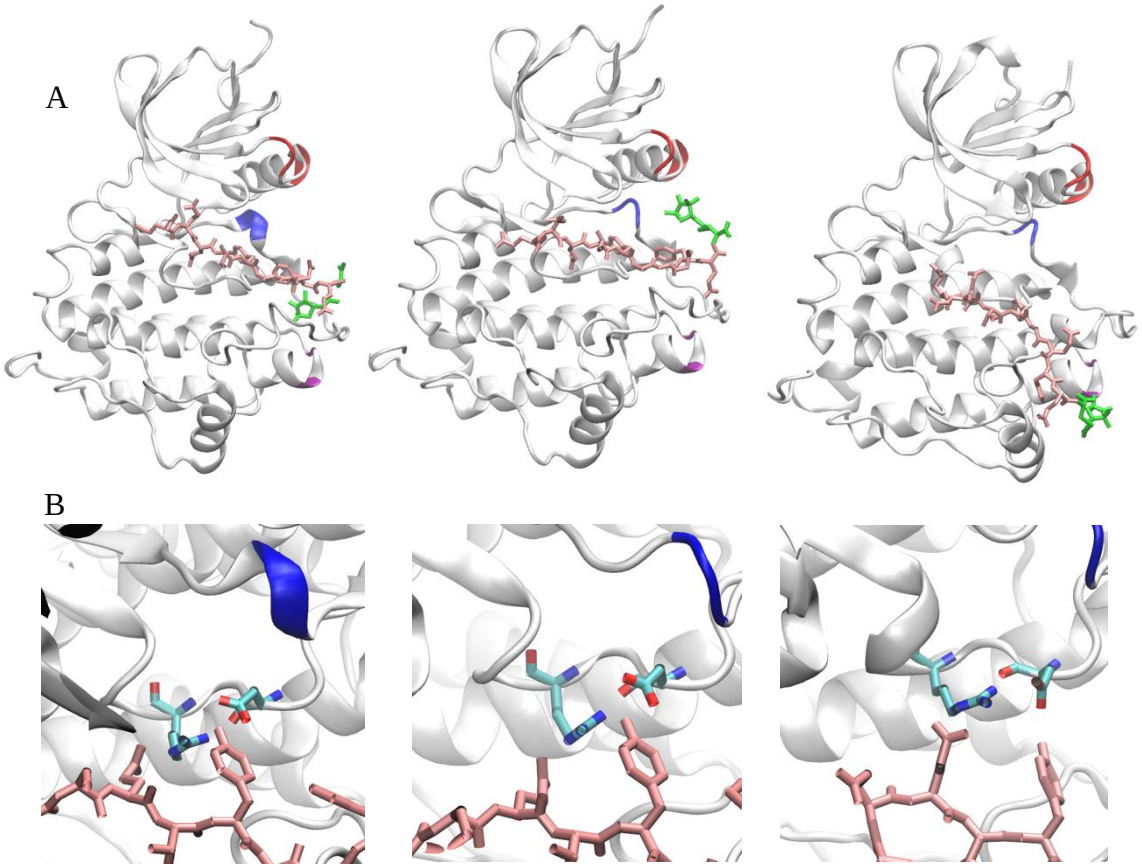


Figure 4.1 The cluster 2, 4, 6 average structure to show cleft-B mode, cleft-A mode and C-lobe binding mode. Red: N-lobe PRE. Blue: cleft PRE. Purple: C-lobe PRE. Pink: SSP except CYP. Green: CYP label. A: the N-lobe PRE residues 300-301, 305-306 are colored in red, the cleft PRE residues 406-407 are colored in blue, and the C-lobe PRE residues 436 and 438 are colored in purple. The peptide is in stick representation and colored pink, except that CYP is in green. B: a zoom in view to show the tyrosine on the peptide in between Asp 386 and Arg 388 for catalysis.

4.2 Methods

4.2.1 System setup in explicit solvent

For Model 1, three groups of systems were set up. The SSP peptide in the simulation only contains 10 residues, without the 1st ALA as shown in the 11-residue sequence. The first group has the peptide with SSP label kept in the model, the 2nd group has ATP and magnesium modeled from Src-ATPgS complex (PDB ID 3DQW[124]); the two protein structures were aligned using PyMOL, and the new saved coordinates for ATP and Mg was added to the Src-SSP complex. The third group has a ALA in place of the CYP label. The ALA was generated that atoms coordinates for the CYP are deleted except the common heavy atoms with ALA.

In addition to the Src-SSP models, two protein kinase crystal structures, tyrosine kinase insulin-like growth factor 1 (IGF1) receptor (PDB ID 1K3A[136]) and a serine/threonine kinase, the catalytic subunit of cAMP-dependent protein kinase (PKA, PDB ID 1ATP[143]) in complex with peptide were also included to explore the stability of protein-peptide complex.

PKA has a 20-residue inhibitor peptide (residues 5-24, chain I in PDB file). The peptide was truncated to start from residue 14 in the original PDB numbering to match the length of SSP label, for a total length of 11 residues. The peptide is in cleft binding mode, consistent with the serine/threonine kinases substrate binding paradigm.

IGF1 receptor has a 14-residue insulin receptor substrate peptide (residue 6-13, chain B in PDB file). The whole peptide was kept in the model. The peptide is in the C-lobe binding mode, consistent with the tyrosine kinase substrate binding paradigm. The loop residues 1069-1076 were missing in the PDB file. The loop was modeled from phosphorylated insulin receptor tyrosine kinase (PDB ID 1IR3[144]) by aligning insulin receptor to IGF1R and save the new coordinates using PyMol. The coordinates of the loop heavy atoms were copied to the missing region of IGF1R. Three residues 1073-1075 are different between the two structures, Gly, Arg and Pro in insulin receptor and Val, Leu, Ala in IGF1 receptor. For the three residues, the atoms except common heavy atoms were built using CHARMM.

The active form of Src kinase domain (PDB ID 1y57) was also included, to compare the protein backbone RMSD values of the crystal structure and the modeled structures on GPU.

For Model 2, CYP label was replaced with ALA, and ATP and magnesium were not included in the system. A full list of the simulations for the two Src-SSP models carried out in explicit solvent is in Table 4.1.

Hydrogens were added to the structures using CHARMM HBUILD. For the crystal structures, the protonation states of the His residues in CHARMM forcefield (HSD, HSE, HSP) were assigned manually by visualizing and deciding the state with the most favorable interaction. For the three clus2, clus4 and clus6 with SSP models, cubic TIP3 water boxes were added that each edge of the protein is at least 12Å to the edge of the box in all three directions. For IGF1R and PKA, the distance was shortened to 10Å to save simulation time. For all the other systems, box length was adjusted to have similar number of water molecules in each system, while the 10Å minimum distance to the edge is maintained.

4.2.2 System setup in implicit solvent

The SSP from cluster 2 was modeled to the active form of Src (PDB ID 1Y57) to explore the possible orientations of the C terminal of the peptide. Two SSP conformations, cleft-down and cleft-up, was observed. Only residues 1-8 was included for the peptide to allow the C terminal of the peptide to move freely to either the cleft-down or up state. The Src structure is taken from Src_e implicit solvent simulation in Chapter 3. Cluster 2 was aligned to Src using PyMOL, then the coordinates of the SSP was taken from the aligned cluster 2 to Src coordinates.

The 9 cluster averages from Model 1 and 8 cluster averages from Model 2 was put in implicit solvent for unbiased simulations. For model 1, the CYP label was kept for each structure. The energy-minimized average structures of the 9 clusters were used for NVE heat up and equilibration, then MD simulations. For model 2, ALA was used to replace the CYP label. A full list of the simulations for the two Src-SSP models in implicit solvent is in Table 4.2. The implicit solvent model FACTS was used for the simulations.

Table 4.1 Src-SSP simulations for the two models, and IGF1R and PKA in explicit solvent.

NO.	System	Number of waters	Volume (cubic box edge length in MD, Å)	Time(ns)
	Model 1			
1	Clus2Cleft-down	29312	96.96550	600
2	Clus4Cleft-up	27031	94.50081	600
3	Clus6C-lobe	20480	86.69860	600
4	IGF1R(1K3A)	15135	79.02618	300+100*3[1]
5	PKA(1ATP)	19150	85.29321	300+100*3
6	Src (1y57)	16433	80.77988	300
7	Clus2+ATP+Mg	25579	92.90044	300
8	Clus4+ATP+Mg	25367	92.65180	300
9	Clus6+ATP+Mg	24075	90.76119	300
10	Clus2(ALA)	25591	92.96788	100*3
11	Clus4(ALA)	25492	92.73978	100*3
12	Clus6(ALA)	23763	90.74478	100*3
	Model 2			
13	Clus1(ALA)	20519	86.64673	300
14	Clus2(ALA)	20317	86.36997	100
15	Clus4(ALA)	19220	84.96075	100
16	Clus6(ALA)	18875	84.42043	100
17	Clus3(ALA)	18967	84.55363	100
18	Clus5(ALA)	22428	89.17764	100
19	Clus7(ALA)	20585	86.75950	100
20	Clus8(ALA)	19614	85.46293	100

[1] +: independent trajectories were started from the equilibrated structure

[2] *3: 3 independent trajectories were started with different velocities

Table 4.2 Src-SSP simulations for the two models in implicit solvent

NO.	System	MD production run time (ns)
	Model 1	
1	1y57+peptide	11
2	Clus1	100
3	Clus2	100
4	Clus3	100
5	Clus4	100
6	Clus5	100
7	Clus6	100
8	Clus7	100
9	Clus8	100
10	Clus9	100
	Model 2	
11	Clus1(ALA)	100
12	Clus2(ALA)	100
13	Clus4(ALA)	100
14	Clus6(ALA)	100
15	Clus3(ALA)	100
16	Clus5(ALA)	100
17	Clus7(ALA)	100
18	Clus8(ALA)	100

4.2.3 Simulation details in explicit solvent

0.15M NaCl ions was added to the solvated system. The simulations were carried out using CHARMM version c40b1. For the systems with CYP label, CHARMM22 all-atom force field with CMAP dihedral angle correction was used. The parameters for the CYP label was added to the force field. For the systems without CYP label, CHARMM36 all-atom force field was used. An energy minimization was done NVE heat up and equilibration was finished on the cluster Halstead using CHARMM DOMDEC.

The energy was minimized using the steepest descent and Powell algorithms to a gradient less than 1.0 in the following stages: 1) with the position of protein heavy atoms fixed, 2) with harmonic restraints on protein heavy atoms, 3) with harmonic restraints on protein backbone (N, C, C_α) atoms, and 4) without restraints. The steepest descent algorithm was used for the first three stages and the Powell algorithm is used for the final stage.

The energy-minimized structures were heated from 100 K to 298 K and equilibrated at 298 K over a total period of 500 ps in NVE ensemble. The initial velocities were generated from Gaussian distributions at the specified temperature. The leapfrog integrator was used to calculate the trajectories with a 2 fs time step. The NVE step was performed on the community cluster computation nodes.

The steps following NVE was finished on GPU NVIDIA GENForce GTX 1080. A NPT equilibration for 1ns was restarted from the end of NVE equilibration. Langevin dynamics was used with a temperature of 298K, and the long-range interactions cutoff distances were set to 8, 10 and 12Å. The pressure was controlled using MC barostat (keyword PRMC), which is a Monte Carlo barostat that uses trial volume changes with a Metropolis-based acceptance criteria, with a reference pressure of 1 Pa.

The production runs in NVT ensemble were restarted from NPT simulations on GPU. Langevin dynamics was used to calculate the trajectories at 298K. Coordinates were saved every 20 picosecond. The time series of temperature and potential energy were monitored to assess the simulations were stable.

4.2.4 Simulation details in implicit solvent

For the active Src-peptide model, the system was first energy minimized following the steps as described for the explicit solvent systems. The energy minimized structures were put in NVE ensemble to heat up to 298K. Then production runs in NVT ensemble at 298K were restarted from the NVE equilibrated structures. Lagevin dynamics was used to calculate the trajectories, and the long-range interactions cutoff distances were set to 10, 12 and 14 Å. Coordinates were saved every 2ps.

A model was built to allow the C-terminus of the peptide to explore possible orientations that would satisfy the PREs. The peptide in Model 1 clus2 was modeled to the active form of Src (1Y57). Only the first 8 residues of the peptide were kept in the simulation to allow the C-terminus to explore possible orientations. The NOE restraints in CHARMM were used to keep the Tyr interactions with Asp 386 and Arg 388. To keep the protein fixed and allow the peptide C-terminus to explore the orientations, CHARMM tpcontrol was used that protein has a temperature of 100K and tau equals 10000, and the peptide has a temperature of 500K and tau=0.1. An vv2 integrator was used for the MD simulations for Langevin is not compatible with tpcontrol. The simulation was carried out for 11 ns, with a 2 fs timestep.

All implicit solvent simulations were finished on the Halstead cluster nodes.

4.2.5 Trajectory analysis

The trajectories were aligned with respect to the corresponding energy minimized structure. For Src, the backbone atoms of the C-lobe excluding A loop, residue 342-403 and 425-521, were selected for alignment. For IGF1R and PKA, the backbone atoms (type C, N and CA) of the protein were selected for alignment. Water molecules were removed for the explicit solvent systems.

For each system, the RMSDs of protein and peptide were calculated from the aligned trajectory using the correl rms command in CHARMM, and only backbone atoms (type C, N and CA) were included for RMSD calculations; the timeseries of the interaction energy between the protein and the peptide were calculated by looping through the frames in the trajectory and using the inte command in CHARMM.

4.3 Results and Discussion

4.3.1 The two crystal structures in explicit solvent (Table 4.1 NO. 4-5)

The two protein-peptide complex crystal structures were used as a reference to study protein-peptide interactions in explicit solvent. As shown in Figure 4.2, for IGF1R, the protein backbone RMSD is less than 3Å. The peptide RMSD is around 2Å and remained stable during the 100 ns simulation time in all three cases. The interaction energy between the protein and peptide fluctuates within the -400 to -200 kcal/mol range without significant trend in a change of direction. Similarly, the PKA protein-peptide complex shows a protein backbone RMSD around 2Å, a peptide backbone RMSD about 4Å. The interaction energy is lower and fluctuates around -400 kcal/mol.

To explore if the complexes are stable in a longer timescale, we ran a simulation for each of the two complexes for 300 ns in explicit solvent. Similar to the results from the 100 ns trajectories, the RMSD values for both the protein and the peptide stayed stable during the 300 ns time period. The interaction energy also fluctuates in a similar range as in the 100 ns simulations, indicating the systems are stable in a longer timescale.

A closer look at the trajectories reveals some details in the protein-peptide interactions. For IGF1R, the tyrosine in the peptide only dissociated briefly 70-72 ns in one of the 100 ns trajectories and stayed in place for the rest of the simulation time; in the 300 ns trajectory, Tyr moved out ~200 ns. For PKA, the central part of the peptide stayed stable in all 100 ns trajectories and 300 ns trajectories, while the N and C termini of the peptide might flip around without the peptide dissociating. These observations confirms that the peptides are positioned well for catalysis in these two crystal structures.

In conclusion, the two crystal structure protein-peptide complexes are stable using our simulation protocol, and can be used as a reference for protein-peptide interactions in explicit solvent. IGF1R has the peptide in the C-lobe binding mode and PKA has the peptide in the cleft binding mode, providing template for how the substrate peptide interacts with the kinase.

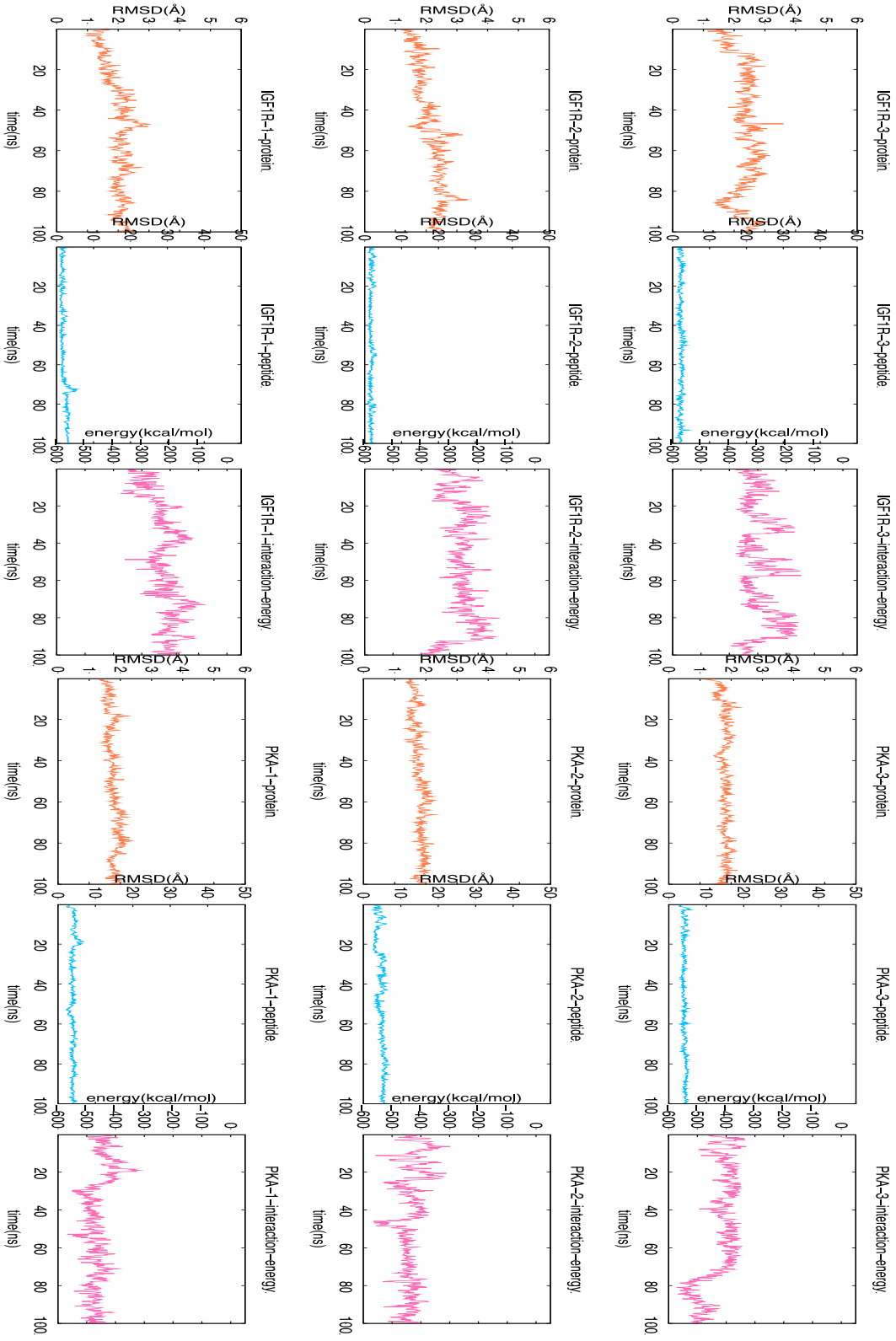


Figure 4.2 The protein backbone RMSD, peptide backbone RMSD, and protein-peptide interaction energies for the two protein-peptide crystal structure complexes. The structures are generally stable during the simulation time period.

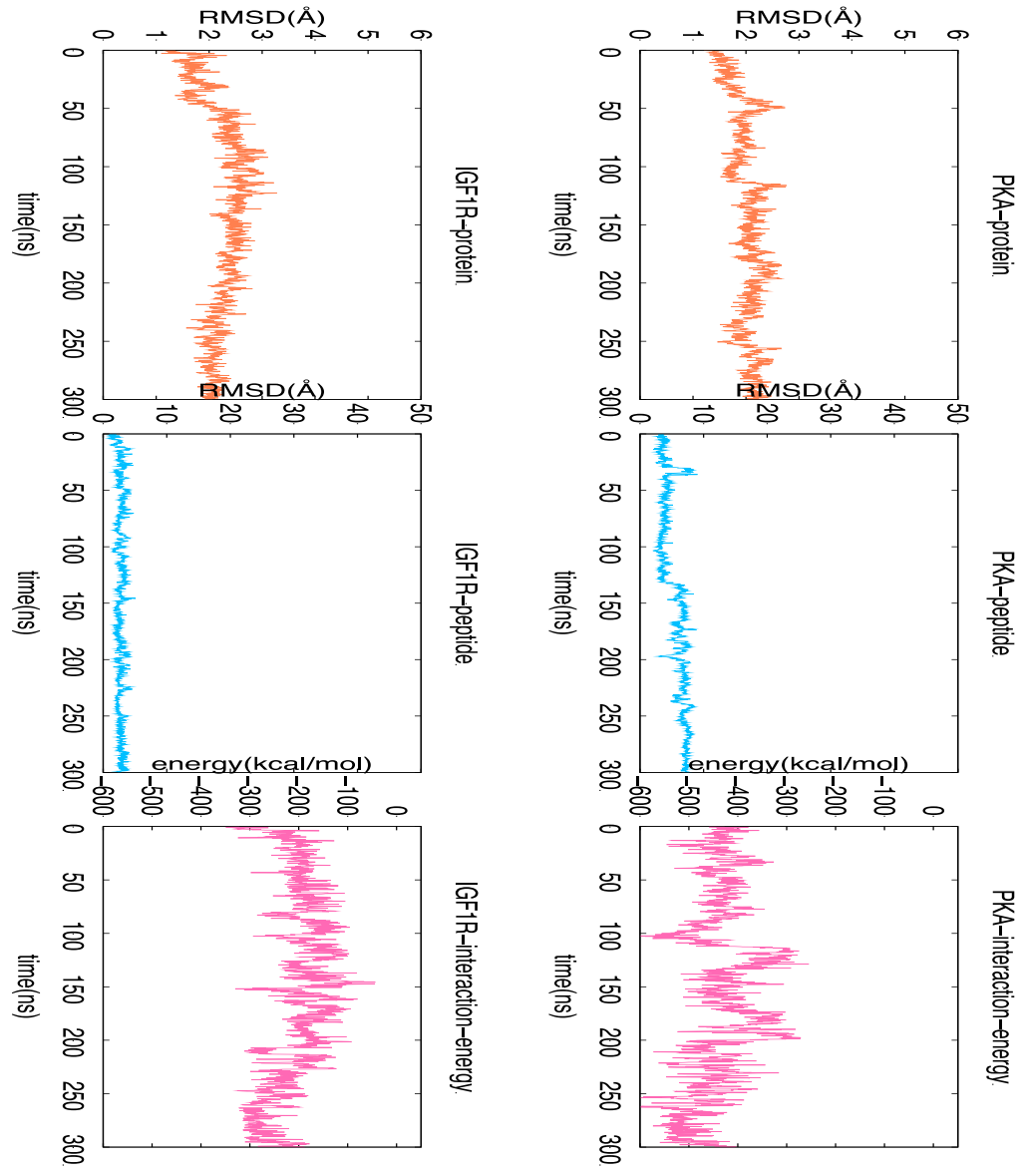


Figure 4.3 The two crystal structures in explicit solvent for 300 ns. The two structures remained stable in a longer simulation time period.

4.3.2 Model 1 in explicit solvent

4.3.2.1 With CYP label (Table 4.1 NO. 1-3)

The results for the three clusters were shown in Figure 4.4. The protein backbone RMSDs are mostly stable during the 600 ns time period. The peptide RMSD values are significantly higher, indicating the movement of the peptides. The trajectories showed that only the peptide in cluster 2 stayed in the cleft, but Tyr moved out from the active site during the MD simulation. For cluster 4 and cluster 6, Tyr moved out from the active site during the equilibration before the start of the MD simulation. Also, the peptide flipped around the binding site without forming stable interactions.

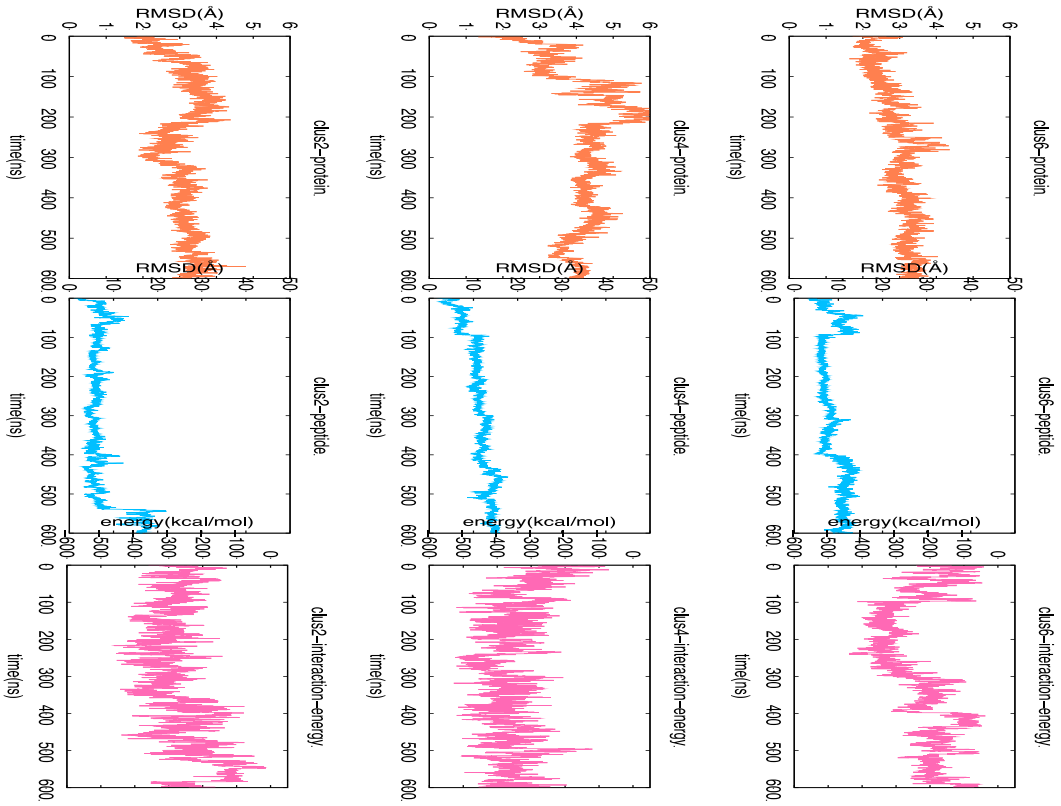


Figure 4.4 Model 1 with CYP label in explicit solvent. The protein backbone RMSD, peptide backbone RMSD and interaction energy are shown for each of three complexes.

4.3.2.2 With ALA replacing CYP (Table 4.1 NO. 10-12)

We reasoned that the instability of the protein-peptide complexes might be due to the CYP label. Therefore, we replaced the CYP label with ALA to see how the complexes behave without the built CYP forcefield. It is worth noting that without CYP label, the

cleft-A and cleft-B peptide conformations have the same orientation at the C-terminus. Three 100 ns simulations were finished for clusters 2, 4 and 6. The results are shown in Figure 4.5. Surprisingly, the simulations without CYP and with shorter time period do not yield more stable complexes. Of the 9 simulations, clus2-1, clus 4-2, and all of the three clus 6 simulations have the peptide dissociate quickly, shown as a high peptide RMSD larger than 20 Å, and the interaction energies that increased to 0 indicating complete dissociation of the peptide. Generally, the simulations do not produce stable complexes but see the peptide dissociate partially or completely in the simulations. These results show that the instability do not come from the CYP label, but the systems themselves.

Also, the model with Mg ion and ATP included in the structure does not help with ligand binding, and the results are not included here.

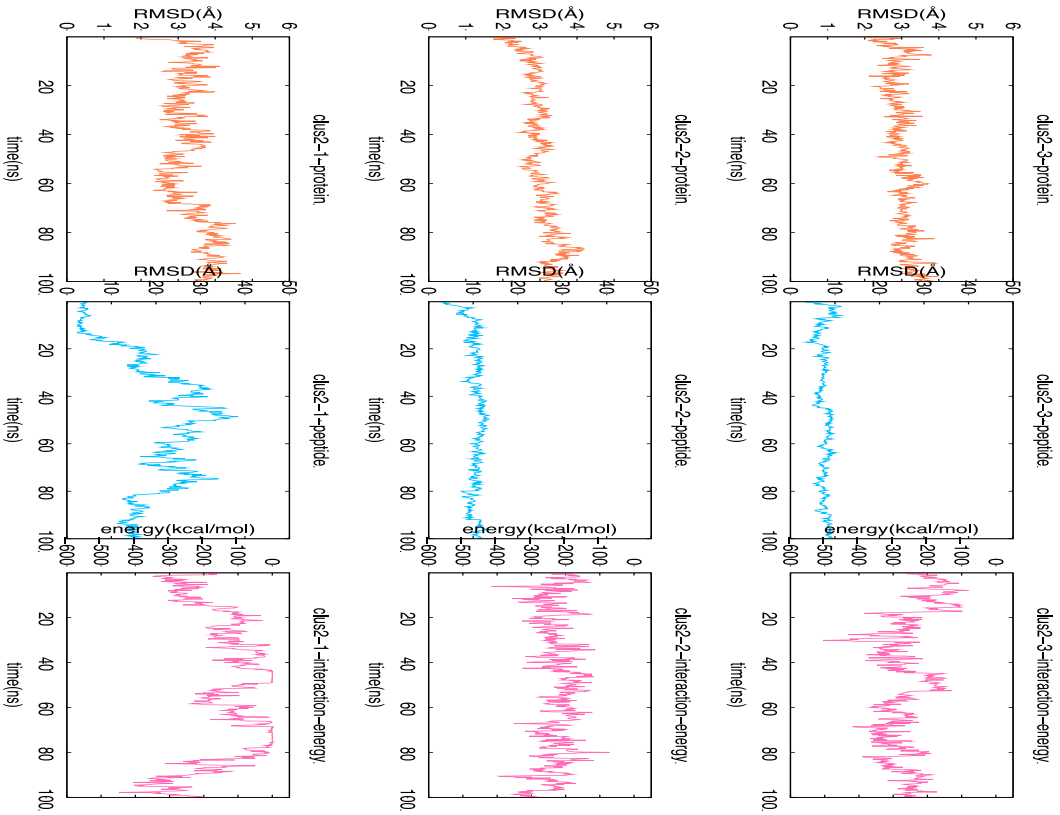
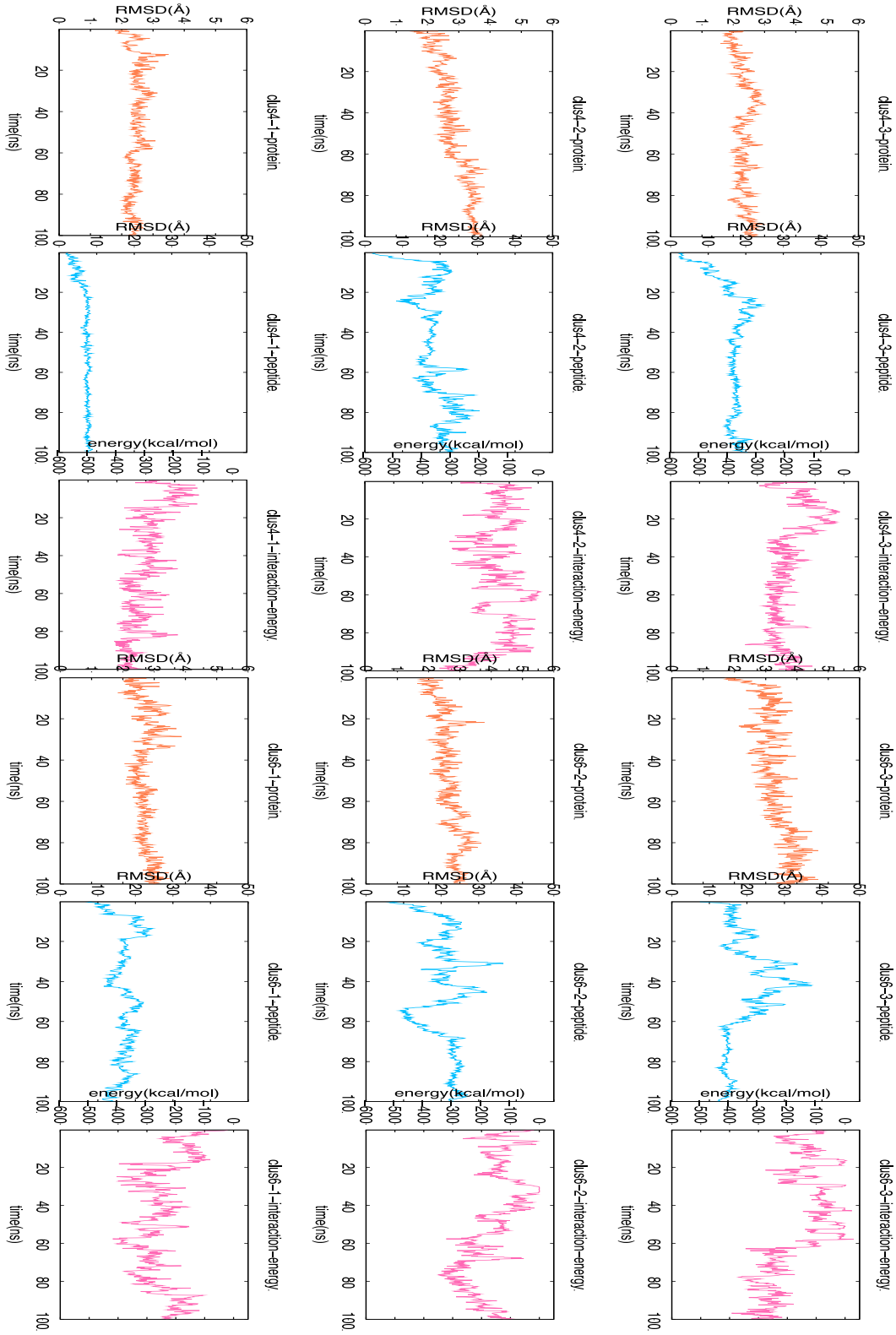


Figure 4.5 The 100 ns by 3 trajectories for clusters 2, 4 and 6. The protein backbone RMSD, peptide backbone RMSD and protein-peptide interaction energy are shown for each 100 ns trajectory.

Figure 4.5 continued



4.3.3 Model 2 in explicit solvent (Table 4.1 NO. 13-20)

Model 2 was built to optimize some of the protein-peptide interactions. 8 clusters were identified from the ensemble MD, and they are divided into two groups by manually examining the protein-peptide interactions. Clusters 1, 2, 4 and 6 are considered to have stronger protein-peptide interactions, and clusters 3, 5, 7 and 8 have relatively weaker interactions. Each cluster average was energy minimized and solvated in explicit solvent TIP3P. The MD simulation was run for 100 ns. The results are shown in Figure 4.6 and Figure 4.7. In clusters 7 and 8, the peptide dissociated as observed from the trajectory and shown in the peptide RMSD data. For clusters 1, 2 and 4, the N terminus of the peptide dissociated from the protein and the Tyr moved out from the active site. Clusters 7 and 8 saw the peptide dissociated quickly, and only some weak interactions remained for cluster 8. Interestingly, in clusters 3 and 6, it is observed from the trajectories that the peptide moved up to the cleft during the simulation; however, the Tyr did not find the interactions with Arg. For cluster 5, the peptide moved out of the cleft during the NPT stage.

These results show that the new model, especially the four clusters with relatively strong interactions, has the protein-peptide complex in a more stable state. However, the simulation did not help with improved interactions, and the Tyr on the peptide moves out of the catalytic site most of the time, indicating that the sampling in explicit solvent is not sufficient to generate a stable protein-peptide complex.

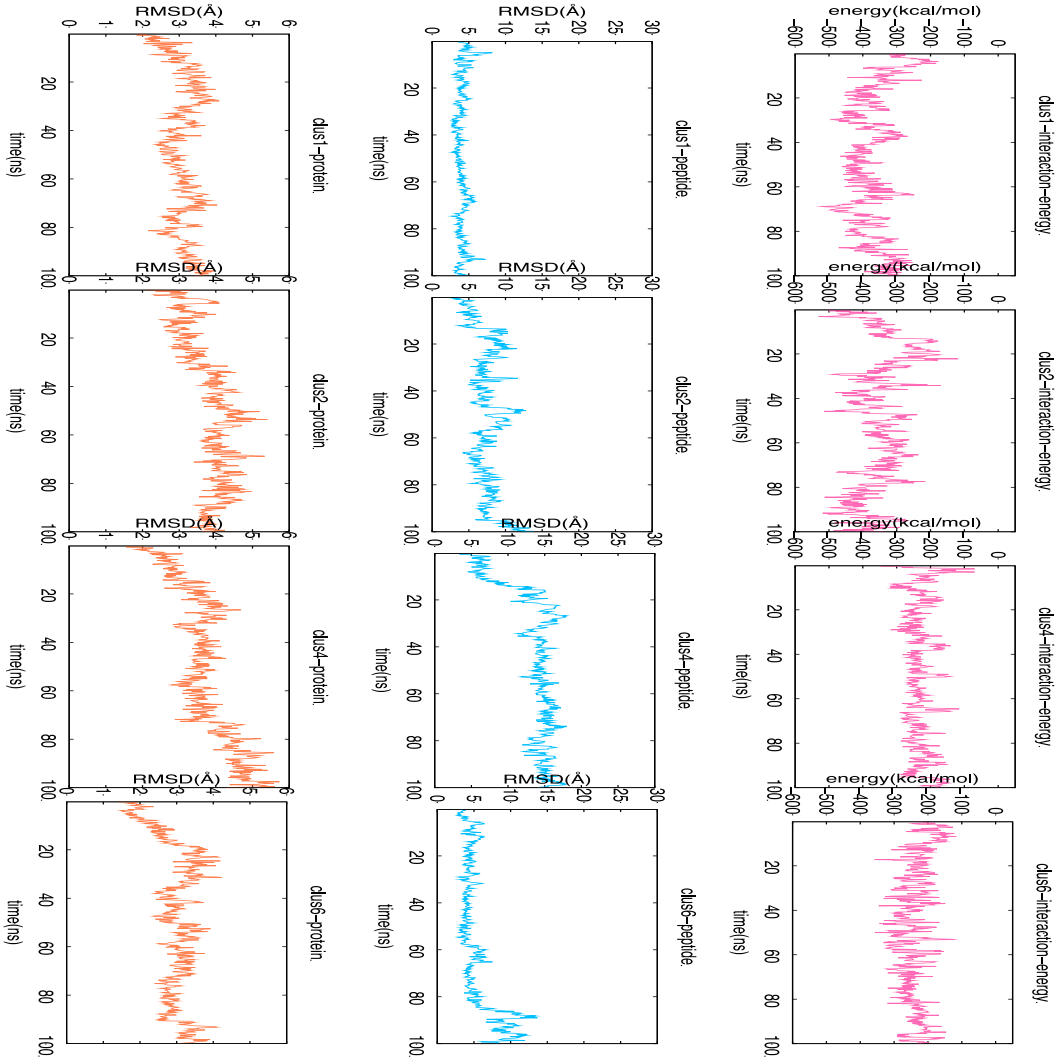


Figure 4.6 The results for the 4 clusters in Model 2 with relatively strong protein-peptide interactions in explicit solvent.

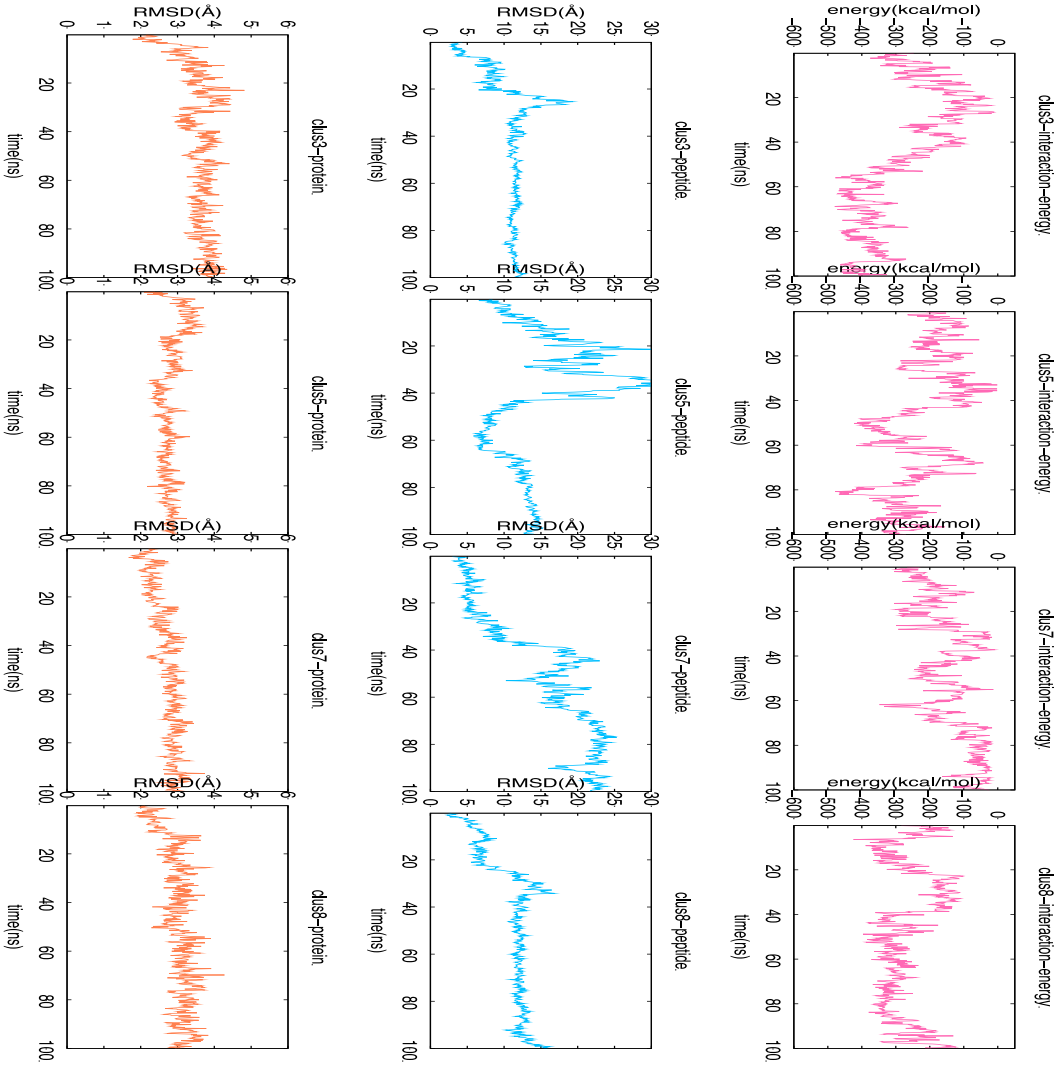


Figure 4.7 The results for the 4 clusters in Model 2 with relatively weak protein-peptide interactions in explicit solvent.

4.3.4 Model 1 in implicit solvent for the 9 cluster averages (Table 4.2 NO. 2-10)

To explore if the instability of the protein-peptide complexes is caused by the effects of the explicit water molecules, we put the systems in implicit solvent model FACTS and analyze the behavior of the complexes. The simulations for the 9 energy-minimized cluster average structures were for 100 ns for each complex. The CYP label was kept in the simulations. The results are shown in Figure 4.8. Of all the 6 clusters 1-5 and 9 in cleft binding mode, the Tyr residue stayed in place during the simulations. Only cluster 1 has the peptide stable as shown in the peptide RMSD data. For the other 5 clusters, the two

ends of the peptide moved around or up and down to different extent. The interaction energies generally remained stable, and in none of the cases increased significantly. For the three clusters 6-8, Tyr moved away from the active site during the simulations. The peptide also had some movement, especially in cluster 7.

In conclusion, the implicit solvent model simulations support the experimental results that the Src substrate peptide might adopt a cleft-binding mode.

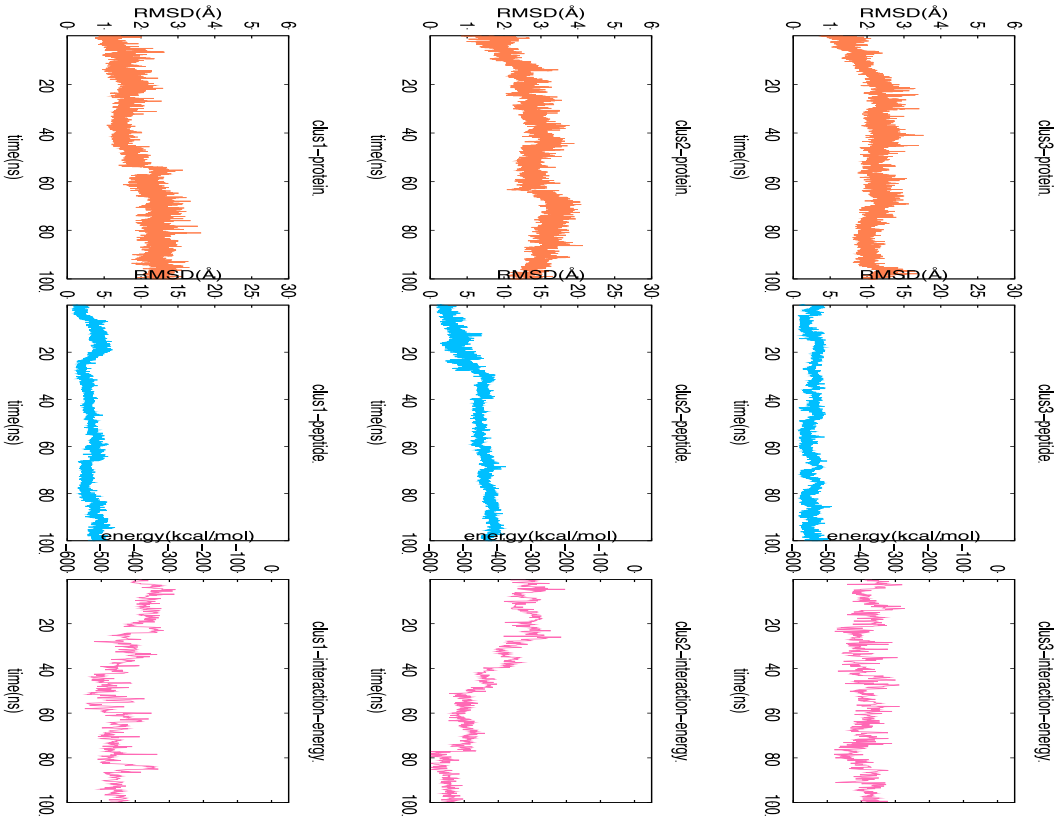
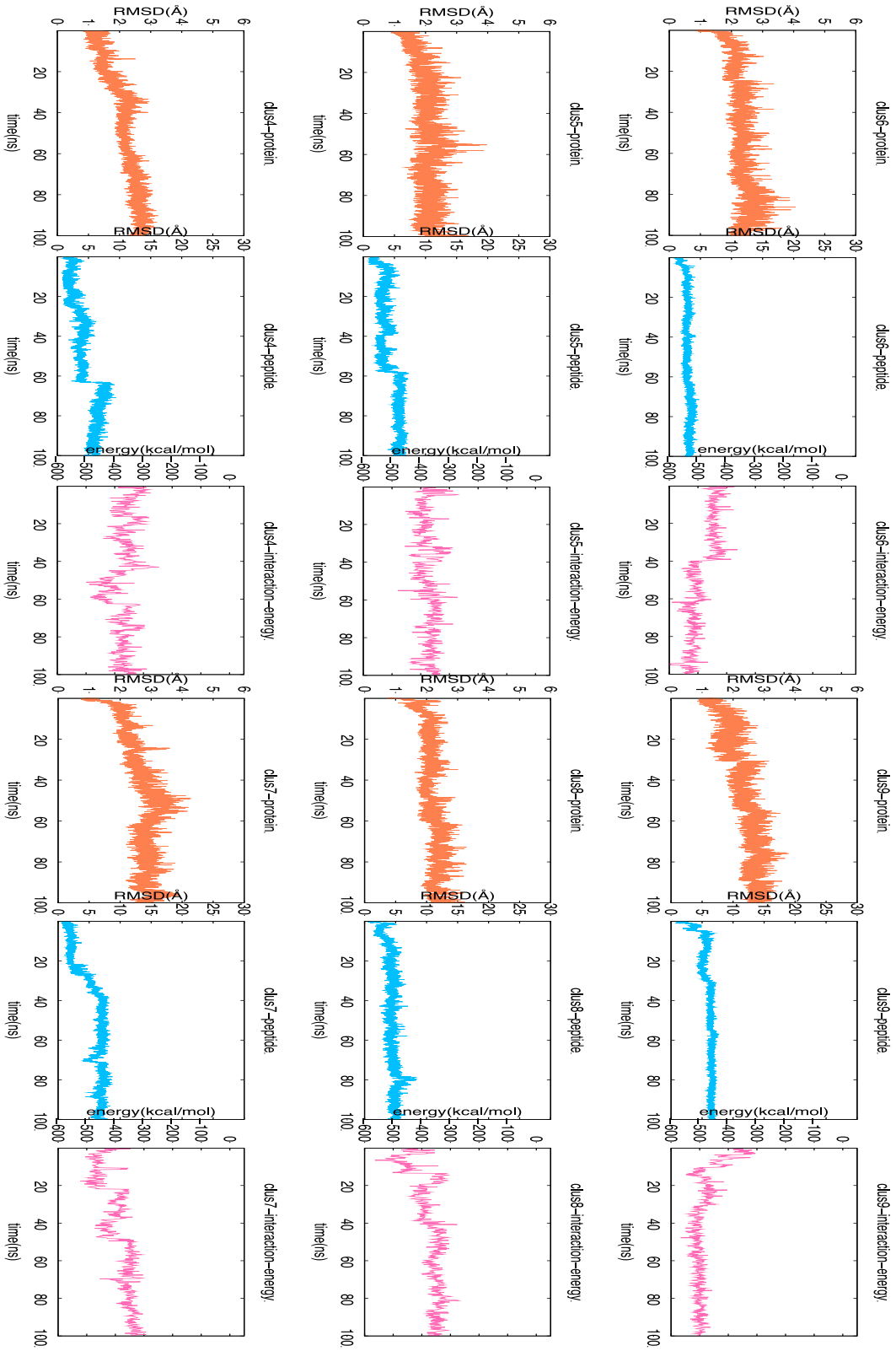


Figure 4.8 The 9 clusters in Model 1 in implicit solvent. The protein backbone RMSD, the peptide backbone RMSD, and the protein-peptide interaction energy are shown for each cluster for a 100 ns simulation.

Figure 4.8 continued



4.3.5 Model 2 in implicit solvent (Table 4.2 11-18)

To see if a better modeling improves the protein-peptide complex stability in implicit solvent, we also ran 100 ns simulations for the 8 cluster averages in Model 2. The ALA residue was used to replace CYP in Model 2. For all the 4 cluster averages where the interactions are determined strong, the Tyr stayed in the catalytic site. Clusters 1 and 2 are in cleft binding mode and clusters 4 and 6 are in C-lobe binding mode. The peptide RMSD values show the peptide in clusters 1 and 2 are more stable than these in clusters 4 and 6. For cluster 4, the N terminus of the peptide moved out of the cleft causing an increase in RMSD value. For cluster 6, the peptide moved away from the C-lobe and was closer to the cleft at the end of the simulation, causing a peptide RMSD of ~15 Å. For the four clusters with relatively weak protein-peptide interactions, clusters 5 and 8 had the Tyr moved away, while clusters 3 and 7 had the Tyr in place. The peptides in the four complexes also moved as indicated from the peptide RMSD values. Among them cluster 3 is relatively stable, with the Tyr in place and the peptide interacting with the protein.

Model 2 provides stable cleft binding mode for the complex. Also, the behavior of the peptide in cluster 6 indicates the cleft mode might be preferred. A C-lobe binding mode cannot be ruled out based on cluster 3 results, although the peptide has the tendency to move around in this complex.

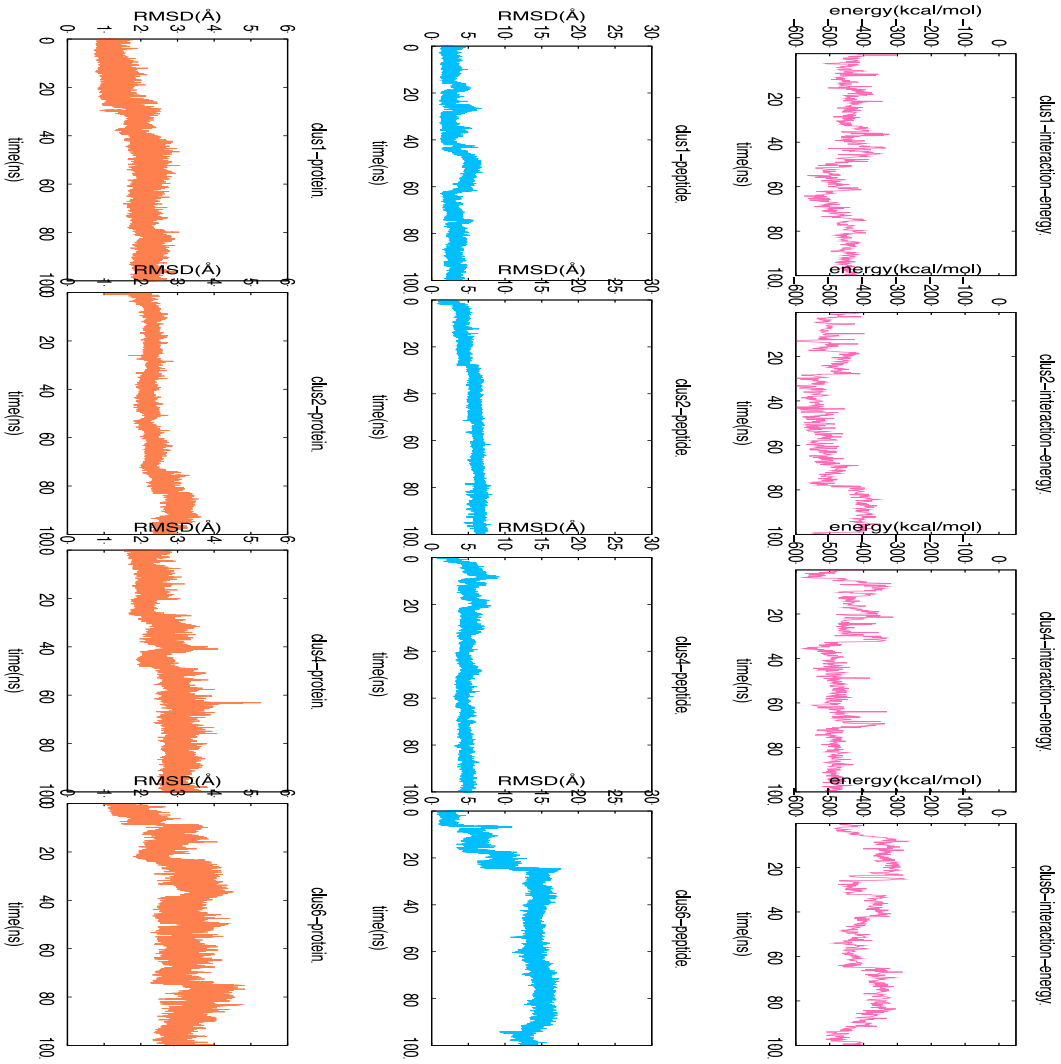


Figure 4.9 The results for the 4 clusters in Model 2 with relatively strong protein-peptide interactions in implicit solvent.

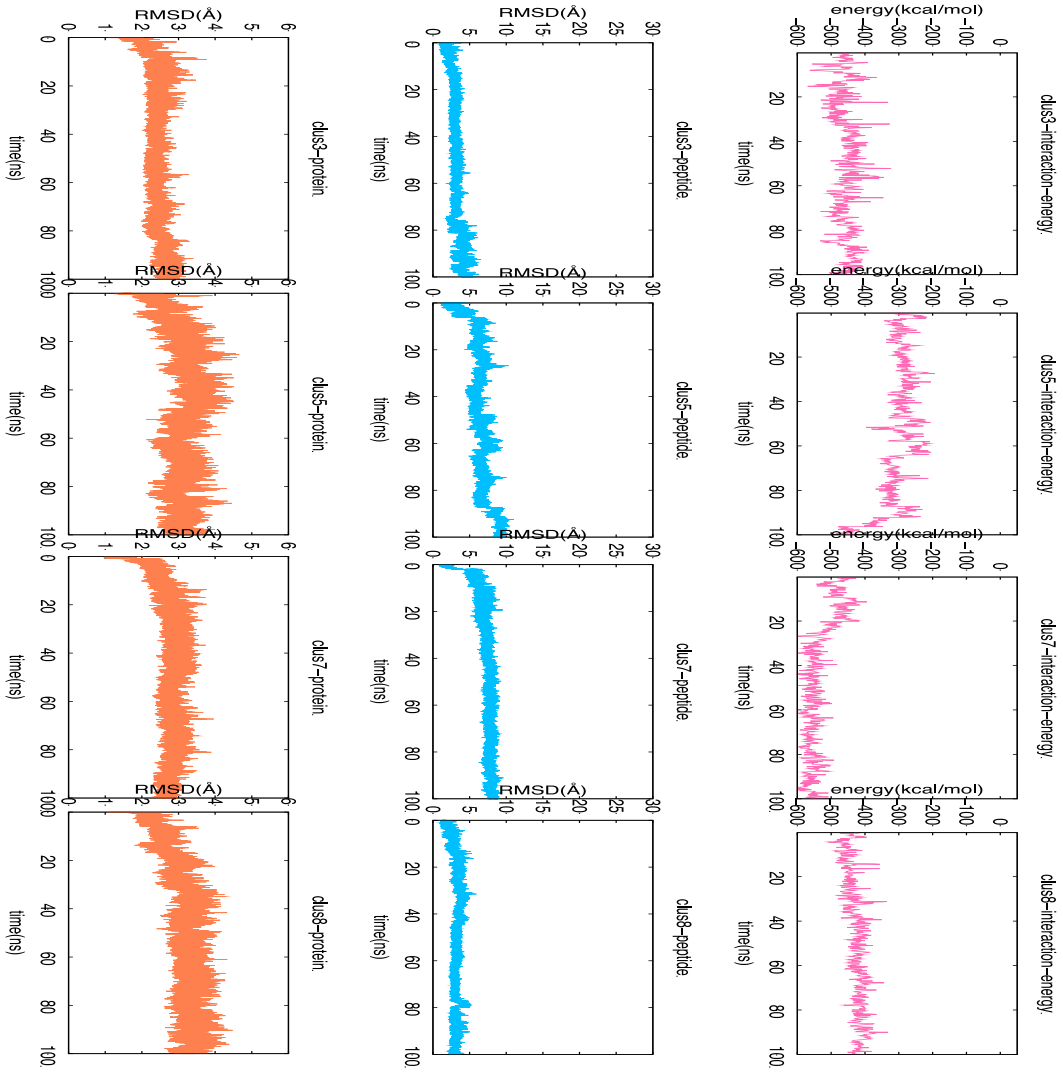


Figure 4.10 The results for the 4 clusters in Model 2 with relatively weak protein-peptide interactions in implicit solvent

4.4 Conclusions

We use MD simulations to model the Src-SSP complex to identify possible binding modes for the Src substrate. To find the binding mode for a small drug molecule with an unguided simulation starting with a free-state ligand would require simulation in μs timescale and MD-specific purpose machines[145]. For a peptide with previously unknown binding mode, the simulation time sufficient to sample the possible conformations would be impractical with current available computation power. Here we

use ensemble MD to get possible binding modes first, then use equilibrated MD to study the complexes.

The stability of the complex is affected by several factors, including the initial quality of the modeling, the solvent environment, and the flexibility of the peptide. Here the substrate we chose has a relatively low K_M = 300 μM. Therefore, the affinity of the peptide is relatively low and make it harder to model the complex. The two models we built from ensemble MD are significantly affected by the explicit water molecules in solution, probably due to the weak protein-peptide interactions.

The protein-peptide complex behavior in implicit solvent proves that the cleft binding mode is possible. Specifically, there are several cases where the peptide dissociates from the C lobe and move up to the cleft. Also, our modeling shows that a C-lobe binding mode is not required to satisfy the C-lobe PRE. A cleft binding mode with a CYP label that moves freely can also satisfy both N-lobe and C-lobe PRE in solution.

Longer simulations might be required to get a more solid solution. In addition, new models might provide more insights into the Src-SSP interactions.

CHAPTER 5. DOCKING FLEXIBLE MOLECULES USING GLIDE

5.1 Introduction

The accurate docking of flexible ligands or short peptides has remained a topic under research progress. The ligand docking process, taking the academically and industrially widely used Schrödinger package as an example, can be divided into several steps. The first step is to prepare the proteins to an all-atom structure with proper protonation states in the physiological pH environment using Protein Preparation Wizard. In the docking process, the protein is treated as rigid except the option to rotation a few side chains, therefore, a careful selection of the initial structure and protein preparation is essential to a successful docking project. The 2nd step is to prepare the ligands using LigPrep. This includes generating conformational, protonation states and rotamer states. Or, a prepared library can be imported for docking. Thirdly, ligand docking that can be viewed as a two-stage process including grid generation and the ligand docking step. The details for each step are described in methods.

The main challenge in flexible ligand docking is to generate ligand conformations. Flexible molecules or short peptides have significantly more rotatable bonds compared to traditional small drug molecules, and the current ligand conformational generation techniques usually cannot exhaust all possible conformations for the ligands. With the increase in number of flexible bonds, the number of possible conformations increases exponentially. Some conformations that would fit the binding site, might never be generated during the ligand preparation process. On the other side, following Schrödinger workflow, the ligand conformations are generated without the knowledge of the binding site. In the case of flexible molecules, some conformations that are far too distorted to fit into the binding site are generated. Besides, the scoring of the protein-ligand complexes can be a problem due to the increased number of protein-ligand interactions, and the much larger number of possible conformations. Currently Schrödinger suites have no independent workflow for flexible molecule docking apart from regular docking.

The Schrödinger package provides an interface named Maestro for molecular modeling tasks. The Glide module is used for docking. Docking is a two-stage process, the first stage is Grid Generation and the second stage is Ligand Docking. The docking has three types of precisions, from the computationally least expensive to the most costly are High Through-put Virtual Screening (HTVS), Standard Precision (SP) and Extra Precision (XP). SP mode is designed to avoid false negatives using a soft potential, and XP mode uses a hard potential to minimize false positives. Glide works by filtering generated conformations through the "Glide funnel". The funnel has four steps namely Rough Scoring, Refine, Grid Minimization and Post-Docking Minimization. At the end, the conformations are scored and ranked by Glide scoring functions as final poses. The binding poses with lower binding scores have better protein-ligand interactions.

Here two projects of docking flexible molecules are presented. One is to dock a pYEEI-like molecule to Src regulatory domain SH2, and the other is to model Src kinase domain-SSP peptide binding. Inhibitors have been designed to target Src SH2 domain to disturb signal transduction[146-147]. The structure of SH2 domain is shown in Figure 5.1A. SH2 is a modular unit with about 100 residues. SH2 domains mediate protein-protein interactions in tyrosine kinase signaling by recognizing and binding to tryrosyl-phosphorylated peptide sequence on the target proteins. SH2 has been identified in over 110 human proteins, and a number of these proteins are over-activated in diseases[148]. C-Src SH2 domain preferentially binds to peptides containing a pYEEI (Figure 5.2 A) motif, and the crystal structure of Src SH2 in complex with pYEEI is solved at 1.9 Å resolution[149]. In the crystal structure, the phosphate binds tightly in the pY pocket, while the side chain of isoleucine fits into the specificity pocket (Figure 5.1B). The details of SH2-pYEEI interaction is shown in Figure 5.1C.

The structures of two small molecule inhibitors (referred to as L1 and L2 in this Chapter) targeting SH2 domain were obtained from Prof. Borch. The structures of pYEEI and the two molecules are shown in Figure 5.2. The two molecules have high conformational flexibility with several rotatable bonds including three chiral centers in the structure. For both molecules, one end of each compound is a phosphate group, which is designed for pY pocket; the other end has a phenyl group, which would ideally fit into the specificity pocket. The two molecules are predicted to mimic the binding mode of pYEEI given the chemical

structural features. The objective of this project is to 1) predict the binding mode of the ligands 2) identify the relationship between the ligand affinity and the three chiral centers.

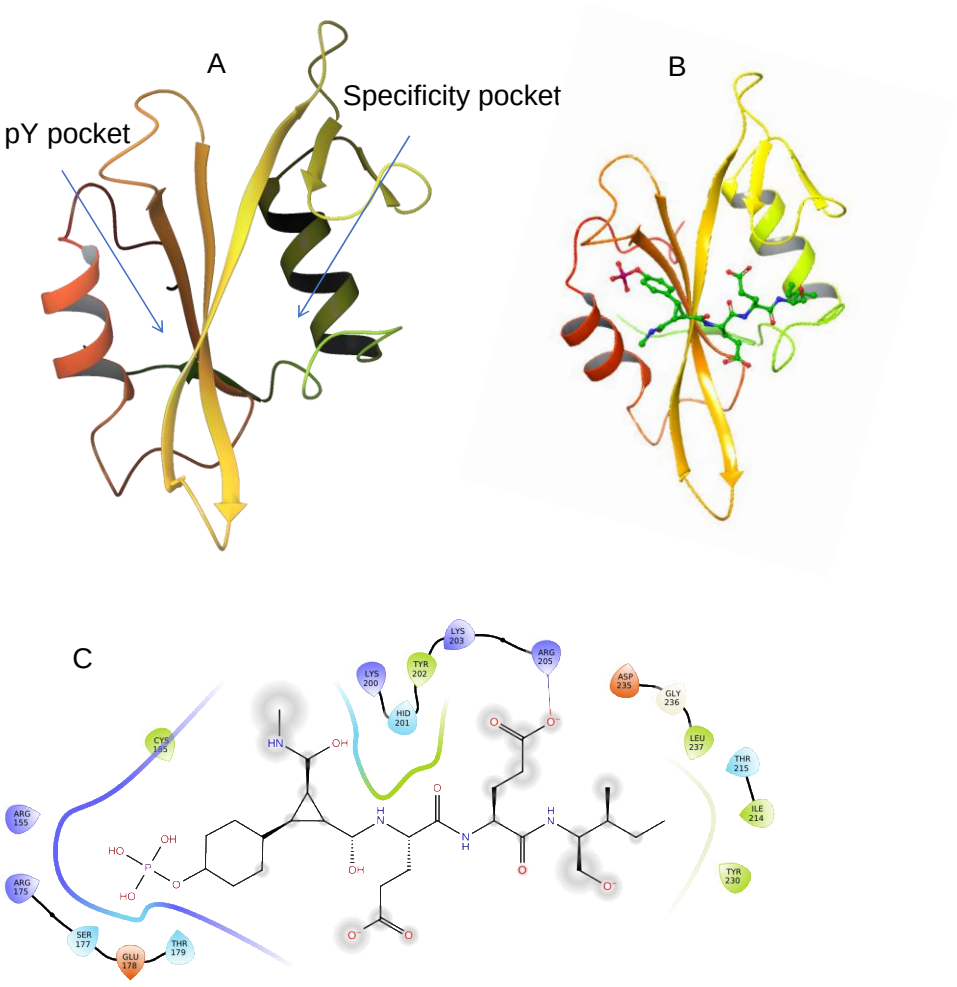


Figure 5.1 An illustration of SH2 domain structure and SH2-pYEEI binding. PDB ID 1IS0. A: Src SH2 domain showing the two binding pockets. B: SH2-pYEEI complex showing the ligand binding mode. The pTyr residue is constrained in this structure. C: the detailed protein-ligand interaction scheme.

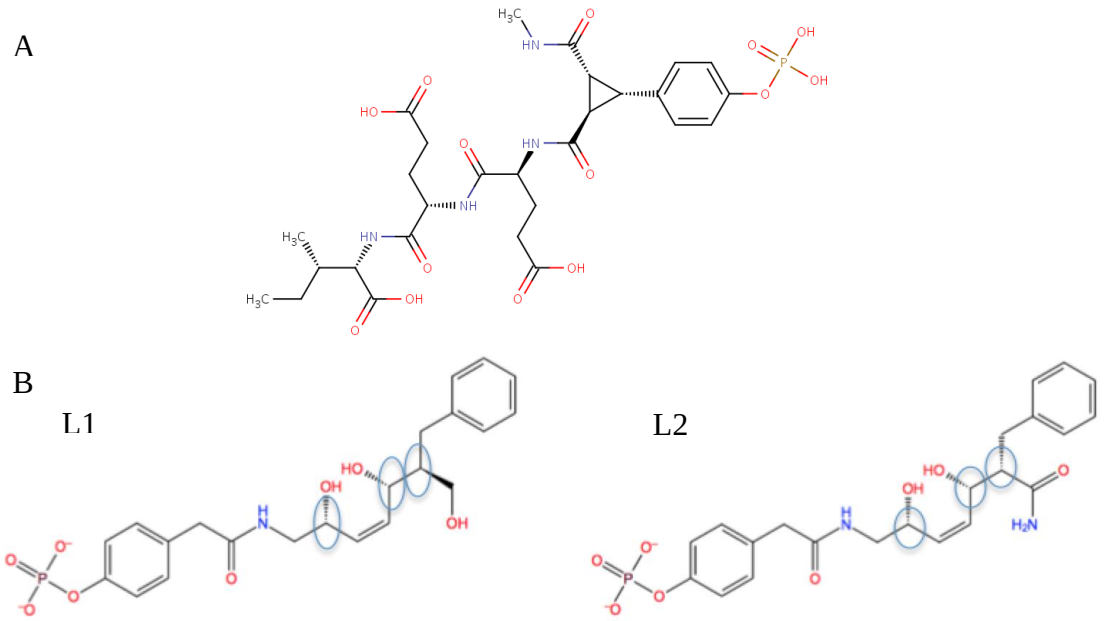


Figure 5.2 The chemical structure of the constrained pYEEI and the two ligands mimicking pYEEI binding mode. A: pYEEI with pTyr constrained. B: two pYEEI-like ligand L1 and L2. Circles indicates the three chiral centers in the molecule.

The other project is to model Src kinase domain-SSP peptide binding. As described in the introduction of Chapter 4, it is not directly shown with crystal structure proof if the Src substrate binds to the kinase domain in the cleft mode like serine/threonine kinases, or, binds to the lobe like other tyrosine kinases. Here we use Glide to dock the 11-residue peptide to Src kinase domain to explore a possible cleft binding mode.

5.2 Methods

5.2.1 Protein preparations

Src SH2 domain

Src SH2 in complex with PYEEI (PDB ID 1IS0) was used. The Src PDB contains 2 SH2 subunits and some crystal water molecules. The Schrödinger's Maestro interface was used for all operations. We overlaid the two subunits to identify the structural difference of the two units and determine if there are any conserved water molecules. Two waters that

function as bridges between protein and ligand by forming hydrogen bonds are potentially important.

The 3D structures of the two SH2 units are almost identical, therefore only chain B that has less missing atoms was kept for future work. The protein was prepared using Schrödinger's Protein Preparation Wizard (PPW). Preprocess was done to assign bond orders, add hydrogens, create disulfide bonds, fill missing side chains using Prime and cap termini. For the protein refinement, the hydrogen bond network was optimized at pH=7. Crystal water molecules are removed except the two identified functionally important water molecules. Three protein systems were generated, with the two water molecules both removed, or one of the water molecules kept, or both kept. Two types of restrained minimization were used: 1. hydrogens only and 2. converge heavy atoms to RMSD 0.3A. The reported problems from PPW were viewed and some bond orders were manually adjusted.

Src kinase domain

The protein was taken from the energy-minimized average of clus 2 (filename 56.67_clus2_min.cor). The protein was originally prepared from the active form of Src kinase domain, PDB ID 1Y57.

5.2.2 Ligand preparations

L1 and L2

The two ligands were built using the chemical building blocks in Maestro. The structures were built in 2D Sketcher then converted to 3D. The module LigPrep was used to generate possible ligand states at pH 7 using Epik. Tautomers were generated. All stereoisomers were generated, resulting in 8 conformers with 3 chiral carbons in each ligand. 16 ligands were generated after preparation, 8 steroisomers for each ligand.

SSP peptide

The peptide structure and coordinates was taken from file 56.67_clus2_min.cor with Src kinase domain as described in protein preparation. The CYP label at the end of the peptide was replaced with ALA as described in Chapter 4. Two runs of LigPrep were attempted. The input parameters are shown in Table 5.1. When chiralities are chosen to be determined (DETERMINE_CHIRALITIES) from 3D structure, only 4 ligands were output

from LigPrep. When all combinations of chiralities are generated, the output increased largely to 128. However, these structures do not retain the natural chiral structures. The ligands from the Run 1 were used for docking, although less structural variations are generated in this set.

Table 5.1 Parameters used for ligand preparation.

Parameter	Run 1	Run 2
FORCE_FIELD	OPLS3	OPLS3
EPIK	yes	yes
USE_DESALTER	no	no
GENERATE_TAUTOMERS	no	no
DETERMINE_CHIRALLITIES	yes	no
IGNORE_CHIRALITIES	no	no
NUM_STEREOISOMERS	64	64

5.2.3 Flexible molecule docking for L1 and L2

In Glide, docking is a two-step process, grid generation and ligand docking.

Grid generation

The parameters are chosen as described here. For the receptor, the default Van der Waals radius scaling was used. For the binding site, the centroid of workspace ligand was chosen as the center of the enclosing box. For the ligand size, dock ligands similar in size to the workspace ligand was selected. Grids with multiple sizes was generated to evaluate its effect on ligand docking. The smallest side length is 24 Å, with the largest being 45 Å. No constraints or excluded volumes were used. The rotatable groups close to the binding pocket were allowed. Three grids were generated with different numbers of waters, one without water, one with one of the waters, and one with both waters.

Ligand docking with and without core pattern

The core constraints option in Glide chooses a chemical structure as a “core” pattern in a reference ligand. The reference ligand is usually in a known well-docked conformation. The core setting would restrict docking of the same core structure in the ligand within the specified distance to the reference core position. Structures that do not contain the core pattern will not be docked.

We first docked the ligands without using core constraints. Standard precision (SP) mode was used. The ligand sampling was chosen to be flexible. Post-docking minimization

was performed. Extra precision (XP) mode was tested and lower docking scores were observed. The docking scores in XP mode were improved partially due to the difference in scoring functions and local optimizations, while the ligand binding mode remained similar. The XP mode is more time-consuming and therefore was not used in further runs.

The core pattern was used to improve the docking results for in regular docking the phenyl group might not find the hydrophobic binding pocket. In this case, two cores were chosen using a well-docked ligand as the reference. The first is the phosphate group, and the tolerance was set to 1 Å. The second is the phenyl group, and the tolerance was set to 3 Å. The phosphate group is well-docked in most cases with specific interactions, thus a lower tolerance was set.

5.2.4 Docking with NOE constraints

Grid generation with NOE constraints

The grid was generated using the clus2 structure as a reference (cleft binding mode). The peptide center was used as the grid center. It is worth noting that for the two cleft and C-lobe binding modes, the peptides share a similar center, so only one grid was generated. The inner cubic box dimension was set to the defaulted 10 Å edge length.

NOE positional constraints was set for the grid. NOE constraint sets a pair of atoms restrained to a certain distance. In Glide, the NOE positional constraints option sets the constraints between an atom in protein and a specified chemical group in ligand, and the choice of atoms is completed in two steps. The first step is to pick an atom in protein and define spherical shells around the atom in grid generation. The center of the sphere is defined as the centroid of the picked atom. The second step is in docking, where specified chemical groups of atoms in ligand should occupy the defined shells. We picked four atoms to define NOE constraints, ASP 386 OD1, ASP 386 OD2, Arg 388 NH2 and Arg 388 NE. The minimum and maximum distance are 1 Å and 3.5 Å respectively, creating a shell around the selected atom. The constraints are to keep the interactions with the tyrosine on the peptide in place. Docking a ligand with similar size to the picked ligand is chosen, ending up in a 39.18 Å edge length of the outer box. The rotatable groups close to the binding site were allowed to rotate. In the auto-generated input file tab, the PEPTIDE option was set to True.

Peptide docking

9 runs were launched. The parameters are listed in Table 5.2. For precision, SP-peptide mode was used. Reward intramolecular hydrogen bonds was chosen. The output poses per ligand was set to 100. For some runs, the Glide funnel was made larger (MAXKEEP and MAXREF in Table 5.2) to retain more conformations in the earlier stages and potentially increase the number of output binding poses. The constraints and ligand features were adjusted to better find the protein-Tyr interaction.

Table 5.2 The constraints and important parameters for docking. Ligand feature needs to be within the specified distance as set in the Grid constraints for the ligand to be kept through the Glide funnel. MAXKEEP: number of poses per ligand to keep in initial phase of docking. MAXREF: number of poses to keep per ligand for energy minimization.

NO. run	Constraints groups	Ligand feature (in chemical identifier format)	MAXKEEP	MAXREF
1	All 4 in one group	O	100,000	1,000
2	None		100,000	1,000
3	All 4 in one group	O	100,000	1,000
4	None		100,000	1,000
5	Two in one group (386 OD1 OD2)	[H]cc(c[H])O[H]	100,000	1,000
6	Two in one group (386 OD1 OD2)	cO[H]	100,000	1,000
7	Two in one group (386 OD1 OD2)	cO[H]	1,000,000	10,000
8	386OD2, 388NH2	cO[H]	1,000,000	10,000
9	Two in one group (386 OD1 OD2)	cO[H]	1,000,000	10,000

5.3 Results and Discussion

5.3.1 Docking flexible ligands L1 and L2 to Src SH2 domain

16 ligands were generated from LigPrep for L1 and L2, 8 for each, for each chiral center would have two possible conformations. We named the ligands 1RRR, 1RRS, 1RSR, 1RSS, 1SRR, 1SRS, 1SSR, 1SSS for L1, and the same naming rule is applied to the 8 ligands for L2.

During docking, the conformations for each ligand are sampled internally, including different rotamer states. Users do not have access to the complete conformation sets

sampled. The numbers in the Glide funnel can be increased in each stage to retain more conformations, with a larger computation time cost.

Three grids were generated with different numbers of water molecules. One without water, one with one water molecule, and the third one with two water molecules. The top 2 scores for each ligand are listed in Table 5.3. The waters generally did not help with docking as shown from the binding poses. In some poses, the two water molecules did function as bridges between the protein and the ligand. This is shown as a lower docking score in the 2W grid compared with 0W grid. However, the two molecules obstructed ligand binding in many cases, for they are treated as part of the rigid receptor in Glide; in the table, the lowest ranked docking scores for 2W grid are higher than those in 0W grid. This obstruction becomes more significant when core constraints are applied to the ligands. For L1 and L2, the part which might interact with the water molecules is highly flexible, and the accurate interactions with the waters were rarely found. Therefore, the grid with no water molecules were used for further work.

Table 5.3 SP scores with different numbers of water

0W	glide gscore		1W	glide gscore		2W	glide gscore	
name	1	2	name	1	2	name	1	2
1SRS	-9.1	-8.69	native	-9.27	-8.45	native	-9.37	-8.06
1RSS	-8.7	-8.54	1RRS	-8.33	-8.33	2SRS	-9.36	-8
2SSS	-8.49	-7.59	2RRS	-8.15	-7.33	2RRS	-8.96	-8.9
2SRS	-8.42	-8.33	1RSR	-8.03	-7.48	2SSR	-8.1	-7.94
native	-8.4	-8.36	2RSR	-8	-7.71	1SSR	-7.93	-7.24
1RSR	-8.34	-8.21	1SSS	-7.93	-7.59	2RRR	-7.89	-7.75
2RRS	-8.25	-8.18	2RSS	-7.8	-7.48	1RRS	-7.82	-7.32
1SSR	-8.17	-7.82	2SRS	-7.75	-7.54	2RSS	-7.81	-7.8
1RRR	-8.05	-8	2SSR	-7.74	-7.71	1RRR	-7.69	-7.26
2SSR	-7.9	-7.68	1RSS	-7.54	-7.52	2RSR	-7.65	-7.64
2SRR	-7.84	-7.83	1RRR	-7.53	-7.48	1RSS	-7.58	-7.33
1SSR	-7.81	N/A	1SRS	-7.48	-7.35	2SRR	-7.53	-6.64
1RRS	-7.74	-7.66	2SSS	-7.45	-7.16	2SSS	-7.49	-7.13
2RSR	-7.67	-7.6	2RRR	-7.3	-7.2	1SSS	-7.41	-7.17
2RRR	-7.67	-7.11	1SSR	-7.29	-7.24	1SRS	-7.4	-7.06
1SSS	-7.59	-7.57	1SRR	-7.2	N/A	1RSR	-6.92	-6.89
2RSS	-7.5	-7.17	2SRR	-6.59	-6.41	1SRR	-6.75	N/A

The Glide grid is structured as follows: two cubic boxes, the inner box and the outer box, share the same geometric center. The inner box size defines the volume that the ligand center explores during the exhaustive site-point search, and defines the volume in which the grid potentials are computed, all ligand atoms of a valid pose must be located within this outer box. The default edge length of the inner box is 10 Å, and the edge length of the outer box is the length of the reference ligand plus the side length of the inner box. How outer box edge length is calculated is not mentioned clearly in Glide manual but learned by testing with parameters. According to Glide manual, a grid should 1) cover the binding site 2) be large enough to hold the whole ligand, and a grid larger than needed would be a waste of computer time. The computation cost that associated with different grid size would be obvious in high-throughput screening where a large ligand library is involved.

The grid size has an affect on the docking results when it is altered within a certain range. Schrödinger does not provide an explicit explanation for this affect. To our knowledge, the reasons for the variations are the numerical algorithms for both searching and scoring, including complex numerical algorithms, complex grid-based potentials and the practical limitations on ligand conformation sampling.

A series of grid size was used to examine its effect on docking results. Our conclusion is that for flexible molecules, a larger grid might be required to get the optimal results. Our ligand length is approximately 19 Å, and we started with an outer box of side length 24 Å, which is theoretically large enough to hold the whole ligand. In the docking results, it is observed that the phosphate group always fit the pY pocket, but the phenyl side usually did not fit into the hydrophobic pocket. Close to the edge of the box, the last flexible bond on the phenyl side would rotate and fold phenyl inward to prevent fitting. One explanation is that the edge of the box is restricting the poses when the phenyl group is close to the edge. To test this interpretation, we increased the grid size gradually to 45 Å. The increase of the grid size alleviates the restriction on the phenyl side, but there still exists bad poses probably due to incomprehensive sampling. Another observation is that the distance between the phosphate and the grid border is smaller than the distance between the phenyl group and the other edge, which means the grid size do not fully account for the poor binding of the phenyl group.

A further explanation would be different functional groups have different priority in the scoring function. The phosphate-pY pocket interaction is much favored than the other interactions and contributed largely to the docking score, thus the poses with the phosphate-pY interactions were always kept. The hydrophobic interaction does not contribute as much to the docking score and can be filtered out in favor of other interactions. The ligands are highly flexible and the three hydroxyl groups in the middle of the ligand can form interactions with the residues between the two binding pockets.

The core settings are used to determine if the phenyl group is poorly docked is due to non-optimized Glide settings that the structures are lost in the funnel, or the conformations were not generated before entering the funnel. We defined two function groups in a well-docked ligand as cores, the first is the phosphate group (core 1), and the second is the phenyl group (core 2). The use of phosphate group did not have an influence on the results, probably for that the phosphate always fits into the pY pocket in the output poses. For core 2, we set tolerance to the reference ligand as 3 Å that only binding poses with phenyl within 3 Å of the reference phenyl will be retained. Compared to the previous docking results, some docked poses with phenyl in the hydrophobic pocket were obtained. Because they appeared in the core-based docking, these conformations were pre-generated within Glide. It is very likely that the core setting changed how Glide ranks the output ligands rather than generate new conformations to satisfy the core restraints.

Larger funnels were used in a 26 Å grid to test if the good ligand conformations are not generated or are filtered out in the funnel. We doubled the number of ligands for initial phase of docking from 5000 to 10000 and best poses for energy minimization from 500 to 1000. A binding pose was found in this setting for a ligand, 2RSR, with no binding pose found in default settings. Surprisingly, the docking score (-8.483) is in the middle range of all the output poses, as opposed in the higher or worse range. This proves that the Glide funnel filters out some ligand conformations that would eventually yield good binding poses. However, this type of experiments is time-consuming. This run costs approximately three folds of time compared with the default setting. This shows that flexible molecule docking is more time consuming.

The chiral centers in the ligands are associated with the binding affinity of the ligands. With different grid size settings, some ligands always find the same binding pose, while

some hardly yield any good poses. Two ligands (1SSS and 2RSR) did not yield binding poses with phenyl in the specificity pocket without core constraints. The scores of the two ligands with phenyl in the specificity pocket was obtained with the phenyl core constraints. The scores are -7.671 and -8.331 respectively. The latter score is in the middle range of all docking scores; even in this case, the phenyl group is close to the hydrophobic pocket without fitting into it. The ligand structure shows that the S conformation of the 2nd chiral center causes the phenyl to flip out rather than fit into the hydrophobic pocket.

Glide results are dependent on initial ligand coordinates input. To test this, we altered only the ligand coordinates by randomly translating the ligands in space. No changes in bond angles/lengths was made. The output from the two runs were not identical.

The final results are integrated from the runs with different grid sizes. We calculate the average scores for each ligand with binding poses that has the phenyl finding the specificity pocket. Only the top 2 scores for each ligand in each run are considered. Then we calculated the average docking scores for each ligand. The results are listed in

Table 5.4.

Based on the docking score and visualization of the binding poses, we divide the ligands into three groups.

Better: 2RRR, 2SRS, 1SRS, all three are easy to dock and have high scores.

Average: 1RRS, 1SSR, 1RSS, 2RSS, 2SSR, 2SSS, 1RRR, 2RRS, 2RSR. 1RRS have less poses found and therefore is categorized into the average group.

Worse: 2SRR, 1SRR, 1SSS. These three ligands have low docking scores, or no good binding pose was found.

Two exceptions that are not grouped are 2RSR and 1RSR. For 2RSR, the binding pose was not obtained by changing grid size, but was found in a wider funnel with a mid-range score. 1RSR has only one pose identified by changing grid size; in larger grids, more poses were found with mid-range scores.

Table 5.4 The integrated results for docking ligands L1 and L2. The average docking score and the standard deviation for the scores are listed for each ligand.

NO.	Ligand	Avg Score	SD
1	1RSR	-8.9	0
2	2RRR	-8.78656	0.289387
3	2SRS	-8.65111	0.085792
4	1RRS	-8.64641	0.273335
5	1SRS	-8.64226	0.281581
6	1SSR	-8.51467	0.292235
7	1RSS	-8.43372	0.23577
8	2RSS	-8.43196	0.079796
9	2SSR	-8.42449	0.213615
10	2SSS	-8.37663	0.248248
11	1RRR	-8.36077	0.339145
12	2RRS	-8.2129	0.170084
13	2SRR	-8.00443	0.215023
14	1SRR	-7.13911	0.660343
15	1SSS		
16	2RSR		

5.3.2 Src-SSP modeling using peptide docking

In Glide, the peptide docking mode has different default values for several parameters to improve docking results for polypeptides. Three parameters are altered for the peptide docking mode (Table 5.5). Several other keywords are also set internally, including the maximum number of conformers, which is increased by a factor of 10; also, an increase about a factor of 3 is applied to the number of diameter directions. These parameter settings are designed to retain more ligand conformations in the Glide funnel.

Table 5.5 The three parameters that differ in different docking modes. MAXKEEP: number of poses per ligand to keep in initial phase of docking. MAXREF: number of poses to keep per ligand for energy minimization. POSTDOCK_NPOSE: number of poses to use in post-docking minimization.

Parameter	Default	XP	Peptide
MAXKEEP	5000	5000	100,000
MAXREF	400	800	1,000
POSTDOCK_NPOSE	5	10	100

The numbers of output and the docking score for the highest ranked ligand from each run are listed in Table 5.6. Docking without restraints yields ~100 docking poses per run.

Both C-lobe and cleft binding mode have been observed in the output binding poses. Two representative binding poses are shown in Figure 5.3. The docking scores for the two structures are -8.432 and -6.400 respectively. The scores are not very high, probably due to the weak interactions between the protein and the peptide. The C-lobe binding mode shows the C terminus of the peptide interacting with the C-lobe, while the cleft-binding mode shows the peptide residing entirely in the cleft. However, the Tyr in the peptide is not situated in between Asp 386 and Arg 388 in the two structures. Pose A has the Tyr flipping up interacting with a loop, while pose B has the Tyr interacting with only Asp 386 but not Arg 388.

To build binding poses with Tyr in place, we applied NOE constraints to the docking process. As shown in the table, adding constraints would greatly reduce the number of output, or even produce no output. No result was generated when both constraints for Asp 386 and Arg 388 were used. The increase of the funnel size does not increase the numbers of output significantly. The highest scored binding pose is from Run 6, with Tyr interacting with Asp 386, and the peptide partially fit into the cleft. These results show that some peptide conformations might never be generated rather than being filtered out in an earlier stage of the Glide funnel.

Table 5.6 The number of output from the docking runs.

NO. run	1	2	3	4	5	6	7	8	9
NO. poses	0	91	6	97	0	4	0	1	0
Top score	N/A	-8.50	-3.55	-8.50	N/A	-6.35	N/A	-5.80	N/A

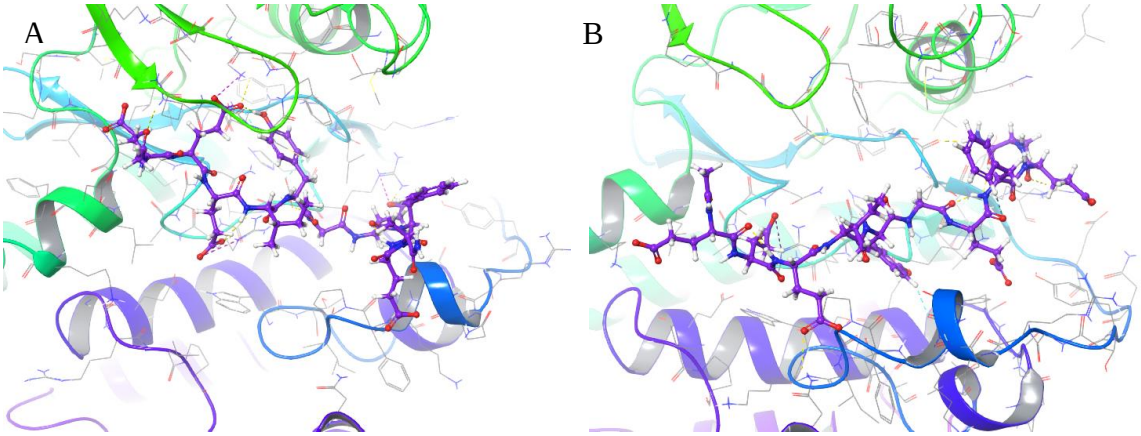


Figure 5.3 The two binding poses generated from docking without constraints. A: a pose close to a C-lobe binding mode. B: cleft-binding mode.

5.4 Conclusions

In conclusion, grid size, Glide funnel width and restraints have been experimented for docking flexible ligands L1 and L2. A proper grid size is required to accommodate the ligand flexibility at two ends in a computationally efficient way. The grid size should be larger for docking flexible ligands. When the grid is small, the edge of the grid would restrain the position of the ligand, and the better poses can be filtered out during the early stages. This is different from small molecule docking where the ligands are mostly rigid. The Glide funnel width is a trade-off between accuracy and computational cost for flexible docking. In peptide docking mode, the Glide funnel is wider than the XP mode. The width should be further increased if no good poses are generated from the default peptide docking settings. The restraints would improve the results when properly used, but the number of output would likely also be reduced with restraints in place.

Two factors need to be considered for future improvement. One is to sample conformations comprehensively for flexible molecules. The current ligand sampling in Glide does not have priority settings for different bonds. The flexible bonds are treated equally that some conformations generated are not "drug-like" while the conformations that would fit the binding site are not sampled sufficiently. A temporary solution would be to increase the grid funnel size to retain more structures. The other option is to use Glide flexible docking protocol. The protocol might only provide limited improvement for only side chain interactions are optimized, while the issue remains in the lack of ligand backbone conformation sampling.

Also, the scoring function needs to be adjusted for flexible molecules. For small molecules that do not fit, a high penalty is applied to the scoring function for steric clashes. For flexible molecules, there are more interactions between the protein and the ligand which contribute to the total score, while important interactions might be lost in favor of a combination of several non-essential interactions. For the same reason, the Glide score range for flexible molecules is relatively narrow, even for very different poses. An example is for the worst scored poses where the phenyl does not find the hydrophobic pocket at all, the scores are around -7, while the best docking score is higher than -9. The current scoring function would call for a more careful manual examination of output binding poses when docking flexible molecules.

Several factors would affect docking results in Glide. Alterations in the initial ligand conformations, including bond lengths and angles, torsional differences, or the identical ligand conformation differ in absolute coordinates would all affect the results. In general, ligands that are good binders will be affected less than the poor binders. To eliminate input dependencies, one option is to use the "Regularize input geometries" feature of the Virtual Screening Workflow. This option converts the input ligands to unique SMILES and back to 3D.

Glide gives good reproducibility with identical input. Two docking jobs with the exact same settings usually give identical results, same scores and same poses.

CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Conclusions

In conclusion, we developed the methodology of using ABPO for conformational transitions in all-atom protein systems. We experimented the types of reduced variables that are suitable for each conformational transition. The dihedral angle reduced variables are efficient for loop movement, while distance-based reduced variables are generally required for transitions that involves breaking and formation of inter-residue interactions. A series of analysis was done to evaluate the convergence and the outputs of the simulations. The ABPO parameters were tested to get the Normalized RV plots are examined to evaluate both the quality of the RVs and convergence of the simulation.

We obtained the conformational transition paths for the aforementioned protein systems. For the systems with simple transitions, the path can be easily described in terms of reduced variables. For large-scale transitions, the structures along the transition paths are analyzed to build the transition trajectory. The details and key events in the Src kinase conformational activation was described.

The Src kinase domain-SSP protein-peptide interactions and behavior was studied by long equilibrium MD in both implicit and explicit solvent. The behavior of the peptide is affected by explicit water molecules, especially when the binding affinity is low. In the explicit water, using the modeling of the protein-peptide complex, the peptide might dissociate quickly. In implicit solvent, the protein-peptide interactions are more stable. Another phenomenon observed is that the peptide can flip from one binding site to the other site. In several simulations, the C-terminus of the peptide dissociates from the C-lobe and move up to the cleft. This is explained by the rotation of the peptide backbone atoms. Our simulation results suggest a possible cleft binding mode for Src substrate peptide, while a C-lobe binding mode is potentially possible.

Flexible molecule docking was applied on Src kinase and regulatory domain SH2. The protein-ligand interactions were visualized and analyzed. Also, the peptide/flexible molecule docking protocol in with Glide was evaluated. Two issues identified are that the

ligand conformations are not sampled exhaustively with the increased number of flexible bonds, and the scoring functions have difficulty distinguishing good binding poses and those with many loose interactions. For the Src regulatory SH2 domain, the binding poses for L1 and L2 with designed purpose were obtained with top-ranked scores. For Src substrate binding, the cleft binding mode is possible from our results while there is still debate on the Src substrate binding mode.

6.2 Future directions

Currently ABPO utilized CHARMM ensemble module to accelerate sampling by launching multiple trajectories for each state. In this work, implicit solvent model FACTS is mostly used to quickly explore the possible sets of RVs and ABPO parameters. Several techniques to improve the simulation performance in explicit solvent has been advanced or developed in the past few years. The most notable are CHARMM DOMDEC[150] and CHARMM/openMM[151][152][153]. DOMDEC uses domain decomposition to accelerate the calculation of non-bonded forces, the most time-consuming step in calculating simulation trajectories. OpenMM utilizes GPU to do fast calculations. Both significantly boost the performance of CHARMM equilibrium explicit solvent simulations. We have tested the feasibility of doing ABPO simulations with explicit solvent, and the preliminary results for ERα LBD show similar profile for the RVs. However, the simulation speed is still traditional, as the ABPO implementation does not include any technique to enhance performance for explicit solvent. How to combine the power of the new techniques and ABPO to further accelerate the sampling can be an interesting topic.

We used equilibrium simulations to study protein-peptide interactions. Although modeling protein-peptide interactions is not a topic of this work, the modeling does have a huge effect on the simulations and should be carefully performed before any simulations. Besides, the simulations can be improved by using a proper forcefield that better models both proteins and peptides. Also, the effect of explicit water molecules need to be considered.

Docking flexible molecules can be difficult due to the increased degrees of freedom of the ligands. Current protocols (taking Glide as an example) focuses on improving the performance based on current framework. While some advancement is observed, problems like comprehensive conformation sampling of flexible molecule, retaining the “drug-like”

conformations in the funnel, and a scoring function that better describes protein-peptide interactions should be further researched. Also, special measurement or methods for flexible molecules might need to be developed to avoid the exponentially increased computational time with the number of flexible bounds and increased funnel width.

APPENDIX A. OTHER ATTEMPTED SYSTEMS

Three other protein systems that have different conformational states have been examined for the possibility of using ABPO for the transition. The systems are adenosine kinase (ADK), nitrogen regulatory protein C (NtrC), and retinoid X receptor (RXR). Each system was prepared as described in Chapter 2, and a 10 ns equilibrium simulation was launched for each state. The conformational difference for the two states for the three systems are shown in Figure A1.

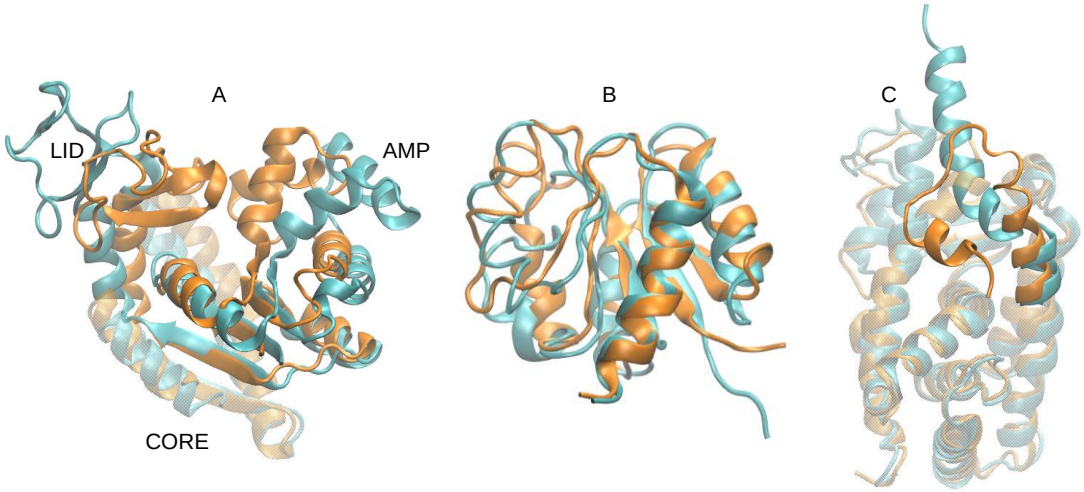


Figure A1 The conformational transition of the three systems, illustrated in ribbon representation. A: ADK. Cyan: open form, PDB ID 4AKE. Orange: closed form, PDB ID 1AKE. The protein has 214 residues in total, residues 160-214 are in transparent representation. B: NtrC. Cyan: the active conformation, PDB ID 1DC8. Orange: the inactive conformation, PDB ID 1DC7. C: RXR. Cyan: the apo form, PDB ID 1LBD. Orange: the agonist-bound form, PDB ID 1FM9. The receptor domain residues 227-458 are included in the crystal structures. Residues 227-429 are shown in transparent, and residues 430-458 are shown in solid colors.

The conformational transition of the open and closed states of ADK can be visualized as domain movement. ADK has three domains namely CORE, LID and AMP binding domain. From the open and closed conformation, the LID and AMP domains moves towards the CORE domain to form the closed three-domain conformation. The system has been chosen for preliminary examination for it is studied in several other methods for computational transitions. However, in our 10 ns simulations, we observed that the closed

state conformation already reached the open state at the end of the simulation. Therefore, we conclude that this system is not suitable for our path calculation purpose.

NtrC conformational transition has been studied using the string method with an elastic network model[66]. Besides, TMD has been used to build a transition pathway for the system[154]. The conformational transition for the active and inactive conformation can be described in terms of a combination of helix movement and rotation. The equilibrium simulation was done for 15 ns for each state. We looked at the dihedral angles of the region that has structural difference. Only a few residues have distinct two-state ϕ-ψ distributions while the other residues have largely overlapping distributions or intermediate sampling between the two states. To determine if NtrC is a two-state system, we computed the pairwise rms deviation as defined in [121]. For the 15 ns trajectory, each state has one single peak but the inactive form is forming another small peak from the main peak (Figure A2A). We extended the simulation for another 15 ns. In the pairwise RMSD for the 30 ns combined trajectory, the inactive form is observed to have two states (Figure A2B). We determined the system is not a good two-state model and did not further pursue the conformational transition.

For the third system RXR, the apo form and the agonist bound form were compared. Upon agonist binding, the C-terminus helix partially unfolds and moves to another position to cover the ligand binding pocket. The structural change would be an interesting conformational change to evaluate, however, it is very likely that the extended form of the helix in the apo form is due to crystal stacking, and the physical apo form remains unknown. Due to this observation, we decided to not to use the apo form, and examined the transition between the agonist and antagonist-bound forms of steroid receptor as shown in Chapter 2.

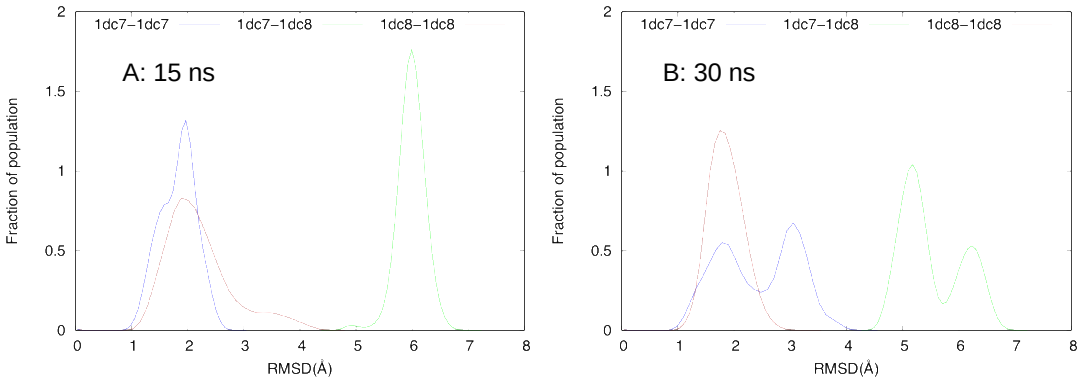


Figure A2 The distributions of pairwise RMSD values calculated using all pairs of snapshots between the two trajectories as labeled. For every pair of snapshot, the RMSD value was computed over all heavy atoms following the superposition of the two protein structures.

APPENDIX B. CHARMM EXECUTABLES BUILD COMMANDS

For community clusters (Halstead as an example) with ABPO support.
Note the version of the software packages might change. Check the available versions using “module avail” before installation.

charmmcompile.sh
#!/bin/csh
module purge
module load intel/16.0.1.150 impi/5.1.2.150
rm -rf ./build/em64t_M ./lib/em64t_M ./exec/em64t_M ./tool/prefx_em64t_M
setenv MPIIFORT YES
./install.com em64t xxlarge FULL M E ABPO X86_64 IFORT

For community clusters (Halstead as example) with DOMDEC support.
charmmcompile.sh
#!/bin/csh
module purge
module load intel/16.0.1.150 impi/5.1.2.150
rm -rf ./build/em64t_M ./lib/em64t_M ./exec/em64t_M ./tool/prefx_em64t_M
setenv MPIIFORT YES
./install.com em64t xxlarge FULL M COLFFT DOMDEC X86_64 IFORT

For local GPU with CHARMM openMM.
charmmcompile.sh
rm -rf ./build/gnu_M ./lib/gnu_M ./exec/gnu_M ./tool/prefx_gnu_M
./install.com gnu openmm M

APPENDIX C. CHARMM INPUT SCRIPTS

B.1 CHARMM input file for D-tensor evaluation for linear distance combination RVs in FACTS

```
* abpo on src with linker active and inactive structure
*

if ?ABPO .ne. 1 then
 echo "test not performed"
 stop
endif

set toppar * ! set your toppar file path here
set io * ! set you io directory path here

! Read topol and param file
read rtf card unit 11 name @toppar/top_all22_prot.rtf
read para card unit 12 name @toppar/par_all22_prot_gbsw.inp
! set up ensemble environment
ensemble nensem 8
set rep ?whoiam
if @rep .lt. 4 then
read psf card unit 13 name @io/1y57.psf
read coor card unit 14 name @io/1y57.cor
coor copy comp
else
read psf card unit 13 name @io/1y57.psf
read coor card unit 14 name @io/2src.cor
coor copy comp
endif
```

```
! NBOND list
set con 10
set coff 12
set cnb 14
set cim 15

!----- NON BONDED OPTIONS ----------------------
nbon inbf -1 elec atom cdie shif vdw vatom vswi -
     eps 1.0 e14f 1.0 wmin 1.5 wrnmxd 0.5 -
     nbxm 5 ctonnb @con ctofnb @coff cutnb @cnb

scalar fbeta set 1.0 sele .not. hydrogen end
energy

scalar wmain = radius

!----- IMPLICIT SOLVENT FACTS -------------------------------------
fact tcps 22 teps 1.0 tkps 8.0 gamm 0.015 -
    conc 0.0 temp 298 tcil @cnb tcic @coff tavw

shake fast bonh para

!setup collective variables
ensemble abpo setcv -
dist NP 8 1 A 311  CA  A 260  NE1 -
-1 A 255  CA  A 308  CA -
-1 A 255  CA  A 311  CA -
-1 A 255  CA  A 312  CA -
-1 A 256  CA  A 311  CA -
-1 A 256  CA  A 312  CA -
-1 A 257  CA  A 311  CA -
```

-1 A 257 CA A 312 CA -
dist NP 6 -1 A 307 CA A 295 CA -
-1 A 307 CA A 296 CA -
-1 A 307 CA A 297 CA -
-1 A 307 CA A 334 CA -
-1 A 307 CA A 335 CA -
-1 A 307 CA A 336 CA -
dist NP 9 -1 A 310 CA A 295 CA -
-1 A 310 CD A 295 CE -
1 A 310 CD A 382 CA -
1 A 310 CA A 382 CA -
1 A 310 CD A 409 CZ -
1 A 310 CA A 410 CA -
-1 A 310 CD A 403 CA -
-1 A 310 CD A 404 CA -
-1 A 311 CA A 325 CA -
dist NP 5 1 A 406 CA A 410 CA -
1 A 407 CA A 410 CA -
1 A 407 CA A 411 CA -
1 A 408 CA A 411 CA -
1 A 409 CA A 412 CA -
dist NP 3 1 A 412 CA A 417 CA -
1 A 413 CA A 416 CA -
1 A 413 CA A 417 CA -
dist NP 6 1 A 414 CA A 417 CA -
1 A 414 CA A 418 CA -
1 A 415 CA A 418 CA -
1 A 415 CA A 419 CA -
1 A 415 CA A 420 CA -
1 A 416 CA A 419 CA -
dist NP 3 1 A 413 CA A 423 CA -

```
 1 A 415  CA  A 423  CA -
 1 A 416  CA  A 424  CA -
dist NP 5 1 A 410  CA  A 302  CA -
 1 A 411  CA  A 278  CA -
 1 A 411  CA  A 301  CA -
 1 A 411  CA  A 302  CA -
 1 A 412  CA  A 278  CA -
dist NP 4 -1 A 410  CA  A 380  CA -
-1 A 410  CA  A 381  CA -
-1 A 410  CA  A 382  CA -
-1 A 411  CA  A 381  CA -
dist NP 4 1 A 416  CA  A 386  CA -
 1 A 416  CA  A 388  CA -
 1 A 416  CA  A 428  CA -
 1 A 417  CA  A 385  CA -
dist NP 5 -1 A 422  CA  A 437  CA -
-1 A 422  CA  A 439  CA -
-1 A 422  CA  A 433  CA -
-1 A 423  CA  A 433  CA -
-1 A 423  CA  A 429  CA -

!evaluate the D matrix
ensemble abpo dtns dsteps 1000000 dfrq 100
dyna leap lang tbath 298 timestep 0.002 -
   FIRSTT 298.0 FINALT 298.0 TSTRUC 298.0 TWINDH 5.0 TWINDL -5.0 -
    nsavc 1000 nprint 1000 IPRFrq 5000

stop
```

B.2 CHARMM input file for initiating ABPO in FACTS

```
* abpo on src with linker active and inactive structure
*

if ?ABPO .ne. 1 then
 echo "test not performed"
 stop
endif

set toppar *
set io *

! Read topol and param file
read rtf card unit 11 name @toppar/top_all22_prot.rtf
read para card unit 12 name @toppar/par_all22_prot_gbsw.inp
! set up ensemble environment
ensemble nensem 16
set rep ?whoiam
if @rep .lt. 8 then
read psf card unit 13 name @io/1y57.psf
read coor card unit 14 name @io/1y57.cor
coor copy comp
else
read psf card unit 13 name @io/1y57.psf
read coor card unit 14 name @io/2src.cor
coor copy comp
endif

! NBOND list

set con 10
```

```
set coff 12
set cnb 14
set cim 15

!----- NON BONDED OPTIONS ----------------------
nbon inbf -1 elec atom cdie shif vdw vatom vswi -
     eps 1.0 e14f 1.0 wmin 1.5 wrnmxd 0.5 -
     nbxm 5 ctonnb @con ctofnb @coff cutnb @cnb

scalar fbeta set 1.0 sele .not. hydrogen end
energy

scalar wmain = radius

!----- IMPLICIT SOLVENT FACTS -----------------------------------
fact tcps 22 teps 1.0 tkps 8.0 gamm 0.015 -
    conc 0.0 temp 298 tcil @cnb tcic @coff tavw

shake fast bonh para

!setup collective variables
ensemble abpo setcv -
dist NP 8 1 A 311  CA  A 260  NE1 -
-1 A 255  CA  A 308  CA -
-1 A 255  CA  A 311  CA -
-1 A 255  CA  A 312  CA -
-1 A 256  CA  A 311  CA -
-1 A 256  CA  A 312  CA -
-1 A 257  CA  A 311  CA -
-1 A 257  CA  A 312  CA -
dist NP 6 -1 A 307  CA  A 295  CA -
```

-1 A 307 CA A 296 CA -
-1 A 307 CA A 297 CA -
-1 A 307 CA A 334 CA -
-1 A 307 CA A 335 CA -
-1 A 307 CA A 336 CA -
dist NP 9 -1 A 310 CA A 295 CA -
-1 A 310 CD A 295 CE -
1 A 310 CD A 382 CA -
1 A 310 CA A 382 CA -
1 A 310 CD A 409 CZ -
1 A 310 CA A 410 CA -
-1 A 310 CD A 403 CA -
-1 A 310 CD A 404 CA -
-1 A 311 CA A 325 CA -
dist NP 5 1 A 406 CA A 410 CA -
1 A 407 CA A 410 CA -
1 A 407 CA A 411 CA -
1 A 408 CA A 411 CA -
1 A 409 CA A 412 CA -
dist NP 3 1 A 412 CA A 417 CA -
1 A 413 CA A 416 CA -
1 A 413 CA A 417 CA -
dist NP 6 1 A 414 CA A 417 CA -
1 A 414 CA A 418 CA -
1 A 415 CA A 418 CA -
1 A 415 CA A 419 CA -
1 A 415 CA A 420 CA -
1 A 416 CA A 419 CA -
dist NP 3 1 A 413 CA A 423 CA -
1 A 415 CA A 423 CA -
1 A 416 CA A 424 CA -

```
dist NP 5 1 A 410  CA  A 302  CA -
 1 A 411  CA  A 278  CA -
 1 A 411  CA  A 301  CA -
 1 A 411  CA  A 302  CA -
 1 A 412  CA  A 278  CA -
dist NP 4 -1 A 410  CA  A 380  CA -
-1 A 410  CA  A 381  CA -
-1 A 410  CA  A 382  CA -
-1 A 411  CA  A 381  CA -
dist NP 4 1 A 416  CA  A 386  CA -
 1 A 416  CA  A 388  CA -
 1 A 416  CA  A 428  CA -
 1 A 417  CA  A 385  CA -
dist NP 5 -1 A 422  CA  A 437  CA -
-1 A 422  CA  A 439  CA -
-1 A 422  CA  A 433  CA -
-1 A 423  CA  A 433  CA -
-1 A 423  CA  A 429  CA -

! Run path optimization
ensemble abpo opti -
  bcyc 1 ecyc 40 temp 298 -
  mnbl 30 bste 40000 minc 200 smoo 0.05 pred 2000 -
  npnt 2000 moll 0.05 rtub 20 ftub 5.0 -
  bfct 0.8  cvfr 1000
! cfct doesn't exist! checked source code
dyna leap lang tbath 298 timestep 0.002 -
  nsavc 1000 nprint 1000 iprfrq 5000 -
  FIRSTT 298.0 FINALT 298.0 TSTRUC 298.0 TWINDH 10.0 TWINDL -10.0 -

stop
```

B.3 CHARMM commands for restarting ABPO

```
! restart from cyc040
! Run path optimization
ensemble abpo opti -
  rest bcyc 40 ecyc 70 temp 298 -
  mnbl 30 bste 40000 minc 200 smoo 0.05 pred 2000 -
  npnt 2000 moll 0.05 rtub 20 ftub 5.0 -
  bfct 0.8  cvfr 1000
! cfct doesn't exist! checked source code
dyna leap lang tbath 298 timestep 0.002 -
  nsavc 1000 nprint 1000 iprfrq 5000 -
  FIRSTT 298.0 FINALT 298.0 TSTRUC 298.0 TWINDH 10.0 TWINDL -10.0 -

stop
```

B.4 CHARMM script for calculating protein-peptide interaction energy time series

```
*FILENAME: interaction-energy.inp
*PURPOSE: compute interaction energies from trajectory
*

!set file directories here
set toppar *
set io *
set ofile *

!read toppar for explicit water systems
read rtf card name @toppar/top_all36_prot.rtf
read para card flex name @toppar/par_all36_prot.prm
stream @toppar/toppar_water_ions.str
read psf card name @io/clus.psf
```

```
open unit 51 read unform name @io/analysis/md1_0.2ns.dcd
! specify how we are going to read the trajectory
traj firstu 51 nunit 1 skip 100000  ! use whole trajectory

open write unit 21 form name @ofile/inte_A_B.dat
write title unit 21
* time prot-ligand
*
set con 8
set coff 10
set cnb 12
set cim 13


set t 0.2 ! keep track of time

label loop
! get next coordinate set according to specifications above
traj read
! we have to update lists every time, things can move a lot in 200ps
update cutim @cim cutnb @cnb ctofnb @coff ctonnb @con cdie shif vdw vato vswi
! protein ligand interaction
inte sele segid A end sele segid B end
set e1 ?ener
write title unit 21
* @t @e1
*

incr t by 0.2
if t le 100 goto loop
```

APPENDIX D. A PRELIMINARY COMPARISON OF EXPLICIT SOLVENT SIMULATIONS BETWEEN CHARMM OPENMM AND DOMDEC

Two types of methods are implemented in CHARMM to get better computational performance of equilibrium simulations in explicit solvent. One is CHARMM openmm that uses GPU for fast calculation of trajectories. The other is CHARMM DOMDEC that uses domain decomposition to update non-bonded list faster. We noticed in Chapter 4 that for the simulations in explicit solvent on GPU, a higher RMSD value is observed for protein systems compared with implicit solvent simulations. Here we compare the simulations of the active form of Src kinase domain in explicit solvent TIP3P using openmm and DOMDEC. The simulations are for 300 ns with openmm mixed precision on GPU, and 30 ns with DOMDEC on the cluster Halstead without GPU support. The DOMDEC simulation uses 64 hours on 80 cores (4 nodes, each with two 10-core Intel Xeon-E5 processors), while the openmm simulation takes ~155 hours. The speedup on GPU is significant while there is no direct comparison for GPU and the Halstead cluster have different hardware infrastructure. The newly developed DOMDEC method incorporates GPU support, and is not compared here. The scalability of openmm is not good on multiple GPUs. Typically, one simulation is performed on a single GPU. DOMDEC scales well to at least 160 cores as we have tested before and would potentially scale on more cores.

The time profiles for protein backbone RMSD values are shown in Figure C1. The DOMDEC simulation has the backbone RMSD increase from 2 to 3 Å during the 30 ns time period. The openmm simulation has the RMSD increase to ~3.5 Å within the first 30 ns and fluctuates around 3.5 Å. From our results, the RMSD in the openmm simulation is slightly higher, however, a longer DOMDEC simulation and more systems need to be examined for a complete comparison.

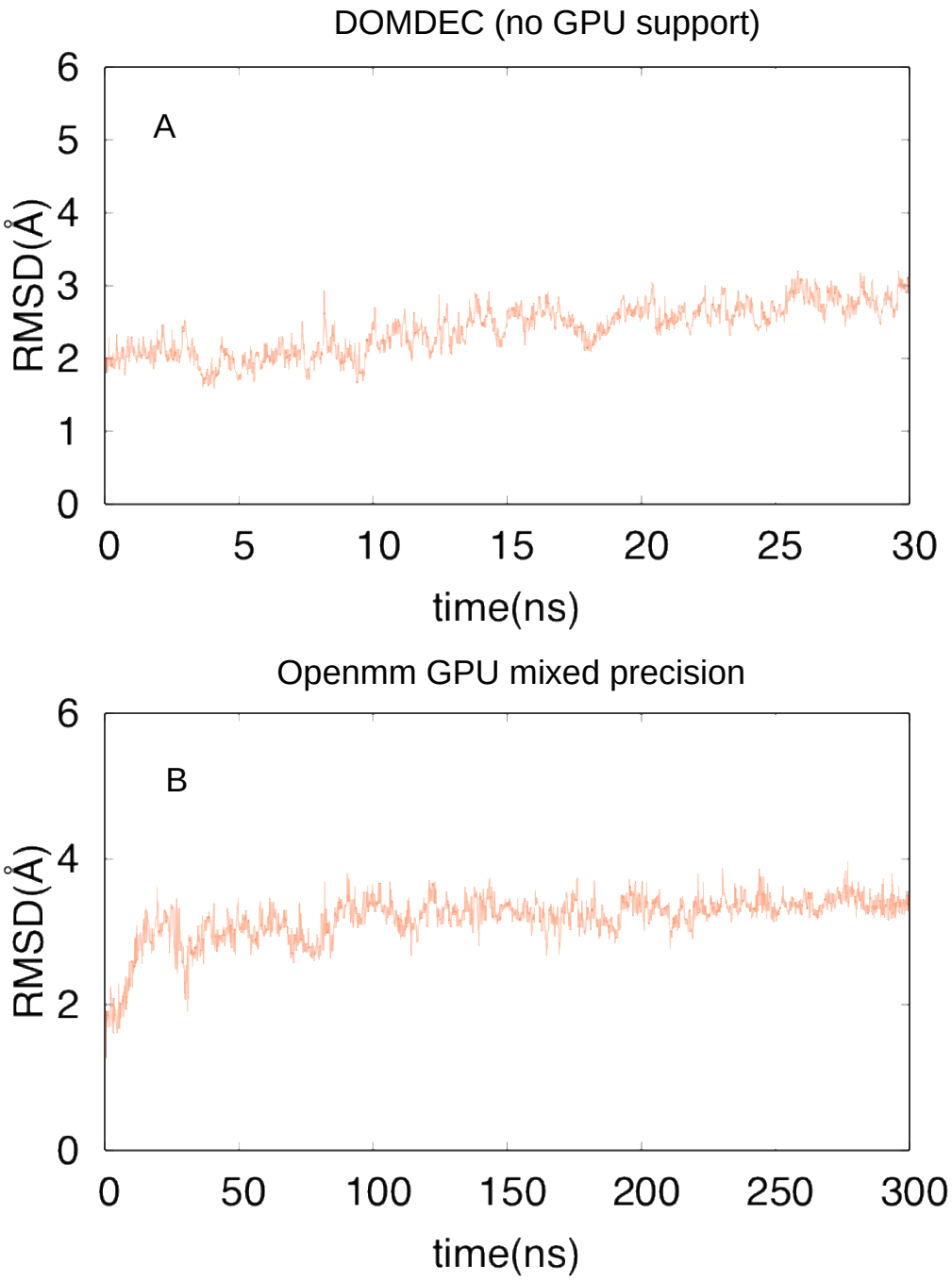


Figure C1 The protein backbone RMSD values for the active conformation of Src kinase domain. A: the simulation with DOMDEC on Halstead cluster. B: the simulation with openmm on GPU.

APPENDIX E. EXTRACTION OF STRUCTURES ALONG THE TRANSITION PATH

To extract structures along the transition path, the frequency to save trajectories (NSAV) and the frequency to save RVs (CVFR) should be set the same. The ABPO module was altered to output the slice indexes corresponding to each frame in block***.cv. The subroutine update_lambda is changed to the following.

```
subroutine update_lambda(p, lambda, dist, global)
! Update lambda to lambda(p), dist to distance(p, path(lambda(p))),
!  use global search if global is True
  real(kind=chm_real), dimension(n_cv + 1) :: p
  integer :: lambda
  real(kind=chm_real) :: dist
  logical :: global
  real(kind=chm_real), dimension(n_cv) :: iD2_p
  integer :: i
  real(kind=chm_real) :: d, d_l, d_r

  i = lambda
  iD2_p = matmul(abpo_iD2, p(1:n_cv))
  if (global .or. i < 1 .or. i > n_point) then
    lambda = 1
    dist = fast_distance(p(1:n_cv), iD2_p, 1)
    do i = 2, n_point
      d = fast_distance(p(1:n_cv), iD2_p, i)
      if (d < dist) then
        lambda = i
        dist = d
      end if
    end do
  else
```

```
      d = fast_distance(p(1:n_cv), iD2_p, i)
      if (i > 1) d_l = fast_distance(p(1:n_cv), iD2_p, i-1)
      if (i == 1 .or. d_l > d) then
        do
          if (i == n_point) exit
          d_r = fast_distance(p(1:n_cv), iD2_p, i+1)
          if (d_r > d) exit
          i = i + 1
          d = d_r
        end do
      else
        do
          if (i == 1) exit
          d_l = fast_distance(p(1:n_cv), iD2_p, i-1)
          if (d_l > d) exit
          i = i - 1
          d = d_l
        end do
      end if
      lambda = i
      dist = d
    end if
    p(n_cv+1)=lambda
  end subroutine update_lambda
```

A Python script (in Python 2) was used to extract structures and write out a trajectory along the transition path. The slice indexes of the extracted frames were also written to a .dat file.

```
#!/usr/bin/env python
import numpy as np
```

```
import MDAnalysis as mda
import os
from time import time
from MDAnalysis.analysis.align import *

rep = 16
block = 4
slicenum = 2000
frame = 30
cv = 12
start_time=time()

# Function to extract slice indexes for a block
def convert(filename):
    datalist = []
    data = np.zeros((frame, cv+1))
    f = open(filename, 'r')
    for line in f:
        temp = line.split()
        for i in temp:
            datalist.append(float(i))
    data = np.array(datalist).reshape(frame, cv+1)
    sliceidx = data[:, cv].astype(int)
    return sliceidx

# iterate the trajectories in all blocks and all reps in the cycle
for i in range(rep):
    for j in range(block):
        psffile = '1y57.psf'
        trajfile = 'rep' + str(i).zfill(3) + '/block' + str(j+1).zfill(3) + '.dcd'
        cvfile = 'rep' + str(i).zfill(3) + '/block' + str(j+1).zfill(3) + '.cv'
```

```
        slice_idx = convert(cvfile)
        u = mda.Universe(psffile, trajfile)
        protein = u.select_atoms('protein')
        m=0
        for ts in u.trajectory:
            tempname = 'slice' + str(slice_idx[m]).zfill(4) + '_temp.dcd'
            if not os.path.isfile(tempname):
                with mda.Writer(tempname, protein.n_atoms) as w:
                    w.write(protein)
            m+=1

# Write out the combined trajectory along the transition path
psffile = '1y57.psf'
ref = mda.Universe('1y57_converted.pdb')
counter = 0
index = []
with mda.Writer('path_full.dcd', n_atoms=4309) as w2:
    for k in range(slicenum):
        trajfile = 'slice' + str(k+1).zfill(4) + '_temp.dcd'
        if os.path.isfile(trajfile):
            counter += 1
            index.append([k+1, counter])
            u2=mda.Universe(psffile, trajfile)
            rms_fit_trj(u2, ref, filename='temp.dcd')
            u3=mda.Universe(psffile, 'temp.dcd')
            for ts in u3.trajectory:
                w2.write(u3.select_atoms('protein'))
            os.remove('temp.dcd')

# Write out the slice indexes for extracted frames
with open('single_struc_index.dat', 'w') as f:
```

```
    for item in index:
        f.write(str(item[0])+'  '+str(item[1])+'\n')
print "Done!(%5.3f seconds)" % (time()-start_time)
```

REFERENCES

[1] J. Schlessinger, “Cell Signaling by Receptor Tyrosine Kinases,” *Cell*, vol. 103, no. 2, pp. 211–225, Oct. 2000.

[2] P. Blume-Jensen and T. Hunter, “Oncogenic kinase signalling,” *Nature*, vol. 411, no. 6835, p. 355, 2001.

[3] M. E. M. Noble, J. A. Endicott, and L. N. Johnson, “Protein Kinase Inhibitors: Insights into Drug Design from Structure,” *Science*, vol. 303, no. 5665, pp. 1800–1805, Mar. 2004.

[4] J. Zhang, P. L. Yang, and N. S. Gray, “Targeting cancer with small molecule kinase inhibitors,” *Nat. Rev. Cancer*, vol. 9, no. 1, p. 28, 2009.

[5] G. Vlahovic and J. Crawford, “Activation of tyrosine kinases in cancer,” *The oncologist*, vol. 8, no. 6, pp. 531–538, 2003.

[6] J. Brognard and T. Hunter, “Protein kinase signaling networks in cancer,” *Curr. Opin. Genet. Dev.*, vol. 21, no. 1, pp. 4–11, 2011.

[7] M. S. Collett and R. L. Erikson, “Protein kinase activity associated with the avian sarcoma virus src gene product,” *Proc. Natl. Acad. Sci.*, vol. 75, no. 4, pp. 2021–2024, Apr. 1978.

[8] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, “The Protein Kinase Complement of the Human Genome,” *Science*, vol. 298, no. 5600, pp. 1912–1934, Dec. 2002.

[9] S. J. Parsons and J. T. Parsons, “Src family kinases, key regulators of signal transduction,” *Oncogene*, vol. 23, no. 48, pp. 7906–7909, Oct. 2004.

[10] T. Erpel and S. A. Courtneidge, “Src family protein tyrosine kinases and cellular signal transduction pathways,” *Curr. Opin. Cell Biol.*, vol. 7, no. 2, pp. 176–182, 1995.

[11] S. M. Thomas and J. S. Brugge, “Cellular functions regulated by Src family kinases,” *Annu. Rev. Cell Dev. Biol.*, vol. 13, no. 1, pp. 513–609, 1997.

[12] L. C. Kim, L. Song, and E. B. Haura, “Src kinases as therapeutic targets for cancer,” *Nat. Rev. Clin. Oncol.*, vol. 6, no. 10, p. 587, 2009.

[13] C. L. Abram and S. A. Courtneidge, “Src Family Tyrosine Kinases and Growth Factor Signaling,” *Exp. Cell Res.*, vol. 254, no. 1, pp. 1–13, Jan. 2000.

[14] D. McGarrigle and X.-Y. Huang, “GPCRs Signaling Directly Through Src-Family Kinases,” *Sci STKE*, vol. 2007, no. 392, pp. pe35–pe35, Jun. 2007.

[15] S. K. Mitra and D. D. Schlaepfer, “Integrin-regulated FAK–Src signaling in normal and cancer cells,” *Curr. Opin. Cell Biol.*, vol. 18, no. 5, pp. 516–523, Oct. 2006.

[16] M. A. Shupnik, “Crosstalk between steroid receptors and the c-Src-receptor tyrosine kinase pathways: implications for cell proliferation,” *Oncogene*, vol. 23, no. 48, pp. 7979–7989, Oct. 2004.

[17] G. Berton, A. Mócsai, and C. A. Lowell, “Src and Syk kinases: key regulators of phagocytic cell activation,” *Trends Immunol.*, vol. 26, no. 4, pp. 208–214, Apr. 2005.

[18] M. C. Maa, T. H. Leu, D. J. McCarley, R. C. Schatzman, and S. J. Parsons, “Potentiation of epidermal growth factor receptor-mediated oncogenesis by c-Src: implications for the etiology of multiple human cancers,” *Proc. Natl. Acad. Sci.*, vol. 92, no. 15, pp. 6981–6985, Jul. 1995.

[19] J. E. Smart, H. Oppermann, A. P. Czernilofsky, A. F. Purchio, R. L. Erikson, and J. M. Bishop, "Characterization of sites for tyrosine phosphorylation in the transforming protein of Rous sarcoma virus (pp60v-src) and its normal cellular homologue (pp60c-src)," *Proc. Natl. Acad. Sci.*, vol. 78, no. 10, pp. 6013–6017, 1981.
[20] T. E. Kmiecik, P. J. Johnson, and D. Shalloway, "Regulation by the autophosphorylation site in overexpressed pp60c-src.," *Mol. Cell. Biol.*, vol. 8, no. 10, pp. 4541–4546, Oct. 1988.
[21] R. J. Boerner, D. B. Kassel, S. C. Barker, B. Ellis, P. DeLacy, and W. B. Knight, "Correlation of the Phosphorylation States of pp60c-src with Tyrosine Kinase Activity: The Intramolecular pY530−SH2 Complex Retains Significant Activity If Y419 Is Phosphorylated," *Biochemistry*, vol. 35, no. 29, pp. 9519–9525, Jan. 1996.
[22] J. A. Cooper, K. L. Gould, C. A. Cartwright, and T. Hunter, "Tyr527 is phosphorylated in pp60c-src: implications for regulation," *Science*, vol. 231, no. 4744, pp. 1431–1434, Mar. 1986.
[23] M. Okada and H. Nakagawa, "A protein tyrosine kinase involved in regulation of pp60c-src function.," *J. Biol. Chem.*, vol. 264, no. 35, pp. 20886–20893, Dec. 1989.
[24] M. P. Lutz *et al.*, "Overexpression and activation of the tyrosine kinase Src in human pancreatic carcinoma," *Biochem. Biophys. Res. Commun.*, vol. 243, no. 2, pp. 503–508, 1998.
[25] R. H. Alvarez, H. M. Kantarjian, and J. E. Cortes, "The role of Src in solid and hematologic malignancies," *Cancer*, vol. 107, no. 8, pp. 1918–1929, 2006.
[26]R. B. Irby and T. J. Yeatman, "Role of Src expression and activation in human cancer," *Oncogene*, vol. 19, no. 49, p. 5636, 2000.
[27] J. M. Summy and G. E. Gallick, "Src family kinases in tumor progression and metastasis," *Cancer Metastasis Rev.*, vol. 22, no. 4, pp. 337–358, 2003.
[28] M. L. Hibbs *et al.*, "Multiple defects in the immune system of Lyn-deficient mice, culminating in autoimmune disease," *Cell*, vol. 83, no. 2, pp. 301–311, Oct. 1995.
[29] C. A. Lowell, "Src-family kinases: rheostats of immune cell signaling," *Mol. Immunol.*, vol. 41, no. 6, pp. 631–643, Jul. 2004.
[30] B. J. Druker *et al.*, "Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr–Abl positive cells," *Nat. Med.*, vol. 2, no. 5, pp. 561–566, May 1996.
[31] R. Capdeville, E. Buchdunger, J. Zimmermann, and A. Matter, "Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug," *Nat. Rev. Drug Discov.*, vol. 1, no. 7, pp. 493–502, Jul. 2002.
[32] A. A. Wylie *et al.*, "The allosteric inhibitor ABL001 enables dual targeting of BCR–ABL1," *Nature*, vol. 543, no. 7647, pp. 733–737, Mar. 2017.
[33] J. Schoepfer *et al.*, "Discovery of Asciminib (ABL001), an Allosteric Inhibitor of the Tyrosine Kinase Activity of BCR-ABL1," *J. Med. Chem.*, Aug. 2018.
[34] M. Mauro *et al.*, "Oral Asciminib (ABL001) vs Bosutinib in Patients With Chronic Myeloid Leukemia in Chronic Phase (CML-CP) Who Received ≥ 2 Prior Tyrosine Kinase Inhibitors (TKIs): A Multicenter, Open-Label, Randomized, Phase 3 Study," *Clin. Lymphoma Myeloma Leuk.*, vol. 17, pp. S315–S316, Sep. 2017.
[35] M. G. Fury *et al.*, "Phase II Study of Saracatinib (AZD0530) for Patients with Recurrent or Metastatic Head and Neck Squamous Cell Carcinoma (HNSCC)," *Anticancer Res.*, vol. 31, no. 1, pp. 249–253, Jan. 2011.

[36] A. Gucalp *et al.*, “Phase II Trial of Saracatinib (AZD0530), an Oral SRC-inhibitor for the Treatment of Patients with Hormone Receptor-negative Metastatic Breast Cancer,” *Clin. Breast Cancer*, vol. 11, no. 5, pp. 306–311, Oct. 2011.
[37] H. J. Mackay *et al.*, “A phase II trial of the Src kinase inhibitor saracatinib (AZD0530) in patients with metastatic or locally advanced gastric or gastro esophageal junction (GEJ) adenocarcinoma: a trial of the PMH phase II consortium,” *Invest. New Drugs*, vol. 30, no. 3, pp. 1158–1163, Jun. 2012.
[38] T. C. Gangadhar, J. I. Clark, T. Karrison, and T. F. Gajewski, “Phase II study of the Src kinase inhibitor saracatinib (AZD0530) in metastatic melanoma,” *Invest. New Drugs*, vol. 31, no. 3, pp. 769–773, Jun. 2013.
[39] J. E. Cortes *et al.*, “Bosutinib Versus Imatinib in Newly Diagnosed Chronic-Phase Chronic Myeloid Leukemia: Results From the BELA Trial,” *J. Clin. Oncol.*, vol. 30, no. 28, pp. 3486–3492, Oct. 2012.
[40] T. H. Brümmendorf *et al.*, “Bosutinib versus imatinib in newly diagnosed chronic-phase chronic myeloid leukaemia: results from the 24-month follow-up of the BELA trial,” *Br. J. Haematol.*, vol. 168, no. 1, pp. 69–81, Jan. 2015.
[41] J. E. Cortes *et al.*, “Bosutinib Versus Imatinib for Newly Diagnosed Chronic Myeloid Leukemia: Results From the Randomized BFORE Trial,” *J. Clin. Oncol.*, vol. 36, no. 3, pp. 231–237, Jan. 2018.
[42] B. J. Mayer, M. Hamaguchi, and H. Hanafusa, “A novel viral oncogene with structural similarity to phospholipase C,” *Nature*, vol. 332, no. 6161, p. 272, 1988.
[43] I. Sadowski, J. C. Stone, and T. Pawson, “A noncatalytic domain conserved among cytoplasmic protein-tyrosine kinases modifies the kinase function and transforming activity of Fujinami sarcoma virus P130gag-fps,” *Mol. Cell. Biol.*, vol. 6, no. 12, pp. 4396–4408, Dec. 1986.
[44] M. T. Brown and J. A. Cooper, “Regulation, substrates and functions of src,” *Biochim. Biophys. Acta BBA - Rev. Cancer*, vol. 1287, no. 2, pp. 121–149, Jun. 1996.
[45] S. A. Courtneidge, A. D. Levinson, and J. M. Bishop, “The protein encoded by the transforming gene of avian sarcoma virus (pp60src) and a homologous protein in normal cells (pp60proto-src) are associated with the plasma membrane,” *Proc. Natl. Acad. Sci.*, vol. 77, no. 7, pp. 3783–3787, Jul. 1980.
[46] M. D. Resh, “Myristylation and palmitylation of Src family members: the fats of the matter,” *Cell*, vol. 76, no. 3, pp. 411–413, 1994.
[47] R. Ren, B. J. Mayer, P. Cicchetti, and D. Baltimore, “Identification of a ten-amino acid proline-rich SH3 binding site,” *Science*, vol. 259, no. 5098, pp. 1157–1161, Feb. 1993.
[48] J. Kuriyan and D. Cowburn, “Modular Peptide Recognition Domains in Eukaryotic Signaling,” *Annu. Rev. Biophys. Biomol. Struct.*, vol. 26, no. 1, pp. 259–288, 1997.
[49] S. S. Taylor and A. P. Kornev, “Protein kinases: evolution of dynamic regulatory proteins,” *Trends Biochem. Sci.*, vol. 36, no. 2, pp. 65–77, 2011.
[50] E. Ozkirimli and C. B. Post, “Src kinase activation: A switched electrostatic network,” *Protein Sci.*, vol. 15, no. 5, pp. 1051–1062, May 2006.
[51] E. Ozkirimli, S. S. Yadav, W. T. Miller, and C. B. Post, “An electrostatic network and long-range regulation of Src kinases,” *Protein Sci.*, vol. 17, no. 11, pp. 1871–1880, 2008.

[52] M. Karplus and J. A. McCammon, “Dynamics of proteins: elements and function,” *Annu. Rev. Biochem.*, vol. 52, no. 1, pp. 263–300, 1983.
[53] N. Sinha and S. J. Smith-Gill, “Protein structure to function via dynamics,” *Protein Pept. Lett.*, vol. 9, no. 5, pp. 367–377, 2002.
[54] M. Karplus and J. Kuriyan, “Molecular dynamics and protein function,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 19, pp. 6679–6685, May 2005.
[55] R. C. Bernardi, M. C. R. Melo, and K. Schulten, “Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems,” *Biochim. Biophys. Acta*, vol. 1850, no. 5, pp. 872–877, May 2015.
[56] P. Tiwary and A. van de Walle, “A Review of Enhanced Sampling Approaches for Accelerated Molecular Dynamics,” in *Multiscale Materials Modeling for Nanomechanics*, Springer, 2016, pp. 195–221.
[57] R. Elber, “Perspective: Computer simulations of long time dynamics,” *J. Chem. Phys.*, vol. 144, no. 6, p. 060901, Feb. 2016.
[58] A. Laio and M. Parrinello, “Escaping free-energy minima,” *Proc. Natl. Acad. Sci.*, vol. 99, no. 20, pp. 12562–12566, Oct. 2002.
[59] V. Leone, F. Marinelli, P. Carloni, and M. Parrinello, “Targeting biomolecular flexibility with metadynamics,” *Curr. Opin. Struct. Biol.*, vol. 20, no. 2, pp. 148–154, Apr. 2010.
[60] B. M. Dickson, “Survey of adaptive biasing potentials: comparisons and outlook,” *Curr. Opin. Struct. Biol.*, vol. 43, pp. 63–67, Apr. 2017.
[61] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, “Adaptive biasing force method for scalar and vector free energy calculations,” *J. Chem. Phys.*, vol. 128, no. 14, p. 144120, Apr. 2008.
[62] E. Darve and A. Pohorille, “Calculating free energies using average force,” *J. Chem. Phys.*, vol. 115, no. 20, pp. 9169–9183, Nov. 2001.
[63] A. K. Faradjian and R. Elber, “Computing time scales from reaction coordinates by milestoning,” *J. Chem. Phys.*, vol. 120, no. 23, pp. 10880–10889, May 2004.
[64] D. Hamelberg, J. Mongan, and J. A. McCammon, “Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules,” *J. Chem. Phys.*, vol. 120, no. 24, pp. 11919–11929, Jun. 2004.
[65] W. E, W. Ren, and E. Vanden-Eijnden, “Finite Temperature String Method for the Study of Rare Events,” *J. Phys. Chem. B*, vol. 109, no. 14, pp. 6688–6693, Apr. 2005.
[66] A. C. Pan, D. Sezer, and B. Roux, “Finding transition pathways using the string method with swarms of trajectories,” *J. Phys. Chem. B*, vol. 112, no. 11, pp. 3432–3440, 2008.
[67] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, “String method in collective variables: minimum free energy paths and isocommittor surfaces,” *J. Chem. Phys.*, vol. 125, no. 2, p. 024106, 2006.
[68] R. Zhao, J. Shen, and R. D. Skeel, “Maximum Flux Transition Paths of Conformational Change,” *J. Chem. Theory Comput.*, vol. 6, no. 8, pp. 2411–2423, Aug. 2010.
[69] Y. Shan, A. Arkhipov, E. T. Kim, A. C. Pan, and D. E. Shaw, “Transitions to catalytically inactive conformations in EGFR kinase,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 18, pp. 7270–7275, 2013.

[70] Z. H. Foda, Y. Shan, E. T. Kim, D. E. Shaw, and M. A. Seeliger, "A dynamically coupled allosteric network underlies binding cooperativity in Src kinase," *Nat. Commun.*, vol. 6, 2015.
[71] W. Gan, S. Yang, and B. Roux, "Atomistic View of the Conformational Activation of Src Kinase Using the String Method with Swarms-of-Trajectories," *Biophys. J.*, vol. 97, no. 4, pp. L8–L10, Aug. 2009.
[72] Y. Meng, Y. Lin, and B. Roux, "Computational study of the 'DFG-flip' conformational transition in c-Abl and c-Src tyrosine kinases," *J. Phys. Chem. B*, vol. 119, no. 4, pp. 1443–1456, 2015.
[73] D. Shukla, Y. Meng, B. Roux, and V. S. Pande, "Activation pathway of Src kinase reveals intermediate states as targets for drug design," *Nat. Commun.*, vol. 5, 2014.
[74] Y. Meng, D. Shukla, V. S. Pande, and B. Roux, "Transition path theory analysis of c-Src kinase activation," *Proc. Natl. Acad. Sci.*, vol. 113, no. 33, pp. 9193–9198, Aug. 2016.
[75] Y. Meng and B. Roux, "Locking the Active Conformation of c-Src Kinase through the Phosphorylation of the Activation Loop," *J. Mol. Biol.*, vol. 426, no. 2, pp. 423–435, Jan. 2014.
[76] S. Yang, N. K. Banavali, and B. Roux, "Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories," *Proc. Natl. Acad. Sci.*, vol. 106, no. 10, pp. 3776–3781, 2009.
[77] Y. Meng and B. Roux, "Computational study of the W260A activating mutant of Src tyrosine kinase," *Protein Sci.*, vol. 25, no. 1, pp. 219–230, 2016.
[78] M. Fajer, Y. Meng, and B. Roux, "The Activation of c-Src Tyrosine Kinase: Conformational Transition Pathway and Free Energy Landscape," *J. Phys. Chem. B*, Oct. 2016.
[79] H. Huang, R. Zhao, B. M. Dickson, R. D. Skeel, and C. B. Post, "αC Helix as a Switch in the Conformational Transition of Src/CDK-like Kinase Domains," *J. Phys. Chem. B*, vol. 116, no. 15, pp. 4465–4475, Apr. 2012.
[80] D. Shukla, Y. Meng, B. Roux, and V. S. Pande, "Activation pathway of Src kinase reveals intermediate states as targets for drug design," *Nat. Commun.*, vol. 5, p. 3397, Mar. 2014.
[81] B. M. Dickson, H. Huang, and C. B. Post, "Unrestrained computation of free energy along a path," *J. Phys. Chem. B*, vol. 116, no. 36, pp. 11046–11055, 2012.
[82] E. Vanden-Eijnden and M. Venturoli, "Revisiting the finite temperature string method for the calculation of reaction tubes and free energies," *J. Chem. Phys.*, vol. 130, no. 19, p. 05B605, 2009.
[83] B. M. Dickson, "Approaching a parameter-free metadynamics," *Phys. Rev. E*, vol. 84, no. 3, p. 037701, Sep. 2011.
[84] B. M. Dickson, "μ-tempered metadynamics: Artifact independent convergence times for wide hills," *J. Chem. Phys.*, vol. 143, no. 23, p. 234109, Dec. 2015.
[85] J. F. Dama, M. Parrinello, and G. A. Voth, "Well-Tempered Metadynamics Converges Asymptotically," *Phys. Rev. Lett.*, vol. 112, no. 24, p. 240602, Jun. 2014.
[86] A. Barducci, G. Bussi, and M. Parrinello, "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method," *Phys. Rev. Lett.*, vol. 100, no. 2, p. 020603, Jan. 2008.

[87] V. Ovchinnikov, M. Karplus, and E. Vanden-Eijnden, "Free energy of conformational transition paths in biomolecules: The string method and its application to myosin VI," *J. Chem. Phys.*, vol. 134, no. 8, p. 085103, Feb. 2011.
[88] N. Gō and T. Noguti, "Collective variable description of small-amplitude conformational fluctuations in a globular protein," *Nature*, vol. 296, no. 5859, p. 776, Apr. 1982.
[89]S. Hayward and N. Go, "Collective Variable Description of Native Protein Dynamics," *Annu. Rev. Phys. Chem.*, vol. 46, no. 1, pp. 223–250, 1995.
[90] R. Elber, "Calculation of the potential of mean force using molecular dynamics with linear constraints: An application to a conformational transition in a solvated dipeptide," *J. Chem. Phys.*, vol. 93, no. 6, pp. 4312–4321, Sep. 1990.
[91] K. Jug, N. N. Nair, and T. Bredow, "Molecular dynamics investigation of oxygen vacancy diffusion in rutile," *Phys. Chem. Chem. Phys.*, vol. 7, no. 13, pp. 2616–2621, 2005.
[92] L. Maragliano and E. Vanden-Eijnden, "On-the-fly string method for minimum free energy paths calculation," *Chem. Phys. Lett.*, vol. 446, no. 1, pp. 182–190, 2007.
[93] C. Clementi, H. Nymeyer, and J. N. Onuchic, "Topological and energetic factors: what determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? an investigation for small globular proteins11Edited by F. E. Cohen," *J. Mol. Biol.*, vol. 298, no. 5, pp. 937–953, 2000.
[94] J. Karanicolas and C. L. Brooks, "The origins of asymmetry in the folding transition states of protein L and protein G," *Protein Sci.*, vol. 11, no. 10, pp. 2351–2361, 2002.
[95] Z. Zhang *et al.*, "Crystal Structure of Recombinant Chicken Triosephosphate Isomerase-Phosphoglycolohydroxamate Complex at 1.8-. ANG. Resolution," *Biochemistry*, vol. 33, no. 10, pp. 2830–2837, 1994.
[96] M. R. Sawaya and J. Kraut, "Loop and subdomain movements in the mechanism of Escherichia coli dihydrofolate reductase: crystallographic evidence," *Biochemistry*, vol. 36, no. 3, pp. 586–603, 1997.
[97] A. K. Shiau *et al.*, "The Structural Basis of Estrogen Receptor/Coactivator Recognition and the Antagonism of This Interaction by Tamoxifen," *Cell*, vol. 95, no. 7, pp. 927–937, Dec. 1998.
[98] M. E. Noble, J. P. Zeelen, and R. K. Wierenga, "Structures of the 'open' and 'closed' state of trypanosomal triosephosphate isomerase, as observed in a new crystal form: implications for the reaction mechanism," *Proteins Struct. Funct. Bioinforma.*, vol. 16, no. 4, pp. 311–326, 1993.
[99] V. Pareek, M. Samanta, N. V. Joshi, H. Balaram, M. Murthy, and P. Balaram, "Connecting Active-Site Loop Conformations and Catalysis in Triosephosphate Isomerase: Insights from a Rare Variation at Residue 96 in the Plasmodial Enzyme," *ChemBioChem*, 2016.
[100] A. R. Katebi and R. L. Jernigan, "The critical role of the loops of triosephosphate isomerase for its oligomerization, dynamics, and functionality," *Protein Sci.*, vol. 23, no. 2, pp. 213–228, 2014.
[101] M. J. Osborne, J. Schnell, S. J. Benkovic, H. J. Dyson, and P. E. Wright, "Backbone Dynamics in Dihydrofolate Reductase Complexes: Role of Loop Flexibility in the Catalytic Mechanism," *Biochemistry*, vol. 40, no. 33, pp. 9846–9859, Aug. 2001.

[102] "A Unified Nomenclature System for the Nuclear Receptor Superfamily," *Cell*, vol. 97, no. 2, pp. 161–163, Apr. 1999.
[103] P. Germain, L. Altucci, W. Bourguet, C. Rochette-Egly, and H. Gronemeyer, "Nuclear receptor superfamily: principles of signaling," *Pure Appl. Chem.*, vol. 75, no. 11–12, pp. 1619–1664, 2003.
[104] D. J. Kojetin and T. P. Burris, "Small molecule modulation of nuclear receptor conformational dynamics: implications for function and drug discovery," *Mol. Pharmacol.*, vol. 83, no. 1, pp. 1–8, 2013.
[105] B. A. Johnson, E. M. Wilson, Y. Li, D. E. Moller, R. G. Smith, and G. Zhou, "Ligand-induced stabilization of PPARγ monitored by NMR spectroscopy: implications for nuclear receptor activation1," *J. Mol. Biol.*, vol. 298, no. 2, pp. 187–194, 2000.
[106] J. Lu, D. P. Cistola, and E. Li, "Analysis of Ligand Binding and Protein Dynamics of Human Retinoid X Receptor Alpha Ligand-Binding Domain by Nuclear Magnetic Resonance," *Biochemistry*, vol. 45, no. 6, pp. 1629–1639, Feb. 2006.
[107] T. S. Hughes *et al.*, "Ligand and receptor dynamics contribute to the mechanism of graded PPARγ agonism," *Structure*, vol. 20, no. 1, pp. 139–150, 2012.
[108] B. R. Brooks *et al.*, "CHARMM: the biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–1614, 2009.
[109] A. D. MacKerell Jr *et al.*, "All-atom empirical potential for molecular modeling and dynamics studies of proteins†," *J. Phys. Chem. B*, vol. 102, no. 18, pp. 3586–3616, 1998.
[110] A. D. MacKerell Jr, M. Feig, and C. L. Brooks, "Improved treatment of the protein backbone in empirical force fields," *J. Am. Chem. Soc.*, vol. 126, no. 3, pp. 698–699, 2003.
[111] U. Haberthür and A. Caflisch, "FACTS: Fast analytical continuum treatment of solvation," *J. Comput. Chem.*, vol. 29, no. 5, pp. 701–715, 2008.
[112] D. P. Hua, H. Huang, A. Roy, and C. B. Post, "Evaluating the dynamics and electrostatic interactions of folded proteins in implicit solvents," *Protein Sci.*, vol. 25, no. 1, pp. 204–218, Jan. 2016.
[113] M. Gangloff, M. Ruff, S. Eiler, S. Duclaud, J. M. Wurtz, and D. Moras, "Crystal structure of a mutant hERα ligand-binding domain reveals key structural features for the mechanism of partial agonism," *J. Biol. Chem.*, vol. 276, no. 18, pp. 15059–15065, 2001.
[114] W. Xu, S. C. Harrison, and M. J. Eck, "Three-dimensional structure of the tyrosine kinase c-Src," *Nature*, vol. 385, no. 6617, p. 595, 1997.
[115] J. C. Williams *et al.*, "The 2.35 å crystal structure of the inactivated form of chicken src: a dynamic molecule with multiple regulatory interactions11Edited by R. Huber," *J. Mol. Biol.*, vol. 274, no. 5, pp. 757–775, Dec. 1997.
[116] W. Xu, A. Doshi, M. Lei, M. J. Eck, and S. C. Harrison, "Crystal structures of c-Src reveal features of its autoinhibitory mechanism," *Mol. Cell*, vol. 3, no. 5, pp. 629–638, 1999.
[117] S. W. Cowan-Jacob *et al.*, "The crystal structure of a c-Src complex in an active conformation suggests possible steps in c-Src activation," *Structure*, vol. 13, no. 6, pp. 861–871, 2005.

[118] M. Overduin, B. Mayer, C. B. Rios, D. Baltimore, and D. Cowburn, "Secondary structure of Src homology 2 domain of c-Abl by heteronuclear NMR spectroscopy in solution," *Proc. Natl. Acad. Sci.*, vol. 89, no. 24, pp. 11673–11677, Dec. 1992.

[119] N. A. Farrow *et al.*, "Backbone dynamics of a free and a phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation," *Biochemistry*, vol. 33, no. 19, pp. 5984–6003, 1994.

[120] T. S. Ulmer, J. M. Werner, and I. D. Campbell, "SH3-SH2 Domain Orientation in Src Kinases: NMR Studies of Fyn," *Structure*, vol. 10, no. 7, pp. 901–911, Jul. 2002.

[121] J. D. Taylor, P. J. Gilbert, M. A. Williams, W. R. Pitt, and J. E. Ladbury, "Identification of novel fragment compounds targeted against the pY pocket of v-Src SH2 by computational and NMR screening and thermodynamic evaluation," *Proteins Struct. Funct. Bioinforma.*, vol. 67, no. 4, pp. 981–990, Jun. 2007.

[122] I. Amata *et al.*, "Multi-phosphorylation of the Intrinsically Disordered Unique Domain of c-Src Studied by In-Cell and Real-Time NMR Spectroscopy," *ChemBioChem*, vol. 14, no. 14, pp. 1820–1827, Sep. 2013.

[123] M. Tong, J. G. Pelton, M. L. Gill, W. Zhang, F. Picart, and M. A. Seeliger, "Survey of solution dynamics in Src kinase reveals allosteric cross talk between the ligand binding and regulatory sites," *Nat. Commun.*, vol. 8, no. 1, p. 2160, Dec. 2017.

[124] M. Azam, M. A. Seeliger, N. S. Gray, J. Kuriyan, and G. Q. Daley, "Activation of tyrosine kinases by mutation of the gatekeeper threonine," *Nat. Struct. Mol. Biol.*, vol. 15, no. 10, p. 1109, 2008.

[125] S. Yang and B. Roux, "Src Kinase Conformational Activation: Thermodynamics, Pathways, and Mechanisms," *PLOS Comput. Biol.*, vol. 4, no. 3, p. e1000047, Mar. 2008.

[126] Z. H. Foda, Y. Shan, E. T. Kim, D. E. Shaw, and M. A. Seeliger, "A dynamically coupled allosteric network underlies binding cooperativity in Src kinase," *Nat. Commun.*, vol. 6, p. 5939, 2015.

[127] S. Gonfloni, J. C. Williams, K. Hattula, A. Weijland, R. K. Wierenga, and G. Superti-Furga, "The role of the linker between the SH2 domain and catalytic domain in the regulation and function of Src," *EMBO J.*, vol. 16, no. 24, pp. 7261–7271, 1997.

[128] M. LaFevre-Bernt, F. Sicheri, A. Pico, M. Porter, J. Kuriyan, and W. T. Miller, "Intramolecular regulatory interactions in the Src family kinase Hck probed by mutagenesis of a conserved tryptophan residue," *J. Biol. Chem.*, vol. 273, no. 48, pp. 32129–32134, 1998.

[129] S. Gonfloni, F. Frischknecht, M. Way, and G. Superti-Furga, "Leucine 255 of Src couples intramolecular interactions to inhibition of catalysis," *Nat. Struct. Mol. Biol.*, vol. 6, no. 8, p. 760, 1999.

[130] S. Gonfloni, A. Weijland, J. Kretzschmar, and G. Superti-Furga, "Crosstalk between the catalytic and regulatory domains allows bidirectional regulation of Src," *Nat. Struct. Mol. Biol.*, vol. 7, no. 4, p. 281, 2000.

[131] D. R. Knighton *et al.*, "Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase," *Science*, vol. 253, no. 5018, pp. 407–414, 1991.

[132] E. D. Lowe, M. E. M. Noble, V. T. Skamnaki, N. G. Oikonomakos, D. J. Owen, and L. N. Johnson, "The crystal structure of a phosphorylase kinase peptide substrate

complex: kinase substrate recognition," *EMBO J.*, vol. 16, no. 22, pp. 6646–6658, 1997.

[133] N. R. Brown, M. E. Noble, J. A. Endicott, and L. N. Johnson, "The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases," *Nat. Cell Biol.*, vol. 1, no. 7, p. 438, 1999.

[134] J. Yang, P. Cron, V. M. Good, V. Thompson, B. A. Hemmings, and D. Barford, "Crystal structure of an activated Akt/protein kinase B ternary complex with GSK3-peptide and AMP-PNP," *Nat. Struct. Mol. Biol.*, vol. 9, no. 12, p. 940, 2002.

[135] A. N. Bullock, J. Debreczeni, A. Amos, S. Knapp, and B. E. Turk, "Structure and substrate specificity of the Pim-1 kinase," *J. Biol. Chem.*, 2005.

[136] S. Favelyukis, J. H. Till, S. R. Hubbard, and W. T. Miller, "Structure and autoregulation of the insulin-like growth factor 1 receptor kinase," *Nat. Struct. Mol. Biol.*, vol. 8, no. 12, pp. 1058–1063, Dec. 2001.

[137] H. Chen *et al.*, "A molecular brake in the kinase hinge region regulates the activity of receptor tyrosine kinases," *Mol. Cell*, vol. 27, no. 5, pp. 717–730, 2007.

[138] T. L. Davis, J. R. Walker, A. Allali-Hassani, S. A. Parker, B. E. Turk, and S. Dhe-Paganon, "Structural recognition of an optimized substrate for the ephrin family of receptor tyrosine kinases," *FEBS J.*, vol. 276, no. 16, pp. 4395–4404, 2009.

[139] M. J. Begley *et al.*, "EGF-receptor specificity for phosphotyrosine-primed substrates provides signal integration with Src," *Nat. Struct. Mol. Biol.*, vol. 22, no. 12, p. 983, 2015.

[140] M. Senften, G. Schenker, J. M. Sowadski, and K. Ballmer-Hofer, "Catalytic activity and transformation potential of v-Src require arginine 385 in the substrate binding pocket.," *Oncogene*, vol. 10, no. 1, pp. 199–203, Jan. 1995.

[141] B. S. Gaul, M. L. Harrison, R. L. Geahlen, R. A. Burton, and C. B. Post, "Substrate Recognition by the Lyn Protein-tyrosine Kinase NMR STRUCTURE OF THE IMMUNORECEPTOR TYROSINE-BASED ACTIVATION MOTIF SIGNALING REGION OF THE B CELL ANTIGEN RECEPTOR," *J. Biol. Chem.*, vol. 275, no. 21, pp. 16174–16182, May 2000.

[142] Z. Songyang *et al.*, "Catalytic specificity of protein-tyrosine kinases is critical for selective signalling," *Nature*, vol. 373, no. 6514, p. 536, 1995.

[143] J. Zheng *et al.*, "2.2 \AA refined crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MnATP and a peptide inhibitor," *Acta Crystallogr. D Biol. Crystallogr.*, vol. 49, no. 3, pp. 362–365, 1993.

[144] S. R. Hubbard, "Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog," *EMBO J.*, vol. 16, no. 18, pp. 5572–5581, Sep. 1997.

[145] Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger, and D. E. Shaw, "How Does a Drug Molecule Find Its Target Binding Site?," *J. Am. Chem. Soc.*, vol. 133, no. 24, pp. 9181–9183, Jun. 2011.

[146] R. S. Bohacek *et al.*, "X-ray structure of citrate bound to Src SH2 leads to a high-affinity, bone-targeted Src SH2 inhibitor," *J. Med. Chem.*, vol. 44, no. 5, pp. 660–663, 2001.

[147] N.-H. Nam, G. Ye, G. Sun, and K. Parang, "Conformationally constrained peptide analogues of pTyr-Glu-Glu-Ile as inhibitors of the Src SH2 domain binding," *J. Med. Chem.*, vol. 47, no. 12, pp. 3131–3141, 2004.

[148] D. Kraskouskaya, E. Duodu, C. C. Arpin, and P. T. Gunning, “Progress towards the development of SH2 domain inhibitors,” *Chem. Soc. Rev.*, vol. 42, no. 8, pp. 3337–3370, 2013.

[149] J. P. Davidson, O. Lubman, T. Rose, G. Waksman, and S. F. Martin, “Calorimetric and structural studies of 1, 2, 3-trisubstituted cyclopropanes as conformationally constrained peptide inhibitors of Src SH2 domain binding,” *J. Am. Chem. Soc.*, vol. 124, no. 2, pp. 205–215, 2002.

[150] A.-P. Hynninen and M. F. Crowley, “New faster CHARMM molecular dynamics engine,” *J. Comput. Chem.*, vol. 35, no. 5, pp. 406–413, Feb. 2014.

[151] P. Eastman and V. S. Pande, “OpenMM: A Hardware Independent Framework for Molecular Simulations,” *Comput. Sci. Eng.*, vol. 12, no. 4, pp. 34–39, Jul. 2015.

[152] P. Eastman *et al.*, “OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation,” *J. Chem. Theory Comput.*, vol. 9, no. 1, pp. 461–469, Jan. 2013.

[153] P. Eastman *et al.*, “OpenMM 7: Rapid development of high performance algorithms for molecular dynamics,” *PLOS Comput. Biol.*, vol. 13, no. 7, p. e1005659, Jul. 2017.

[154] M. Lei, J. Velos, A. Gardino, A. Kivenson, M. Karplus, and D. Kern, “Segmented Transition Pathway of the Signaling Protein Nitrogen Regulatory Protein C,” *J. Mol. Biol.*, vol. 392, no. 3, pp. 823–836, Sep. 2009.

VITA

Heng Wu was born in Nantong, Jiangsu province, China. After completing high school at Nantong Middle School, she attended East China Normal University, Shanghai, China and received her B.S. in Biological Sciences in July, 2012. She started graduate studies in the Purdue University Life Sciences interdisciplinary program (PULSe) in August, 2012. She joined the lab of Dr. Post in May, 2013.