MULTI-SCALE SIMULATIONS OF INTRINSICALLY DISORDERED PROTEINS AND DEVELOPMENT OF ENHANCED SAMPLING TECHNIQUES

by

WEIHONG ZHANG

B.S., Beijing University of Chemical Technology, 2009

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Biochemistry and Molecular Biophysics Graduate Group

Kansas State University
Manhattan, Kansas

2014

ABSTRACT

Intrinsically disordered proteins (IDPs) are functional proteins that lack stable tertiary structures under physiological conditions. IDPs are key components of regulatory networks that dictate various aspects of cellular decision-making, and are over-represented in major disease pathways. For example, about 30% of eukaryotic proteins contain intrinsic disordered regions, and over 70% of cancer-associated proteins have been identified as IDPs. The highly heterogeneous nature of IDPs has presented significant challenge for experimental characterization using NMR, X-ray crystallography, or FRET. These challenges represent a unique opportunity for molecular modeling to make critical contributions. In this study, computer simulations at multiple scales were utilized to characterize the structural properties of unbound IDPs as well as to obtain a mechanistic understanding of IDP interactions. These studies of IDPs also reveal significant limitations in the current simulation methodology. In particular, successful simulations of biomolecules not only require accurate molecular models, but also depend on the ability to sufficiently sample the complex conformational space. By designing a realistic yet computationally tractable coarse-grained protein model, we demonstrated that the popular temperature replica exchange enhanced sampling is ineffective in driving faster reversible folding transitions for proteins. The second original contribution of this dissertation is the development of novel simulation methods for enhanced sampling of protein conformations, specifically, replica exchange with guided-annealing (RE-GA) method and multiscale enhanced sampling (MSES) method. We expect these methods to be highly useful in generating converged conformational ensembles.

MULTI-SCALE SIMULATIONS OF INTRINSICALLY DISORDERED PROTEINS AND DEVELOPMENT OF ENHANCED SAMPLING TECHNIQUES

by

WEIHONG ZHANG

B.S., Beijing University of Chemical Technology, 2009

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Biochemistry and Molecular Biophysics Graduate Group

Kansas State University
Manhattan, Kansas

2014

Approved by:

Major Professor

Associate Professor, Jianhan Chen

ABSTRACT

Intrinsically disordered proteins (IDPs) are functional proteins that lack stable tertiary structures under physiological conditions. IDPs are key components of regulatory networks that dictate various aspects of cellular decision-making, and are over-represented in major disease pathways. For example, about 30% of eukaryotic proteins contain intrinsic disordered regions, and over 70% of cancer-associated proteins have been identified as IDPs. The highly heterogeneous nature of IDPs has presented significant challenge for experimental characterization using NMR, X-ray crystallography, or FRET. These challenges represent a unique opportunity for molecular modeling to make critical contributions. In this study, computer simulations at multiple scales were utilized to characterize the structural properties of unbound IDPs as well as to obtain a mechanistic understanding of IDP interactions. These studies of IDPs also reveal significant limitations in the current simulation methodology. In particular, successful simulations of biomolecules not only require accurate molecular models, but also depend on the ability to sufficiently sample the complex conformational space. By designing a realistic yet computationally tractable coarse-grained protein model, we demonstrated that the popular temperature replica exchange enhanced sampling is ineffective in driving faster reversible folding transitions for proteins. The second original contribution of this dissertation is the development of novel simulation methods for enhanced sampling of protein conformations, specifically, replica exchange with guided-annealing (RE-GA) method and multiscale enhanced sampling (MSES) method. We expect these methods to be highly useful in generating converged conformational ensembles.

TABLE OF CONTENTS

Table of Contents **v**

List of Figures **xii**

List of Tables **xviii**

Acknowledgements **xix**

Dedication **xx**

1 Introduction **1**

1.1 Intrinsically Disordered Proteins 1

1.1.1 Experimental methods for characterizing structural properties of unbound IDPs 2

1.1.2 Experimental methods for mechanistic characterization of IDP interactions 4

1.1.3 Physics-based atomistic molecular modeling for characterization of unbound IDP 5

1.1.4 Molecular modeling for mechanistic characterization of IDP interactions 7

1.2 Enhanced Sampling Techniques for Molecular Modeling 9

1.2.1 Replica exchange method 9

1.2.2 Representative competing enhanced sampling techniques 13

1.3 Organization of Dissertation 14

2 Residual Structures, Conformational Fluctuations, and Electrostatic Interactions in the Synergistic Folding of Two Intrinsically Disordered Proteins 16

2.1 Introduction 18

2.2 Results 23

2.2.1 Convergence and validation of the disordered structure ensembles 23

2.2.2 Folded-like conformations in the unbound state of NCBD 25

2.2.3 Induced folding-like mechanism on the baseline level 27

2.2.4 Binding and folding intermediates involving the NCBD and ACTR C-terminal segments 30

2.2.5 Native and non-native salt-bridges in encounter complexes and intermediates 33

2.3 Discussion 38

2.4 Methods 41

2.4.1 REX/GBSW simulations of free NCBD and ACTR 42

2.4.2 Room temperature and high-temperature simulations of the wild-type and mutant NCBD/ACTR complexes 43

2.4.3 Explicit solvent high-temperature simulations 44

2.4.4 Free energy calculations 45

2.5 Supplemental Materials 47

3 Synergistic Folding of Two Intrinsically Disordered Proteins: Searching for Conformational Selection 55

3.1 Introduction 57

3.2 Methods 62

3.2.1 Topology-based modeling of NCBD/ACTR 62

3.2.2 Simulation protocols 63

3.2.3 Data analysis 64

3.3 Results and Discussion 65

3.3.1 Calibration of the sequence-flavored Gō-like model 65

3.3.2 The baseline mechanism: induced folding vs. conformational selection . 67

3.3.3 A key intermediate state of the NCBD/ACTR interaction 68

3.3.4 Mechanism of coupled binding and tertiary folding of NCBD 69

3.3.5 Folding and binding of individual helical segments of NCBD and ACTR 72

3.3.6 Network analysis of the complex pathways of coupled binding and folding 74

3.4 Conclusions 77

3.5 Supplemental Materials 81

4 Electrostatically Accelerated Encounter and Folding for Facile Recognition of Intrinsically Disordered Proteins 89

4.1 Introduction 91

4.2 Results 95

4.2.1 Topology-based modeling of IDP coupled binding and folding 95

4.2.2 Baseline mechanisms of coupled binding and folding: effects of electrostatic forces 98

4.2.3 Kinetic effects of long-range and nonspecific electrostatic forces 100

4.2.4 Mechanism of electrostatically accelerated folding upon encounter 102

4.3 Discussion 103

4.4 Methods 104

4.4.1 Calibration of topology-based coarse-grained models with and without explicit charges 104

4.4.2 Simulation protocols 105

4.4.3 Free energy and kinetic analysis 106

4.5 Supplemental Materials 112

4.5.1 Sequences of IDPs 112

5 Efficiency of Adaptive Temperature-based Replica Exchange for Sampling Large-scale Protein Conformational Transitions 120

5.1 Introduction 122

5.2 Methods 126

5.2.1 Adaptive temperature-based REMD protocols 126

5.2.2 A topology-based coarse-grained protein model with rugged energy landscape 128

5.2.3 Simulation protocols and analysis 129

5.3 Results and Discussion 132

5.3.1 Convergence of the EE- and CM-optimized temperature distributions . . 132

5.3.2 Exchange acceptance rates and numbers of round trips in the temperature space 133

5.3.3 Number of folding and unfolding transitions 134

5.3.4 Convergence of thermodynamic properties 135

5.4 Conclusions 138

5.5 Supplemental Materials 141

6 Replica Exchange with Guided Annealing for Accelerated Sampling of Disordered Protein Conformations 142

6.1 Introduction 143

6.2 Methods 145

6.2.1 Theory of replica exchange with guided annealing (RE-GA) 145

6.2.2 Numerical RE simulations using a kinetic network model 148

6.2.3 Coarse-grained and atomistic protein simulations 149

6.3 Results 151

6.3.1 Evaluation of RE-GA schemes based on coarse-grained protein simulations 151

6.3.2 Numerical simulations of RE-GA schemes 152

6.3.3 Atomistic simulation of intrinsically disordered KID protein 155

6.4 Discussion 157

6.5 Supplemental Materials 160

7 Accelerate Sampling in Atomistic Energy Landscapes Using Topology-based Coarse-grained Models **166**

7.1 Introduction 167

7.2 Methods 169

7.2.1 Multi-scale enhanced sampling (MSES) method 169

7.2.2 Simulation protocols 171

7.3 Results and Discussion 173

7.3.1 MSES simulations recover the correct canonical ensembles 173

7.3.2 Efficiency of MSES simulations 174

7.3.3 Convergence of MSES simulations 176

7.4 Conclusions 177

7.5 Supplemental Materials 179

8 Conclusions and Future Directions **182**

References **186**

A Copyright Clearance

LIST OF FIGURES

2.1 Two representative folded conformations of NCBD 22

2.2 Calculated residue helicity of NCBD and ACTR in the free and bound states . . . 25

2.3 Orientations of NCBD $C_{\alpha 1}$ and $C_{\alpha 3}$ with respect to $C_{\alpha 2}$ in the unbound state . . 27

2.4 Averaged structures of the six most populated clusters of free NCBD 28

2.5 Evolutions of various contact fractions of the NCBD/ACTR complex at 475 K . . 31

2.6 Free energy surfaces of (un)binding and (un)folding of the wild-type and mutant NCBD/ACTR complex 32

2.7 Evolutions of the fractions of native intermolecular interactions of individual helices of the NCBD/ACTR complex 33

2.8 Native and non-native salt-bridges in the NCBD/ACTR interaction 35

2.9 Potential of mean forces of the Arg-Asp interactions in implicit and explicit solvents 36

2.10 Overall mechanism of synergistic folding of NCBD and ACTR 40

S2.1 Convergence of the calculated residue helicity of free NCBD 47

S2.2 Additional conformational properties of free NCBD 47

S2.3 Orientations of NCBD $C_{\alpha 1}$ and $C_{\alpha 3}$ with respect to $C_{\alpha 2}$ 48

S2.4 Summary of control simulations of the NCBD/ACTR complex in GBSW and GBSW/MS2 49

S2.5 Trial unfolding simulations in GBSW/MS2 at different temperatures 50

S2.6 Tertiary contacts of the NCBD/ACTR complex 51

S2.7 Evolution of various contact fractions in GBSW/MS2 simulations at 450 and 475 K 52

S2.8 Distances between key charged residues during three representative unfolding simulations at 475 K 52

S2.9 Evolution of various contact fractions during unfolding simulations in TIP3P at 500 K 53

S2.10 Evolution of various contact fractions during unfolding simulations of the mutant NCBD/ACTR complex at 450 K 54

3.1 The NMR structure of the NCBD/ACTR complex 60

3.2 Calibration of Gō-like model 66

3.3 2D free energy surfaces of the synergistic binding and folding of NCBD and ACTR 68

3.4 Probability distributions of the fraction of native intermolecular contacts formed . 70

3.5 2D free energy surfaces of coupled binding and tertiary folding of NCBD 71

3.6 2D free energy surfaces of as functions of Q_{inter} and the fraction of intra-molecular native interactions formed within individual helices of NCBD and ACTR 73

3.7 2D free energy surfaces as functions of the fractions of native intermolecular contacts formed by various segments of ACTR 74

3.8 2D free energy surfaces as functions of the fractions of native intermolecular contacts formed by various segments of NCBD 75

3.9 The CSN of the synergetic folding of NCBD and ACTR 76

S3.1 Residue-residue intra-molecular contact maps of ACTR and NCBD in the complex 81

S3.2 REX calibration simulations and production simulations of NCBD and ACTR . . 84

S3.3 PMFs and C_v of NCBD/ACTR complex 85

S3.4 The RMSF profiles of NCBD and ACTR in the bound state at 300 K 85

S3.5 Distributions of the fractions of intra-molecular contacts of various ACTR and NCBD helical segments in the unbound, intermediate and bound states 86

S3.6 Distributions of the distances and cross-angles between NCBD helices in the unbound, intermediate and bound states 87

S3.7 CSN of the synergetic folding of NCBD and ACTR 88

4.1 Structures and surface electrostatic potentials of three complexes 95

4.2 Free-energy surfaces at T_m as a function of the fractions of intra- and intermolecular contacts formed 108

4.3 Free energy as a function intermolecular contact fraction at T_m 109

4.4 Effective rate constants for transitions between the unbound, collision complex and bound states 109

4.5 Distributions of IDPs on the substrate surfaces in the collision complexes derived from simulations using models with and without explicit charges 110

4.6 Free-energy surfaces at T_m as a function of binding RMSD of the IDP and center of mass separation between two peptides (R_{CM}) 111

S4.1 Residual helicities of p53-TAD1, HIf-1α, and ACTR in the unbound states 113

S4.2 2D free energy surfaces at T_m of HIF1-α 116

S4.3 2D free energy surfaces at T_m of ACTR 117

S4.4 2D free energy surfaces at T_m of NCBD 118

S4.5 Representative snapshots along the binding and folding pathways of p53-TAD1/TAZ2 119

S4.6 Representative snapshots along the binding and folding pathways for HIF-1α/TAZ1 119

5.1 REMD simulations of GB1p β-hairpin 124

5.2 The structure of B1 domain of streptococcal protein G 130

5.3 The heat capacity and free energy of model protein 132

5.4 Temperature distributions optimized by EE and CM methods 133

5.5 Exchange acceptance rates extracted from different RE protocols 134

5.6 2D free energy surfaces of folding of protein GB1 136

5.7 Self-convergence of thermodynamic properties 138

5.8 Self-convergence of thermodynamic properties 139

S5.1 Time series of the temperature, total energy, RMSD and the numbers of native and total contacts for a single replica 141

6.1 Folded structures of protein GB1 and KID domain 150

6.2 The time evolution and free energy profiles of protein GB1 152

6.3 Numerical simulations of RE-SA 154

6.4 Convergence of calculated residue helicity profiles of KID using RE-SA and RE . 156

6.5 Probability distributions of helical substates of KID 157

6.6 Distributions of the end-to-end distance of KID 158

S6.1 2D Free energy surface of protein GB1 160

S6.2 The ratio between calculated and expected equilibrium constants as a function of guided annealing frequency for different models 162

S6.3 Number of reversible transitions as a function of guided annealing frequency . . 163

S6.4 Representative structures of KID domain from RE simulation 164

S6.5 Representative structures of KID domain from RE-SA simulation 165

7.1 Folded structure and simulated ensembles of GB1p 173

7.2 Numbers of native hydrogen bonds as a function of time 175

7.3 Conformational space sampled by individual replicas in the T-RE and MSES . . 177

7.4 Probability distributions of the number of native hydrogen bonds for GB1p, GB1m1, and GB1m3 178

S7.1 Co-evolution of the numbers of native H-bonds of the CG and atomistic copies for a representative replica from the MSES control simulation of GB1p 179

S7.2 Representative trajectories from constant temperature MSES folding simulations of peptide GB1p 180

S7.3 Conformational space during MSES control simulations of GB1m1 and GB1m3 . 181

LIST OF TABLES

2.1 Unfolding and unbinding kinetic constants at 475 K 30

S3.1 Native intermolecular interactions in NCBD/ACTR complex 82

S3.2 Summary of all 11 production simulations at 315 K. 83

4.1 Key properties of three IDP complexes 94

4.2 Summary of coarse-grained models 97

S4.1 MFPTs and numbers of transitions of the p53-TAD1/TAZ2 complex 114

S4.2 MFPTs and numbers of transitions of the HIF-1α/TAZ1 complex 114

S4.3 MFPTs and numbers of transitions of the NCBD/ACTR complex 115

5.1 Summary of production REMD and MD simulations 131

6.1 Summary of parameters of the numerical RE and RE-SA simulations 153

S6.1 Key parameters of kinetic models used in numerical RE simulations 161

7.1 Summary of T-RE and MSES simulations of GB1p, GB1m1 and GB1m3 174

S7.1 GB1p folding efficiency and rates at 270 K 180

Acknowledgments

I would like to express my sincere appreciation to my research advisor Dr. Jianhan Chen, who has been a great mentor to me. His guidance and support help me become an independent research scientist. His patience, motivation, enthusiasm and immense knowledge have a profound influence to me. I could not have imagined having a better advisor for my Ph.D study.

Besides my advisor, I also want to thank the rest of my committee: Dr. Ramaswamy Krishnamoorthi, Dr. Paul E. Smith, Dr. Michal Zolkiewski, and Dr. John Schlup, for their encouragement, insightful comments, and great suggestions to my thesis.

I also appreciate KSU computing cluster "Beocat" for offering me great computing power to perform simulations for biomolecules.

Last but not least, I thank my fellow labmates in Chen Group: Dr. Debabani Ganguly, Dr. Kuo Hao Lee, Dr. Zhiguang Jia, Alex Beugelsdijk, and Michael Mohan for the stimulating discussion, the close collaboration, and all the fun we have had in the last five years. I am very luck to have a group of brilliant people by my side during the adventure in science.

Dedication

This dissertation is lovingly dedicated to my parents, Xinning Zhang and Nailin Shen, and my wife Xuan Zhou. Their unconditional love, spiritual support and meticulous caring have sustained me throughout my life.

Chapter 1

Introduction

1.1 Intrinsically Disordered Proteins

Intrinsically disordered proteins (IDPs) refer to functional proteins that lack stable tertiary structures under physiological conditions[1–6]. The discovery of IDPs has challenged the traditional structure-function paradigm, which states that protein function is closely related to its three-dimensional structure. The existence of functional disordered proteins has been recognized for many years. For example, many small polypeptide hormones, such as monomeric glucagon, are unstructured in aqueous solution and fold upon binding to their receptors[7]. The prevalence of IDPs, however, have only been recently recognized since the availability of vast library of gene sequences, which allows us to perform gene-based functional analysis and intrinsic disorder prediction[8,9]. IDPs have low content of bulky hydrophobic amino acid residues (Val, Leu, Ile, Met, Phe, Trp and Tyr) and high content of polar / charged (Gln, Ser, Glu, Lys) and structural breaking (Pro) amino acid residues[10,11]. By capturing this strong sequence signature of IDPs, it is predicted, surprisingly, more than 30% of eukaryotic proteins contain extended disordered segments[12,13]. This apparent prevalence of IDPs suggests that some proteins require the absence of stable 3D structure to carry out functions, and the heterogeneity can offer great advantages in functions. Indeed, the majority of known disordered proteins or regions are involved in cell signaling or regulation[14]. Importantly, IDPs are overly presented in different diseases including cancer, cardiovascular diseases, neurodegenerative diseases and diabetes[15–17]. A large number of disease-related mutations are found in the predicted disordered regions[18]. About 20% of these mutations could elevate the level of residual structures, thus causing dysfunctionality of these

proteins in cell signaling and regulation pathways[17].

The disordered properties have several advantages for the functional roles of IDPs[19–25]. For instance, IDPs often fold into stable structures to carry out functions, and the folding process is associated with specific binding[26]. This process is referred to as "coupled binding and folding" of IDPs. The delicate enthalpy-entropy compensation of the coupled binding and folding allows IDPs to achieve high-specificity and low affinity binding, which is considered as a crucial advantage in cell signaling and regulation. In addition, IDPs possess structural plasticity, which could allow IDPs to bind different targets and adopt different structures[8]. This makes IDPs the hubs in the cell signaling network, and enables one-to-many (one disordered region could bind to many partners) and many-to-one (multiple disordered regions could binding to one site) signaling. The very C-terminal segment of p53 (residue 374 to 388) is a good example of structural plasticity. It is capable of forming all three major secondary structure types (α-helix, β-sheet, irregular structure) upon binding to four different targets, namely, S100$\beta\beta$, sirtuin, CBP and cyclin A2[17].

Because of the high involvement in cell functions and diseases, there has been an increasing interest in studying structures and functions of IDPs using both experimental and theoretical methods. However, little is known about the structural properties of IDPs, and how intrinsic disorder mediates protein functions.

1.1.1 Experimental methods for characterizing structural properties of unbound IDPs

One major research focus is what structural properties of the disordered unbound IDPs are important for them to efficiently interact with multiple targets. Pre-existing folded-like structures have been observed in unbound IDPs[3]. It has been suggested that these folded-like structures might provide initial binding interface to facilitate the coupled binding and folding of IDPs[27–29]. In contrast, recent mechanistic studies emphasize the crucial role of conformational flexibility in coupled binding and folding of IDPs[30–34]. Better understanding of interaction mechanisms

requires performing detailed structural characterizations for unbound disordered IDPs. These characterizations include describing key sub-states in the ensemble, calculating the population of each sub-states, measuring the timescale of inter-conversion between these states, and estimating the free energy barriers separating different interconverting sub-states[35].

Many experimental methods can be used to characterize the ensemble properties of unbound IDPs. Spectroscopic methods such as nuclear magnetic resonance (NMR) have been extensively used to study the structural properties of IDPs[26,36–38]. NMR is one of the most comprehensive techniques that could offer detailed insights into the structures and dynamics of disordered proteins. Recent advances in NMR technology have allowed the resonances of several unfolded proteins and partial folded proteins to be fully assigned[39–41]. Many other techniques also give important structural information. Circular dichroism (CD) can provide secondary structure information of non-globular proteins and can be used to detect disorder-to-ordered transitions[42]. Small angle X-ray scattering (SAX) data describe the overall shape of the protein, and hydrodynamic experiments can determine the size-distribution over the protein[43,44]. Single molecular techniques such as single molecular Förster resonance energy transfer (sm-FRET) are also unique techniques for providing distance information in disordered structures[45–47]. Indeed, sm-FRET has been successfully used to study several large IDPs, such as α-synuclein[48–50], prions[45], and p53[51].

However, faithful structural interpretation based on the experimental measurements still remain an important challenge in structure properties characterization of unbound IDPs[52]. It is almost impossible to crystalize disordered unbound form of IDPs, making it difficult to obtain detailed structural information from X-ray crystallography methods. The structure characterizations from NMR are also limited. This is mainly because of the heterogeneous nature of IDPs, and it remains a challenging problem to interpret the individual structural properties of flexible systems based on the ensemble average measurements[53]. Single molecular techniques such as sm-FRET, are limited to the requirement of attached donors and acceptors. These covalently attached chromophores could have large impact on the sensitive disordered ensembles of IDPs. In addition, most sm-

FRET can detect the distances that are greater than 20 Å, limiting the application of sm-FRET to small IDPs, where it is the case for most of the regulatory IDPs. As a result, it is very difficult to determine IDP structural ensembles based on experimental data alone. This has become a fundamental limitation for understanding structure basis of free IDPs.

1.1.2 Experimental methods for mechanistic characterization of IDP interactions

Besides the structural properties of unbound IDPs, a mechanistic understanding needs to be established based on how the structural properties of IDPs can contribute to the robust interactions with their targets, and how theses properties are modulated by cells, e.g. through post-translational modification, switch in environmental conditions, and interactions with salt and membranes. However, only a few published studies aim to directly investigate the mechanism of IDP interactions. Most of the IDPs mechanistic studies focused on coupled binding and folding mechanisms of regulatory IDPs[26]. It is well recognized that coupled binding and folding of IDPs is highly associated with their activity and is modulated by many cellular events such as phosphorylation, acetylation, and interactions with other proteins[54]. This indicates coupled binding and folding of IDPs is an inherent feature of IDP function and regulation. Therefore, understanding the mechanism of IDPs interactions is of fundamental importance and high biological relevance.

Several experimental techniques have been proven to be very useful to detect the coupled binding and folding events of IDPs. NMR relaxation dispersion experiment is a useful technique to determine the rates of exchange and chemical shift differences between different conformational states. And it can provide information about intermediate states based on the chemical shifts[55–57]. For example, in the study of kinase-inducible activation (KID) domain of cAMP response element-binding protein (CREB), NMR relaxation dispersion analysis complemented with NMR titration, and provided unparalleled details on the mechanism of KID binding to KIX domain of CREB-binding protein (CBP)[32]. Hydrogen-deuterium exchange combined with mass

spectrometry is another powerful tool to characterize the backbone flexibility and key intermediate states during the coupled binding and folding process[58]. ϕ-value analysis that probes the effects of point mutations on the folding kinetics can also be used to understand the folding pathway at the structural level involves only one single transition state[59]. Other techniques such as temperature-jump spectroscopy and stop-flow analysis can also be performed to infer the baseline mechanisms of coupled binding and folding based on the important kinetic information[31,60,61].

Limitations exist for mechanistic investigation of IDPs using experimental methods. The disorder to order transitions along the coupled binding and folding pathway significantly complicate the interaction mechanisms of IDPs. For example, NMR relaxation dispersion analysis is limited to slow exchange in microsecond to millisecond timescale. It is not applicable to fast inter-conversion between conformational states. Another important limitation is that only a few states can be fitted with dispersion curves. For complicated interactions with multiple intermediate states, relaxation dispersion could lead to uncertainty in interpretation of derived kinetic parameters[62]. ϕ-value analysis can be laborious to map out key interactions in the transition state using point mutations. Furthermore, ϕ-value analysis is strictly based on a two-state model system. The interpretation can be complicated by the presence of any intermediate states.

1.1.3 Physics-based atomistic molecular modeling for characterization of unbound IDP

The significant challenges in experimental methods for studying IDPs represents a unique opportunity for molecular modeling to make critical contributions. One of my key research objectives is applying molecular modeling techniques to obtain detailed structural properties of unbound IDPs and to acquire the mechanistic understanding of IDPs interactions.

The reliability and accuracy of molecular modeling and simulations of IDPs highly depend on two major components, namely, force field and conformational sampling. Most of the physics-based force fields are parameterized using large range of experimental data and with important approx-

imations. Therefore, they are referred to as empirical force fields. Different force fields such as CHARMM force fields[63], the AMBER force field[64], the OPLS force fields[65] and the GROMOS force fields[66], have been demonstrated successful applications on simulations of proteins. However, for highly heterogeneous and flexible IDPs, they tend to sample a broad conformational space with numerous weakly populated subspaces separated by relatively small free energy barriers. Therefore, successful and reliable simulations of free IDPs are pushing the limits of both the force field accuracy and conformational sampling efficiency.

In recent years, important advances have been achieved in optimization of protein force fields for simulations of protein conformational transitions[67–71]. Among all the successful protein force fields, implicit solvent force field reaches a balance between accuracy and computational efficiency for atomistic simulation of IDPs[1,72]. Traditional explicit solvent force field explicitly includes all the water molecules around the protein solute, dramatically increasing the size of the system and the computational cost of the simulation. The idea of implicit solvent is to capture the mean influence of water on proteins by directly estimating the solvation free energy[73]. Only protein molecules need to be represented in atomistic level, thus reducing the system size about 10-fold and improving the simulation efficiency. For instance, using generalized Born (GB) approximation[74], GBSW implicit solvent force field is about 30 times faster than explicit solvent force field for simulations of small proteins (less than 100 residues)[75]. The increase in efficiency has allowed the optimization of implicit solvent force fields to be more realistic and reduce certain systematic errors in the explicit solvent force fields[76–80]. For example, the optimized GBSW protein force field could reproduce the structures and different stabilities of several α-helical and β-hairpin peptides[77,78,81]. However, limitations also exist for implicit solvent force fields. In implicit solvent force fields, the water is treated as continuum, which ignores some important non-bulk water-protein interactions. Protein-water boundary in the continuum electrostatics is also ambiguous to choose. Nevertheless, the compact system size of implicit solvent force field is a great advantage for simulations of free disordered IDPs, where the more extended structures of IDPs required larger water box and more water molecules to be accommodated in explicit solvent

force fields.

1.1.4 Molecular modeling for mechanistic characterization of IDP interactions

In principle, direct simulations of IDPs coupled binding and folding using atomistic models can provide a detailed description on IDPs interactions. Many computational techniques can be performed to obtain mechanistic understanding on coupled binding and folding of IDPs. Free energy calculations, in particular, have played an important role in understanding protein folding[82]. Free energy landscapes constructed by using proper reaction coordinates allow thermodynamic clarification of coupled binding and folding mechanism[30]. A range of important mechanistic features can be obtained from such calculations, including most probable pathways, intermediate states, and transition states. However, free energy calculation are computationally expensive to perform using atomistic force fields, because in atomistic model exhaustive sampling of the IDP conformational space is difficult to achieve, especially in presence of a specific target. In addition, limited number of reaction coordinates might be difficult to describe the complex binding and folding interactions. For this, transition path sampling (TPS) techniques[83–85] calculate an ensemble of transition paths, and can provide the most detailed molecular description of such coupled binding and folding. An important limitation is that TPS requires a predefined initial state, which is difficult to choose for the heterogeneous free IDPs. Alternatively, temperature driven unfolding/unbinding simulations started from a bound and folded complex structure can be used to study the coupled binding and folding process of IDPs. The assumption is that binding and folding is a reversed process of unbinding and unfolding. Unfolding and unbinding process in the simulation at high temperature can be reversed back to infer the coupled binding and folding pathway. However, a concern is that most probable transition state might depend strongly on the temperature[86]. Despite such a caveat, high temperature unfolding simulations have been successful for understanding interactions of many proteins including IDPs[34,87–90].

Coarse-grained representation of protein molecules (e.g. one amino acid residue is represented by on bead) can significantly reduce the degrees of freedom of the system and can be applied to study IDP binding and folding[91]. The reversible process sampled using this technique enables mechanistic analysis such as calculation of free energy surfaces and transition of IDP interactions. Specifically, topology-based coarse-grained modeling allows efficient simulation of IDP coupled binding and folding process[62,91–93]. Topology-based modeling is based on the minimal frustration theory of folded proteins, which argues that native contacts dictate protein energy landscape[94]. Given a folded topology of a protein, the true protein energy landscape can be approximated by constructing the effective energy based only on the native interactions[95]. Those effective energy functions are often referred to as Gō or Gō-like models. Gō models are simple yet very successful for study of protein folding mechanisms. It has demonstrated impressive correspondence between experiment and theory[96–98]. Importantly, there are significant differences of sequences and binding interfaces between IDPs and globular proteins[21]. Original Gō-like models require careful calibration to balance intermolecular and intramolecular interactions, so that they do not inherently favor particular type of events (e.g. binding vs. folding)[62]. Such calibration requires integration with experimental data, including binding affinity and residual secondary structure. Moreover, topology-based coarse-grained modeling can be readily extended and modified to study how additional factors may affect the folding and binding[92,99]. This can be used to investigate how different forces modulate IDPs interactions, such as electrostatic interactions, post-translational modifications, and solution conditions. However, limitations of Gō-like models need to be understood. For example, detailed atomistic information for the protein molecule is lost during the coarse-graining process. Coarse-grained model is a poor description of the unbound state of IDPs. In addition, Gō-like models are difficult to model structural plasticity of IDPs. Therefore, the observation and hypothesis generated by topology-based coarse-grained models need to be further investigated using more detailed model such as atomistic models or experiments. This multi-scale computational approach, integrated with experiments, would be an effective and powerful tool in mechanistic studies of IDPs.

1.2 Enhanced Sampling Techniques for Molecular Modeling

Reliable computer simulations of proteins not only require accurate molecular models but also require sufficient sampling of the conformational space. My second major focus is to develop novel methods for enhanced sampling of protein conformations. At present, generating atomistic structure ensembles that are statistically representative of the accessible conformations under a given set of thermodynamic conditions remains a challenging problem in simulations of proteins[100]. The difficulty arises not only because of the large and complex conformational space of proteins, but also due to significant free energy barriers that separate different conformational subspaces. Traditional molecular dynamics (MD)[101] and Monte Carlo (MC)[102] sampling methods are limited because proteins will frequently be trapped in numerous local energy minima. MD is a very straightforward sampling technique. In MD, the forces and potential energy of the system are defined by the force field. The movements of all the atoms are determined by solving the Newton's equation of motion numerically, which typically requires a very small time step on the order of 10^{-15} s. MC method propagates the conformation of the simulated system by making random moves. The newly generated configuration will be subject to test using Metropolis criteria to decide whether it is accepted or rejected.

1.2.1 Replica exchange method

The replica exchange (RE) method[103–105], also known as parallel tempering, has emerged as a relatively straightforward and powerful approach that can enhance conformational sampling. The basic idea is to simulate multiple replicas of the system at different temperatures independently using either MC or MD. Periodically, replicas attempt to exchange simulation temperatures according to a Metropolis criterion that preserves the detailed balance and ensures canonical distributions at all temperatures. The resulting random walk in the temperature space helps the

replicas to escape the energy local minima and sample wider range in the conformational space. Replica exchange molecular dynamics (REMD) in particular has been successfully applied to protein simulations[106–110]. On the other hand, the true efficiency of REMD in sampling large-scale protein conformational transitions needs to be rigorously benchmarked, and the dependence of REMD simulations on the protein system and key parameters needs to be explored.

Extensive research based on both theories[111–116] and simulations of small peptides[110,117–119] have generally confirmed that RE can accelerate sampling as long as there is a positive activation enthalpy associated with conformational transitions of the protein. However, it is also recognized that RE can be severely limited by the presence of sharp, cooperative conformational transitions such as protein folding[111,120]. Replicas often segregate and only sampling distinct major conformational states (e.g. protein folded state and unfolded state), and this segregation in the conformational space often lead to exchange bottlenecks near melting temperature (T_m) where the transition occurs[121]. One possible strategy to resolve this issue is to increase the temperature diffusion of replicas in RE simulations. Many methods have been previously proposed for this purpose, such as more frequent exchange attempts[119], global energy reassignment[122], and non-equilibrium switches[123]. More aggressive approaches attempt to reduce to number of replicas required for effective temperature diffusion by directly reducing the number of particles that participate in temperature exchanges[124–126], but sometimes with undesirable consequences[127].

Adaptive replica exchange methods

Adaptive RE methods are variations of original RE that aim to enhance the temperature diffusion by recursively adjusting the distribution of simulation windows along the temperature space[128]. They can effectively reduce or eliminate exchange bottlenecks by either equalize the exchange acceptance rate (exchange equalization or EE)[129] or maximize temperature diffusion current (current maximization or CM)[128,130]. EE method assumes that REMD will provide the

highest sampling efficiency if all the replicas spend equal amount of time at each temperature. As proposed by Sugita and Yokomoto[104] and Kofke[129], this may be achieved through allocating more simulation windows around Tm of the protein to equalize the exchange rate. Nonetheless, it is not clear whether consideration of exchange efficiency alone is sufficient to optimize the sampling efficiency, which is more directly measured by the frequency of conformational transitions, and ultimately by the convergence of various thermodynamic properties of interest. The goal of CM method, on the other hand, is to maximize the number of times that clients progress from the lowest temperature to the highest temperature and back again, i.e., "round trips". Arguably the number of "round trips" is more directly related to the ability of RE to drive conformational transitions, and thus CM could substantially improve the sampling efficiency[130].

Replica exchange with guided annealing

Another strategy to overcome the replica segregation in conformational space is to design temperature exchange schemes that could guide the trapped replicas at lower temperatures to visit higher temperatures more frequently[121]. Such schemes may be generally referred to as RE with guided annealing (RE-GA). A similar idea is the recently described convective RE method (cRE)[131]. In cRE, the replicas take turn to be the "stick replica" and undergo convective annealing cycles. Specifically, the exchange attempts are chosen such that the stick replica will first only try to exchange to higher temperatures until reaching the highest temperature. Then, the stick replica will only attempt to exchange to lower temperatures. When the stick replica reaches the lowest temperature, the direction of exchange attempts is reversed again until the stick replica returns to its original temperature. In principle, this modified exchange attempt patterns could guide the folded replicas to visit higher temperatures more frequently and accelerate unfolding and to guide the unfolded replicas towards lower temperatures to encourage folding transitions. As formally proved by Spill *et al.*[131], it appears that exchanging according Metropolis criteria and balanced exchange biases towards both higher temperatures and lower temperatures would

guarantee detailed balance and ensure the correct canonical ensemble to be generated by cRE simulations. However, in the presence of slow cooperative conformational transitions, it will take many more attempts for replicas occupying low-energy states to exchange to higher temperatures than returning to lower temperatures, and vice versa for replicas occupying high-energy states. Therefore, such RE-GA schemes lead to unbalanced exchange attempt probabilities and break the detailed balance condition when conformational transitions are slow compared to the temperature diffusion timescales, which is almost always the case for RE simulations of proteins.

Multiscale enhanced sampling

Fundamentally, the limited efficiency of RE in sampling cooperative transitions such as protein folding can be attributed to large entropic component in the free energy barriers[132,133]. Protein folding rate only weakly depends on temperature and often displays anti-Arrhenius behaviors[114,134], which makes tempering very ineffective in drive protein conformational transitions. One way to enhance the crossing over entropic barriers is to use coarse-grained (CG) representation, because CG models can significantly reduce the degrees of freedom of the system, shrink the conformational space and allow faster reversible transitions. But this can only be achieved at the expense of reduced details and accuracy. An ideal approach should involve combining both CG model with atomistic model at the same time so that it could benefit from both the faster transitions of the CG model and the ultimate details offered by the atomistic model. The promise of such multiscale approach has been well recognized, and several clever ideas have been proposed towards this goal[125,135–139]. Lwin and Luo purposed a dual resolution sampling method[136], which couples atomistic model for accuracy and coarse-grained model for efficiency, aiming to cross the entropic barrier in protein folding process. However, this method uses two sets of replicas, one for atomistic model and one for CG model, limited its use on large systems. Lyman et al. developed the resolution exchange which exchange the shared essential coordinates of CG model with a subset of coordinates in atomistic model (e.g. coordinates of

protein backbone atoms)[137]. Although this exchange algorithm could allow novel conformations sampled by CG model to be injected into atomistic ones, the method suffers from very low exchange acceptance rate by directly exchanging the coordinates. Liu and Voth designed the smart resolution replica exchange to achieve efficient exchange between different resolution and only two replicas are required for simulating small peptides[140]. In smart resolution replica exchange, before the coordinate-swapping between configurations at different resolutions, the coarse-grained model is mapped to the atomistic configuration and the relaxation simulation is performed to avoid "bad-contact" after exchange. However, the coarse-grained model lack information of the missing atoms in atomistic model, and the relaxation process will apparently cause the detailed balance condition to fail.

1.2.2 Representative competing enhanced sampling techniques

Besides RE-based enhanced sampling techniques, many methods were developed aiming to prevent biomolecules such as proteins from being trapped at local energy minima. The conformational transitions of proteins can be treated as a series of rare events that climb up the free energy barrier and moving from one energy basin to another. These rare events are directly related to free energy barrier heights. Instead of using temperature to enhance the barrier crossing, some enhanced sampling techniques utilize a bias potential to drive the transition across the energy barriers. Umbrella sampling is one approach to bridge the gaps between free energy barriers[141]. The Boltzmann-weighted distribution is inefficient for adequate sampling across free energy barriers, thus is replaced by a specific bias potential to cancel the free energy barriers. This bias potential will drive the molecular system from one state to another along the reaction coordinate. In umbrella sampling, there are a series of windows each restrains the reaction coordinate to a target value by a bias potential. From the improved sampling, all the simulation windows can be combined and analyzed by weighted histogram analysis method (WHAM)[142], and the free energy landscape can be calculated. One limitation of umbrella sampling is that

the compensating bias potential energy requires the prior knowledge of different conformational states. Accelerated molecular dynamics[143] is an efficient way to simulate transitions between high energy barriers without any advance knowledge of the location of conformational states. It alters the potential energy landscape by adding bias potentials, such that the effective energy barriers are reduced and the barrier crossing rates are enhanced. While the bias potentials are added at the potential energy minima, the transition states (saddle points on the potential energy landscape) remain unaltered. The statistics sampled on the bias potential are then corrected to recover the canonical distribution. Similarly, metadynamics methods also include bias potentials in the energy landscape[144]. In metadynamics simulations, the location of the system on a certain reaction coordinate is determined and a positive Gaussian bias potential is added to the true energy landscape at this location. This bias potential will discourage the system to come back to the previous point. The accumulation of the Gaussian potentials is pushing the system to explore the full energy landscape. The true free energy landscape along the reaction coordinate can be calculated as the opposite of sum of the Gaussian potentials. Other sampling methods such as Conformational Space Annealing (CSA)[145] focuses on a population of conformations rather than a single conformation to obtain the global minimum-energy structures. CSA method searches the whole conformational space in the early stage of the sampling process by building up "seed conformations"[145]. All the newly sampled conformations are modified from the "seed conformations". Then CSA narrows the search to smaller regions while maintaining the diversity of the sampling. Such annealing in conformational space would find not only the global minimum but also many other distinct low energy regions in the conformational space.

1.3 Organization of Dissertation

The overall structure of the study takes the form of eight chapters, including this introductory chapter. The first contribution of my research is to utilize multiscale molecular modeling and

simulation to understand the physical basis of how intrinsic disorder mediates IDPs functions. Chapter 2, 3 and 4 present the simulations for a specific IDP complex, namely, nuclear co-activator binding domain of the cyclic-AMP-response-element-binding protein binding protein (NCBD) and the activation domain of the activator for thyroid hormone and retinoid receptors of the p160 co-activator (ACTR). Specifically, chapter 2 focuses on characterization of structural properties of unbound IDPs as well as utilizing high-temperature induced unfolding/unbinding simulations to gain insights on the coupled binding and folding mechanism of NCBD and ACTR. Both simulations were performed using atomistic models in implicit solvent force field. In chapter 3, topology-based coarse-grained model was used to directly simulate coupled binding and folding events of NCBD and ACTR. Chapter 4 investigates the electrostatic interactions that affect coupled binding and folding kinetics. The second section of the dissertation including chapter 5, 6 and 7 is about the development of advanced conformational sampling methodologies for more efficient simulations of proteins. Chapter 5 begins by rigorously benchmarking the sampling efficiency of conventional and adaptive RE simulations, and identifies the potential limitations in temperature-based RE simulations. Chapter 6 examines the RE-GA schemes for enhanced sampling of protein conformational based on theoretical and numerical analysis. It then further discusses the potential usage of RE-GA on simulations of IDPs. Chapter 7 describes a multiscale enhanced sampling (MSES) method where efficient topology-based coarse-grained models are coupled with all-atom ones to enhance the sampling of atomistic protein energy landscape. The bias from the coupling is removed by Hamiltonian replica exchange, thus allowing one to benefit simultaneously from faster transitions of CG model and accuracy of atomistic force field. Finally, chapter 8 gives a brief summary and critiques of the findings. Directions for further research are identified.

Chapter 2

Residual Structures, Conformational Fluctuations, and Electrostatic Interactions in the Synergistic Folding of Two Intrinsically Disordered Proteins*

Abstract

To understand the interplay of residual structures and conformational fluctuations in the interaction of intrinsically disordered proteins (IDPs), we first combined implicit solvent and replica exchange sampling to calculate atomistic disordered ensembles of the nuclear co-activator binding domain (NCBD) of transcription coactivator CBP and the activation domain of the p160 steroid receptor coactivator ACTR. The calculated ensembles are in quantitative agreement with NMR-derived residue helicity and recapitulate the experimental observation that, while free ACTR largely lacks residual secondary structures, free NCBD is a molten globule with a helical content similar to that in the folded complex. Detailed conformational analysis reveals that free NCBD has an inherent ability to substantially sample all the helix configurations that have been previously observed either unbound or in complexes. Intriguingly, further high-temperature unbinding and unfolding simulations in implicit and explicit solvents emphasize the importance of conformational fluctuations in synergistic folding of NCBD with ACTR. A balance between preformed elements and conformational fluctuations appears necessary to allow NCBD to interact with different targets and fold into alternative conformations. Together with previous topology-based

*Reprinted under the Creative Commons Attribution (CC BY) license. Zhang, W., Ganguly, D., and Chen, J. (2012) "Residual structures, conformational fluctuations, and electrostatic interactions in the synergistic folding of two intrinsically disordered proteins", PLoS Computational Biology 8, e1002353.

Author Contributions: Conceived and designed the experiments: JC WZ. Performed the experiments: WZ JC. Analyzed the data: WZ JC DG. Contributed reagents/materials/analysis tools: WZ JC. Wrote the paper: WZ JC DG.

modeling and existing experimental data, the current simulations strongly support an "extended conformational selection" synergistic folding mechanism that involves a key intermediate state stabilized by interaction between the C-terminal helices of NCBD and ACTR. In addition, the atomistic simulations reveal the role of long-range as well as short-range electrostatic interactions in cooperating with readily fluctuating residual structures, which might enhance the encounter rate and promote efficient folding upon encounter for facile binding and folding interactions of IDPs. Thus, the current study not only provides a consistent mechanistic understanding of the NCBD/ACTR interaction, but also helps establish a multi-scale molecular modeling framework for understanding the structure, interaction, and regulation of IDPs in general.

2.1 Introduction

It is now widely recognized that many functional proteins lack stable tertiary structures under physiological conditions[1–3,5,6]. Importantly, such intrinsically disordered proteins (IDPs) are highly prevalent in proteomes[12], play crucial roles in cellular areas such as signaling and regulation[146,147], and are often associated with human diseases such as cancers[15,17,148]. The concept that intrinsic disorder can confer functional advantages has been discussed extensively[19,20,22–24]. For example, the disordered nature of IDPs could offer several unique benefits for signaling and regulation, including high specificity/low affinity binding, inducibility by posttranslational modifications, and structural plasticity for binding multiple partners. The last property appears to be particularly advantageous, and could support one-to-many and many-to-one signaling[22,149]. Nonetheless, the physical basis of these proposed phenomena remains largely elusive. Specifically, how IDP recognition and regulation are supported by the interplay of residual structures, conformational fluctuations and other physical properties as encoded in the peptide sequence is poorly understood.

The current limit in mechanistic understanding of how intrinsic disorder supports function might be attributed to two key challenges in characterizing IDPs. These challenges are broadly shared by mechanistic studies of protein folding, misfolding, and aggregation in general[150–153]. The first one is related to the difficulty in deriving detailed structural information of the disordered unbound states[37,154–156]. In general, only ensemble-averaged properties can be measured for disordered proteins except with single-molecule techniques (which have their own limitations in spatial resolution, labeling need, and protein size[46,47,157]). Recovering the underlying structural heterogeneity using averaged properties is a severely underdetermined problem[52,53,158–160]. It is generally not feasible to construct a unique disordered structure ensemble that is consistent with the available data. This fundamental limitation leads to significant ambiguity in the current knowledge of the conformational nature of unbound IDPs. The second challenge is to further clarify the functional roles of any putative conformational sub-states or other properties of an

IDP in its recognition and regulation (i.e., "function"). In particular, whereas some IDPs remain disordered in complexes[38,161], many fold into stable structures upon binding to specific targets[26]. The roles of intrinsic disorder vs. residual structures in such coupled binding and folding interactions have been under much debate[26]. On one hand, residual structures have been observed frequently in unbound IDPs, and intriguingly, such residual structures often resemble those in the folded complexes[27–30,162]. These observations have led to an attractive hypothesis that preformed structural elements might provide initial binding sites to facilitate efficient recognition (i.e., conformational selection-like mechanisms)[24,27]. On the other hand, evidence has accumulated in recent years, from computation as well as experimentation, to support a central role of nonspecific binding and emphasize the importance of disordered nature itself in promoting facile IDP recognition[30–33,59,62,90,93,163–165]. In fact, all published studies that extend beyond examining the unbound states alone have suggested induced folding-like mechanisms, at least at the baseline level.

Precisely how the disordered nature contributes to binding, however, is less clear. One proposal is that nonspecific binding of unstructured and presumably more extended conformations can increase the capture radii to enhance the binding kinetics[166,167]; however, such "fly-casting" effects is small with a theoretical maximum of $\sim$1.6-fold acceleration. Recent studies have shown that unbound IDPs tend to be much more compact than previously assumed[168–171], further reducing the proposed fly-casting affects. In addition, the rate-enhancing effect due to increased size is likely offset by slower diffusion[172]. Alternatively, the unbound state of IDPs is presumed heterogeneous and strongly fluctuating. More specifically, conformational sub-states in the unbound IDPs should be marginally stable and separated by small free energy barriers (e.g., a few kcal/mol or less). These conformational fluctuations could contribute to efficient IDP recognition by allowing the peptide to fold rapidly upon (nonspecific) binding[163,172], which is required for achieving the diffusion-controlled maximum binding rate (otherwise folding becomes rate-limiting)[173]. It should be noted that cellular events frequently modify the folding of IDPs to modulate their activities, such as through phosphorylations or by binding of other proteins[54]. Therefore, in

contrast to globular proteins where folding often serves only to achieve the native structures, folding and unfolding appears to be direct and inherent aspects of IDP function. This underpins the importance and biological relevance of obtaining a mechanistic understanding of binding-induced folding of IDPs beyond a subject of theoretical curiosity.

The challenge in detailed characterization of IDPs represents a unique opportunity for molecular modeling to make critical contributions[1]. In particular, atomistic simulations could provide the ultimate level of detail necessary for understanding the structure and interaction of IDPs. At the same time, the dynamic and heterogeneous nature of IDPs also pushes the limits of both the force field accuracy and conformational sampling capability. So-called implicit solvent is arguably an optimal choice for de novo simulations of IDPs because of its necessary balance of accuracy and speed[73,174–176]. The basic idea of implicit solvent is to capture the mean influence of water by direct estimation of the solvation free energy, therefore reducing the system size about 10-fold. Important advances have been made to greatly improve the efficiency and achievable accuracy of implicit solvent, such as via the popular generalized Born (GB) theory[175]. With reduced system size, implicit solvent is also particularly suitable for replica exchange (REX) simulations[104,177,178], an enhanced sampling technique that has proven highly effective in sampling protein conformational equilibria[179]. Importantly, improved efficiency with implicit solvent also allows careful optimization to suppress certain systematic biases that have plagued explicit solvent approaches[180,181]. For example, we have previously optimized the generalized Born with smooth switching (GBSW) model[75,76] together with the underlying CHARMM22/CMAP protein force field[69,182–184]. The resulting GBSW protein force field not only recapitulates the structures and stabilities of helical and β-hairpin model peptides with a wide range of stabilities[77,78], but also allows calculation of the conformational equilibria of small proteins under stabilizing and destabilizing conditions[81,185,186]. Although inherent and methodological limitations remain in implicit solvent[74], initial applications of implicit solvent to modeling small IDPs have been reasonably successful[30,33,72,169,187–189], substantiating the notion that it is a viable approach for atomistic simulations of IDPs.

The current work focuses on the nuclear-receptor co-activator binding domain (NCBD) of the transcription coactivator CREB-binding protein (CBP) and its interaction with the p160 steroid receptor co-activator ACTR. CBP and its paralogue p300 are general transcriptional coactivators that play critical roles in transcriptional regulation and participate in cell cycle control, differentiation, transformation, and apoptosis[190,191]. The NCBD domain (residues 2059-2117 in mouse CBP) is also known as interferon regulatory factor (IRF) binding domain (iBID) or the SRC1 interaction domain (SID). It mediates the interaction of CBP with a number of important proteins, including steroid receptor coactivators, p53 and IRFs[3,192]. The interaction of CBP with p160 coactivators in particular is important for recruitment of CBP/p300 to transmit the hormonal signal to the transcription machinery[193]. Besides the biological and medical significance, the NCBD/ACTR interaction also offers unique opportunities for understanding the molecular principles of IDP recognition. Both NCBD and the activation domain of ACTR that it interacts with (residues 1018-1088 in human ACTR; hereafter referred to as ACTR) are IDPs. Their interaction is an example of the "synergistic folding" mechanism[194] (the other known example also involves NCBD, but with the p53 transactivation domain, TAD[195]). In addition, four folded structures of NCBD have been solved in complex with various protein targets besides ACTR[194–197]. In these complexes, NCBD adopts two distinct tertiary folds that involve three similar helices, represented by the NCBD/ACTR and NCBD/IRF3 complexes (see Figure 2.1). Therefore, NCBD represents one of the few experimentally validated examples of structural plasticity, which is believed to be a key functional advantage of intrinsic disorder[22].

Interestingly, although free ACTR is largely devoid of residual structures, free NCBD contains one the highest levels of residual structures with folded-like helical content and molten globule characteristics[198,199]. In addition, even though nuclear magnetic resonance (NMR) relaxation analysis has established that free NCBD is highly dynamic on picosecond (ps) to nanosecond (ns) timescales[199], it appears to have a strong tendency to adopt marginally stable tertiary folds, allowing two NMR structures of the unbound state determined to date[28,192]. These structures are presumably obtained by stabilizing various conformational sub-states under specific solution

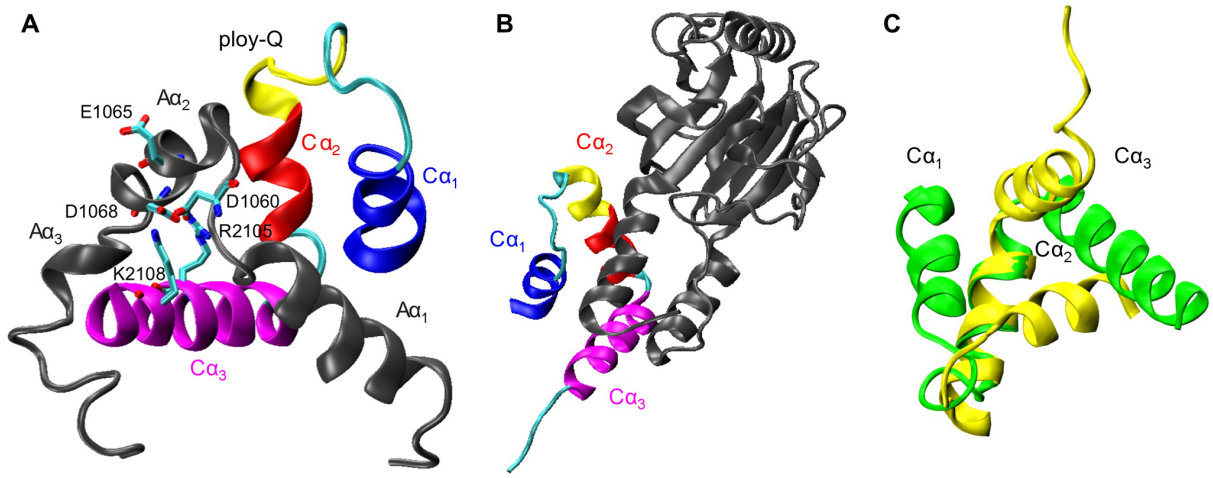


Figure 2.1: Two representative folded conformations of NCBD. (A) NMR structure of the NCBD/ACTR complex (PDB: 1kbh[194], model 1). Both NCBD and ACTR contain 3 helical structure segments, labeled as $C_{\alpha 1}$ (blue), $C_{\alpha 2}$ (red) and $C_{\alpha 3}$ (magenta) in NCBD and $A_{\alpha 1}$, $A_{\alpha 2}$ and $A_{\alpha 3}$ in ACTR. See Methods for specific residue ranges of these helical segments. Several key structural features are also marked, including the poly-Q loop (yellow) linking $C_{\alpha 1}$ and $C_{\alpha 2}$, a buried salt-bridge between NCBD R2105 and ACTR D1068, and several key charged residues adjacent to this buried salt-bridge. (B) X-ray crystal structure of the NCBD/IRF3 complex (PDB: 1zoq[196]). (C) Overlay of the folded structures of NCBD in complex with IRF3 (yellow) and ACTR (green). Only the structured segment (residues 2066-2112) is shown, and the two structures are aligned using the backbone atoms of $C_{\alpha 2}$.

conditions. Particularly intriguing is that the latest NMR structure of free NCBD turns out to be similar to the folded conformation observed when bound to ACTR[28]. Although such pre-existence of folded-like conformations should be considered only as a necessary but insufficient condition for conformational selection-like mechanisms, the unusually high level of residual structures of NCBD strongly suggests a functional role of pre-folding in its coupled binding and folding interactions. In this work, we first exploit implicit solvent-based atomistic simulations and REX enhanced sampling to characterize the conformational properties of free NCBD and ACTR. The roles of preformed structures vs. conformational fluctuation in the NCBD/ACTR interaction are then directly probed using high-temperature unfolding and unbinding simulations in both implicit and explicit solvents. Combined with our recent coarse-grained simulations and existing experimental data, we aim to obtain a detailed mechanistic picture of how residual structures, conformational fluctuations, and electrostatic interactions contribute to efficient synergistic folding of NCBD and ACTR.

2.2 Results

2.2.1 Convergence and validation of the disordered structure ensembles

De novo calculation of the disordered ensembles for IDPs is challenging[1], especially for NCBD that is both of moderate size and apparently with a complex, solution condition-sensitive conformational equilibrium. Our previous works have suggested that implicit solvent coupled with REX enhanced sampling could generate reasonably accurate disordered ensembles for small IDPs, including a 28-residue segment of the kinase inducible domain (KID) of transcription factor CREB[169]. In Figure S2.1, we first test the convergence of the calculated disordered ensembles by examining the dependence of residue helicity on REX simulation time and by comparing results from independent simulations initiated from dramatically different conformations (folding vs. control; see Methods). The sequences of both domains are provided in Methods. Free ACTR appears to be highly disordered with marginal residual helicity. The calculated residual helicity profiles from the control and folding runs converge to similar ones (data not shown). For NCBD, while the time evolution of the calculated residual helicity appears to stabilize over the course of 100 ns in either the control or folding REX simulation, the final profiles from these two independent calculations differ substantially, suggesting that the actual convergence is rather limited. Nonetheless, both the folding and control simulations clearly suggest significant residual helicity in all three helical segments that become stably folded upon binding to various specific targets. Detailed analysis of the conformational ensemble (see below) demonstrates that free NCBD is compact and contains substantial tertiary contacts. These conformational properties of NCBD, coupled with the larger size, contribute to the difficulty of achieving better convergence using the REX/GB protocol. In addition, the current surface area-based treatment of nonpolar solvation can over-stabilize non-specific collapsed states[74,200]. This problem further limits the ability to sufficiently sample accessible tertiary organizations of free NCBD and their

inter-conversions, which is required for achieving good convergence.

Given the limited convergence achieved in the REX simulations of free NCBD and apparent difficulties in substantially improving the level of convergence, we focus on semi-quantitative or qualitative analysis of the conformational properties of NCBD. That is, although significant conformational sub-states sampled by REX may be genuine, the relative stability (population) is not likely to be reliable. Considering that NCBD is experimentally known to be highly helical, the folding simulations (initiated with a fully extended conformation) should take longer to converge, and the disordered ensemble calculated from the control simulation is likely more realistic. Therefore, all the subsequent analysis is based on the ensemble of conformations sampled during the last 60 ns of the 100 ns control REX simulation. In Figure 2.2, we compare the residue helicity of NCBD and ACTR in the free and bound states. The results appear to be fully consistent with the previous NMR secondary chemical shift analysis (Figure 2 of Ref.[199]), showing that all three NCBD helices are largely formed in the unbound state and ACTR is largely free of residual helices. Interestingly, the poly-Q segment of NCBD (residues 2082-2086), although disordered in the NCBD/ACTR complex, is largely helical in the unbound state and extends $C_{\alpha 2}$. This is fully consistent in the NMR chemical shift analysis[199]. Recent sequence correlation analysis has revealed a link between sequence order and binding promiscuity[201,202]. One might expect that the length of the poly-Q stretch might affect conformational flexibility, and furthermore, the ability to interact with diverse targets. We also have analyzed the ensemble distribution of the radius of gyration of free NCBD. The results, shown in Figure S2.2A, confirm that free NCBD is highly compact. Despite a clear lack of convergence, the control and folding simulations appear to sample a set of conformation sub-states with similar characteristic sizes. Direct comparison of the calculated size profiles to one derived from a recent small-angle X-ray scattering (SAXS) study[28] is complicated by the different constructs used and uncertainty in proper inclusion of the solvation shell for a heterogeneous ensemble. Nonetheless, one can estimate that including the disordered N- and C-terminal tails (13 residues total) truncated in the current simulations would increase the radius by 2-3 Å, and that the solvation shell may add another 2-3 Å (estimated

by comparing results from HydroPro[203] and CHARMM). These corrections together bring the calculated radius of the gyration profile close to the SAXS-derived profile that centers around 15.2 Å under "native-like" conditions[28]. Apparent agreement between NMR and SAXS on these ensemble-averaged properties is not sufficient to validate the reliability of the simulations, but it suggests that the simulated ensemble may offer a qualitative or even semi-quantitative characterization of the conformational properties of free NCBD.

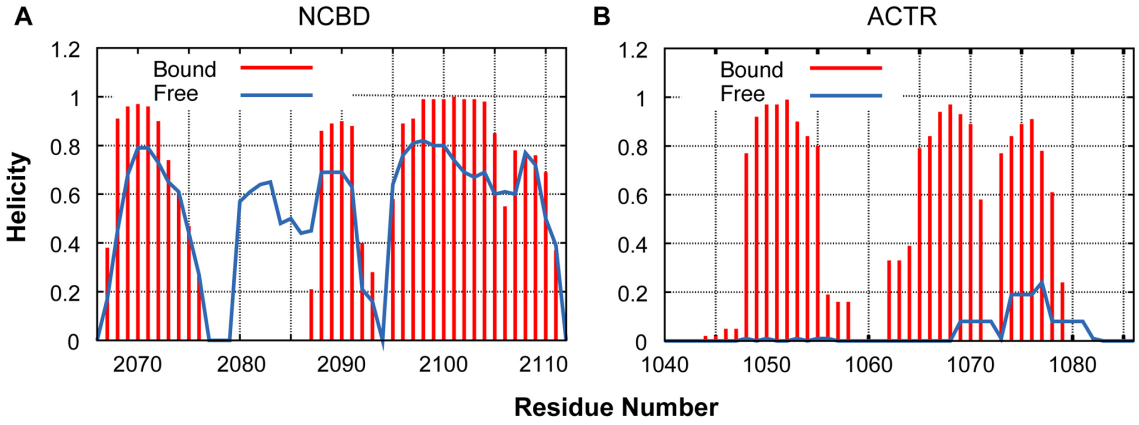


Figure 2.2: Calculated residue helicity of NCBD and ACTR in the free and bound states. Results for the bound state were calculated from a 100 ns control simulation of the complex (see Methods), and those for the free peptides were calculated based the conformationals sampled at 305 K during the last 60 ns of the control REX simulations.

2.2.2 Folded-like conformations in the unbound state of NCBD

Because all three NCBD helices are largely formed in the unbound state, the conformational fluctuation of free NCBD mainly involves tertiary packing of these helices. For example, as shown in Figure 2.1, when aligned using the central helix $C_{\alpha 2}$, the two representative folded conformations of NCBD differ mainly in the orientation of $C_{\alpha 1}$ and slightly less so in that of $C_{\alpha 3}$. Therefore, all conformations of the calculated ensemble first re-oriented by aligning $C_{\alpha 2}$ (to the -z axis) before the orientations of $C_{\alpha 1}$ and $C_{\alpha 3}$ were calculated. Note such analysis also provides an effective description of the tertiary packing even when one or more of the three NCBD segments are not in helical states. The results, shown in Figure 2.3, illustrate that NCBD is

strongly fluctuating and samples a large number of helix configurations, as expected for a molten globule. Intriguingly, free NCBD appears to substantially sample all three distinct conformations that have been observed experimentally so far, either in complexes or in isolation. These folds are represented by PDB structures 1kbh, 1zoq, and 1jjs, respectively. The $C_{\alpha 1}$ orientation of 1kbh and $C_{\alpha 3}$ orientation of 1jjs appear to be least sampled. Nonetheless, conformational sub-states exist with similar orientations, as marked by arrows in panels (C) and (D) of Figure 2.3. Specifically, for 1kbh-like $C_{\alpha 1}$ orientation, the adjacent sub-state contains more parallel (with smaller helix cross angles), and thus tighter, packing of $C_{\alpha 1}$ with $C_{\alpha 2}$, but with a helix interface similar to that of 1kbh. Further structural analysis (see the following paragraph) suggests that such tighter packing is likely a result of helix formation in the poly-Q segment (e.g., see Figure 2.2), which shortens the $C_{\alpha 1}$-$C_{\alpha 2}$ loop and promotes tighter packing.

Clustering analysis was performed to further analyze the structural properties of the major conformational sub-states of free NCBD. The average structures of the six most populated clusters identified using K-means clustering with a 3.0 Å radius are shown in Figure 2.4. Helix configurations for all members of these clusters are shown in Figure S2.3. Interestingly, even though one of the clusters (Figure 2.4D) is similar to the fold observed in 1kbh, most clusters are different from either 1zoq or 1kbh on the whole domain level, as suggested by the large RMSD values. Therefore, even though both individual $C_{\alpha 1}$-$C_{\alpha 2}$ and $C_{\alpha 2}$-$C_{\alpha 3}$ helix pairs sample all three distinct PDB folds, these folded-like configurations of individual helix pair generally do not occur at the same time. Notably, the folded conformations of NCBD in 1kbh and 1zoq have relatively similar $C_{\alpha 2}$-$C_{\alpha 3}$ helix packing (see Figure 2.1C). The packing of $C_{\alpha 2}$ and $C_{\alpha 3}$ also appears to be more restricted in free NCBD compared to that of $C_{\alpha 1}$ and $C_{\alpha 2}$ (e.g., as indicated by a larger "inhibited" red area in Figure 2.3D compared to Figure 2.3C). NCBD has a strong inherent propensity to adopt $C_{\alpha 2}$-$C_{\alpha 3}$ configurations analogous to those in 1kbh and 1zoq. Such persistent folded-like conformations of free NCBD could contribute to recruitment of specific targets such as ACTR and IRF3, allowing NCBD to adopt different final structures by docking the more flexibly linked $C_{\alpha 1}$ into different positions. Another interesting observation is that the poly-Q segment appears to

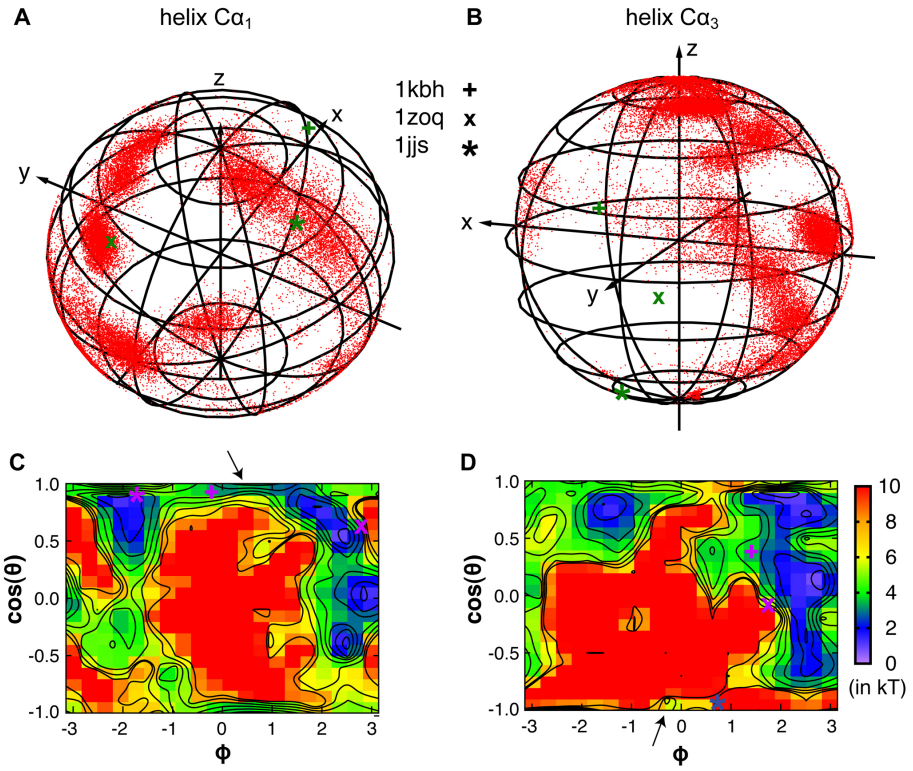


Figure 2.3: Orientations of NCBD $C_{\alpha 1}$ and $C_{\alpha 3}$ with respect to $C_{\alpha 2}$ in the unbound state. All conformations sampled at 305 K during the last 60 ns of the control REX simulation were first aligned using the backbone atoms of $C_{\alpha 2}$, and then reoriented such that $C_{\alpha 2}$ was aligned with -z axis. The orientations as observed in three distinct folds of NCBD, represented by PDB 1kbh, 1zoq, and 1jj (also see Figure 2.1), are marked with "+", "×", and "∗", respectively. Note that different colors for the same symbol may be used in different panels for clarity. In panels (C) and (D), ϕ and θ are the inclination and azimuth angles of the spherical coordinate system. Note that the PMFs were computed using $\sin(\theta)$ instead of θ itself as an order parameter to remove the Jacobian entropy contribution. The range shown corresponds to θ = 0 (top) to π (bottom). Contours are drawn at every kT up to 7 kT, with k being the Boltzmann factor.

be capable of readily switching between helical and coil states. Such conformational fluctuations could allow NCBD to adapt to different substrates, extending the $C_{\alpha 2}$ helix when bound with IRF3 but becoming more disordered when in complex with ACTR (see Figure 1).

2.2.3 Induced folding-like mechanism on the baseline level

Although the REX simulations provide intriguing insights into the possible residual structures of free NCBD, how these conformational properties contribute to synergistic folding of NCBD with

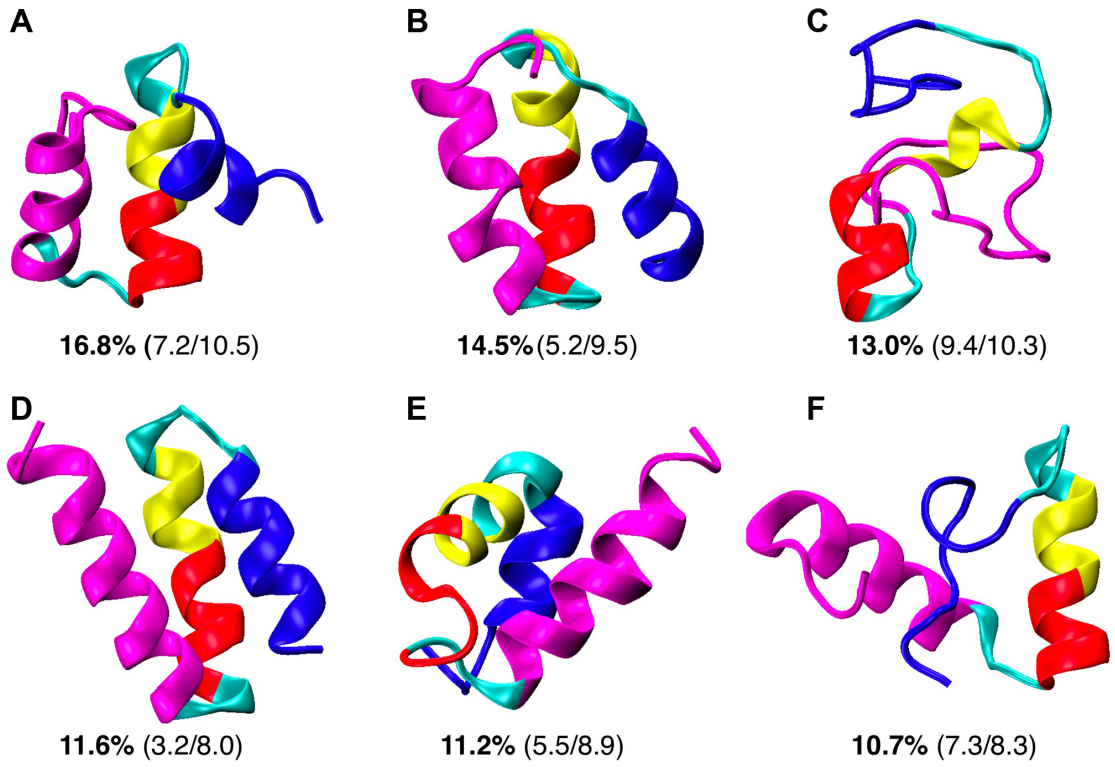


Figure 2.4: Averaged structures of the six most populated clusters of free NCBD. These clusters are identified based on the conformations sampled at 305 K during the last 60 ns of the 100 ns REX control simulation. All structures were aligned by minimizing the backbone RMSD of $C_{\alpha 2}$ (the red segments) and visualized in the same view. The numbers below each structure are the population of the cluster and backbone RMSD values from the folded conformations in 1kbh and 1zoq (see Figure 2.1). The protein is colored in the same fashion as in Figure 2.1A-B.

ACTR is not obvious based on these equilibrium simulations alone. For this, one could calculate the coupled binding and folding free energy surfaces[30,164] or transition paths[33] to more directly clarify the recognition mechanism and probe the roles of residual structures vs. pre-folding in specific finding. However, given the moderate size and relatively complex topology, such calculations can be extremely demanding using an atomistic physics-based force field for the NCBD/ACTR complex. Instead, temperature-induced unfolding and unbinding simulations may be used to effectively infer the molecular processes of coupled binding and folding. A key assumption is that binding/folding is largely a reverse of unbinding/unfolding. An important concern is that the transition states or the most probable transition paths might depend on temperature[86]. Nonetheless, high-temperature unfolding simulations have so far proven quite successful

for studying folding and interaction of many proteins, including IDPs[87–90].

A 100 ns equilibrium simulation of the complex was first performed at 300 K, which confirms that the native fold (model 1 of PDB:1kbh) is very stable in the GBSW/MS2 implicit solvent (see Figure S2.4). Subsequent pilot simulations suggest 475 K to be optimal for simulating unbinding and unfolding of the NCBD/ACTR complex in GBSW/MS2 (e.g., see Figure S2.5). In Figure 2.5, we compare the time evolution of various fractions of native contacts computed from 50 independent unfolding simulations at 475 K. The fraction of native intermolecular interactions (Q_{inter}) is used to describe binding, and the fraction of native tertiary intramolecular interactions (Q_{NCBD}) is used for folding of NCBD. As shown in Figure S2.6, ACTR is completely devoid of any inter-helix tertiary contacts in the NCBD/ACTR complex. Because ACTR is largely free of residual structures in the unbound state, the overall helicity (α_{ACTR}) is used to effectively monitor (binding-induced) folding of ACTR. On the baseline level, all unfolding and unbinding kinetics appear to be reasonably well represented by single exponential functions. The fitted kinetic data is summarized in Table 2.1. The secondary (helix) unfolding of NCBD is predicted to be the slowest process (α_{NCBD}; green traces in Figure 2.5), which is expected given the high level of residual structures in unbound NCBD; however, both the ACTR (helix) and NCBD tertiary unfolding appear to be significantly faster than unbinding. This result suggests that binding occurs prior to the folding of both ACTR and NCBD; that is, both ACTR and NCBD follow induced-folding-like mechanisms on the baseline level in the GBSW/MS2 implicit solvent. Considering the apparent tendency of NCBD to pre-fold (see above), this result is somewhat surprising, but it highlights the importance of conformational fluctuations and nonspecific binding in specific recognition of IDPs, even for IDPs with significant residual structures like NCBD. Significant heterogeneity is apparent in the unfolding/unbinding pathways of NCBD/ACTR and is partially reflected in substantial ruggedness that remains in the curves shown in Figure 2.5 (e.g., compared with a previous explicit solvent unfolding simulation of the p53-MDM2 complex, where 10 independent 10-ns simulations at 498 K were sufficient to yield much smoother curves[87]). The complex fully disassociates within 10 ns in only 6 out of the 50 independent runs. In examining the

Table 2.1: Unfolding and unbinding kinetic constants at 475 K.

	τ (ns)	A	B	R^2
Q_{inter}	1.61	0.23	0.17	0.94
$C_{\alpha 1}$	0.26	0.37	0.00	0.97
$C_{\alpha 2}$	1.52	0.25	0.034	0.84
$C_{\alpha 3}$	2.20	0.24	0.23	0.94
$A_{\alpha 1}$	0.80	0.34	0.037	0.94
$A_{\alpha 2}$	1.39	0.26	0.26	0.86
$A_{\alpha 3}$	2.93	0.32	0.30	0.86
Q_{NCBD}	0.94	0.27	0.055	0.90
α_{NCBD}	1.76	0.31	0.15	0.96
α_{ACTR}	1.38	0.20	0.13	0.89

All curves were fitted by single exponentials $Ae^{(-t/\tau)} + B$. R is the correlation coefficient of fitting. See the captions of Figures 2.5 and 2.7 for the definitions of various contact fractions.

unbinding/unfolding characteristics at a lower temperature of 450 K (see Figure S2.7), we found the heterogeneity of unfolding/unbinding pathways to be even more evident. In addition, the complex appears trapped in some intermediate states and does not fully unfold/unbind even after 20 ns. Nonetheless, unfolding of either ACTR or NCBD appears to lag behind unbinding, which is consistent with the induced-folding baseline mechanisms predicted at 475 K.

2.2.4 Binding and folding intermediates involving the NCBD and ACTR C-terminal segments

Indications are that binding-induced folding of NCBD and ACTR is not simply 2-state-like. For example, decay of Q_{NCBD} and α_{ACTR} appears to pause at $\sim$2 ns (red and blue traces in Figure 2.5), which could suggest a common intermediate state where ACTR and NCBD are partially bound and folded. The decay curves are too noisy (partially due to underlying heterogeneity) for

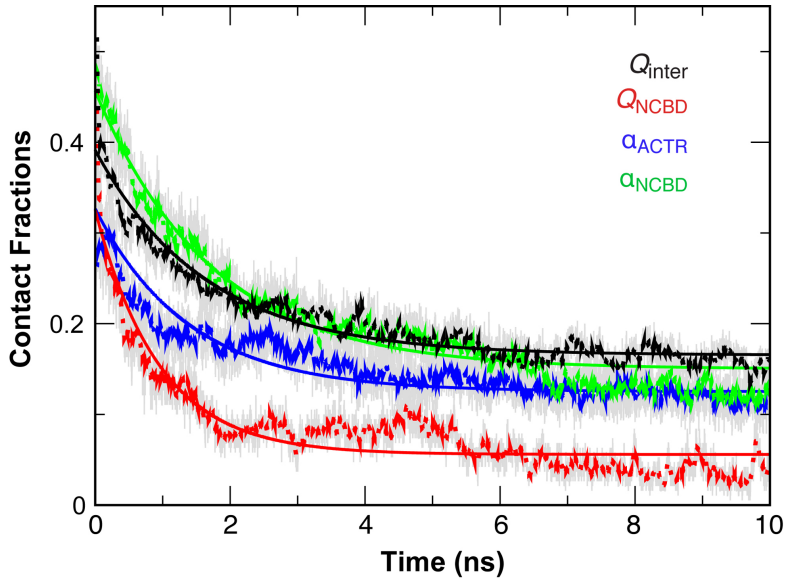


Figure 2.5: Evolutions of various contact fractions of the NCBD/ACTR complex at 475 K. All contact factions and helicities were computed by averaging results from 50 independent unfolding simulations. Q_{inter} and Q_{NCBD} denote the native fractions of intermolecular and NCBD tertiary intramolecular contacts, and α_{ACTR} and α_{NCBD} denote the overall helicities of ACTR and NCBD. Significant fluctuation remains in the raw averaged contact fraction traces (grey traces), and thus 50-ps running averages (dotted traces) are plotted for clarity. The solid traces correspond to the best single exponential fits (see Table 2.1 for the fitted kinetic constants). Note that both Q_{inter} and Q_{NCBD} quickly decrease from above 0.8 to $\sim$ 0.4 during the first 10-20 ps. The initial decay is out of the plotting range and not shown for clarity.

reliable kinetic fitting using double exponential functions. Therefore, we constructed (pseudo) unbinding and unfolding free energy surfaces based on statistics collected from the first 5 ns of the unfolding simulations. Note that the system is not at equilibrium during this time frame, so the resulting free energy profiles are not equilibrated (and thus strongly dependent on initial conditions). Nonetheless, the profiles provide qualitative approximations of the true free energy surfaces[204]. As shown in Figure 2.6A, an intermediate state is evident at $Q_{inter} \sim 0.25$ and Q_{NCBD} $\sim$ 0.15. Interestingly, a similar key intermediate state also has been predicted in our recent topology-based modeling of the NCBD/ACTR complex[163]. A strong resemblance between the free energy surface is shown in Figure 2.6A and the result derived from topology-based modeling (Figure 5A of reference[163]). Both the atomistic simulations (see further analysis detailed in the following paragraph) and topology-based modeling predict that the intermediate state mainly involves the C-terminal segments of NCBD and ACTR. Such a prediction appears highly consistent

with a recent H/D exchange mass spectrometry (H/D-MS) study[58], where peptide segments within the C-terminal regions of both NCBD and ACTR were found to have much larger protection factors compared with those mapped into other folded regions of the complex.

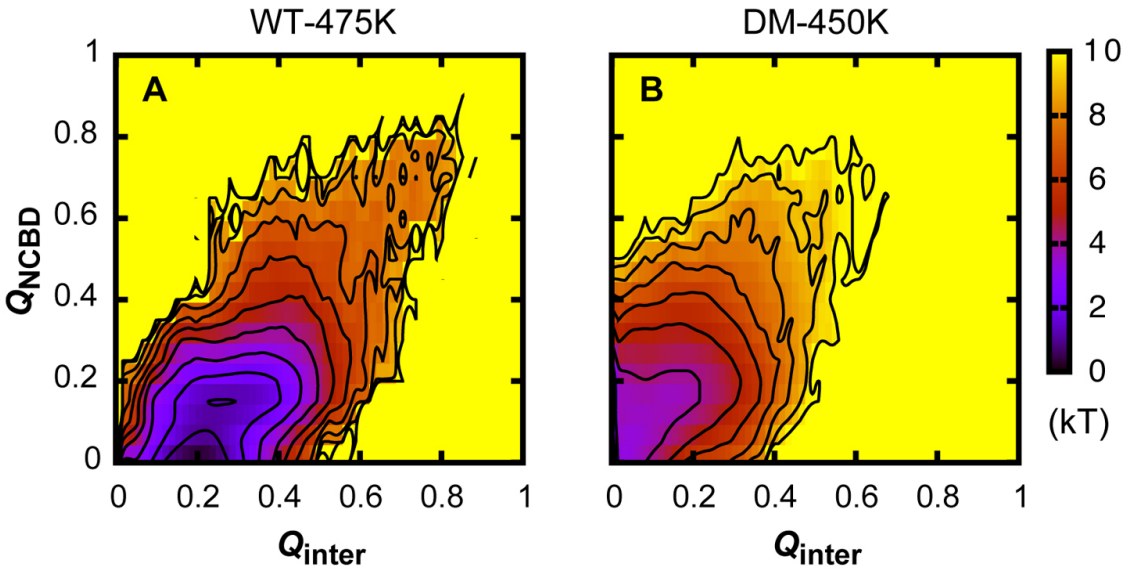


Figure 2.6: Free energy surfaces of (un)binding and (un)folding of the wild-type and mutant NCBD/ACTR complex. (A) Computed from the first 5 ns of 50 independent simulations of the wild-type NCBD/ACTR complex (WT) at 475 K. (B) Computed from the first 5 ns of 50 independent simulations of the double-Leu mutant complex (NCBD:R2105L/ACTR:D1068L; DM) at 450 K. Contours are drawn at every kT.

In Figure 2.7, we further examined the binding kinetics of individual NCBD and ACTR helices. The kinetic data derived from fitting to single exponential functions is summarized in Table 2.1. The analysis shows that $A_{\alpha 3}$ and $C_{\alpha 3}$ unbind with the largest half times, τ = 2.93 ns and 2.20 ns, respectively, which are greater than that of the overall intermolecular interaction formation (τ = 1.61 ns). This result indicates that binding is mainly initiated by the C-terminal helices. In contrast, the first helices of NCBD and ACTR unbind much faster then the second and third helices. In fact, unbinding of $A_{\alpha 1}$ and $C_{\alpha 1}$ occurs even faster than folding of either NCBD or ACTR (as described by Q_{NCBD} and α_{ACTR}, see Table 2.1). These kinetic rates are consistent with a multi-stage synergistic folding process, where NCBD and ACTR first bind rapidly through the C-terminal segments, forming intermediates that are mainly stabilized by native-like interactions between α_2 and α_3 helices. This first step appears to be highly cooperative

(e.g., see Figure 2.6A), although indications are that both induced folding and conformational selection might contribute[163]. Interestingly, the transition between the intermediate and bound states appears largely conformational selection-like where NCBD and ACTR folding precedes $A_{\alpha 1}$ and $C_{\alpha 1}$ binding. Formation of the partially folded core appears to facilitate the rest of NCBD to fold into native-like conformations, allowing $A_{\alpha 1}$ and $C_{\alpha 1}$ to rapidly form native intermolecular interactions en route to the fully folded bound state. Taken together, even though the synergistic folding of NCBD and ACTR follows an induced folding-like baseline mechanism (where binding precedes folding on the overall level), detailed analysis reveals multiple stages of induced folding and conformational selection. Such a mechanism closely resembles an "extended conformational selection" recently proposed by Csermely *et al.*[205,206] and is remarkably consistent with our recent topology-based modeling of the NCBD/ACTR complex[163].

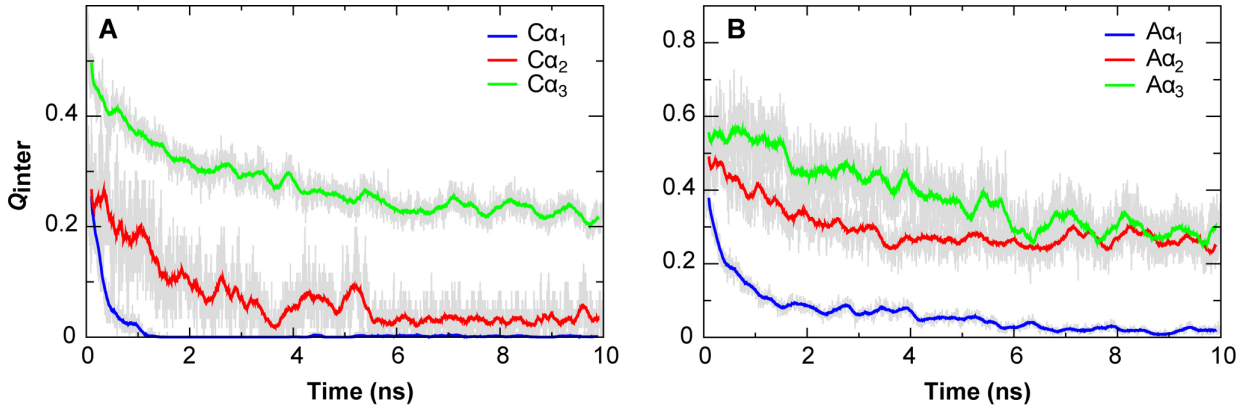


Figure 2.7: Evolutions of the fractions of native intermolecular interactions of individual helices of the NCBD/ACTR complex. The grey traces were calculated from averaging 50 unfolding simulations at 475 K and the colored traces are 50-ps running averages.

2.2.5 Native and non-native salt-bridges in encounter complexes and intermediates

One of the most notable features of the NCBD/ACTR complex is a buried salt-bridge between NCBD R2105 and ACTR D1068[194] (see Figure 2.1A), which is also conserved in the interaction of NCBD with p53 TAD[195]. Interestingly, this buried salt-bridge is part of a local network of salt-bridges that could form between multiple complementary charges, including R2105 and

K2108 of NCBD and D1060, E1065, and D1068 of ACTR (see Figure 2.1A). This network of native and non-native salt-bridges appears to play a significant role in stabilizing the putative intermediate state, either thermodynamically or kinetically. Although most individual salt-bridges frequently break and reform during individual unfolding simulations (see Figure S2.8), on average they largely persist throughout the 10 ns unfolding simulations at 475 K and hinder the transition from the partially bound intermediates to fully disassociated ones (see Figure 2.8). Out of the 50 unfolding simulations at 475K, the complexes fully dissociate only by the end of 10 ns simulations in six cases. The native salt-bridges, between NCBD R2105 and ACTR D1068 and D1060, are the most protected. As shown in Figure 2.8, they are the most preserved and remain formed over 80% of the time throughout the simulations (blue and black traces in Figure 2.8A). NCBD K2108 is adjacent to R2015 and close enough to interact with ACTR D1068 and D1060, but these salt-bridges are more solvent-exposed and thus slightly less preserved during high-temperature simulations. The side chain of ACTR E1065 is positioned away from NCBD in the native structure. Partial unfolding of $A_{\alpha 2}$ allows E1065 to rotate and participate in the salt-bridge network with 10-30% probability by the end of the 10 ns simulation at 475 K (purple and red traces in Figure 2.8A).

The conformational heterogeneity of the intermediate state does not permit reliable free energy calculations to quantify the contribution of salt-bridge interactions to stability. Nonetheless, previous mutagenesis studies have suggested that the buried salt-bridge between NCBD R2105 and ACTR D1068 contributes minimally to binding affinity[198]. The salt-bridge network likely could not significantly stabilize the intermediate state thermodynamically, either, which raises a concern that the observed persistence of the local salt-bridge network is artificial, such as due to over-stabilization of charge-charge interactions in the GBSW/MS2 implicit solvent. To address this concern, we first examine the potential of mean forces (PMFs) between Arg and Asp side chain analogs in TIP3P and GBSW/MS2. The results, summarized in Figure 2.8, show that GBSW/MS2 actually slightly under-stabilizes the Arg-Asp interaction compared with TIP3P, either in a constrained head-to-head configuration (which was used in the force field optimization[76]) or when fully unconstrained. In particular, configurationally unconstrained Arg-Asp

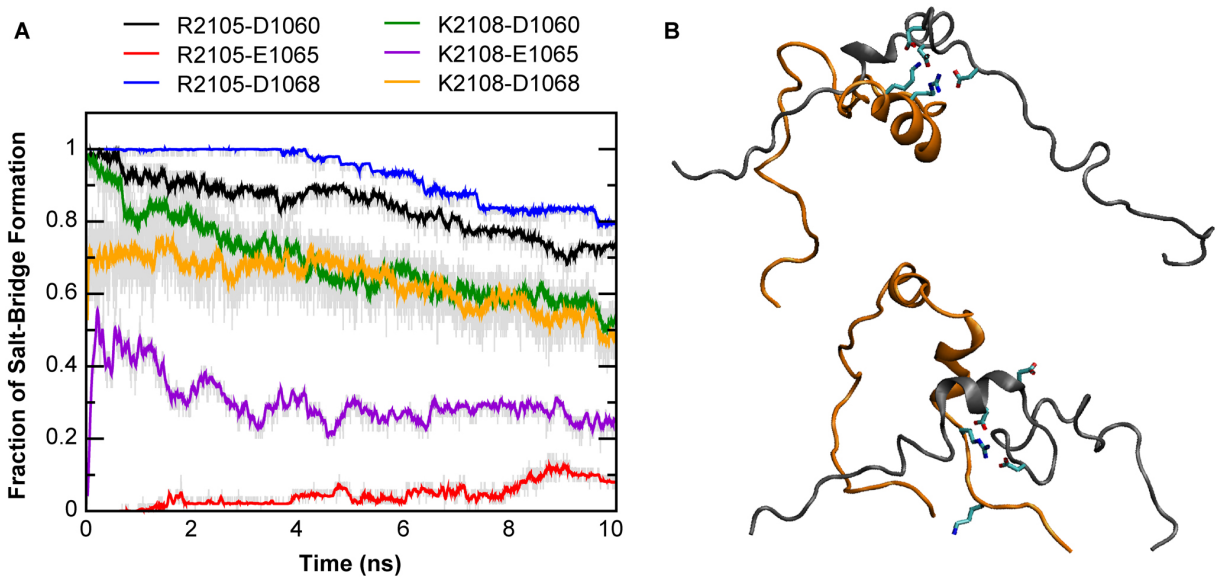


Figure 2.8: Native and non-native salt-bridges in the NCBD/ACTR interaction. (A) Evolutions of average probabilities of various salt-bridge interactions during unfolding simulations at 475 K. Arg and Glu/Asp residues were considered in contact if the carbonyl carbon and Arg CZ distance was no greater than 5 Å, and Lys and Glu/Asp residues were considered in contact if the side chain carbonyl carbon and amide nitrogen distance is no greater than 4 Å. (B) Two representative final conformations after 10 ns simulations at 475 K. NCBD and ACTR are colored orange and gray, respectively. The side chains of key charged residues are also shown, including NCBD R2105 and K2108 and ACTR D1060, E1065, and D1068. The snapshot on the top represents a case where all six possible salt-bridges are formed, and the one at the bottom represents a case where only the native ones, between NCBD R2105 and ACTR D1068 and D1060, are formed.

interaction is unstable in GBSW/MS2 (Figure 2.9B). Therefore, the observed stabilization effects of salt-bridges on the intermediates are likely of a kinetic nature. Such kinetic stabilization arises from substantial desolvation barriers in disassociation of salt-bridges, particularly in partially folded protein environments where the side chain configurations are restricted (e.g., see Figure 2.9A). With a concentrated local network of salt-bridges, very large desovaltion barriers can be expected for complete dissociation of NCBD and ACTR, which explains why only a small fraction of the high-temperature simulations (6 out of 50) successfully reached the fully unbound state in 10 ns.

To further confirm that the observed salt-bridge network is not an artifact of implicit solvent, a set of 10 unfolding simulations was performed in TIP3P explicit solvent at 500 K. Most simulations

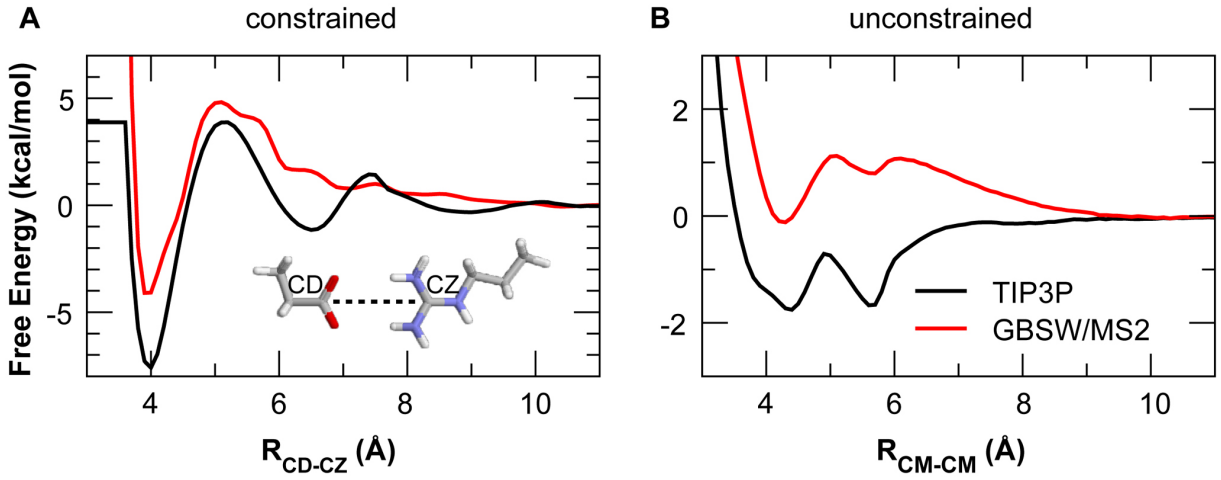


Figure 2.9: Potential of mean forces of the Arg-Asp interactions in implicit and explicit solvents. All profiles were calculated using umbrella sampling and WHAM (see Methods). The center-of-mass (CM) separation is used as the order parameter in the unconstrained PMF.

were terminated between 3 to 4 ns when the complex size exceeded the periodic box dimensions. The lengths of these simulations are insufficient to capture degrees of unfolding and unbinding similar to implicit solvent simulations, and the number of trials is insufficient to obtain smooth curves for kinetic fitting. Nonetheless, visual inspection of simulation trajectories as well as examination of the evolution of various contact fractions support an unbinding and unfolding mechanism that is consistent with the one derived from implicit solvent simulations (see Figure S2.9). The same set of native and non-native interactions, particularly the buried one between NCBD R2105 and ACTR D1068 (blue trace in Figure S2.9B), persist and appear to stabilize the partially unbound and unfolded intermediates. Note that the helical secondary structures are substantially over-stabilized in these explicit solvent simulations (e.g., see the blue trace in Figure S2.9A). This is a known artifact of the current version CHARMM22/CMAP explicit solvent force field[78,207,208].

A double-Leu mutant complex follows a similar unfolding and unbinding mechanism

A control simulation of the double-Leu mutant complex, NCBD:R2105L/ACTR: D1068L, at 300 K suggests that the native fold remains stable in the GBSW/MS2 implicit solvent (data not shown). A set of 50 unfolding simulations was carried out at 450 K to further investigate the role of the buried salt-bridge in synergistic folding. The heterogeneity of the unfolding/unbinding pathway observed in the wild-type complex (e.g., see Figure 2.5) is even more pronounced without the buried salt-bridge. All averaged time traces of contact fractions remain very noisy (e.g., see Figure S2.10). Most traces cannot be satisfactorily fitted to either single or double exponential functions, preventing quantitative analysis of unfolding and unbinding kinetics. Nonetheless, the pseudo binding and folding free energy surface computed from the first 5 ns of the unfolding trajectories appears to resemble that from simulations of the wild-type complex (see Figure 2.6). In particular, a similar intermediate state exists at $Q_{inter} \sim 0.2$ and $Q_{NCBD} \sim 0.15$; however, the small free energy barrier separating the intermediate and fully unbound states in Figure 2.6A is largely absent in Figure 2.6B. Removal of NCBD:R2105L largely disrupts the local salt-bridge network. The intermediate state appears to have much shorter resident times, and can quickly fluctuate to the fully unbound state. Importantly, examination of the evolution of intermolecular contact factions of individual NCBD and ACTR helices, shown in Figure S2.10, supports that the mutant complex largely follows a similar, albeit more heterogeneous, unbinding and unfolding mechanism, with the N-terminal α_1 helices disassociated first (black traces in Figure S2.10B-C). These results suggest the local salt-bridge network does not appear to fundamentally modulate the recognition mechanism. Instead, it mainly augments a productive synergistic folding mechanism inherent in (the topology of) the NCBD/ACTR complex, by transiently stabilizing a key on-pathway intermediate state to facilitate complete folding en route to the specific complex.

2.3 Discussion

With one of the highest levels of residual structures, NCBD is an intriguing model system for understanding the roles of residual structure vs. conformational fluctuations in coupled binding and folding of IDPs. We have combined equilibrium and non-equilibrium simulations using physics-based, atomistic protein force fields to characterize the conformational properties of unbound NCBD and ACTR and to understand how these properties facilitate efficient synergistic folding of these two IDPs. The calculation recapitulates that free NCBD has folded-like helical content, is strongly fluctuating, and samples a wide range of tertiary configurations, which is consistent with the previous notion that free NCBD is a molten globule[199]. Intriguingly, the calculated disordered ensemble of NCBD contains significant populations with helical packings that are highly similar to all those previously observed experimentally in isolation and in complex with various targets. Observations of such pre-folded conformations, especially for IDPs with significant residual structures like NCBD, could be considered strong evidence for conformational selection-like mechanisms, where such preformed structural elements provide initial binding sites. Direct examination of the unfolding and unbinding pathways in high-temperature simulations, however, shows that both ACTR and NCBD tend to unfold first before unbinding, suggesting an induced folding-like baseline mechanism for their synergistic folding. This seemingly surprising result appears to be consistent with the observation that, although individual $C_{\alpha1}$/$C_{\alpha2}$ and $C_{\alpha2}$/$C_{\alpha3}$ helical pair samples folded-like packing with substantial probability, these configurations rarely occur simultaneously. Therefore, population of folded-like tertiary conformations on the whole domain level is insufficient to support conformational selection-like mechanisms on the baseline level. Further analysis reveals an on-pathway intermediate state that mainly involves the C-terminal helices of ACTR and NCBD, which also has been predicted by a recent coarse-grained simulation study using topology-based models[163]. Importantly, existence of such a major intermediate state also appears to be consistent with a recent H/D-MS experiments showing that peptide segments within the C-terminal regions of NCBD and ACTR have much larger

protection factors compared with those mapped into other regions of the complex[58]. Our kinetic analysis suggests that, once the initial mini folding core is formed, the N-terminal helix of NCBD folds rapidly (Table 2.1), allowing subsequent facile binding and folding the ACTR N-terminal helix en route to the final specific complex. Therefore, although the baseline mechanism is induced folding-like, conformational selection actually occurs at local levels. Together with our recent topology-based modeling study[163], the atomistic simulations strongly support the prediction that synergistic folding of NCBD and ACTR follows the "extended conformational selection" mechanism[205]. Our topology-based modeling of the NCBD/ACTR interaction[163] has revealed a separate, albeit less prevalent, pathway where binding is initiated by the N-terminal α_1 helices. These mechanistic insights on synergistic folding of NCBD and ACTR, derived from the atomistic and coarse-grained simulations, are summarized in Figure 2.10.

An intriguing interplay appears to exist among residual structures, conformational fluctuations, and electrostatic interactions to facilitate the rate-limiting step of forming the partially folded intermediates. The NCBD $C_{\alpha 2}$/$C_{\alpha 3}$ helix-turn-helix motif appear to be conformationally more restricted (Figure 2.2D), whereas the C-terminus of $C_{\alpha 3}$ retains the least amount of helical content and is considerably more heterogeneous (Figure S2.2B). Both features were also observed in the previous NMR chemical shift and relaxation analysis[199]. Such a balance of residual structures and conformational fluctuations is likely important for the NCBD C-terminal to act as a key initiation point for coupled folding and binding to ACTR and other proteins. Another novel insight provided by the current atomistic simulations is the role of a local network of native and non-native salt-bridges in transiently stabilizing the intermediates. These salt-bridge interactions likely do not contribute substantially to the thermodynamic stability of either the intermediates or the final specific complex[198], but substantial desolvation barriers involved in breaking up these interactions in a conformationally restricted protein environment (e.g., Figure 2.9A) can extend the resident time of the intermediates to allow the rest of the complex to fold with higher efficiency. As demonstrated using a dual-transition state kinetic model[173], efficient folding upon encounter is necessary for achieving facile binding at or near the diffusion-limited basal binding rate, a

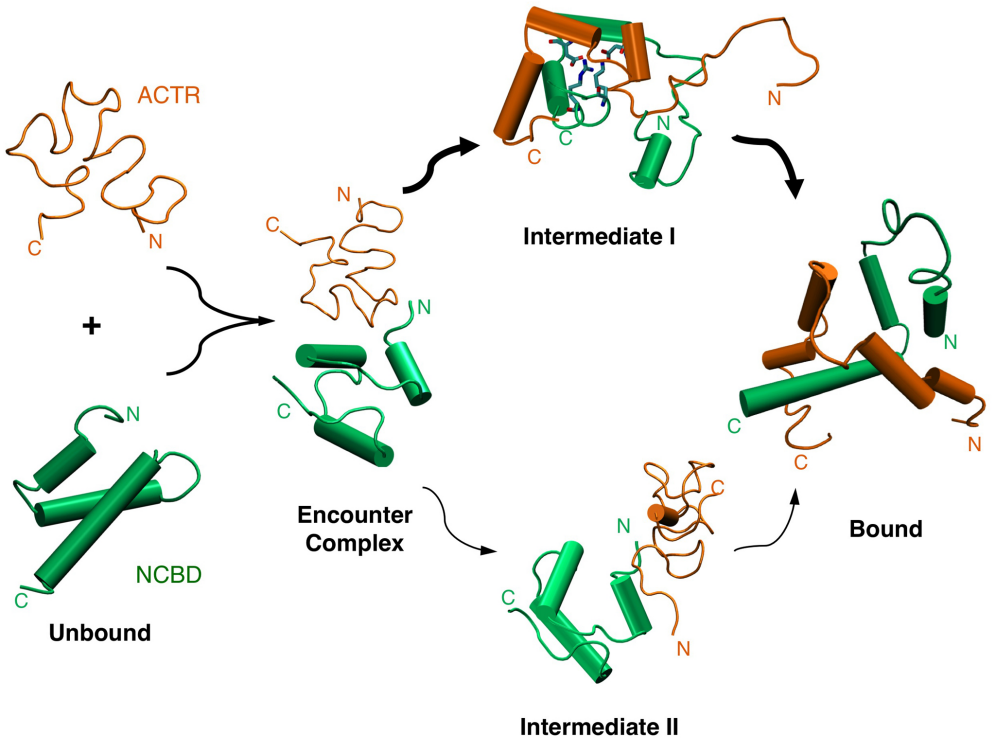


Figure 2.10: Overall mechanism of synergistic folding of NCBD and ACTR. The schematic view is based on the current atomistic simulations as well as the previous coarse-grained modeling[163]. It illustrates that unbound ACTR is largely unstructured and NCBD has significant helical structures. The nonspecific encounter complexes could evolve toward the bound state through two parallel pathways. The salt-bridge network that plays a key role in stabilizing Intermediate I along the prevalent pathway is also illustrated.

highly desirable property for signaling and regulatory IDPs that need to constantly evade protein degradation machinery in cell. IDPs are known to be enriched with charges[12]. NCBD and ACTR are no exceptions, with +6 and -8 net charges, respectively (including the flanking loops that remain disordered in the complex[194]). These enriched charges hinder (independent) folding and can protest against aggregation. In addition, long-range electrostatic interactions between these large numbers of complementary charges on NCBD and ACTR could dramatically enhance the encounter rate, similar to electrostatic steering, which is known to be important in interactions of globular protein[209]. Furthermore, the complementary pattern of charge, especially within the predicted mini folding core involving the C-termini (Figure 2.1), suggests that long-range electrostatic interactions could further promote folding-competent encounter complexes before

transiently stabilizing the on-pathway intermediates via formation of short-range salt-bridge network. These effects can enhance the efficiency of folding upon encounter to promote facile recognition.

The current study also reveals important limitations in both the protein force field accuracy and sampling capability, especially for modeling IDPs of moderate sizes and with complex residual structures. These limitations underscore the importance of continual development of the protein force field, with increased focus on balancing various competing interactions to allow an accurate description of not only a few (native) folds but also the whole conformational equilibrium[74,210]. Sampling methodologies clearly need to improve. The standard temperature REX-MD has failed to achieve convergence for the disordered ensemble of NCBD within 100 ns. Besides limited simulation timescale, certain limitations of the implicit solvent protein force field also contributed. In particular, current empirical protein models have been shown to contain a systematic bias to over-stabilize protein-protein interactions[211,212]. Furthermore, simple surface area-based estimation of the nonpolar solvation free energy employed in most current implicit solvent models also tends to over-stabilize nonspecific compact protein states[74]. The standard temperature REX-MD clearly has limited ability to sample alternative deeply trapped low energy states with high efficiency. These limitations together have also prevented us from more directly investigating the proposed mechanistic roles of electrostatic interactions using atomistic simulations. Despite these outstanding limitations, the key mechanistic features derived from atomistic physics-based simulations, coarse-grained topology-based modeling, and various biophysical measurements are remarkably consistent, which suggests that an integration of multi-scale modeling and experimentation can provide a viable approach for understanding the functional and control of IDPs.

2.4 Methods

2.4.1 REX/GBSW simulations of free NCBD and ACTR

Only segments of the NCBD and ACTR domains that are structured in the complex are included in the current simulations, which include residues 2066-2112 for NCBD (in mouse CBP numbering; SALQD LLRTL KSPSS PQQQQ QVLNI LKSNP QLMAA FIKQR$_{2105}$ TAKYV AN) and residues 1040-1086 for the ACTR domain (in human ACTR numbering; E GQSDE RALLD QLHTL LSNTD ATGLE EID$_{1068}$RA LGIPE LVNQG QALEP K). The peptide termini are neutralized using with either acetyl (Ace) or amine (NH2) groups. A previously optimized GBSW/MS2 model was used in all implicit solvent simulations unless otherwise noted[76]. This model adopts an effective approximation of the molecular surface for defining the solute-solvent boundary, which is believed to be more physical compared to the van der Waals-like surface used in the original GBSW model[213,214]. Importantly, the GBSW/MS2 model has also been carefully optimized to balance solvation and intramolecular interactions and can reasonably capture the competition between α and β secondary structures. Specifically for NCBD/ACTR, the structure of the complex (PDB: 1kbh[194]) remains stable in the GBSW/MS2 force field for over 100 ns, but substantially deviates from the native conformation in the original GBSW protein force field (see Figure S2.4).

REX was used to enhance the sampling of the accessible conformational space of free NCBD and ACTR. For this, the Multiscale Modeling Tools for Structural Biology (MMTSB) toolset[215] (http://www.mmtsb.org) was used in conjunction with CHARMM[63,216]. The basic idea of REX is to simulate multiple non-interacting replicas at different temperatures simultaneously. Periodically, one attempts to exchange the simulation temperatures between pairs of replicas based on a Metropolis criterion derived from the detail balance principle. As such, not only the resulting random walk in the temperature space facilitates the system to cross the energy barriers and exploit the conformational space more efficiently, but proper canonical ensembles are also generated at all temperatures, allowing direct calculation of thermodynamic properties for comparison with

experiments. We performed two independent REX simulations for each peptide, initiated from the folded structure extracted from the complex (control) and a fully extended conformation (folding), respectively. Comparison of the calculated structure ensembles from these independent control and folding runs with dramatically different initial conditions allows rigorous assessment of the convergence. In each REX simulation, 16 replicas were simulated at temperatures exponentially distributed from 270 to 500 K. SHAKE[217] was applied to fix the lengths of all hydrogen-related bonds, allowing a 2.0 fs molecular dynamics (MD) time step. Temperature exchanges between neighboring replicas were attempted every 2 ps, and the total length of each REX simulation was 100 ns (50,000 REX cycles). Similar REX/GBSW protocols have proven effective in calculating the disordered structural ensembles for other IDPs (albeit of smaller sizes than NCBD and ACTR studied in the current work)[30,169]. All analysis was performed based on the conformations sampled during the last 60 ns of the control simulation at 305 K (where most existing experimental data were acquired), unless otherwise noted. The orientations of helical segments (1044-1058, 1063-1071, 1072-1080 in ACTR; 2067-2076, 2086-2091, 2095-2110 in NCBD) were calculated using the Chothia-Levitt-Richardson algorithm[218] as implemented in CHARMM. The K-means clustering algorithm as implemented in the MMTSB toolset was used to cluster the calculated disordered ensembles based on mutual C_α RMSD distances. Various clustering radii ranging from 1.5 to 4.5 Å were tested before an optimal radius of 3.0 Å was used for the final clustering results presented. All molecular visualizations were generated using the VMD software[219].

2.4.2 Room temperature and high-temperature simulations of the wild-type and mutant NCBD/ACTR complexes

The same peptide segments defined above were included the simulations of the complex. The model 1 from the NMR ensemble (PDB: 1kbh) was first equilibrated in the GBSW/MS2 implicit solvent using energy minimization and short MD with weak harmonic positional restraints imposed on all backbone heavy atoms. Subsequently, a 160 ns unrestrained simulation was

performed at 300 K to examine the structural stability and dynamics of the complex near its native basin. The native structure of the NCBD:R2105L/ACTR:D1068L double-Leu mutant complex was prepared by computational mutagenesis and then equilibrated using a similar protocol as described above. To identify the optimal temperatures for unbinding/unfolding simulations, a series of pilot simulations was performed at temperatures ranging from 350 K to 500 K (e.g., see Figure S2.5). At the optimal temperature, the complex should unfold/unbind within tractable time scales (e.g., 10-20 ns) while retaining important details of the unfolding/unbinding pathways. Once such optimal temperatures were chosen (450-475 K for the wild-type and 450 K for the mutant), 50 independent high-temperature simulations of 10-20 ns in length were initiated from the equilibrated native structures with different initial velocities. The results presented in this work are averages computed from 50 unfolding simulations unless otherwise noted. For native fraction analysis, a list of native tertiary contacts (shown in Figure S2.6) was first identified using the equilibrated native structure based on side chain minimal heavy atom distances with a 4.2 Å cutoff. The native contacts were then divided into inter-molecular and intra-molecular categories. In analysis of the high-temperature simulation trajectories, a contact was considered formed when the minimal heavy atom distance between two side chains was no greater than 4.5 Å. Helicity of various helical segments was calculated based on the hydrogen bonding patterns using the COOR SECS module of CHARMM.

2.4.3 Explicit solvent high-temperature simulations

Additional high-temperature unfolding and unbinding simulations of the wild-type complex were performed in TIP3P water to examine the unfolding/unbinding pathway and in particular the putative role of the buried salt-bridge between NCBD:R2105 and ACTR:D1068 in (transiently) stabilizing the intermediate state(s). For this, the equilibrated NCBD/ACTR complex was placed in a cubic water box with periodic boundary conditions imposed. The final solvated system contains 9176 TIP3P water molecules and the box size is $\sim$65 Å. Two potassium ions were

added to neutralize the total charge. The proteins were described by the CHARMM22/CMAP protein force field[69,182–184]. The particle mesh Ewald method was used for long-range electrostatic interactions[220], and the van de Waals interactions were smoothly switched off from 12 to 13 Å. Lengths of all hydrogen-related bonds were kept constant with SHAKE[217], and the MD time step was 2 fs. After 10 ps of NPT equilibration at 300 K, a set of 10 independent NVT productions was carried out at 500 K up to 10 ns until the dimensions of the proteins exceed those of the periodic box. The dynamic time step was reduced to 1 fs in the NVT production simulations for numerical stability.

2.4.4 Free energy calculations

An umbrella sampling protocol[77] was used to compute the PMFs between the side chains of Asp and Arg, either constrained in a head-to-head configuration[77] (see Figure 2.9) or allowed to freely rotate. In the constrained setup, the side chains were allowed to move only in fixed orientations along the reaction coordinate (indicated by a dashed line in Figure 2.9), enforced using the MMFP module in CHARMM. For explicit solvent simulations, solutes were solvated by either $\sim$710 TIP3P waters in a rectangular box (for the constrained PMF) or by $\sim$1040 TIP3P waters in a truncated octahedral box (for the unconstrained PMF). Periodic boundary conditions were imposed. Non-bonded and other setups are identical to those described above for explicit solvent high-temperature simulations. Harmonic restraint potentials were placed every 0.5 Å along the reaction coordinate with a force constant of 5.0 kcal/mol/Å^2. For each umbrella-sampling window, the system was first equilibrated for 60 ps, followed by 2 ns (constrained PMF) or 4 ns (unconstrained PMF) NPT production at 300 K and 1 atm. The final PMFs were calculated using the weighted histogram analysis method (WHAM)[142]. The constrained PMF in GBSW/MS2 was computed by direct translation of the side chains along the reaction coordinate, and the unconstrained PMF in GBSW/MS2 was computed in the same umbrella sampling protocol except that implicit solvent was used instead of TIP3P waters. Convergence of the PMFs was examined

by comparing results from the first and second halves of the data and was shown to be on the order of 0.2 kcal/mol.

2.5 Supplemental Materials

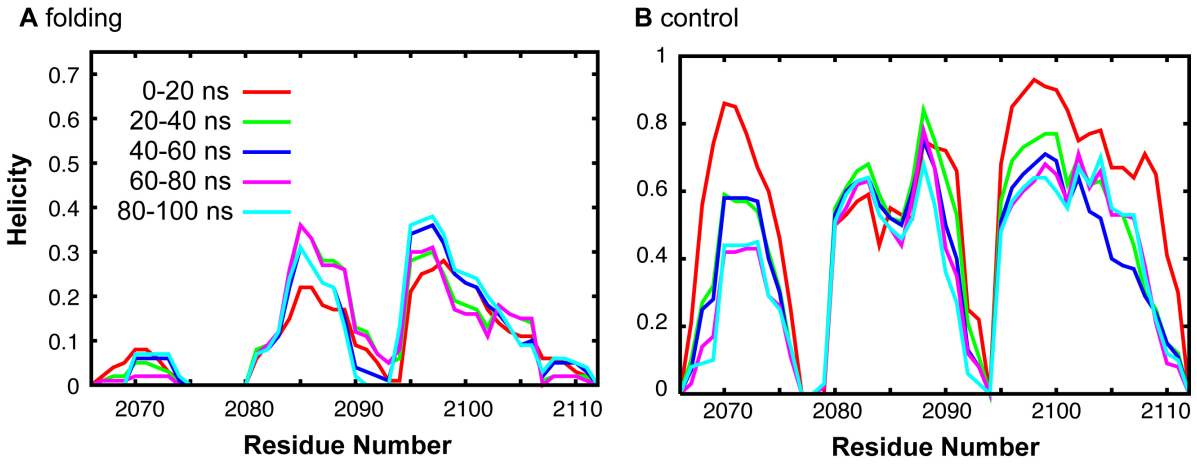


Figure S2.1: Convergence of the calculated residue helicity of free NCBD. Residue helicities calculated using different segments of the folding (A) and control (B) REX simulations are shown. Only conformations sampled at 305 K were included in the analysis.

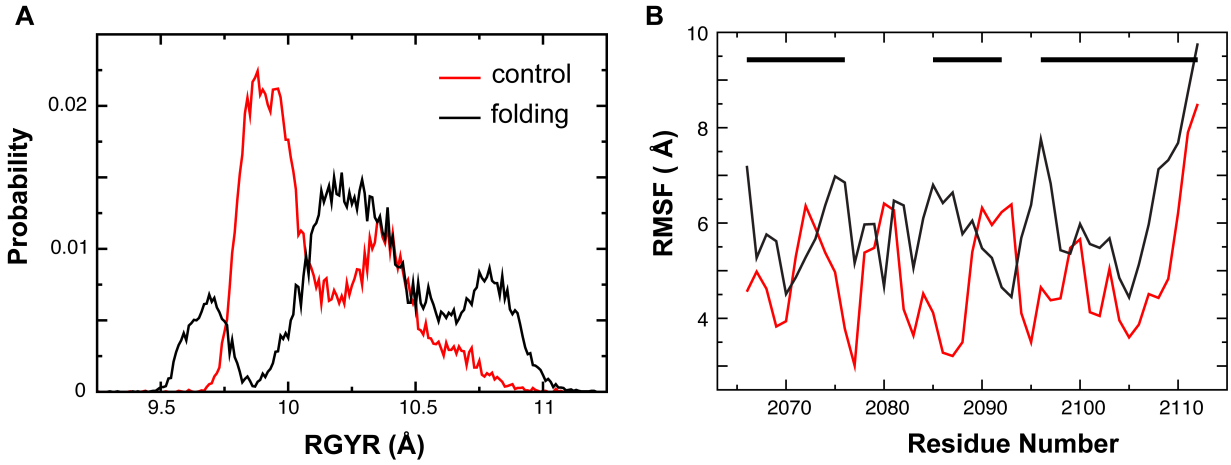


Figure S2.2: Additional conformational properties of free NCBD. (A) Distributions of the radius of gyration, and (B) C_α root mean square fluctuation (RMSF) profiles at 305 K, calculated from the last 60 ns of control (red traces) and folding (black traces) simulations.

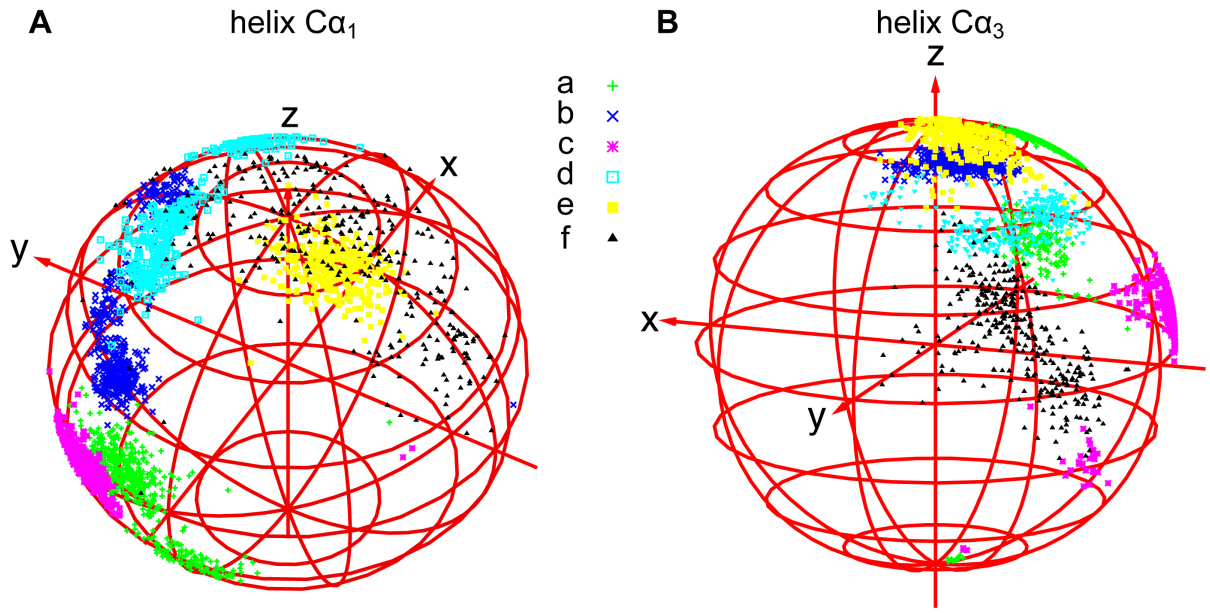


Figure S2.3: Orientations of NCBD $C_{\alpha 1}$ and $C_{\alpha 3}$ with respect to $C_{\alpha 2}$. Conformations that belong to the six most populated clusters of free NCBD sampled at 305 K are color-coded. See the caption of Figure 2.3 in the main text for additional information.

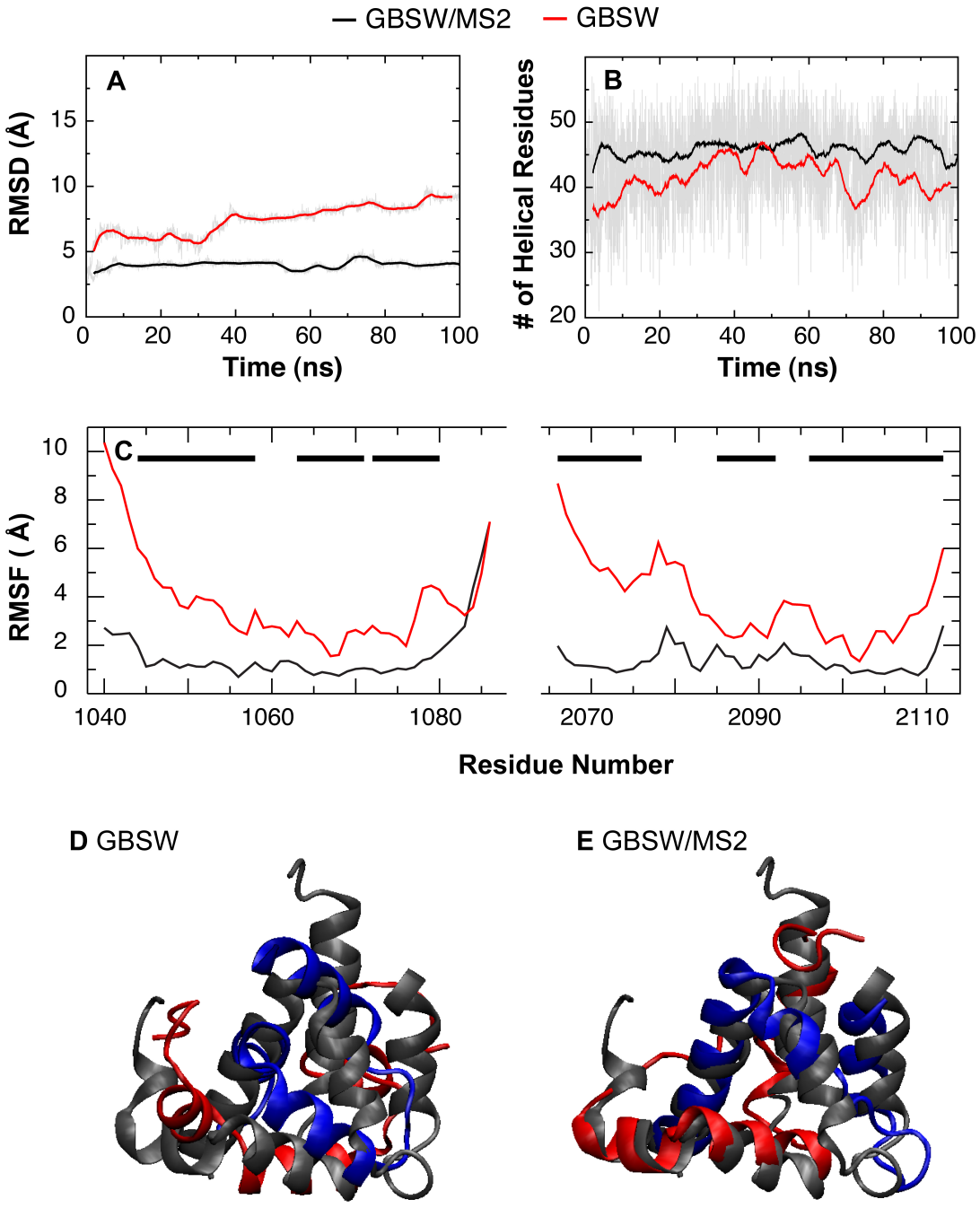


Figure S2.4: Summary of control simulations of the NCBD/ACTR complex in GBSW and GBSW/MS2. (A) Backbone RMSD as a function of time. (B) Number of helical residues as a function of time. (C) The C_α RMSF profiles computed from the last 50 ns of the 100 ns control simulations. Helical segments of ACTR and NCBD are marked. (D)-(E) The final snapshots overlaid with the PDB structure (shown in gray cartoon). The results suggest that the NCBD/ACTR complex is unstable in GBSW both at the secondary and tertiary levels. In contrast, the complex remains reasonably stable in GBSW/MS2, with significant fluctuations mainly observed in the C-terminal segment of ACTR, and to a lesser extent in the NCBD C-terminus (see panel C).

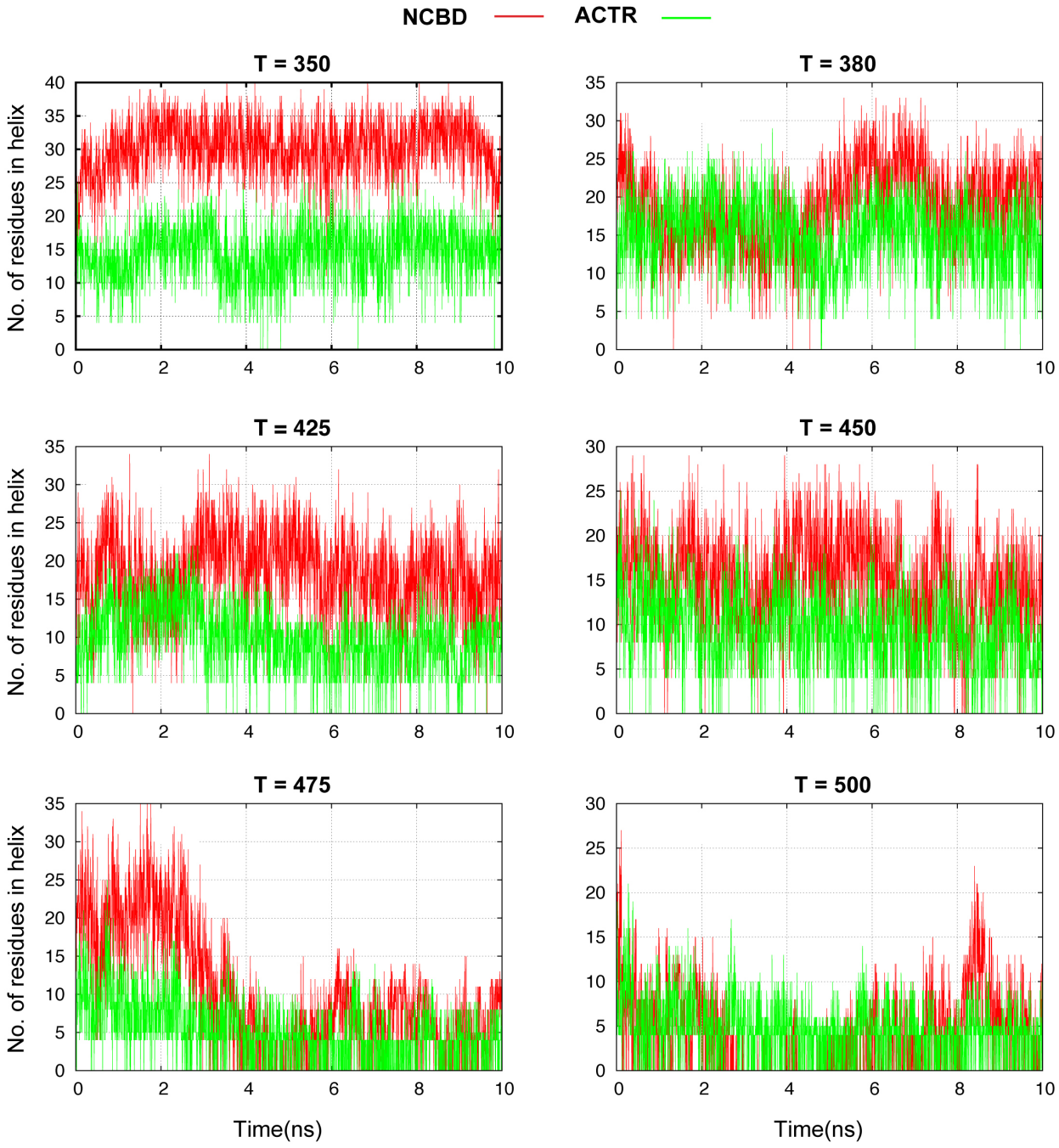


Figure S2.5: Trial unfolding simulations in GBSW/MS2 at different temperatures. The numbers of helical residues of NCBD and ACTR are monitored to detect the unbinding/unfolding of the complex.

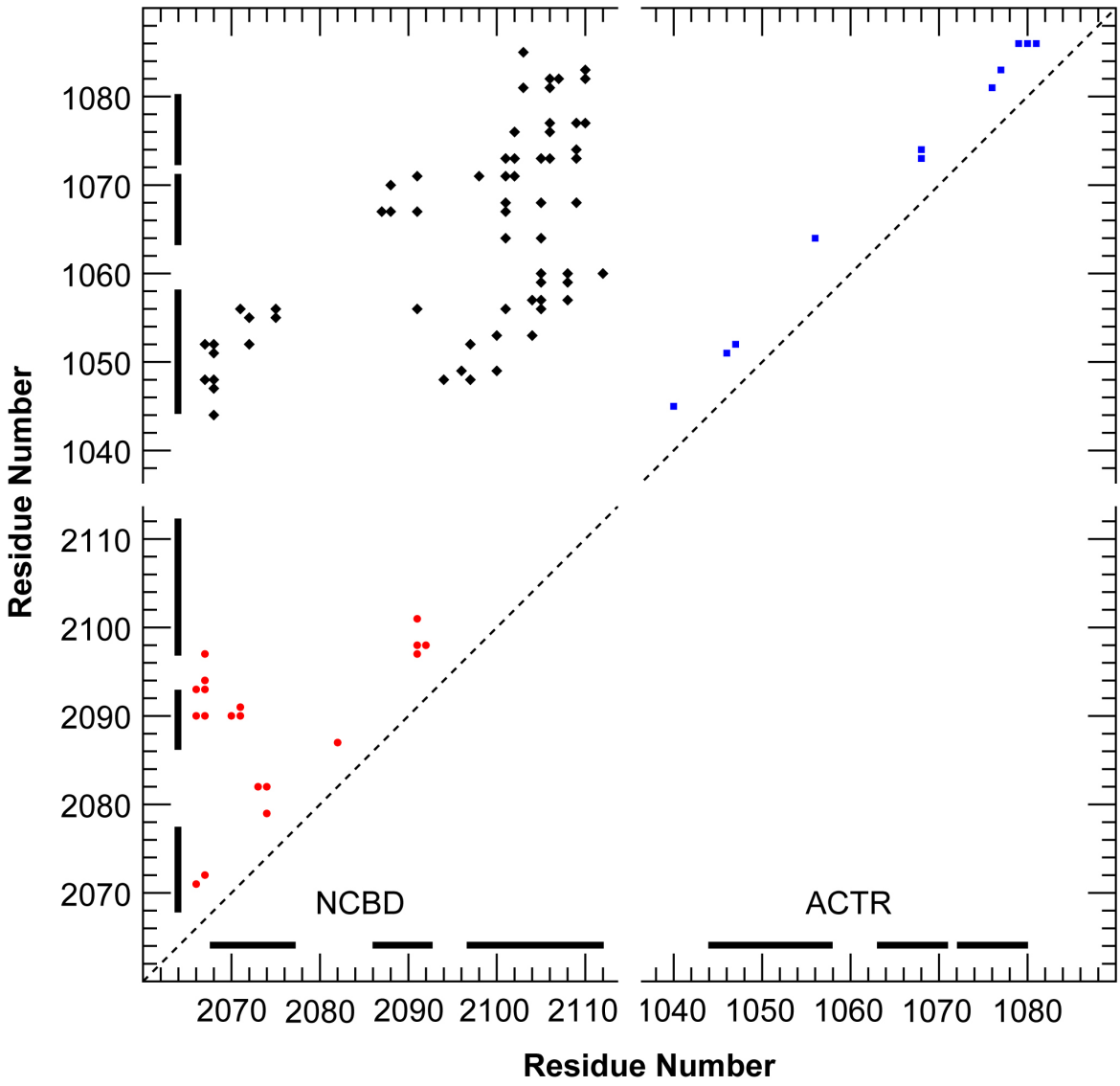


Figure S2.6: Tertiary contacts of the NCBD/ACTR complex. The contacts were derived based on the first model of PDB:1kbh. Residues are considered in contact if the minimal heavy atom distance is no more than 4.2 Å. The black bars indicate the ranges of all helical segments in NCBD and ACTR. Although there are a large number of intermolecular contacts (62; black dots), there are only 11 (blue dots) and 19 (red dots) tertiary intramolecular contacts for ACTR and NCBD, respectively.

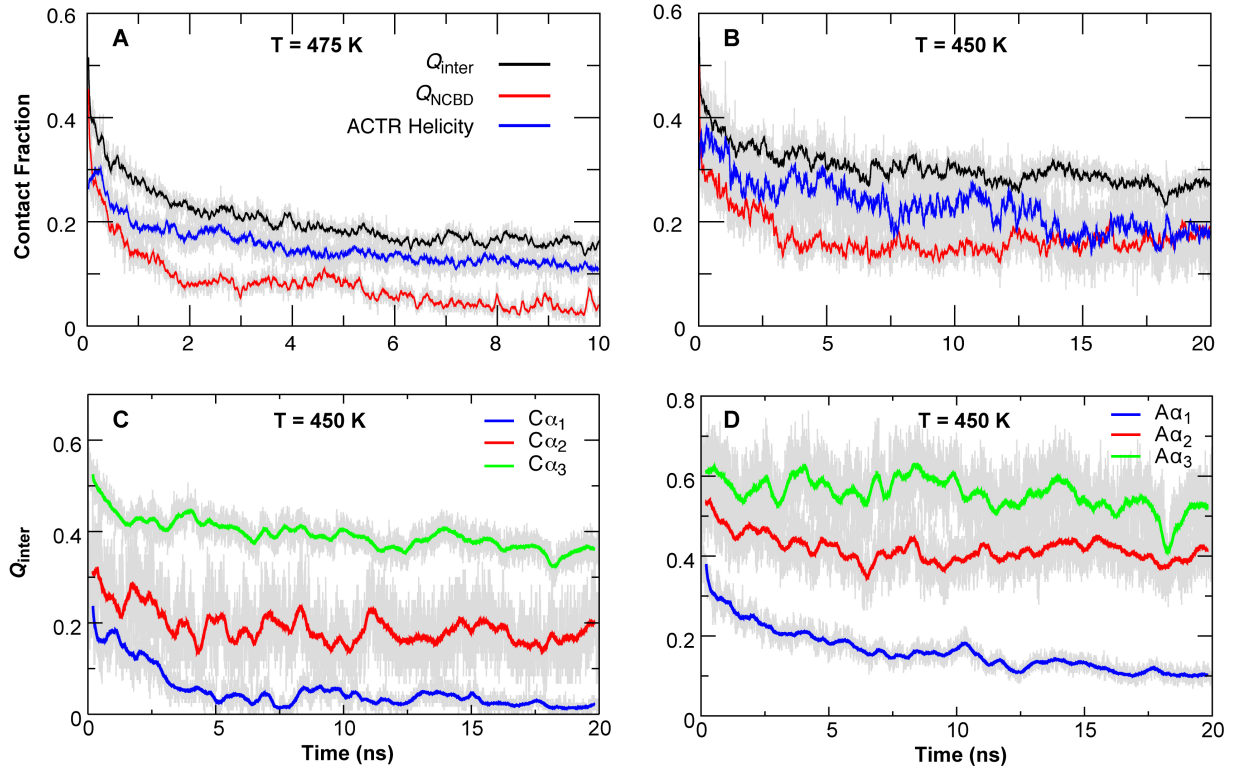


Figure S2.7: Evolution of various contact fractions in GBSW/MS2 simulations at 450 and 475 K. The grey traces were calculated from averaging 50 independent simulations at corresponding temperatures, and the colored traces are 50-ps running averages.

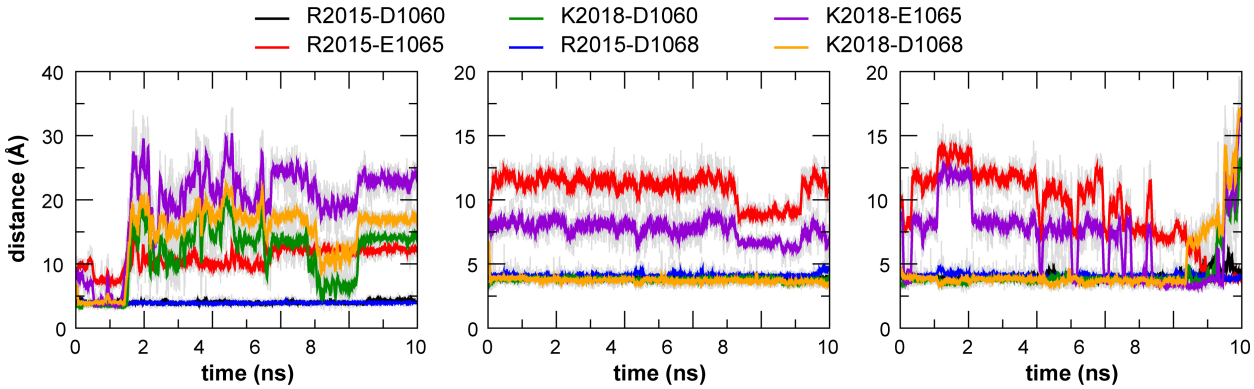


Figure S2.8: Distances between key charged residues during three representative unfolding simulations at 475 K. For Arg and Glu/Asp pairs, the distance between the side chain carbonyl carbon and Arg C_Z distance is shown. For Lys and Glu/Asp pairs, the distance between the side chain carbonyl carbon and amide nitrogen is shown.

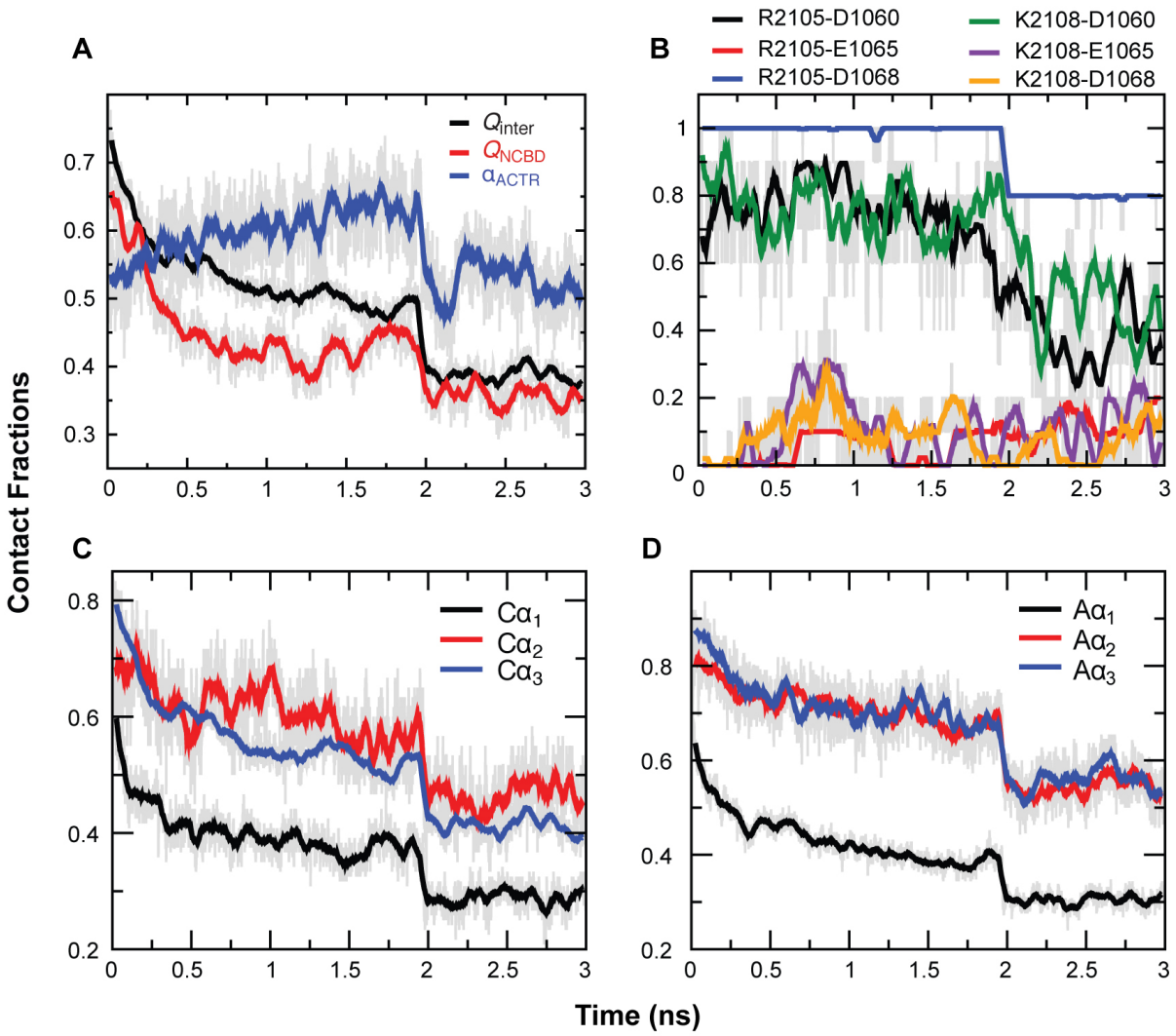


Figure S2.9: Evolution of various contact fractions during unfolding simulations in TIP3P at 500 K. All curves were calculated from averaging 10 independent simulations of 3 to 4 ns in length (only the first 3 ns are shown). The grey traces were calculated from averaging 50 independent simulations, and the colored traces are 50-ps running averages. The results are consistent with key observations derived from GBSW/MS2 simulations. Specifically, 1) the baseline mechanism for coupled binding and folding of NCBD is an induced folding-like one, where binding precedes folding (Panel A); Fitting of Q_{inter} and Q_{NCBD} traces to single exponential functions yields half times, τ = 0.35 ns and 0.25 ns, respectively. 2) The C-terminal segments initiate binding (thus the first helices unbind the first; see black traces in Panels C-D); 3) the local native and non-native salt-bridges persist in the partially unfolded and partially unbound intermediate state (Panel B). Note that the helical secondary structures appear to be over-stabilized (e.g., see the blue trace in Panel A), which is a known artifact of the current version CHARMM22/CMAP explicit solvent force field.

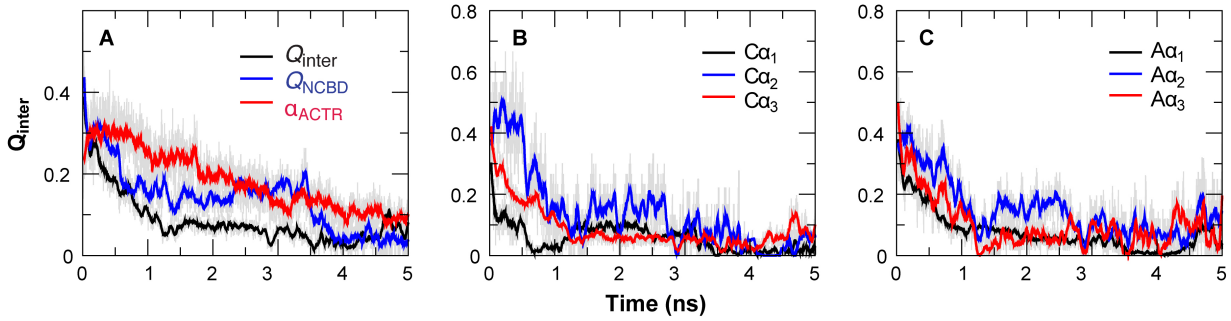


Figure S2.10: Evolution of various contact fractions during unfolding simulations of the mutant NCBD/ACTR complex at 450 K. The grey traces were calculated from averaging 50 independent simulations, and the colored traces are 50-ps running averages. The simulations were 15 ns in length. The complex unfolds rapidly and thus only results from the first 5 ns are shown.

Chapter 3

Synergistic Folding of Two Intrinsically Disordered Proteins: Searching for Conformational Selection*

Abstract

Intrinsically disordered proteins (IDPs) lack stable structures under physiological conditions but often fold into stable structures upon specific binding. These coupled binding and folding processes underlie the organization of cellular regulatory networks, and a mechanistic understanding is thus of fundamental importance. Here, we investigated the synergistic folding of two IDPs, namely, the NCBD domain of transcription coactivator CBP and the p160 steroid receptor coactivator ACTR, using a topology-based model that was carefully calibrated to balance intrinsic folding propensities and intermolecular interactions. As one of the most structured IDPs, NCBD is a plausible candidate that interacts through conformational selection-like mechanisms, where binding is mainly initiated by pre-existing folded-like conformations. Indeed, the simulations demonstrate that, even though binding and folding of both NCBD and ACTR is highly cooperative on the baseline level, the tertiary folding of NCBD is best described by the "extended conformational selection" model that involves multiple stages of selection and induced folding. The simulations further predict that the NCBD/ACTR recognition is mainly initiated by forming a mini folded core that includes the second and third helices of NCBD and ACTR. These predictions

*Reproduced with permission of Royal Society of Chemistry, from "Synergistic folding of two intrinsically disordered proteins: searching for conformational selection", D. Ganguly, W. Zhang and J. Chen, Mol. BioSyst., 2012, 8, 198, Copyright 2012, Royal Society of Chemistry. http://pubs.rsc.org/en/content/articlelanding/2011/mb/c1mb05156c.
Author Contributions: Conceived and designed the experiments: JC DG WZ. Performed the experiments: DG WZ JC. Analyzed the data: DG WZ JC. Contributed reagents/materials/analysis tools: DG JC. Wrote the paper: DG JC WZ.

are fully consistent with independent physics-based atomistic simulations as well as a recent experimental mapping of the H/D exchange protection factors. The current work thus adds to the limited number of existing mechanistic studies of coupled binding and folding of IDPs, and provides a first direct demonstration of how conformational selection might contribute to efficient recognition of IDPs. Interestingly, even for highly structured IDPs like NCBD, the recognition is initiated by the more disordered C-terminal segment and with substantial contribution from induced folding. Together with existing studies of IDP interaction mechanisms, this argues that induced folding is likely prevalent in IDP-protein interaction, and emphasizes the importance of understanding how IDPs manage to fold efficiently upon (nonspecific) binding. Success of the current study also further supports the notion that, with careful calibration, topology-based models can be effective tools for mechanistic study of IDP interaction and regulation, especially when combined with physics-based atomistic simulations and experiments.

3.1 Introduction

Cellular signaling and regulation frequently involve proteins or protein segments that lack stable tertiary folds under physiological conditions and instead exist as heterogeneous and presumably dynamic ensembles of disordered structures[1,3,15,146,147]. Such intrinsically disordered proteins (IDPs) often fold into stable structures upon binding to specific targets. It is important to understand the mechanisms of these coupled binding and folding interactions, as they underlie the organization of regulatory networks for cellular signaling and decision-making. IDPs are also extensively implicated in various human diseases including cancer, neurodegenerative diseases and diabetes[17]. Mechanistic understanding of IDP interactions and regulation can thus aid in assessing related human diseases and devising rational strategies to modulate IDP functions for therapeutic purposes. In particular, signaling and regulatory IDPs arguably represent a novel class of potential drug targets[221]. Several small molecules have been successfully developed to bind IDPs and interfere with their interactions using high-throughput screening[148,222,223]. However, the structural plasticity that allows IDPs to function as versatile regulators poses a significant challenge for rational optimization of the potential drug molecules. The structure of the bound IDP complex alone is not likely going to be sufficient. Instead, an in-depth understanding of how coupled binding and folding occurs and how this process might be modulated by drug molecules is expected to be necessary.

At the baseline level, coupled binding and folding could follow two ideal mechanisms, namely, induced folding and conformational selection. These two extreme mechanisms differ in the kinetic ordering of the binding and folding events: (nonspecific) binding precedes folding in induced folding, and vice versa in conformational selection. Importantly, these mechanisms emphasize different conformational properties of IDPs for interaction. Conformational selection requires the pre-existence of folded-like conformations in the unbound state, and further argues that such preformed structural elements play a main role in initiating recognition[24,27,156]. In

contrast, induced folding emphasizes intrinsic flexibility and nonspecific binding for efficient interaction. Under induced folding scenario, the specific features of the residual structures in the unbound state do not directly affect recognition. Instead, it is the overall level of residual structures that plays a functional role, which is to modulate the binding thermodynamics through the entropic cost of folding. Therefore, such a seemingly semantic classification of the baseline mechanism provides a necessary starting point for understanding how recognition of a specific IDP may be regulated or modulated, such as by post-translational modifications, amino acid replacements, cellular environment, and drug molecules. Note that actual IDP interactions are not expected to follow either ideal mechanism exclusively. Both mechanisms could play roles, such as at different stages of coupled binding and folding[205,224]. There might also be dependence on the solution conditions[225] and even the nature of the specific target.

Residual structures often persist in unbound IDPs[3]. Intriguingly, these residual structures often resemble the folded conformations adopted in complexes[27–29]. Such observations have been frequently considered as evidence for conformational selection-like mechanisms of IDP interactions[24,27–29,162]. However, pre-existence of folded-like conformations is not sufficient evidence for conformational selection. Instead, one needs to further clarify whether the preformed structures play a significant role in initiating binding, such as by examining the free energy surfaces and transition state ensembles of coupled binding and folding, or, more directly, by comparing the time-scales (or equivalently rate constants) of binding and folding transitions[173,225]. For example, previous atomistic simulations of the extreme C-terminus of tumor suppressor p53 reveal that, while the free peptide appears to sample several distinct folded-like conformations observed experimentally in various complexes, its interaction with one of its specific targets, S100B($\beta\beta$), is mainly initiated by nonspecific binding of unfolded conformations[30]. Interestingly, the p53 peptide does not appear to be an unusual case, and evidence has recently accumulated to suggest that induced folding is likely prevalent in IDP-protein interactions[1,26]. Induced folding has been consistently observed in mechanistic studies of IDP interaction from experiments[31,32,59] and simulations[33,62,90,93,165]. Additional evidence of induced folding comes from kinetic data showing

that stabilizing native-like structures in unbound IDPs actually reduce the binding rate[226,227]. Theoretical considerations based on the dynamic energy landscape view have predicted that induced folding would prevail with stronger and longer-range intermolecular interactions[228]. This appears to be the case for IDP-protein interactions: structural plasticity for adopting distinct folded states is considered a hallmark of regulatory IDPs[22,149]; therefore, intermolecular interactions do overwhelm intrinsic folding prior to binding and dictate binding-induced folding of IDPs.

Despite the compelling arguments that can be made above for the prevalence of induced folding, conformational selection could play important or even dominant roles for some IDPs. One such possible example is the nuclear-receptor co-activator binding domain (NCBD) of transcription coactivator CREB-binding protein (CBP). It is one of the most structured IDPs that have been characterized so far. Free NCBD is highly helical with molten globule characteristics[198,199]. Four folded structures of NCBD have been determined, in complex with the trans-activation domain (TAD) of tumor suppressor p53[195], the p160 steroid receptor co-activator ACTR[194], the steroid receptor co-activator 1 (SRC1)[197], and the interferon regulatory factor 3 (IRF3)[196], respectively. In these complexes, NCBD adopts two distinct folds, which mainly differ in the tertiary packing of three similar helices. Two representative folded structures of NCBD, as observed in the NCBD/ACTR and NCBD/IRF3 complexes, are shown in Figure 3.1. The structures of NCBD in complex with SRC1 and p53 are similar to that with ACTR. NCBD appears to have a strong tendency to pre-fold, and it is possible to stabilize various conformational sub-states of the unbound NCBD by tuning the solution conditions. For example, two structures of free NCBD have been determined by solution nuclear magnetic resonance (NMR)[28,192]. Intriguingly, the recent NMR structure of free NCBD turns out to be very similar to the folded structure in the NCBD/ACTR complex, and this was considered strong evidence for conformational selection in coupled binding and folding of NCBD[28]. However, as we previously demonstrated in the case of the p53 extreme C-terminus, pre-existence of folded-like conformations is only a necessary but insufficient condition of conformational selection. Nonetheless, given the highly helical nature and apparent tendency to pre-fold, NCBD does seem to represent one of the most probable cases

of conformation selection, if any IDP could rely on preformed structures for efficient initiation of specific recognition.

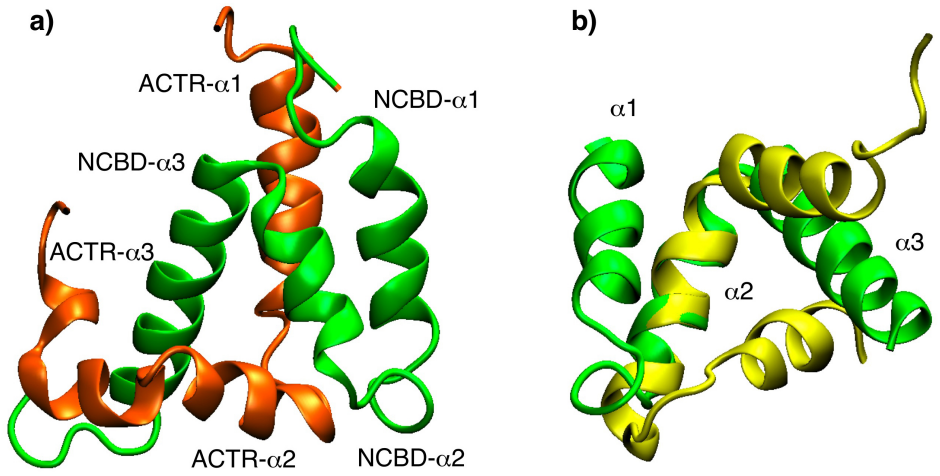


Figure 3.1: (A) The NMR structure of the NCBD/ACTR complex (PDB: 1kbh[194], model 1). NCBD is shown in green and ACTR in orange. All helices of NCBD and ACTR are labeled. (B) An overlay of two representative folded structures of NCBD. The conformation in complex with ACTR is shown in green, and the one with IRF3 in yellow (PDB: 1zoq[196]). Only the structured segment (residues 2066-2112) is shown, and the two structures are aligned using the backbone atoms of the second helix (residues 2085-2093).

This work exploits topology-based modeling as an effective means to determine the mechanism of NCBD/ACTR interaction and to test whether conformation selection indeed could play a dominant role for highly structured IDPs like NCBD. The NCBD/ACTR interaction is particularly interesting also because ACTR is an IDP as well. Such synergistic folding of two IDPs has not yet been investigated in detail. Topology-based modeling is based on the conceptual framework of minimally frustrated energy landscape for natural proteins[98], which argues natural proteins achieve efficient and robust folding by evolving to possess smooth, funneled underlying free energy landscapes. There is a strong correlation between the free energy and fraction of native contacts. In other words, native interactions largely shape the protein energy landscape and non-native ones do not play significant roles. Therefore, given the folded topology, one can derive a list of native contacts and construct effective energy functions that capture the gross features of the true energy landscape. These energy functions are often referred to as Gō- or Gō-like models. These models are extremely efficient and allow direct simulation of folding and unfolding

transitions to characterize both kinetics and thermodynamics of folding. Indeed, topology-based modeling has provided impressive correspondence between experiment and theory for many proteins[98,229]. In principle, it should be applicable to binding-induced folding of IDPs, as binding and folding are analogous processes[230,231] and the topology of the folded complex ought to dictate the gross aspects of recognition mechanism. However, there do exist important differences between sequence and interfacial characteristics of IDPs and globular proteins. For example, IDPs are enriched with charged and polar residues and lack large hydrophobic residues[12]. At the same time, IDPs rely on more on hydrophobic contacts for interfacial interactions[21]. These differences can translate into significant shift in the balance of local folding and intermolecular binding, which subsequently determines important aspects of coupled binding and folding, such as whether the baseline mechanism follows induced folding or conformational selection. Therefore, existing Gō-like models designed for globular proteins might not be directly applied to IDP complexes.

Using well-characterized model IDP complexes[62], we have recently illustrated that, even with sequence-flavoring, existing Gō-like models need to be re-calibrated to balance the intrinsic folding propensities and the intermolecular interaction strength. Such calibration requires additional (experimental) information including the binding affinity and the level of residual structures in the unbound states. We have further shown that, once calibrated, topology-based models do not only appear to predict the correct baseline mechanism of interaction, but are also capable of capturing nontrivial specific details of binding-induced folding. For example, the calibrated Gō-like model predicts that the phosphorylated kinase inducible domain (pKID) of transcription factor CREB initiates binding to the KIX domain of CBP via the C-terminus in disordered conformations, followed by binding and folding of the rest of the C-terminal helix and finally the N-terminal helix. This multi-step sequential binding-induced folding mechanism of pKID is surprisingly consistent with several key observations derived from a recent NMR study[32], and provides a molecular interpretation of key NMR-derived kinetic rates. In this work, we applied a similar approach to construct a balanced topology-derived model of the NCBD/ACTR complex and investigate the mechanism of the synergistic folding of NCBD and ACTR. While important limitations clearly

exist with such simplistic proteins models derived from the folded topology[62], these models can be expected to capture important aspects of the NCBD/ACTR recognition and provide an effective means to generate initial insights that may be further investigated by detailed simulations and/or experiments.

3.2 Methods

3.2.1 Topology-based modeling of NCBD/ACTR

An initial sequenced-flavored Gō-like model was first derived from the PDB structure of the NCBD/ACTR complex (PDB: 1kbh[194]) (see Figure 3.1A), using the Multiscale Modeling Tools for Structural Biology (MMTSB) Gō-Model Builder (http://www.mmtsb.org)[232,233]. The model represents each residue using a single Cα bead and treats the Cα-based native interactions using the Miyazawa-Jernigan (MJ) statistical potentials[234] to provide residue-specific energetic biases. In addition, it includes knowledge-based sequence-dependent, but native-structure independent, pseudo-torsional potentials. The underlying idea is that sequence could provide differing statistical weights to the populations of structural elements during folding to modulate their prevalence as observable intermediates and affect folding kinetics. The sequenced-flavored Gō-like models have been shown to recapitulate subtle differences in folding mechanisms and kinetics that arise from sequence differences in topologically analogous proteins[235,236]. Therefore, it is particularly suitable for extension to modeling IDPs. The initial model was then calibrated by first uniformly scaling the strengths of sets of intra-molecular native contact interaction strengths based on experimental knowledge of the overall level of residual structures in unbound NCBD and ACTR. The strengths of inter-molecular contacts were then scaled to match the simulated and experimental binding affinities of the complex. Both NCBD and ACTR fold into three helices in the complex. The three NCBD helices are (in mouse CBP numbering): α1 (2066-2076; N_{intra}=12, N_{inter}=13),

α2 (2085-2092; N_{intra}=8, N_{inter}=8), and α3 (2094-2112; N_{intra}=18, N_{inter}=41); the three ACTR helices are (in human ACTR numbering): α1 (1044-1058; N_{intra}=18, N_{inter}=32), α2 (1063-1071; N_{inter}=9, N_{inter}=16), and α3 (1072-1080; N_{intra}=9, N_{inter}=12). N_{inter} denotes the numbers of native inter-molecular contacts, and N_{intra} is the number of native contacts within the individual helix. All 76 native intermolecular contacts and the corresponding strengths of interactions from the original sequenced-flavored Gō-like model are listed in the Supplemental Materials Table S3.1. The total number of intra-molecular contacts is 49 for ACTR and 78 for NCBD. As shown in Figure S3.1, while NCBD contains a small number of tertiary contacts that define the α1- α2 interface and the short α2- α3 turn, ACTR largely lacks tertiary contacts.

3.2.2 Simulation protocols

The complex was simulated in a 105 Å cubic box with periodic boundary conditions using CHARMM[63,216]. Langevin dynamics simulations were performed with a dynamic time step of 15 fs and a friction coefficient of 0.1 ps^{-1}. Lengths of all virtual bonds were fixed with SHAKE[217], and the cut-off distance for non-bonded interactions was 25 Å. For the calibration of the intra-molecular interactions, free NCBD and ACTR were simulated at 300 K for 750 ns. Due to the tight binding, enhanced sampling with replica exchange (REX)[104] is necessary for reliable calculation of K_D to calibrate the intermolecular interactions. All REX simulations were performed with the MMSTB Toolset[232,233] with eight replicas spanning 270 to 370 K. The lengths of calibration REX simulations range from 2 to 5 μs. Once the model was properly calibrated, a 30-μs production simulation was initiated from the PDB structure near the melting temperature ($T_m \sim$ 315 K), which was used to calculate all the free energy profiles shown in the rest of this paper. Ten additional productions simulations were initiated from randomly selected folded and unfolded conformations sampled in the REX calibration run (see Figure S3.2A). These simulations allow better transition statistics for the construction of the conformational space network (CSN). As summarized in Table S3.2, a total of 268 folding/binding and unfolding/unbinding transitions

were sampled in all production simulations. Representative time traces of the fractions of inter- and intra-molecular contacts are shown in Figure S3.2B.

3.2.3 Data analysis

All the analysis was carried out using CHARMM and additional in-house scripts. A given native contact is considered formed if the inter-Cα distance is no more than 1 Å greater than the distance in the PDB structure. For equilibrium simulations of free NCBD and ACTR, the helicity was calculated as the fraction of 1-5 (backbone) native contacts formed. For REX simulations of the complex, weighted histogram analysis method (WHAM) was used to combine information from all temperatures to compute either C_v curves or unbiased probability distributions[142]. The unbound state was identified as the one without any native intermolecular contacts formed, and the dissociation constants were calculated from the bound and unbound probabilities as,

$$K_D = \frac{1660}{V_0} \frac{{p_{ub}}^2}{1 - p_{ub}} \tag{3.1}$$

where V_0 is the periodic box volume in unit of Å^3. For production simulations at T_m, all free energy surfaces are converted directly from the corresponding histograms. The surfaces were then shifted such that the bound minima were at zero. Helix cross angles were calculated using the Chothia-Levitt-Richardson algorithm[218] as implemented in CHARMM.

To construct the CSN, all conformations sampled during all 11 production simulations at 315 K were first assigned to discrete microscopic states (nodes) using 8 fractions of native contacts as descriptors, including the fraction of intra-molecular contacts of ACTR (Q_{intra}^{ACTR}), the fraction of tertiary contacts of NCBD ($Q_{intra-tert}^{NCBD}$), the fractions of intermolecular contacts made by the three ACTR helices ($Q_{inter}^{ACTR-\alpha1}$, $Q_{inter}^{ACTR-\alpha2}$ and $Q_{intrer}^{ACTR-\alpha3}$), and the fractions of inter-molecular contacts made by the three NCBD helices ($Q_{inter}^{NCBD-\alpha1}$, $Q_{inter}^{NCBD-\alpha2}$ and $Q_{intrer}^{NCBD-\alpha3}$). Distribution

along each descriptor was divided evenly into 5 bins except for $Q_{intra-tert}^{NCBD}$, where five non-uniform bins were used with 0.1, 0.2, 0.4 and 0.8 as the dividing values. The reason for using non-uniform sub-states is to resolve natural conformational states along $Q_{intra-tert}^{NCBD}$ without having to using more bins. The total possible number of conformational states using the above setup is 5^8 = 390625. For clarity, only the most populated nodes with additional nodes from the transition paths were included in the CSN. The transition paths were defined as those where the system left either the bound or unbound state and entered the other state without revisiting the originating state. The fraction of the total native intermolecular contacts formed (Q_{inter}) was used as the order parameter for defining the bound and unbound state for transition path identification. Q_{inter}=0.15 was used as the upper bound of the unbound state, and Q_{inter}=0.4 as the lower bound of the bound state. Inclusion of nodes on the transition paths is necessary to preserve different transition pathways when a limited number of nodes are used to construct the CSN. The resulting network was visualized with stress minimization using visone (www.visone.de). The node sizes and link thickness in the final CSN reflect the statistical weights in logarithmic scale.

3.3 Results and Discussion

3.3.1 Calibration of the sequence-flavored Gō-like model

Previous NMR secondary chemical shift analysis has estimated that the free NCBD has native-like helical content and the free ACTR is highly disordered with low residual helicity[199]. Figure 3.2A compares the overall helicity distributions of unbound ACTR with different levels of scaling of the strengths of all intra-molecular interactions. Clearly, it shows that the original sequence-flavored Gō-like model overestimates the residual structure level. The scaling factor of ACTR intra-molecular interaction strengths was chosen to be 0.4 in the final model, which yields an average helicity of $\sim$30%. Note that, due to the coarse-grained nature, the Cα-only model has a

limiting helicity of near 20% even without any specific intramolecular interactions (e.g., see the 0.1 trace in Figure 3.2A). A helicity of ∼30% is thus near the "random coil" limit within the context of the peptide model. For NCBD, it turned out that no scaling of the intra-molecular interaction strengths was necessary. As shown in Fig 2b, all three helices of NCBD in the unbound state are nearly as stable as in the bound state. It is interesting that sequence-flavoring alone correctly predicts NCBD-α3 to be the least stable helix in the unbound state. This is consistent with the results of NMR secondary chemical shift analysis[199].

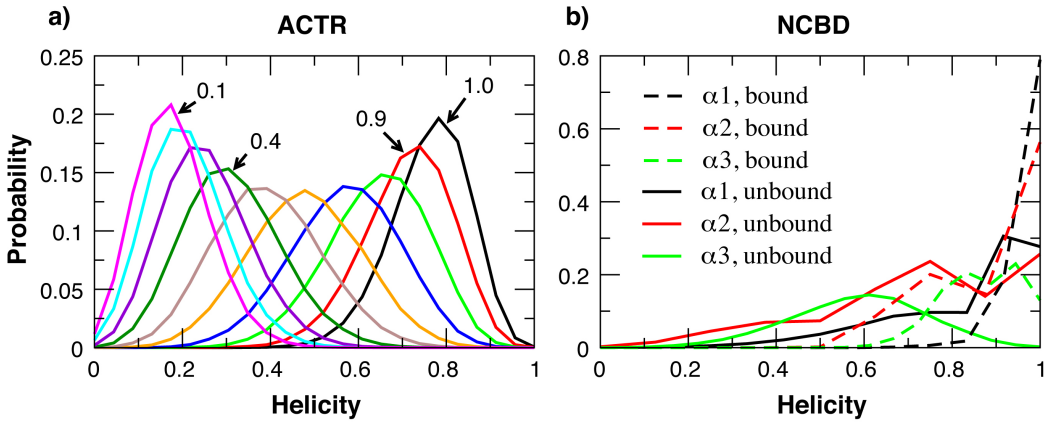


Figure 3.2: (A) Probability distributions of the overall helicity of the unbound ACTR, calculated with various uniform scaling of the intra-molecular interaction strengths. (B) Probability distributions of the helicities of three NCBD helical segments in the unbound and bound states. The unbound state was calculated without any scaling of the intra-molecular interaction strengths, and the bound state distributions were calculated from a 1-μs simulation of the complex using the final calibrated model (see main text).

Once the scaling factors of the intra-molecular interaction strengths were determined, multiple REX simulations were carried out using different scaling of the intermolecular interaction strengths. The free energy profiles as a function of Q_{inter} with a few different scaling factors are provided in Figure S3.3A. The original model yields $K_D \sim \mu$M, nearly two orders of magnitude weaker than the experimental value of K_D = 34 $\pm$ 8 nM[198]. The optimal scaling of the intermolecular interaction strengths turns out to be 1.1, which yields $K_D \sim$ 23 nM and $T_m \sim$ 315 K. The heat capacity as a function of temperature calculated from a 4.9-μs REX simulation using the final model is shown in Figure S3.3B. Surprisingly, with sequence flavoring, the topology-derived models appear to consistently predict strong structural fluctuations within the folded complex,

such that the folded minimum centers at $Q_{inter} \sim 0.6$ even with substantial strengthening of the intermolecular interactions (e.g., with scaling factors up to 1.5; data not shown). Further examination of the list of all native intermolecular contacts (see Table S3.1) reveals that it contains many contacts involving small hydrophobic residues and/or charged ones. These contacts are weak in the MJ scale[234], and frequently involve the C-termini of ACTR and NCBD. Indeed, the root-mean-squared fluctuation (RMSF) profiles computed from a control simulation of the complex at 300 K using the calibrated model reveal significantly elevated fluctuation at the C-termini of both ACTR-α3 and NCBD-α3 (see Figure S3.4). Interestingly, a previous NMR relaxation analysis has also revealed fluctuating contacts between ACTR-α3 and NCBD-α3[199]. In addition, a recent H/D exchange mass spectrometry (H/D-MS) study[58] showed that, within the folded regions of NCBD and ACTR, peptide segments that map to the C-termini of both ACTR-α3 and NCBD-α3 had the smallest protection factors. Therefore, it appears that the strong structural fluctuations predicted by the calibrated sequence-flavored model is realistic, and no adjustment to the model was applied to further stabilize the complex.

3.3.2 The baseline mechanism: induced folding vs. conformational selection

With careful calibration, the final sequence-flavored Gō-like model is able to reproduce the experimental data on the binding affinity and the level of residual structures in the unbound proteins. Therefore, the model properly reflects the balance between the intrinsic folding propensities of NCBD and ACTR and the strength of their interactions. This balance should allow a reliable prediction of the baseline mechanism. For this, we examine the free energy surfaces along appropriate binding and folding reaction coordinates, where the most probable transition paths can be identified as the minimum free energy paths connecting various basins. In the context of topology-based modeling, the fractions of native contacts provide natural reaction coordinates for describing folding, and analogously, binding[237]. Figure 3.3 examines the 2D binding and folding free energy surfaces of NCBD and ACTR, using the total fractions of inter- and intra-molecular

contacts as order parameters. Apparently, both NCBD and ACTR bind and fold in a highly cooperative fashion, as Q_{intra}^{ACTR} and Q_{intra}^{NCBD} gradually increase together with Q_{inter}. In particular, even though the free NCBD is highly helical (see Figure 3.2B), Q_{intra}^{NCBD} does not appear to increase any faster than Q_{inter}, i.e., folding does not precede binding on the whole protein level. Therefore, on the baseline level, neither NCBD nor ACTR follows either induced folding or conformational selection. Not surprisingly, folding of NCBD and ACTR are highly synergistic. As shown in Figure 3.3C, neither protein displays any significant folding without binding (and folding) of the partner.

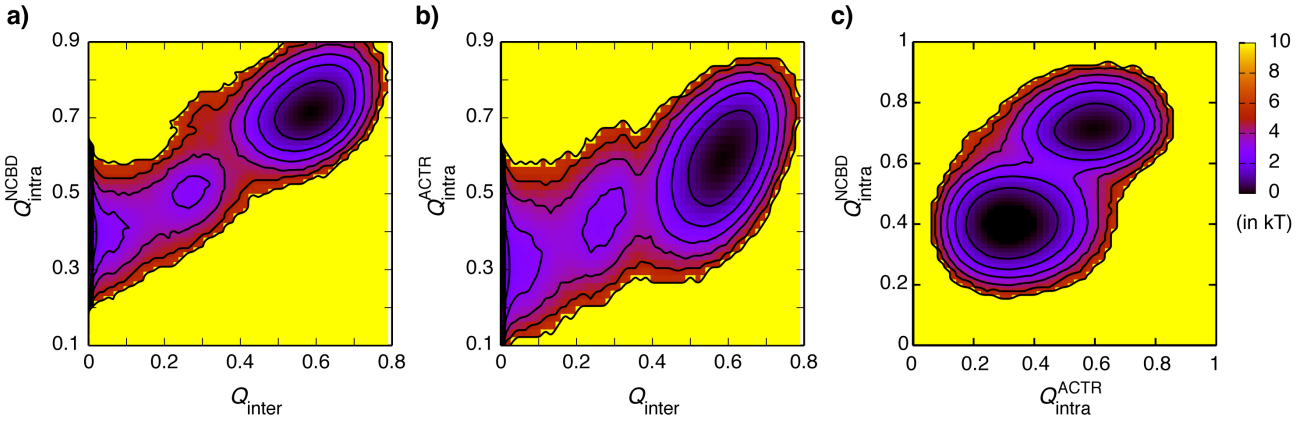


Figure 3.3: 2D free energy surfaces of the synergistic binding and folding of NCBD and ACTR. Q_{intra}^{ACTR} and Q_{intra}^{NCBD} are the fractions of native intra-molecular contacts formed by ACTR and NCBD, respectively. Contour levels are drawn at every kT. Figure prepared by DG.

3.3.3 A key intermediate state of the NCBD/ACTR interaction

The free energy surfaces in Figure 3.3 also reveal a key intermediate state of the NCBD/ACTR interaction, at $Q_{inter} \sim 0.25$. To further characterize the nature of this state, conformations sampled during the production simulation were grouped to three states: $Q_{inter} = 0$ for the unbound state, $0.21 < Q_{inter} < 0.32$ for the intermediate state (corresponding to 16 to 24 native contacts formed), and $Q_{inter} > 0.5$ for the bound state. Structural analysis of the resulting ensembles reveals that in the intermediate state NCBD and ACTR mainly interact through the C-terminal

segments that include both $\alpha 2$ and $\alpha 3$, while $\alpha 1$ helices from both proteins are largely unbound (see Figure 3.4). At the intermediate state, $\alpha 2$ and $\alpha 3$ from both proteins are similarly folded compared to the bound state, while $\alpha 1$ helices remain as (un)structured as in the unbound state (see Figure S3.5). Further analysis of the helix-helix packing geometry including helix center distances and cross angles (see Figure S3.6) demonstrates that the C-terminal segments of NCBD and ACTR adopt highly folded-like tertiary conformations in this intermediate state, which is consistent with the $Q_{inter}^{NCBD\alpha 2\alpha 3-ACTR\alpha 2\alpha 3}$ distributions shown in Figure 3.4B. Therefore, the C-terminal segments of NCBD and ACTR appear to serve as a mini folding core prior to complete binding and folding (e.g., see Figure 3.4C). As discussed above, NMR, H/D-MS and the current simulations all suggest significant structural fluctuation in interactions between NCBD-$\alpha 3$ and ACTR-$\alpha 3$. The observation that the C-terminal $\alpha 2$ and $\alpha 3$ regions of these two proteins form the key folding core, and play a major role in initiating specific recognition can thus be surprising. Nonetheless, this prediction is fully consistent with independent atomistic unfolding and unbinding simulations using physics-based explicit and implicit solvent protein force fields[34]. Furthermore, it also appears to be consistent with the recent H/D-MS study[58], where peptide segments within the $\alpha 2$ and $\alpha 3$ regions of both NCBD and ACTR were shown to have much larger protection factors compared to those mapped into other folded regions of the complex.

3.3.4 Mechanism of coupled binding and tertiary folding of NCBD

NCBD is highly helical in the unbound state (see Figure 3.2B), and only forms a limited number of tertiary intra-molecular contacts upon folding and binding to ACTR (see Figure S3.1B). The total fraction of intra-molecular contacts (Q_{intra}^{NCBD}) is thus not a sensitive measure of NCBD tertiary folding. To better understand the interplay between binding and NCBD tertiary folding, Figure 3.5A examines the free energy surface as a function of Q_{inter} and the fraction of tertiary intra-molecular contacts of NCBD, $Q_{intra-tert}^{NCBD}$. At the baseline level (e.g., assuming an inability to resolve the details along the pathways connecting the unbound and bound states), it appears

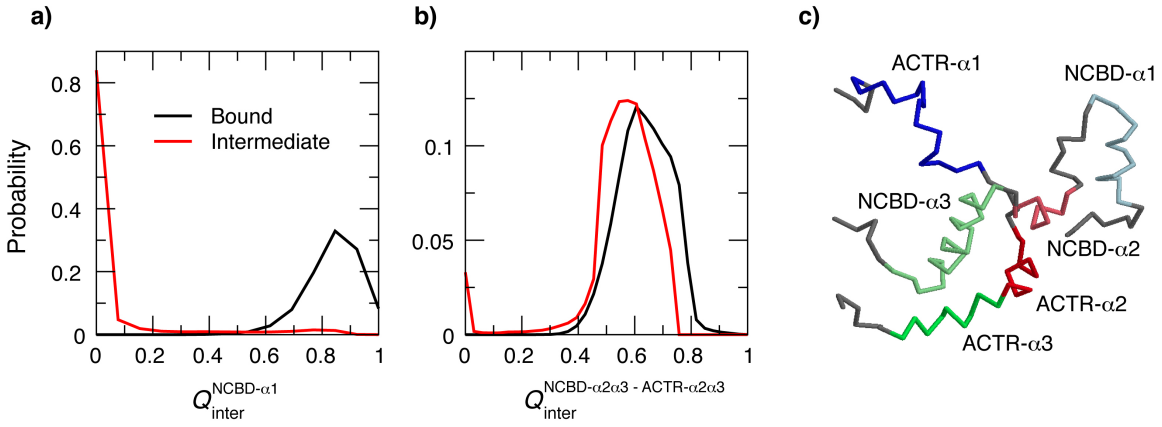


Figure 3.4: Probability distributions of the fraction of native intermolecular contacts formed (A) by NCBD-α1, $Q_{inter}^{NCBD-\alpha1}$, and (B) between the C-terminal segments of NCBD and ACTR, $Q_{inter}^{NCBD-\alpha2\alpha3-ACTR-\alpha2\alpha3}$. The unbound state has no native intermolecular contact by definition and is thus not shown. (C) A representative snapshot of the intermediate state, with all helical segments colored and marked. Figure prepared by DG.

that the increase in Q_{inter} precedes and thus presumably drives that of $Q_{intra-tert}^{NCBD}$, i.e., an induced folding-like mechanism. However, such a baseline mechanistic classification appears to break down once the additional details of the transition pathways are taken into consideration. Instead, conformational selection appears to play key roles during different stages of binding and tertiary folding of NCBD. Specifically, the transition between the unbound and intermediate states follows both induced folding and conformational selection-like pathways, as indicated by the yellow and dark green dashed lines connecting states U and I in Figure 3.5A. Furthermore, the conformational selection-like pathway has lower free energy barrier (by $\sim$ 1 kT), and is thus slightly favored. More notably, the intermediate-bound transition appears to mainly follow conformational selection on the tertiary level, where NCBD quickly folds before forming additional native contacts with ACTR (e.g., see the green dashed line connecting states I and B in Figure 3.5A). Such a staged mechanism of coupled binding and tertiary folding of NCBD resembles the extended conformational selection model recently discussed by Csermely, Palotai and Nussinov[205], which emphasizes a multi-stage mutual adjustment process that involves both induced folding and conformational selection.

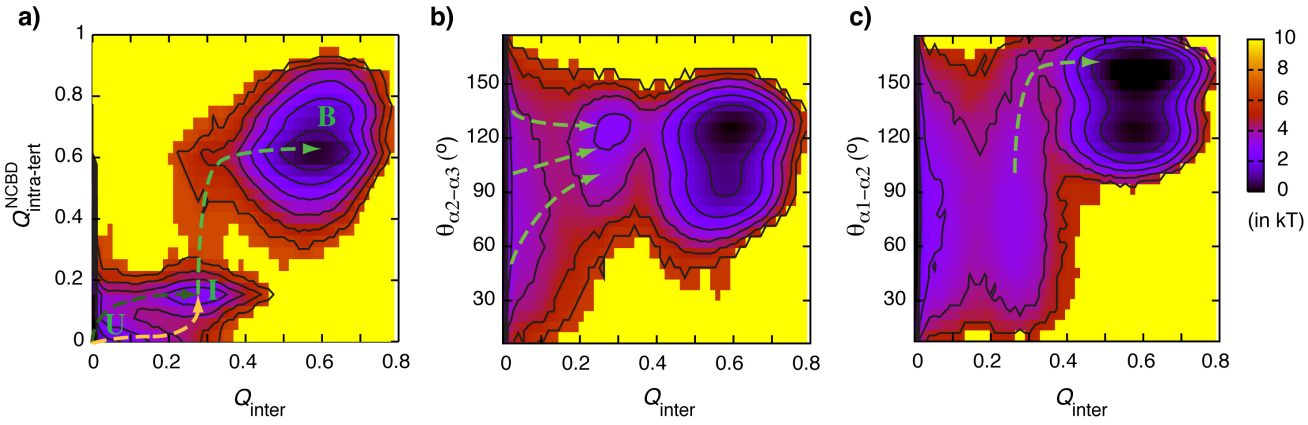


Figure 3.5: 2D free energy surfaces of coupled binding and tertiary folding of NCBD. $Q^{NCBD}_{intra-tert}$ is the fraction of native tertiary intra-molecular contacts formed by NCBD. $\theta_{\alpha2-\alpha3}$ is the cross angle between NCBD-α2 and α3, and $\theta_{\alpha1-\alpha2}$ is that between NCBD-α1 and α2. In panel (A), the unbound, intermediate and bound states are marked with U, I and B, respectively. Contour levels are drawn at every kT. Figure prepared by DG WZ.

With largely folded helices in the unbound state, the tertiary folding of NCBD mainly involves packing of the three helical segments. The analysis above (e.g., see Figure 3.4) has shown that the unbound-intermediate transition mainly involves the folding of NCBD-α2 and α3 and the next step involves that of NCBD-α1. In Figure 3.5B and C, we directly examine the coupling between intermolecular interactions and formation of native-like helix-helix packing as reflected in the helix-helix cross-angles. The analysis shows that the transition pathways between the unbound and intermediate states indeed have a very broad distribution, and there is a continuum between two extreme mechanisms of induced folding and conformational selection for (binding-induced) tertiary packing of NCBD-α2 and α3 (as indicated by multiple dashed lines in Figure 3.5B). In the intermediate state, NCBD-α1 remains nearly as dynamic as in the unbound state, but with a slight enrichment of folded-like conformations (also see Figure S3.6, red traces). These folded-like conformations appear to play a key role in initiating the binding and folding of the rest of the complex. One way to understand the conformational selection-like transition between the intermediate and bound states is that, as the most stable helix, NCBD-α1's packing with the folded core of NCBD-α2 and α3 is defined by only a few degrees of freedom. Thus, NCBD-α1 can readily adopt native-like packing upon making of a few additional intermolecular contacts,

which appears to drive the formation of the remaining intermolecular contacts.

3.3.5 Folding and binding of individual helical segments of NCBD and ACTR

We have further examined coupled binding and folding of individual helices of NCBD and ACTR. As shown in Figure 3.6A-B, NCBD-α1 and α2 are very stable in the unbound state, and bind largely as pre-folded helices as expected. The least stable helix of NCBD, α3 appears to fold concurrently with binding (Figure 3.6C). In contrast to NCBD helices, all ACTR helices are largely unstructured in the unbound state, and they appear to mainly follow induced folding-like mechanisms. As shown in Figure 3.6D-F, Q_{inter} increases faster than various Q_{intra} of individual helices during transitions, either between the unbound and intermediate states (ACTR-α2 and α3) or between the intermediate and bound states (ACTR-α1). In other words, intermolecular interactions drive the (secondary) folding of ACTR. Taken together, the current topology-based simulation suggests that NCBD provides pre-folded structural elements on both secondary and tertiary levels, which allow efficient binding of ACTR in unstructured conformations and drive specific folding of ACTR during different stages of the recognition.

Figs. 7 and 8 further examine the sequence of binding of all ACTR and NCBD helices by comparing the free energy projections along various combinations of the fractions of native intermolecular contacts formed by different helical segments. Examination of the minimum free energy paths connecting various basins along these projections reveals detailed (kinetic) ordering of binding and folding of individual segments. The analysis supports the above observation that α2 and α3 from both proteins drive the recognition by forming the folded core at the intermediate state. Specifically, binding of ACTR-α2 and α3 precedes that of ACTR-α1 (Figure 3.7D and E), and binding of NCBD-α2 and α3 precedes that of NCBD-α1 (Figure 3.8D and E). Furthermore, NCBD-α3 and ACTR-α2 appear to be the most frequently involved in initiating the recognition. NCBD-α3 has the largest number of native intermolecular contacts (N_{inter}=41) and its role in

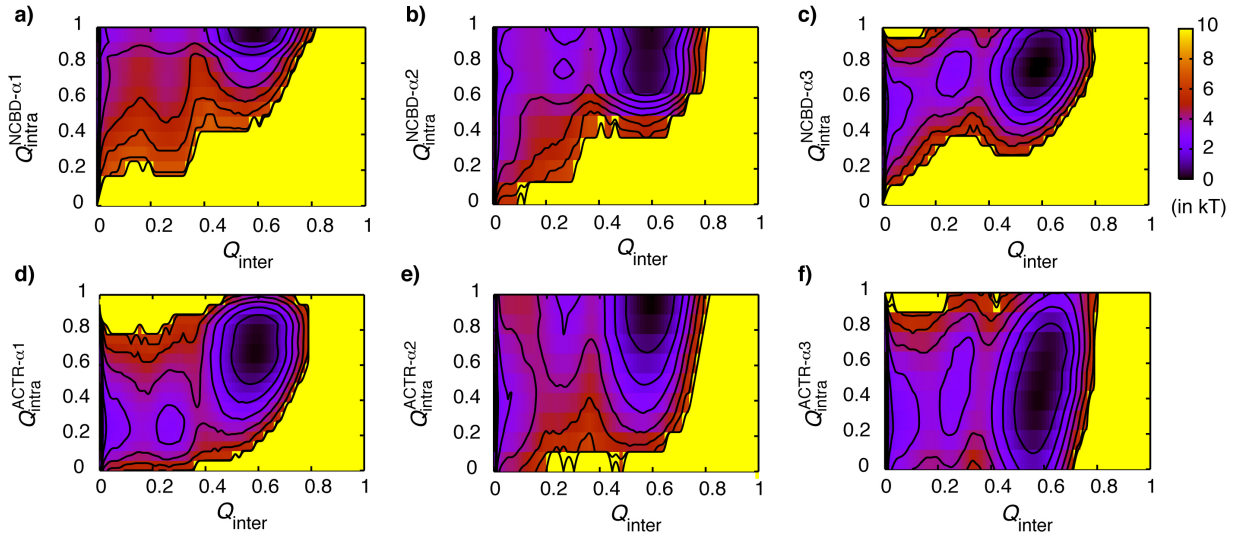


Figure 3.6: 2D free energy surfaces of as functions of Q_{inter} and the fraction of intra-molecular native interactions formed within individual helices of NCBD and ACTR. Contour levels are drawn at every kT. Figure prepared by DG WZ.

initiating binding and folding may thus be expected[238]. However, ACTR-α2 does not have the highest density of native contacts and its role in initiating recognition is unexpected from simple consideration of native contact density. Interestingly, these free energy surfaces also reveal a co-existence of many parallel pathways of the NCBD/ACTR recognition. For example, Figure 3.8A shows that, along the dominant pathway (indicated by the yellow dashed line), $Q_{inter}^{NCBD-\alpha1}$ does not increase from near zero until Q_{inter} exceeds $\sim$ 0.25. That is, NCBD-α1 binds only after a signification number of native intermolecular interactions are formed). At the same time, there exists a minor pathway where binding is initiated by NCBD-α1 (indicated by the green dashed line in Figure 3.8A). These parallel pathways are also evident in Figure 3.8D. In fact, the free energy surfaces shown in Figure 3.7 and 3.8 suggest that all helices of NCBD and ACTR could initiate binding, albeit with different levels of prevalence. Such diversity in folding and binding pathway is not surprising, and is actually expected to be generally true based on the funneled energy landscape theory[239]. The importance of examining the recognition mechanism using multiple sets of order parameters should also be emphasized. For example, the $Q_{intra-tert}^{NCBD-\alpha1}$ - Q_{inter} free energy surface shown in Figure 3.5A alone could lead to an overly simplified view that

the recognition occurs through a well-defined pathway that involves folding and binding of $\alpha 2$ and $\alpha 3$, followed by binding and folding of $\alpha 1$ helices. This is a limitation of free energy analysis along pre-selected order parameters, which can mask important heterogeneity and complexity along orthogonal degrees of freedom.

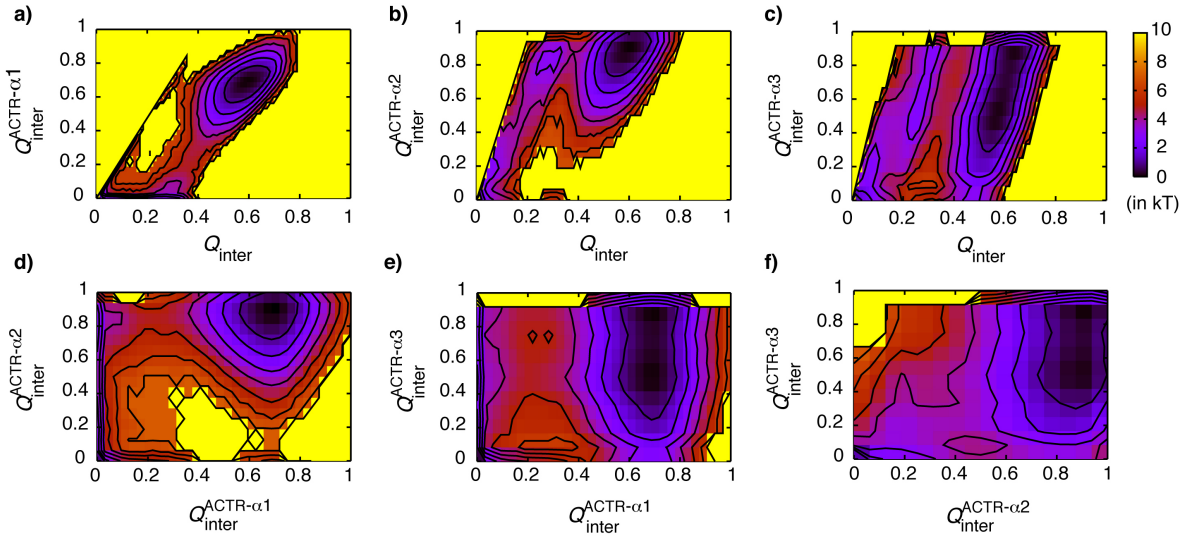


Figure 3.7: 2D free energy surfaces as functions of the fractions of native intermolecular contacts formed by various segments of ACTR. Contour levels are drawn at every kT. Figure prepared by DG WZ.

3.3.6 Network analysis of the complex pathways of coupled binding and folding

CSN analysis does not rely on pre-determined order parameters as required in the traditional free energy analysis, and can thus allow better visualization of the heterogeneous pathways of protein folding and binding[240–242]. One of key challenges in constructing the CSN is the need to divide the continuous protein conformational space into discrete microstates. The discretization has been mainly achieved either by conformational clustering[243–246] or by using a reduced set of (structural) descriptors[242,247]. In the context of topology-based modeling, various fractions of native contexts do provide natural reaction coordinates and are thus appropriate for defining microstates. Figure 3.9 shows a CSN of the synergetic folding of NCBD and ACTR derived from all 11 production simulations, by including only the most populated 100 nodes and additional 200 nodes from the

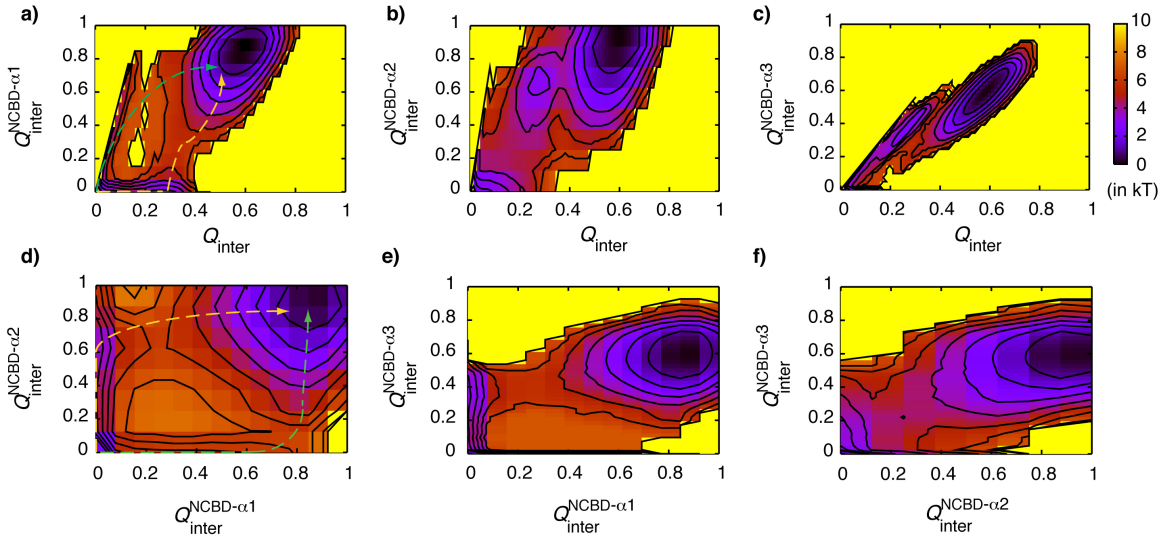


Figure 3.8: 2D free energy surfaces as functions of the fractions of native intermolecular contacts formed by various segments of NCBD. Contour levels are drawn at every kT. Figure prepared by DG WZ.

transition paths. The total number of links is 15161. Including additional nodes does not change the appearance of the CSN (e.g., see Figure S3.7). Even though powerful analysis can be done to further analyze the kinetic portioning and connectivity of the conformational space, the goal here is mainly to illustrate and visualize the complexity of multiple pathways of binding and folding of NCBD and ACTR. With the nodes distributed with minimized stress (as implemented in visone), the CSN shows a natural segregation of different (meta-)stable free energy states that include the unbound (labeled in blue), intermediate (green), and fully bound (red) states. It furthers illustrates the co-existence two main groups of recognition pathways. While the peptides mostly initiate binding through the C-terminal $\alpha2$ and $\alpha3$ and go through the intermediate state toward the bound state, they can also initiate binding through $\alpha1$ helices (e.g., the link between nodes 37 and 218) and reach the bound state through an intermediate state that is mainly stabilized by interactions between $\alpha1$ helices (purple nodes). Interestingly, it appears that NCBD needs to pre-fold with 0.4$< Q^{NCBS}_{intra-tert} <$0.8 to initiate binding through $\alpha1$. Along the major pathway, the CSN shows three key routes initiated by nodes 202, 248 and 270. These routes appear to correspond to conformational selection (node 202) and induced folding initiated by NCBD-$\alpha2$ (node 248) and $\alpha3$ (node 270) for the unbound-intermediate transitions observed from the free

energy analysis (e.g., see Figure 3.5A). Transitions from the intermediate to bound state mainly go through an intermediate state where NCBD become pre-folded (orange nodes), even though highly cooperative binding and folding of $\alpha 1$ helices also appear possible (e.g., see direct links between green and red nodes). Taken together, the CSN appears to provide a clear and concise illustration of the heterogeneous pathways of the NCBD/ACTR recognition that is fully consistent with the observations derived from analysis of multiple free energy surfaces.

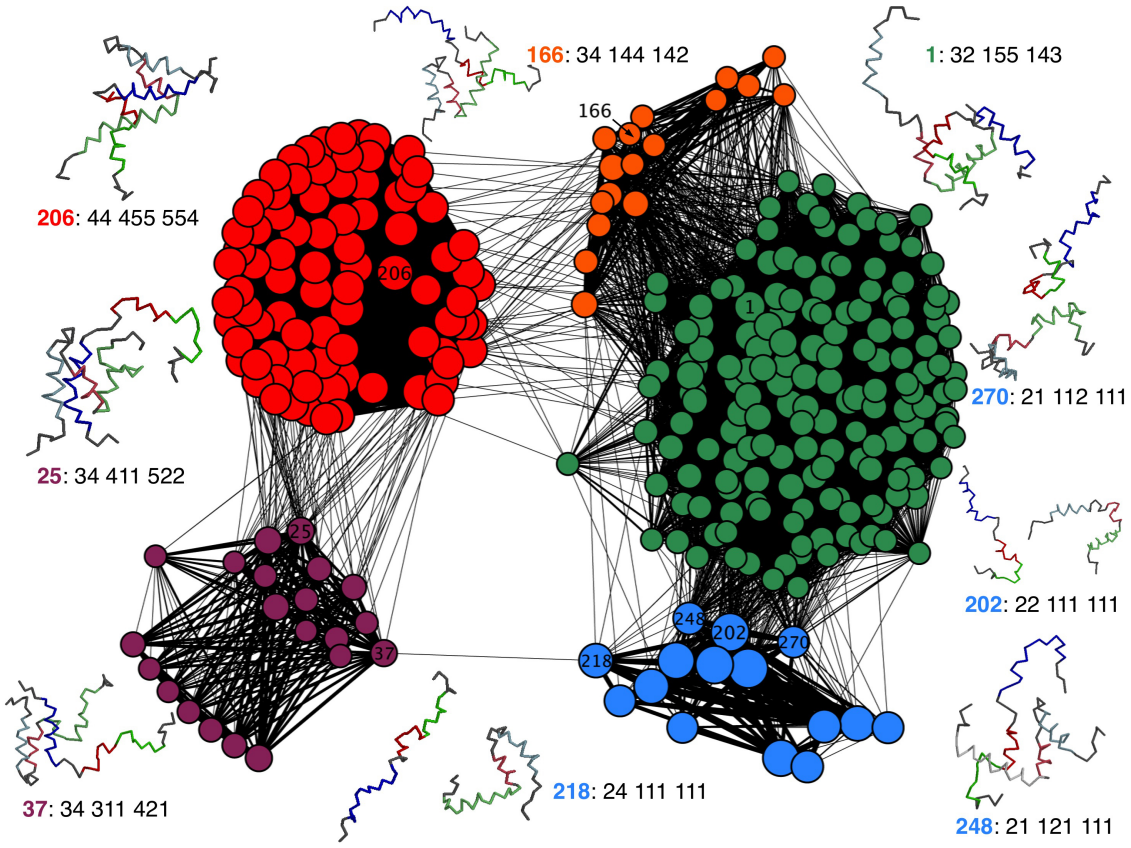


Figure 3.9: The CSN of the synergetic folding of NCBD and ACTR. The nodes represent the conformational microstates, and the links represent the transitions between them. The node sizes and link widths reflect the statistical weights in logarithmic scale. The colors of the nodes are assigned according to states of NCBD: blue: unfolded and unbound; green: partially folded and bind with ACTR through NCBD-$\alpha 2$ and $\alpha 3$; orange, folded and bind with ACTR through NCBD-$\alpha 2$ and $\alpha 3$; red, folded and bound; and, purple: partially folded and bind to ACTR through NCBD-$\alpha 1$. Representative snapshots are shown for selected nodes, where NCBD and ACTR helices are colored using the same scheme as in Fig. 4. The notation is the node ID (in bold fonts) followed by the bin indices (1 through 5) along the 8 structural descriptors (Q_{intra}^{ACTR}, $Q_{intra-tert}^{NCBD}$, $Q_{inter}^{ACTR-\alpha 1}$, $Q_{inter}^{ACTR-\alpha 2}$, $Q_{inter}^{ACTR-\alpha 3}$, $Q_{inter}^{NCBD-\alpha 1}$, $Q_{inter}^{NCBD-\alpha 2}$, $Q_{inter}^{NCBD-\alpha 3}$; see Methods for detail).

3.4 Conclusions

Topology-based modeling has been successfully applied to investigate the synergistic folding of two IDPs with drastically different residual stabilities in the unbound states. Through careful calibration based on additional experimental data besides the complex structure, the topology-based model was able to properly capture the balance between the intrinsic folding propensities of NCBD and ACTR and the strength of their intermolecular interaction. Subsequent simulations revealed several important mechanistic features of the coupled binding and folding processes. Despite a drastic difference in residual structural level, both NCBD and ACTR bind and fold in a highly cooperative fashion on the baseline level that involves a key intermediate state. In the intermediate state, the C-terminal helices $\alpha 2$ and $\alpha 3$ of NCBD and ACTR form a mini folding core that allow rapid folding and binding of $\alpha 1$ helices. Interestingly, due to the highly structured nature of the unbound NCBD, conformational selection appears to play significant roles in the formation of both the intermediate state and the final specific complex. The binding-induced tertiary folding of NCBD involves multiple stages of selection and induced folding, and is clearly an example of "extended conformational selection"[205,206]. Importantly, key mechanistic features predicted by the current topology-based modeling, such as regarding individual helix folding and binding, tertiary folding, and intermolecular interactions, are surprisingly consistent with independent atomistic simulations using implicit solvent protein force fields[34]. Several key aspects of the predicted mechanism are also consistent with the protection factor mapping derived from a recent H/D-MS study of NCBD/ACTR[58].

Nonetheless, it is interesting to note that, even for an unusually structured IDP like NCBD, the recognition is initiated by the more flexible C-terminal segment and with substantial contribution from induced folding. Formation of the meta-stable mini folding core appears necessary for conformational selection to play an even larger role during later stages of recognition, where NCBD-$\alpha 1$ readily form native-like packing with the folded core and allows rapid binding and

folding of the rest of the complex. Combined with existing experimental and theoretical evidence (see Introduction), the current work further supports the notion that induced folding is very likely the prevalent mechanism of specific IDP-protein interactions. Even when conformational selection does play a role, it will likely be limited to the local (secondary) structure level and later stages of the recognition process. A fundamental question is then why and how induced folding might confer functional advantages for IDP recognition. The need for proteins to remain unstructured in the unbound state is believed to arise from certain functional constraints, particularly in signaling and regulation, such as to allow high specificity coupled with low affinity binding, inducibility by posttranslational modifications, structural plasticity for binding multiple targets, and thermo-instability for alloteric regulation[6,20]. It has also been proposed that disordered proteins could enhance the (nonspecific) binding rate up to 1.6 fold due to larger capture radii (i.e., the fly-casting effects[166,167]). However, recent studies show that unbound IDPs tend to be compact[171,188,248] and thus may not have much greater capture radii to have the full fly-casting effects. Furthermore, the rate-enhancing affect due to increased capture radii will be largely offset by slower diffusion[172]. Therefore, it is not obvious that intrinsic disorder itself could provide any significant kinetic advantages.

Instead, it appears that while required for satisfying other functional constraints, intrinsic disorder could lead to a kinetic bottleneck that must be overcome to allow facile recognition in signaling and regulation. This bottleneck arises from the requirement of (partial or full) folding during specific binding, as protein folding is usually a slow process (compared to translational and orientational diffusion) with an estimated "speed limit" of μs[249]. Indeed, the recent dual-transition state model developed by Zhou[173] predicts that the diffusion-limited binding rate provides an upper bound of the binding rate, which is achieved only if the protein can rapidly undergo folding transition upon nonspecific binding. This limit corresponds to the case of induced folding. In contrast, conformational selection arises in the limit of slow conformational transitions and actually defines the lower bound of the binding rate. Interestingly, existing experimental binding rates show that IDPs bind no slower than globular proteins[172]. This suggests that IDPs are able to

overcome the kinetic bottleneck of folding and achieve rates near or at the diffusion limit. This is consistent with the notion that induced folding is the prevalent mechanism for coupled binding and folding of IDPs. A key question is then how IDPs manage to fold so rapidly upon nonspecific binding, often at rates beyond the traditional folding speed limit. The constraint of rapid folding could explain why the interaction motifs of IDPs are usually short and often fold into simple topologies with low contact orders upon binding. Furthermore, it is likely that IDPs (and their binding targets) may exploit additional physical properties to achieve rapid folding. For example, previous studies of IDP interactions[30] and protein-DNA interactions[250,251] have suggested that long-range electrostatic interactions may play an important role.

While it is encouraging that simple models derived from the folded complex topology can reliably predict important features of coupled binding and folding, several inherent limitations of such models should not be overlooked. For example, topology-derived models can not faithfully describe specific details of the unbound states, particularly non-native-like residual structures[252], or properly model the encounter complexes, a critical step that often involve transient nonspecific contacts[31,32]. Importantly, non-native interactions can play an important role in stabilizing nonspecific encounter complexes and/or folding intermediates, leading to nontrivial consequences in binding and folding pathway and kinetics[253,254]. Given the prevalence of charges in IDPs, long-range electrostatic interactions do not only modulate the conformational properties of the unbound states[188,255], but can also play a key role in the binding and folding interactions[30]. Explicit charges could be introduced into the conventional topology-derived models to account for long-range electrostatic interactions[99,256]. Nonetheless, even though more sophisticated Gō-like models might be exploited[96], contributions of specific yet non-native interactions are not encoded in the topology *per se* and can not be expected to be properly accounted for in topology-based modeling in general. It is also important to emphasize that detail characterization of disordered protein states and transient structures represent a broader challenge beyond topology-based modeling. Due to the heterogeneous and dynamic nature of such states, experiments alone generally do not provide sufficient restraints for unambiguous determination of the unfolded ensembles[37,53,155]. Arguments

can be made that *de novo* molecular simulations are necessary to provide the missing structural detail of free IDPs[1,72,169], even though such simulations are limited by both sample capability and force field accuracy. At present, only small free IDPs could be modeled using physics-based force fields with reasonable reliability, and direct simulations of the coupled binding and folding processes are largely out of reach. As such, it is important to tightly integrate hypothesis-driven topology-based modeling, physics-based *de novo* simulation, and various biochemical and biophysical characterizations to obtain better understanding of how the structure and interaction of IDPs are precisely controlled and regulated.

3.5 Supplemental Materials

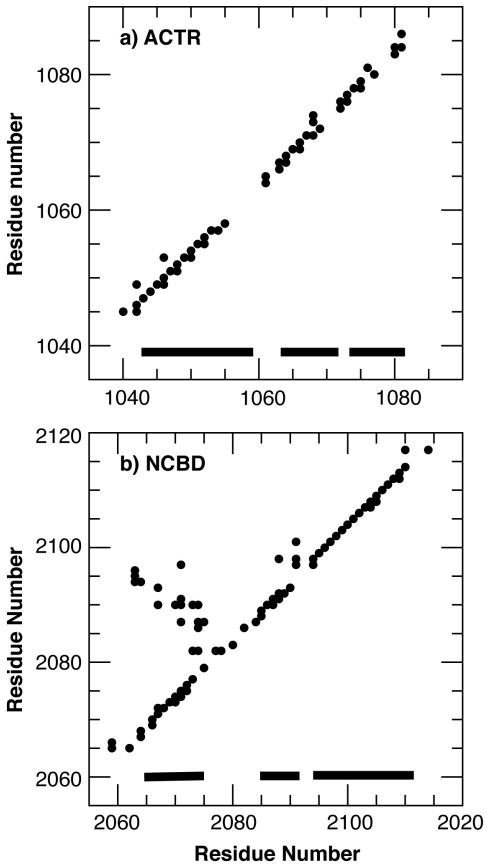


Figure S3.1: Residue-residue intra-molecular contact maps of ACTR and NCBD in the complex. The contact maps are derived from the PDB:1kbh, model 1. The solid bars mark the locations of helical segments formed in the complex. Figure prepared by DG.

Table S3.1: All 76 native intermolecular contacts identified from PDB:1kbh, model 1. Residues 1 - 47 correspond to ACTR residues 1040 - 1086, and residues 48 - 106 correspond to NCBD residues 2059 - 2117. Table prepared by DG JC.

Contacting Residues	ϵ_{ij} **(kcal/mol)**	**Contacting Residues**	ϵ_{ij} **(kcal/mol)**
GLU 01 ILE 52	0.514847	ALA 22 SER 68	0.316466
GLU 01 GLN 85	0.223573	THR 23 SER 69	0.308593
GLN 03 GLN 85	0.242466	LEU 25 LEU 64	1.160374
SER 04 ARG 50	0.255062	LEU 25 PHE 90	1.146204
ASP 05 ARG 50	0.36055	LEU 25 ARG 94	0.634506
ASP 05 SER 51	0.256636	GLU 27 GLN 74	0.223573
GLU 06 ARG 50	0.357402	ILE 28 VAL 76	0.952546
GLU 06 SER 51	0.233019	ILE 28 LEU 77	1.108417
LEU 09 SER 51	0.617187	ILE 28 LEU 80	1.108417
LEU 09 SER 53	0.617187	ILE 28 PHE 90	1.076928
LEU 09 LEU 57	1.160374	ASP 29 PHE 90	0.547911
LEU 10 GLN 85	0.63608	ASP 29 ARG 94	0.36055
LEU 10 LEU 86	1.160374	ASP 29 TYR 98	0.43455
LEU 10 ALA 89	0.773058	ALA 31 LEU 77	0.773058
GLN 12 LEU 57	0.63608	LEU 32 LEU 77	1.160374
GLN 12 LEU 61	0.63608	LEU 32 MET 87	1.009226
LEU 13 SER 53	0.617187	LEU 32 PHE 90	1.146204
LEU 13 ALA 56	0.773058	LEU 32 ILE 91	1.108417
LEU 13 LEU 60	1.160374	ILE 34 PHE 90	1.076928
LEU 13 LEU 61	1.160374	ILE 34 ILE 91	1.029694
LEU 13 LEU 86	1.160374	ILE 34 ARG 94	0.571528
HSD 14 LEU 86	0.714803	ILE 34 THR 95	0.634506
HSD 14 ALA 89	0.379444	ILE 34 TYR 98	0.826589
LEU 16 LEU 61	1.160374	PRO 35 TYR 98	0.502251
LEU 16 LEU 64	1.160374	LEU 37 ILE 91	1.108417
LEU 16 LYS 65	0.530592	LEU 37 THR 95	0.683314
LEU 17 LEU 60	1.160374	VAL 38 THR 95	0.544762
LEU 17 LEU 64	1.160374	VAL 38 TYR 98	0.727399
LEU 17 VAL 76	1.020248	VAL 38 VAL 99	0.8691
LEU 17 LEU 80	1.160374	ASN 39 MET 105	0.464465
LEU 17 LEU 86	1.160374	GLN 42 LYS 92	0.203105
LEU 17 PHE 90	1.146204	GLN 42 THR 95	0.299147
SER 18 GLN 93	0.234594	ALA 43 THR 95	0.365274
SER 18 ARG 94	0.255062	ALA 43 ALA 96	0.428252
ASN 19 LEU 64	0.588847	ALA 43 VAL 99	0.63608
ASP 21 ARG 94	0.36055	LEU 44 GLN 106	0.63608
ASP 21 LYS 97	0.264509	GLU 45 LYS 92	0.283402
ALA 22 LEU 64	0.773058	LYS 47 LYS 92	0.018893

Table S3.2: Summary of all 11 production simulations at 315 K. The average fraction of the unbound state is above 0.5, indicating that T_m is actually slightly below 315 K. The five initial bound (b1 to b5) and unbound (u1 to u5) conformations are shown in Figure S3.2a. Table prepared by DG.

Initial Conformation	**Time (μs)**	N_{trans}	P_{ub}
1kbh Model 1	30	26	0.62
1100.pdb (b1)	30	29	0.5
6200.pdb (b2)	30	23	0.73
6800.pdb (b3)	30	21	0.6
700.pdb (b4)	30	16	0.69
7200.pdb (b5)	30	26	0.55
15230.pdb (u1)	30	29	0.61
4200.pdb (u2)	30	20	0.8
2900.pdb (u3)	30	20	0.78
19113.pdb (u4)	30	30	0.65
23510.pdb (u5)	30	28	0.62

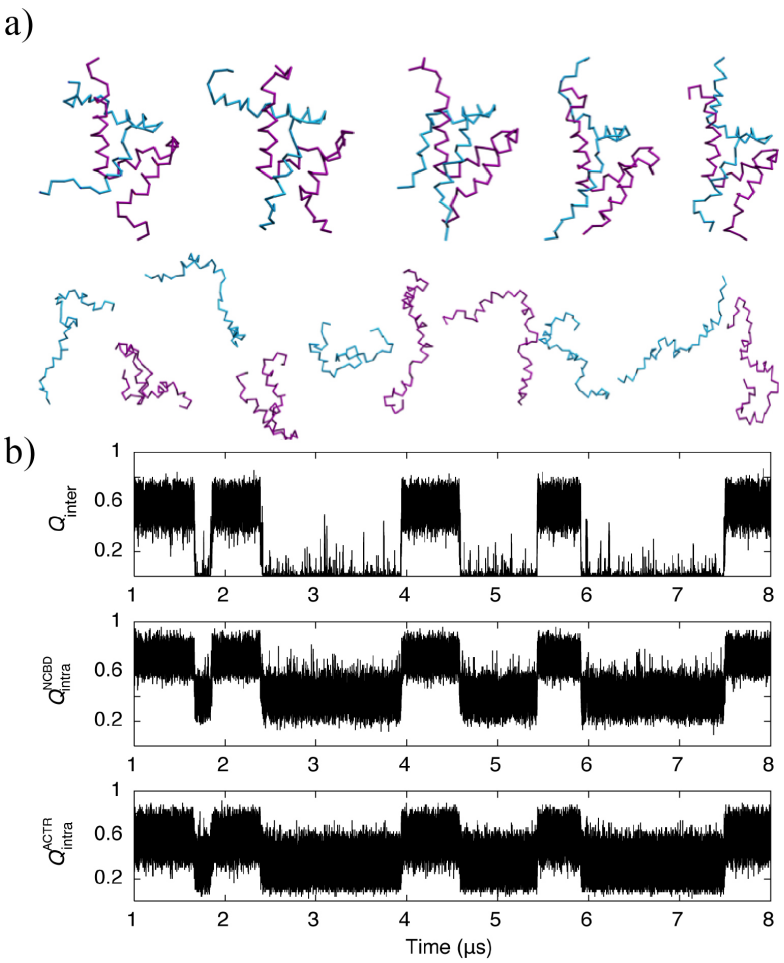


Figure S3.2: (A) Five bound and five unbound conformations randomly selected from the REX calibration simulation using the final calibrated sequenced-flavored Gō-like model. NCBD is shown in purple trace and ACTR in cyan. (B) Representative time traces of the fractions of inter- and intra-molecular contacts from one of eleven independent 30-μs production simulations at T_m = 315 K. Figure prepared by DG WZ.

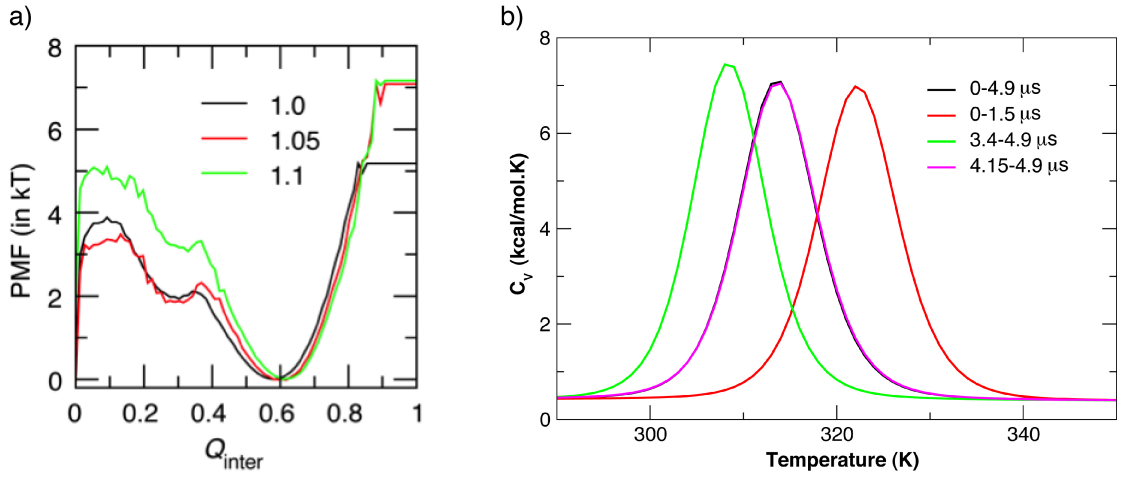


Figure S3.3: (A) The PMFs as functions of Q_{inter} with three of scaling values of the intermolecular interaction strength searched during model calibration. (B) The heat capacity as a function of temperature. The curves were calculated from various segments of a 4.9-μs REX simulation of the NCBD/ACTR complex using the final calibrated model. The simulation appears to reasonably converged after 4 μs. Figure prepared by DG WZ.

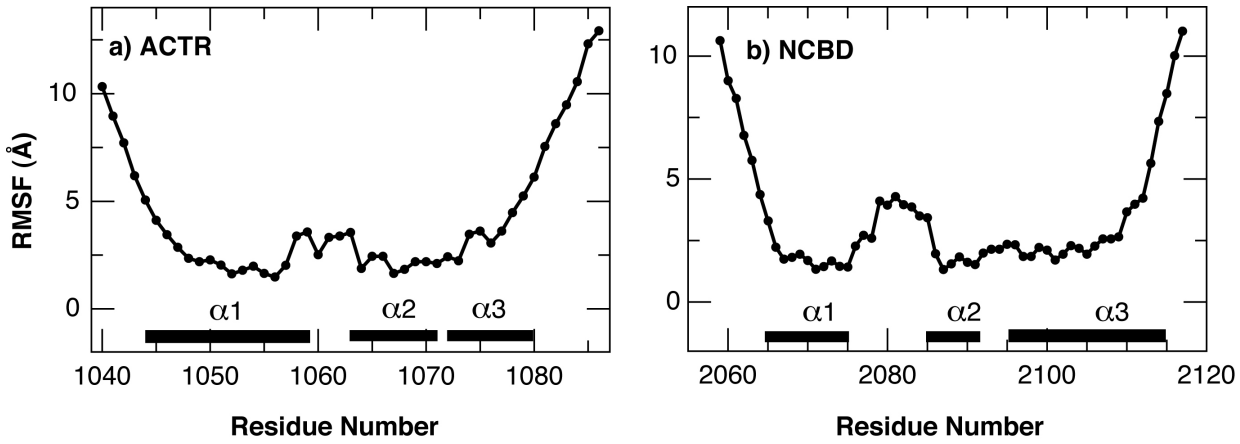


Figure S3.4: The RMSF profiles of NCBD and ACTR in the bound state at 300 K. These profiles were calculated from a 1-μs simulation of the folded complex using the final calibrated model.Figure prepared by DG.

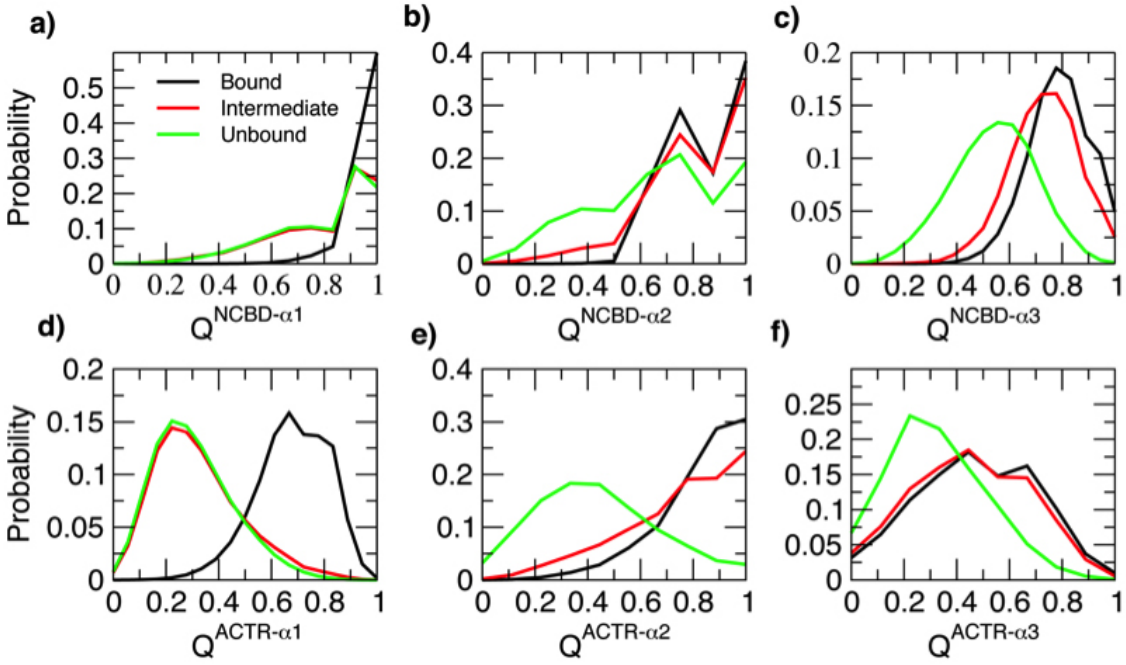


Figure S3.5: Distributions of the fractions of intra-molecular contacts of various ACTR and NCBD helical segments in the unbound, intermediate and bound states. See the main text for the state assignment criteria. Figure prepared by DG.

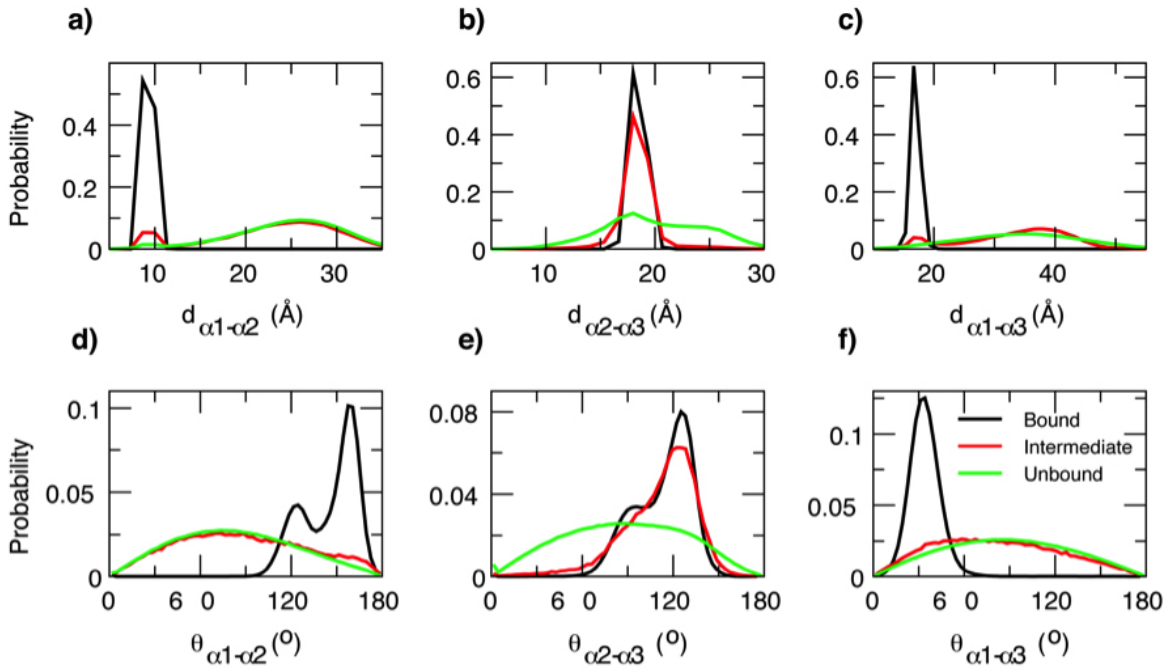


Figure S3.6: Distributions of the distances and cross-angles between NCBD helices in the unbound, intermediate and bound states. The helix-helix distances were calculated as the distances between the Cα atoms at the middle of the helices. See the main text for the state assignment criteria. Figure prepared by DG.

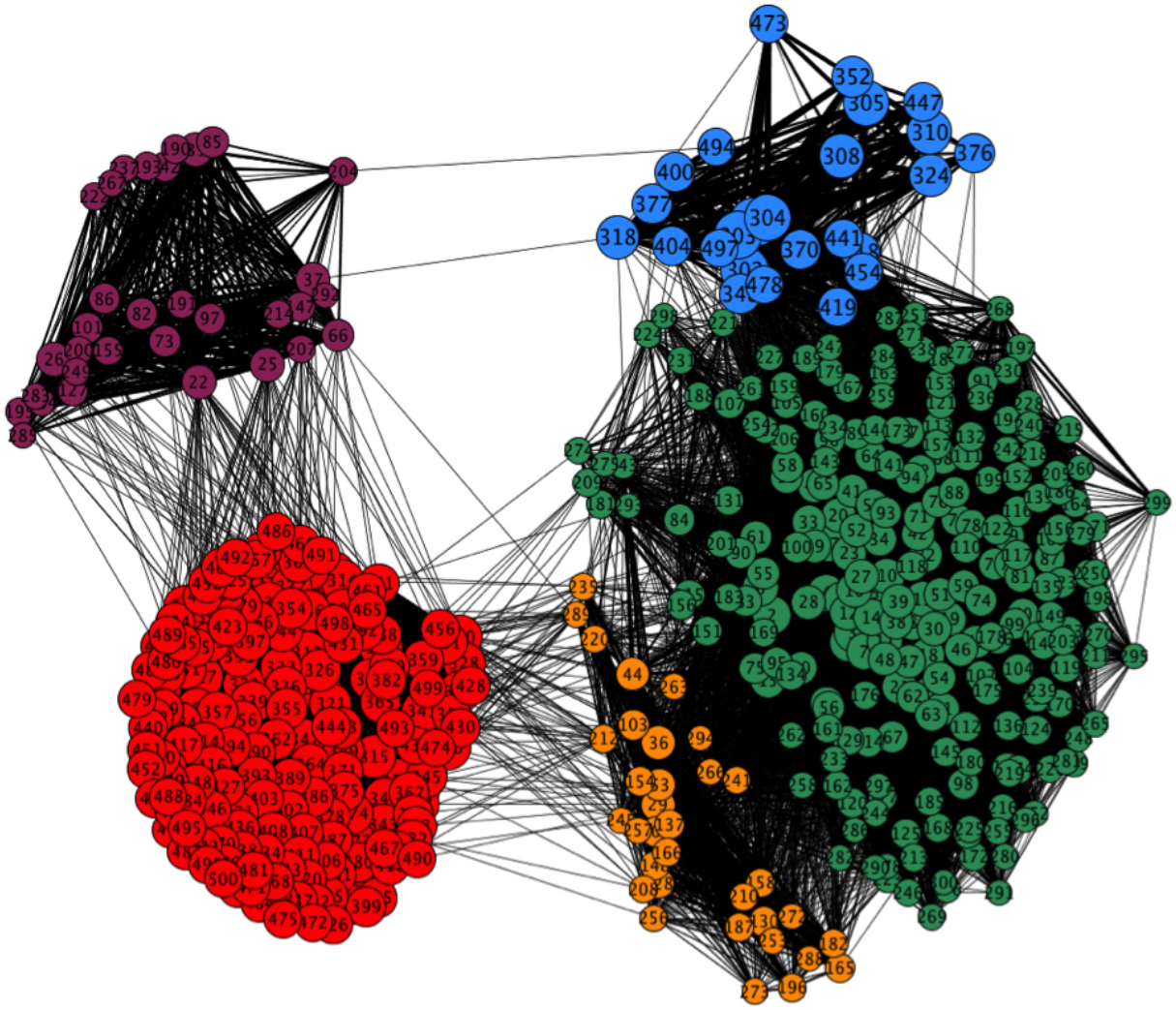


Figure S3.7: CSN of the synergetic folding of NCBD and ACTR, constructed by including the most populated 200 nodes and additional 300 nodes from the transition paths. The nodes are colored in the same fashion as in Figure 3.9 of the main text.

Chapter 4

Electrostatically Accelerated Encounter and Folding for Facile Recognition of Intrinsically Disordered Proteins*

Abstract

Achieving facile specific recognition is essential for intrinsically disordered proteins (IDPs) that are involved in cellular signaling and regulation. Consideration of the physical time scales of protein folding and diffusion-limited protein-protein encounter has suggested that the frequent requirement of protein folding for specific IDP recognition could lead to kinetic bottlenecks. How IDPs overcome such potential kinetic bottlenecks to viably function in signaling and regulation in general is poorly understood. Our recent computational and experimental study of cell-cycle regulator p27 (Ganguly *et al.*, J. Mol. Biol. (2012)) demonstrated that long-range electrostatic forces exerted on enriched charges of IDPs could accelerate protein-protein encounter via "electrostatic steering" and at the same time promote "folding-competent" encounter topologies to enhance the efficiency of IDP folding upon encounter. Here, we further investigated the coupled binding and folding mechanisms and the roles of electrostatic forces in the formation of three IDP complexes with more complex folded topologies. The surface electrostatic potentials of these complexes lack prominent features like those observed for the p27/Cdk2/cyclin A complex to directly suggest the ability of electrostatic forces to facilitate folding upon encounter. Nonetheless, similar

*Reprinted under the Creative Commons Attribution (CC BY) license.

D. Ganguly, W. Zhang and J. Chen (2013), "Electrostatically Accelerated Encounter and Folding for Facile Recognition of Intrinsically Disordered Proteins." PLoS Comput. Biol. 9(11): e1003363. doi:10.1371/journal.pcbi.1003363.
Author Contributions: Conceived and designed the experiments: JC DG WZ. Performed the experiments: DG WZ JC. Analyzed the data: DG WZ JC. Contributed reagents/materials/analysis tools: DG WZ JC. Wrote the paper: DG WZ JC.

electrostatically accelerated encounter and folding mechanisms were consistently predicted for all three complexes using topology-based coarse-grained simulations. Together with our previous analysis of charge distributions in known IDP complexes, our results support a prevalent role of electrostatic interactions in promoting efficient coupled binding and folding for facile specific recognition. These results also suggest that there is likely a co-evolution of IDP folded topology, charge characteristics, and coupled binding and folding mechanisms, driven at least partially by the need to achieve fast association kinetics for cellular signaling and regulation.

4.1 Introduction

Cellular signaling and regulation are frequently mediated by proteins that, in part or as a whole, lack stable structures under physiological conditions[3,146,147]. Such intrinsically disordered proteins (IDPs) are highly prevalent in proteomes[12] and over-represented in diseases pathways[15,17]. For example, nearly one-third of eukaryotic proteins have been predicted to contain extended disordered regions[257], and about 25% of disease-associated missense mutations can be mapped into predicted disordered regions[18] (although cancer mutations appear to prefer ordered regions[16]). The prevalence of intrinsic disorder suggests that protein conformational heterogeneity could provide crucial functional advantages, for which many concepts have been proposed[20,23,24,166,224]. Understanding the physical basis of how intrinsic disorder mediates protein function (and how such functional mechanism may fail in human diseases[258]) is of fundamental significance and has attracted intense interests in recent years[35]. Important progresses have been made on characterizing the conformational properties of unbound IDPs and determining how these conformational properties contribute to efficient and reliable interactions[1,26,35,52,160,259,260].

A key recent recognition is that frequent requirement of protein folding for specific recognition of IDPs could lead to kinetic bottlenecks[34,173,261]. As predicted by the dual-transition-state theory[173], the diffusion-limited encounter rate constant represents the upper bound for that of a coupled binding and folding interaction. Importantly, the upper bound can be achieved only if the IDP readily folds upon encounter, which requires folding rates on the order of 10 μs^{-1} or greater[173]. That is, IDPs need to achieve folding rates beyond the typical μs^{-1} "speed limit" estimated for folding of isolated proteins[249] to maximize association kinetics. Therefore, the putative functional advantages of intrinsic disorder, especially structural plasticity for specific interactions with numerous partners[22], come with a potential cost of slow binding kinetics. Such kinetic bottleneck must be resolved for IDPs to be viable in cellular signaling and regulation. Interestingly, a recent survey of binding kinetic data revealed that IDP binding was not systematically slower than that of

globular proteins[172]. The implication is that most IDPs do manage to fold rapidly upon nonspecific binding, and this is apparently consistent with the accumulating observations that IDP coupled binding and folding tends to follow induced folding-like baseline mechanisms (i.e., bind then fold)[26,35]. Several factors could contribute to efficient folding of IDPs upon binding, in particular small interacting (and folding) domains and simple folded topologies with low contact orders. There also appears to be a delicate balance between pre-folding and conformational flexibility that allows an IDP to quickly fluctuate among accessible conformational states, especially upon encounter[35,205,262]. Nonetheless, it is not yet clear how in general IDPs may achieve fast folding at rates beyond the traditional μs^{-1} folding "speed limit" upon encountering their specific targets.

An important characteristics of IDPs is that they are enriched with charged and polar residues[263]. Electrostatics can thus be expected to play key roles in IDP structure and function. For example, the charge content can modulate compaction and other conformational properties of free IDPs[188,264]; DNA search efficiency is controlled by charge composition and distribution in disordered tails of DNA-binding proteins[250,251]. It has been also observed or speculated in a few cases that electrostatics might be important for fast IDP recognition[265–268]. However, these discussions have been often based on the classic electrostatic steering effects[209], and the actual underlying mechanisms of putative electrostatic acceleration were not known. Our recent computational and experimental study of the p27-Cdk2/cyclin A interaction revealed that long-range electrostatic forces could promote facile IDP recognition via an "electrostatically accelerated encounter and folding mechanism"[261]. Specifically, the measured p27/Cdk2/cyclin A association rate constants showed a strong salt-dependence, increased $\sim$12 fold when the ionic strength was reduced from 0.6 to 0.075 M. However, the salt-dependence is poorly described by an approximate Debye-Hückel relation[269] that mainly captures the electrostatic steering effects. Instead, simulations using a series of topology-based coarse-grained models suggested that long-range electrostatic forces exerted on a large number of charges on p27 did not only accelerate the encounter rate (via the classical electrostatic steering effect[209]), but enhance the efficiency of p27 folding upon encounter by promoting native-like encounter topologies.

Analysis of surface charges in a set of existing IDP complexes further revealed that the vicinity of IDP binding sites tended to be enriched with charges to complement those on IDPs[261] (even though the IDP binding interface itself is more hydrophobic than the rest of the protein surface as previously observed[21]). Electrostatic forces are known to be a dominant long-range force that can guide protein orientation in protein-DNA interactions[270,271] and/or modulate early stages of protein folding[99,272,273]. One implication of enriched charges near IDP binding sites is thus that the electrostatically accelerated encounter and folding mechanism observed for p27 may be prevalent in signaling and regulatory IDPs. Nonetheless, the ability for long-range electrostatic forces to enhance folding upon binding can be surprising, as nonspecific interactions (electrostatic or van der Waals) have been generally expected to accelerate binding but slow down folding[253,274]. It has also been predicted that, while inter-chain electrostatic interactions facilitate binding of disordered chaperone Chz1 to histone variant H2A.Z-H2B, intra-chain electrostatic interactions could lead to premature collapse of Chz1 under low salt conditions and hinder the overall rate of forming the specific complex[275].

In the present work, we investigated the recognition mechanisms and the roles of long-range electrostatic interactions in forming of three IDP complexes, namely, p53-TAD1/TAZ2, HIF-1α/TAZ1, and NCBD/ACTR (Table 4.1). All these complexes have important biological functions. For example, tumor suppressor p53 is considered one of the most important proteins in cancer[276]; NCBD and TAZ1/2 are key regulatory domains of CBP, a key component of the general transcriptional machinery that plays critical roles in cell fate regulation[190]. For understanding IDP recognition, these systems involve more complex folded topologies than that of p27 in the p27/Cdk2/cyclin A complex. As shown in Figure 4.1, both HIF-1α/TAZ1 and NCBD/ACTR possess extensive binding interfaces, whereas the binding interface in p53-TAD1/TAZ2 is more localized. Importantly, while strong charge complementary exists near the binding interface (as expected), the surface electrostatic potentials of the folded substrates do not show prominent features like those observed on Cdk2/cyclin A (e.g., see Figure 1 of reference[261]) to directly suggest that long-range electrostatic forces could promote native-like (and thus more folding-

Table 4.1: Key properties of three IDP complexes.*

Name[a]	Length	K_D[b]	k_{on}($M^{-1}s^{-1}$)	PDB	IDP Fold	Charges[c]
p53-TAD1/TAZ2	39/90	2.7μM[277]	$\sim10^{8}$ [d]	2k8f	helix/loops	8(-6),9(+9),4(+2)
HIF-1α/TAZ1	51/99	7nM[278]	1.3×10^{9} [279]	1l8c	helices/loops	11(-5),11(+7),10(+5)
NCBD/ACTR	59/47	34nM[194]	3.0×10^{7} [280]	1kbh	helices	(both IDPs)

*. Data of p53-TAD1/TAZ2 and HIF-1α/TAZ1 are obtained by DG.
a. Abbreviations: ACTR: the activation domain of p160 steroid receptor co-activator; HIF-1α: hypoxia-inducible factor 1 α subunit; NCBD: the nuclear-receptor co-activator binding domain of CREB binding protein (CBP); p53-TAD1: the transactivation domain 1 of tumor suppressor p53; TAZ1/2: the TAZ domains of CBP. The sequences of all IDPs involved (highlighted in bond fonts) are provided in the Supporting Information.
b. The experimental K_D values were measured at 308 K for p53-TAD1/TAZ2, 298 K for HIF-1α/TAZ1, and 304 K for NCBD/ACTR. Note that K_D only weakly depends on temperature for p53-TAD1/TAZ2 (doubled when the temperature is increased from 288K to 308K[277]).
c. Numbers of charged residues and the net charges (in parentheses) of the IDP, its binding site, and the vicinity of the binding site. Residues at the IDP binding interface are identified as those with greater than 1.0 Å^2 solvent accessible surface area changes upon complex formation. Surface residues are identified as those with $>5\%$ solvent accessibility. All surface residues within 15 Å Cα-Cα distance from the bound IDP but not directly involved in intermolecular contacts are considered to be within the vicinity of the IDP binding site.
d. Estimated based on the association rate constant of p53-TAD2/TAZ2 (10^{10} $M^{-1}s^{-1}$ [265]), assuming that TAD1 and TAD2 have similar off rates. TAD2 binds to the TAZ2 primary site with K_D 32 nM[265], about two orders of magnitude stronger than TAD1.

competent) encounter complexes. The NCBD/ACTR complex involves synergistic folding of two IDPs and thus offers a particularly interesting opportunity to understand whether and how electrostatic interactions may modulate the formation of nontrivial folded topologies. Amazingly, all three complexes associate with on-rates in excess of 10^7 $M^{-1}s^{-1}$ (see Table 4.1), a regime that is typically considered "diffusion-limited" and can only be accessed in the limit of ultrafast conformational transitions[209].

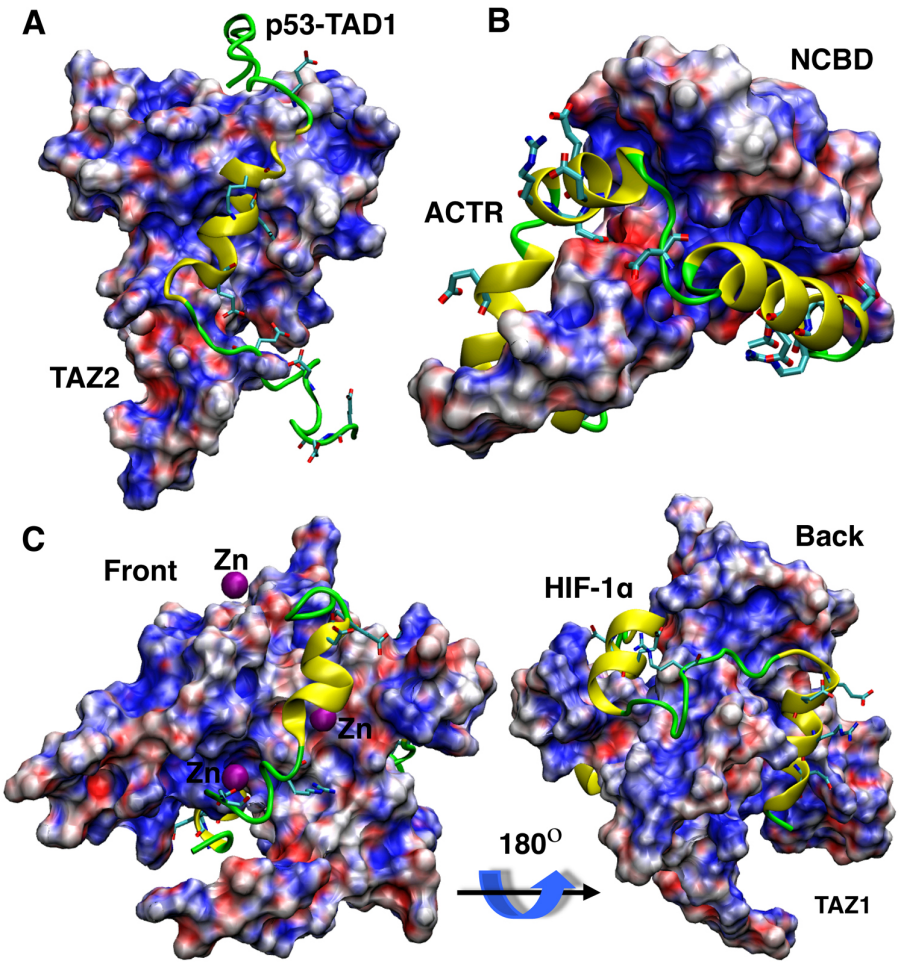


Figure 4.1: Structures and surface electrostatic potentials of three complexes. (A) p53-TAD1/TAZ2, (B) NCBD/ACTR, and (C) HIF-1α/TAZ1. TAZ2, NCBD and TAZ1 are shown in molecular surface and colored based on the surface electrostatic potential calculated using PBEQ module of CHARMM[281,282]. Red indicates negative and blue indicate positive charge. p53-TAD1, ACTR and HIF-1α are shown in cartoons, with charged side chains shown in stick. Figure (A) and (C) are prepared by DG.

4.2 Results

4.2.1 Topology-based modeling of IDP coupled binding and folding

Series of topology-based coarse-grained models were first derived based on the complex structures to allow direct simulation of reversible binding and folding with tractable computational

cost. Topology-based modeling is based on the theoretical framework of minimally frustrated energy landscapes for natural proteins[98], and has been highly successful in predicting essential features of protein folding mechanisms[97,98,283]. Formation of stable IDP complexes such as those studied in this work should also satisfy minimal frustration, and thus topology-based modeling is applicable. Indeed, it has been successfully applied to several IDP complexes[33,92,93,163,165], with many key predictions substantiated by independent experimental studies. Nonetheless, important differences do exist between IDPs and structured proteins in sequence compositions and binding interface characteristics[21]. We have previously demonstrated that traditional topology-based models need to be carefully calibrated to ensure proper balance among competing intramolecular and intermolecular interactions (see Methods for detail on the calibration protocol)[62]. We note that the importance of model calibration was also illustrated in a recent study of the HIF-1α/TAZ1 complex[92].

Table 4.2 summarizes the final calibrated models for all three complexes. The calculated residual helicity distributions of the unbound states are show in Figure S4.1. Three independent models were constructed for each complex: one without explicit charges (mimicking high salt concentration with fully screened long-range electrostatic interactions), one with explicit charges (mimicking low salt concentration with unscreened long-range electrostatic interactions), and a third one with explicit charges and 0.05 M salt (mimicking physiological conditions). All models reproduce the experimental K_D to the same order of magnitude, except that the no charge model for HIF-1α/TAZ1 yields a K_D value about one order of magnitude too large. We note that calculated K_D values can be very sensitive to small changes of in the scaling of intermolecular interactions during model calibration (see Methods). It is computationally expensive to use REX simulations to systematically search for the parameter space, especially for models without explicit charges due to slower transitions. Nonetheless, by performing production simulations at the corresponding melting temperatures, remaining imperfections in the balance of various interactions should be further suppressed, allowing reliable comparative studies of the mechanistic roles of electrostatic interactions in coupled binding and folding.

Table 4.2: Dissociation constants, melting temperatures, average reversible coupled binding and folding transition rates calculated using various coarse-grained models with and without explicit charges and/or 0.05 M salt.

Models	**Calc. K_D**	**T_m (K)**	**k_{TS} (ms^{-1})**	**k_{cap} (ns^{-1})**	**k_{esc} (ns^{-1})**	**k_{evo} (ns^{-1})**
TAD1/TAZ2						
No charge	1.4 $\pm$ 2.0 μM	327	4.3 $\pm$1.5	1.4	8	0.049
Charged, 0.05M salt	1.6 $\pm$ 1.6 μM	340	14.5 $\pm$ 1.1	3.2	4.1	0.08
Explicit charges	4.9 $\pm$ 3.2 μM	335	27.0 $\pm$ 0.2	32.1	0.1	0.16
HIF-1α/TAZ1						
No charge	64 $\pm$ 64 nM	327	6.1 $\pm$ 0.5	2.8	6.3	0.022
Charged, 0.05M salt	9.4 $\pm$ 9.6 nM	340	10.2 $\pm$ 1.8	3.4	5.6	0.039
Explicit charges	1.3 $\pm$ 1.6 nM	345	29.4 $\pm$ 3.7	5	0.69	0.048
NCBD/ACTR						
No charge	67 $\pm$ 99 nM	318	0.53 $\pm$ 0.2	0.13	0.61	0.0043
Charged, 0.05M salt	96 $\pm$ 92 nM	315	1.7 $\pm$ 0.1	0.29	0.31	0.0074
Explicit charges	39 $\pm$ 14 nM	322	5.2 $\pm$ 0.7	0.79	0.02	0.012

K_D was calculated from REX simulations at 300 K(see Table 4.1 for the experimental values); k_{TS} was calculated from the production Langevin simulations at the corresponding T_m, as k_{TS} = N_{TS}/t_{tot}, where N_{TS} is the number of reversible binding and folding transitions observed during the total simulation time span t_{tot}. As all simulations were performed at T_m, k_{TS} as defined is half of the binding and unbinding rates. k_{cap}, k_{esc} and k_{evo} are defined in Equation 4.1-4.4. The effective concentrations of these simulations are 1.66 mM, 1.66 mM and 1.43 mM for p53-TAD1/TAZ2, HIF-1α/TAZ1 and NCBD/ACTR, respectively. All uncertainties were estimated as the differences between results calculated from the first and second halves of the data. Data of p53-TAD1/TAZ2 and HIF-1α/TAZ1 are obtained by DG.

4.2.2 Baseline mechanisms of coupled binding and folding: effects of electrostatic forces

Free energy surfaces were constructed using various combinations of folding and binding order parameters to understand the baseline mechanisms of coupled binding and folding and to dissect the effects of long-range electrostatic forces. In particular, the fractions of native contacts formed have been shown to provide natural reaction coordinates for such mechanistic analysis[237]. Figure 4.2 compares the free energy surfaces as a function of intra- and inter-molecular native contact factions for all three complexes, calculated using calibrated Gō-like models with and without explicit charges and/or salt (see Table 4.2). Both p53-TAD1 and HIF-1α recognitions follow induced folding-like mechanisms, where the peptides only gain structures after forming significant numbers of native intermolecular contacts. For example, Figure 4.2A shows that p53-TAD1 does not start to fold until Qinter reaches $\sim$0.5. Free NCBD is a molten globule with folded-like secondary structures[199], and its synergistic folding with ACTR has been previously shown to involve multiple stages of selection and induced folding[34,163], reminiscent of the "extended conformational selection" mechanism[205]. Nonetheless, neither protein gains significant secondary (for ACTR) or tertiary (for NCBD) structures until over 20% of native intermolecular contacts are formed (Figure 4.2 G and J).

Interestingly, formation of all three complexes involves intermediates, even though the intermediate in p53-TAD/TAZ2 interaction only become pronounced in the presence of nonspecific electrostatic forces (see Figure 4.2 A vs. C). Detailed examination of the simulation trajectories and various free energy surfaces using fractions of native contacts formed by different IDP segments (e.g., see Figure S4.2-S4.4) revealed the existence of multiple parallel pathways for forming HIF-1α/TAZ1 and NCBD/ACTR. While these mechanistic details are not the focus of the current work, they appear to be highly consistent with previous experimental and computational studies. For example, as shown in Figure S4.2, both the first and third helices of HIF-1α could initiate recognition, with the pathway initiated by the third helix binding being much more prevalent.

Similar observations were also made in a separate computational study[92]. Specific recognition of NCBD/ACTR appears to be primarily initiated by the C-terminal segments of these two peptides (Figure S4.3-S4.4), which forms a key intermediate that was also suggested by an H/D exchange mass spectrometry study[58]. Kinetic data from a recent stop-flow study of the NCBD/ACTR interaction[280] are consistent with the prediction of induced folding as a baseline mechanism and have further confirmed the existence of parallel pathways and multiple folding intermediates. Representative snapshots along the dominant binding and folding pathways of p53-TAD1/TAZ2 and HIF-1α/TAZ1 are shown in Figure S4.5-S4.6.

Explicit inclusion of charges does not significantly perturb the baseline mechanisms of coupled binding and folding. As shown in Figure 4.2 and Figure S4.2-S4.4, long-range electrostatic forces do not lead to fundamental changes in any of the free energy surfaces examined. The baseline mechanisms for the formation of all three complexes remain induced folding-like. Furthermore, nonspecific electrostatic interactions do not change the relative prevalence of the parallel pathways that exist. For example, HIF-1α still initiates binding mainly through the third helix (Figure S4.2); synergistic folding NCBD and ACTR is still mainly initiated through their C-terminal segments (Figure S4.3-S4.4). The key effect of electrostatic forces appears to be substantial reductions in the free energy barriers that separate various basins. That is, even under the no salt condition, strong nonspecific electrostatic interactions do not appear to add to the ruggedness of coupled binding and folding free energy surfaces. An implication is that there exists a level of self-consistency between the charge distribution and folded topology in the bound states, despite a lack of apparent complementary between folding topologies and surface electrostatic potentials for these IDP complexes (see Figure 4.1).

4.2.3 Kinetic effects of long-range and nonspecific electrostatic forces

Kinetics of coupled binding and folding was derived directly from production Langevin dynamics simulations performed using the calibrated Gō-like models at their corresponding T_m. The results, summarized in Table 4.2, show that long-range electrostatic forces accelerate the reversible binding/unbinding transition rates for all three complexes. The overall electrostatic acceleration, estimated by comparing the average transition rates (k_{TS}) calculated using models with and without explicit charges, ranges from $\sim$5 fold for HIF-1α to 10 fold for NCBD/ACTR. The magnitude of acceleration is similar to what was previously measured for other IDPs including p27[261] and PUMA[267] (both $\sim$10 fold). The presence of 0.05 M salt significantly attenuates the predicted electrostatic acceleration, to only about two fold. However, the effect of salt screening on electrostatic acceleration is likely over-predicted[261], which is due to the Cα-only model used in this work and may be corrected with more detailed protein models[99]. Consistent with the kinetic analysis, there are significant reductions in the free energy barriers along Q_{inter} (see Figure 4.3), which has been shown to be a good binding reaction coordinate[62]. In addition, the magnitude of barrier reduction correlates well with the degree of rate acceleration calculated directly from Langevin dynamics simulations, with the largest barrier reduction observed for NCBD/ACTR and the smallest reduction observed from HIF-1α/TAZ1.

To further analyze the effects of electrostatic interactions on different stages of coupled binding and folding, the recognition process was divided into two generic steps, including an encounter step followed by an evolving (folding) step to final bound and folded state (Equation 4.1 in Methods). Such generic decomposition ignores the details of IDP-specific folding pathways, to allow on to focus on the net effects of electrostatic forces on the overall efficiency of IDP folding upon encounter. For this, three general states were identified during production simulations, including the unbound (U), collision complex (CC), and bound (B) states (see Methods for specific criteria for state assignment). The mean first passage times (MFPT) and numbers of transitions

(N_{trans}) among these states were then calculated. The results, summarized in Table S4.1-S4.3, show that long-range electrostatic forces greatly reduce the average encounter time, from 0.72 to 0.03 ns for p53-TAD, from 0.37 to 0.20 ns for HIF-1α, and from 7.71 to 1.26 ns for NCBD. At the same time, long-range electrostatic forces also significantly enhance the efficiency of IDP folding upon encounter, allowing much larger fractions of the encounter complexes to eventually evolve to the bound states. For example, for NCBD/ACTR, only 16 out $\sim$2300 encounter events evolved to the bound state in absence of long-range electrostatic forces (0.7%); whereas with explicit charges, there was $\sim$37% probability (108 out of 288) of forming the specific complex once the proteins were captured into the collision complex state (Table S4.3). For the HIF-1α/TAZ1 complex, the percentages of collision to specific complex transition are 0.4% without and 7% with explicit charges (Table S4.2); for p53-TAD1/TAZ2, the production percentages are 0.6% without and 60% with explicit charges (Table S4.1). It should be emphasized that nonspecific electrostatic interactions significantly stabilize the collision complexes, due to large and complementary net charges of the interacting proteins (see Table 4.1). As such, much fewer fully unbinding events were observed during production simulations using the charged models. This effect also led to more reversible transitions between the bound and collision complex states and thus an overestimation of the true folding efficiency of IDPs upon collision as estimated above. We also note that the collision complexes as defined in our analysis were not intended to represent so-called "encounter complexes" that have been often considered key intermediates of protein-protein association[284], although encounter complexes are also believed to be mainly stabilized by nonspecific electrostatic interactions.

The enhanced apparent efficiency of folding upon encounter appears to be frequently achieved at the cost of longer folding times. For example, the MFPTs of transitions from the collision complexes to the bound states increase from 0.26 to 3.94 ns for the p53-TAD1/TAZ2 complex (Table S4.1) and from 8.14 to 44.56 ns for the NCBD/ACTR complex (Table S4.3). The net effects on the kinetics of encounter and folding stages can be quantified by calculating three effective rate constants as defined in Equation 4.2-4.4 (see Methods)[172]. The results, summarized in Table 4.2

and plotted in Figure 4.4, clearly demonstrate that nonspecific electrostatic interaction enhance the encounter rates and reduce the escape rates of the collision complexes. Importantly, the effective evolution rates are always faster, by about three fold, in the presence of long-range electrostatic forces, despite longer MFPTs for the transitions from the collision complexes to the bound state observed for the p53-TAD1/TAZ2 and NCBD/ACTR complexes. The magnitude of electrostatic acceleration of folding upon encounter is similar to what was previously observed for folding and binding of p27 to the Cdk2/cyclin A complex[261].

4.2.4 Mechanism of electrostatically accelerated folding upon encounter

Inspection of the conformational properties of the collision complexes provides further insights into the molecular basis for enhanced efficiency of IDP folding upon encounter due to long-range electrostatic forces. As shown in Figure 4.5, without nonspecific electrostatic interactions (models without explicit charges), the initial contacts between two binding partners are largely random, and the distributions of IDP initial contact points on the substrate surface in the collision complexes are relatively uniform (left column). In contrast, with the inclusion of explicit charges, the probabilities of IDP encountering near the native binding interface are dramatically increased. Coupled with reduced escape rates, this allows much higher efficiency of IDP folding upon encounter to achieve higher overall association rate constants (Table 4.2). The ability of long-range electrostatic forces to guide the recognition process is also reflected in the free energy surfaces as a function of binding RMSD of the IDP and center of mass separation between two peptides. As shown in Figure 4.6, long-range electrostatic forces generate a strong free energy gradient that extends over 10-15 Å away from the native bound positions, without creating over-stabilized misfolded states at short separation distances. It is intriguing that, even though both NCBD and ACTR are disordered in the unbound state, nonspecific long-range electrostatic forces between complementary charges on these two proteins can still manage to promote native-like topologies in the collision complexes. In particular, there is a much higher probability of NCBD and ACTR

initiating contacts via the C-terminal helix of NCBD and the second helix of ACTR (Figure 4.5 E-F). This is part of a key pathway of synergistic folding inherent to the NCBD/ACTR complex that was predicted by coarse-grained and atomistic simulations[34,163] and later substantiated by H/D exchange mass spectrometry[58]. Therefore, nonspecific electrostatic interactions appear to mainly augment existing folding pathways inherent to the folded topologies to facilitate efficient folding of IDPs upon encounter. Coupled with the previous observation that the vicinity of the IDP binding site tends to be enriched with charges to complement those on IDPs[261], the current results suggest that there is likely a co-evolution of IDP folded topology, charge characteristics, and coupled binding and folding mechanisms. Furthermore, the co-evolution is likely driven by the important need to achieve facile IDP recognition for cellular signaling and regulation.

4.3 Discussion

While fulfilling important functional constraints such as structural plasticity for binding numerous specific targets, protein intrinsic disorder can lead to potential kinetic bottlenecks to be viable in cellular signaling and regulation. Our previous work on the p27/Cdk2/cyclin A complex has revealed a mechanism where nonspecific electrostatic interactions not only enhance the protein-protein encounter kinetics but also promote folding-competent encounter topologies to increase the efficiency of IDP folding upon encounter[261]. Using carefully calibrated topology-based coarse-grained models, we have now further demonstrated that similar electrostatically accelerated encounter and folding mechanisms also underlie the formation of three IDP complexes with more complexed folded structures, namely, p53-TAD1/TAZ2, HIF-1α/TAZ1, and NCBD/ACTR. Importantly, these complexes lack apparent features on the electrostatic surface potentials to directly suggest the ability of nonspecific long-range electrostatic forces to promote native-like encounter topologies to enhance the IDP folding efficiency upon encounter. Nonetheless, there seems to exist a sufficient level of self-consistency between the charge distributions and folded topologies in the

bound state to allow accelerated recognition in presence of nonspecific electrostatic interactions. Therefore, enriched charges on IDPs not only play key roles in modulating the conformational properties of the unbound state, but also likely play general and important roles in regulating efficient interactions of IDPs with specific partners. We note that IDPs are frequently regulated by post-translational modifications that add or remove charges. Improved mechanistic understanding of electrostatic forces in IDP recognition derived from the current work will thus help to dissect the profound impacts of post-translational modifications and disease-related mutations on IDP structure and interaction.

4.4 Methods

4.4.1 Calibration of topology-based coarse-grained models with and without explicit charges

Cα-only sequence-flavored Gō-like models[232] were first derived from the complex structures of p53-TAD1/TAZ2, HIF1-α/TAZ1 and NCBD/ACTR (see Table 4.1) using the Multiscale Modeling Tools for Structural Biology (MMTSB) Gō-Model Builder (http://www.mmtsb.org)[215]. The 3 zinc ions bound to TAZ1 in the HIF1-α/TAZ1 complex were modeled explicitly with distance restraints to the coordinating residues. All three models were then calibrated to balance the intrinsic folding propensity and the strength of intermolecular interactions using a previously described protocol[62]. Briefly, the strengths of intra-molecular native contact were uniformly scaled to reproduce the experimentally measured residual helicity of unbound IDPs, which are mainly based on NMR secondary chemical shift and/or circular dichroism analysis (p53-TAD1[285], NCBD/ACTR[199], and HIF1-α[278]). The residual helicity distributions calculated using the final models listed in Table 4.2 are provided in Figure S4.1. Then, the strengths of intermolecular contacts were adjusted, such that binding affinities calculated from replica exchange molecular dynamics (REX-MD) simulations approximately match the experimental values (see Table 4.1).

Following the previously described procedure[261], the calibrated sequence-flavored Gō-like models were then further modified by assigning proper explicit charges to all charged residues (Lys, Arg, Glu and Asp) as well as zinc ions in the HIF1-α/TAZ1 complex. The charged models were then re-calibrated to reproduce the experimental residual structure level (Figure S4.1) and binding affinity (Table 4.2). Such calibration is critical to avoid inherent bias for particular types of interactions, e.g., intra- vs. inter-molecular or native vs. nonspecific electrostatic. Nonspecific electrostatic interactions were modeled using the Debye-Hückel potential to account for ionic screening. The dielectric constant was set at 80.

4.4.2 Simulation protocols

The complexes were simulated in cubic boxes with periodic boundary conditions imposed in CHARMM[63,216]. The box sizes are 100, 100 and 105 Å for p53-TAD1/TAZ2, HIF-1α/TAZ1 and NCBD/ACTR, respectively. Langevin dynamics was performed with 15 fs time steps and a friction coefficient of 0.1 ps^{-1}. SHAKE was used to fix all virtual bond lengths[217]. Non-bonded interactions were cut off at 25 Å. Unbound IDPs were simulated at 300 K for 750 ns to calibrate the intramolecular interactions. REX-MD was performed using the MMTSB Toolset[215] for calibration of the intermolecular interactions. For this, eight replicas spanning 270 to 400 K were used. The lengths of REX calibration simulations ranged from 1.05 μs (for p53-TAD1/TAZ2) up to 10 μs (for NCBD/ACTR), as needed for achieving sufficient convergence. Temperature weighted histogram analysis method (WHAM)[286] was used to compute the heat capacity (C_V) curves and generate unbiased probability distributions for free energy and thermodynamic analysis. In particular, the dissociation constants (K_D) were calculated from the bound and unbound probabilities at 300 K[62], where the unbound state was defined as the state without any native intermolecular contacts formed. For NCBD/ACTR complex, the 1D free energy profile lack significant barriers between the unbound and partially bound intermediate states (Figure 4.3C, red trace). Therefore, the unbound probability was calculated as 1 - P_{bound}, where P_{bound} is the bound probability (see below

for the specific criteria of state assignments). Once calibrated, production simulations of 30-40 μs in lengths were performed using all models at the corresponding T_ms (see Table 4.2). The T_m value was first identified based on the C_V curve and then fine tuned to ensure that similar probabilities of sampling the bound and unbound states were observed in the production simulation.

4.4.3 Free energy and kinetic analysis

All free energy profiles were calculated from the REX simulations and the kinetic analysis was performed based on the production simulations, unless otherwise stated. For calculation of contact fractions, a given native contact was considered as formed if the inter-Cα distance was within 1.0 Å of the distance in the native complex. Nonspecific intermolecular contacts are considered as formed when the inter-Cα distance is within 10 Å cutoff. Three general conformational states were defined for each complex, including the unbound (U), collision complex (CC) and bound (B) states, to understand the effects of electrostatic forces on protein-protein encounter and subsequent folding upon encounter. The unbound state includes conformations with no specific or nonspecific contacts formed between IDP and substrate, and the collision complex state includes conformations with at least one nonspecific but no specific intermolecular contact formed. The bound states are defined as following: 1) for p53-TAD1/TAZ2: $N_{inter} \geq 11$; 2) for HIF-1α/TAZ1: $N_{inter} \geq 26$ for the no charge model, $N_{inter} \geq 23$ for the charged model, and $N_{inter} \geq 24$ for the charged model with 0.05 M salt; 3) for ACTR/NCBD: $N_{inter} \geq 30$. N_{inter} is the total number of native intermolecular contacts formed. Note that slightly different criteria were used to define the bound state of HIF-1α/TAZ1 due to small shifts of the bound free energy basins calculated using different models (see Figure 4.3). 15-ps running averages were used for assigning states, to avoid including fictitious transitions due to rapid small fluctuations in the calculated contact counts (especially between the U and CC states). MFPTs and numbers of transitions among all three states were derived from the production simulation trajectories, and various rates were

calculated as defined in Equation 4.2-4.4.

$$U \xleftrightarrow{k_{cap}/k_{esc}} CC \xrightarrow{k_{evo}} B \tag{4.1}$$

$$k_{cap} = MFPT_{cap}^{-1} \tag{4.2}$$

$$k_{esc} = (\frac{MFPT_{esc} \times N_{esc} + MFPT_{evo} \times N_{eov}}{N_{esc} + N_{evo}})^{-1} \times \frac{N_{esc}}{N_{esc} + N_{evo}} \tag{4.3}$$

$$k_{evo} = (\frac{MFPT_{esc} \times N_{esc} + MFPT_{evo} \times N_{eov}}{N_{esc} + N_{evo}})^{-1} \times \frac{N_{evo}}{N_{esc} + N_{evo}} \tag{4.4}$$

Here, k_{cap}, k_{esc}, and k_{evo} are the capture, escape (to the unbound state) and evolution (to the bound state) rates of the collision complex, respectively; N_{esc} and N_{evo} are the numbers of escape and evolution transitions. Note that the MFPTs calculated correspond to the average times spent in an initial state before a transition to the final state. Ideally, the average lifetime of CC should be independent of whether the trajectory ends up in either the U or B state for a true three-state model as shown in Equation 4.1. However, The actual transitions between the CC and B states involve several intermediates that are not represented in Equation 4.1, and the effective MFPTs as calculated thus depend on both the initial and final states (e.g., see Table S4.1-S4.3). Analytical expressions on similar MFPTs involved in amyloid fibril templating can be found a recent theoretical analysis by Schmit[287]. All molecular visualizations were prepared using VMD[219].

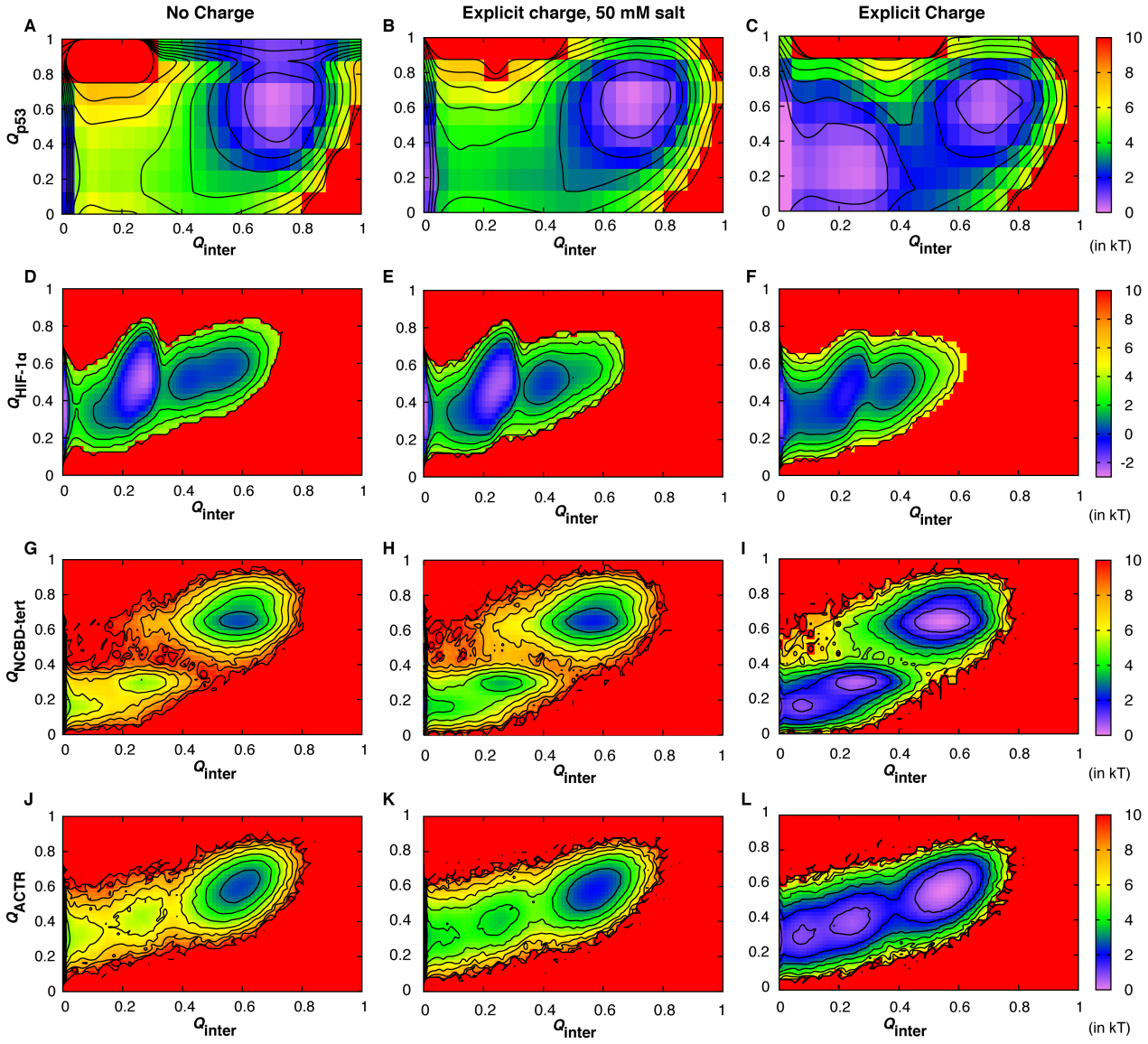


Figure 4.2: Free-energy surfaces at T_m as a function of the fractions of intra- and intermolecular contacts formed, computed using various Gō-like models with and without explicit charges and/or 50 mM salt (see Table 4.2). Rows A-C, D-F and G-L are for the p53-TAD1/TAZ2, HIF-1α/TAZ1 and NCBD/ACTR complexes, respectively. Q_{inter} is the fraction of intermolecular contacts formed; Q_{p53}, $Q_{HIF-1\alpha}$ and Q_{ACTR} are the fractions of intramolecular contacts formed by p53-TAD1, HIF-1α and ACTR, respectively; $Q_{NCBD-tert}$ is the fraction of tertiary intramolecular contacts formed by NCBD (the helical content of NCBD remain similar during coupled binding and folding). Contours are drawn every kT, where k is Boltzmann constant and T is the absolute temperature. Figure (A)-(F) are prepared by DG.

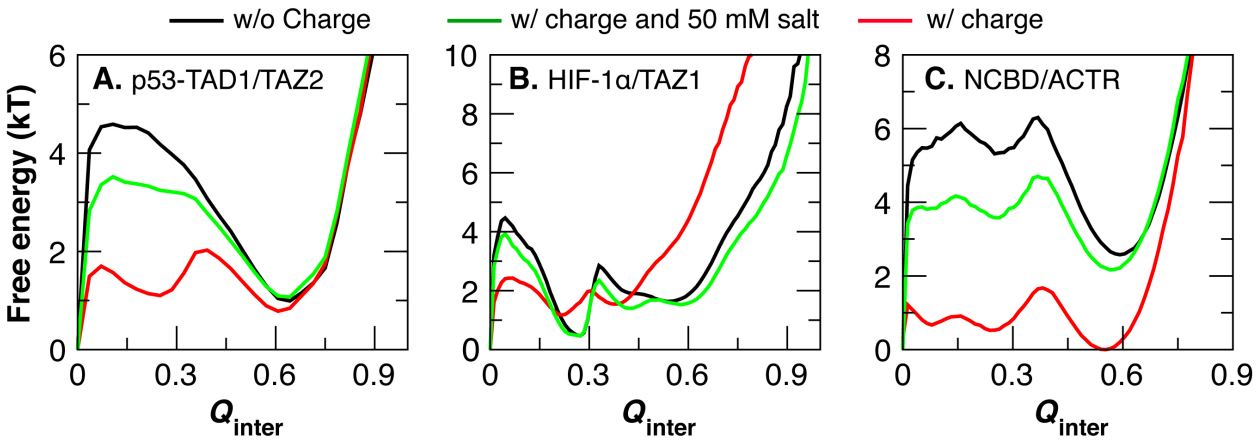


Figure 4.3: Free energy as a function intermolecular contact fraction at T_m, calculated from the REX simulations using WHAM, for (A) TAD1/TAZ2, (B) HIF-1α/TAZ1 and (C) NCBD/ACTR. Figure (A) and (B) are prepared by DG.

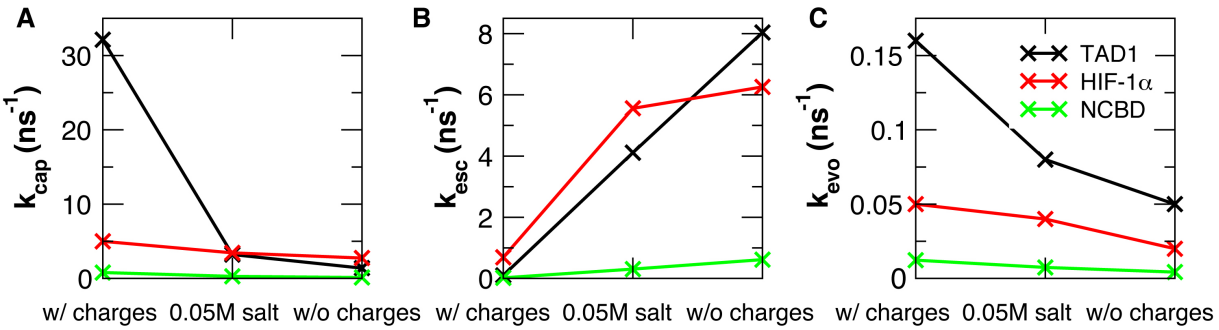


Figure 4.4: Effective rate constants for transitions between the unbound, collision complex and bound states, calculated using models with and without explicit charges and/or 50mM salt, for (A) TAD1/TAZ2, (B) HIF-1α/TAZ1 and (C) NCBD/ACTR. These rates were defined in Methods, Equation 4.1-4.4. The results demonstrate that long-range electrostatic forces increase both the capture and evolution rates and at the same time reduce the escape rates. Figure (A) and (B) are prepared by DG.

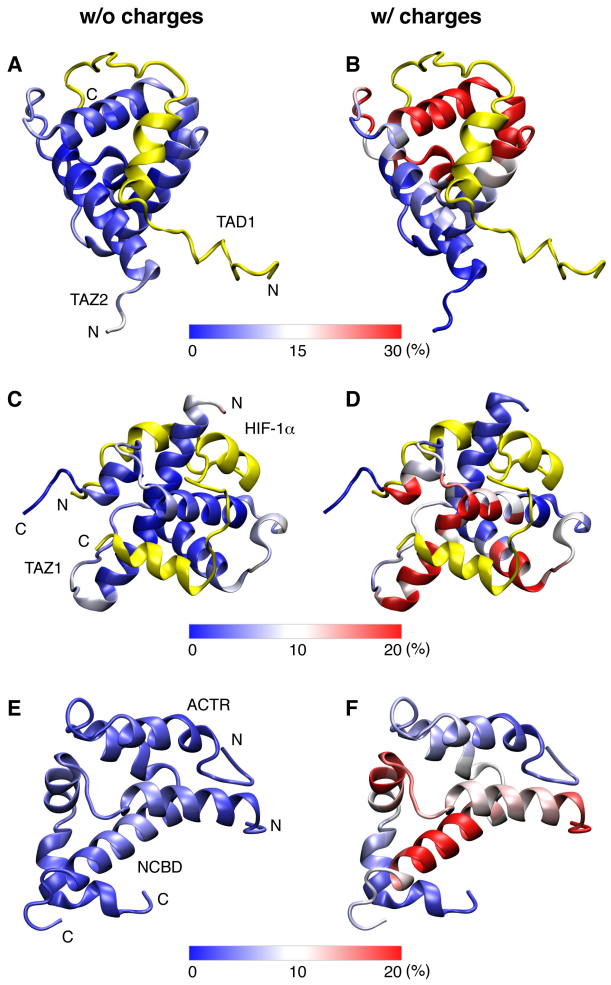


Figure 4.5: Distributions of IDPs on the substrate surfaces in the collision complexes derived from simulations using models with and without explicit charges. For the p53-TAD1/TAZ2 (A)-(B) and HIF-1α/TAZ1 (C)-(D) complexes, TAZ2 and TAZ1 are colored based on the probability of each residue in contact with the IDPs in the collision complex ensembles, and p53-TAD1 and HIF-1α are shown only in the fold and bound conformations (yellow cartoon) for reference. For the NCBD/ACTR complex (E)-(F), both IDPs are shown in the bound and folded conformations and colored based on the probability of each residue involved (nonspecific) intermolecular contacts in the collision complex ensemble. Figure (A)-(D) are prepared by DG.

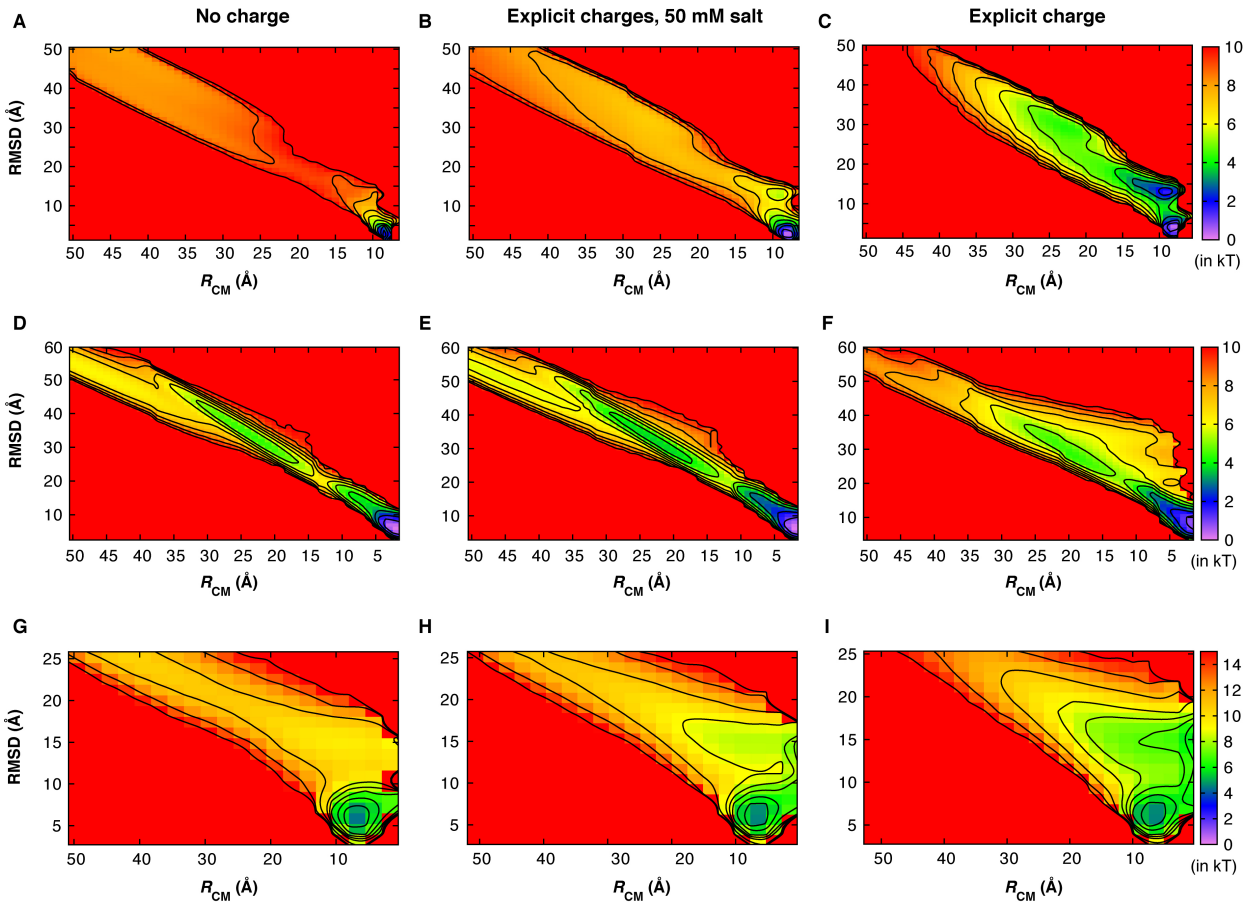


Figure 4.6: Free-energy surfaces at T_m as a function of binding RMSD of the IDP and center of mass separation between two peptides (R_{CM}), computed using various Gō-like models with and without explicit charges and/or 50 mM salt (see Table 4.2). The binding RMSD (of the IDP) was calculated by first aligning the snapshot with respect to the folded structure using only the folded substrate. For NCBD/ACTR, both proteins are IDPs and the (regular) RMSD was calculated using the whole complex. Rows (A)-(C), (D)-(F) and (G)-(I) are for the p53-TAD1/TAZ2, HIF-1α/TAZ1 and NCBD/ACTR complexes, respectively. Contours are drawn every kT. Figure (A)-(F) are prepared by DG.

4.5 Supplemental Materials

4.5.1 Sequences of IDPs

Charged residues are highlighted in red/blue fonts, and the folded segments in the complexes are underlined. Sequence plots of p53-TAD1 and HIF-1α are prepared by DG.

p53-TAD1:

MET GLU GLU PRO GLN SER ASP PRO SER VAL GLU PRO PRO LEU SER GLN GLU THR PHE SER ASP LEU TRP LYS LEU LEU PRO GLU ASN ASN VAL LEU SER PRO LEU PRO SER GLN ALA

HIF-1α:

SER ASP LEU ALA CYS ARG LEU LEU GLY GLN SER MET ASP GLU SER GLY LEU PRO GLN LEU THR SER TYR ASP CYS GLU VAL ASN ALA PRO ILE GLN GLY SER ARG ASN LEU LEU GLN GLY GLU GLU LEU LEU ARG ALA LEU ASP GLN VAL ASN

NCBD:

PRO ASN ARG SER ILE SER PRO SER ALA LEU GLN ASP LEU LEU ARG THR LEU LYS SER PRO SER SER PRO GLN GLN GLN GLN GLN VAL LEU ASN ILE LEU LYS SER ASN PRO GLN LEU MET ALA ALA PHE ILE LYS GLN ARG THR ALA LYS TYR VAL ALA ASN GLN PRO GLY MET GLN

ACTR:

GLU GLY GLN SER ASP GLU ARG ALA LEU LEU ASP GLN LEU HIS THR LEU LEU SER ASN THR ASP ALA THR GLY LEU GLU GLU ILE ASP ARG ALA LEU GLY ILE PRO GLU LEU VAL ASN GLN GLY GLN ALA LEU GLU PRO LYS

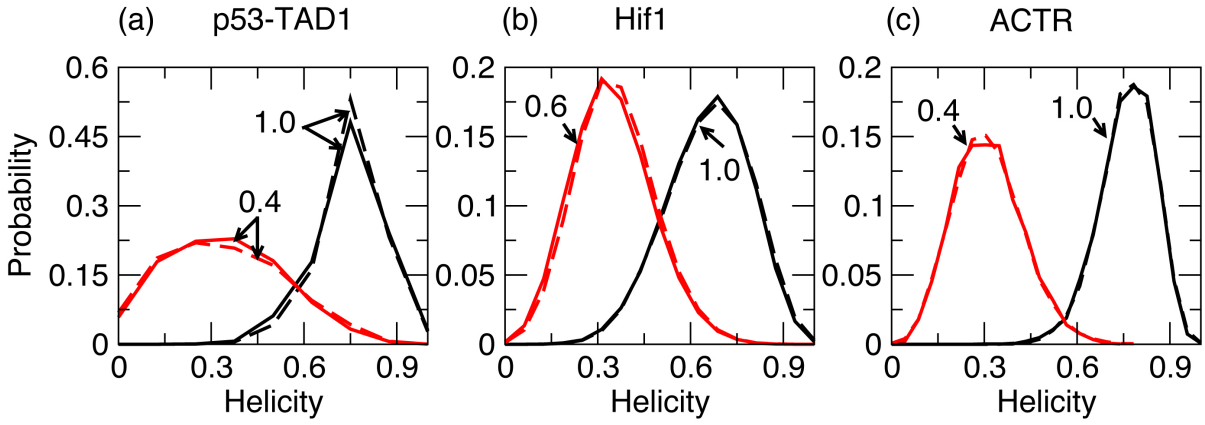


Figure S4.1: Residual helicities of (a) p53-TAD1, (b) HIf-1α, and (c) ACTR in the unbound states calculated using different Gō-like models. The solid traces correspond to models without explicit charges and the dashed traces are from the charged models. The black traces were computed from models with no adjustment of the intramolecular interaction strengths (i.e., scale = 1.0), which significantly over-stabilized the helices. The red traces were calculated using the final calibrated models with optimal scaling of intramolecular interactions (see Table 4.2 of the main text). The residual helicity showed minimal dependence on the salt concentration for all peptides and the corresponding profiles are thus not shown. Free NCBD has helical content similar to the bound state, and no further calibration to the original sequence-flavored Gō-like model appears necessary (see Chapter 3 for additional details). The helicity distribution of NCBD is thus not shown. Figure (a) and (b) are prepared by DG.

Table S4.1: Mean-first-passage-times (MFPTs) and numbers of transitions (in parenthesis) between conformational sub-states of the p53-TAD1/TAZ2 complex computed from the production Langevin simulations. The state assignment criteria are provided in the main text, and the transitions shown are from the one in the row to that in the column. All MFPTs are in ns. Table is prepared by DG.

w/o charge **w/charge+0.05M salt** **w/ charge**	**U**	**CC**	**B**
U		0.72 (10026) 0.31 (13735) 0.03 (1098)	
CC	0.12 (10026) 0.24 (13735) 3.47 (1098)		0.26 (60) 0.43 (256) 3.94 (1694)
B		109.02 (60) 28.86 (256) 2.63 (1695)	

Table S4.2: Mean-first-passage-times (MFPTs) and numbers of transitions (in parenthesis) between conformational sub-states of the HIF-1α/TAZ1 complex computed from the production Langevin simulations. The state assignment criteria are provided in the main text, and the transitions shown are from the one in the row to that in the column. All MFPTs are in ns. Table is prepared by DG.

w/o charge **w/charge+0.05M salt** **w/ charge**	**U**	**CC**	**B**
U		0.37 (19428) 0.29 (19091) 0.2 (5932)	
CC	0.14 (19428) 0.16 (19091) 1.20 (5932)		14.33 (67) 3.26 (133) 3.65 (416)
B		137.41 (68) 45.14 (133) 12.45 (417)	

Table S4.3: Mean-first-passage-times (MFPTs) and numbers of transitions (in parenthesis) between conformational sub-states of the NCBD/ACTR complex computed from the production Langevin simulations. The state assignment criteria are provided in the main text, and the transitions shown are from the one in the row to that in the column. All MFPTs are in ns.

w/o charge **w/charge+0.05M salt** **w/ charge**	**U**	**CC**	**B**
U		7.71 (2282) 3.50 (2161) 1.26 (181)	11.06 (5) 27.20 (2) 22.60 (1)
CC	1.57 (2287) 2.95 (2160) 22.10 (180)		8.14 (16) 13.13 (52) 44.56 (108)
B	1420.5 (1) 368.77 (3) 165.00 (2)	818.89 (21) 274.24 (51) 190.80 (108)	

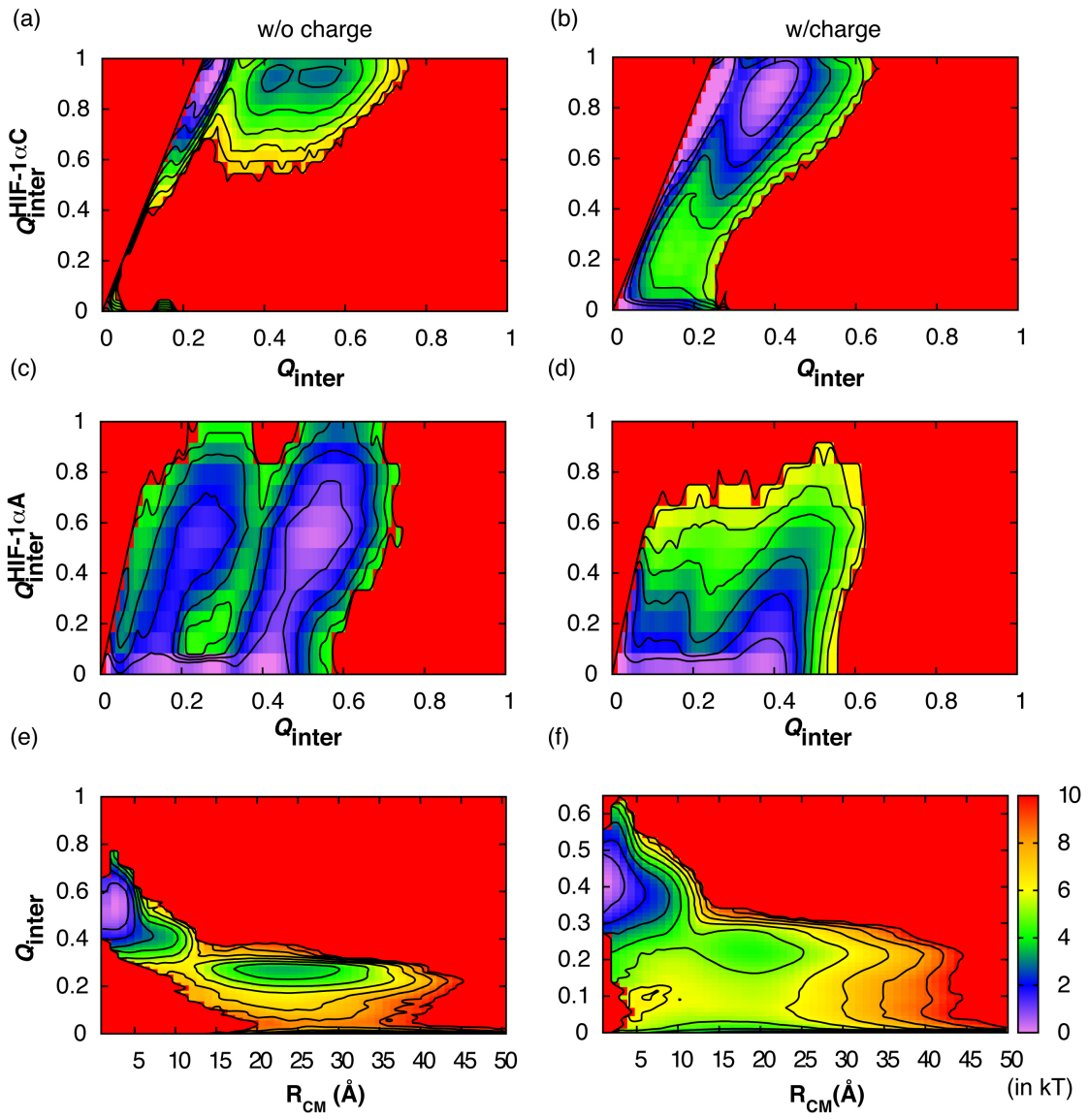


Figure S4.2: 2D free energy surfaces at T_m calculated using models with and without explicit charges (see Table 4.2 of the main text). $Q_{inter}^{HIF-1\alpha A}$ and $Q_{inter}^{HIF-1\alpha C}$ are the fractions of native intermolecular contacts formed by the first and third helices of HIF-1α, respectively. R_{CM} is the distance between the centers of mass of HIF-1α and TAZ1. Contours are drawn every kT. Figures are prepared by DG.

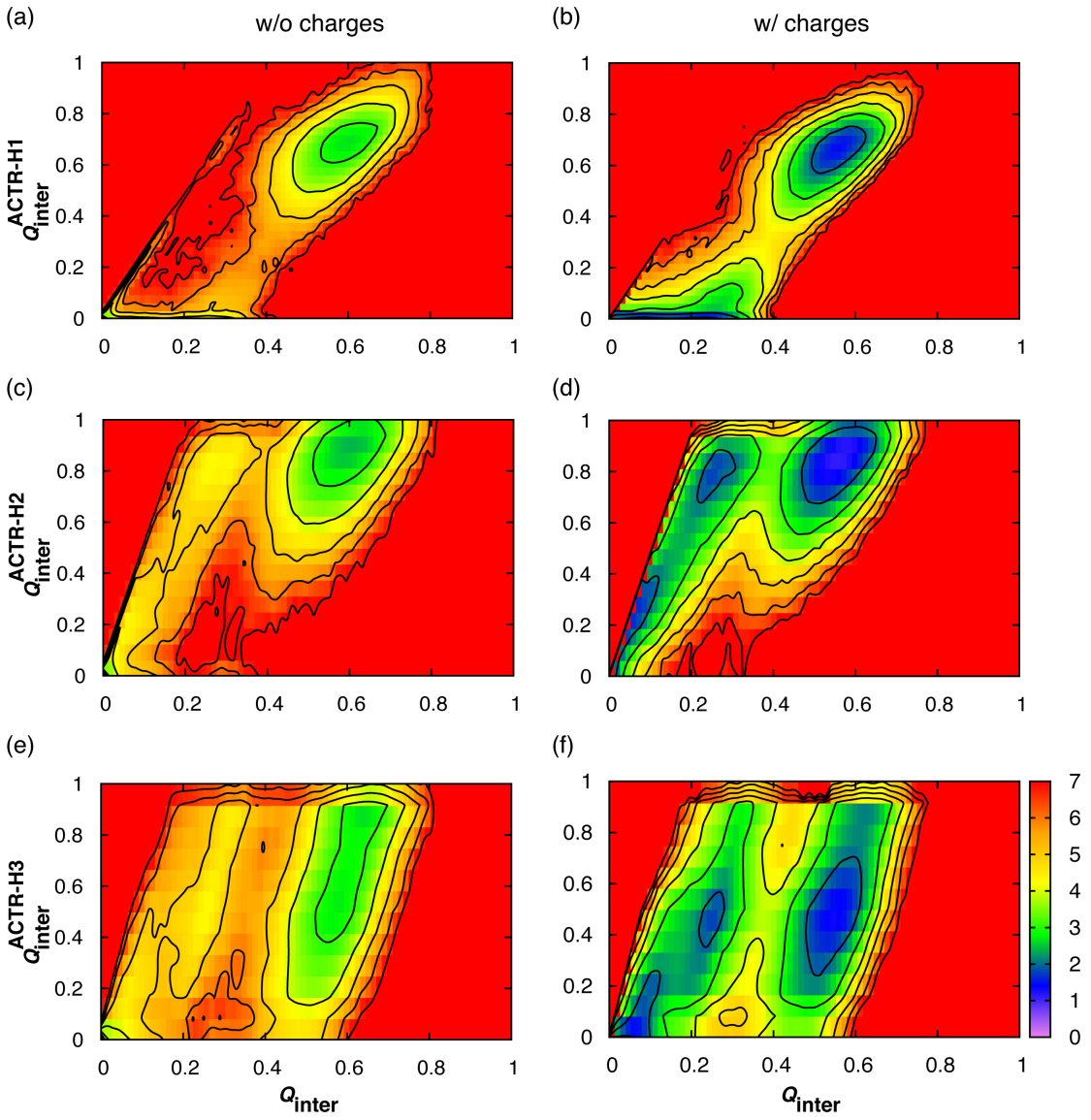


Figure S4.3: 2D free energy surfaces at T_m calculated using models with and without explicit charges (see Table 4.2 of the main text). $Q_{inter}^{ACTR-H1}$, $Q_{inter}^{ACTR-H2}$ and $Q_{inter}^{ACTR-H3}$ are the fractions of native intermolecular contacts formed by the first, second and third helices of ACTR, respectively. Contours are drawn every kT.

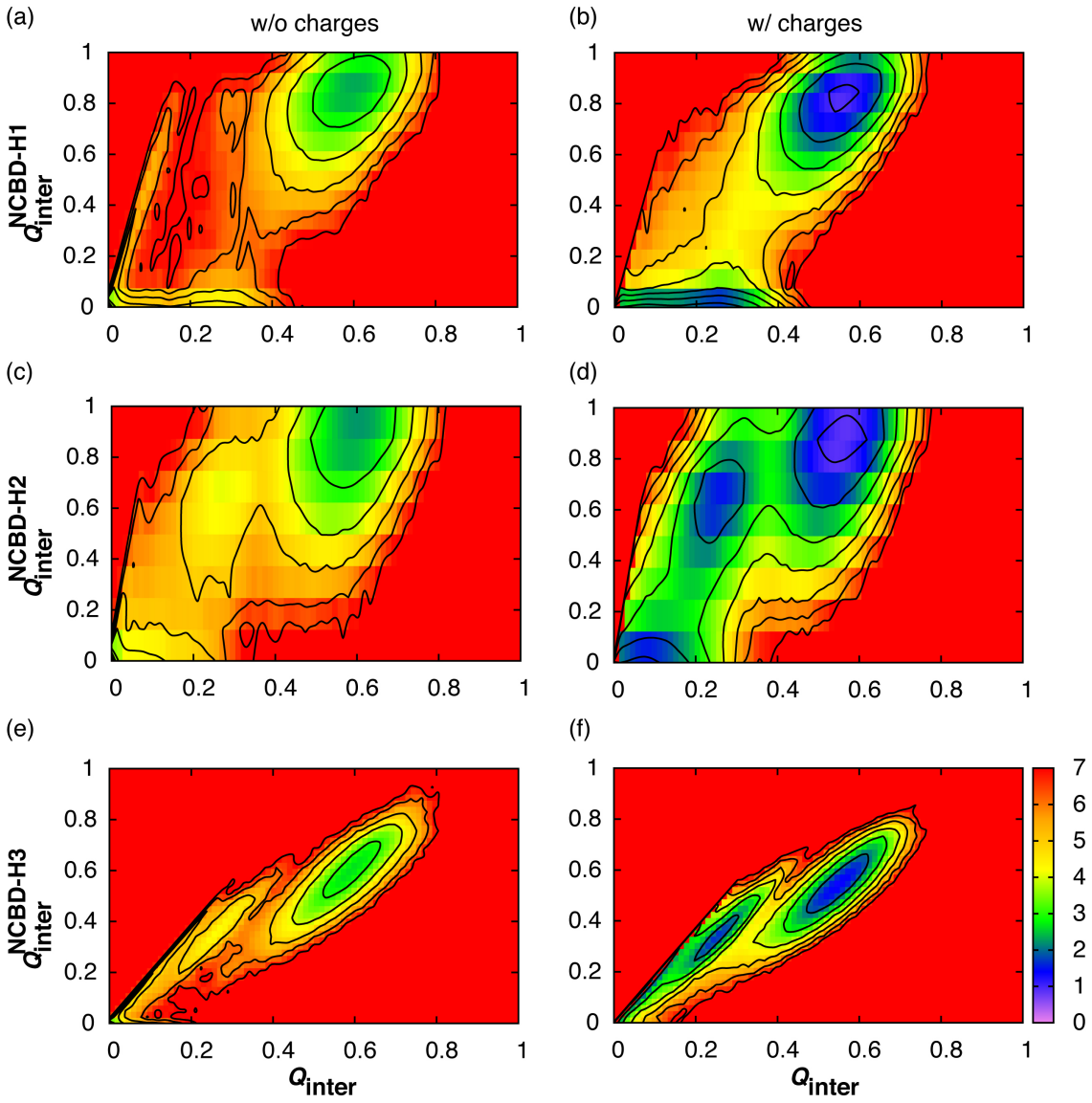


Figure S4.4: 2D free energy surfaces at T_m calculated using models with and without explicit charges (see Table 4.2 of the main text). $Q_{inter}^{NCBD-H1}$, $Q_{inter}^{NCBD-H2}$ and $Q_{inter}^{NCBD-H3}$ are the fractions of native intermolecular contacts formed by the first, second and third helices of NCBD, respectively. Contours are drawn every kT.

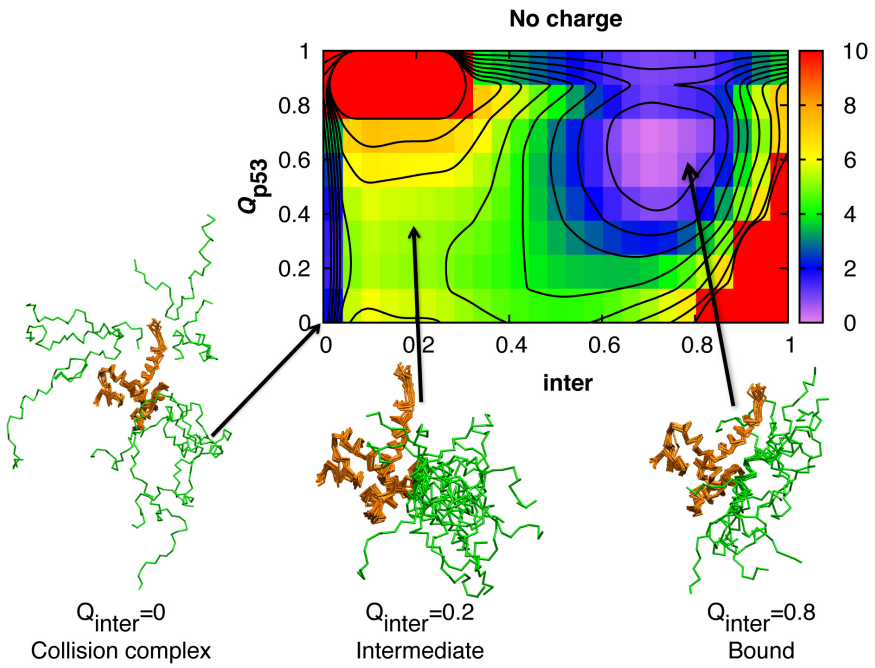


Figure S4.5: Representative snapshots along the binding and folding pathways of p53-TAD1/TAZ2. Figure is prepared by DG.

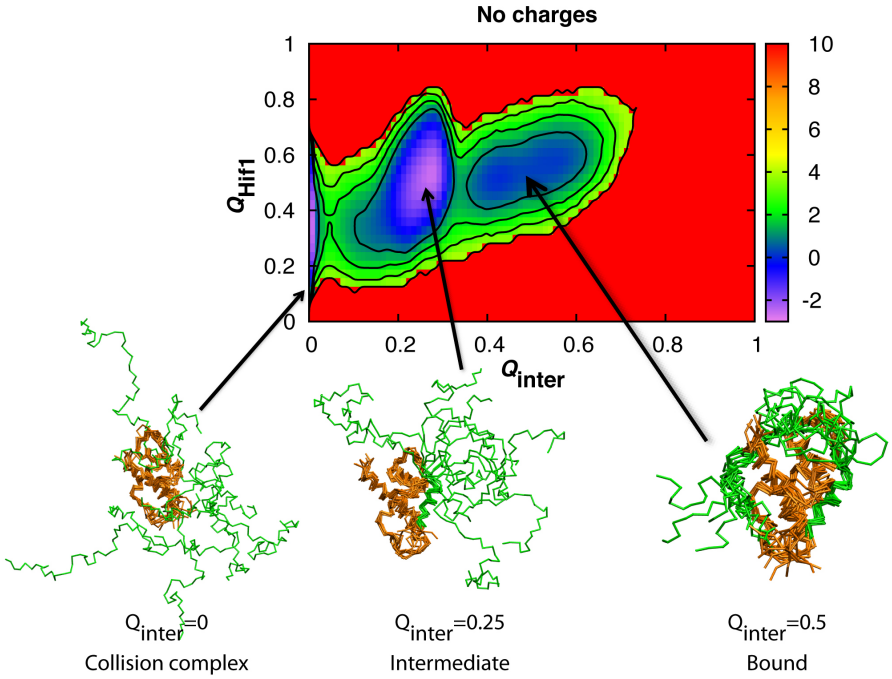


Figure S4.6: Representative snapshots along the binding and folding pathways for HIF-1α/TAZ1. Figure is prepared by DG.

Chapter 5

Efficiency of Adaptive Temperature-based Replica Exchange for Sampling Large-scale Protein Conformational Transitions*

Abstract

Temperature-based replica exchange (RE) is now considered a principal technique for enhanced sampling of protein conformations. It is also recognized that existence of sharp cooperative transitions (such as protein folding/unfolding) can lead to temperature exchange bottlenecks and significantly reduce the sampling efficiency. Here, we revisit two adaptive temperature-based RE protocols, namely, exchange equalization (EE) and current maximization (CM), that were previously examined using atomistic simulations (Lee and Olson, J. Chem. Physics, 134, 24111 (2011)). Both protocols aim to overcome exchange bottlenecks by adaptively adjusting the simulation temperatures, either to achieve uniform exchange rates (in EE) or to maximize temperature diffusion (CM). By designing a realistic yet computationally tractable coarse-grained protein model, one can sample many reversible folding/unfolding transitions using conventional constant temperature molecular dynamics (MD), standard REMD, EE-REMD, and CM-REMD. This allows rigorous evaluation of the sampling efficiency, by directly comparing the rates of folding/unfolding transitions and convergence of various thermodynamic properties of interest. The results demonstrate that both EE and CM can indeed enhance temperature diffusion com-

*Reprinted (adapted) with permission from W. Zhang and J. Chen (2013), "Efficiency of adaptive temperature-based replica exchange for sampling large-scale protein conformational transitions", J. Chem. Theory Comput. 9, 2849-2856. Copyright (2013), American Chemical Society.
Author Contributions: Conceived and designed the experiments: JC WZ. Performed the experiments: WZ JC. Analyzed the data: WZ JC. Contributed reagents/materials/analysis tools: WZ JC. Wrote the paper: WZ JC.

pared to standard RE, by $\sim$3- and over 10-fold, respectively. Surprisingly, the rates of reversible folding/unfolding transitions are similar in all three RE protocols. The convergence rates of several key thermodynamic properties, including the folding stability and various 1D and 2D free energy surfaces, are also similar. Therefore, the efficiency of RE protocols does not appear to be limited by temperature diffusion, but by the inherent rates of spontaneous large-scale conformational re-arrangements. This is particularly true considering that virtually all RE simulations of proteins in practice involve exchange attempt frequencies ($\sim$ps^{-1}) that are several orders of magnitude faster than the slowest protein motions ($\sim\mu$s^{-1}). Our results also suggest that the efficiency of RE will not likely be improved by other protocols that aim to accelerate exchange or temperature diffusion. Instead, protocols with some types of guided tempering will likely be necessary to drive faster large-scale conformational transitions.

5.1 Introduction

Successful computer simulations of protein conformational equilibrium and transitions not only require accurate description of protein energetics in complex heterogeneous environments, but also require sufficient sampling of the relevant conformational space. At present, generating atomistic structure ensembles that are statistically representative of the accessible conformations of a protein under a given set of thermodynamic conditions remains a challenging problem[100]. The difficulty arises not only because of the large and complex conformational space of biomolecules, but also due to significant energy barriers that might separate different conformational subspaces. The efficiency of conventional Monte Carlo (MC) or molecular dynamics (MD) is limited due to frequent trapping of protein in numerous local energy minima. The replica exchange (RE) method[103–105,178], also known as parallel tempering, has emerged as a relatively straightforward but powerful approach for enhanced conformational sampling. The basic idea is to simulate multiple replicas of the system at different temperatures independently using either MC or MD. Periodically, replicas attempt to exchange simulation temperatures according to a Metropolis criterion that preserves the detailed balance and ensures canonical distributions at all temperatures. The resulting random walk in the temperature space helps the system to escape states of local energy minima and thus facilitate conformational sampling. Replica exchange molecular dynamics (REMD) in particular has been widely applied to and shown to be successful in protein simulations[106–110]. Nonetheless, questions remain regarding the true efficiency of RE in sampling large-scale (protein) conformational transitions and its dependence on the properties of the system and key RE parameters, such as the number of replicas, exchange attempt frequency, and choice of simulation temperatures, etc.

The key parameters of REMD of protein simulations has been a subject of substantial research interest in recent years, examined both based on theoretical considerations[111–116] and through actual simulations of small peptides[110,117–119]. These studies generally confirm that RE can en-

hance the conformational sampling as long as the activation enthalpies are positive. In particular, recent theoretical analysis[112] and kinetic network models[114,115] of RE for two-state systems have emphasized the importance of choosing a maximum temperature slightly above the temperature where the folding rate is maximal because of the anti-Arrhenius behavior of protein folding at high temperatures. Once the maximum temperature is chosen, other REMD parameters can be set to achieve effective diffusion of replicas across the temperature ladder, arguably with the smallest possible number of replicas. Many strategies have been previously proposed for the later purpose, such as more frequent exchange attempts[119], global energy reassignment[78], and non-equilibrium switches[123]. More aggressive approaches attempt to reduce to number of replicas required for effective temperature diffusion by directly reducing the number of particles that participate in temperature exchanges[124–126], but sometimes with undesirable consequences[127].

It is typically assumed that REMD will provide the highest sampling efficiency if all the replicas spend equal amount of time at each temperature. As proposed by Sugita and Yokomoto[104] and Kofke[129], this may be achieved through optimizing the allocation of simulation temperatures to provide efficient and uniform exchanges. Nonetheless, it is not clear whether consideration of exchange efficiency alone is sufficient to optimize the sampling efficiency, which is more directly measured by the frequency of conformational transitions, and ultimately by the convergence of various thermodynamic properties of interest. For example, REMD simulations were previously performed in the GBSW implicit solvent to examine the conformational equilibrium of a β-hairpin derived from protein GB1 domain (GB1p; residues 41-56)[77]. The exchange (acceptance) rates of both folding and control REMD simulations were very uniform and $>60\%$ for all pairs (Figure 5.1A). Nonetheless, there was an apparent segregation of replicas in the temperature space: a few of replicas dominated the lowest temperature and had much lower averaged temperatures (e.g., replicas 2, 6 and 10; see Figure 5.1B). Closer examination suggests that the observed temperature trapping reflects a lack of reversible folding and unfolding transitions during these simulations. The few replicas that happened to fold first during the folding run (or those did not unfold at first during the control run) remained folded throughout the simulation timespan and thus had much

higher tendency to visit the lower temperature windows (e.g. replicas 2 and 10; the black and red traces in Figure 5.1C). Even though the folded replicas did visit higher temperatures frequently, they did not spend sufficient time to unfold spontaneously and thus quickly reversed back to lower temperatures. Vice versa was true for the unfolded replicas. As a result, there was rapid mixing of the replicas due to uniform and high exchange rates, but all replicas remained trapped in various free energy basins. Similar observations have also been made by Periole and Mark[117].

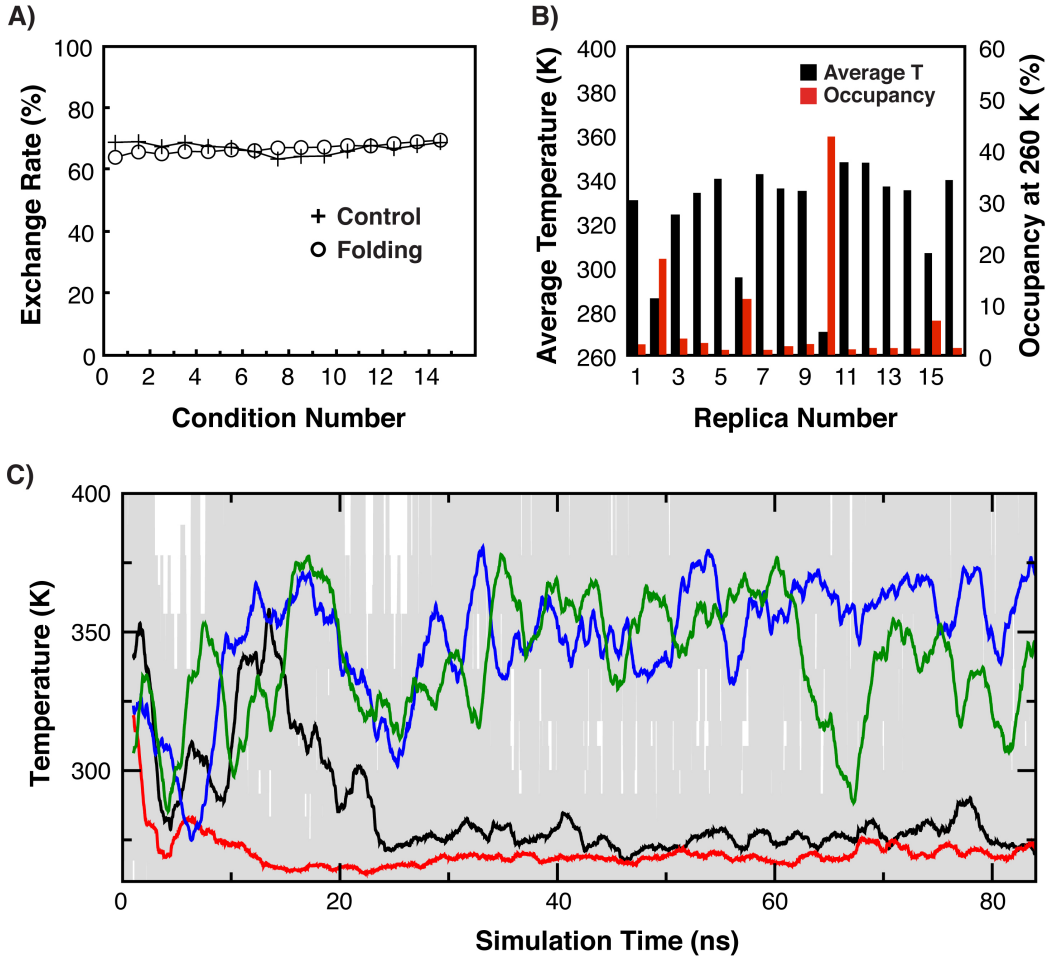


Figure 5.1: REMD simulations of GB1p β-hairpin. (A) Exchange acceptance rates between neighboring temperatures for the control (initiated from the native structure) and folding (from a fully extended structure) REMD simulations. Condition numbers correspond to different simulation temperatures. (B) Average temperatures and percentages of occupancy at the lowest temperature (260 K) of all replicas in the folding REMD simulation. (C) 50-ps running averages of the temperature history of four representative replicas (2: black, 10: red, 12: blue, 16: green) during the folding REMD simulation. The grey traces in the background are the raw time traces. Details of these simulations were described in Reference[77].

In light of the inadequacy if uniform exchange to promote folding/unfolding transitions, the recent proposed current maximization (CM) protocol[130] is very attractive. CM aims to systematically optimize temperature diffusion such that the number of "round trips" in the temperature space of each replica is maximized. Arguably the number of "round trips" is more directly related to the ability of RE to drive conformational transitions, and thus CM could substantially improve the sampling efficiency[130]. Lee and Olson recently compared the sampling efficiency of CM and another adaptive temperature-based REMD protocol, namely, exchange equalization (EE)[128], using the 57-residue SH3 domain of α-spectrin in implicit solvent as a model system. However, due to the system size and associated computational cost, no reversible folding/unfolding transitions were sampled. As a result, how effectively the improved temperature diffusion in CM can translate into sampling efficiency remains to be established.

Identification of appropriate protein model systems for rigorous benchmarking of sampling efficiency has actually been difficult. Due to the computational cost constraint, small peptides have been mainly used as model systems[110,117,119]. REMD simulations of the 21-residue Fs-21 peptide in implicit solvent demonstrated that REMD could enhance the sampling efficiency by 14 to 72 times at different temperatures compare to conventional MD based on examination of helicity auto-correlation functions[110]. Periole and Mark simulated a β-heptapeptide in explicit solvent to examine the convergence and sampling efficiency, which suggested REMD to be at least one order of magnitude more efficient than MD[117]. While these studies have provided important insights into the efficacy of REMD, small model peptides lack long-range ordering that exist in actual globular proteins, and, more importantly, do not fully reflect globular proteins where folding/unfolding transitions often occur at much longer timescales compare to local conformational fluctuations. Our previous simulations of small proteins[1,35,74] (also see Figure 5.1) have suggested that the ability to drive slow global conformational transitions is a key bottleneck in REMD sampling. Therefore, there is a need to develop realistic yet computationally tractable protein models for more rigorous benchmarking of REMD sampling efficiency.

In this work, we first construct a protein model derived from coarse-grained topology-based models[98], which are based on the principle of minimal frustration for evolved proteins and allow direct simulations of folding and unfolding transitions to characterize folding mechanisms. In particular, impressive correspondence has been demonstrated between experiment and theory for many proteins[98,229], supporting the notion that topology-based models capture the essence of folding behaviors of globular proteins. Specifically, the original sequence-flavored Gō-like model[232] was modified to include non-specific hydrophobic interactions to mimic the presence of non-native interactions in real proteins. Inclusion of non-specific interactions also increases the complexity of the energy landscape, and thus the sampling challenge. Such a model arguably provides a reasonable balance between the level of realisticity and computational tractability for more rigorous benchmarking of how the above-discussed adaptive temperature-based REMD protocols enhance sampling of protein conformations.

5.2 Methods

5.2.1 Adaptive temperature-based REMD protocols

A typical REMD involves multiple replicas simulated simultaneously at different temperatures. For given choices of the number of replicas N, the minimum temperature, T_{min}, and the maximum temperature, T_{max}, the temperature distribution is usually assigned as:

$$T_1 = T_{min}, T_{i+1} = T_i(\frac{T_{max}}{T_{min}})^{[\frac{1}{N-1}]}. \tag{5.1}$$

Exchange of simulation temperatures are attempted periodically and accepted according to a Metropolis criterion that maintains the detailed balance[104].

$$p(i \leftrightarrow j) = min[1, e^{(\beta_j - \beta_i)(E_j - E_i)}] \tag{5.2}$$

where $\beta_i = 1/k_B T_i$, k_B is Boltzmann's constant, and T_i and E_i are the temperature and potential energy of replica i. After a successful exchange of simulation temperatures, the velocities of swapped replicas are rescaled according to,

$$P(T_j) = \sqrt{\frac{T_j}{T_i}} P(T_i), P(T_i) = \sqrt{\frac{T_i}{T_j}} P(T_j). \tag{5.3}$$

Exchange rate equalization (EE)

For systems displaying strong phase transitions such as protein folding and unfolding, the standard REMD can suffer from depressed exchange rate near the transition temperature, T_m. One approach to resolve this limitation is to dynamically adapt the temperature spacing and populate more simulation windows near T_m. This may be achieved by equalizing the exchange rates through out the REMD temperature range. Briefly, as described by Lee and Olson[288], the simulation temperatures, $\{T_i\}$, can be updated periodically during REMD to maximize the sum of exchange acceptance rates, $S_{i \to i+1}$, raised to a properly chosen negative power, r,

$$\Omega = max[\sum_{i=1}^{N-1} (s_{i \to i+1})^r] \tag{5.4}$$

In the limit of $r \to \infty$, the exchange rates would be uniform with maximized Ω. The value of r was assigned to be -4 in the current work as suggested by Lee and Olson[288]. Assuming that the energy distributions at all temperatures are Gaussian, the exchange rates can be estimated by integrating the acceptance probability over the energy distributions[128]. This allows numerical

optimization of $\{T_i\}$ according to Equation 5.4 to achieve approximate equalization of exchange rates between all neighboring replica pairs.

Current maximization (CM)

The CM method aims to maximize the number of round trips that the replicas travel between the lowest and highest simulation temperatures[130]. For this, a replica is marked as cold or hot depending on its last visit to either temperature extremes. The cold fraction can be then calculated for each temperature as

$$f(T) = \frac{n_{cold}(T)}{n_{cold}(T) + n_{hot}(T)} \tag{5.5}$$

where $n_{cold}(T)$ and $n_{hot}(T)$ are the number of cold and hot replicas at temperature T. It has been shown that the current of temperature diffusion can be maximized by adjusting the temperatures, T_i, such that $f(T)$ increases linearly as a function of the temperature index, i[289]. This can be achieved by first interpolating a continuous $f(T)$ from the computed values of f at the current set of temperatures and then using it to search for new temperatures within the pre-specified temperature ranges where $f(T) = i/(N-1)$. To avoid crowding of replicas near T_m, a constraint is introduced such that no neighboring temperatures can be more than two geometric spacing units apart[128]

$$\frac{T_{i+1}}{T_i} \leq (\frac{T_{max}}{T_{min}})^{\frac{2}{N-1}} \tag{5.6}$$

5.2.2 A topology-based coarse-grained protein model with rugged energy landscape

The sequence-flavored Gō-like model[232] is an advanced topology-based model that exploits the idea that sequence can provide differing statistical weights to alternative folding pathways. By including knowledge-based pseudo-torsional potentials and using the Miyazawa-Jernigan (MJ) statistical potentials[234] for residue-specific Cα-based native interactions, the model can recapitu-

late subtle differences in folding mechanisms that arise from sequence differences in topologically analogous proteins[235,236]. In this work, we further extend this model to include nonspecific hydrophobic interactions. Specifically, an initial model was first generated for the B1 domain of streptococcal protein G (GB1) (PDB: 3gb1)[290] (see Figure 5.2) using the Multiscale Modeling Tools for Structural Biology (MMTSB) Gō-Model Builder (http://www.mmtsb.org)[215]. We then recalibrated the model to have $T_m \sim$ 350K by uniformly adjusting the strength of all native contacts. Non-specific van der Waals (vdW) interactions were then introduced between all Cα beads with strengths similar to those of the native contacts. Specifically, the well depth was empirically set to ε = -1.1 kcal/mol, which appeared to provide a reasonable compromise between retaining fast folding rates and increasing energy landscape complexity based on pilot REMD simulations. The pilot simulations also suggested that non-specific 1-4 vdW interactions as well as those in two short loops (see Figure 5.2) needed to be turned off to avoid steric restriction of rotation around Cα-Cα bonds and facilitate chain diffusion. The final model retains a sharp folding transition (Figure 5.3A), but displays more complex energy landscape with multiple intermediate states (Figure 5.3B; also see Section 3.4). Compared to the original sequence-flavored Gō-like model, the folding and unfolding rates are about 15 times slower at corresponding T_m's.

5.2.3 Simulation protocols and analysis

All the REMD simulations were performed using the MMTSB Toolset[215] with 8 replicas spanning 300 to 400K. For each replica, Langevin dynamics simulation was performed using CHARMM[63,216] with a dynamic time step of 10 fs and a friction coefficient of 0.1 ps^{-1}. The SHAKE algorithm[217] was applied to constrain all virtual bond lengths. Replica exchanges were attempted between neighboring replicas every 5000 MD steps (50 ps). The geometrically spaced distribution (Equation 5.1) was used as the initial temperature profile. Pilot runs were first performed to obtain the optimized temperature distributions according to either EE or CM protocol. Specifically, multiple cycles of 1-μs REMD were performed, each followed by optimization of the temperature

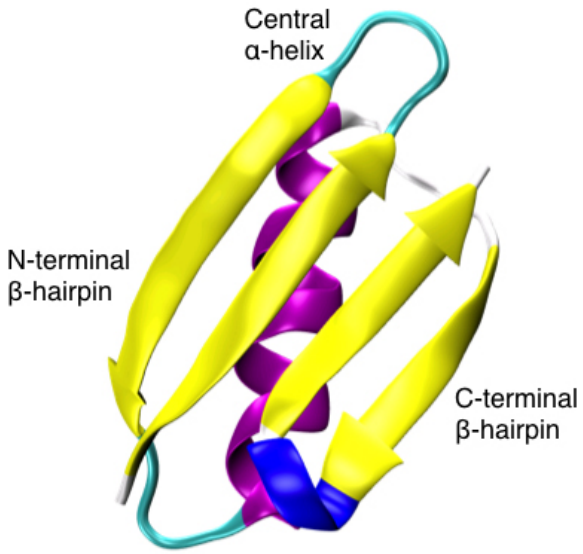


Figure 5.2: The structure of B1 domain of streptococcal protein G (PDB ID: 3gb1)[290]. The center helix is colored in violet, β-strands in yellow, and loops in cyan.

distributions using the EE and CM methods until the whole profile stabilized. Additional 5-μs REMD simulations were then performed verify that the EE and CM temperature distributions were stable (for the given protein model). Finally, the converged EE and CM temperature profiles were used for independent production REMD simulations, initiated from the native structure (control run) and a fully extended structure (folding run), respectively. The length of all REMD simulations was 100 μs. The parallel control and folding runs allow additional diagnosis of convergence. For control, standard REMD and constant-temperature MD simulations (at 300 K and $T_m \approx$ 354K) were also performed (initiated from the native structure). These simulations are summarized in Table 5.1. We note that T_m calculated from the production simulations are slightly different from the value estimated using the pilot run ($\sim$354K).

The sampling efficiencies of various simulation protocols are judged mainly based on: the number of folding/unfolding transitions N_{TS}, average time of conformational transitions, τ_{TS}, and convergence of the calculated folding free energy, ΔG, and several 1D and 2D potential of mean forces (PMFs). In the context of topology-based modeling, the fractions of native contacts provide natural reaction coordinates for describing folding and unfolding transitions[237]. The model for protein

Table 5.1: Summary of production REMD and MD simulations.

Method	Initial State	Length (μs)	T_m(K)	N_{TR}	N_{TS}	τ_{TS} (μs)
EE	Folded	100	351	4423	28	28.6
	Extended	100	352	4003	24	33.3
CM	Folded	100	354	18533	18	44.4
	Extended	100	352	19570	23	34.8
Standard	Folded	100	352	1411	21	38.1
MD @ 354K	Folded	200	-	-	14	14.3
MD @ 300K	Folded	100	-	-	0	Undef

T_m values shown were estimated based on the peak of the final calculated C_V curves. N_{TR} is the total number of round trips that all replicas travel between temperature extremes.

GB1 includes 75 native contacts. Based on the 1D PMF as a function of native contact fraction (see Figure 5.3B), conformations with $<$25 native contacts were assigned to be in the unfolded state, and those with $>$55 native contacts to the folded state. These definitions of the folded and unfolded states were then used to calculate the number of folding and unfolding transitions sampled in MD or by all replicas in REMD simulations (see Table 5.1). We further divided the native contacts into those involving the N-terminal residues 1-20 (Q-nt), the C-terminal residues 42-55 (Q-ct), or the central α-helical residues 23-36 (Q-h), and calculated 2D various PMFs. Achieving convergence on these 2D PMFs should be more challenging than converging on folding stability alone or 1D PMF. Thus, it provides a more stringent test on various REMD protocols' ability to enhance sampling. The weighted histogram analysis method (WHAM) was used to combine information from all temperatures to compute either the C_V curves or various unbiased 1D and 2D probability distributions[142]. All the analysis was performed using a combination of the MMTSB toolset, CHARMM and in-house scripts. The numbers of temperature round trips in REX simulations were calculated between the lowest two temperature and highest two temperatures.

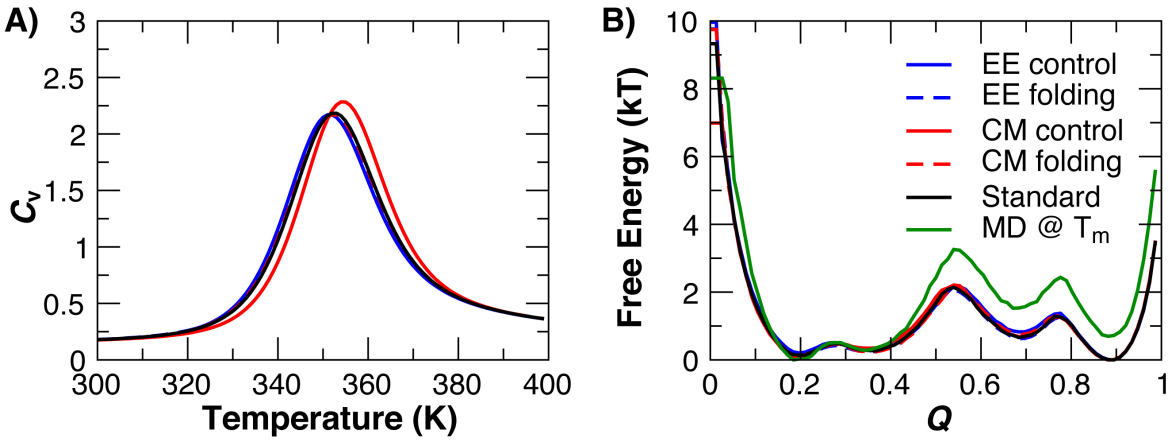


Figure 5.3: (A) The heat capacity, C_V, as a function of temperature and (B) free energy as a function of the native contact fraction (Q) at T_m = 354 K, calculated from various 100-μs EE- and CM-REMD and standard REMD simulations. The PMF derived from a 200-μs constant temperature MD at T_m = 354K is also shown (green trace in panel B).

5.3 Results and Discussion

5.3.1 Convergence of the EE- and CM-optimized temperature distributions

For given temperature range and number of replicas, the EE- or CM-optimized temperature distributions should depend solely on the nature of the system (e.g., the density of state distribution). Indeed, as shown in Figure 5.4, both EE- and CM-optimized temperature distributions quickly deviated from the initial geometrically spaced profile (black traces), but started to stabilize by the fourth 1-μs cycles of iterative REMD runs. In the next couple 1-μs cycles, the temperature profiles varied only slightly between cycles (<5 K), which appears to arise mainly due to the finite length of the pilot REMD cycles. Additional 5-μs pilot REMD runs further validate that the final EE- and CM-optimized temperature profiles are well converged. As previously observed[128], the EE-optimized temperature distribution sets large intervals at lower temperatures and closely places simulation windows at higher temperatures; whereas CM optimization tends to place dense windows near the melting temperature of the protein. The final EE- and CM-optimized tempera-

ture distributions (red traces in Figure 5.4) were used in the 100-μs control and folding production REMD simulations.

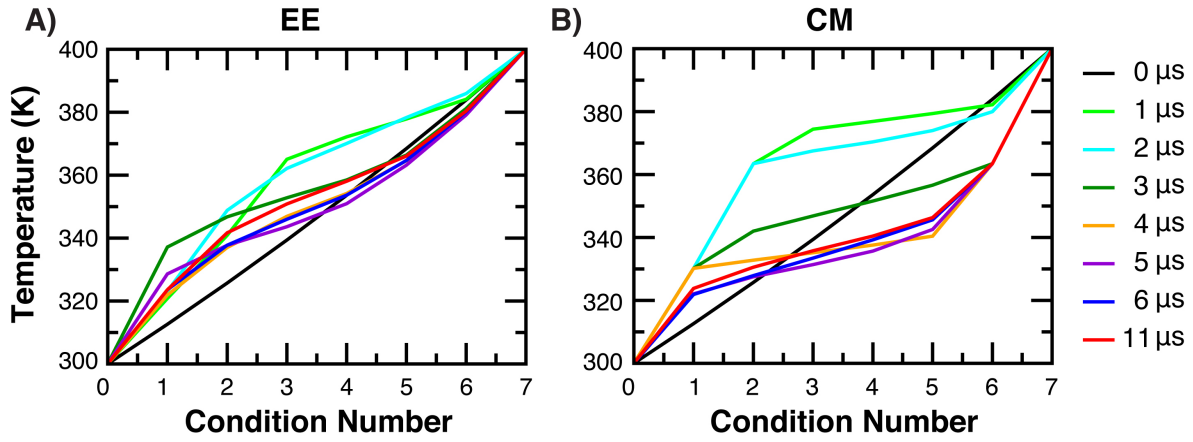


Figure 5.4: Temperature distributions optimized by (A) EE and (B) CM at the end of multiple cycles of 1-μs and 5-μs pilot REMD runs (see Methods).

5.3.2 Exchange acceptance rates and numbers of round trips in the temperature space

As a first step towards examining the efficiency of EE and CM, key REMD metrics, including exchange acceptance rates and numbers of temperature round trips, were calculated to ensure that the adaptive temperature distributions achieved the goals of EE and CM. As shown in Figure 5.5, the exchange acceptance rates are indeed approximately equalized ($\sim$60%) in both control and folding EE-REMD production simulations. Note that the exchange acceptance rates between control and folding simulations are similar, indicating that simulations were well converged. In contrast, the geometrically spaced temperature distribution used in standard REMD leads to an exchange bottleneck near T_m, albeit less pronounced compared to what was observed in atomistic simulations[128]. CM-REMD yielded higher exchange acceptance rates near T_m ($>$80%), which is consistent of denser windows in this region. The exchange acceptance rates in CM are lower at both low and high temperature extremes, which is also similar to the previous study using atomistic model with more replicas[128]. The numbers of round trips in the temperature space are summarized in Table 5.1. The results show that the CM-optimized temperature distribution does enhance the

diffusion in the temperature space, about 13-fold compared to the standard REMD. In comparison, equalizing the exchange rates also enhances temperature diffusion, but only about 3-fold compared to standard REMD. In the following, we further examine whether such equalized exchange rates and enhanced temperature diffusion can effectively translate into more conformational transitions and faster convergence in various thermodynamic properties.

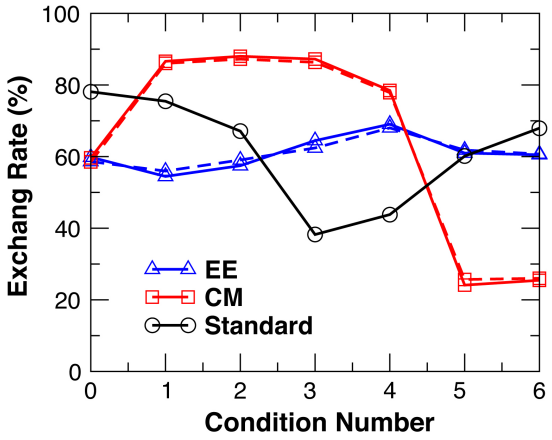


Figure 5.5: Exchange acceptance rates extracted from the 100-μs production REMD simulations. Solid and dash lines correspond to results from the control and folding simulations, respectively.

5.3.3 Number of folding and unfolding transitions

The sampling efficiency of REMD is arguably more directly related to the number of major conformational transitions sampled, such as between folded and unfolded states. As summarized in Table 5.1, 28 and 24 folding-unfolding transitions were observed during the control and folding EE-REMD runs, respectively. Interestingly, despite enhanced temperature diffusion, the numbers of conformational transitions sampled by CM-REMD simulations are actually smaller, 23 in the folding run and only 18 in the control run. Using the differences between the control and folding runs as rough estimates of the uncertainties, EE- and CM-REMD generated 26 ± 4 and 20.5 ± 5 reversible folding/unfolding transitions in an aggregated length of 800 μs. These numbers are not significantly higher than a total of 21 reversible folding/unfolding events sampled in the standard REMD simulation. Therefore, enhanced temperature diffusion, using either EE or CM, does not

appear to accelerate large-scale conformational transitions. The implication is that temperature diffusion is not limiting in driving large-scale conformational transitions even with the standard REMD. We note that similar observations have also been made by Kouza and Hansmann in their recent study of so-called rejection free replica exchange simulations[291].

Theoretical studies based on two-state model systems have predicted that REMD cannot drive transitions faster than the maximal rates at all the temperatures sampled (which should occur at slightly above T_m for actual proteins)[114]. Indeed, the average time per conformational transition, τ_{TS}, is $\sim$30 μs or more in all REMD simulations, which over twice longer than τ_{TS} of $\sim$14.3 μs observed in regular MD at T_m (see Table 5.1). We note that a key advantage of REMD, however, is the ability to generate correct thermodynamic dynamic ensembles at all temperatures, such that conformations sampled at all temperatures can be combined together using WHAM[286] to calculate thermodynamic properties any temperature of interest (e.g., the lowest temperature sampled). In contrast, recovering thermodynamic properties at the room temperature using a single simulation at T_m is generally unreliable. Therefore, slower average τ_{TS} in REMD compared to MD at T_m does not suggest that REMD is less efficient than regular MD (the contrary has been shown to be true extensively by previous works as discussed in Introduction). We also note that at 300 K no reversible folding and unfolding transition was sampled in multiple 100-μs MD runs and REMD is thus necessary for efficient generation of converged ensembles.

5.3.4 Convergence of thermodynamic properties

Arguably, the ultimate goal of REMD simulations is to generate statistically representative structure ensembles such that well-converged thermodynamic properties can be derived. Here, we focus on several typical thermodynamic properties frequently involved in protein folding studies, including T_m, CV, folding stability (ΔG_{fold}, "zero-dimensional" free energy), 1D PMF, and several 2D PMFs. Importantly, both CV and 1D PMF as a function of the total native contact

fraction (Q) are well-converged among all REMD simulations (e.g., see Figure 5.3). ΔG_{fold} values at $T_m \approx$ 354K derived from all REMD and MD simulations are within 0.25 kT from each other (except the CM-REMD control run, which deviates by ∼0.5 kT from other runs). Sample 2D PMFs, calculated from the control EE-REMD run, are shown in Figure 5.6. We note that inclusion of non-specific hydrophobic interactions does not change the fundamental features of these free energy surfaces compared to the original sequence-flavored model[232]. For example, the folding of the C-terminal hairpin precedes that of the N-terminal hairpin (see Figure 5.6D). However, all the free energy surfaces are significantly more rugged with several local minima that are not present in the surfaces derived from the original model (e.g., see Figure 5.6B).

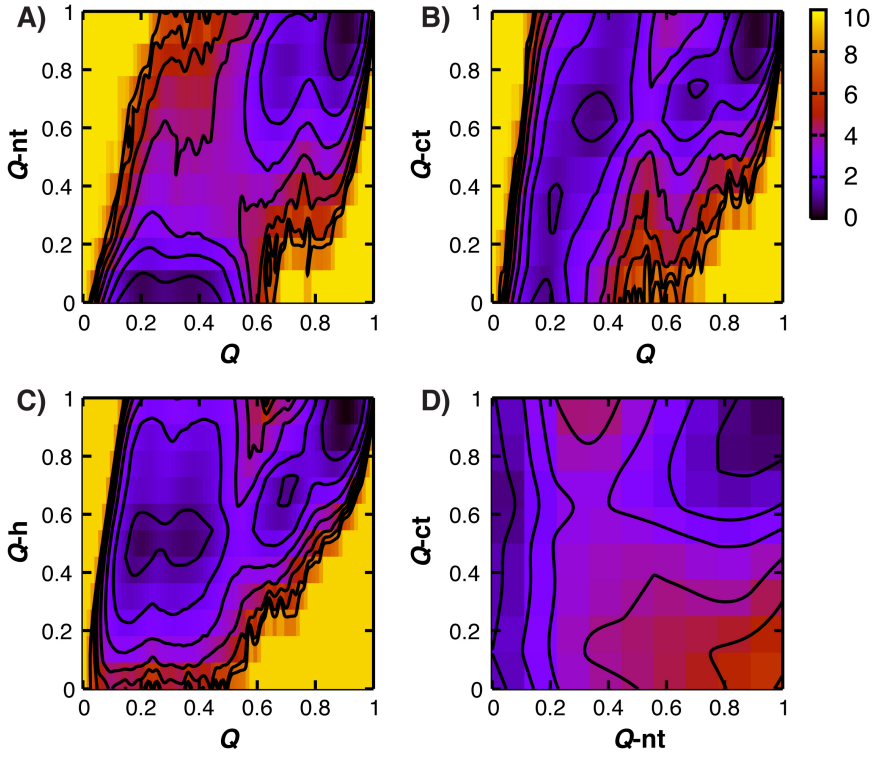


Figure 5.6: 2D free energy surfaces of folding of protein GB1 along various order parameters at T_m calculated from control EE-REMD simulations. Q, Q-nt, Q-ct and Q-h are fraction of native contacts of the whole protein, the N-terminal hairpin, the C-terminal hairpin and central helix respectively.

Figure 5.7 compares the self-convergence of 1D PMF and a representative 2D free energy (Q vs. Q-nt) as a function of simulation time, measured using the root mean square deviation (RMSD)

from the final profiles calculated using all 100 μs data. A few observations can be made. First, among all three simulations initiated from the folded state (EE control, CM control and standard REMD), both EE- and CM-REMD runs appear to converge faster compared to the standard REMD. For example, the 1D PMF quickly converged within about 5 μs in both EE- and CM-REMD control runs while it took $\sim$15 μs to achieve a similar level of convergence for the standard REMD (Figure 5.7A). Similar observations were also made in 2D PMFs (e.g., see Figure 5.7B). Importantly, the apparent faster convergence of EE- and CM-REMD control simulations within short timescales does not appear to arise from more efficient conformational sampling. Instead, it is mainly attributed to faster mixing of conformation states sampled at different temperatures due to more efficient temperature diffusion in EE- and CM-REMD. Second, comparing the results of the control and folding runs suggests that the true convergence of thermodynamic properties is much slower than the apparent rates of self-convergence in the control runs. Interestingly, EE-REMD appears to be slightly more efficient than CM-REMD despite less efficient temperature diffusion, which is consistent with the larger number of folding/unfolding transitions sampled. Third, the convergence of higher dimensional PMFs expectedly is much slower. All three REMD protocols perform very similarly, with the RMSD from the final profiles gradually decreasing throughout the 100 μs simulation span.

To further investigate how the number of actual conformational transitions sampled affects the accuracy of calculated thermodynamic properties, we divided both EE and CM control simulations into uniform fragments of various lengths and calculated the average numbers of conformational transitions and standard deviations of calculated ΔG_{fold}. The results are summarized in Figure 5.8. It shows a clear inverse correlation of the standard deviation of ΔG_{fold} and the number of conformational transitions sampled. Importantly, due to rapid mixing of conformations sampled at different temperatures (as noted above), reasonable estimates of ΔG_{fold}, with $\sim$1 kT uncertainty, appear feasible even with "ultra-short" REMD simulations on the order of 100 ns (which is actually the typical length of atomistic REMD simulations), even though virtually no reversible transitions could be sampled in such a short simulation timespan. In addition, the uncertainty in

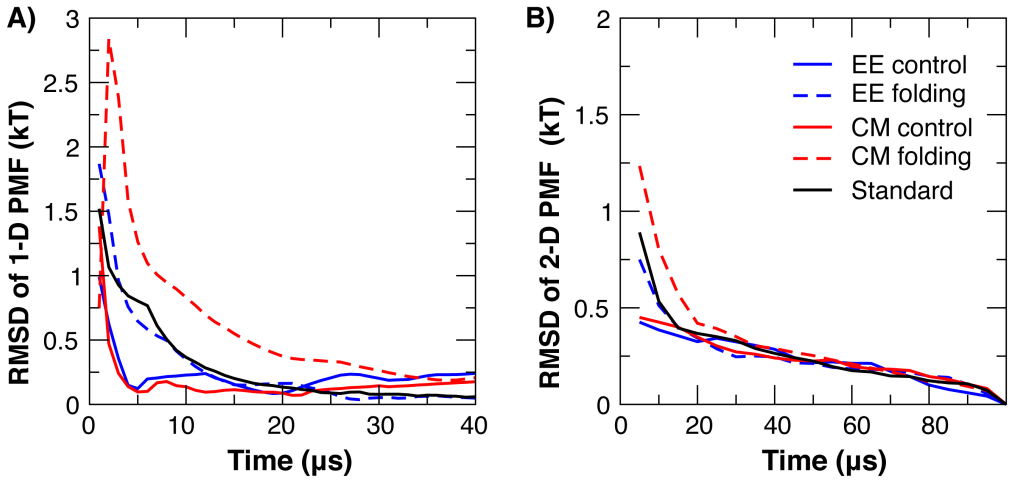


Figure 5.7: Self-convergence of (A) 1D PMF (as a function of Q), and (B) a representative 2D free energy surface (as a function of Q and Q-nt) during various 100-μs REMD simulations. The RMSD values shown were calculated with reference to the final 1D or 2D free energy profiles extracted using all 100 μs data of corresponding REMD runs.

the calculated ΔG_{fold} does not decrease significantly with longer simulations, until significant numbers of reversible folding/unfolding transitions are sampled. The later requires simulation timescales a few fold of the inherent folding/unfolding time, which is $>$20 μs for the current protein model. Along this line, it appears that equalized exchange achieved by EE-REMD is slightly advantageous (e.g., compare the black vs. red traces in Figure 5.8), in allowing more rapid mixing of different temperatures to achieve better convergence with simulations much shorter than the folding/unfolding timescales (which is the case for most atomistic simulations).

5.4 Conclusions

Existence of sharp cooperative transitions in proteins can lead to temperature exchange bottleneck and subsequently limit the sampling efficiency of the popular REMD method. Several approaches have been proposed to address this bottleneck, including two adaptive temperature REMD protocols previously examined by Lee and Olson[128]. In this work, we constructed a reasonably

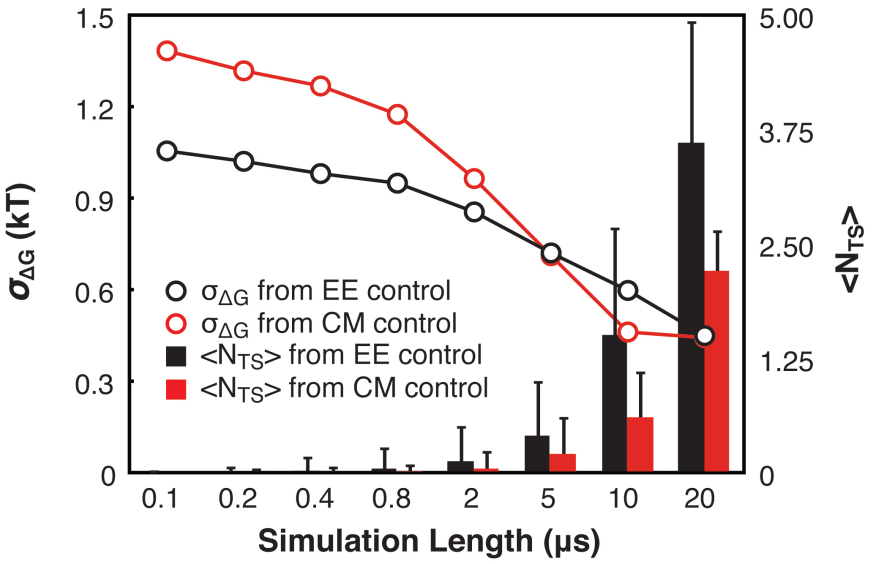


Figure 5.8: Standard deviation of calculated folding free energy ($\sigma\Delta G$) and average number of conformational transitions ($<N_{TS}>$) calculated using various length of EE- and CM-REMD control simulations.

realistic yet computationally tractable protein model to re-evaluate how effective equalizing the exchange rates or maximizing the current of temperature diffusion can enhance the sampling of large-scale protein conformational transitions. The results demonstrate that, despite substantially enhanced temperature diffusion, neither EE- nor CM-REMD could generate significantly more folding/unfolding transitions. As the result, the convergence of key thermodynamic properties is similar with EE-, CM- or standard REMD, except for very short simulations. In the later case, the apparent convergence can be mainly attributed to mixing of conformations sampled at different temperatures and thus can benefit from enhanced temperature diffusion from either EE- or CM-REMD. We note that typical atomistic REMD simulations are on the order of 100 ns and thus do fall into the "very short" category. A key lesson from the current work is also that, with exchange attempt frequency ($\sim$ps^{-1}) several orders of magnitude faster than the inherent timescale of slowest protein motions ($\sim\mu$s^{-1}), temperature diffusion does not appear to be limiting in the ability of REMD to drive large-scale conformational transitions. Therefore, it is unlikely that any protocol that aims to accelerate temperature exchange or temperature diffusion will lead to substantial enhancement in true sampling efficiency of REMD protein simulations. One will

likely need to explore protocols that involve some types of guided tempering to allow the protein to take advantage of faster unfolding at high temperatures simultaneously with rapid folding near the transition temperature.

5.5 Supplemental Materials

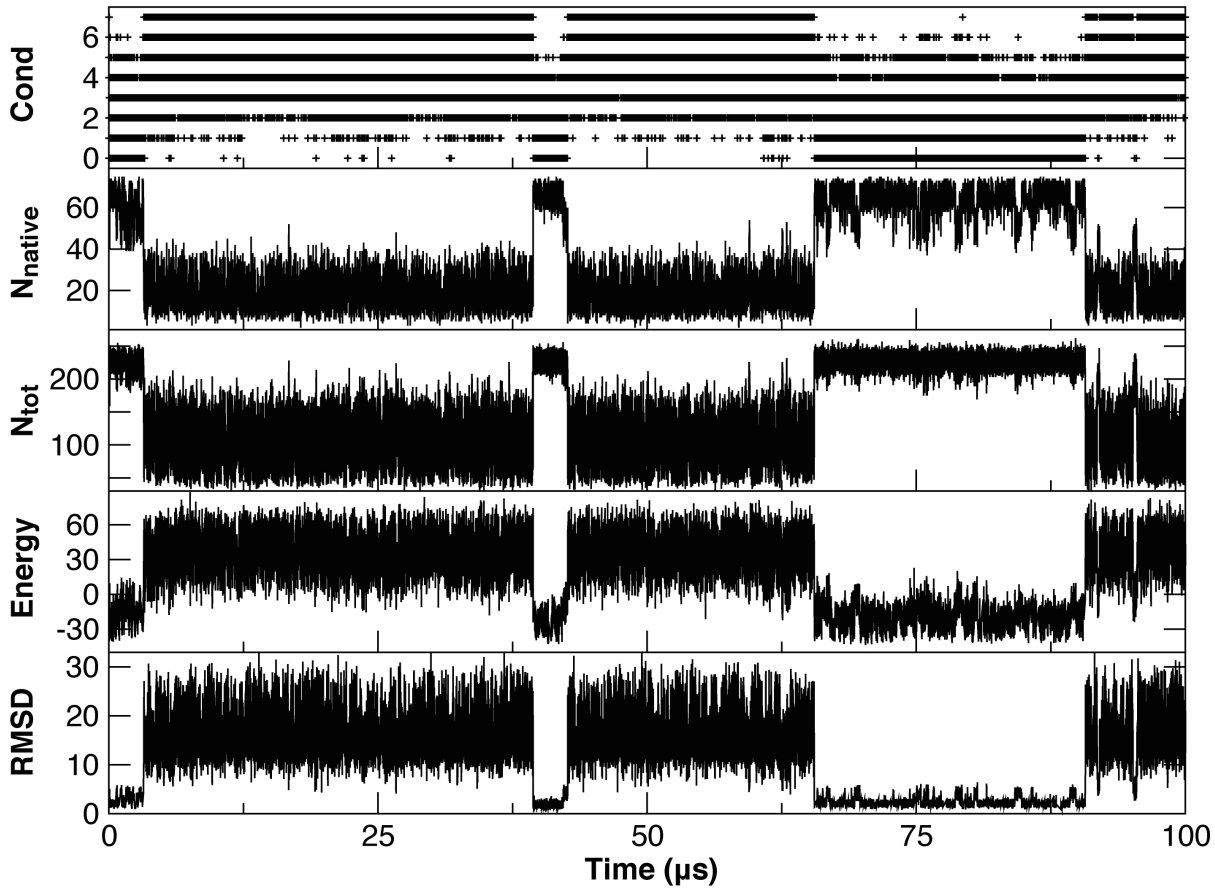


Figure S5.1: Time series of the temperature (condition number), total energy, RMSD and the numbers of native and total (nonspecific) contacts for a single replica (replica 1 from the EE control run; see main text). Nonspecific contacts are considered formed whenever the Cα-Cα distance is below 10 Å.

Chapter 6

Replica Exchange with Guided Annealing for Accelerated Sampling of Disordered Protein Conformations *

Abstract

We critically examine a recently proposed convective replica exchange (cRE) method for enhanced sampling of protein conformation based on theoretical and numerical analysis. The results demonstrate that cRE and related replica change with guided annealing (RE-GA) schemes lead to unbalanced exchange attempt probabilities and break detailed balance whenever the system undergoes slow conformational transitions (relative to the temperature diffusion timescale). Nonetheless, numerical simulations suggest that approximate canonical ensembles can be generated for systems with small conformational transition barriers. This suggests that RE-GA maybe uniquely suitable for simulating intrinsically disordered proteins (IDPs), an important class of newly recognized functional proteins. The efficacy of RE-GA is demonstrated by calculating the conformational ensembles of intrinsically disordered KID protein. The results show that RE-GA helps the protein to escape nonspecific compact states more efficiently and provides several fold speedups in generating correct and converged ensembles compared to the standard temperature RE.

*Accepted by *Journal of Computational Chemistry* as a full paper.
Author Contributions: Conceived and designed the experiments: JC WZ. Performed the experiments: WZ JC. Analyzed the data: WZ JC. Contributed reagents/materials/analysis tools: WZ JC. Wrote the paper: WZ JC.

6.1 Introduction

Temperature-based replica exchange (RE) is now widely accepted as a premier method for enhanced sampling of biomolecular conformation[104,105,292]. In RE, multiple replicas of the system are simulated independently at different temperatures, and they periodically attempt to exchange simulation temperatures according to a Metropolis criterion that preserves detailed balance and ensures proper canonical ensembles at all temperatures. The resulting random walk in the temperature space could help each replica to escape local energy minima and thus facilitate conformational sampling. Extensive theoretical analysis[111–116] and simulation studies[110,117–119] have established that RE can accelerate sampling as long as the activation enthalpies of conformational transitions are positive. At the same time, it has also been recognized that temperature RE can be severely limited by the presence of sharp, cooperative conformational transitions such as protein folding[111,120]. In such cases, replicas often segregate into distinct conformational subspaces (e.g., folded/compact and unfolded), and the segregation may often be accompanied with apparent temperature exchange bottlenecks[121]. Importantly, so-called adaptive RE methods can effectively reduce or eliminate exchange bottlenecks, by adjusting the RE temperature distribution to either equalize exchange success rates[129] or maximize temperature diffusion current[128,130]. However, improved temperature distributions were found to be limited in improving actual sampling[293–295]. In particular, a recent benchmark study using coarse-grained protein models directly demonstrated that improved temperature exchange did not always lead to better sampling of major conformational transitions, judged by rates of global folding transitions as well as the convergence of various key thermodynamic properties of interest[254].

A common symptom of RE in the challenging case of simulating cooperative transitions is that a few replicas occupy the lowest temperatures for extensive period of time, if not for the entire span of RE after initial relaxation (e.g., see Figure 1 of Ref.[121]). These low temperature replicas always occupy conformational states that are enthalpically much more stable than other states sampled at

higher temperatures. Even though the trapped replicas do visit higher temperatures occasionally, especially with adaptive RE schemes, they usually exchange back to lower temperatures very rapidly and thus can not take sufficient advantage of higher temperatures to escape the stable energy minima. This observation has led us to suggest that an effective way to overcome the apparent conformational/temperature trapping may be to design temperature exchange schemes that would guide the trapped replicas to visit higher temperatures more frequently[121]. Such schemes may be generally referred to as RE with guided annealing (RE-GA). A similar idea is the recently described convective RE method (cRE)[131], where the replicas take turns to steadily raise or lower their temperatures in "convective" cycles. We have also been independently exploring several alternative RE-GA schemes. For example, one could first detect trapped replicas such as by examining the average temperatures and then guide the identified trapped replicas to steadily raise temperature by modifying the exchange attempt patterns. Alternatively, one can periodically chose the lower-temperature replica and subject it to guided annealing. Since these RE-GA schemes only modify the exchange attempt patterns and retain the use of Metropolis criteria for accepting exchange attempts, it appears that detailed balance should be maintained and correct canonical ensemble generated. In fact, a formal proof of detailed balance was offered by Spill et al. in their description of the cRE method[131].

In this work, we rigorously evaluate the correctness and effectiveness of RE-GA schemes including cRE, based on theoretical analysis, numerical simulations, and proteins simulations at coarse-grained and atomistic levels. The results reveal that all RE-GA schemes inevitably lead to unbalanced exchanged attempt probabilities and do not rigorously maintain detailed balance in cases of slow conformational transitions. Nonetheless, numerical simulations suggest that the systematic bias induced by GA cycles is minimal when the activation enthalpy of conformational transitions is small (e.g., $<$ a few kcal/mol). Therefore, RE-GA appears to be suitable for accelerated sampling of disordered protein states using atomistic force fields. It is particularly effective in assisting the system to escape numerous nonspecific compact conformational states that are present in the current atomistic force fields, which in turn allow more efficient sampling

of transient local protein conformations. Application to the simulation of the disordered ensemble of a 28-residue intrinsically disordered protein (IDP) demonstrates that RE-GA is several fold more efficient compared to the standard temperature RE.

6.2 Methods

6.2.1 Theory of replica exchange with guided annealing (RE-GA)

In temperature RE simulations, N replicas of the system are simulated in parallel at different temperatures distributed between two pre-selected values, T_{min}, which is often the temperature of interest, and T_{max}, which should be slightly above the temperature of maximal reversible transition kinetics[114]. Periodically, replicas attempt to exchange simulation temperatures with an acceptance probability given by the Metropolis criterion,

$$p(i \leftrightarrow j) = min[1, e^{(\beta_j - \beta_i)(E_j - E_i)}] \tag{6.1}$$

where $\beta_i = 1/k_B T_i$, k_B is Boltzmann's constant, and T_i and E_i are the current temperature and total potential energy of replica i, respectively. Most standard RE implementations only involve exchange attempts between replicas occupying neighboring temperatures, even though several more aggressive global exchange schemes have been proposed[122,296]. Importantly, the exchange attempts alternate between even and odd neighboring pairs, such that all replicas have equal probabilities to attempt to exchange to either higher or lower temperatures except when they are occupying T_{min} or T_{max}. The balanced exchange attempt probabilities together with the Metropolis exchange acceptance criterion (Equation 6.1) allow proper canonical ensembles to be generated at all temperatures during RE simulations. It should also be emphasized that Equation 6.1 is valid only when all replicas are fully relaxed and their conformational states satisfy the

canonical distributions prior to exchange attempts. Some aggressive exchange schemes such as the rejection-free RE[291], while formally correct, do not allow sufficient importance sampling between exchanges and will fail to generate canonical ensembles. This principle was thoroughly discussed by Kouza and Hansmann[291].

Given the Metropolis criterion of Equation 6.1, replicas that sample lower energy states naturally tend to occupy lower temperatures, and those sampling higher energy states tend to occupy high temperatures. For systems with sharp cooperative transitions, the low energy replicas will remain at low temperatures until spontaneous large-scale conformational transitions occur between folded/compact and unfolded states. This can lead to apparent segregation of replicas in the temperature space. The conformational transitions that allow replicas to escape trapping (at either low or high temperature regions) will involve either spontaneous unfolding of folded/compact states at low temperature or folding of disordered states at high temperature. Both transitions are rare events and this further worsens the replica segregation problem as frequently observed in protein RE simulations[111,117,121,297,298]. Arguably, this may be considered a fundamental limitation of temperature RE, and one may have to consider Hamiltonian RE to overcome this limitation[135,299,300]. Alternatively, RE-GA attempts to directly address the apparent replica segregation through GA cycles, which modify the exchange attempt patterns to guide the folded/compact replicas to visit higher temperatures more frequently and accelerate unfolding and/or to guide the unfolded replicas towards lower temperatures to encourage folding transitions.

In this work, we mainly focus on three GA schemes. The first scheme is the cRE method previously described by Spill et al.[131], where replicas take turn to be the "stick replica" and undergo convective annealing cycles. For that, the exchange attempts are chosen such that the stick replica will first only try to exchange to higher temperatures until reaching T_{max}. Then, the stick replica will only attempt to exchange to lower temperatures. When the stick replica reaches T_{min}, the direction of exchange attempts is reversed again until the stick replica returns to its original temperature. In this work, the cRE algorithm was slightly modified to include a convective

annealing frequency, which specifies the number of regular RE steps between convective annealing cycles. The original algorithm[131] corresponds to the limit of annealing frequency zero. In the second scheme, referred to as RE with selective annealing (RE-SA) hereafter, we periodically select the lowest temperature replica and subject it to a GA cycle where only exchanges to higher temperatures are attempted until the selected replica reach T_{max}. The main difference between RE-SA and cRE is that the replica is not guided to exchange toward T_{min} right after reaching T_{max}, which allows the replica to spend more time at higher temperatures to escape the stable energy minima. An alternative implementation of RE-SA is first to detect so-called trapped replicas, whose average temperatures during last N_{trap} RE cycles are below a preselected threshold, and then periodically selected the trapped replica occupying the lowest temperature to undergo GA cycles. These two implementations appear to be equivalent in practice, and we only present the results for the implementation without explicit detection of temperature trapping. The third scheme, denoted RE with random annealing (RE-RA), periodically selects a random replica, regardless of its temperature, and subjects it to a GA cycle that guides it to reach T_{max} before being released. All RE-GA schemes have been implemented in MMTSB[301].

Spill et al. have previously offered a formal proof of detailed balance for cRE[131]. The key conclusion from the proof is that an algorithm driving the stick replica through complete round trips in the temperature space should guarantee detailed balance. However, the proof contains an error in reordering the double summation in Equation 5 of Ref.[131]. The reordering is only correct if the numbers of actual exchange attempts required for successful exchange to higher and lower temperatures are the same. However, in the presence of slow cooperative conformational transitions, it will takes many more attempts for replicas occupying low energy states to exchange to higher temperatures than returning to lower temperatures, and vice versa for replicas occupying high energy states. As such, the net drift in the stationary distributions, B, as defined in Ref.[131], will be non-zero whenever spontaneous conformational transitions significantly lag (convective) temperature cycles. In other words, cRE (and other RE-GA schemes) lead to unbalanced exchange attempt probabilities and break the detailed balance condition when the conformational transitions

are slow compared to the temperature diffusion timescales. The theoretical analysis will be validated using both numerical and coarse-grained protein simulations.

6.2.2 Numerical RE simulations using a kinetic network model

Numerical simulations based on a two-state kinetic network model were performed to characterize the properties of various RE-GA schemes and study the dependence of their performance on inherent conformational dynamics of the system. The system is assumed to undergo ideal two-state transition between folded and unfolded states. As summarized in Table S6.1, four models with different folding and unfolding activation enthalpy and entropy, denoted as $\Delta H_f^\dagger$, $\Delta H_u^\dagger$, $\Delta S_f^\dagger$ and $\Delta S_u^\dagger$, respectively, are constructed. Given the folding and unfolding rates at different temperatures,

$$\begin{aligned} k_f(T) &= e^{-(\Delta H_f^\dagger - T\Delta S_f^\dagger)/k_B T}, \\ k_u(T) &= e^{-(\Delta H_u^\dagger - T\Delta S_u^\dagger)/k_B T}, \end{aligned} \tag{6.2}$$

the probability of spontaneous conformational transitions during between replica exchange attempts can be calculated as,

$$\begin{aligned} p_{f\to u}(T) &= 1 - e^{-\frac{\Delta t}{\tau_u(T)}}, \\ p_{u\to f}(T) &= 1 - e^{-\frac{\Delta t}{\tau_f(T)}}, \end{aligned} \tag{6.3}$$

where Δt is the time between exchange attempts, and $\tau = 1/k$. The fifth model was design to mimic the anti-Arrhenius behavior of protein folding[114], where the folding activation enthalpy and entropy vary linearly with respect to the temperature and its logarithm, respectively:

$$\begin{aligned} \Delta H_f^\dagger(T) &= \Delta H_f^\dagger(T_0) + (T - T_0)\Delta C_p^\dagger, \\ \Delta S_f^\dagger(T) &= \Delta S_f^\dagger(T_0) + ln(\frac{T}{T_0})\Delta C_p^\dagger. \end{aligned} \tag{6.4}$$

When the activation heat capacity, $\Delta C_p^\dagger$, is negative, the activation enthalpy becomes negative and the folding rate decreases as a function of temperature above $T = T_0 - \Delta H_f^\dagger(T_0)/\Delta C_p^\dagger$. Table S6.1 summarizes the key parameters including temperature dependence of folding and unfolding times of all five models. Note the parameters for model 5 were designed to mimic the Gō-like model of protein G B1 domain, where the folding rate displays anti-Arrhenius behavior from 300 K (that is $\Delta H_f^\dagger$ (T_0=300K)<0). The numerical simulations involve 4 or 8 replicas distributed exponentially between 300 and 450 K (see Table S6.1). Temperature exchanges were attempted every 10 ps, and the total length of each numerical simulation is 1.0 ms (100 million RE cycles). For each model, 10 independent numerical simulations were performed to analyze the conformational transition statistics and thermodynamic properties.

6.2.3 Coarse-grained and atomistic protein simulations

Both coarse-grained and atomistic protein simulations were performed to further evaluate the correctness and efficiency of various RE-GA schemes. Following our previous work[121], a sequence-flavored Gō-like model was derived from the PDB structure of protein G B1 domain (PDB: 3gb1[290]; see Figure 6.1A), using the Multiscale Modeling Tools for Structural Biology (MMTSB) Gō-Model Builder (http://www.mmtsb.org)[232,233]. The model was recalibrated to to have T_m $\sim$ 350K by uniformly adjusting the strength of all contacts. Standard RE, cRE and RE-SA simulations were carried out using 8 replicas spanning 300 to 400K. For each replica, Langevin dynamics simulation was performed using CHARMM[63,216] with a dynamic time step of 10 fs and a friction coefficient of 0.1 ps^{-1}. We note that the use of small friction coefficient expedites conformational dynamics and that test simulations using higher friction coefficients of 1.0 and 5.0 ps^{-1} generated virtually identical conformational ensembles. The SHAKE algorithm[217] was used to constrain all virtual bond lengths. Exchanges were attempted between neighboring replicas every 5000 dynamics steps (50 ps). For RE-SA, the replica occupying the lowest temperature was selected for GA every 1000 RE cycles (50 ns) upon the completion of the previous annealing

cycle. The lengths of all simulations were 10 μs per replica. An additional RE-SA simulation was initiated from a fully extended conformation (referred to as "folding" run) to examine the level of convergence.

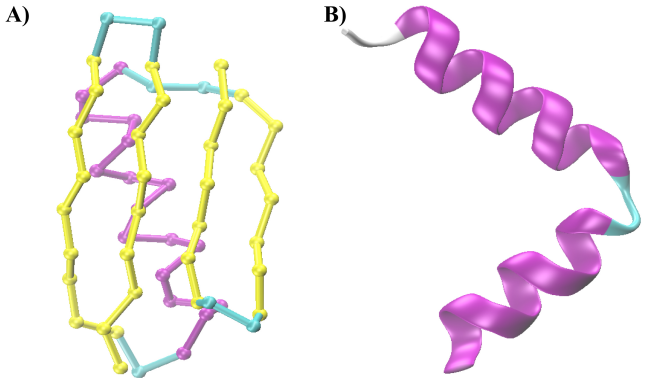


Figure 6.1: Folded structures of (A) protein GB1, shown in Cα-only trace representation, and (B) KID domain, shown in the cartoon representation. The helical, sheet and loop regions are colored purple, yellow and cyan, respectively.

Atomistic simulations of intrinsically disordered kinase inducible domain (KID; residues 119-146, TD SQKRR EILSR RPSYR KILND LSSDA P) of the transcription factor CREB were performed in the GBSW implicit solvent protein force field[75,77]. The folded conformation of KID was derived from the pKID/KIX complex structure (PDB ID: 1kdx[302]; see Figure 6.1B). We have previous characterized the disordered ensembles of KID using two independent 200-ns standard RE simulations[169]. In the current work, two independent RE-SA simulations (control and folding) were carried out using 16 replicas spanning 270 to 500 K. Langevin simulations were performed with a friction coefficient of 0.1 ps^{-1}, and temperature exchanges were attempted every 2 ps. The SHAKE algorithm[217] was used to constrain the lengths all bonds involving hydrogen atoms and the dynamic time step was 2 fs. The GA frequency is every 200 RE steps (0.4 ns). The lengths of both control and folding simulations are 50 ns per replica. All analysis was performed using a combination of the MMTSB toolset, CHARMM and in-house scripts.

6.3 Results

6.3.1 Evaluation of RE-GA schemes based on coarse-grained protein simulations

We first utilized the sequence-flavored Gō-like model of protein G B1 domain (Figure 6.1A), to investigate if the introduction of GA cycles could indeed enhance the sampling efficiency and to examine if RE-GA could generate correct canonical ensembles. Gō-like models are based on the principle of minimal frustration for evolved proteins[98], and have been shown to be capable of generating realistic reversible folding transitions[229]. These models thus provide a good balance between the level of realisticity and computational tractability for rigorous evaluation of the proposed RE-GA schemes. As illustrated in Figure 6.2A, introduction of GA cycles, in both cRE and RE-SA schemes, dramatically accelerated folding and unfolding transitions compared to the standard RE. Specifically, about 110 and 350 reversible folding transitions were observed during the 10-μs cRE and RE-SA simulations, respectively, compared to only 8 such transitions during the standard RE simulation. Therefore, GA cycles seem effective in overcoming the apparent segregation of replicas in the temperature space and dramatically enhance conformational sampling. Furthermore, not requiring the replica to return to low temperatures upon reaching T_{max} in RE-SA appears more efficient in driving sampling, generating over 3-fold more reversible transitions compared to cRE.

Further analysis of the thermodynamic properties revealed systematic differences in the ensembles derived from cRE and RE-SA simulations compared to the standard RE. For example, Figure 6.2B compares the free energy profiles as a function of the total native contact fraction. Even though all free energy profiles are well converged (e.g., comparing RE-SA control vs. folding in Figure 6.2B), both cRE and RE-SA significantly under-estimate the stability of the folded state with respect to the unfolded state. The folding temperature, T_m, calculated from RE-SA is 17 K lower than the reference value calculated from RE. Similar observations can also be made

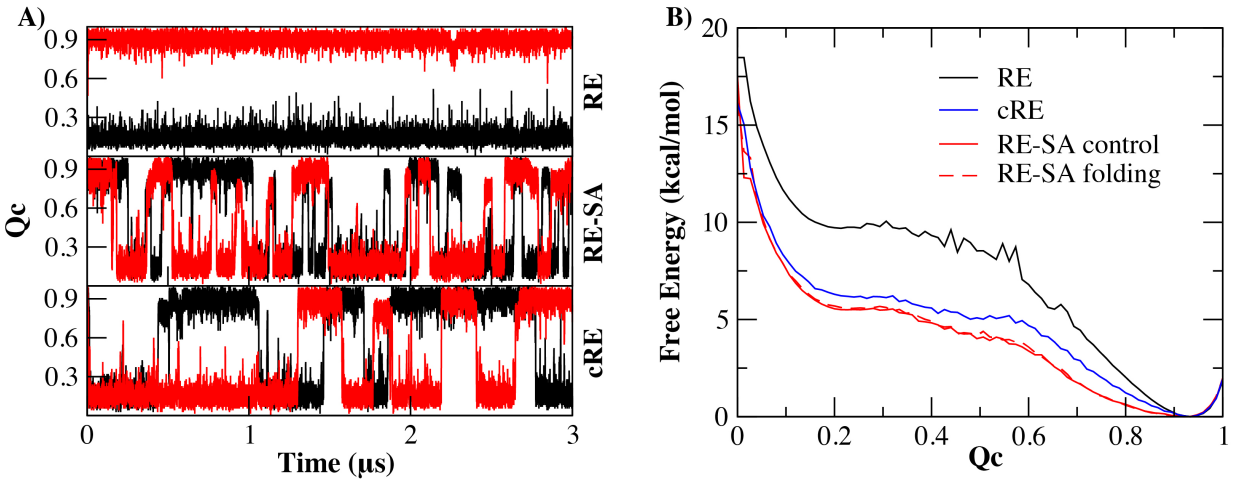


Figure 6.2: (A) The time evolution of the native contact fraction for three representative replicas, selected from the standard RE, RE-SA, and cRE control simulations of protein GB1, respectively. (B) Free energy profiles as a function of the total native contact fraction of protein GB1 at 300 K, calculated from the last 8 μs of the 10-μs standard RE, cRE and RE-SA simulations.

when comparing various multi-dimensional free energy surfaces, which consistently show that RE-SA over-estimates the unfolded population in the simulated ensembles (e.g., see Figure S6.1). The observation that cRE and RE-SA, even though apparently effective in driving faster folding transitions, lead to systematic biases in the simulated ensembles is consistent with the theoretical analysis, which concludes that exchange attempt probabilities are not balanced with GA cycles in the presence of slow cooperative conformational transitions (see Method). Spill et al.[131] aslo observed systematic over-representation of high-energy states in their cRE simulations (e.g., see Table 1 of Ref.[131]), even though this was thought to merely reflect a natural consequence of more extensive sampling of cRE.

6.3.2 Numerical simulations of RE-GA schemes

Numerical simulations were performed to further investigate how the systematic bias of various RE-GA schemes depends on inherent conformational dynamics of the system and key GA parameters. As summarized in Table 6.1 and S6.1, four models with increasing folding enthalpy and

Table 6.1: Summary of model parameters and key statistics of the numerical RE and RE-SA simulations. Only results from RE-SA simulations with an annealing frequency of every 1000 RE cycles are shown. Results using different GA frequencies and the additional RE-RA scheme are shown in the Supporting Materials (Figure S6.2 and Figure S6.3). Full model parameters are shown in Table S6.1.

Model	ΔH_f	ΔS_f	RE			RE-SA		
	(kcal/mol)	(kcal/mol)	K_{eq}	N_{TS}	K_{eq}/K_{eq0}	K_{eq}	N_{TS}	K_{eq}/K_{eq0}
1	-1.5	-3.75	0.52±0.02	250.3±19	0.99	0.54±0.02	251.2±19	1.01
2	-6	-15	0.082±0.03	12.2±3.6	1.01	0.076±0.01	12.5±3.3	0.95
3	-9	-22.5	0.021±0.006	3.5±1.7	0.91	0.017±0.007	3.9±1.4	0.75
4	-12	-30	$6.5\pm0.1\times10^{-3}$	3.0±1.5	1.00	$2.7\pm0.1\times10^{-3}$	3.6±1.7	0.41
5	-38	-111	$3.7\pm0.1\times10^{-4}$	58.8±8.6	1.07	$2.1\pm0.1\times10^{-3}$	178.5±16	6.05

ΔH_f and ΔS_f denote the folding enthalpy and entropy, respectively. The ideal unfolding equilibrium constant is calculated as $K_{eq0} = e^{[(\Delta H_f - T\Delta S_f)/RT]}$. Simulated K_{eq} values were calculated directly from the statistics collected at T = 300 K. N_{TS} denotes the number of reversible transitions sampled per replica. The average and standard deviations from 10 independent numerical simulations are shown.

entropy were designed to mimic ideal 2-state proteins of increasing size and stability. Model 1 has the lowest folding energy barrier and conformational transitions occur at fast ns timescales. It represents highly flexible proteins with mostly local structural transitions, such as intrinsically disordered proteins (IDPs)[1,3,35]. In contrast, model 4 has sizable folding enthalpy and entropy, and the folding and unfolding transitions occur at μs to ms timescales. This model mimics moderately sized proteins with slow corporative folding transitions. The results of numerical simulations of the standard RE and three RE-GA schemes (cRE, RE-SA and RE-RA) using these four models are summarized in Table 6.1 and Figure 6.3. More results are summarized in Figure S6.2 and S6.3. The results confirm that all RE-GA schemes lead to systematic errors in calculated equilibrium constants for models with slow conformational transitions (e.g., models 3 and 4; Figure S6.2 C-D). The systematic errors decreases with lower folding/unfolding barriers and faster conformational transitions, and become minimal for model 1. In contrast, the standard RE always generates proper canonical ensembles and recovers the expected equilibrium constants for all models (also see Table 6.1). Therefore, RE-GA does break detail balance and fails to generate proper canonical

ensembles. As expected, all RE-GA schemes eventually recover the correct behavior of the standard RE with less frequent annealing cycles.

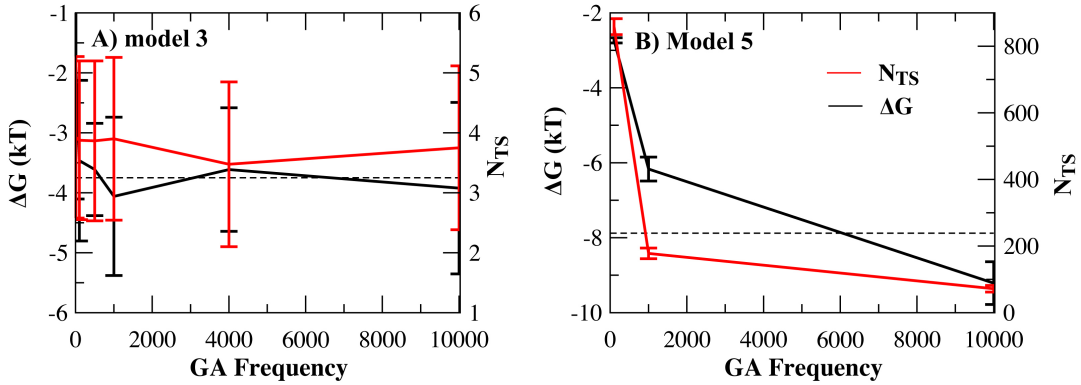


Figure 6.3: Numerical simulations of RE-SA using (A) model 3 and (B) model 5. The specific model parameters are provided in Table S6.1. The red traces plot the number of reversible transitions per replica (N_{TS}) and the black traces plot the calculated folding free energies. The dashed lines mark the ideal folding free energies. Errors bars shown correspond to the deviations of 10 independent 1 ms numerical simulations.

A surprising observation from the numerical simulations is that, as summarized in Figure 6.3A and S6.3, introduction of GA cycles does not lead to any significant increase in the numbers of conformation transitions sampled for models 1 through 4. The implication is that the standard RE protocol is already as efficient as possible for enhanced sampling of systems that have positive activation enthalpies and display Arrhenius temperature dependence. This observation apparently contradicts the results from coarse-grained protein simulations (e.g., see Figure 6.2A). To reconcile this discrepancy, we designed an ideal 2-state model with folding and unfolding barriers similar to those derived from the Gō-like model of protein G B1 domain (model 5 in Table 6.1). Model 5 is anti-Arrhenius with a negative folding activation enthalpy and a large folding entropic barrier. Numerical simulations showed that RE-SA did accelerate reversible transitions about 3-fold compared to RE, albeit at the same time over-estimating K_{eq} about 6-fold (Table 6.1).

6.3.3 Atomistic simulation of intrinsically disordered KID protein

The observation that RE-GA leads to minimal systematic errors for systems with small transition free energy barriers suggests that RE-GA may be uniquely suitable for simulating disordered ensembles of IDPs. In particular, we have previously observed that the tendency of current implicit solvent protein force fields to over-stabilize nonspecific collapsed states severely limits the ability to generate well-converged ensembles for moderately sized IDPs using the standard RE protocol[1,34]. Here, we evaluate the applicability and efficiency of RE-SA for simulating the KID domain in the GBSW implicit solvent[77]. Our previous work showed that it required over 160 ns for the standard RE simulations to achieve satisfactory convergence in the simulated conformational ensembles[169] (also see Figure 6.4D-F). In contrast, RE-SA appears to achieve convergence well within 50 ns. As shown in Figure 6.4, the residue helicity profiles from both the control and folding RE-SA simulations fully stabilize within 40 ns (Panels A-B), while in RE simulations the profiles continue to fluctuate up to the full simulation time span of 200 ns (especially in the control run; Panels D-E). Comparing results derived from the converged segments of RE and RE-SA simulations (last 40%), summarized in Panels C and F, shows that control and folding RE-SA runs yielded essentially identical helicity profile with a root-mean-square difference (RMSD) of 0.022, while substantial differences persist in the RE-derived profiles (RMSD = 0.088) despite four times longer sampling time. Importantly, the final helicity profiles derived from RE and RE-SA are similar, and both agree with previous NMR chemical shift analysis[303]. We have further examined the distributions of helical substates in the simulated ensembles. The results, shown in Figure 6.5, demonstrate that the same states were sampled by RE-SA with similar probabilities (e.g., compared to Figure 5a of Ref.[169]), but much higher efficiency. In particular, the 2D distributions derived from the control and folding runs are better converged with shorter RE-SA simulations (RMSD = 0.0012) compared to those derived from longer RE simulations (RMSD = 0.0047; Figure S1 of Ref.[169]). Therefore, RE-SA was able to generate apparently correct disordered conformational ensembles of KID with several fold higher efficiency compared to RE.

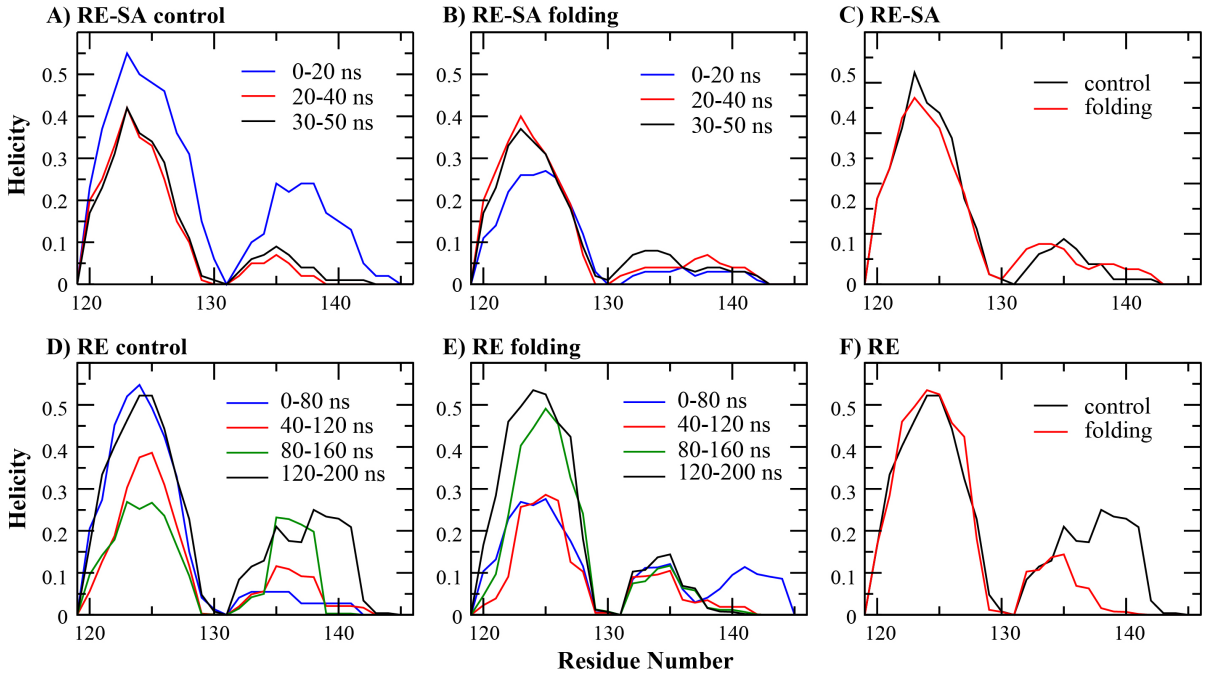


Figure 6.4: Convergence of calculated residue helicity profiles of KID using RE-SA (A)-(C) and RE (D)-(F). Only conformations sampled at 285 K were considered in the analysis. Panel (C) compared results from the last 20 ns of control and folding RE-SA simulations, and Panel (F) compared results from the last 80 ns of the control and folding RE simulations previously reported[169].

Interestingly, the most important differences between the conformational ensembles generated by RE and RE-SA are observed in the biopolymer properties. The end-to-end distance distributions derived from RE and RE-SA simulations, shown in Figure 6.6, cover similar ranges between 5 to 25 Å. However, RE-SA yields much smoother and more continuous distributions, while the RE-derived distribution contains more features and is dominated by a few compact states with small end-to-end distances. The interpretation, derived largely based on direct examination of the temperature and conformational evolution of replicas as a function of simulation time, is that KID were frequently trapped in nonspecific collapsed states during the RE simulation. While the tapping did not appear to prevent the peptide to visit various local (helical) structures, it clearly reduces the efficiency of local sampling in RE. In contrast, GA cycles appeared to be highly effective in assisting the peptide to escape the collapse states, which allows the peptide to visit a wide range of global configurations and facilitate the sampling of local structures in RE-SA. The

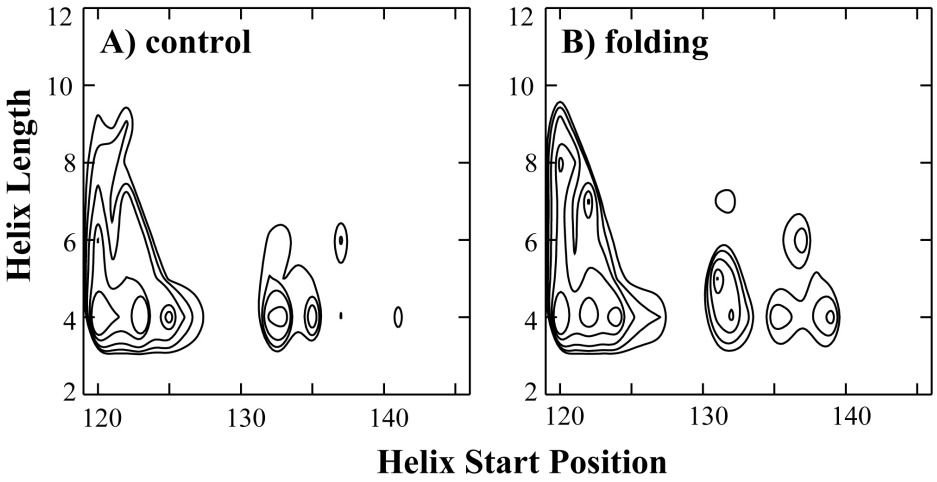


Figure 6.5: Probability distributions of helical substates of KID derived from the last 20 ns of 50-ns RE-SA control (A) and folding (B) simulations. The helical states are defined based on the helix starting position and length. Contours are drawn at 0.002, 0.005, 0.01, 0.02, 0.04 and 0.06 levels.

caveat is that RE-SA may generate ensembles that are overly expanded.

6.4 Discussion

We have examined the ability of various RE-GA schemes including the recently developed cRE method[131] to enhance conformational sampling and generate proper canonical ensembles. The introduction of GA cycles has been motivated by the observation that replicas in the standard RE tend to segregate in the temperature space in the presence of sharp cooperative transitions such as protein folding. Theoretical analysis argues that introduction of GA cycles results in unbalanced exchange attempt probabilities and thus breaks the detailed balance condition whenever spontaneous conformational transitions lag temperature diffusion. This conclusion is validated by both coarse-grained protein simulations and numerical RE simulations using 2-state kinetic models. The numerical simulations further reveal that GA cycles only accelerate sampling for systems that display anti-Arrhenius kinetics. For systems with positive folding activation enthalpies, the standard RE appears to be already maximally efficient. Another key observation from the

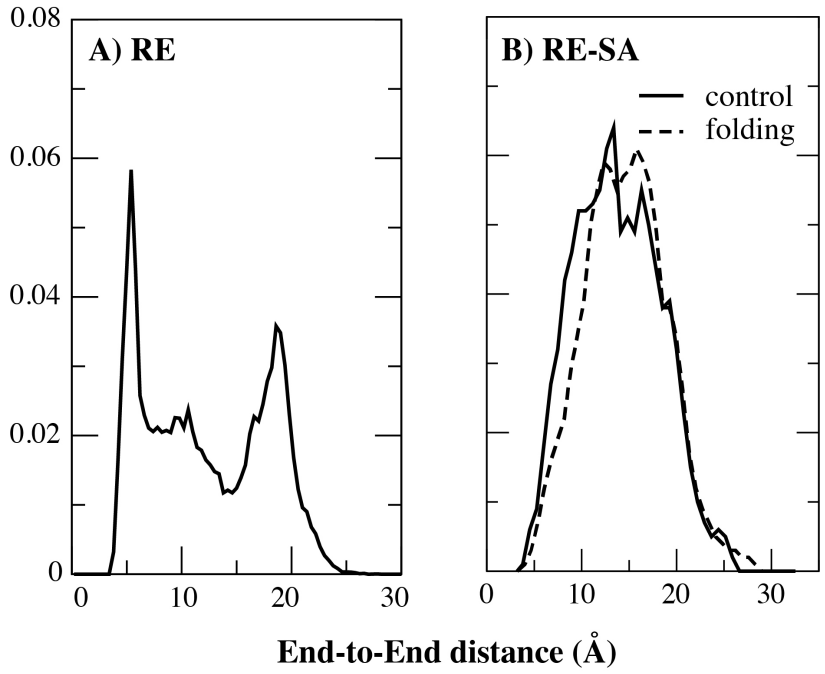


Figure 6.6: Distributions of the end-to-end distance of KID, calculated from (A) the last 80 ns of 200-ns RE folding simulation[169] and (B) the last 20 ns of the control and folding RE-SA simulations.

numerical simulations is that systematic errors introduced by GA cycles diminish for systems with low transition barriers. This observation suggests that RE-GA may be uniquely suitable for disordered protein states such as IDPs, an important class of recently recognized functional proteins[3,35]. In particular, the peptide is frequently trapped in nonspecific collapsed states during the RE atomistic simulations[74], and GA cycles should be highly effective in assisting the peptide to escape these trapped states and better sample local conformational states. Indeed, application to a prototypical model IDP demonstrates that RE-GA achieves superior convergence in conformational properties with merely one-fourth of simulation time compared to the standard RE. Importantly, no systematically difference was observed between the conformational ensembles derived from RE and RE-GA simulations, except certain biopolymer properties where improved sampling of RE-GA yields smooth and continuous distributions expected for highly disordered peptides. At present, quantitative estimate of the potential systematic errors introduced by the GA cycles does not appear feasible for an unknown system. This is an important caveat of the proposed RE-GA protocols. Nonetheless, the numerical simulations based on ideal 2-state models

suggest that the error in the calculated folding equilibrium constant is only $\sim$5% for model 2 with modest barriers ($\Delta H_f^\dagger$ = 1 kcal/mol and $\Delta H_u^\dagger$ = 7 kcal/mol; see Table 6.1 and Table S6.1). The implication is that RE-GA should be applicable to most intrinsically disordered proteins, where the activation barriers for transitions between conformational states are expected to be small. For these systems, the systematic error introduced by the GA cycles is likely much smaller than the achievable level of convergence (e.g., see Figure 6.3A). Therefore, we anticipate that RE-GA will be very useful in simulation of disordered protein states in general, despite its inherent tendency to introduce systematic errors for simulating folded proteins.

6.5 Supplemental Materials

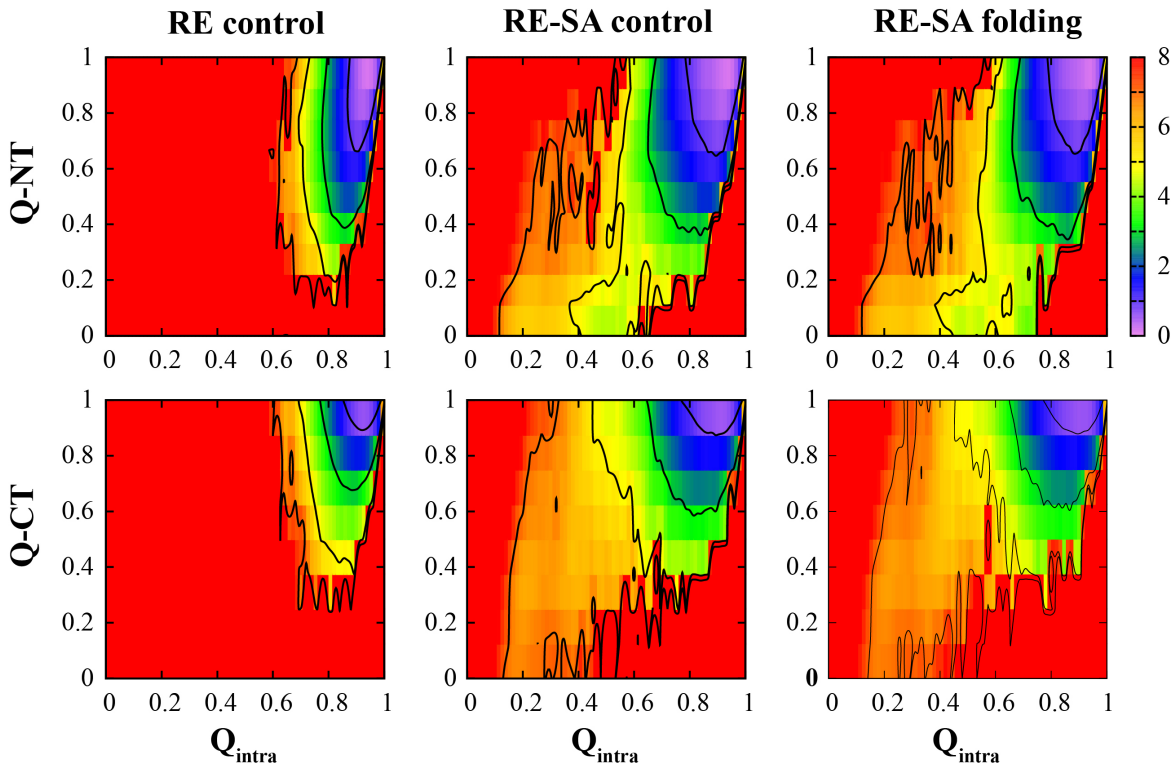


Figure S6.1: 2D Free energy surface as a function of the total fraction of native contacts formed (Q_{intra}) and the fractions of native contacts formed in either the C-terminal β-hairpin (Q-CT) or the N-terminal β-hairpin (Q-NT). The free energy surfaces were derived based on conformations sampled at 300 K during the last 8 μs of the standard RE and RE-GA coarse-grained simulations of protein GB1. Comparison of results from the control and folding RE-GA runs demonstrates that the free energy surfaces are very well converged but deviate significantly from those derived from the standard RE run.

Table S6.1: Key parameters of kinetic models used in numerical RE simulations.

Condition	Temperature (K)	ΔG_f (kcal/mol)	τ_f(ns)	τ_u(ns)
Model 1				
$\Delta H_f^\dagger$ = 0.25 kcal/mol, $\Delta S_f^\dagger$ = -1.25 kcal/K/mol				
$\Delta H_u^\dagger$ = 1.75 kcal/mol, $\Delta S_u^\dagger$ = 2.5 kcal/K/mol				
0	300.0	-0.375	2.86	5.35
1	343.4	-0.212	2.71	3.69
2	393.1	-0.026	2.58	2.67
3	450.0	0.188	2.48	2.01
Model 2				
$\Delta H_f^\dagger$ = 1 kcal/mol, $\Delta S_f^\dagger$ = -5.0 kcal/K/mol				
$\Delta H_u^\dagger$ = 7 kcal/mol, $\Delta S_u^\dagger$ = 10 kcal/K/mol				
0	300.0	-1.5	66.28	820.81
1	343.4	-0.849	53.61	186.02
2	393.1	-0.103	44.55	50.86
3	450.0	0.75	37.89	16.37
Model 3				
$\Delta H_f^\dagger$ = 1.5 kcal/mol, $\Delta S_f^\dagger$ = -7.5 kcal/K/mol				
$\Delta H_u^\dagger$ = 10.5 kcal/mol, $\Delta S_u^\dagger$ = 15 kcal/K/mol				
0	300.0	-2.25	539.64	23515.74
1	343.4	-1.273	392.6	2537.17
2	393.1	-0.155	297.34	362.62
3	450.0	1.125	233.25	66.28
Model 4				
$\Delta H_f^\dagger$ = 2 kcal/mol, $\Delta S_f^\dagger$ = -10 kcal/K/mol				
$\Delta H_u^\dagger$ = 14 kcal/mol, $\Delta S_u^\dagger$ = 20 kcal/K/mol				
0	300.0	-3	4393.38	673718.08
1	343.4	-1.698	2874.79	34604.32
2	393.1	-0.207	1984.6	2585.81
3	450.0	1.5	1435.8	268.24
Model 5				
$\Delta H_f^\dagger$ = -18 kcal/mol, $\Delta S_f^\dagger$ = -61 kcal/K/mol				
$\Delta H_u^\dagger$ = 20 kcal/mol, $\Delta S_u^\dagger$ = 50 kcal/K/mol				
0	300.0	-4.7	0.02	43.93
1	317.9	-2.71	0.09	6.648
2	336.8	-0.61	0.45	1.12
3	356.9	1.62	2.04	2.083
4	378.2	3.98	8.53	0.042
5	400.8	6.49	32.82	0.01
6	424.7	9.14	117.07	0.002
7	450.0	11.95	388.98	0.001

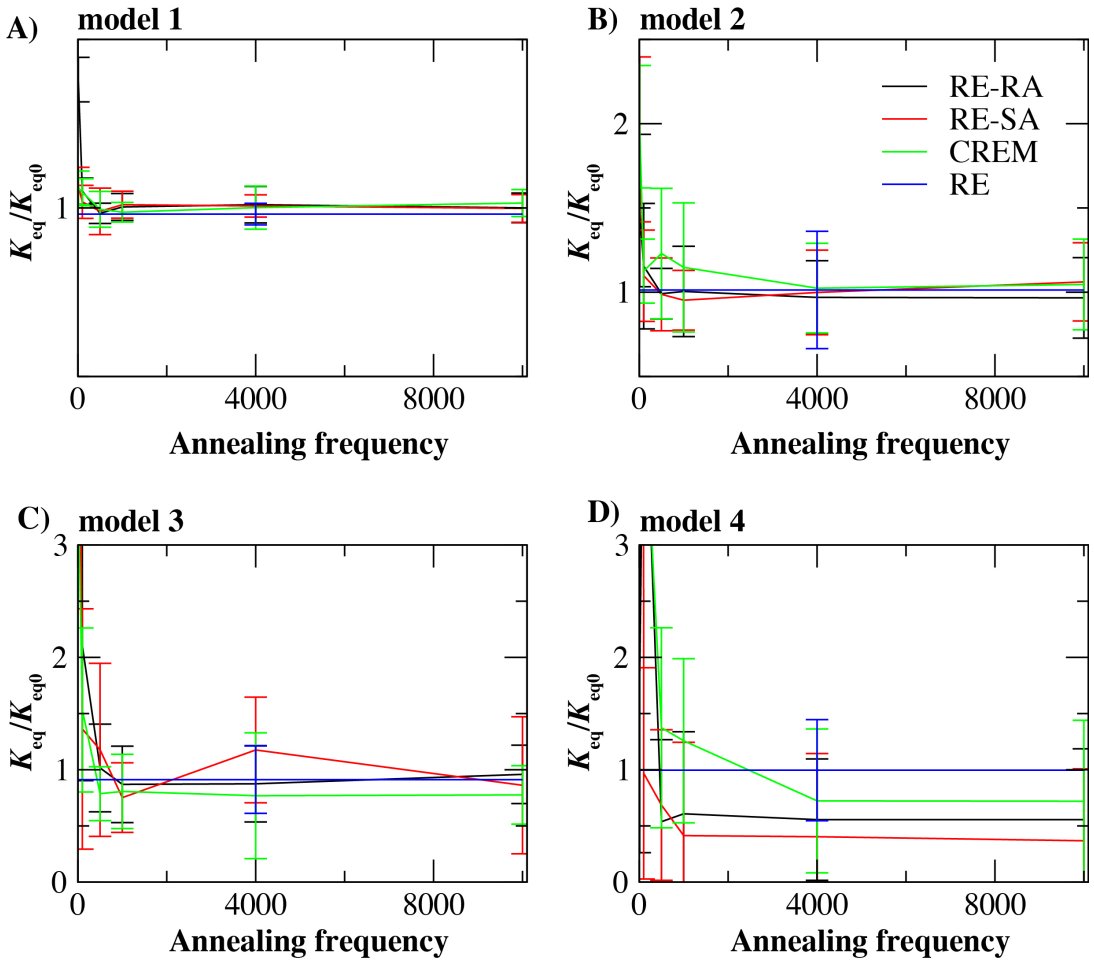


Figure S6.2: The ratio between calculated and expected equilibrium constants as a function of guided annealing frequency for different models. The error bars are not shown for clarity. Note that the RE-GA numerical simulations were not well converged within 1 ms numerical simulation timescale for model 4 due to very slow conformational transitions.

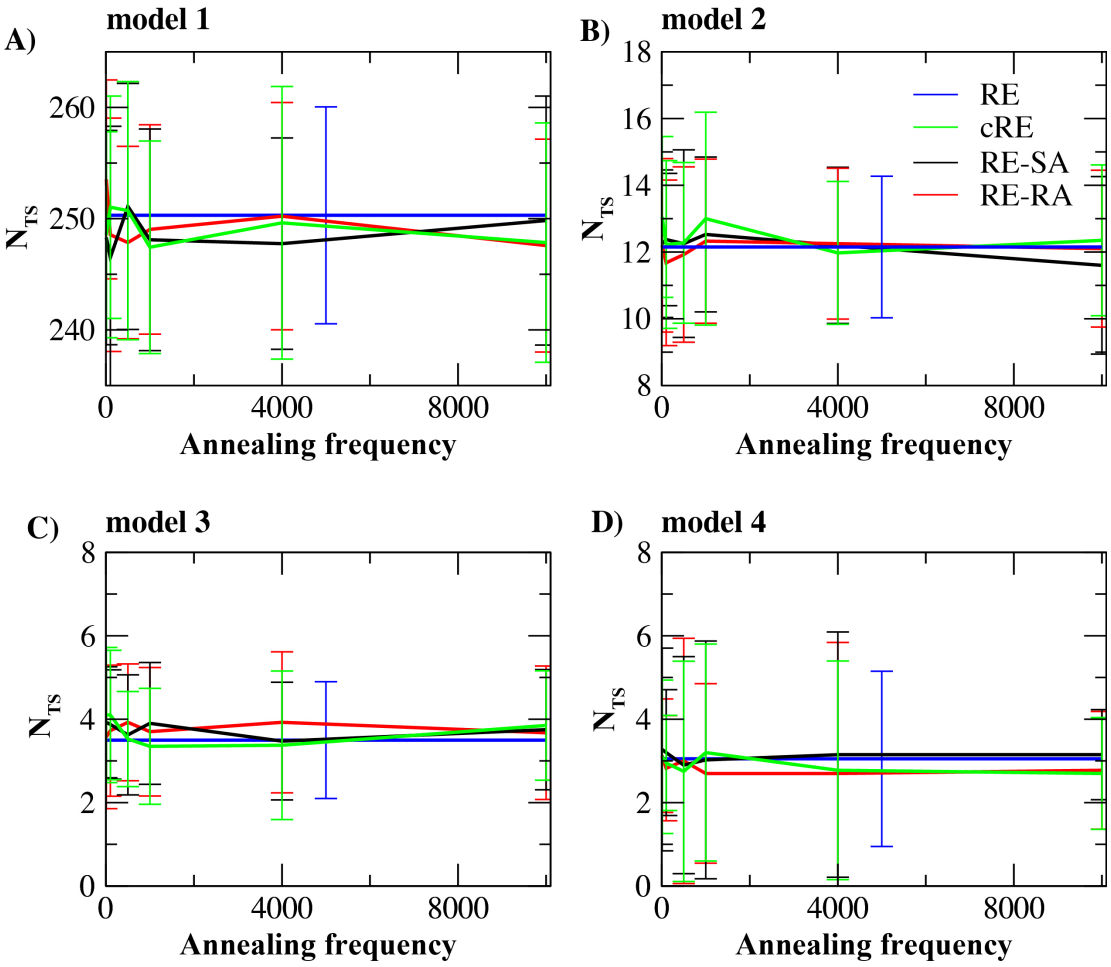


Figure S6.3: Number of reversible transitions as a function of guided annealing frequency calculated from various RE simulations using different models.

RE

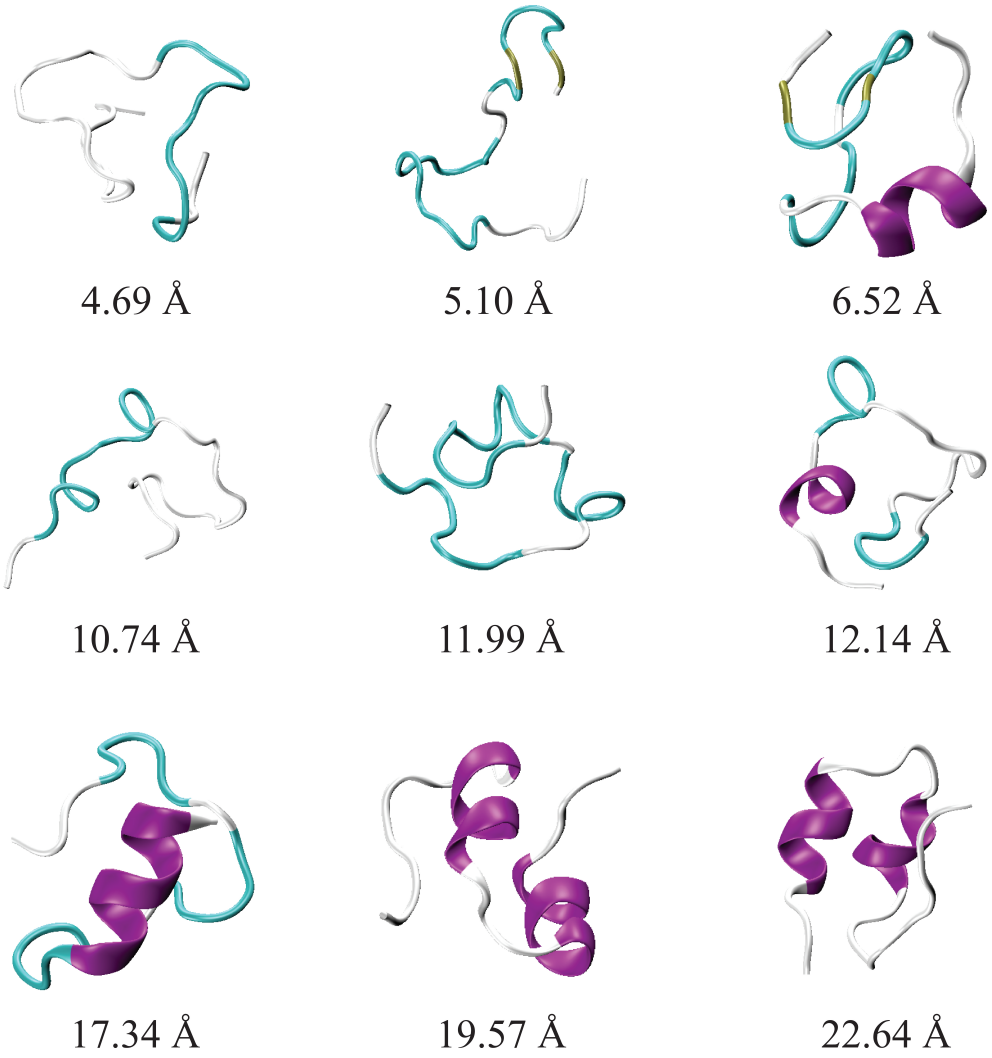


Figure S6.4: Representative structures of KID domain from RE simulation. Corresponding end-to-end distance is listed below each structure.

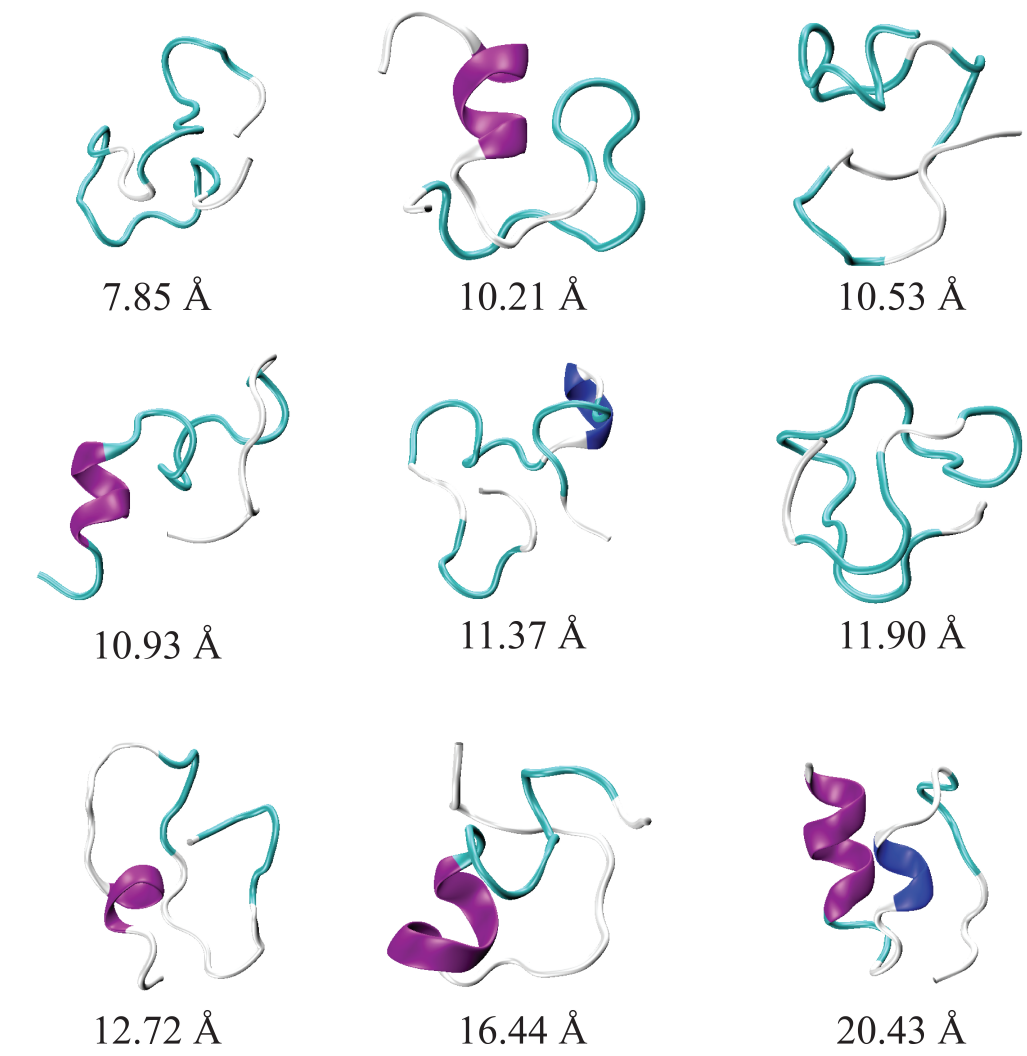


Figure S6.5: Representative structures of KID domain from RE-SA simulation. Corresponding end-to-end distance is listed below each structure.

Chapter 7

Accelerate Sampling in Atomistic Energy Landscapes Using Topology-based Coarse-grained Models*

Abstract

We describe a multi-scale enhanced sampling (MSES) method where efficient topology-based coarse-grained models are coupled with all-atom ones to enhance the sampling of atomistic protein energy landscape. The bias from the coupling is removed by Hamiltonian replica exchange, thus allowing one to benefit simultaneously from faster transitions of coarse-grained modeling and accuracy of atomistic force fields. The method is demonstrated by calculating the conformational equilibria of several small but nontrivial β-hairpins with varied stabilities.

*Reprinted (adapted) with permission from W. Zhang and J. Chen (2014), "Accelerate sampling of atomistic energy landscapes using topology-based coarse-grained models" J. Chem. Theory Comput. 10, 918-923. Copyright (2014), American Chemical Society.
Author Contributions: Conceived and designed the experiments: JC WZ. Performed the experiments: WZ JC. Analyzed the data: WZ JC. Contributed reagents/materials/analysis tools: WZ JC. Wrote the paper: WZ JC.

7.1 Introduction

Generating statistically representative conformational ensembles remains a major challenge in atomistic simulation of biomolecules[179]. This is not only due to the large and complex conformational space, but also because of significant energy barriers that frequently separate different subspaces. Temperature replica exchange (T-RE)[104,177] is now widely accepted as a relatively straightforward yet powerful technique for enhanced sampling. Multiple replicas of the system are simulated independently at different temperatures, and periodically attempt to exchange simulation temperatures according to a Metropolis criterion that preserves the detailed balance. The resulting random walk in the temperature space helps each replica to escape local energy minima and thus facilitate conformational sampling. Extensive theoretical and simulation studies have confirmed that T-RE enhances sampling compared to constant temperature simulations as long as the activation enthalpies of conformational transitions are positive[114,116,117,119]. Nonetheless, the efficiency of T-RE can be severely limited by the presence of sharp cooperative conformational transitions such as protein folding[120]. Importantly, this limitation cannot be overcome by various T-RE variants designed to accelerate either exchanges or diffusion in temperature space[121]. In practice, virtually all T-RE protein simulations involve exchange attempt frequencies ($\sim$ps^{-1}) that are several orders of magnitude faster than the slowest protein motions (folding; μs^{-1} or slower[249]). As such, the efficiency of T-RE sampling of large-scale conformational transitions is rarely limited by diffusion in temperature space, but mainly by the inherent rates of spontaneous processes.

Fundamentally, the limited efficiency of T-RE in sampling cooperative transitions such as protein folding can be attributed to large entropic components in the free energy barriers[132,133]. The folding rate only depends weakly on temperature and often displays anti-Arrhenius behaviors[114,134], such that tempering is ineffective in driving transitions. Instead, coarse-grained (CG) models[304] are often utilized to significantly reduce the conformational space and allow faster reversible

transitions; but this is achieved at the expense of reduced detail and accuracy. An ideal approach could involve CG modeling to overcome major entropic barriers and at the same time seamlessly propagates the transitions to atomistic simulations for detailed sampling of different conformational subspaces. A key requirement is that one must be able to recover canonical ensembles at the atomistic level. The promise of such multi-scale enhanced sampling (MSES) has been well recognized, and several clever ideas have been proposed towards this goal[125,135–139]. The resolution exchange approach is particularly interesting[137]. It involves independent simulations of the system at two or more resolutions and attempts to directly swap shared coordinates of low- and high-resolution models according to Metropolis criteria. Resolution exchange in principle allows the possibility of injecting high-resolution simulation into novel conformational subspaces sampled at low-resolution. A key limitation is that conformations sampled at different resolutions must be similar to be exchangeable for large biomolecules, as excessive bad contacts would effectively prohibit successful exchange. A smart resolution replica exchange was proposed to improve exchange acceptance, where a ladder of mixed atomistic and CG potentials are used and CG-sampled configurations are relaxed prior to exchange attempt[305]. However, the detailed balance is broken due to relaxation and only canonical sampling is achieved. More importantly, the aggressive approach of direct coordinate swap in resolution exchange, with or without relaxation, requires configurations sampled at different resolutions to closely track each other, which counteracts the purpose of rapid hopping among drastically different conformational subspaces sampled at the CG level.

The efficacy of an MSES scheme, regardless of how sampling at different resolutions is coupled, depends critically on the ability of CG simulations to generate transitions that are consistent with inherent conformational dynamics of the atomistic model. Otherwise, CG simulations would attempt to engage the atomistic model along unfavorable pathways and become ineffective in accelerating atomistic transitions. For proteins, topology-based CG modeling has been highly successful and demonstrated impressive correspondence between experiment and theory for folding mechanisms[229]. This approach is based on to the minimal frustration theory of protein folding[98],

which argues that native interactions dictate the protein free energy landscape and that stabilization due to non-native contacts ("frustration") should be minimal. Accordingly, the true protein energy landscape can be approximated by effective energy functions that only include native interactions. These effective energy functions, commonly referred to as Gō or Gō-like models, are highly efficient, and yet powerful enough to generate realistic reversible folding pathways. As such, they should be ideally suitable for driving conformational sampling on atomistic protein energy landscapes. Here, we describe an MSES approach that utilizes the efficient sequence-flavored Gō-like model[232] to accelerate the sampling of atomistic protein conformational equilibria, and demonstrate its efficiency using a series of small but nontrivial β-hairpins with varied stabilities.

7.2 Methods

7.2.1 Multi-scale enhanced sampling (MSES) method

Our approach is inspired by Moritsugu et al.'s multi-scale essential sampling method[139], where both the CG and atomistic representations of the protein are simulated simultaneously in a hybrid system. The atomistic and CG copies do not interact, and are only coupled through restraint potentials designed to restrict the divergence between CG and atomistic configurations. The overall potential function of the hybrid system is

$$U_{mix}(r_{AT}, r_{CG}, \lambda) = U_{AT}(r_{AT}) + U_{AT}(r_{AT}) + \lambda U_{MSES}(r_{AT}, r_{CG}), \tag{7.1}$$

where U_{AT} and U_{CG} are the atomistic and CG potential functions, respectively. Only the coupling potential U_{MSES} depends on both atomistic (r_{AT}) and CG (r_{CG}) coordinates. Given a proper coupling potential, the atomistic and CG copies can be restrained to track one and another when

the overall coupling scaling factor λ = 1. Hamiltonian RE can be performed to communicate coupled conformational dynamics to the limit of λ = 0, where the CG and atomistic copies are fully independent and proper canonical ensembles are generated simultaneously at both resolutions. Our MSES approach incorporates both Hamiltonian and temperature RE to further accelerate conformational sampling. Specifically, N replicas of the hybrid system are simulated independently with increasing couple scaling factors and temperatures, $\{\lambda_i, T_i\}$, i = 1, 2, ..., N. λ_1 = 0 and λ_N = 1. Replicas periodically attempt to exchange simulation conditions according to

$$P_{m\leftrightarrow n} = min(1, e^{\Delta_{mn}}), \tag{7.2}$$

where $\Delta_{mn} = \beta_m[U_{mix}(r_{AT,m}, r_{CG,m}, \lambda_m) - U_{mix}(r_{AT,n}, r_{CG,n}, \lambda_m)] + \beta_n[U_{mix}(r_{AT,n}, r_{CG,n}, \lambda_n) - U_{mix}(r_{AT,m}, r_{CG,m}, \lambda_n)]$, and $\beta = 1/k_BT$ (k_B is the Boltzmann constant). When the same temperature is used for all replicas, the exchange probability is determined by the coupling term only, allowing excellent scalability to large systems[139].

Coupling the CG and atomistic models using restraint potentials is a significant advantage compared to direct coordinate swapping in resolution exchange. It allows one to control the impacts of large divergences between CG and atomistic configurations on the total energy, which dramatically improves exchange efficiency and provides superior scalability to large systems. Motivated by the notion that native contacts dictate protein folding transitions[98], the CG and atomistic copies are coupled by penalty functions that depend on the differences in Cα-Cα distances of residues involved in all native contacts:

$$U_{MSES} = \sum E_i(\Delta d_i) = \sum \frac{1}{2} k_i (d_i^{AT} - d_i^{CG})^2, \tag{7.3}$$

where $\Delta d_i = d_i^{AT} - d_i^{CG}$,and d_i^{AT} and d_i^{CG} are the Cα-Cα distances of the ith native contact in

atomistic and CG copies, respectively. The force constant, k_i, can be different for various subsets of native contacts, for example, to emphasize the relative importance of tertiary contacts vs. local secondary ones. For large proteins with hundreds of native contacts, the simple harmonic potential of Equation 7.3 can lead to large penalty energies and result in severe exchange bottlenecks, particularly between the uncoupled (λ_1 = 0) and coupled ($\lambda > 0$) conditions. For this, a restraint potential with a soft asymptote is used at large Δd_i,

$$E_i(\Delta d_i) = A + \frac{B}{(\Delta d_i)^s} + f_{max}\Delta d_i, \text{ if } \Delta d_i > d_s. \tag{7.4}$$

In Equation 7.4, d_s is the distance threshold where the penalty function $E_i(\Delta d_i)$ switches from the harmonic form (Equation 7.3) to the soft asymptote. The switching exponent s controls how quickly the limiting slope, f_{max}, is approached at large Δd_i. The parameters A and B are identified by requiring both $E_i(\Delta d_i)$ and its first derivative to be continuous at the switching distance ($\Delta d_i = d_s$). We note that similar penalty functions with soft asymptotes are widely employed in NMR structure calculations[306]. The purpose is to avoid premature structural collapse due to a few large distance restraint violations and thus allow better conformational sampling to generate structures that are maximally consistent with all structural restraints. The simulations in the current work only involve exchanging a ladder of models with different λ_i. More sophisticated Hamiltonian RE schemes can be devised that involve additional parameters of the restraint potentials, particularly k_i, d_s and f_{max} , to more carefully control how conformational transitions are communicated between CG and atomistic copies for larger proteins.

7.2.2 Simulation protocols

The MSES method has been implemented in CHARMM[216] and MMTSB[215]. Its efficacy is examined here using a series of β-hairpins derived from the protein G B1 domain, including GB1p

(GEWTYD DATK TFTVTE), GB1m1 (GEWTYD DATK T**A**TVTE), and, GB1m3 (KKWTYN **P**ATG KFTVQE) (loop regions underlined and key mutations highlighted in bold fonts). The wild-type GB1p is ∼42% folded at 278K based on NMR chemical shift analysis[307]. The Phe to Ala mutation in GB1m1 reduces the hairpin stability to ∼6% folded, and the more rigid proline-containing loop increases the stability of GB1m3 to ∼86% folded[307]. These β-hairpins, albeit small, resemble larger proteins in many essential aspects including cooperative folding transitions and microsecond folding timescales. The optimized generalized born with smooth switching (GBSW) atomistic implicit solvent force field appears to recapitulate both the structures and stabilities of these β-hairpins[77]. However, the previous T-RE simulations failed to generate converged conformational ensembles for GB1p[77]. In fact, no reversible folding/unfolding transition was sampled by T-RE for any of these β-hairpins, and the apparent convergence for GB1m1 and GB1m3 ensembles was mainly due to mixing of conformational states sampled at different temperatures.

A sequence-flavored Gō-like model was first generated by the MMTSB Gō-Model Builder[232] and then used in simulations of all three hairpins. The model represents each residue using a single Cα bead and adopts the Miyazawa-Jernigan (MJ) statistical potential[234] for residue-specific native interactions. The Gō-like model was coupled to the GBSW atomistic implicit solvent model (Figure 7.1a), by imposing the restraint potentials on all nine native contacts (see Equation 7.3-7.4). All MSES simulations were performed using 8 replicas with (λ_i, T_i)= (0, 270), (0.05, 290), (0.1, 312), (0.25, 336), (0.4, 361), (0.6, 388), (0.8, 418), and (1.0, 450). These conditions were assigned by having roughly exponential distributions for both λ and T. The exchange acceptance rates were uniform and ∼25% for all MSES simulations. Other parameters of the MSES restraint potential were: k = 1.0 kcal/mol/Å^2, s = 1, d_s = 2.0 Å and f_{max} = 0.1 kcal/mol/Å. We note that the soft asymptote has minimal impacts on the exchange and sampling efficiencies for these small hairpins. Langevin simulations were performed with a friction coefficient of 0.1 ps^{-1}, and exchanges were attempted every 2 ps. For each peptide, two independent simulations were performed, starting from the folded (control) and fully extended (folding) structures, respectively. The length of

all MSES simulations was 100 ns per replica. Control and folding T-RE simulations were also performed for GB1p as a reference. These simulations are summarized in Table 7.1.

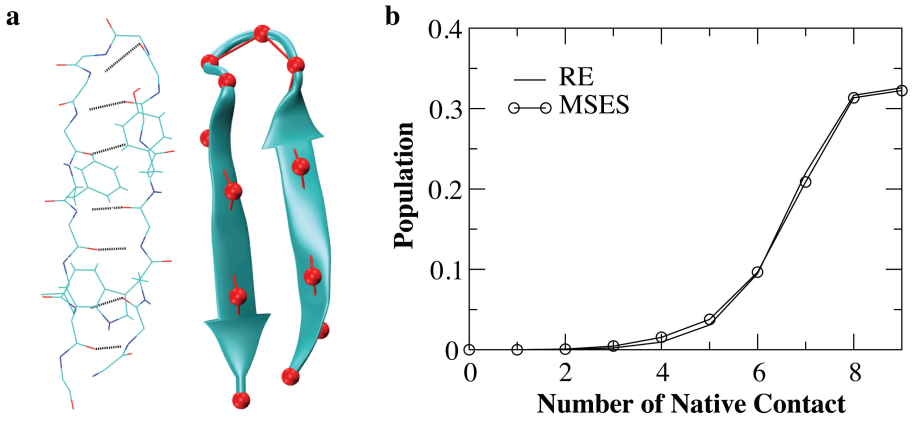


Figure 7.1: Folded structure and simulated ensembles of GB1p. (a) Left: Atomistic structure of GB1p. Dashed lines indicate all seven native hydrogen bonds. Right: Hybrid model containing atomistic (cyan cartoon) and Cα-only CG (red beads) copies. (b) Probability distributions of the number of native contacts for CG ensembles derived from a reference T-RE simulation of the Gō model alone (solid line) and the MSES control simulation of GB1p (solid line with circles).

7.3 Results and Discussion

7.3.1 MSES simulations recover the correct canonical ensembles

We first validate if MSES can indeed recover the correct canonical ensembles at the uncoupled condition. Reliable reference atomistic ensembles are not available for these hairpins due to apparent difficulty in achieving convergence using T-RE simulations[77]. Therefore, we compare the CG ensembles derived from MSES simulations to a reference ensemble obtained from a 4-μs T-RE simulation of the Gō-like model alone. Over 1300 reversible folding transitions were sampled in the T-RE simulation and the resulting reference CG ensemble at 270 K is fully converged. As illustrated in Fig 1b, MSES simulations generated CG ensembles virtually identical to the reference. That is, bias due to coupling between atomistic and CG models is completely removed

Table 7.1: Summary of T-RE and MSES simulations of GB1p, GB1m1 and GB1m3. All simulations involve 8 replicas and last 100 ns. N_{TS} is the number of reversible folding transitions sampled by all replicas during the entire course of the simulation. The effective reversible transition rate k_{TS} is calculated as N_{TS} divided by total simulation time (800 ns in all cases).

Sequence	**Protocol**	**Run**	N_{TS}	k_{TS}(ns^{-1})
GB1p	T-RE	Control	1	0.0013
		Folding	0	-
GB1p	MSES	Control	43	0.054
		Folding	33	0.041
GB1m1	MSES	Control	32	0.04
		Folding	20	0.025
GB1m3	MSES	Control	34	0.043
		Folding	26	0.033

via Hamiltonian RE as expected. The co-evolution of the CG and atomistic copies during MSES simulations is illustrated in Figure S7.1. As designed, the CG and atomistic conformations can diverge at low temperature/weak coupling conditions, but strongly track one and another at high temperature/strong coupling conditions.

7.3.2 Efficiency of MSES simulations

The efficacy of MSES in accelerating large-scale atomistic conformational transitions is first evaluated by calculating the number of reversible transitions between the folded ($\geq$5 native hydrogen bonds) and unfolded (no hydrogen bond) states at the atomistic level. Summarized in Table 7.1, the results show that T-RE simulations rarely sampled reversible folding transitions, with only one such event observed in the control and folding runs of GB1p. In contrast, many reversible transitions ($\sim$31 on average) were sampled in MSES simulations of all β-hairpins, reflecting almost two orders of magnitude enhancement in the efficiency of sampling atomistic transitions.

Figure 7.2 depicts the evolution of the number of native hydrogen bonds from representative T-RE and MSES replicas. The T-RE replicas remain at either the folded or unfolded states throughout the simulation (Figure 7.2a); whereas all MSES replicas undergo rapid reversible transitions between the folded and unfolded states (Figure 7.2b), apparently driven by coupling to the efficient Gō-like model. To further evaluate the effects of MSES coupling on folding kinetics and efficiency, we performed multiple sets of folding simulations of GB1p at 270 K using the GBSW, Gō-like, and hybrid potentials (Equation 7.1). 100 50-ns simulations were initiated from unfolded atomistic and CG structures randomly selected from pre-generated equilibrium ensembles (see Table S7.1 and Figure S7.2). The results show that coupling with the CG model allowed the atomistic model to reach the folded state in 34% of the trajectories within <18 ns on average, while all simulations in GBSW alone were trapped in various compact states and failed to fold within 50 ns. The dramatic improvement in atomistic folding efficiency was achieved with only a moderate 50% increase in the average folding time of the CG model (from 0.83 to 1.19 ns). Curiously, we observed that the CG model actually also underwent faster reversible folding transitions in MSES simulations, with an average rate of $k_{TS} \sim 0.084$ ns^{-1} compared to $k_{TS} \sim 0.041$ ns^{-1} in the T-RE simulation of the Gō-like model alone. This is likely because that coupling with the atomistic model accelerates unfolding transitions at the CG level.

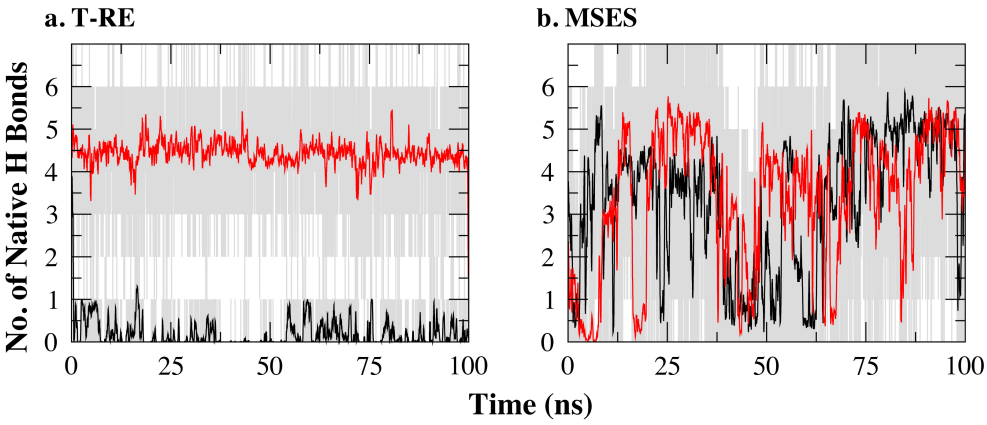


Figure 7.2: Numbers of native hydrogen bonds as a function of time for representative replicas from (a) T-RE and (b) MSES control simulations of GB1p. Gray lines depict the raw data, and black and red lines show the 50-ps running averages.

7.3.3 Convergence of MSES simulations

Ideally, a fully converged RE simulation should involve all replicas sampling the same entire accessible conformational space. This limit is almost never achieved in atomistic T-RE simulations of nontrivial peptides and proteins, where individual replicas generally sample separate major conformational states throughout the simulations (e.g., see Figure 7.3a-c). Amazingly, MSES simulations of all three β-hairpins display many characteristics of full convergence, with all replicas sampling similar, complete atomistic conformational spaces (Figure 7.3e-g and Figure S7.3). When combining information from all replicas, T-RE and MSES simulations appear to cover similar conformational spaces (Figure 7.3d and 7.3h), even though T-RE under samples most regions except the major free energy basins. Atomistic conformational ensembles derived from independent control and folding runs were compared to further evaluate the ability of MSES to achieve convergence in key thermodynamic properties. Figure 7.4 compares the probability distributions of the number of native hydrogen bonds. The results show that MSES does generate largely consistent ensembles from the control and folding runs. Especially for GB1p, large discrepancies that persist between control and folding T-RE runs (Figure 7.4a) are greatly reduced (Figure 7.4b). We also note that, although the same Gō-like model was used for all three β-hairpins, the MSES simulations were able to recapitulate varied stabilities of these sequences in the GBSW implicit solvent as expected. Nonetheless, despite impressive conformational space coverage by all replicas in MSES simulations (Figure 7.3e-g and Figure S7.3), substantial differences persist in ensembles derived from control and folding MSES runs, except for GB1m3 with a more rigid proline-containing loop. This illustrates the formidable challenges in generating converged equilibrium ensembles even for small but flexible peptides.

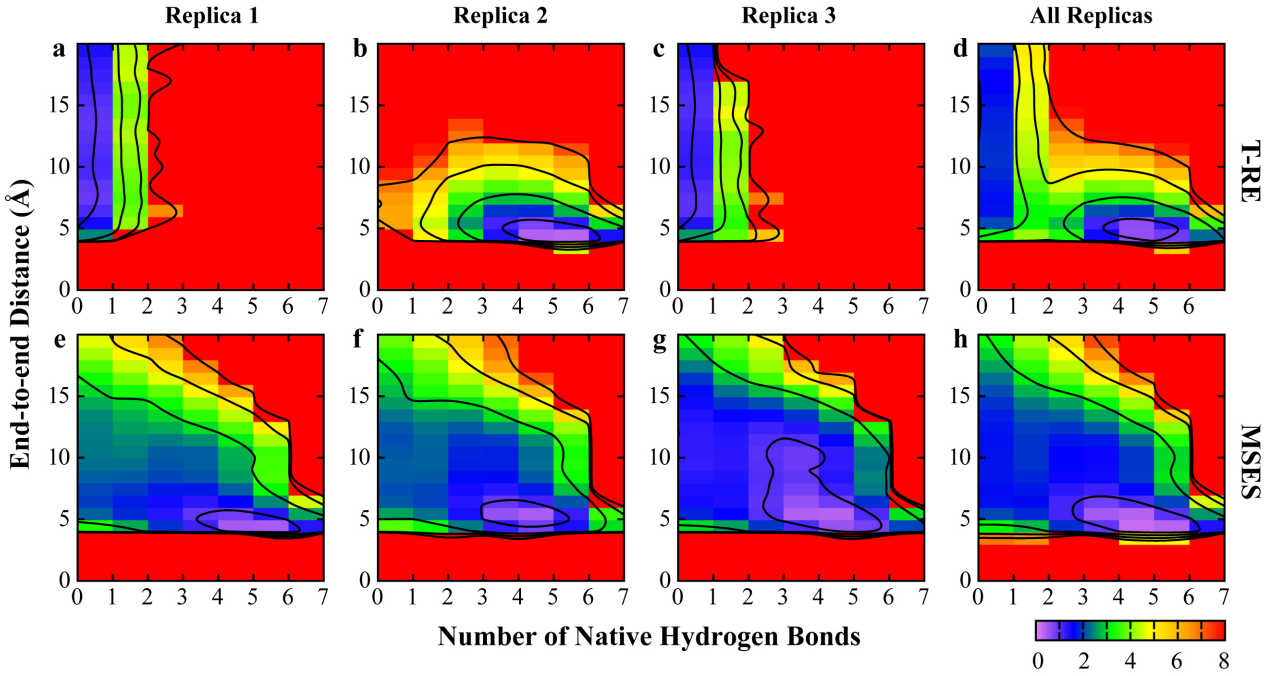


Figure 7.3: Conformational space sampled by individual replicas in the T-RE (a)-(c) and MSES (e)-(g) control simulations of GB1p. Panels (d) and (h) were calculated by including all replicas. All atomistic conformations sampled during the last 80 ns of these simulations, regardless of the temperature or coupling scaling factor, were included to compute the pseudo-free energy surfaces, which are plot in the unit of kT. Contours are shown at 1, 3, 5, and 7 kT levels.

7.4 Conclusions

In conclusion, we have developed an effective MSES approach that utilizes efficient topology-based CG models to accelerate the sampling of complex and rough atomistic energy landscapes. The CG and atomistic model are coupled using native contact-based restraint potentials that are motivated by the current understanding of protein folding mechanisms as well as lessons from NMR structure calculations and allow excellent scalability to larger and more complex systems. The bias from the coupling potential is removed by performing Hamiltonian/temperature RE, allowing one to benefit simultaneously from faster transitions of the CG model and the accuracy of the atomistic force field. Application to implicit solvent simulations of small but nontrivial GB1p series of β-hairpins demonstrates that MSES dramatically accelerate atomistic

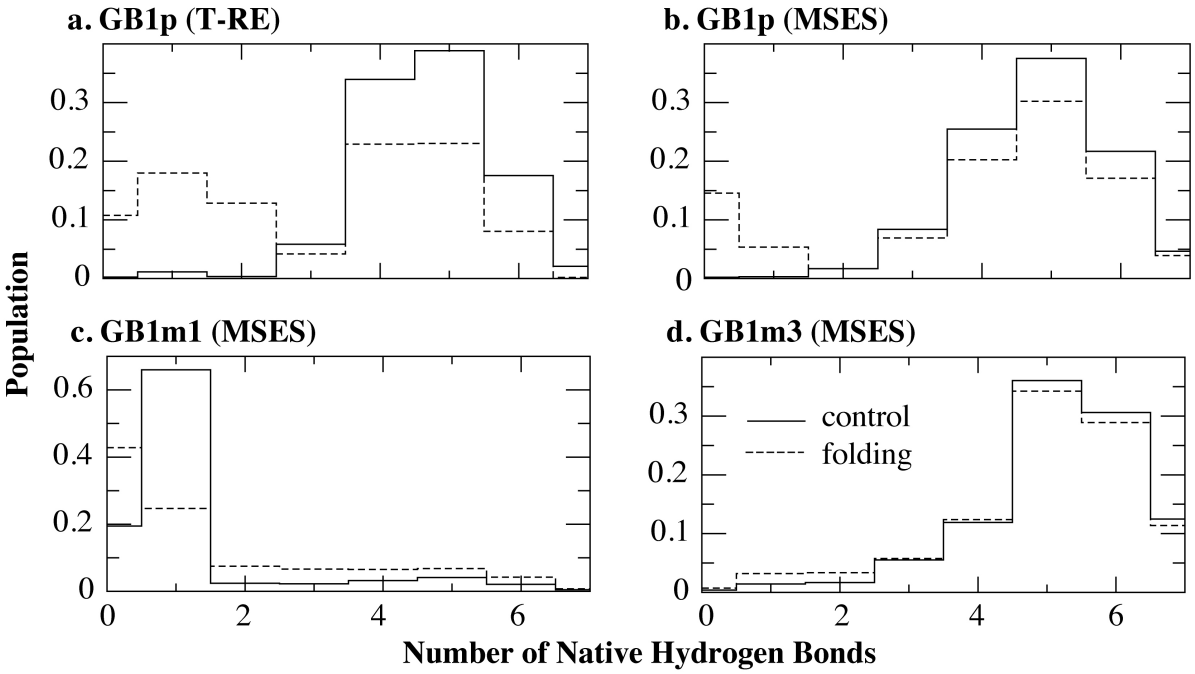


Figure 7.4: Probability distributions of the number of native hydrogen bonds for GB1p (a)-(b), GB1m1 (c), and GB1m3 (d). These distributions were calculated from structure ensembles extracted from the last 80 ns of T-RE or MSES simulations at T = 270K (and $\lambda = 0$) (see Table 7.1).

folding/unfolding transitions and improves the convergence of various thermodynamic properties of interest. We anticipate MSES to be highly useful whenever generation of well-converged protein conformational ensembles is critical, including intensive current efforts that rely on peptide simulations to optimize implicit and explicit solvent protein force fields[76,308].

7.5 Supplemental Materials

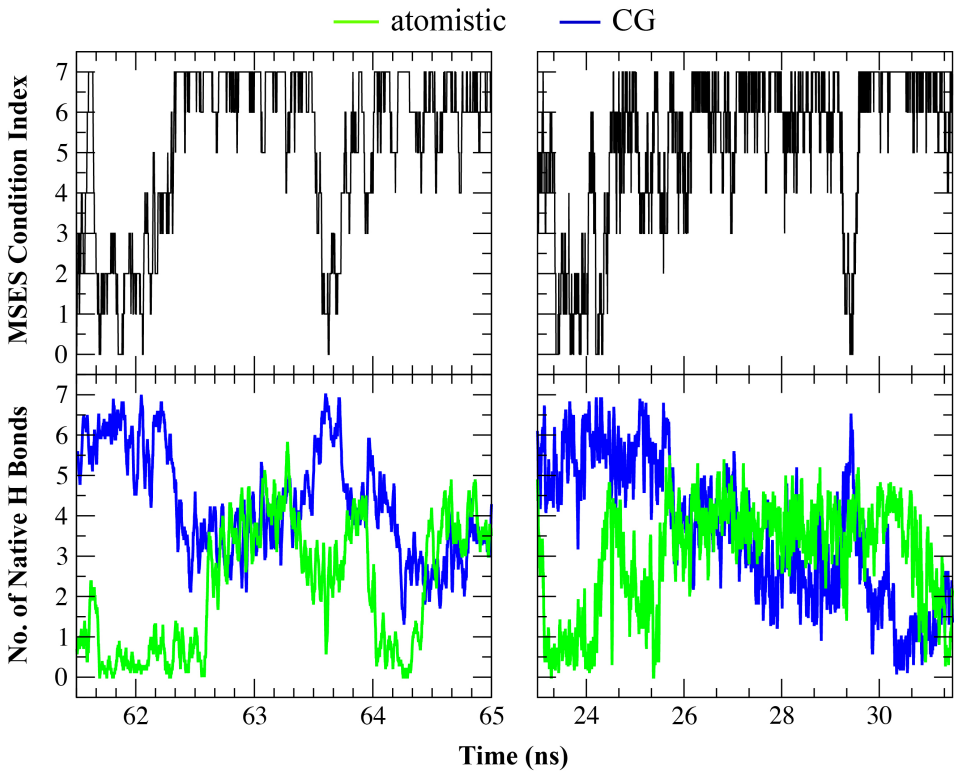


Figure S7.1: Co-evolution of the numbers of native H-bonds of the CG and atomistic copies for a representative replica from the MSES control simulation of GB1p. The black and red traces are the raw data and green and blue are 20-ps running averages in the lower panels. It illustrates that the CG model often undergoes conformational changes first, which in turn drives the atomistic model through similar transitions. It also illustrates how the CG and atomistic conformations tend to diverge at low temperature/weak coupling conditions (small indices) but strongly track one and another at high temperature/strong coupling conditions (large indices).

Table S7.1: GB1p folding efficiency and rates at 270 K using the GBSW, Gō-like, and hybrid potentials (Equation 7.1 of the main text). The MSES constant temperature runs were performed with full coupling strength (λ = 1, k = 1.0 kcal/mol/Å^2, s = 1, d_s = 2.0 Å and f_{max} = 0.1 kcal/mol/Å). The results show that MSES coupling only slightly reduces the folding kinetics of the Gō-like model (∼50%), but dramatically improve the folding rate and efficiency of the atomistic model. In contrast, none of the 100 independent atomistic simulations initiated from a representative ensemble of unfolded structures reached the folded state within the 50-ns simulation time span.

Model	Folded%	Fastest folding time (ns)	Average folding time (ns)
Gō-like alone	100	0.019	0.83
Atomistic alone	0	N/A	N/A
MSES-coupled Gō-like	100	0.021	1.19
MSES-coupled atomistic	34	0.52	17.59

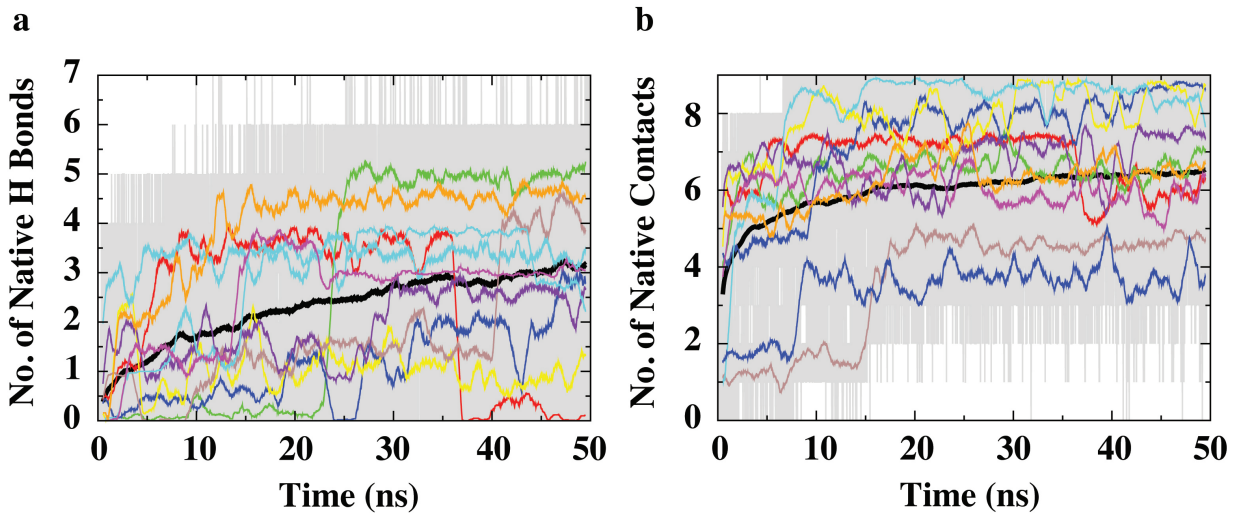


Figure S7.2: Representative trajectories from constant temperature MSES folding simulations of peptide GB1p (see Table S7.1). (a) Number of native hydrogen bonds as a function of time for the atomistic model in MSES simulation. (b) Number of native contacts as a function of time of the Gō-like model in MSES simulation. Colored lines correspond to individual folding trajectories, and the black lines are the results of averaging over all trajectories that successfully reach the folded states.

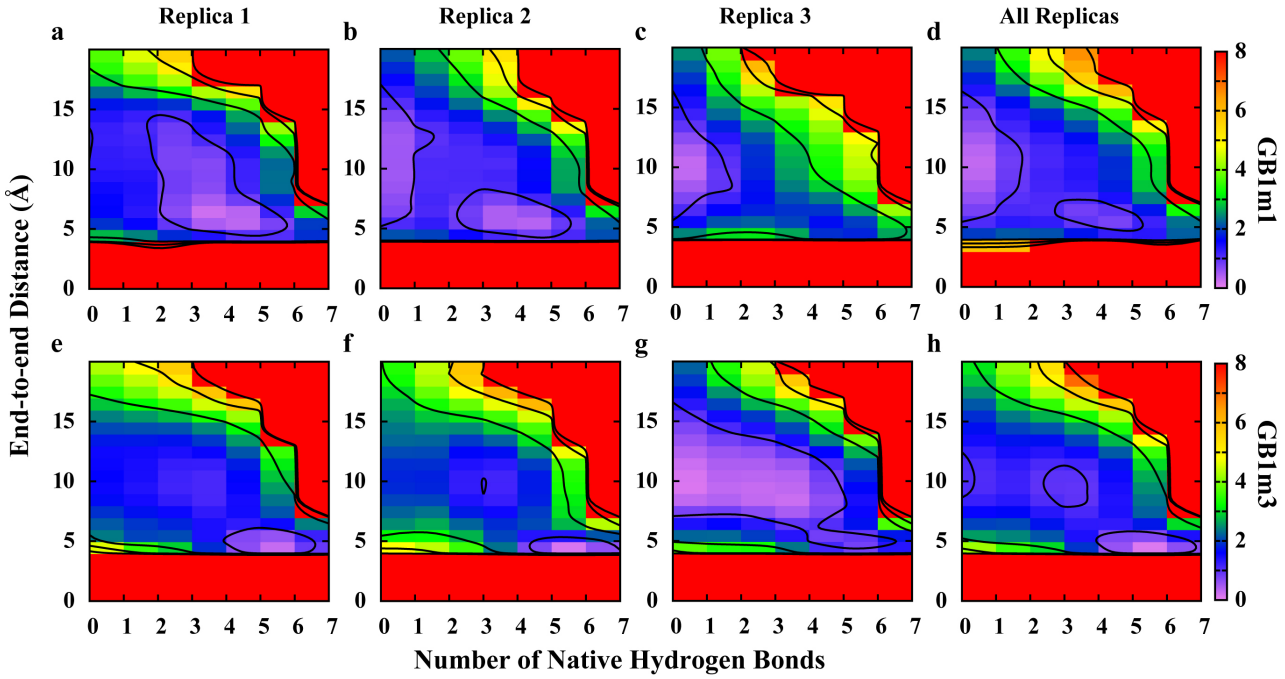


Figure S7.3: Conformational space during MSES control simulations of GB1m1 (a)-(d) and GB1m3 (e)-(h) (see Table 7.1 in the main text). Conformations sampled during the last 80 ns of these simulations, regardless of the temperature, were included to compute the pseudo-free energy surfaces, which are plot in the unit of kT. Contours are shown at 1, 3, 5, and 7 kT levels.

Chapter 8

Conclusions and Future Directions

The first section of this dissertation (chapter 2, 3 and 4) established a multi-scale molecular modeling framework for understanding the structure properties and interactions of intrinsically disordered proteins (IDPs) in general. Chapter 2 demonstrated that the implicit solvent force field, coupled with advanced sampling, provides a viable approach for atomistic simulations of IDPs. These simulations at atomistic level could offer the ultimate details of structure information of unbound IDPs. Specifically, detailed structural characterizations of disordered ensemble of unbound NCBD indicate there is a substantial amount of pre-existing folded-like populations. This observation, along with the fact that NCBD is one of the most structured IDPs, is a strong evidence for NCBD interacting with ACTR through a conformational selection mechanism. Intriguingly, direct examination of the unfolding and unbinding pathway in high-temperature simulations suggests an induced folding mechanism at baseline level. Further analysis revealed an intermediate state on the coupled binding and folding pathway that mainly involves in the C-terminal helices of NCBD and ACTR. The simulations showed in chapter 2 also reveal important limitations in both protein force field accuracy and sampling efficiency. The GBSW implicit solvent force field tends to favor non-specific collapsed states. The future directions would include optimization of the implicit solvent force field for disordered proteins and reduce the systematic bias towards compact states.

Chapter 3 utilize the topology-based coarse-grained model to directly simulate the synergistic folding of NCBD and ACTR. Through careful calibration based on additional experimental data, the topology-based model was able to properly capture the balance between the intrinsic folding propensities of NCBD and ACTR and the strength of their intermolecular interaction. The simula-

tions provided important mechanistic understandings for the coupled binding and folding between NCBD and ACTR. Both NCBD and ACTR have cooperative binding and folding which involves in a key intermediate state. In consistency with the atomistic simulation in chapter 2, the intermediate state is formed by C-terminal helices of NCBD and ACTR. More detailed analysis suggests that the binding-induced tertiary folding of NCBD follows step-wised pathway involves in both induced folding and conformational selection. A limitation of topology-based modeling is that it ignores non-specific interactions which could contribute in stabilizing non-specific encounter complexes or intermediate states.

In Chapter 4, we continue to investigate the recognition mechanisms and the roles of long-range electrostatic interactions in forming of three IDP complexes, namely, p53-TAD1/TAZ2, HIF-1α/TAZ1, and NCBD/ACTR. Using carefully calibrated topology-based coarse-grained models with explicitly added charges, we demonstrated that nonspecific electrostatic interactions not only enhance the protein-protein encounter kinetics but also promote folding-competent encounter topologies to increase the efficiency of IDP folding upon encounter. The enriched charges on IDPs not only play key roles in modulating the conformational properties of the unbound state, but also likely play general and important roles in regulating efficient interactions of IDPs with specific partners. In fact, IDPs are frequently regulated by post-translational modifications that add or remove charges. The continuing work involves in thoroughly investigation of other factors such as residual structure and mutations that could possibly impact on the kinetics or even mechanisms of coupled binding and folding of IDPs. Such knowledge will aid in assessing related human diseases and devising rational strategies to modulate IDP functions for therapeutic purposes.

This dissertation also focus on development of novel methods for enhanced sampling of protein conformations. Chapter 5 examined and benchmarked two adaptive replica exchange (RE) protocols, namely, exchange equalization (EE) and current maximization (CM). Both adaptive RE simulations aim to enhance the sampling efficiency by increase the temperature diffusion. By designing a realistic yet computationally tractable coarse-grained protein model, we performed

simulations using conventional constant temperature molecular dynamics (MD), standard REMD, EE-REMD, and CM-REMD. The results demonstrate that both EE and CM can indeed enhance temperature diffusion compared to standard RE, but both of them failed to increase the reversible folding/unfolding transitions. The convergence rates of several key thermodynamic properties are also similar among all three RE simulations. The conclusion form the benchmark is that the efficiency of RE protocols does not appear to be limited by temperature diffusion, but by the inherent rates of spontaneous large-scale conformational re-arrangements. A key lesson is that it is unlikely that any protocol aims to accelerate temperature exchange will lead to substantial enhancement in conformational sampling efficiency of REMD protein simulations.

In chapter 6, we demonstrated RE protocols that involve in guided annealing (GA) steps to overcome the apparent conformational/temperature trapping and guide the trapped replicas to visit higher temperatures more frequently. We examined the ability of various RE-GA schemes to enhance conformational sampling and generate proper canonical ensembles. Theoretical analysis suggests that introduction of GA cycles results in unbalanced exchange attempt probabilities and thus breaks the detailed balance condition whenever spontaneous conformational transitions lag temperature diffusion. This conclusion is validated by both coarse-grained protein simulations and numerical RE simulations using 2-state kinetic models. However, systematic errors introduced by GA cycles diminish for systems with low transition barriers. This observation suggests that RE-GA may be uniquely suitable for disordered protein states such as IDPs. Specifically, due to the bias of the force field, IDPs are frequently trapped in collapsed states during the RE atomistic simulations, and GA cycles should be highly effective in assisting the peptide to escape these trapped states and better sample local conformational states. Indeed, application to a model IDP, KID domain, demonstrated that RE-GA achieves superior convergence in conformational properties with merely one-fourth of simulation time compared to the standard RE. Therefore, we anticipate that RE-GA will be very useful in simulation of disordered protein states in general, despite its systematic errors for simulating folded proteins.

Chapter 7 describes a multi-scale enhanced sampling (MSES) method where efficient topology-based coarse-grained (CG) models are coupled with atomistic models to enhance the sampling of atomistic protein energy landscape. The CG and atomistic model are coupled using native contact-based restraint potentials that are motivated by the current understanding of protein folding mechanisms as well as lessons from NMR structure calculations and allow excellent scalability to larger and more complex systems. The bias from the coupling is removed by Hamiltonian replica exchange, thus allowing one to benefit simultaneously from faster transitions of coarse-grained modeling and accuracy of atomistic force fields. Application to implicit solvent simulations of small but nontrivial GB1p series of β-hairpins demonstrates that MSES dramatically accelerate atomistic folding/unfolding transitions and improves the convergence of various thermodynamic properties of interest. Ongoing work related to chapter 7 involves in application of MSES method to moderate sized/large proteins, including villin headpiece, protein A, protein GB1 and Trp-cage. The preliminary MSES simulations of larger proteins further reveal the bias towards non-specific collapsed states in implicit solvent force fields. The future work would involve in applying MSES method to enhance the sampling for force field optimizations.

References

[1] T. H. Click, D. Ganguly, and J. Chen. Intrinsically disordered proteins in a physics-based world. *International Journal of Molecular Sciences*, 11:5292–309, 2010.

[2] A. K. Dunker, I. Silman, V. N. Uversky, and J. L. Sussman. Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology*, 18:756–64, 2008.

[3] H. J. Dyson and P. E. Wright. Intrinsically unstructured proteins and their functions. *Nature Reviews: Molecular Cell Biology*, 6:197–208, 2005.

[4] J. Gsponer, M. E. Futschik, S. A. Teichmann, and M. M. Babu. Tight regulation of unstructured proteins: From transcript synthesis to protein degradation. *Science*, 322:1365–8, 2008.

[5] P. Tompa. Intrinsically unstructured proteins. *Trends in Biochemical Sciences*, 27:527–33, 2002.

[6] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293:321–31, 1999.

[7] C. Boesch, A. Bundi, M. Oppliger, and K. Wuthrich. 1h nuclear-magnetic-resonance studies of the molecular conformation of monomeric glucagon in aqueous solution. *European Journal of Biochemistry*, 91:209–14, 1978.

[8] A. K. Dunker, E. Garner, S. Guilliot, P. Romero, K. Albrecht, J. Hart, Z. Obradovic, C. Kissinger, and J. E. Villafranca. Protein disorder and the evolution of molecular recognition: Theory, predictions and observations. *Pacific Symposium on Biocomputing*, pages 473–84, 1998.

[9] E. Garner, P. Cannon, P. Romero, Z. Obradovic, and A. K. Dunker. Predicting disordered regions from amino acid sequence: Common themes despite differing structural characterization. *Genome Inform Ser Workshop Genome Inform*, 9:201–213, 1998.

[10] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker. Sequence complexity of disordered protein. *Proteins*, 42:38–48, 2001.

[11] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic. Flavors of protein disorder. *Proteins*, 52:573–84, 2003.

[12] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic. Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, 19:26–59, 2001.

[13] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology*, 337:635–45, 2004.

[14] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic. Intrinsic disorder and protein function. *Biochemistry*, 41:6573–82, 2002.

[15] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic, and A. K. Dunker. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology*, 323:573–84, 2002.

[16] M. Pajkos, B. Meszaros, I. Simon, and Z. Dosztanyi. Is there a biological cost of protein disorder? analysis of cancer-associated mutations. *Molecular BioSystems*, 8:296–307, 2012.

[17] V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Intrinsically disordered proteins in human diseases: Introducing the d2 concept. *Annual Review of Biophysics*, 37:215–46, 2008.

[18] V. Vacic and L. M. Iakoucheva. Disease mutations in disordered regions–exception to the rule? *Molecular BioSystems*, 8:27–32, 2012.

[19] K. Gunasekaran, C. J. Tsai, S. Kumar, D. Zanuy, and R. Nussinov. Extended disordered proteins: Targeting function with less scaffold. *Trends in Biochemical Sciences*, 28:81–5, 2003.

[20] V. J. Hilser and E. B. Thompson. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proceedings of the National Academy of Sciences, USA*, 104:8311–5, 2007.

[21] B. Meszaros, P. Tompa, I. Simon, and Z. Dosztanyi. Molecular principles of the interactions of disordered proteins. *Journal of Molecular Biology*, 372:549–61, 2007.

[22] C. J. Oldfield, J. Meng, J. Y. Yang, M. Q. Yang, V. N. Uversky, and A. K. Dunker. Flexible nets: Disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics*, 9 Suppl 1:S1, 2008.

[23] P. Tompa, C. Szasz, and L. Buday. Structural disorder throws new light on moonlighting. *Trends in Biochemical Sciences*, 30:484–9, 2005.

[24] C. J. Tsai, B. Ma, Y. Y. Sham, S. Kumar, and R. Nussinov. Structured disorder and conformational selection. *Proteins*, 44:418–27, 2001.

[25] H. X. Zhou. Intrinsic disorder: Signaling via highly specific but short-lived association. *Trends in Biochemical Sciences*, 37:43–8, 2012.

[26] P. E. Wright and H. J. Dyson. Linking folding and binding. *Current Opinion in Structural Biology*, 19:31–8, 2009.

[27] M. Fuxreiter, I. Simon, P. Friedrich, and P. Tompa. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *Journal of Molecular Biology*, 338:1015–26, 2004.

[28] M. Kjaergaard, K. Teilum, and F. M. Poulsen. Conformational selection in the molten globule state of the nuclear coactivator binding domain of cbp. *Proceedings of the National Academy of Sciences, USA*, 107:12535–40, 2010.

[29] S. G. Sivakolundu, D. Bashford, and R. W. Kriwacki. Disordered p27kip1 exhibits intrinsic structure resembling the cdk2/cyclin a-bound conformation. *Journal of Molecular Biology*, 353:1118–28, 2005.

[30] J. Chen. Intrinsically disordered p53 extreme c-terminus binds to s100b(betabeta) through "fly-casting". *Journal of the American Chemical Society*, 131:2088–9, 2009.

[31] R. Narayanan, O. K. Ganesh, A. S. Edison, and S. J. Hagen. Kinetics of folding and binding of an intrinsically disordered protein: The inhibitor of yeast aspartic proteinase ypra. *Journal of the American Chemical Society*, 130:11477–85, 2008.

[32] K. Sugase, H. J. Dyson, and P. E. Wright. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*, 447:1021–5, 2007.

[33] J. Wang, Y. Wang, X. Chu, S. J. Hagen, W. Han, and E. Wang. Multi-scaled explorations of binding-induced folding of intrinsically disordered protein inhibitor ia3 to its target enzyme. *PLoS Computational Biology*, 7:e1001118, 2011.

[34] W. Zhang, D. Ganguly, and J. Chen. Residual structures, conformational fluctuations, and electrostatic interactions in the synergistic folding of two intrinsically disordered proteins. *PLoS Computational Biology*, 8:e1002353, 2012.

[35] J. Chen. Towards the physical basis of how intrinsic disorder mediates protein function. *Archives of Biochemistry and Biophysics*, 524:123–31, 2012.

[36] H. J. Dyson and P. E. Wright. Nuclear magnetic resonance methods for elucidation of structure and dynamics in disordered states. *Methods in Enzymology*, 339:258–70, 2001.

[37] H. J. Dyson and P. E. Wright. Unfolded proteins and protein folding studied by nmr. *Chemical Reviews*, 104:3607–22, 2004.

[38] T. Mittag, S. Orlicky, W. Y. Choy, X. Tang, H. Lin, F. Sicheri, L. E. Kay, M. Tyers, and J. D. Forman-Kay. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proceedings of the National Academy of Sciences, USA*, 105:17772–7, 2008.

[39] D. Eliezer, J. Yao, H. J. Dyson, and P. E. Wright. Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding. *Nature Structural Biology*, 5:148–55, 1998.

[40] T. M. Logan, Y. Theriault, and S. W. Fesik. Structural characterization of the fk506 binding protein unfolded in urea and guanidine hydrochloride. *Journal of Molecular Biology*, 236:637–48, 1994.

[41] O. Zhang, L. E. Kay, J. P. Olivier, and J. D. Forman-Kay. Backbone 1h and 15n resonance assignments of the n-terminal sh3 domain of drk in folded and unfolded states using enhanced-sensitivity pulsed field gradient nmr techniques. *Journal of Biomolecular NMR*, 4:845–58, 1994.

[42] J. Zhang and C. R. Matthews. Ligand binding is the principal determinant of stability for the p21(h)-ras protein. *Biochemistry*, 37:14881–90, 1998.

[43] P. Bernado, L. Blanchard, P. Timmins, D. Marion, R. W. Ruigrok, and M. Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proceedings of the National Academy of Sciences, USA*, 102:17002–7, 2005.

[44] P. Bernado, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun. Structural characterization of flexible proteins using small-angle x-ray scattering. *Journal of the American Chemical Society*, 129:5656–64, 2007.

[45] A. C. Ferreon, C. R. Moran, Y. Gambin, and A. A. Deniz. Single-molecule fluorescence studies of intrinsically disordered proteins. *Methods in Enzymology*, 472:179–204, 2010.

[46] W. J. Greenleaf, M. T. Woodside, and S. M. Block. High-resolution, single-molecule measurements of biomolecular motion. *Annual Review of Biophysics and Biomolecular Structure*, 36:171–90, 2007.

[47] B. Schuler and W. A. Eaton. Protein folding studied by single-molecule fret. *Current Opinion in Structural Biology*, 18:16–26, 2008.

[48] A. C. Ferreon, Y. Gambin, E. A. Lemke, and A. A. Deniz. Interplay of alpha-synuclein binding and conformational switching probed by single-molecule fluorescence. *Proceedings of the National Academy of Sciences, USA*, 106:5645–50, 2009.

[49] S. Mukhopadhyay, R. Krishnan, E. A. Lemke, S. Lindquist, and A. A. Deniz. A natively unfolded yeast prion monomer adopts an ensemble of collapsed and rapidly fluctuating structures. *Proceedings of the National Academy of Sciences, USA*, 104:2649–54, 2007.

[50] A. J. Trexler and E. Rhoades. Single molecule characterization of alpha-synuclein in aggregation-prone states. *Biophysical Journal*, 99:3048–55, 2010.

[51] F. Huang, S. Rajagopalan, G. Settanni, R. J. Marsh, D. A. Armoogum, N. Nicolaou, A. J. Bain, E. Lerner, E. Haas, L. Ying, and A. R. Fersht. Multiple conformations of full-length p53 detected with single-molecule fluorescence resonance energy transfer. *Proceedings of the National Academy of Sciences, USA*, 106:20758–63, 2009.

[52] C. K. Fisher and C. M. Stultz. Constructing ensembles for intrinsically disordered proteins. *Current Opinion in Structural Biology*, 21:426–31, 2011.

[53] D. Ganguly and J. Chen. Structural interpretation of paramagnetic relaxation enhancement-derived distances for disordered protein states. *Journal of Molecular Biology*, 390:467–77, 2009.

[54] A. Besson, S. F. Dowdy, and J. M. Roberts. Cdk inhibitors: Cell cycle regulators and beyond. *Developmental Cell*, 14:159–69, 2008.

[55] D. D. Boehr, D. McElheny, H. J. Dyson, and P. E. Wright. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science*, 313:1638–42, 2006.

[56] E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, and D. Kern. Intrinsic dynamics of an enzyme underlies catalysis. *Nature*, 438:117–21, 2005.

[57] J. P. Loria, M. Rance, and 3rd Palmer, A. G. A trosy cpmg sequence for characterizing chemical exchange in large proteins. *Journal of Biomolecular NMR*, 15:151–5, 1999.

[58] T. R. Keppel, B. A. Howard, and D. D. Weis. Mapping unstructured regions and synergistic folding in intrinsically disordered proteins with amide h/d exchange mass spectrometry. *Biochemistry*, 50:8722–32, 2011.

[59] A. Bachmann, D. Wildemann, F. Praetorius, G. Fischer, and T. Kiefhaber. Mapping backbone and side-chain interactions in the transition state of a coupled protein folding and binding reaction. *Proceedings of the National Academy of Sciences, USA*, 108:3952–7, 2011.

[60] B. K. Muralidhara, R. Rathinakumar, and P. Wittung-Stafshede. Folding of desulfovibrio desulfuricans flavodoxin is accelerated by cofactor fly-casting. *Archives of Biochemistry and Biophysics*, 451:51–8, 2006.

[61] M. Onitsuka, H. Kamikubo, Y. Yamazaki, and M. Kataoka. Mechanism of induced folding:

Both folding before binding and binding before folding can be realized in staphylococcal nuclease mutants. *Proteins*, 72:837–47, 2008.

[62] D. Ganguly and J. Chen. Topology-based modeling of intrinsically disordered proteins: Balancing intrinsic folding and intermolecular interactions. *Proteins*, 79:1251–66, 2011.

[63] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.

[64] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117:5179–5197, 1995.

[65] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118:11225–11236, 1996.

[66] L. D. Schuler, X. Daura, and W. F. van Gunsteren. An improved gromos96 force field for aliphatic hydrocarbons in the condensed phase. *Journal of Computational Chemistry*, 22, 2001.

[67] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins*, 65:712–25, 2006.

[68] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins*, 78:1950–8, 2010.

[69] Jr. Mackerell, A. D., M. Feig, and 3rd Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry*, 25:1400–15, 2004.

[70] J. Mittal and R. B. Best. Dependence of protein folding stability and dynamics on the density and composition of macromolecular crowders. *Biophysical Journal*, 98:315–20, 2010.

[71] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. How robust are protein folding simulations with respect to force field parameterization? *Biophysical Journal*, 100:L47–9, 2011.

[72] A. Vitalis and R. V. Pappu. Absinth: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *Journal of Computational Chemistry*, 30:673–99, 2009.

[73] B. Roux and T. Simonson. Implicit solvent models. *Biophysical Chemistry*, 78:1–20, 1999.

[74] J. Chen and 3rd Brooks, C. L. Implicit modeling of nonpolar solvation for simulating protein folding and conformational transitions. *Physical Chemistry Chemical Physics*, 10:471–81, 2008.

[75] W. Im, M. S. Lee, and 3rd Brooks, C. L. Generalized born model with a simple smoothing function. *Journal of Computational Chemistry*, 24:1691–702, 2003.

[76] J. Chen. Effective approximation of molecular volume using atom-centered dielectric functions in generalized born models. *Advances in Protein Chemistry*, 6:2790–2803, 2010.

[77] J. Chen, W. Im, and 3rd Brooks, C. L. Balancing solvation and intramolecular interactions: Toward a consistent generalized born force field. *Journal of the American Chemical Society*, 128:3728–36, 2006.

[78] W. Im, J. Chen, and 3rd Brooks, C. L. Peptide and protein folding and conformational

equilibria: Theoretical treatment of electrostatics and hydrogen bonding with implicit solvent models. *Advances in Protein Chemistry*, 72:173–98, 2005.

[79] S. Jang, E. Kim, and Y. Pak. Direct folding simulation of alpha-helices and beta-hairpins based on a single all-atom force field with an implicit solvation model. *Proteins*, 66:53–60, 2007.

[80] A. Okur, B. Strockbine, V. Hornak, and C. Simmerling. Using pc clusters to evaluate the transferability of molecular mechanics force fields for proteins. *Journal of Computational Chemistry*, 24:21–31, 2003.

[81] J. Khandogin, J. Chen, and 3rd Brooks, C. L. Exploring atomistic details of ph-dependent peptide folding. *Proceedings of the National Academy of Sciences, USA*, 103:18546–50, 2006.

[82] E. M. Boczko and 3rd Brooks, C. L. First-principles calculation of the folding free energy of a three-helix bundle protein. *Science*, 269:393–6, 1995.

[83] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry*, 53:291–318, 2002.

[84] G. Hummer. From transition paths to transition states and rate coefficients. *Journal of Chemical Physics*, 120:516–23, 2004.

[85] A. C. Pan, D. Sezer, and B. Roux. Finding transition pathways using the string method with swarms of trajectories. *Journal of Physical Chemistry B*, 112:3432–40, 2008.

[86] A. R. Dinner and M. Karplus. Is protein unfolding the reverse of protein folding? a lattice simulation analysis. *Journal of Molecular Biology*, 292:403–19, 1999.

[87] H. F. Chen and R. Luo. Binding induced folding in p53-mdm2 complex. *Journal of the American Chemical Society*, 129:2930–7, 2007.

[88] V. Daggett. Protein folding-simulation. *Chemical Reviews*, 106:1898–916, 2006.

[89] R. D. Schaeffer, A. Fersht, and V. Daggett. Combining experiment and simulation in protein folding: Closing the gap for small model systems. *Current Opinion in Structural Biology*, 18:4–9, 2008.

[90] G. M. Verkhivker, D. Bouzida, D. K. Gehlhaar, P. A. Rejto, S. T. Freer, and P. W. Rose. Simulating disorder-order transitions in molecular recognition of unstructured proteins: Where folding meets binding. *Proceedings of the National Academy of Sciences, USA*, 100:5148–53, 2003.

[91] I. Staneva and S. Wallin. Binding free energy landscape of domain-peptide interactions. *PLoS Computational Biology*, 7:e1002131, 2011.

[92] D. De Sancho and R. B. Best. Modulation of an idp binding mechanism and rates by helix propensity and non-native interactions: Association of hif1alpha with cbp. *Molecular BioSystems*, 8:256–67, 2012.

[93] A. G. Turjanski, J. S. Gutkind, R. B. Best, and G. Hummer. Binding-induced folding of a natively unstructured transcription factor. *PLoS Computational Biology*, 4:e1000060, 2008.

[94] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254:1598–603, 1991.

[95] N. Go. Theoretical studies of protein folding. *Annual Review of Biophysics and Bioengineering*, 12:183–210, 1983.

[96] C. Clementi. Coarse-grained models of protein folding: Toy models or predictive tools? *Current Opinion in Structural Biology*, 18:10–5, 2008.

[97] K. A. Dill and H. S. Chan. From levinthal to pathways to funnels. *Nature Structural Biology*, 4:10–9, 1997.

[98] P. G. Wolynes. Recent successes of the energy landscape theory of protein folding and function. *Quarterly Reviews of Biophysics*, 38:405–10, 2005.

[99] A. Azia and Y. Levy. Nonnative electrostatic interactions can modulate protein folding: Molecular dynamics with a grain of salt. *Journal of Molecular Biology*, 393:527–42, 2009.

[100] A. Liwo, C. Czaplewski, S. Oldziej, and H. A. Scheraga. Computational techniques for efficient conformational sampling of proteins. *Current Opinion in Structural Biology*, 18:134–9, 2008.

[101] B. J. Alder and T. E. Wainwright. Studies in molecular dynamics. i. general method. *Journal of Chemical Physics*, 31:459–466, 1959.

[102] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. N. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087, 1953.

[103] K. Hukushima and K. Nemoto. Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65:1604–1608, 1996.

[104] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314:141–151, 1999.

[105] R. H. Swendsen and J. S. Wang. Replica monte carlo simulation of spin glasses. *Physical Review Letters*, 57:2607–2609, 1986.

[106] D. J. Earl and M. W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7:3910–6, 2005.

[107] E. Gallicchio, R. M. Levy, and M. Parashar. Asynchronous replica exchange for molecular simulations. *Journal of Computational Chemistry*, 29:788–94, 2008.

[108] H. Nymeyer, S. Gnanakaran, and A. E. Garcia. Atomic simulations of protein folding, using the replica exchange algorithm. *Methods in Enzymology*, 383:119–49, 2004.

[109] A. E. Roitberg, A. Okur, and C. Simmerling. Coupling of replica exchange simulations to a non-boltzmann structure reservoir. *Journal of Physical Chemistry B*, 111:2415–8, 2007.

[110] W. Zhang, C. Wu, and Y. Duan. Convergence of replica exchange molecular dynamics. *Journal of Chemical Physics*, 123:154105, 2005.

[111] R. Denschlag, M. Lingenheil, and P. Tavan. Efficiency reduction and pseudo- convergence in replica exchange sampling of peptide folding-unfolding equilibria. *Chemical Physics Letters*, 458:244–248, 2008.

[112] H. Nymeyer. How efficient is replica exchange molecular dynamics? an analytic approach. *Journal of Chemical Theory and Computation*, 4:626–636, 2008.

[113] C. Predescu, M. Predescu, and C. V. Ciobanu. On the efficiency of exchange in parallel tempering monte carlo simulations. *Journal of Physical Chemistry B*, 109:4189–96, 2005.

[114] W. Zheng, M. Andrec, E. Gallicchio, and R. M. Levy. Simulating replica exchange simulations of protein folding with a kinetic network model. *Proceedings of the National Academy of Sciences, USA*, 104:15340–5, 2007.

[115] W. Zheng, M. Andrec, E. Gallicchio, and R. M. Levy. Simple continuous and discrete models for simulating replica exchange simulations of protein folding. *Journal of Physical Chemistry B*, 112:6083–93, 2008.

[116] D. M. Zuckerman and E. Lyman. A second look at canonical sampling of biomolecules using replica exchange simulation. *Journal of Chemical Theory and Computation*, 2:12001202, 2006.

[117] X. Periole and A. E. Mark. Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. *Journal of Chemical Physics*, 126:014903, 2007.

[118] F. Rao and A. Caflisch. Replica exchange molecular dynamics simulations of reversible folding. *Journal of Chemical Physics*, 119:4035–4042, 2003.

[119] D. Sindhikara, Y. Meng, and A. E. Roitberg. Exchange frequency in replica exchange molecular dynamics. *Journal of Chemical Physics*, 128:024103, 2008.

[120] J. Kim, T. Keyes, and J. E. Straub. Generalized replica exchange method. *Journal of Chemical Physics*, 132:224107, 2010.

[121] W. Zhang and J. Chen. Efficiency of adaptive temperature-based replica exchange for sampling large-scale protein conformational transitions. *Journal of Chemical Theory and Computation*, 9:2849–2856, 2013.

[122] X. Li, C. P. O'Brien, G. Collier, N. A. Vellore, F. Wang, R. A. Latour, D. A. Bruce, and S. J. Stuart. An improved replica-exchange sampling method: Temperature intervals with global energy reassignment. *Journal of Chemical Physics*, 127:164116, 2007.

[123] A. J. Ballard and C. Jarzynski. Replica exchange with nonequilibrium switches. *Proceedings of the National Academy of Sciences, USA*, 106:12224–9, 2009.

[124] X. Cheng, G. Cui, V. Hornak, and C. Simmerling. Modified replica exchange simulation methods for local structure refinement. *Journal of Physical Chemistry B*, 109:8220–30, 2005.

[125] P. Liu, B. Kim, R. A. Friesner, and B. J. Berne. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. *Proceedings of the National Academy of Sciences, USA*, 102:13749–54, 2005.

[126] A. Okur, L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak, and C. Simmerling. Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *Journal of Chemical Theory and Computation*, 2:420–433, 2006.

[127] X. Huang, M. Hagen, B. Kim, R. A. Friesner, R. Zhou, and B. J. Berne. Replica exchange with solute tempering: Efficiency in large scale systems. *Journal of Physical Chemistry B*, 111:5405–10, 2007.

[128] M. S. Lee and M. A. Olson. Comparison of two adaptive temperature-based replica exchange methods applied to a sharp phase transition of protein unfolding-folding. *Journal of Chemical Physics*, 134:244111, 2011.

[129] D. A. Kofke. Comment on "the incomplete beta function law for parallel tempering sampling of classical canonical systems" [j. chem. phys. 120, 4119 (2004)]. *Journal of Chemical Physics*, 121:1167, 2004.

[130] S. Trebst, M. Troyer, and U. H. Hansmann. Optimized parallel tempering simulations of proteins. *Journal of Chemical Physics*, 124:174903, 2006.

[131] Y. G. Spill, G. Bouvier, and M. Nilges. A convective replica-exchange method for sampling new energy basins. *Journal of Computational Chemistry*, 34:132–40, 2013.

[132] D. J. Bicout and A. Szabo. Entropic barriers, transition states, funnels, and exponential protein folding kinetics: A simple model. *Protein Science*, 9:452–65, 2000.

[133] B. Schuler, E. A. Lipman, and W. A. Eaton. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature*, 419:743–7, 2002.

[134] D. A. Beck, G. W. White, and V. Daggett. Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations. *Journal of Structural Biology*, 157:514–23, 2007.

[135] W. Kwak and U. H. Hansmann. Efficient sampling of protein structures by model hopping. *Physical Review Letters*, 95:138102, 2005.

[136] T. Z. Lwin and R. Luo. Overcoming entropic barrier with coupled sampling at dual resolutions. *Journal of Chemical Physics*, 123:194904, 2005.

[137] E. Lyman, F. M. Ytreberg, and D. M. Zuckerman. Resolution exchange simulation. *Physical Review Letters*, 96:028105, 2006.

[138] J. H. Meinke and U. H. Hansmann. Protein simulations combining an all-atom force field with a go term. *Journal of Physics: Condensed Matter*, 19:285215, 2007.

[139] K. Moritsugu, T. Terada, and A. Kidera. Scalable free energy calculation of proteins via multiscale essential sampling. *Journal of Chemical Physics*, 133:224105, 2010.

[140] B. Zagrovic, E. J. Sorin, and V. Pande. Beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *Journal of Molecular Biology*, 313:151–69, 2001.

[141] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation - umbrella sampling. *Journal of Computational Physics*, 23:187–199, 1977.

[142] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules .1. the method. *Journal of Computational Chemistry*, 13:1011–1021, 1992.

[143] D. Hamelberg, J. Mongan, and J. A. McCammon. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *Journal of Chemical Physics*, 120:11919–29, 2004.

[144] A. Laio and M. Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences, USA*, 99:12562–6, 2002.

[145] J. Lee, H. A. Scheraga, and S. Rackovsky. New optimization method for conformational energy calculations on polypeptides: Conformational space annealing. *Journal of Computational Chemistry*, 18:1222–1232, 1997.

[146] R. G. Smock and L. M. Gierasch. Sending signals dynamically. *Science*, 324:198–203, 2009.

[147] V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Showing your id: Intrinsic disorder as an id for recognition, regulation and cell signaling. *Journal of Molecular Recognition*, 18:343–84, 2005.

[148] S. J. Metallo. Intrinsically disordered proteins are potential drug targets. *Current Opinion in Chemical Biology*, 14:481–8, 2010.

[149] A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, and V. N. Uversky. Flexible nets. the roles of intrinsic disorder in protein interaction networks. *FEBS Journal*, 272:5129–48, 2005.

[150] R. L. Baldwin. The search for folding intermediates and the mechanism of protein folding. *Annual Review of Biophysics*, 37:1–21, 2008.

[151] A. I. Bartlett and S. E. Radford. An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. *Nature Structural and Molecular Biology*, 16:582–8, 2009.

[152] G. R. Bowman, V. A. Voelz, and V. S. Pande. Taming the complexity of protein folding. *Current Opinion in Structural Biology*, 21:4–11, 2011.

[153] C. M. Dobson. Protein folding and misfolding. *Nature*, 426:884–90, 2003.

[154] C. Bracken, L. M. Iakoucheva, P. R. Romero, and A. K. Dunker. Combining prediction, computation and experiment for the characterization of protein disorder. *Current Opinion in Structural Biology*, 14:570–6, 2004.

[155] D. Eliezer. Biophysical characterization of intrinsically disordered proteins. *Current Opinion in Structural Biology*, 19:23–30, 2009.

[156] V. Receveur-Brechot, J. M. Bourhis, V. N. Uversky, B. Canard, and S. Longhi. Assessing protein disorder and induced folding. *Proteins*, 62:24–45, 2006.

[157] R. Roy, S. Hohng, and T. Ha. A practical guide to single-molecule fret. *Nature Methods*, 5:507–16, 2008.

[158] C. K. Fisher, A. Huang, and C. M. Stultz. Modeling intrinsically disordered proteins with bayesian statistics. *Journal of the American Chemical Society*, 132:14919–27, 2010.

[159] C. K. Fisher and C. M. Stultz. Protein structure along the order-disorder continuum. *Journal of the American Chemical Society*, 133:10022–5, 2011.

[160] T. Mittag and J. D. Forman-Kay. Atomic-level characterization of disordered protein ensembles. *Current Opinion in Structural Biology*, 17:3–14, 2007.

[161] T. Mittag, J. Marsh, A. Grishaev, S. Orlicky, H. Lin, F. Sicheri, M. Tyers, and J. D. Forman-Kay. Structure/function implications in a dynamic complex of the intrinsically disordered sic1 with the cdc4 subunit of an scf ubiquitin ligase. *Structure*, 18:494–506, 2010.

[162] J. Song, L. W. Guo, H. Muradov, N. O. Artemyev, A. E. Ruoho, and J. L. Markley. Intrinsically disordered gamma-subunit of cgmp phosphodiesterase encodes functionally relevant transient secondary and tertiary structure. *Proceedings of the National Academy of Sciences, USA*, 105:1505–10, 2008.

[163] D. Ganguly, W. Zhang, and J. Chen. Synergistic folding of two intrinsically disordered proteins: Searching for conformational selection. *Molecular BioSystems*, 8:198–209, 2012.

[164] J. Higo, Y. Nishimura, and H. Nakamura. A free-energy landscape for coupled folding and binding of an intrinsically disordered protein in explicit solvent from detailed all-atom computations. *Journal of the American Chemical Society*, 133:10448–58, 2011.

[165] Q. Lu, H. P. Lu, and J. Wang. Exploring the mechanism of flexible biomolecular recognition with single molecule dynamics. *Physical Review Letters*, 98:128105, 2007.

[166] B. A. Shoemaker, J. J. Portman, and P. G. Wolynes. Speeding molecular recognition by

using the folding funnel: The fly-casting mechanism. *Proceedings of the National Academy of Sciences, USA*, 97:8868–73, 2000.

[167] E. Trizac, Y. Levy, and P. G. Wolynes. Capillarity theory for the fly-casting mechanism. *Proceedings of the National Academy of Sciences, USA*, 107:2746–50, 2010.

[168] S. Brocca, L. Testa, F. Sobott, M. Samalikova, A. Natalello, E. Papaleo, M. Lotti, L. De Gioia, S. M. Doglia, L. Alberghina, and R. Grandori. Compaction properties of an intrinsically disordered protein: Sic1 and its kinase-inhibitor domain. *Biophysical Journal*, 100:2243–52, 2011.

[169] D. Ganguly and J. Chen. Atomistic details of the disordered states of kid and pkid. implications in coupled binding and folding. *Journal of the American Chemical Society*, 131:5214–23, 2009.

[170] D. P. Teufel, C. M. Johnson, J. K. Lum, and H. Neuweiler. Backbone-driven collapse in unfolded protein chains. *Journal of Molecular Biology*, 409:250–62, 2011.

[171] H. T. Tran, A. Mao, and R. V. Pappu. Role of backbone-solvent interactions in determining conformational equilibria of intrinsically disordered proteins. *Journal of the American Chemical Society*, 130:7380–92, 2008.

[172] Y. Huang and Z. Liu. Kinetic advantage of intrinsically disordered proteins in coupled folding-binding process: A critical assessment of the "fly-casting" mechanism. *Journal of Molecular Biology*, 393:1143–59, 2009.

[173] H. X. Zhou. From induced fit to conformational selection: A continuum of binding mechanism controlled by the timescale of conformational transitions. *Biophysical Journal*, 98:L15–7, 2010.

[174] D. Bashford and D. A. Case. Generalized born models of macromolecular solvation effects. *Annual Review of Physical Chemistry*, 51:129–52, 2000.

[175] J. Chen, 3rd Brooks, C. L., and J. Khandogin. Recent advances in implicit solvent-based methods for biomolecular simulations. *Current Opinion in Structural Biology*, 18:140–8, 2008.

[176] M. Feig and 3rd Brooks, C. L. Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Current Opinion in Structural Biology*, 14:217–24, 2004.

[177] U. H. E. Hansmann and Y. Okamoto. Numerical comparisons of three recently proposed algorithms in the protein folding problem. *Journal of Computational Chemistry*, 18:920–933, 1997.

[178] U. H. Hansmann and Y. Okamoto. New monte carlo algorithms for protein folding. *Current Opinion in Structural Biology*, 9:177–83, 1999.

[179] H. Lei and Y. Duan. Improved sampling methods for molecular simulation. *Current Opinion in Structural Biology*, 17:187–91, 2007.

[180] Jr. Mackerell, A. D. Empirical force fields for biological macromolecules: Overview and issues. *Journal of Computational Chemistry*, 25:1584–604, 2004.

[181] J. W. Ponder and D. A. Case. Force fields for protein simulations. *Advances in Protein Chemistry*, 66:27–85, 2003.

[182] M. Feig, A. D. Mackerell, and C. L. Brooks. Force field influence on the observation of pi- helical protein structures in molecular dynamics simulations. *Journal of Physical Chemistry B*, 107:2831–2836, 2003.

[183] A. D. Mackerell, D. Bashford, M. Bellot, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera,

D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, 102:3586–3616, 1998.

[184] Jr. MacKerell, A. D., M. Feig, and 3rd Brooks, C. L. Improved treatment of the protein backbone in empirical force fields. *Journal of the American Chemical Society*, 126:698–9, 2004.

[185] J. Khandogin and 3rd Brooks, C. L. Linking folding with aggregation in alzheimer's beta-amyloid peptides. *Proceedings of the National Academy of Sciences, USA*, 104:16880–5, 2007.

[186] J. Khandogin, D. P. Raleigh, and 3rd Brooks, C. L. Folding intermediate in the villin headpiece domain arises from disruption of a n-terminal hydrogen-bonded network. *Journal of the American Chemical Society*, 129:3056–7, 2007.

[187] R. Halfmann, S. Alberti, R. Krishnan, N. Lyle, C. W. O'Donnell, O. D. King, B. Berger, R. V. Pappu, and S. Lindquist. Opposing effects of glutamine and asparagine govern prion formation by intrinsically disordered proteins. *Molecular Cell*, 43:72–84, 2011.

[188] A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, and R. V. Pappu. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences, USA*, 107:8183–8, 2010.

[189] K. P. Wu, D. S. Weinstock, C. Narayanan, R. M. Levy, and J. Baum. Structural reorganization of alpha-synuclein at low ph observed by nmr and remd simulations. *Journal of Molecular Biology*, 391:784–96, 2009.

[190] R. H. Goodman and S. Smolik. Cbp/p300 in cell growth, transformation, and development. *Genes and Development*, 14:1553–77, 2000.

[191] N. Vo and R. H. Goodman. Creb-binding protein and p300 in transcriptional regulation. *Journal of Biological Chemistry*, 276:13505–8, 2001.

[192] C. H. Lin, B. J. Hare, G. Wagner, S. C. Harrison, T. Maniatis, and E. Fraenkel. A small domain of cbp/p300 binds diverse proteins: Solution structure and functional studies. *Molecular Cell*, 8:581–90, 2001.

[193] B. York and B. W. O'Malley. Steroid receptor coactivator (src) family: Masters of systems biology. *Journal of Biological Chemistry*, 285:38743–50, 2010.

[194] S. J. Demarest, M. Martinez-Yamout, J. Chung, H. Chen, W. Xu, H. J. Dyson, R. M. Evans, and P. E. Wright. Mutual synergistic folding in recruitment of cbp/p300 by p160 nuclear receptor coactivators. *Nature*, 415:549–53, 2002.

[195] C. W. Lee, M. A. Martinez-Yamout, H. J. Dyson, and P. E. Wright. Structure of the p53 transactivation domain in complex with the nuclear receptor coactivator binding domain of creb binding protein. *Biochemistry*, 49:9964–71, 2010.

[196] B. Y. Qin, C. Liu, H. Srinath, S. S. Lam, J. J. Correia, R. Derynck, and K. Lin. Crystal structure of irf-3 in complex with cbp. *Structure*, 13:1269–77, 2005.

[197] L. Waters, B. Yue, V. Veverka, P. Renshaw, J. Bramham, S. Matsuda, T. Frenkiel, G. Kelly, F. Muskett, M. Carr, and D. M. Heery. Structural diversity in p160/creb-binding protein coactivator complexes. *Journal of Biological Chemistry*, 281:14787–95, 2006.

[198] S. J. Demarest, S. Deechongkit, H. J. Dyson, R. M. Evans, and P. E. Wright. Packing, specificity, and mutability at the binding interface between the p160 coactivator and creb-binding protein. *Protein Science*, 13:203–10, 2004.

[199] M. O. Ebert, S. H. Bae, H. J. Dyson, and P. E. Wright. Nmr relaxation study of the complex formed between cbp and the activation domain of the nuclear hormone receptor coactivator actr. *Biochemistry*, 47:1299–308, 2008.

[200] J. Chen and 3rd Brooks, C. L. Critical importance of length-scale dependence in

implicit modeling of hydrophobic interactions. *Journal of the American Chemical Society*, 129:2444–5, 2007.

[201] A. Afek, E. I. Shakhnovich, and D. B. Lukatsky. Multi-scale sequence correlations increase proteome structural disorder and promiscuity. *Journal of Molecular Biology*, 409:439–49, 2011.

[202] D. B. Lukatsky, A. Afek, and E. I. Shakhnovich. Sequence correlations shape protein promiscuity. *Journal of Chemical Physics*, 135:065104, 2011.

[203] J. Garcia De La Torre, M. L. Huertas, and B. Carrasco. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophysical Journal*, 78:719–30, 2000.

[204] Y. Zhang, J. Zhang, and W. Wang. Atomistic analysis of pseudoknotted rna unfolding. *Journal of the American Chemical Society*, 133:6882–5, 2011.

[205] P. Csermely, R. Palotai, and R. Nussinov. Induced fit, conformational selection and independent dynamic segments: An extended view of binding events. *Trends in Biochemical Sciences*, 35:539–46, 2010.

[206] B. Ma and R. Nussinov. Enzyme dynamics point to stepwise conformational selection in catalysis. *Current Opinion in Chemical Biology*, 14:652–9, 2010.

[207] P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten. Ten-microsecond molecular dynamics simulation of a fast-folding ww domain. *Biophysical Journal*, 94:L75–7, 2008.

[208] P. L. Freddolino, S. Park, B. Roux, and K. Schulten. Force field bias in protein folding simulations. *Biophysical Journal*, 96:3772–80, 2009.

[209] G. Schreiber, G. Haran, and H. X. Zhou. Fundamental aspects of protein-protein association kinetics. *Chemical Reviews*, 109:839–60, 2009.

[210] R. B. Best and J. Mittal. Balance between alpha and beta structures in ab initio protein folding. *Journal of Physical Chemistry B*, 114:8790–8, 2010.

[211] J. Chen, 3rd Brooks, C. L., and H. A. Scheraga. Revisiting the carboxylic acid dimers in aqueous solution: Interplay of hydrogen bonding, hydrophobic interactions, and entropy. *Journal of Physical Chemistry B*, 112:242–9, 2008.

[212] M. Kang and P. E. Smith. A kirkwood-buff derived force field for amides. *Journal of Computational Chemistry*, 27:1477–85, 2006.

[213] M. S. Lee and M. A. Olson. Evaluation of poisson solvation models using a hybrid explicit/implicit solvent method. *Journal of Physical Chemistry B*, 109:5223–36, 2005.

[214] J. M. J. Swanson, J. A. Wagoner, N. A. Baker, and J. A. McCammon. Optimizing the poisson dielectric boundary with explicit solvent forces and energies: Lessons learned with atom- centered dielectric functions. *Journal of Chemical Theory and Computation*, 3:170–183, 2007.

[215] M. Feig, J. Karanicolas, and 3rd Brooks, C. L. Mmtsb tool set: Enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of Molecular Graphics and Modelling*, 22:377–95, 2004.

[216] B. R. Brooks, 3rd Brooks, C. L., Jr. Mackerell, A. D., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. Charmm: The biomolecular simulation program. *Journal of Computational Chemistry*, 30:1545–614, 2009.

[217] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical-integration of cartesian

equations of motion of a system with constraints - molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23:327–341, 1977.

[218] C. Chothia, M. Levitt, and D. Richardson. Helix to helix packing in proteins. *Journal of Molecular Biology*, 145:215–50, 1981.

[219] W. Humphrey, A. Dalke, and K. Schulten. Vmd: Visual molecular dynamics. *Journal of Molecular Graphics*, 14:33–8, 27–8, 1996.

[220] T. Darden, D. York, and L. Pedersen. Particle mesh ewald - an n.log(n) method for ewald sums in large systems. *Journal of Chemical Physics*, 98:10089–10092, 1993.

[221] A. K. Dunker and V. N. Uversky. Drugs for 'protein clouds': Targeting intrinsically disordered transcription factors. *Current Opinion in Pharmacology*, 10:782–8, 2010.

[222] D. I. Hammoudeh, A. V. Follis, E. V. Prochownik, and S. J. Metallo. Multiple independent binding sites for small-molecule inhibitors on the oncoprotein c-myc. *Journal of the American Chemical Society*, 131:7390–401, 2009.

[223] G. Mustata, A. V. Follis, D. I. Hammoudeh, S. J. Metallo, H. Wang, E. V. Prochownik, J. S. Lazo, and I. Bahar. Discovery of novel myc-max heterodimer disruptors with a three-dimensional pharmacophore model. *Journal of Medicinal Chemistry*, 52:1247–50, 2009.

[224] D. D. Boehr, R. Nussinov, and P. E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology*, 5:789–96, 2009.

[225] G. G. Hammes, Y. C. Chang, and T. G. Oas. Conformational selection or induced fit: A flux description of reaction mechanism. *Proceedings of the National Academy of Sciences, USA*, 106:13737–41, 2009.

[226] E. A. Bienkiewicz, J. N. Adkins, and K. J. Lumb. Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27(kip1). *Biochemistry*, 41:752–9, 2002.

[227] T. Zor, B. M. Mayr, H. J. Dyson, M. R. Montminy, and P. E. Wright. Roles of phosphorylation and helix propensity in the binding of the kix domain of creb-binding protein by constitutive (c-myb) and inducible (creb) activators. *Journal of Biological Chemistry*, 277:42241–8, 2002.

[228] K. Okazaki and S. Takada. Dynamic energy landscape view of coupled binding and protein conformational change: Induced-fit versus population-shift mechanisms. *Proceedings of the National Academy of Sciences, USA*, 105:11182–7, 2008.

[229] Jr. Hills, R. D. and 3rd Brooks, C. L. Insights from coarse-grained go models for protein folding and dynamics. *International Journal of Molecular Sciences*, 10:889–905, 2009.

[230] Y. Levy, P. G. Wolynes, and J. N. Onuchic. Protein topology determines binding mechanism. *Proceedings of the National Academy of Sciences, USA*, 101:511–6, 2004.

[231] C. J. Tsai, B. Ma, and R. Nussinov. Folding and binding cascades: Shifts in energy landscapes. *Proceedings of the National Academy of Sciences, USA*, 96:9970–2, 1999.

[232] J. Karanicolas and 3rd Brooks, C. L. The origins of asymmetry in the folding transition states of protein l and protein g. *Protein Science*, 11:2351–61, 2002.

[233] J. Karanicolas and 3rd Brooks, C. L. Improved go-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *Journal of Molecular Biology*, 334:309–25, 2003.

[234] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, 256:623–44, 1996.

[235] Jr. Hills, R. D. and 3rd Brooks, C. L. Subdomain competition, cooperativity, and topological frustration in the folding of chey. *Journal of Molecular Biology*, 382:485–95, 2008.

[236] J. Karanicolas and 3rd Brooks, C. L. The structural basis for biphasic kinetics in the folding of the ww domain from a formin-binding protein: Lessons for protein design? *Proceedings of the National Academy of Sciences, USA*, 100:3954–9, 2003.

[237] S. S. Cho, Y. Levy, and P. G. Wolynes. P versus q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proceedings of the National Academy of Sciences, USA*, 103:586–91, 2006.

[238] Jr. Hills, R. D. and 3rd Brooks, C. L. Coevolution of function and the folding landscape: Correlation with density of native contacts. *Biophysical Journal*, 95:L57–9, 2008.

[239] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. Theory of protein folding: The energy landscape perspective. *Annual Review of Physical Chemistry*, 48:545–600, 1997.

[240] A. Caflisch. Network and graph analyses of folding free energy surfaces. *Current Opinion in Structural Biology*, 16:71–8, 2006.

[241] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao. Complex network analysis of free-energy landscapes. *Proceedings of the National Academy of Sciences, USA*, 104:1817–22, 2007.

[242] S. Muff and A. Caflisch. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a beta-sheet miniprotein. *Proteins*, 70:1185–95, 2008.

[243] M. Andrec, A. K. Felts, E. Gallicchio, and R. M. Levy. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proceedings of the National Academy of Sciences, USA*, 102:6801–6, 2005.

[244] G. R. Bowman, K. A. Beauchamp, G. Boxer, and V. S. Pande. Progress and challenges in the automated construction of markov state models for full protein systems. *Journal of Chemical Physics*, 131:124101, 2009.

[245] G. R. Bowman and V. S. Pande. Protein folded states are kinetic hubs. *Proceedings of the National Academy of Sciences, USA*, 107:10890–5, 2010.

[246] V. A. Voelz, V. R. Singh, W. J. Wedemeyer, L. J. Lapidus, and V. S. Pande. Unfolded-state dynamics and structure of protein l characterized by simulation and experiment. *Journal of the American Chemical Society*, 132:4702–9, 2010.

[247] F. Rao and A. Caflisch. The protein folding network. *Journal of Molecular Biology*, 342:299–306, 2004.

[248] X. Wang, A. Vitalis, M. A. Wyczalkowski, and R. V. Pappu. Characterizing the conformational ensemble of monomeric polyglutamine. *Proteins*, 63:297–311, 2006.

[249] J. Kubelka, J. Hofrichter, and W. A. Eaton. The protein folding 'speed limit'. *Current Opinion in Structural Biology*, 14:76–88, 2004.

[250] Y. Levy, J. N. Onuchic, and P. G. Wolynes. Fly-casting in protein-dna binding: Frustration between protein folding and electrostatics facilitates target recognition. *Journal of the American Chemical Society*, 129:738–9, 2007.

[251] D. Vuzman and Y. Levy. Dna search efficiency is modulated by charge composition and distribution in the intrinsically disordered tail. *Proceedings of the National Academy of Sciences, USA*, 107:21004–9, 2010.

[252] E. Shakhnovich. Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. *Chemical Reviews*, 106:1559–88, 2006.

[253] Y. Huang and Z. Liu. Nonnative interactions in coupled folding and binding processes of intrinsically disordered proteins. *PLoS One*, 5:e15375, 2010.

[254] Z. Zhang and H. S. Chan. Competition between native topology and nonnative interactions in simple and complex folding kinetics of natural and designed proteins. *Proceedings of the National Academy of Sciences, USA*, 107:2920–5, 2010.

[255] J. A. Marsh and J. D. Forman-Kay. Sequence determinants of compaction in intrinsically disordered proteins. *Biophysical Journal*, 98:2383–90, 2010.

[256] P. Weinkam, E. V. Pletneva, H. B. Gray, J. R. Winkler, and P. G. Wolynes. Electrostatic effects on funneled landscapes and structural diversity in denatured protein ensembles. *Proceedings of the National Academy of Sciences, USA*, 106:1796–801, 2009.

[257] C. J. Oldfield, Y. Cheng, M. S. Cortese, C. J. Brown, V. N. Uversky, and A. K. Dunker. Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, 44:1989–2000, 2005.

[258] V. Vacic, P. R. Markwick, C. J. Oldfield, X. Zhao, C. Haynes, V. N. Uversky, and L. M. Iakoucheva. Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Computational Biology*, 8:e1002709, 2012.

[259] R. K. Das, A. Mittal, and R. V. Pappu. How is functional specificity achieved through disordered regions of proteins? *Bioessays*, 35:17–22, 2013.

[260] C. A. Galea, Y. Wang, S. G. Sivakolundu, and R. W. Kriwacki. Regulation of cell division by intrinsically unstructured proteins: Intrinsic flexibility, modularity, and signaling conduits. *Biochemistry*, 47:7598–609, 2008.

[261] D. Ganguly, S. Otieno, B. Waddell, L. Iconaru, R. W. Kriwacki, and J. Chen. Electrostatically accelerated coupled binding and folding of intrinsically disordered proteins. *Journal of Molecular Biology*, 422:674–84, 2012.

[262] H. X. Zhou, X. Pang, and C. Lu. Rate constants and mechanisms of intrinsically disordered proteins binding to structured targets. *Physical Chemistry Chemical Physics*, 14:10466–76, 2012.

[263] A. M. Brown and N. J. Zondlo. A propensity scale for type ii polyproline helices

(ppii): Aromatic amino acids in proline-rich sequences strongly disfavor ppii due to proline-aromatic interactions. *Biochemistry*, 51:5041–51, 2012.

[264] S. Muller-Spath, A. Soranno, V. Hirschfeld, H. Hofmann, S. Ruegger, L. Reymond, D. Nettels, and B. Schuler. From the cover: Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences, USA*, 107:14609–14, 2010.

[265] M. Arai, J. C. Ferreon, and P. E. Wright. Quantitative analysis of multisite protein-ligand interactions by nmr: Binding of intrinsically disordered p53 transactivation subdomains with the taz2 domain of cbp. *Journal of the American Chemical Society*, 134:3792–803, 2012.

[266] C. A. Galea, A. Nourse, Y. Wang, S. G. Sivakolundu, W. T. Heller, and R. W. Kriwacki. Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27 kip1. *Journal of Molecular Biology*, 376:827–38, 2008.

[267] J. M. Rogers, A. Steward, and J. Clarke. Folding and binding of an intrinsically disordered protein: Fast, but not 'diffusion-limited'. *Journal of the American Chemical Society*, 135:1415–22, 2013.

[268] S. G. Sivakolundu, A. Nourse, S. Moshiach, B. Bothner, C. Ashley, J. Satumba, J. Lahti, and R. W. Kriwacki. Intrinsically unstructured domains of arf and hdm2 form bimolecular oligomeric structures in vitro and in vivo. *Journal of Molecular Biology*, 384:240–54, 2008.

[269] M. Vijayakumar, K. Y. Wong, G. Schreiber, A. R. Fersht, A. Szabo, and H. X. Zhou. Electrostatic enhancement of diffusion-controlled protein-protein association: Comparison of theory and experiment on barnase and barstar. *Journal of Molecular Biology*, 278:1015–24, 1998.

[270] H. Liu, Y. Shi, X. S. Chen, and A. Warshel. Simulating the electrostatic guidance of

the vectorial translocations in hexameric helicases and translocases. *Proceedings of the National Academy of Sciences, USA*, 106:7449–54, 2009.

[271] A. Marcovitz and Y. Levy. Frustration in protein-dna binding influences conformational switching and target search kinetics. *Proceedings of the National Academy of Sciences, USA*, 108:17957–62, 2011.

[272] A. D. Stoycheva, 3rd Brooks, C. L., and J. N. Onuchic. Gatekeepers in the ribosomal protein s6: Thermodynamics, kinetics, and folding pathways revealed by a minimalist protein model. *Journal of Molecular Biology*, 340:571–85, 2004.

[273] A. D. Stoycheva, J. N. Onuchic, and C. L. Brooks. Effect of gatekeepers on the early folding kinetics of a model beta-barrel protein. *Journal of Chemical Physics*, 119:5722–5729, 2003.

[274] A. Toth-Petroczy, I. Simon, M. Fuxreiter, and Y. Levy. Disordered tails of homeodomains facilitate dna recognition by providing a trade-off between folding and specific binding. *Journal of the American Chemical Society*, 131:15084–5, 2009.

[275] X. Chu, Y. Wang, L. Gan, Y. Bai, W. Han, E. Wang, and J. Wang. Importance of electrostatic interactions in the association of intrinsically disordered histone chaperone chz1 and histone h2a.z-h2b. *PLoS Computational Biology*, 8:e1002608, 2012.

[276] J. P. Kruse and W. Gu. Modes of p53 regulation. *Cell*, 137:609–22, 2009.

[277] H. Feng, L. M. Jenkins, S. R. Durell, R. Hayashi, S. J. Mazur, S. Cherry, J. E. Tropea, M. Miller, A. Wlodawer, E. Appella, and Y. Bai. Structural basis for p300 taz2-p53 tad1 binding and modulation by phosphorylation. *Structure*, 17:202–10, 2009.

[278] S. A. Dames, M. Martinez-Yamout, R. N. De Guzman, H. J. Dyson, and P. E. Wright. Structural basis for hif-1 alpha /cbp recognition in the cellular hypoxic response. *Proceedings of the National Academy of Sciences, USA*, 99:5271–6, 2002.

[279] K. Sugase, J. C. Lansing, H. J. Dyson, and P. E. Wright. Tailoring relaxation dispersion experiments for fast-associating protein complexes. *Journal of the American Chemical Society*, 129:13406–7, 2007.

[280] J. Dogan, T. Schmidt, X. Mu, A. Engstrom, and P. Jemth. Fast association and slow transitions in the interaction between two intrinsically disordered protein domains. *Journal of Biological Chemistry*, 287:34316–24, 2012.

[281] S. Jo, T. Kim, V. G. Iyer, and W. Im. Charmm-gui: A web-based graphical user interface for charmm. *Journal of Computational Chemistry*, 29:1859–65, 2008.

[282] S. Jo, M. Vargyas, J. Vasko-Szedlar, B. Roux, and W. Im. Pbeq-solver for online visualization of electrostatic potential of biomolecules. *Nucleic Acids Research*, 36:W270–5, 2008.

[283] J. E. Shea, J. N. Onuchic, and 3rd Brooks, C. L. Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment b of protein a. *Proceedings of the National Academy of Sciences, USA*, 96:12512–7, 1999.

[284] M. Ubbink. The courtship of proteins: Understanding the encounter complex. *FEBS Letters*, 583:1060–6, 2009.

[285] H. Lee, K. H. Mok, R. Muhandiram, K. H. Park, J. E. Suk, D. H. Kim, J. Chang, Y. C. Sung, K. Y. Choi, and K. H. Han. Local structural elements in the mostly unstructured transcriptional activation domain of human p53. *Journal of Biological Chemistry*, 275:29426–32, 2000.

[286] E. Gallicchio, M. Andrec, A. K. Felts, and R. M. Levy. Temperature weighted histogram analysis method, replica exchange, and transition paths. *Journal of Physical Chemistry B*, 109:6722–31, 2005.

[287] J. D. Schmit. Kinetic theory of amyloid fibril templating. *Journal of Chemical Physics*, 138:185102, 2013.

[288] N. Rathore, M. Chopra, and J. J. de Pablo. Optimal allocation of replicas in parallel tempering simulations. *Journal of Chemical Physics*, 122:024111, 2005.

[289] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer. Feedback-optimized parallel tempering monte carlo. *Journal of Statistical Mechanics-Theory and Experiment*, page 22, 2006.

[290] J. Kuszewski, A. M. Gronenborn, and G. M. Clore. Improving the packing and accuracy of nmr structures with a pseudopotential for the radius of gyration. *Journal of the American Chemical Society*, 121:2337–2338, 1999.

[291] M. Kouza and U. H. Hansmann. Velocity scaling for optimizing replica exchange molecular dynamics. *Journal of Chemical Physics*, 134:044124, 2011.

[292] U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281:140, 1997.

[293] W. Nadler and U. H. Hansmann. Dynamics and optimal number of replicas in parallel tempering simulations. *Phys Rev E Stat Nonlin Soft Matter Phys*, 76:065701, 2007.

[294] W. Nadler and U. H. Hansmann. Generalized ensemble and tempering simulations: A unified view. *Phys Rev E Stat Nonlin Soft Matter Phys*, 75:026109, 2007.

[295] W. Nadler, J. H. Meinke, and U. H. Hansmann. Folding proteins by first-passage-times-optimized replica exchange. *Phys Rev E Stat Nonlin Soft Matter Phys*, 78:061905, 2008.

[296] H. X. Kondo and M. Taiji. Enhanced exchange algorithm without detailed balance condition for replica exchange method. *Journal of Chemical Physics*, 138:244113, 2013.

[297] J. Chen and 3rd Brooks, C. L. Can molecular dynamics simulations provide high-resolution refinement of protein structure? *Proteins*, 67:922–30, 2007.

[298] J. Chen, W. Im, and 3rd Brooks, C. L. Refinement of nmr structures using implicit solvent and advanced sampling techniques. *Journal of the American Chemical Society*, 126:16038–47, 2004.

[299] J. D. Faraldo-Gomez and B. Roux. Characterization of conformational equilibria through hamiltonian and temperature replica-exchange simulations: Assessing entropic and environmental effects. *Journal of Computational Chemistry*, 28:1634–47, 2007.

[300] W. Zhang and J. Chen. Accelerate sampling of atomistic energy landscapes using topology-based coarse-grained models. *Journal of Chemical Theory and Computation*, 10:918–923, 2014.

[301] M. Feig, W. Im, and 3rd Brooks, C. L. Implicit solvation based on generalized born theory in different dielectric environments. *Journal of Chemical Physics*, 120:903–11, 2004.

[302] I. Radhakrishnan, G. C. Perez-Alvarado, D. Parker, H. J. Dyson, M. R. Montminy, and P. E. Wright. Solution structure of the kix domain of cbp bound to the transactivation domain of creb: A model for activator:coactivator interactions. *Cell*, 91:741–52, 1997.

[303] I. Radhakrishnan, G. C. Perez-Alvarado, H. J. Dyson, and P. E. Wright. Conformational preferences in the ser133-phosphorylated and non-phosphorylated forms of the kinase inducible transactivation domain of creb. *FEBS Letters*, 430:317–22, 1998.

[304] G. S. Ayton, W. G. Noid, and G. A. Voth. Multiscale modeling of biomolecular systems: In serial and in parallel. *Current Opinion in Structural Biology*, 17:192–8, 2007.

[305] P. Liu and G. A. Voth. Smart resolution replica exchange: An efficient algorithm for exploring complex energy landscapes. *Journal of Chemical Physics*, 126:045106, 2007.

[306] A. T. Brunger and M. Nilges. Computational challenges for macromolecular structure determination by x-ray crystallography and solution nmr-spectroscopy. *Quarterly Reviews of Biophysics*, 26:49–125, 1993.

[307] R. M. Fesinmeyer, F. M. Hudson, and N. H. Andersen. Enhanced hairpin stability through loop design: The case of the protein g b1 domain hairpin. *Journal of the American Chemical Society*, 126:7238–43, 2004.

[308] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and Jr. Mackerell, A. D. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *Journal of Chemical Theory and Computation*, 8:3257–3273, 2012.

Appendix A Copyright Clearance

From: **CONTRACTS-COPYRIGHT (shared)** Contracts-Copyright@rsc.org
Subject: RE: Request of reuse article in my dissertation
Date: April 25, 2014 at 9:32 AM
To: Weihong Zhang zhangwh@ksu.edu

Dear Weihong

The Royal Society of Chemistry (RSC) hereby grants permission for the use of your paper(s) specified below in the printed and microfilm version of your thesis. You may also make available the PDF version of your paper(s) that the RSC sent to the corresponding author(s) of your paper(s) upon publication of the paper(s) in the following ways: in your thesis via any website that your university may have for the deposition of theses, via your university's Intranet or via your own personal website. We are however unable to grant you permission to include the PDF version of the paper(s) on its own in your institutional repository. The Royal Society of Chemistry is a signatory to the STM Guidelines on Permissions (available on request).

Please note that if the material specified below or any part of it appears with credit or acknowledgement to a third party then you must also secure permission from that third party before reproducing that material.

Please ensure that the thesis states the following:

Reproduced by permission of The Royal Society of Chemistry

and include a link to the paper on the Royal Society of Chemistry's website.

Please ensure that your co-authors are aware that you are including the paper in your thesis.

Regards

Gill Cockhead
Publishing Contracts & Copyright Executive

Gill Cockhead
Publishing Contracts & Copyright Executive
Royal Society of Chemistry,
Thomas Graham House,
Science Park, Milton Road,
Cambridge, CB4 0WF, UK
Tel +44 (0) 1223 432134

Follow the Royal Society of Chemistry:
www.rsc.org/follow

Winner of The Queen's Award for Enterprise, International Trade 2013

DISCLAIMER:

This communication (including any attachments) is intended for the use of the addressee only and may contain confidential, privileged or copyright material. It may not be relied upon or disclosed to any other person without the consent of the RSC. If you have received it in error, please contact us immediately. Any advice given by the RSC has been carefully formulated but is necessarily based on the information available, and the RSC cannot be held responsible for accuracy or completeness. In this respect, the RSC owes no duty of care and shall not be liable for any resulting damage or loss. The RSC acknowledges that a disclaimer cannot restrict liability at law for personal injury or death arising through a finding of negligence. The RSC does not warrant that its emails or attachments are Virus-free: Please rely on your own screening. The Royal Society of Chemistry is a charity, registered in England and Wales, number 207890 - Registered office: Thomas Graham House, Science Park, Milton Road, Cambridge CB4 0WF



plos.org create account sign in

PLOS | COMPUTATIONAL BIOLOGY

Browse For Authors About Us Search

advanced search

PLOS Editorial and Publishing Policies

The following editorial and publishing policies apply to all of PLOS journals, unless otherwise noted. For additional policy information, please also refer to the PLOS Terms of Use, and the PLOS Privacy Statement.

Contents:

1. Relationship between PLOS Journals
2. Publication Charges
3. Copyright and License Policies
4. Author Status
5. Competing Interests
6. Human and Animal Research
7. Sharing of Data, Materials, and Software
8. Reporting Guidelines for Specific Study Designs
9. Policies Regarding Submission of a New Taxon Name
10. Submission of Related Manuscripts
11. Reviewer and Editor Exclusions
12. Confidentiality
13. Corrections and Additions
14. Publication Ethics
15. Blogs, Wikis, Embargoes, and the Media
16. Studies Sponsored by Specific Funders
17. Biosecurity and Dual Use Research of Concern
18. FAQs for Editorial Policy

1. Relationship between PLOS Journals

All journals published by PLOS are autonomous publications.

PLOS Biology **and** ***PLOS Medicine***. Each of these two journals has a team of professional and academic editors who evaluate papers for publication. If a manuscript that falls between basic and clinical research is submitted to *PLOS Biology* or *PLOS Medicine*, the professional editors of both journals may consult each other on such manuscripts to ensure the fairest possible treatment. As a result, authors who have submitted their manuscript to *PLOS Medicine* may be encouraged instead to submit their paper to *PLOS Biology* (and vice versa).

PLOS Community Journals. *PLOS Computational Biology*, *PLOS Genetics*, *PLOS Pathogens*, and *PLOS Neglected Tropical Diseases* are run solely and independently by their own Editors-in-Chief and community-based editorial boards.

PLOS ONE**.** This journal is run independently from the other journals by a community-based editorial board.

Transfer of papers. Authors can request that papers (with referee reports, if relevant) rejected from one PLOS journal be transferred to another PLOS journal for further consideration there. We trust that reviewers for any PLOS journal are willing to have their reviews considered by the editors of another PLOS journal. Papers will never be transferred between the journals without an author's consent.

Back to Contents

2. Publication Charges

To provide open access, PLOS journals use a business model in which our expenses—including those of peer review, journal production, and online hosting and archiving—are recovered in part by charging a publication fee to the authors or research sponsors for each article they publish. The fees vary by journal.

PLOS is committed to the widest possible global participation in open access publishing. To determine the appropriate fee, we use a country-based pricing model, which is based on the country that provides 50% or more of the primary funding for the research that is being submitted. Research articles funded by Upper Middle and High Income Countries incur our standard publication fees. Corresponding authors who are affiliated with one of our Institutional Members are eligible for a discount on this fee. Such authors will be informed of the discount applicable after submission of their manuscript.

Fees for Low and Lower Middle Income Countries are calculated according to the PLOS Global Participation Initiative pricing program for manuscripts submitted after 9am Pacific Time on September 4, 2012 (this program is not retroactive).

- Group One: Countries from this list will not be charged for publishing
- Group Two: Countries from this list will be charged a flat $500

Our fee waiver policy, whereby PLOS offers to waive or further reduce the payment required of authors who cannot pay the full amount charged for publication, remains in effect. Editors and reviewers have no access to whether authors are able to pay; decisions to publish are only based on editorial criteria.

Back to Contents

3. Copyright and License Policies

Open access agreement. Upon submission of an article, its authors are asked to indicate their agreement to abide by an open access Creative Commons Attribution (CC BY) license. Under the terms of this license, authors retain ownership of the copyright of their articles. However, the license permits any user to download, print out, extract, reuse, archive, and distribute the article, so long as appropriate credit is given to the authors and source of the work. The license ensures that the authors' article will be available as widely as possible and that the article can be included in any scientific archive.

Open access agreement: US government authors. Papers authored by one or more US government employees are not copyrighted, but are licensed under a CC0 Public Domain Dedication, which allows unlimited distribution and reuse of the article for any lawful purpose. Authors should read about CC-BY or CC0 before submitting papers.

Archiving in PubMed Central. Upon publication, PLOS also deposits all articles in PubMed Central. This complies with the policies of funding agencies, such as the NIH in the USA, the Wellcome Trust, and the Research Councils in the UK, and the Deutsche Forschungsgemeinschaft in Germany, which request or require deposition of the published articles that they fund into publicly available databases.

Back to Contents

4. Author Status

Process. All authors will be contacted by email at submission to ensure that they are aware of and approve the submission of the manuscript, its content, and its authorship. Some PLOS journals also require that all co-authors then confirm their assent to publication by email.

Professional medical writers. The involvement of any professional medical writer in the publication process must be declared. We encourage authors to consult the European Medical Writers' Association Guidelines on the role of medical writers.

Authorship criteria. All PLOS journals base their criteria for authorship on those outlined in the International Committee of Medical Journal Editors (ICMJE) Uniform Requirements for Manuscripts Submitted to Biomedical Journals, which are excerpted below. The contributions of all authors must be described. Contributions that fall short of authorship should be mentioned in the Acknowledgments section of the paper.

- *Authorship credit should be based on 1) substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; and 3) final approval of the version to be published. Authors should meet conditions 1, 2, and 3.*
- *When a large, multicenter group has conducted the work, the group should identify the individuals who accept direct responsibility for the manuscript (3). These individuals should fully meet the criteria for authorship/contributorship defined above, and editors will ask these individuals to complete journal-specific author and conflict-of-interest disclosure forms. When submitting a manuscript authored by a group, the corresponding author should clearly indicate the preferred citation and identify all individual authors as well as the group name. Journals generally list other members of the group in the Acknowledgments. The NLM indexes the group name and the names of individuals the group has identified as being directly responsible for the*

manuscript; it also lists the names of collaborators if they are listed in Acknowledgments.

- *Acquisition of funding, collection of data, or general supervision of the research group alone does not constitute authorship.*
- *All persons designated as authors should qualify for authorship, and all those who qualify should be listed.*
- *Each author should have participated sufficiently in the work to take public responsibility for appropriate portions of the content.*

Changes in authorship. PLOS journals follow the COPE guidelines covering changes in authorship. If any changes to the list of authors of a manuscript are necessary after the initial submission of a manuscript to a PLOS journal but before its publication, the corresponding author must first contact the journal staff and provide a clear reason for the change(s). If the change to the authorship list is appropriate and in keeping with the guidelines given above, the corresponding author will be asked to provide written confirmation that all other authors listed on the manuscript at that time consent to the change(s). Any individuals who the corresponding author asks to be added or removed from the list of authors will be contacted by PLOS.

Back to Contents

5. Competing Interests

PLOS policy on competing interests. PLOS is committed to ensuring that research is as free from bias as possible and is seen to be so. It is increasingly recognized that everyone — authors, authors' employers (e.g., an academic institution, government department, commercial company, or other), sponsors of the work, reviewers, editors, and publishers — has competing interests of some sort. It is difficult for individual readers to assess objectively whether competing interests could have biased the presentation of, peer review of, or decision to publish a given work. Transparency of competing interests allows readers to better evaluate the possibility of such bias. Journals and their editors must take all competing interests into account during the review process and ensure that any relevant ones are declared in the published article. PLOS journals therefore have the following three requirements:

- Authors must declare all relevant competing interests for consideration during the review process.
- Editors (professional or academic, paid or unpaid) and reviewers must declare their own competing interests and if necessary disqualify themselves from involvement in the assessment of a paper.
- Anyone who comments on or rates published papers in any PLOS journal must declare their competing interests at the time of posting their comments and/or rating.

Explanation of competing interests.

What is a competing interest?
PLOS defines a competing interest as anything that interferes with, or could reasonably be perceived as interfering with, the full and objective presentation, peer review, editorial decision-making, or publication of research or non-research articles submitted to one of the journals. Competing interests can be financial or non-financial, professional, or personal. Competing interests can arise in relationship to an organization or another person. See below for definitions and examples of various competing interests.

Who needs to declare competing interests?
Everyone involved in authorship, funding, review, and editorial decision-making of submitted articles, or who wishes to comment on or rate published articles, must declare any and all relevant competing interests.

How do PLOS journals treat competing interests in their review and publication process?
No decision is made to publish any paper submitted to a PLOS journal until a competing interests statement has been submitted for all authors. The editors may ask for clarification about declarations. The role of all funding sources in the work must be described and authors must state explicitly whether the funder was involved in the: study design; collection, analysis, and interpretation of data; writing of the paper; and/or decision to submit for publication. If the funder was involved, the role/contribution must be described explicitly.

The editors of a PLOS journal might decide not to publish a paper if they believe the competing interests declared by the authors or funders might have compromised the objectivity or validity of the research, analyses, or interpretations in the paper. PLOS editors will not publish a commissioned or any other non-research article if they are aware of a competing interest that, in their judgment, could introduce bias or a reasonable perception of bias. PLOS editors do not consult reviewers who have competing interests that, in the editors' judgment, could interfere with unbiased review. Failure to declare competing interests at submission, or when an article is commissioned, can result in immediate rejection of the paper. If a competing interest comes to light after publication, PLOS journal will issue a formal correction to or retraction of the whole paper, as appropriate.

Examples of competing interests.

Financial

Financial competing interests include but are not limited to:

- ownership of stocks or shares;
- paid employment or consultancy;
- board membership;

- patent applications (pending or actual), including individual applications or those belonging to the institution to which the authors are affiliated and from which the authors may benefit;
- research grants (from any source, restricted or unrestricted);
- travel grants and honoraria for speaking or participation at meetings;
- gifts.

Authors must declare all potential financial competing interests involving people or organizations that might reasonably be perceived as relevant. Similarly, reviewers and academic and professional editors, paid or unpaid, must declare any financial relationships that could reasonably be perceived as relevant and/or could influence their objective review of the paper; if a financial competing interest exists, these individuals should recuse themselves from handling the paper. Anyone wishing to comment on or rate a published paper must also disclose any relevant financial interests. As a guide, any competing interest that arose within the five years either before or after the commencement of the research described, or within five years of the article being written, or within five years of events described in the article, should be declared. However, interests outside this time-frame might also be relevant; if so, they should also be declared so that their relevance can be judged by the journal editorial team.

Non-financial

Non-financial competing interests include but are not limited to:

Professional:

- acting as an expert witness;
- membership in a government or other advisory board;
- relationship (paid or unpaid) with organizations and funding bodies including nongovernmental organizations, research institutions, or charities;
- membership of lobbying or advocacy organizations;
- writing or consulting for an educational company.

Personal:

- personal relationships (i.e. friend, spouse, family member, current or previous mentor, adversary) with individuals involved in the submission or evaluation of a paper, such as authors, reviewers, editors, or members of the editorial board of a PLOS journal;
- personal convictions (political, religious, ideological, or other) related to a paper's topic that might interfere with an unbiased publication process (at the stage of authorship, peer review, editorial decision-making, or publication).

Authors, reviewers, editors, and anyone wishing to comment on a published paper must disclose any non-financial interests that might influence their reporting, handling, or review of the paper, or that might be negatively or positively affected by publication of the paper.

For example, authors are required to declare whether they have served or currently serve on the editorial board of the journal to which they are submitting, have acted as an expert witness in relevant legal proceedings, or have sat or currently sit on a committee for an organization that may benefit from publication of the paper.

Reviewers are required to declare whether they have held grants, co-authored papers, or worked in the same institution or organization with the authors of the study they are asked to review, or if they are in an adversarial relationship with the authors.

Similarly, editors — academic or professional, paid or unpaid — are required to recuse themselves from deliberations if they cannot evaluate a paper in an objective way because of personal relationships with the authors.

Finally, anyone who comments on or rates published papers for PLOS must declare non-financial competing interests at the time of posting their comments and/or rating.

See this editorial for further clarification: PLOS Medicine Editors (2008) Making Sense of Non-Financial Competing Interests. PLOS Med 5(9): e199 doi:10.1371/journal.pmed.0050199.

Back to Contents

6. Human and Animal Research

Research involving human participants. All research involving human participants must have been approved by the authors' institutional review board or equivalent committee(s), and that board must be named in the manuscript. For research involving human participants, informed consent must have been obtained (or the reason for lack of consent explained — for example, that the data were analyzed anonymously) and all clinical investigation must have been conducted according to the principles expressed in the *Declaration of Helsinki*. Authors should be able to submit, upon request, a statement from the research ethics committee or institutional review board indicating approval of the research. PLOS editors also encourage authors to submit a sample of a patient consent form, and might require submission on particular occasions.

For studies involving humans categorized by race/ethnicity, age, disease/disabilities, religion, sex/gender, sexual orientation, or other socially constructed

groupings, authors should, as much as possible:

- make explicit their methods of categorizing human populations;
- define categories in as much detail as the study protocol allows;
- justify their choices of definitions and categories, including for example whether any rules of human categorization were required by their funding agency;
- explain whether (and, if so, how) they controlled for confounding variables such as socioeconomic status, nutrition, environmental exposures, etc.

In addition, outmoded terms and potentially stigmatizing labels should be changed to more current, acceptable terminology — for example: 'Caucasian' should be changed to 'white' or 'of [western] European descent' (as appropriate); 'cancer victims' should be changed to 'patients with cancer.'

Reporting of animal studies and ethical treatment of animals. For studies involving animals, all work must have been conducted according to applicable national and international guidelines. Prior approval must have been obtained for all protocols from the relevant author's institutional or other appropriate ethics committee, and the institution name and permit numbers must be provided at submission (see example below). For research involving non-human primates, all studies must be performed in accordance with the recommendations of the Weatherall (2006) report, *The use of non-human primates in research*. Where unregulated animals are used or ethics approval is not required by a specific committee, the article should include a clear statement of this fact and the reasons why ethical approval is not required.

We also strongly encourage all authors to comply with the '*Animal Research: Reporting In Vivo Experiments*' (ARRIVE) guidelines, developed by NC3Rs to improve standards of reporting to ensure that the data from animal experiments can be fully scrutinized and utilized. Relevant information should be included in the appropriate section of the article (e.g. title, abstract, or method), as outlined in the ARRIVE guidelines. The ARRIVE guidelines can be applied to any area of bioscience research using laboratory animals. Where research could be confused as pertaining to human clinical research, the animal model should also be noted in the article title.

Example of statement of ethical approval. *This study was performed in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health. The protocol was approved by the Committee on the Ethics of Animal Experiments of the University of Minnesota (Permit Number: 27-2956). All surgery was performed under sodium pentobarbital anesthesia, and every effort was made to minimize suffering.*

Back to Contents

7. Sharing of Data, Materials, and Software

Publication is conditional upon the agreement of the authors to make freely available any materials and information described in their publication that may be reasonably requested by others.

Data Availability

PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception[1].

When submitting a manuscript online, authors must provide a *Data Availability Statement* describing compliance with PLOS's policy. If the article is accepted for publication, the data availability statement will be published as part of the final article.

Refusal to share data and related metadata and methods in accordance with this policy will be grounds for rejection. PLOS journal editors encourage researchers to contact them if they encounter difficulties in obtaining data from articles published in PLOS journals. If restrictions on access to data come to light after publication, we reserve the right to post a correction, to contact the authors' institutions and funders, or in extreme cases to retract the publication.

Methods acceptable to PLOS journals with respect to data sharing are listed below, accompanied by guidance for authors as to what must be indicated in their data availability statement and how to follow best practices in reporting. If authors did not collect data themselves but used another source, this source must be credited as appropriate. Authors who have questions or difficulties with the policy, or readers who have difficulty accessing data, are encouraged to contact the relevant journal office or data@plos.org.

Acceptable data-sharing methods:

Data deposition (strongly recommended). All data and related metadata underlying the findings reported in a submitted manuscript should be deposited in an appropriate public repository[2], unless already provided as part of the submitted article. Repositories may be either subject-specific (where these exist) and accept specific types of structured data, or generalist repositories that accept multiple data types, such as Dryad. Guidance on acceptable repositories is included below[2]. The *Data Availability Statement* must specify that data are deposited publicly and list the name(s) of repositories along with digital object identifiers or accession numbers for the relevant datasets. In some cases authors may not be able to obtain DOIs or accession numbers until the manuscript is accepted; in these cases, the authors must provide these numbers at acceptance. In all other cases, these numbers must be provided at submission.

Data in Supporting Information files. For smaller datasets and certain data types, authors may upload data as Supporting Information files accompanying the manuscript. (See also additional information regarding appropriate use of Supporting Information files.) Authors should take care to maximize the accessibility and reusability of the data by selecting a file format from which data can be efficiently extracted (for example, spreadsheets are preferable to PDF when providing tabulated data).

If data deposition or provision in Supporting Information is not ethical or legal (e.g., underlying data pose privacy or legal concerns, or include human participants[3]), the following two methods may be acceptable alternatives, subject to case-by-case evaluation:

Data made available to all interested researchers upon request. The Data Availability Statement must specify "Data available on request" and identify the group to which requests should be submitted (e.g., a named data access committee or named ethics committee). The reasons for restrictions on public data deposition must also be specified. Note that it is not acceptable for the authors to be the sole named individuals responsible for ensuring data access.

Data available from third party. In the case of a primary dataset that was not originally generated by the authors of the submitted manuscript, appropriate data sharing may require that interested researchers obtain third-party data independently from the named original source. In this case, the *Data Availability Statement* must state the source of the data with full citation and, if the dataset cannot be provided, indicate "Data available from (named source)." The reasons for restrictions on public data deposition must also be specified.

Unacceptable data access restrictions:

PLOS journals will not consider manuscripts for which the following factors influence ability to share data:

- Authors will not share data because of personal interests, such as patents or potential future publications.
- The conclusions depend solely on the analysis of proprietary data (e.g., data owned by commercial interests, or copyrighted data). If proprietary data are used, the manuscript must include an analysis of public data that validates the conclusions so that others can reproduce the analysis and build on the findings.

Explanatory notes and guidance:

A compilation of frequently asked questions about the PLOS Data Policy is available and is updated periodically.

1. Definition of data that must be shared

PLOS defines the "minimal dataset" to consist of the dataset used to reach the conclusions drawn in the manuscript with related metadata and methods, and any additional data required to replicate the reported study findings in their entirety. Core descriptive data, methods, and study results should be included within the main paper, regardless of data deposition. PLOS does not accept references to "data not shown". Editors and reviewers may require particular data types for certain articles on a case-by-case basis. Authors who have datasets too large for sharing via repositories or uploaded files should contact the relevant journal for advice.

2. Guidance on data repositories

PLOS requires that authors comply with field-specific standards for preparation and recording of data and select repositories appropriate to their field, for example deposition of microarray data in ArrayExpress or GEO; deposition of gene sequences in GenBank, EMBL or DDBJ; clinical trials data in ClinicalTrials.gov; and deposition of ecological data in Dryad. Authors are encouraged to select repositories that meet accepted criteria as trustworthy digital repositories, such as criteria of the Centre for Research Libraries or Data Seal of Approval. Large, international databases are more likely to persist than small, local ones. Copyright licensing for data held in repositories may be unclear. If authors use repositories with stated licensing policies the policies should not be more restrictive than CC-BY.

3. Guidance on sharing datasets that derive from clinical studies or other work involving human participants

For studies involving human participants, data must be handled so as to not compromise study participants' privacy. PLOS recommends that researchers follow established guidance and applicable local laws in ensuring they do not compromise participant privacy. Resources which researchers may consult for guidance include:

US National Institutes of Health: Protecting the Rights and Privacy of Human Subjects
Canadian Institutes of Health Research Best Practices for Protecting Privacy in Health Research
UK Data Archive: Anonymisation Overview
Australian National Data Service: Ethics, Consent and Data Sharing

Steps necessary to protect privacy may include de-identification, blocking portions of the database, or license agreements directed specifically at privacy concerns. Authors should indicate, as part of the ethics statement, the ways in which the study participants' privacy was preserved. If license agreements apply, authors should note the process necessary for other researchers to obtain a license.

Availability of materials

PLOS is committed to ensuring the availability of materials that underpin any articles published in PLOS journals. PLOS's ideal is to make all readily replaceable materials relevant to a given article immediately available without restrictions (while not compromising confidentiality in the context of human-subject research).

PLOS journal editors encourage researchers to contact them if they encounter difficulties in obtaining materials from articles published in PLOS journals. PLOS reserves the right to post corrections on articles, to contact authors' institutions and funders, and in extreme cases to withdraw publication, if unreasonable restrictions on access to materials come to light after publication of a PLOS journal article.

Availability of Software

PLOS supports the development of open source software and believes that, for submissions in which software is the central part of the paper, adherence to appropriate open source standards will ensure that the submission conforms to (1) our requirements that methods be described in sufficient detail that another researcher can reproduce the experiments described, (2) our aim to promote openness in research, and (3) our intention that all work published in PLOS journals can be built upon by future researchers. Therefore, if new software or a new algorithm is central to a PLOS paper, the authors must confirm that the software conforms to the Open Source Definition, have deposited the following three items in an open software archive, and included in the submission as Supporting Information:

- **The associated source code of the software described by the paper.** This should, as far as possible, follow accepted community standards and be licensed under a suitable license such as BSD, LGPL, or MIT (see http://www.opensource.org/licenses/alphabetical for a full list). Dependency on commercial software such as Mathematica and MATLAB does not preclude a paper from consideration, although complete open source solutions are preferred.
- **Documentation for running and installing the software.** For end-user applications, instructions for installing and using the software are prerequisite; for software libraries, instructions for using the application program interface are prerequisite.
- **A test dataset with associated control parameter settings.** Where feasible, results from standard test sets should be included. Where possible, test data should not have any dependencies — for example, a database dump.

Acceptable archives should provide a public repository of the described software. The code should be easy to locate and download without the requirement for creating user accounts, logging in or otherwise registering personal details. The repository must have been in existence for over five years or be hosting more than 1,000 projects. Examples of such archives are: SourceForge, Bioinformatics.Org, Open Bioinformatics Foundation (O|B|F), Google Code, BerliOS Developer, Savannah, GitHub and the Codehaus. Authors should provide a direct link to the deposited software from within the paper.

Deposition with the journal and in an open source archive ensures that the original source associated with the paper is available as well as any enhancements made after the paper is published. An article can be considered for publication if it covers a well-established project that has been providing an open source code repository for an extended amount of time. A condition of acceptance is that the software can be run by reviewers accessing the public software and that the results presented in the paper are reproducible. The software need run on only one hardware-software platform in common use by the readership (including MATLAB), although it must run without dependencies on proprietary or otherwise unobtainable ancillary software. Articles describing software that requires access to databases and other resources whose persistence is not guaranteed (e.g. individual laboratory databases without funding support) will not be considered. In addition, the results described in the paper must be reproducible when peer reviewers, editors, or readers run the software on the deposited dataset and with the provided control parameters.

When the software or algorithm is not central to the paper, we nevertheless encourage authors to make all relevant materials freely available.

Back to Contents

8. Reporting Guidelines for Specific Study Designs

Authors should check the EQUATOR Network site for any reporting guidelines that apply to their study design, and ensure that any required Supporting Information (checklists, protocols, flowcharts, etc.) be included in the article submission.

Clinical trials. PLOS follows the WHO definition of a clinical trial: "...any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes...Interventions include but are not restricted to drugs, cells and other biological products, surgical procedures, radiologic procedures, devices, behavioural treatments, process-of-care changes, preventive care, etc."

PLOS supports the position of the International Committee of Medical Journal Editors (ICMJE) on trial registration. All trials initiated from 1 July 2005 must be registered prospectively in a publicly accessible registry (i.e., before patient recruitment has begun) or they will not be considered for publication. For trials initiated before 1 July 2005, all trials must be registered before submission to any PLOS journal. The ICMJE FAQ on trial registration has further details, and WHO provides a list of approved registries. PLOS editors reserve the right to inform authors' institutions or ethics committees if they become aware of unregistered trials.

Authors of trials must adhere to the CONSORT reporting guidelines appropriate to their trial design, available on the CONSORT Statement Website. Before the paper can enter peer review, authors must: (1) record, in the paper trial registry, the trial registration number and institutional review board, and (2) provide a copy of the trial protocol and a completed CONSORT checklist as supporting information (these documents will also be published alongside the paper, if accepted). The CONSORT flow diagram must be included as the manuscript's "Figure 1." Any deviation from the trial protocol must be explained in the paper. Authors must explicitly discuss informed consent in their paper, and PLOS reserves the right to ask for a copy of the patient consent form. Information on statistical methods or participants, beyond that indicated in the CONSORT statement, should be reported in the Methods section.

PLOS supports the public disclosure of all clinical trial results, as mandated, for example, by the FDA Amendments Act, 2007. Prior disclosure of results on a clinical trial registry site will not affect the decision to peer review or accept papers in PLOS journals.

Systematic reviews and meta-analyses. Reports of systematic reviews and meta-analyses should use the PRISMA statement as a guide, and include a completed PRISMA checklist and flow diagram to accompany the main text. Blank templates of the checklist and flow diagram can be downloaded from the PRISMA Web site. Authors must also state within their Methods section whether a protocol exists for their systematic review, and if so, provide a copy of the protocol as Supporting Information. The journal supports the prospective registration of systematic reviews. Authors whose systematic review was prospectively

registered (e.g. in a registry such as PROSPERO) should also provide the registry number in their abstract. Registry details and protocols will be made available to editors and reviewers, and included alongside the paper for readers if the report is ultimately published.

Diagnostic studies. Reports of studies of diagnostic accuracy should conform to the STARD requirements.

Epidemiological studies. For reports of epidemiological studies, authors should consult the STROBE initiative.

Microarray experiments. Reports of microarray experiments should conform to the MIAME guidelines published by the Functional Genomics Data Society (FGED), and the data from the experiments must be deposited in a publicly accessible database.

Checklists for biological and biomedical research investigations. PLOS recommends that authors refer to the MIBBI Portal (Minimum Information for Biological and Biomedical Investigations) for prescriptive checklists for reporting biological and biomedical research where applicable.

Back to Contents

9. Policies Regarding Submission of a New Taxon Name

Zoological names. When publishing papers that describe a new zoological taxon name, PLOS aims to comply with the requirements of the International Commission on Zoological Nomenclature (ICZN). However, the ICZN does not yet recognize online-only journals, and so, unless PLOS adapts its publication process for taxonomic papers (which it does, as detailed below), any scientific animal name published by PLOS would not be considered 'available' under the rules of the Code ('available' is the formal term for legally published under the Code, and is equivalent to the term 'nomenclaturally valid' in botanical literature). There is a proposal to amend the Code to accommodate online-only publication, which, as of February 2011, awaits an ICZN vote.

Until acceptance of this amendment, the ICZN has proposed an interim solution for authors publishing in PLOS journals that allows PLOS to comply with the code by providing a limited hardcopy print run of the article and making it publicly obtainable. Therefore, for all papers that include the naming of a new zoological taxon, PLOS will make a printed version available for outside parties (at a cost of $10, to cover postage and printing) at the same time as the publication of the online open access article. This additional printed version of the article will contain text in the footer of the first page. This text will be added by PLOS staff, and, apart from this new footer, the printed version will be identical to the PDF of the online version. Footer text: *"This printed document was produced by a method that assures numerous identical and durable copies, and those copies were simultaneously obtainable for the purpose of providing a public and permanent scientific record, in accordance with Article 8.1 of the International Code of Zoological Nomenclature. Date of publication: XXXXXXXX. This document is otherwise identical to DOI: XXXXX."*

In addition, we ask that authors add the following information about archiving and ZooBank registration to the online version of the article before it is finally accepted.

- **Digital archiving.** Please include the names of the digital archives where PLOS articles are deposited (currently, PubMedCentral and LOCKSS). If an author's institute (or those of co-authors) has its own repository, we recommend that authors also deposit the published online article there and include the names of such repositories in the article.
- **ZooBank registration.** Authors should contact ZooBank about the registration of the new species name and request a unique digital identifier (a Life Science Identifier [LSID]). The LSID also needs to be included in the published paper. ZooBank has not yet been officially recognized by the ICZN, but it is still important to register there, because the ICZN are likely to request mandatory registration of new names in ZooBank for all online-only journals.

The following paragraphs provide an example of the type of wording that we recommend for a new taxon description in a PLOS journal. The first paragraph below is required for the online-only version but is overruled by the footer note above, added to the print-only edition. The names of the libraries and institutional repositories should be included, as appropriate.

> **Please insert a modification of the following example text into the Methods section, in a subsection called 'Nomenclatural Acts'.**
> *The electronic version of this document does not represent a published work according to the International Code of Zoological Nomenclature (ICZN), and hence the nomenclatural acts contained in the electronic version are not available under that Code from the electronic edition. Therefore, a separate edition of this document was produced by a method that assures numerous identical and durable copies, and those copies were simultaneously obtainable (from the publication date noted on the first page of this article) for the purpose of providing a public and permanent scientific record, in accordance with Article 8.1 of the Code. The separate print-only edition is available on request from PLOS by sending a request to PLOS, 1160 Battery Street, Suite 100, San Francisco, CA 94111, USA along with a check for $10 (to cover printing and postage) payable to 'PLOS'. In addition, this published work and the nomenclatural acts it contains have been registered in ZooBank, the proposed online registration system for the ICZN. The ZooBank LSIDs (Life Science Identifiers) can be resolved and the associated information viewed through any standard web browser by appending the LSID to the prefix 'http://zoobank.org/'. The LSID for this publication is: [insert here] Anochetus boltoni Fisher sp. nov. urn:lsid:zoobank.org:act:B6C072CF-1CA6-40C7-8396-534E91EF7FBB*

Botanical names. When publishing papers that describe a new botanical taxon name, PLOS aims to comply with the requirements of the International Code of Botanical Nomenclature (ICBN). Following from a test case (Knapp S [2010] PLOS ONE. doi:10.1371/journal.pone.0010502) and in association with the International Plant Names Index (IPNI), the following guidelines for publication in an online-only journal have been agreed such that any scientific botanical name published by PLOS is considered effectively published under the rules of the Code. These guidelines differ from those for zoological nomenclature (see above) and will change in concert with the Code. The online version of the article in itself does not represent a published work according to the ICBN, and hence the new names contained in the online version are not effectively published under that Code from the online edition alone.

To comply with Article 29 of the code, therefore, we ask all authors (one author if the paper has multiple authors) to print out copies (in a single print run) of their paper from the relevant PLOS journal website on the day of publication and ensure that they post these copies on that day to multiple relevant institutions (ten are recommended in ICBN Rec. 30.2 of the Vienna Code), including the names indexing centre for the group in question (i.e. IPNI for flowering plants and ferns, MycoBank and Index Fungorum for fungi). Confirmation that the article has been posted can be added as a comment to the online publication.

PLOS will also make a printed version available for outside parties (at a cost of $10, to cover postage and printing) of the online-only article (which remains freely available).

In addition, we ask that authors add the following information about archiving to the online version of the article before it is finally accepted.

- **Digital archiving.** Please include the names of the digital archives where PLOS articles are deposited (currently, PubMedCentral and LOCKSS). If an author's institute (or those of co-authors) has its own repository, PLOS recommends that authors also deposit the published online article there and include the name of such repositories in their article.
- **Unique digital identifier.** Please provide a unique digital identifier (a Globally Unique Identifier [GUID]; currently an LSID) that indicates listing in IPNI for flowering plants, ferns, and fern allies, or MycoBank/Index Fungorum number for fungi. The IPNI GUID/MB number/IF number also must be included in the published paper. We recommend that authors contact IPNI or MycoBank/Index Fungorum to request a unique digital identifier.

The following paragraphs provide an example of the type of wording that is recommended for a new botanical taxon description in a PLOS journal. The names of the libraries and institutional repositories should be included, as appropriate.

> **Please insert a modification of the following example text into the Methods section, in a subsection called 'Nomenclatural Acts'** (taken from Knapp S [2010] PLOS ONE. doi:10.1371/journal.pone.0010502).
>
> *Nomenclature*
> *The electronic version of this document in itself does not represent a published work according to the International Code of Botanical Nomenclature, and hence the new names contained in the electronic version are not effectively published under that Code from the electronic edition alone. Therefore, a separate edition of this document was produced by a method that assures numerous identical printed copies, and those copies were simultaneously distributed (on the publication date noted on the first page of this article) for the purpose of providing a public and permanent scientific record, in accordance with Article 29 of the Code. Copies of the print-only edition of this article were distributed on the publication date to botanical or generally accessible libraries of the following institutions (BM, COL, GH, HUA, K, MEXU, MO, NY, QCA, QCNE, USM). The separate print-only edition is available on request from PLOS (Public Library of Science) by sending a request to PLOS, 1160 Battery Street, Suite 100, San Francisco, CA 94111, USA along with a check for $10 (to cover printing and postage) payable to 'PLOS'.*
> *In addition, new names contained in this work have been submitted to IPNI, from where they will be made available to the Global Names Index. The IPNI LSIDs (Life Science Identifiers) can be resolved and the associated information viewed through any standard web browser by appending the LSID contained in this publication to the prefix http://ipni.org/. The online version of this work is archived and available from the following digital repositories: PubMedCentral and Solanaceae Source: a web resource for the nightshade family.**

The institutions listed are to be chosen by the author, and GUIDs are likely to replace LSIDs and should be cited where available. Also, neither MycoBank nor Index Fungorum has a local resolution service, so for fungal names please use, and refer to, the TDWG LSID resolver.

Globally unique identifier. In the results section, the GUID (LSID) should be listed under the new species name:

> ***Solanum aspersum*** S. Knapp, sp. Nov. [urn:lsid:ipni.org:names:77103633-1] Type: Colombia. Putumayo: vertiente oriental de la Cordillera, entre Sachamates y San Francisco de Sibundoy, 1600-1750 m, 30 Dec 1940, *J. Cuatrecasas* 11471 (holotype, COL; isotypes, F [F-1335119], US [US-1799731]).

Back to Contents

10. Submission of Related Manuscripts

When submitting an article, all authors are asked to indicate that they do not have a related or duplicate manuscript under consideration (or accepted) for publication elsewhere. If related work has been submitted elsewhere, then a copy must be included with the article submitted to PLOS. Reviewers will be asked to comment on the overlap between related submissions.

Back to Contents

11. Reviewer and Editor Exclusions

Upon submission of a manuscript, authors are asked whether they wish to exclude any specific academic editors or reviewers from the peer review of their article. The editorial team will respect these requests so long as this does not interfere with the objective and thorough assessment of the article. See the relevant guidelines for reviewers and more general information on PLOS policy regarding competing interests.

Back to Contents

12. Confidentiality

Editors and reviewers are required to treat all submitted manuscripts in strict confidence.

Back to Contents

13. Corrections and Additions

PLOS publishes corrections, retractions, and expressions of concern as appropriate, and as quickly as possible. We follow the ICMJE (http://www.icmje.org/) and COPE (http://publicationethics.org/) guidelines where applicable.

A notice of correction will be issued by PLOS to document and correct substantial errors that appear in online articles when these errors significantly affect the content or understanding of the work reported (e.g., error in data presentation or analysis) or when the error affects the publication's metadata (e.g., misspelling of an author's name). In these cases, PLOS will publish a correction that will be linked to the original article.

In very rare cases, we may choose to correct the article itself and re-post it online. If that course is taken, a correction notice will also be created to document the changes to the original article.

Authors who wish to alert PLOS to a situation where a correction may be warranted are requested to contact us with the relevant details (journal, full citation of the article, and description of the error) at: corrections@plos.org or the respective journal office.

Authors are encouraged to post comments to their articles to note typographical errors, and other problems that do not significantly affect the scientific integrity of the work.

Back to Contents

14. Publication Ethics

All PLOS Journals are members of the Committee on Publication Ethics (COPE), abide by its Code of Conduct and aim to adhere to its Best Practice Guidelines.

We will vigorously investigate allegations of publication misconduct in PLOS journals (both before and after publication) and reserve the right to contact authors' institutions, funders or regulatory bodies if needed. If we find conclusive evidence of misconduct we will take steps to correct the scientific record, which may include issuing a correction or retraction. PLOS journals have a Journal Ethics Committee composed of representatives of all the journals, which sets ethical policies for PLOS journals and also investigates specific issues.

The following list outlines some key issues in Publication Ethics. It is not an exhaustive list. For further details authors should consult the journals' specific policies and the references below.

- Authors are expected to be aware of, and comply with, best practice in publication ethics specifically with regard to authorship (for example avoidance of ghost or guest authorship), dual submission, plagiarism, manipulation of figures, competing interests and compliance with policies on research ethics
- PLOS has incorporated CrossCheck (http://www.crossref.org/crosscheck.html), powered by iThenticate, into its journal-wide submission system in order to screen submitted content for originality before publication. Each PLOS journal undertakes screening on a proportion of submitted papers. Authors will be contacted if needed following the screening process.
- Reviewers and Editors are required to treat manuscripts fairly and in confidence, and to declare any competing interests.

In cases of suspected or alleged misconduct, the COPE flowcharts will be followed. We may also seek advice on specific cases at the COPE forum.

Any concerns about the above should be addressed to the Editor in Chief, Chief Editor, or Executive Editor as appropriate of the respective PLOS journal. Please direct inquiries to these individuals via the respective journal address.

More extensive resources are available here: COPE, WAME.

Back to Contents

15. Blogs, Wikis, Embargoes, and the Media

Authors may present and discuss their findings ahead of publication: at medical or scientific conferences, on preprint servers, in public databases, and in blogs, wikis, tweets, and other informal communication channels. We recommend, however, that authors not contact the media or respond to such contact unless an article has been accepted for publication and an embargo date has been established. Respect for press embargoes will help to ensure that the work is reported accurately in the popular media and that the full peer-reviewed paper is freely available to any interested reader when the news item is published. However, if a journalist has covered a piece of work ahead of publication, this will not affect consideration of the work for publication. See also our embargo guidelines for journalists and scientists.

Back to Contents

16. Studies Sponsored by Specific Funders

Pharmaceutical companies. We support GPP2 Good Publication Practice for Communicating Company-Sponsored Medical Research.

Tobacco industry. *PLOS Medicine*, *PLOS Biology*, and *PLOS ONE* will not consider for publication papers in which any of the research costs or authors' salaries have been funded, in whole or in part, by a tobacco company. For an editorial giving the reasoning behind these journals' policy, see: The *PLOS Medicine* Editors (2010) doi:10.1371/journal.pmed.1000237.

Back to Contents

17. Biosecurity and Dual Use Research of Concern

PLOS recognizes that certain research may fall into the category of "dual use research of concern". This is defined by the NSABB as any "biological research with legitimate scientific purpose that may be misused to pose a biologic threat to public health and/or national security." As an Open Access publisher, PLOS remains committed to the widespread dissemination of research while being sensitive to the issues of responsible publication standards. We expect that the potential risks of publishing a scientific paper will outweigh the benefits in only the rarest circumstances. On occasion, PLOS reserves the right to consider manuscript submissions within this context. In addition to the usual scientific scrutiny, such submissions may also be referred to an internal PLOS Dual Use Committee for further deliberation. Authors submitting to any PLOS journal are obligated to disclose potential bioethics/dual use concerns to the journal office at the time of initial submission.

Back to Contents

18. FAQs for Editorial Policy

Q. Why does PLOS require that data underlying research published in PLOS journals be made publicly available?

A. PLOS strongly believes that, to best foster scientific progress, the underlying data from an article should be made freely available for researchers to use, wherever this is legal and ethical. Data availability allows validation, replication, reanalysis, new analysis, reinterpretation, or inclusion into meta-analyses, facilitates reproducibility of research and extends the value of the investment made in funding scientific research. Thus, PLOS believes that ensuring access to the underlying data should be an intrinsic part of the scientific publishing process. Furthermore, by getting data into the right place on publication we can reduce the burden on authors in unearthing old data, retaining old hard drives and answering email requests.

Effective data sharing leads to more citations and should lead to more offers of co-authorship and greater opportunities. We understand that some authors may not want to share data, just as some choose not to make their articles available Open Access, but believe that most authors publish their work precisely in order to allow others to benefit from it. More importantly researchers want to see their work used and cited by others. Making that easier can benefit everyone.

Q. To what data does this policy apply?

A. The policy applies to the dataset used to reach the conclusions drawn in the manuscript with related metadata and methods, and any additional data required to replicate the reported study findings in their entirety. You need not submit your entire dataset, or all raw data collected during an investigation, but you must provide the portion that is relevant to the specific study.

Q. What should I do if the PLOS Data Policy does not appear to take into account my type of data?

A. If after reviewing the policy and FAQs, you have questions about how the policy applies to your data, please contact the relevant PLOS journal or data@plos.org. We will consider each situation on a case-by-case basis.

Q. How should the data be made available?

A. PLOS strongly recommends that data be made available in a public repository. Repositories may be subject-specific (eg, GenBank for sequences, clinicaltrials.gov for clinical trials data, and PDB for structures), general (Dryad or FigShare), or institutional, as long as DOIs or accession numbers are provided and the data are at least as open as CCBY. Alternatively, data may be made available in supporting information files (preferably in a file format from which data

can be efficiently extracted) or in the manuscript itself.

Q: What if my dataset is too large to submit to a repository?

A: For very large datasets that are difficult to deposit in a repository, we encourage you to note details of your situation when submitting your data availability information to PLOS and we will work with you to find a solution.

Q. What are the exceptions to making the data publicly available?

A: The exceptions to making the data publicly available are:

- Data cannot be publically available for ethical or legal reasons, e.g., public availability would compromise patient privacy, or could represent some other threat such as specific locations of fossil deposits or endangered species. Please contact us via the journal office or at data@plos.org if you have concerns about the ethics or legality of sharing your data and they do not fit one of these criteria.
- Data were obtained from a third party, i.e., the authors did not generate the primary dataset themselves. (This may also apply to component sets of data used as part of a meta-analysis, see below.) If this is the case, details of the source and contact information should be provided that others can use to request access.

Q. My funding agency or government law only permits sharing of human participants data with researchers with whom they have a written agreement. What should I do?

A. Please provide the contact information of the individuals and institution(s) where an interested researcher would need to apply to start the process to gain access to the data. If you have questions, please contact the relevant journal or data@plos.org.

Q. The national privacy standards that apply to my research would seem to prevent my publishing the research in PLOS. What should I do?

A. None of the policy is intended to over-rule local regulations, legislation or ethical frameworks. However, note that authors are probably already committing to providing access to data on request when they sign agreements with many journals, not just PLOS. It is these kinds of issues that we would be very keen to work with the relevant bodies to help educate researchers on their local obligations and how they might need to adapt or declare limitations on data access when they publish their work. Where these frameworks prevent or limit data release, these limitations should be made clear at the time of publication to anyone who reads the paper.

Q: I need to make my data "available on request" because of privacy concerns but my institution does not have a Data Access Committee and the IRB is not willing to take this on. What should I do?

A. While PLOS strongly believes that data should be freely available, we recognize that in some instances patient privacy or other concerns may preclude making data freely available to all, and that not all institutions have Data Access Committees at this time. If that is the case, please note details of your situation when submitting your data availability information to PLOS. We are still investigating potential solutions to this issue and until we have determined a standard course of action, we will work with authors for whom this presents a challenge. We encourage academic institutions without a data access committee to consider the importance of this function for their research and researchers.

Q: I want to run additional analyses on my dataset for future studies. What portion of my dataset do I need to make freely accessible?

A. The PLOS Data Policy states that the minimal dataset needs to be made available. The "minimal dataset" consists "of the dataset used to reach the conclusions drawn in the manuscript with related metadata and methods, and any additional data required to replicate the reported study findings in their entirety." This does not mean that you must submit your entire dataset, or absolutely all raw data collected during an investigation, but that you must provide the portion that is relevant to the specific study.

Q. I don't want to share my data because another researcher may “scoop” me using my dataset.

A. PLOS believes that after publication (in particular, after publication in an Open Access journal), data underlying a study should be available for re-use by others. This is not just our view: it was the principle underlying the very first journal: the motto of the UK's Royal Society can be roughly translated as “never take someone's word for it”. Many institutions and funding agencies (e.g., NIH, Research Councils UK) already require data relating to publications be stored for up to ten years and to be shared with the minimal possible restrictions.

Q. What if I cannot provide accession numbers or DOIs at submission?

A. We understand that accession numbers or DOIs may not be available until a paper has been accepted, although once a decision to accept has been made, we cannot proceed with publishing the paper until the accession numbers or DOIs are received.

Q. What if the data are needed for peer review but are not yet publicly available?

A. Many repositories permit private access for review purposes, and have policy for public release at publication. If this is not possible, authors can provide the data via other means, such as zipped files via email, dropbox etc. Please contact the relevant journal office or data@plos.org for assistance.

Q: What if I am submitting an individual patient data meta-analysis?

A: For an individual patient data meta-analysis, authors should make freely available data that can be shared legally and ethically. If authors cannot share the data, they should provide information on how to access the datasets used in the study as specified under “data available on request” or “data available from

third parties."

Q. How can I make a very large amount of data available through a repository without a DOI?

A. Authors are encouraged to provide the information they have regarding accessing the data from a repository (eg, URLs, registration numbers, and other identifying information). Authors should submit their manuscript and PLOS will work with them to determine how best to present the information.

Q. I cannot afford the cost of depositing a very large amount of data. What should I do?

A. PLOS encourages authors to contact their institutions if they have difficulty providing access to the data underlying the research. Many institutions and funding agencies (e.g., NIH, Research Councils UK) share this view and make data sharing a requirement. Regardless, authors facing these challenges are encouraged to submit their manuscript and PLOS will work with them to help find a solution.

Q. I don't have the time or institutional infrastructure to deposit all my data. What should I do?

A. We encourage you to contact your institution if you have difficulty. If you have questions, please contact the relevant journal office or data@plos.org.

Q. What if data are found to not be accessible or other issues are found after publication?

A. As is our current policy, we will follow up with the authors and take action as necessary. We reserve the right to issue corrections, notifications, or retractions when authors do not comply with our policies.

Q. My paper was under consideration at PLOS before March 3, 2014. Does it need to adhere to the new policy?

A. Manuscripts submitted prior to March 3rd will not be required to include the Data Availability information or include a Data Availability statement (although authors who would like to include are welcome to).

Q. What if my question is not addressed here?

A. Authors who have questions about the policy, or readers who have difficulty accessing data, are encouraged to contact the relevant journal office or data@plos.org.

Back to Contents

PLOS

Ambra 2.9.15 Managed Colocation provided by Internet Systems Consortium.

Privacy Policy | Terms of Use | Advertise | Media Inquiries

Publications

PLOS Biology
PLOS Medicine
PLOS Computational Biology
PLOS Currents
PLOS Genetics
PLOS Pathogens
PLOS ONE
PLOS Neglected Tropical Diseases

plos.org

Blogs

Collections

Send us feedback

California (US) corporation #C2354500, based in San Francisco



RightsLink®

Home | Account Info | Help



Title:	Efficiency of Adaptive Temperature-Based Replica Exchange for Sampling Large-Scale Protein Conformational Transitions
Author:	Weihong Zhang and Jianhan Chen
Publication:	Journal of Chemical Theory and Computation
Publisher:	American Chemical Society
Date:	Jun 1, 2013

Copyright © 2013, American Chemical Society

Logged in as:
Weihong Zhang

LOGOUT

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK CLOSE WINDOW

Copyright © 2014 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com



RightsLink®

Home | Account Info | Help



Title:	Accelerate Sampling in Atomistic Energy Landscapes Using Topology-Based Coarse-Grained Models
Author:	Weihong Zhang and Jianhan Chen
Publication:	Journal of Chemical Theory and Computation
Publisher:	American Chemical Society
Date:	Mar 1, 2014

Copyright © 2014, American Chemical Society

Logged in as:
Weihong Zhang

LOGOUT

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK CLOSE WINDOW

Copyright © 2014 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com