ABSTRACT

Title of dissertation: UNCOVERING BIOPHYSICAL PROPERTIES AND FUNCTIONS OF DISORDERED HISTONES USING COMPUTER SIMULATIONS

Hao Wu
Doctor of Philosophy, 2020
Biophysics Program

Dissertation directed by: Professor Garegin Papoian
Institute of Physical Science and Technology
Department of Chemistry and Biochemistry

It is a crucial task for the continuation of every species to safely store genetic information and precisely pass it on to the next generation. For all the eukaryotes including humans, this mission is carried out by chromatin, a polymer chain consisting of repeating structural units called the nucleosome, in which 146 bp of DNA wraps around a histone protein octamer. In a typical eukaryotic cell, about two meters of DNA is compacted into a micrometer-sized nucleus, where transcription and replication activities are regulated in part via modulating chromatin's condensation. A comprehensive understanding of chromatin structure and dynamics provides the necessary foundation for explaining the genome organization, which, for example, will help better understand the mechanisms of diseases caused by epigenetic modifications. As the building blocks of chromatin and nucleosome, the histone proteins are the key players in chromatin structure regulation and epigenetic control. How-

ever, studying histones has been challenging in part because histone tails lack well-defined structures, staying disordered when carrying out many functions. In this dissertation, we focus on exploring the biophysical mechanisms related to these intrinsically disordered histones using computer simulations, carefully comparing our results with related experiments. We present recent progress in the development and applications of state-of-art molecular dynamics force fields for disordered histones and histone-DNA interactions. We used these force fields to investigate the structural, dynamical, and thermodynamical properties of various disordered histones, including histone tails, linker histones, and histone monomers, in the nucleosomal environment. Our investigations have uncovered the structural preferences and binding/folding dynamics of these disordered histones, which provide novel insights into how they aid chromatin condensation.

UNCOVERING BIOPHYSICAL PROPERTIES AND FUNCTIONS OF DISORDERED HISTONES USING COMPUTER SIMULATIONS

by

Hao Wu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Garegin Papoian, Chair
Dr. Yamini Dalal
Professor David Fushman
Professor Silvina Matysiak
Professor Sergei Sukharev, Dean's Representative
Professor Pratyush Tiwary

Dedication

To my dearest parents

for their love and support in my whole life

Acknowledgments

Seven years ago, a young, simple, and naive Chinese physics undergraduate student flew across the Pacific Ocean to the University of Maryland, purely out of curiosity for life sciences and American lifestyle. Today, after seven years of efforts and struggle, success and failure, laughter, and tears, he is approaching the end of the entire doctoral journey. As Bob Dylan sang, "How many roads must a man walk down, before you call him a man?" I believe Ph.D. is such a challenging yet beautiful road and would like to thank all the people who supported me towards its final destination.

First of all, I would like to thank my advisor Prof. Garegin Papoian. Garyk, thank you so much for taking me as your graduate student and providing opportunities for doing all these exciting research projects at your lab. I cannot imagine finishing this thesis without your warm-hearted advice and suggestions. Your motivation for solving big scientific questions and the ability to integrate interdisciplinary sciences are always the best example for me to learn. Your instructions are not limited to becoming a good scientist, but also on how to be an honest, humble, and dedicated person, which is the goal I will try my best to achieve for my whole life.

Next, I would like to appreciate the world-class academic mentorship from all the amazing faculties at UMD and beyond. I would like to thank Prof. Sergei Sukharev for bringing me into the world of cell biology and all the insightful advice on my research directions. I am grateful to Dr. Yamini Dalal for her warm-hearted help on my main thesis projects and suggestions on the way of a qualified biophysics

Ph.D. I would like to thank Prof. David Fushman for taking me as a rotating student and helping me learn more about NMR experiments and histone folding. I would like to appreciate Profs. Silvina Matysiak and Pratyush Tiwary for all the inspiring discussions on computational biophysics and efforts in serving as my committee members. I am grateful to Prof. Wolfgang Losert for admitting me to the biophysics program. I would like to thank Profs. John Weeks and Chris Jarzynski for teaching me the important principles of thermodynamics and statistical physics. I would like to thank Prof. Jeffrey Klauda for his kind guidance and instruction during my candidacy oral exam. I am grateful to Prof. Peter Wolynes for his dedicated help and amazing ideas on the AWSEM-IDP project. I would like to thank Prof. Daniel Butts for teaching me computational neuroscience and accepting me to do the rotational research at his lab. I would like to appreciate Profs. Carter Hall, Matt Severson, and David Buehrle for their patient supervision in my teaching assistant experiences. I am grateful to Prof. Yujie Wang for opening the door to the physical sciences as my undergraduate research advisor. I feel so honored to be mentored and work with these knowledgeable and respectable scholars.

I also want to thank all our outstanding administrative staff for their dedicated work and assistance. I am grateful to Souad Nejjar for her devoted help in my defense, candidacy, fellowship, and so many essential things on the way of my graduation. I would like to thank Star Jackson, Debbie Jenkins, and Teri Schuler for welcoming my arrival at UMD, arranging biophysics seminars, and processing travel reimbursements. I would like to thank Paulina Alejandro and Pauline Rirksopa for taking care of my TA stipends, health benefit, and tax documentation. You have

been indispensable in my Ph.D. experience.

Next, I would love to acknowledge my dearest lab members and peer students at UMD. I would like to express my sincere appreciation to Haiqing Zhao, who has been helping and collaborating with me as a more experienced labmate and biophysics peer student for seven years. Thanks, Haiqing. Your motivation for pure science and interesting ideas will always inspire me. I am grateful to David Winogradoff for his kind guidance and advice as a senior labmate and a true friend, especially during my hard time at the beginning of the first project. I would like to thank Aravind Chandrasekaran, Mary Pitman, Qin Ni, Carlos Floyd, Haoran Ni, James Komianos, Ignacia Echeverria, and Konstantin Popov for all the insightful discussions and suggestions for all my thesis projects. I would like to thank Aram Davtyan and Davit Potoyan for their dedicated efforts in developing AWSEM and explaining its basic concepts to me. I would like to thank my outstanding peer students at IPST: Hongcheng Xu, Guang Shi, Yang Shen, Zhiyue Lu, Ruiliang Bai, Yuwei Cui, Xunnong Xu, Siddharth Sharma, and Tsung-Jen Liao for their considerate help in coursework, lab choice, research problems, job hunting and so on. I appreciate the time learning and working together with these adorable young scientists.

Outside the scientific field, there are also many people helping me along my doctoral journey. I am grateful to Wenbo Yu and Rui Wang for being my best friends in College Park and giving me so much physical and emotional support. I would like to appreciate Jitao Zhang and Kunyi Zhang for tolerating my waywardness and keeping caring for me in so many ways. I am grateful to Jie Peng, Guang

Chen, and Rujun Wang for giving me their warm hands in my most difficult time three years ago. I would like to thank Chao Shen for teaching me how to be a good researcher and a cook. I would like to thank Ramy ElDelgawy and Qiansen Yang for dragging me out of loneliness and fear during my first days in America. I am grateful to Hsiang-Ling Hsiao, Yi-Hsieh Wang, Yurong He, Linda Qin, and all the members at Maryland Bible Study Group for their generous gift and help. I appreciate Paula Huang and Michael Huang for all the advice and suggestions on my research direction and job search. I am grateful to all my friends at the unofficial UMD badminton club, SJTU Alumni Association, and Washinkan Kendo club for organizing so many amazing recreations. I am grateful to my dearest old friends from all over the world: Shuyan Wang, Fangheyue Ma, and Xin Huan, for closely contacting and concerning about me for more than ten years. Thank you all again, my friends. Let us toast to all tomorrow's parties.

Besides, I would like to thank the important people in many fields who give me huge emotional relief and courage during my Ph.D., although they might not know me at all. I would like to appreciate Prof. Philip Warren Anderson for showing me the beauty of statistical physics by writing "More is Different". I would like to thank Éric Rohmer and Hong Sang-soo for directing great movies and TV shows that make me realize the truth of the interpersonal relationship. I would like to appreciate everyone at Kyoto Animation for making "K-On!", "Lucky Star", and "Nichijou" that bring me joy after intense work. I appreciate Bethesda Game Studios for producing "The Elder Scrolls V: Skyrim" that saved my real life by immersing me into a virtual life three years ago. I am grateful to everyone at Leaf

for creating “White Album 2” which makes me introspect and bravely face all the choices I made. I would like to appreciate everyone at Gadio and Busang Podcast for accompanying me every day on the radio and preventing me from self-isolation. I would like to thank all my online friends at douban.com for sharing your ideas and concerning about my mental health. Although we have never met with each other in person, I enjoy encountering all of them and their marvelous works that comfort and inspire me.

At the very end of the acknowledgments, I would like to express my deepest appreciation to my family: my dearest parents, grandparents, aunt, and uncle. You raised me and provided me with valuable opportunities to explore the world. I feel guilty for being so far away from home for these years. But please remember, you are always in my heart. Once thinking of your unconditional love, I shall not fear anything. Because home, wherever you are, is my most peaceful harbor forever.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3SPN.2	3-Site-Per-Nucleotide.2 DNA model
Å	Angstrom (10^{-10} m)
AFM	Atomic Force Spectroscopy
AWSEM	The Associative Memory, Water Mediated, Structure and Energy Model
AA	Amino Acid
bp	base pair
BMRB	Biological Magnetic Resonance Bank
CD	Circular Dichroism
CG	Coarse-Grained
cryo-EM	cryogenic Electron Microscopy
COM	Center Of Mass
DNA	DeoxyriboNucleic Acid
de novo	"start from beginning" modeling
D_{e2e}	end-to-end Distance
FM	Fragment Memory
FRET	Förster Resonance Energy Transfer
GROMACS	GROningen MAchine for Chemical Simulations
H4 tail	H4 histone tail
HA	Homologue Allowed
HE	Homologue Excluded
HFP	Histone Fold Protein
IDP/IDR	Intrinsically Disordered Protein/Region
in vivo	"within the living", in living organisms
in vitro	"within the glass", in a laboratory environment
in silico	"within silicon", performed on a computer
LAMMPS	Large-scale Atomic/Molecular Massively Parallel Simulator
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
NRL	Nucleosome Repeat Length
NTD/CTD	N/C-terminal Domain
PaaA2	Pare2-Associated Antitoxin 2
PDB	Protein Data Bank
PMF	Potential of Mean Force
PTM	Post-Translational Modification
REMD	Replica-Exchange Molecular Dynamics
R_g	Radius of gyration
RMSD	Root-Mean-Square Deviation
RMSIP	Root-Mean Square Inner Product
SAXS	Small-Angle X-ray Scattering

sm-FRET	single molecule Fluorescence Resonance Energy Transfer
STRIDE	secondary STRuctural IDEntification
VMD	Visual Molecular Dynamics
WHAM	Weighted Histogram Analysis Method

Chapter 1: Introduction

Genetic inheritance is a fundamental process for all species enabling transmission of their key characteristics to their offspring generation after generation. Eukaryotes or more complex organisms, including human beings, store their genetic material - deoxyribonucleic acid (DNA) - in the form of chromatin, a polymer chain composed of histone proteins and DNA. In a typical eukaryotic cell, $\sim$ 2 meters of DNA are packaged into the micrometer-sized nucleus via some hierarchy of chromatin structures. Understanding this complicated chromatin condensation process is an important step for uncovering the biophysical mechanisms of the genome organization, including epigenetic diseases. Using extensive computer simulations, compared with related experiments, this dissertation investigates histone proteins, the essential building blocks of chromatin, to elucidate their role in regulating chromatin structure and also explores post-translational modifications from the biophysical perspective.

In this introduction, I first present an overview of the biological significance of chromatin and describe its hierarchy of structures. In the second part, I introduce histones based on their subtype, sequence, structure, and function. In the third part, I discuss the main methodology used in this dissertation, the AWSEM protein force

field, and review its applications in chromatin and histone related studies. Lastly, the contents of the subsequent chapters are outlined.

1.1 Overview of Chromatin Structure and Functions

Over a century ago, Walther Flemming observed some small fibers in cell division and named them as "chromatin" based on their staining by certain dyes [1]. The behavior of the most condensed chromatin complex - chromosomes - were soon found to be directly connected with rules of inheritance by a series of experiments and theories by Theodor Boveri, Walter Sutton, and many other pioneer scientists [2]. Since then, many biochemical studies discovered that chromatin consists of DNA and histones, suggesting that their functions are related to genetic inheritance [3]. However, a closer look at the chromatin structure for a better understanding of its biological significance was limited by experimental techniques for a long time. In 1973, 20 years after the milestone discovery of DNA's double-helix structure [4], the first microscopy-based visualization of chromatin revealed its repeating subunits in a "beads on the string" manner [5–7]. The structural insights resulting from the discovery of these subunits, called the nucleosome, established the fundamental understanding of chromatin structure at an atomic length scale [8,9]. With the rapid progress of modern microscopy techniques, especially cryogenic electron microscopy (cryo-EM) [10] and stochastic optical reconstruction microscopy (STORM) [11], now researchers can visualize three-dimensional chromatin structure in different phases of the cell division cycle at very high spatial resolutions up to nanometer length

scale [12–14], marking a new era of chromatin research.

Based on these extensive prior works, now we know that eukaryotic chromatin is condensed in complex ways (Figure 1.1). At the smallest scale, about 146 DNA base pairs (bp) wrap around a histone octamer, which consists of four types of histones H3, H4, H2A, and H2B, forming a nucleosome with a diameter of 11 nm. The nucleosome particles are connected by linker DNA in a "beads-on-a-string" manner. Another histone called H1 binds to the nucleosome to help fold the nucleosomal arrays into a condensed chromatin fiber by regulating the short-range inter-nucleosomal interactions. The diameter of this compacted chromatin fiber has been assumed to be $\sim$ 30 nm [15]. However, whether this "30 nm fiber" really exists *in vivo* has been highly doubted [16–18], with a new cryo-EM study finding that most chromatin fibers are thinner in interphase and mitotic cells [13]. The thin fibers are further compressed and folded as thicker fibers, which are tightly coiled to create a chromatid of the chromosome [19].

These hierarchies of higher-order structure not only compact the long chromatin fibers so that they fit in the small cell nucleus, but also control genome accessibility, the degree to which nuclear proteins can contact DNA, to regulate gene expression and replication in different phases of the cell cycle. During interphase, chromatin is generally less condensed, allowing DNA to be readily accessed by transcription factors and translated into proteins. When cells start to divide in metaphase, chromatin becomes the highly condensed chromosome, which ensures genetic material is safely passed to the offspring cells. Even within a single phase, chromatin accessibility can be extensively regulated by post-translational modifi-

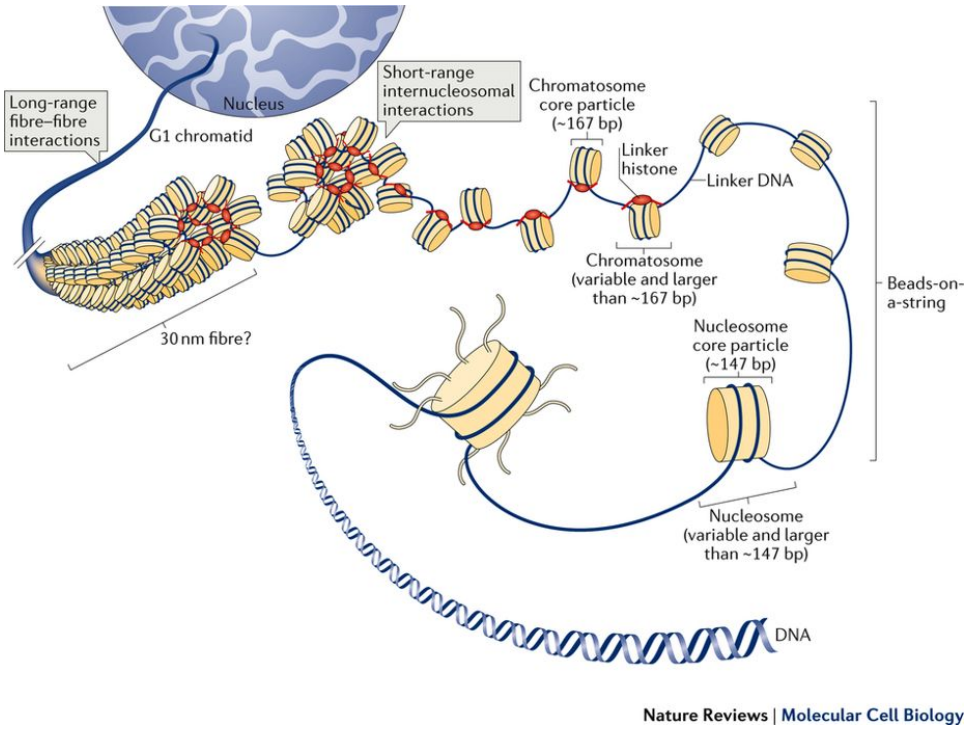


Figure 1.1: **Chromatin is folded via multiple condensation levels.** The different levels of chromatin structure are illustrated from smaller to larger length scales (from bottom to up). Reprinted from [20], with permission from Springer Nature, Nature Reviews Molecular Cell Biology, Copyright (2017).

cations (PTM) [21, 22]. A recently released database of chromatin accessibility in human tumor cells paves the way for future cancer research on a large epigenetic scale [23].

1.2 Histones: Chromatin Scaffold Proteins

Eukaryotic DNA cannot condense into chromatin without binding to the histone proteins. Histones proteins are highly alkaline and conserved across many species. There are five major types of histones: H1, H2A, H2B, H3, and H4, where H1 is the linker histone and the other four are core histones. In this section, we introduce the structure and functions of different histone subtypes.

1.2.1 Core Histones

Eukaryotic core histones do not exist as stable monomers at physiological conditions [24, 25]. After translated into a polypeptide chain, one histone monomer needs to interact with another monomer to fold into a well-defined structure, which is assisted by multiple folding chaperones via intricate pathways [26]. The resulting heterodimer has a unique structural motif called "histone fold" [27, 28], where a disordered N-terminal tail is followed by three α-helices connected by two loops. In particular, H3 binds to H4 and H2A binds to H2B in a head-to-tail fashion, called "hand-shake" motif, to form the H3-H4 and H2A-H2B dimers. Two H3-H4 dimers come together to form a tetramer, serving as a target for two H2A-H2B dimers to bind and form a histone octamer. Approximately 146 bp of DNA wraps around this

histone core complex 1.65 times to produce the nucleosomal particle [29]. Figure 1.2 shows a graphical illustration of the structural hierarchy of nucleosome assembly. Meanwhile, the actual nucleosome assembly process during DNA replication *in vivo* is very complicated [30], involving many other chaperones and enzymes as revealed by a series of experimental studies [31–38].

As the fundamental building blocks for chromatin condensation and organization in the nucleus, the core histones have highly conserved sequences and structures among most eukaryotic species [40]. Sequences of these four histone types are not homologous, but their structures all adopt the histone fold motif [28]. Further experiments and evolutionary studies reveal that the histone fold is not only restricted to eukaryotic histones, but can also be found in some archaeal histones [41], transcription factors [42], and other DNA-binding proteins [43]. It was proposed that all these proteins share a common ancestor and belong to the same "histone fold superfamily", with a similar dimerization pattern and DNA compaction functions [28]. Then during evolution, they differentiated into distinct types of proteins, serving specific functions in more complex organisms, but their characteristic histone fold structures were conserved.

1.2.2 Histone Tails

The core histones are not fully ordered: their terminal regions lack well-defined secondary or tertiary structures. The amino-terminus (N-terminal domain) of H3, H4, and H2B, and the carboxyl-terminus (C-terminal domain) of H2A and H2B are

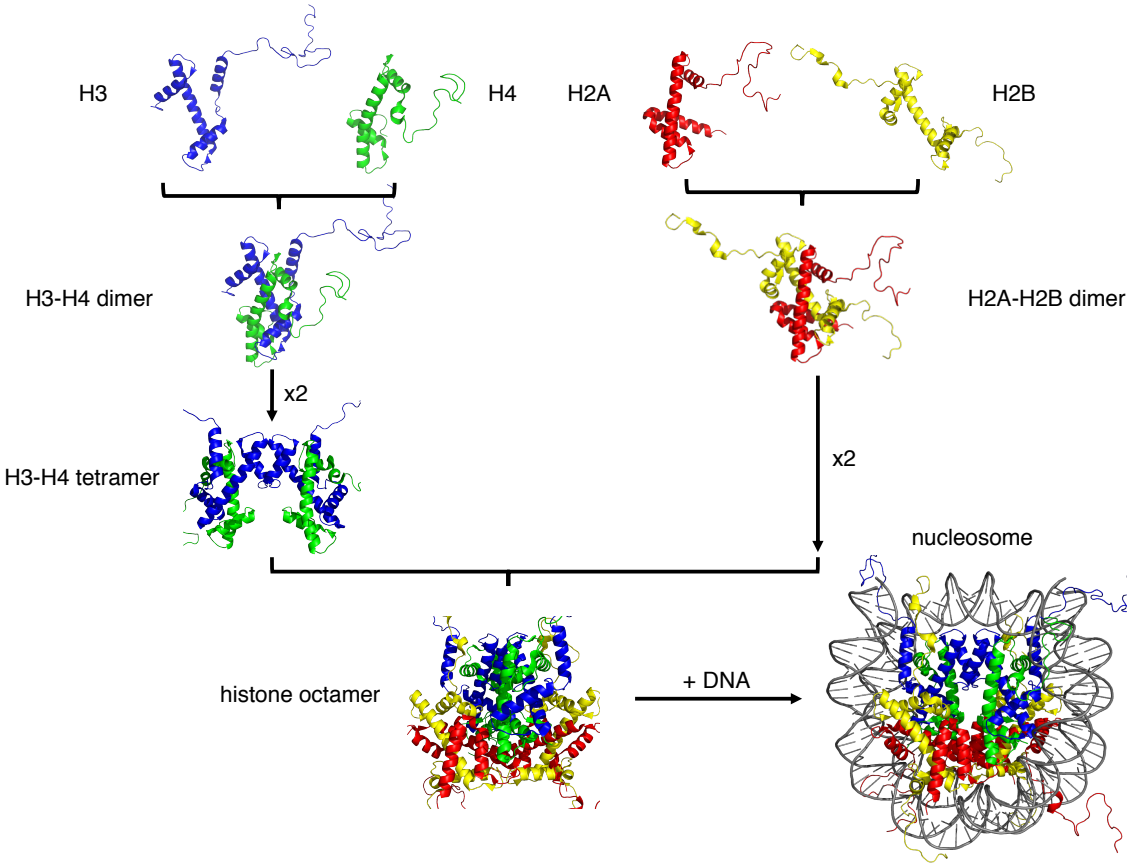


Figure 1.2: **Histone and nucleosome structure.** Four types of core histones form a complex and are wrapped around by DNA to form a nucleosome. The figure only represents the structural hierarchy of nucleosome based on the crystal structure at 1.9 Å resolution (PDB: 1KX5) [39]. It does not include detailed information on how the nucleosome is assembled.

disordered at the physiological condition. These parts, therefore, are categorized as "intrinsically disordered proteins/regions" (IDP/IDR), although they can form partially folded structures at some specific conditions (temperature, pH, counterions, presence of membrane, etc) [44]. These relatively short ($\sim$ 10 - 40 amino acids) regions are comprised of numerous positively charged residues, mainly lysines and arginines. Histone tails contain also many glycines that tend to disrupt helical structure formation. They tend to protrude outside the nucleosomal core and contact negatively-charged DNA and also neighboring nucleosomes primarily through electrostatically mediated interactions. These contacts regulate inter-nucleosomal distances and higher-order chromatin structures [45].

Many major biological functions of histone tails are regulated through post-translational modifications (PTM) [46]. Since the first histone acetylation was reported by Allfrey *et al.* [47], more and more types of PTMs, including acetylation, phosphorylation, methylation, and ubiquitination, among others, are mostly found on histone tails [48] (see Figure 1.3 for histone tail sequences and possible PTM sites). Post-translational modifications change the physical and chemical properties of the side chains on the involved residues, such as their size, hydrophobicity, and partial charge. These modifications will result in changes in affinity for chromatin binders to the nucleosome, extending genetic information content and finely regulating gene transcription and expression [49, 50]. For instance, the acetylation on H4K16 inhibits the formation of compact chromatin fiber and activity of remodeling enzyme ACF, thus modulating chromatin structure and transcription [51]. Enhanced, reduced, or incorrect PTM on histone tails are commonly found in various

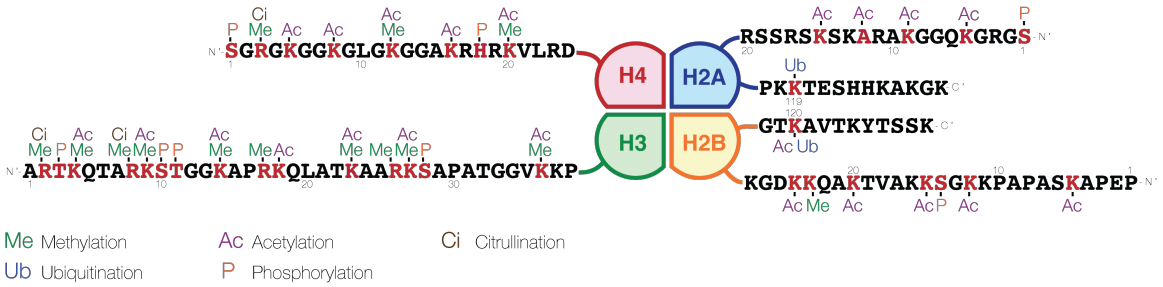


Figure 1.3: **Histone tail sequences and possible PTM sites.** Sequences of core histone tails (N-terminal tails for all four core histones, C-terminal tails for H2A and H2B). Residues with possible PTM sites are in red. Five different PTM are labeled with different colors as in the legend. Adapted from [52], with permission from Springer Nature, Nature Medicine, Copyright (2011).

tumor cells and can cause epigenetic diseases [52]. Hallmark PTMs, such as H4K20 trimethylation and H4K16 acetylation, lead to CpG island hypermethylation and global genomic hypomethylation, imparting aberrant transcription patterns [53].

1.2.3 Linker Histone H1

Along with the other four histones mentioned above, the linker histone H1 is another essential type of histone in eukaryotic chromatin. Although falling into the "histone" category, H1 markedly differs in structure, sequence, and function from the core histones [54]. The H1 consists of ∼200 amino acids, including an ordered globular domain and highly disordered N- and C-terminal domains (NTD and CTD) [55]. H1 globular domain adopts a "winged helix" motif, where three α-helices are connected by β-strands or turns [56] (see Figure 1.4), as opposed to the well-known histone fold motif in the core histones. The sequences of H1 are much

less conserved than the other four histones, especially in the terminal domains [57]. A handful of H1 subtypes with relatively low sequence similarity exist in different sematic cells even within the same species, including one called H5 that is only found in avian red blood cells [58].

The distinct features of H1 serve for its unique functions in the formation of chromatin higher-order structure, gene expression, and DNA repair [54]. Instead of being part of the nucleosome core particle, H1 binds with the nucleosome near the entry/exit site of the linker DNA to form a fundamental complex in metazoan chromatin called chromatosome [59] (Figure 1.4). This binding location may vary for different histone variants and environmental conditions and has a strong regulatory effect for the chromatin condensation [60–62]. Moreover, the number of H1 molecules bound to one nucleosomal particle (H1-nucleosome stoichiometry) affects nucleosome repeat length (NRL) and gene expression [63, 64] in distinct species and cell types. Most H1 variants, especially the disordered NTD and CTD, are also frequent targets of major types of post-translational modifications, which are associated with many epigenetic cellular mechanisms [65].

1.3 Computationally Modeling Chromatin with AWSEM

The ubiquitous and important chromatin has been inspiring great research interest in the last sixty years, during which the fundamental understanding of its structure and function has greatly improved [3]. As reviewed above, many pioneering scientists have revealed the structure of chromatin at different length

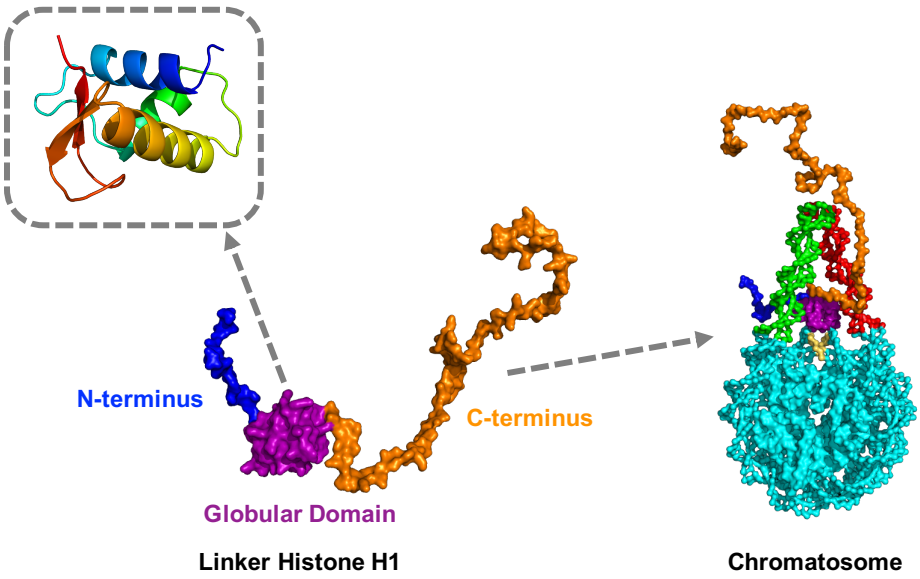


Figure 1.4: **Structure of H1 and chromatosome.** The tripartite structure of linker histone H1 is shown as a surface representation with corresponding color labels. The zoomed-in secondary structure elements of the globular domain are displayed in the upper-left box. H1 can bind with a nucleosome particle to form a chromatosome (right panel). All the molecular structures except for the H1 NTD/CTD are from the crystal structure (PDB ID: 5NL0) [66], where the disordered NTD and CTD are homology models generated with MODELLER [67].

scales, benefiting from the rapid progress of instruments and technology. Various experimental methods, such as X-ray crystallography [9], nuclear magnetic resonance (NMR) [68], cryogenic electron microscopy (cryo-EM) [69], single-molecule fluorescence resonance energy transfer (sm-FRET) [70], atomic force spectroscopy (AFM) [71], stochastic optical reconstruction microscopy (STORM) [14], and Hi-C map [72], have been successfully applied to study the structure, dynamics, and functions of histones, nucleosome, and chromatin. However, it is still challenging for the current experimental techniques to report on the dynamics of nucleosomal arrays at near-atomic resolution, which could shed light on the mechanisms of chromatin condensation at different conditions. Moreover, although the disorder histone regions have proven to be important for chromatin condensation and transcription regulation, their structural preferences and dynamical behaviors are poorly characterized, because it is difficult to track and identify them experimentally. It is therefore desirable to develop novel methods to overcome these obstacles, enabling new advances in chromatin biophysics.

While experimental studies remain the most direct and empirical approach for chromatin research, computer simulations have risen as important complementary and predictive tools to push the field boundaries [73]. As more accurate chromatin models emerge and with the revolution of computing power, theoretical and computational scientists have made significant progress in solving a wide range of chromatin-related problems at different time and length scales [74]. Recent improvements of parallel algorithms combined with powerful hardware even make it possible to carry out short simulations of a chromatin system of a billion atoms,

including around 427 nucleosomes (83 kilobases of DNA) [75]. However, the current chromatin atomistic simulations are limited by huge computational expense, while mesoscale models largely simplify the structure of chromatin and lack near-atomic chemical accuracy. It is therefore desirable to develop a coarse-grained chromatin model that is computationally efficient, enabling simulations of several nucleosomal particles at sufficiently long timescales. Accurate modeling of disordered histone regions is particularly important. Below, I will first focus on introducing one of the most efficient and accurate computational models for this purpose - AWSEM - and discuss its achievements and future challenges.

1.3.1 AWSEM: A Coarse-Grained Protein Model

The associative memory, water-mediated, structure and energy model (AWSEM) [76] is a coarse-grained (CG) protein force field. As the latest successor of a series of protein models [77–80] based on the funneled free energy landscape theory [81], AWSEM represents one amino acid with three beads (C_α, C_β, and O) and uses the implicit solvent to accelerate simulations. The Hamiltonian of AWSEM is designed to mimic the necessary intra- and inter-molecular interactions for a realistic protein backbone and side chain geometry, which is elaborated below:

$$V_{\text{AWSEM}} = V_{\text{backbone}} + V_{\text{contact}} + V_{\text{burial}} + V_{\text{Hbond}} + V_{\text{FM}}, \tag{1.1}$$

$$V_{\text{backbone}} = V_{\text{contact}} + V_{\text{chain}} + V_{\chi} + V_{\text{rama}} + V_{\text{excl}}, \tag{1.2}$$

where the V_{backbone} is responsible for peptide's backbone potential, consisting of V_{contact} for backbone atoms connectivity, V_{chain} for the bond angle, V_{χ} for chirality,

V_{rama} for desired dihedral angle distribution (*i.e.* Ramachandran plot [82]), and V_{excl} for excluded-volume effect. V_{contact} dictates contact interactions between side chains far away from each other and takes protein- and water-mediated effect into consideration [80]. V_{burial} accounts for the effect of an amino acid being buried inside the protein or exposed on the surface. V_{Hbond} is responsible for the formation and stabilization of secondary structures. V_{FM} introduces a bioinformatic structural bias called "fragment memory" to help local structure formation. See the supporting information of Davtyan *et al.* [76] for detailed formulae of each term above.

The original purpose of AWSEM and its predecessors was to study how an amino acid sequence folds into a 3D protein structure, one of the most important and enigmatic biological processes from the past century to today [83–85]. To predict protein structure accurately, the parameters of AWSEM were optimized based on the celebrated folding funnel energy landscape theory [81,86–88]. This theory states that globular proteins have a single "native state", located at a deeply "funneled" free energy minimum, surrounded by a somewhat rugged energy landscape consisting of non-native local minima. With this fundamental theory as the guideline, the parameters in AWSEM are trained to maximize the ratio of folding temperature over the glass transition temperature (T_f/T_g), so that the energy landscape of each training protein becomes most deeply funneled and least frustrated [89]. The coarse-grained (CG) representation of AWSEM, using three beads per residue, greatly accelerates the simulation while still keeping a near-atomic structural resolution. Compared with other CG protein models, a special feature of AWSEM is the water-mediated interactions, which realistically mimic the behaviors of different side chains

within an implicit solvent environment [80]. Another innovation was the fragment memory term, which applies short structural bias on the target protein based on prior knowledge of protein structures with similar sequences. This idea was inspired by associative neural networks [90]. Combining all these features above, AWSEM proves to be one of the most accurate and efficient force fields for protein 3D structure prediction [91].

As an open-source package based on the widely-used molecular simulation platform LAMMPS [92], AWSEM is highly extendable and allows for further modifications. To cope with different systems, many versions of AWSEM have been developed in combination with numerous advanced tools, such as direct coupling analysis [93], atomistic simulations [94], small-angle X-ray scattering experiments [95], and structural refinements [96], to predict protein structure at high precision. In addition to protein folding, AWSEM is also extensively used to study many other protein-related problems, including membrane proteins with an implicit lipid bilayer [97], dimer interface association [98], multidomain protein misfolding [99], amyloid aggregation [100], and protein-DNA interaction [101]. The major features and applications of AWSEM are summarized in Figure 1.5.

1.3.2 Using AWSEM to Simulate Chromatin: Progress and Challenges

As an accurate and efficient protein model, AWSEM also has great potential for rational chromatin simulations. It is very expensive to simulate chromatin with

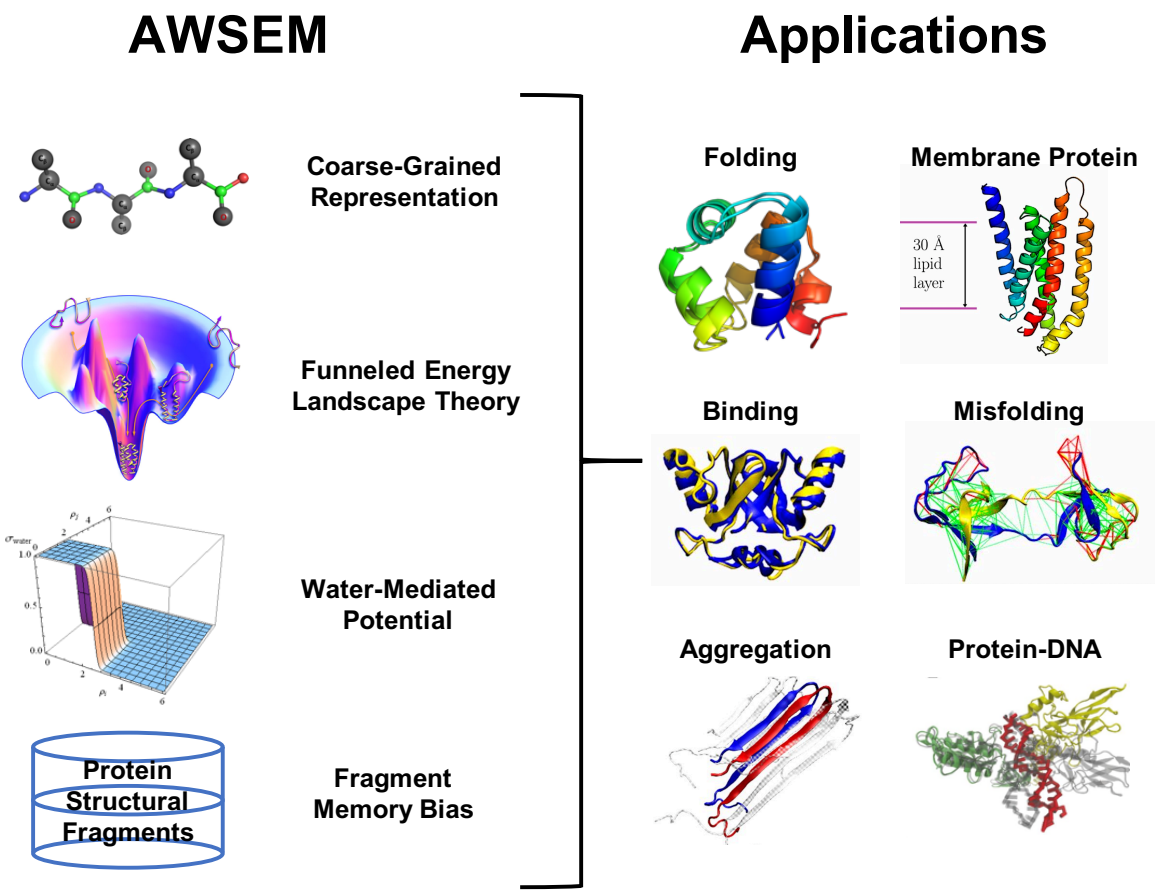


Figure 1.5: **The major features and applications of AWSEM.** As a protein force field with main features listed on the left panel, AWSEM has been successfully used to study a wide range of protein-related scientific questions listed on the right panel. Most of the subfigures, except for the coarse-grained representation and fragment memory bias, are reprinted or adapted from [76, 85, 97–101], with permission from multiple publishers. Copyright (2012-2016), American Association for the Advancement of Science, American Chemical Society, and National Academy of Sciences.

atomistic force fields, especially with the explicit solvent model, for time scales that are of biological interest [102, 103], because even the fundamental unit of chromatin - the nucleosome - is a large system in atomistic representation (∼300,000 atoms in total). On larger length scales, some mesoscale models are suitable to simulate nucleosomal arrays and chromatin fibers [74, 104, 105], but detailed molecular information is ignored by extreme simplifications in these models. Benefiting from its well-chosen coarse-grained representation, AWSEM can simulate nucleosomes at affordable computational cost, at near-atomic resolution.

Among different parts of chromatin, histone assembly can be directly modeled by the standard AWSEM since DNA is not involved. Zhao *et al.* performed a series of computational studies on histone assemblies with AWSEM to investigate the different dynamic and thermodynamic features of canonical and centromere histone dimers and octamers [106, 107]. AWSEM can also be reasonably combined with a CG DNA model [108, 109] to simulate a protein-DNA complex. Along with other progress in developing AWSEM-DNA potential [101, 110, 111] that uses the 3SPN.2 CG DNA model [109], Zhang *et al.* simulated a single nucleosomal particle and explored its unfolding and DNA unwrapping free energy landscape [111]. These chromatin-related works shed light on various mechanisms of chromatin from unique computational perspectives, meanwhile supported by or even predicting wet-lab experimental results [112, 113]. The success of these studies indicates an intriguing possibility of simulating larger scale nucleosomal arrays, possibly including other chromatin related proteins such as transcription factors and chromatin remodelers [114].

As a continuation of previous successful AWSEM simulations of histones and the nucleosome, we would like to move forward to investigate the structure and dynamics of large-scale nucleosomal arrays. To comprehensively understand the mechanism of chromatin condensation, we need to simulate an array of more than one nucleosomal particles, including globular/disordered histones and DNA. With these simulation trajectories of nucleosomal arrays with high-resolution structural details, it will be promising to investigate the functions of each component of chromatin in regulating inter-nucleosomal conformations and the resulting changes of the chromatin fiber structure. However, there have been still two major obstacles to effectively simulating nucleosomal arrays with AWSEM.

The first bottleneck was the inability of AWSEM to model the disordered parts of chromatin because AWSEM was initially designed to predict globular protein structure and has some potentials that are incompatible with IDP. The histone tails and linker histone NTD/CTD belong to intrinsically disordered proteins (IDP) [115], having very different structural and dynamic features from globular proteins. For instance, most IDPs are found in more extended global conformations and contain less ordered secondary structure elements than globular proteins. Without additional modification and calibration, AWSEM will incorrectly treat IDP as globular proteins, reducing the accuracy of IDP and chromatin simulations. Therefore, it was desirable to develop a new version of AWSEM that could treat both globular and disordered proteins.

The second challenge to overcome has been the lack of an accurate and efficient protein-DNA potential in AWSEM. The current development of AWSEM-DNA is

still in an early stage, with many practical limitations such as the lack of systematic calibration for nucleosomes and lack of parallel computing efficiency. Previous related studies [101,110,111] of various protein-DNA systems used different strategies to develop their own specially modified version of AWSEM. It is crucial to converge on a universal and well-calibrated protein-DNA interaction potential in AWSEM for simulating general protein-DNA complexes, including nucleosomal arrays. More importantly, AWSEM-DNA has only been used to model relatively small molecular complexes. To simulate nucleosomal arrays is a much more computationally expensive task, which requires additional parallel computing improvements such that biologically meaningful timescales can be reached.

This dissertation research has striven to overcome the two obstacles mentioned above by developing a new generation of the AWSEM force field and applying it to study different chromatin related problems. We anticipate that the body of this work will pave the way for more accurate and efficient simulations of large-scale nucleosomal arrays.

1.4 Outline of Chapters

In this dissertation, we have developed and applied the AWSEM force field to investigate the structure and dynamics of disordered histones and reveal their relevant biological functions in chromatin folding. Chapter 2-4 are all independent, yet logically connected research projects on different types of histones, from the H4 histone tail to the linker histone H1 and histone monomers/dimers.

In Chapter 2, we report on a new generation of the AWSEM force field, AWSEM-IDP, that can be used to study disordered proteins. This new model can simulate IDPs with high accuracy and efficiency. It was validated by comparing with various experimental and computational studies of H4 histone tail and other IDPs. AWSEM-IDP serves as a basis for further IDP researches, especially the disordered histone subdomains in the nucleosomal environment.

Chapter 3 is a comprehensive study on the dynamics and functions of the disordered linker histone H1 when bound to the nucleosome. Using a state-of-art hybrid protein-DNA model, including AWSEM-IDP, we simulated the full-length H1 in complex with a nucleosomal particle to investigate their binding conformational preferences and dynamics. We found that the H1 disordered terminal domains particularly confine the conformation and dynamics of both the globular domain and linker DNA arms, resulting in a more compact and rigid H1-nucleosome complex. Our results uncover the dynamics of H1 disordered domains at a near-atomic resolution for the first time.

In Chapter 4, we report on the fundamental mechanism of histones folding from dynamical and thermodynamical perspectives. Using molecular dynamics simulations and NMR/circular dichroism experiments, we found that the histone monomers cannot fold independently. Instead, two histone monomers undergo a "coupled-folding upon binding" process to fold as a dimer. Based on our further simulations and analyses, we propose that this folding mechanism may be operational for other proteins with the histone fold structure, enhancing the complexity of corresponding functions in higher organisms during evolution.

Finally, Chapter 5 summarizes these studies and suggests future research directions in computationally modeling chromatin condensation using AWSEM. Appendices A, B, and C elaborate on the supporting information of Chapters 2, 3, and 4.

Chapter 2: Development and Application of AWSEM-IDP: A Force Field for Intrinsically Disordered Proteins

This chapter is based on the published work of the authors: *Hao Wu, Peter G. Wolynes and Garegin A. Papoian; AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins; The Journal of Physical Chemistry B, 122(49):11115-11125 (2018)* [116]

2.1 Introduction

Intrinsically disordered proteins (IDPs) and structured proteins containing intrinsically disordered regions (IDRs) are ubiquitously found in the proteomes of higher organisms. These elements carry out a variety of critical biological functions despite their structurally disordered nature [115, 117–121]. Extensive progress in investigating IDPs in the last twenty years has highlighted the limits of the classical fixed structure-function paradigm of molecular biology [122], suggesting several new mechanisms, including "coupled folding and binding" along with others [123–128]. While many studies treat IDPs as conformational ensembles having no well-defined secondary and tertiary structures, it has more and more been recognized that certain types of structural order are encoded in the overall disor-

dered state [102, 106, 129–132]. Hence, numerous experimental studies, employing solution-based techniques, such as nuclear magnetic resonance (NMR) [133] and small-angle X-ray scattering (SAXS) [126], have investigated various classes of IDPs. Generally, these methods provide only ensemble-averaged measurements, although more recently complete distributions of structural variables have become available through single-molecule observations based on Förster resonance energy transfer (FRET) [133]. Nevertheless, it remains challenging to capture experimentally the detailed structural dynamics of IDPs. To address this blind spot, and also help guide further experiments, computational approaches have come to play an increasingly important role, in illuminating the molecular nature of IDPs' conformational ensembles [130, 131, 134, 135].

Different computational methodologies have been used to study IDPs, including models that require specific experimental inputs as well as *de novo* molecular dynamics (MD) simulations at the all-atom level which do not use experimental data [136–138]. Methods such as the energy-minima mapping and weighting method (EMW) [139], ASTEROIDS [133], and the ENSEMBLE program [140], usually take experimental constraints to generate a best-fitted IDP ensemble through back-calculation from specific experiments on each system. These methods almost exclusively rely on fitting the experimental data, using only relatively simple energy functions to describe the chain. In contrast, traditional atomistic MD simulations attempt to model IDPs at the same high structural resolution employed in developing X-ray crystal structures. *De novo* approaches provide the possibility of discovering new conformations of completely unexpected types. Yet, obtaining adequate sta-

tistical sampling of the configuration space of IDPs utilizing purely atomistic MD simulations is quite challenging, requiring very long runs at a high computational expense. These studies have also already given the impression that atomistic force-field inaccuracies are significant for modeling IDPs [141–144]. Coarse-grained (CG) models, on the other hand, replace atomistic details with a coarser description that can be more rapidly simulated. By having higher computational efficiency coarse-grained models dramatically broaden the exploration of the conformational space of IDPs and IDRs [98, 145–150]. Most of the earlier CG models used for IDP simulations have employed generic polymer physics approaches and have neither been systematically benchmarked against experiments reporting on the properties of the structural ensembles sampled by IDPs nor, alternatively, by being benchmarked against comprehensive atomistic simulations. Therefore, it is desirable to develop a transferable CG force field for IDPs that aims to reproduce the salient structural features of IDPs, namely the extended geometry of the chain and the nature of their conformational disorder, in particular, by taking into account sequence-specific effects.

We might ask: Why has it been so hard to model "intrinsically disordered proteins"? Two fundamental reasons arising from their statistical physics:

First: "Intrinsic disorder" implies, by its definition, large fluctuations in structure. If one were to use thermodynamic perturbation theory to treat any errors in individual terms in the force field as perturbations on the perfect model, one would see that the sensitivity of any average structural feature to small errors in the potential directly depends on the correlated fluctuations in the unperturbed ensemble

of those structural features to be monitored and the energetic error terms (which are themselves structural variables!). Thus, *ipso facto*, an intrinsic disorder with its large structural fluctuations then implies there will be high sensitivity of the average structure of an IDP to modeling errors.

Second: Sequence disordered polymers exhibit many phase transitions that all meet near multiple points in the phase diagram: collapse, folding, liquid crystal order, etc [151]. Making a small error in the force field that fails to locate properly which part of the diagram the molecule is in thus has big consequences. This difficulty is unlike what happens for simulations of well-ordered globular proteins where it can be assumed, at the start, that the system is weakly fluctuating in the fully ordered part of the phase diagram and thus in the most insensitive part of the phase diagram! Simulations can be started from near-native conformations, biasing the system to fluctuate less for example. Remember Pauling was led to the main themes of secondary structure in well-ordered proteins without even considering such important forces as hydrophobicity! His force field was poor, but his structures were excellent.

At the same time that we see that the structure of an intrinsically disordered protein must be sensitive to the details of the force field owing to the big fluctuations of IDPs, likewise thermodynamic perturbation theory by the same token also suggests that the thermodynamic consequences of making these structural errors are small because the system is soft! In other words, there will be entropy/energy compensation. In this way, we see that very often less than perfect structural simulations will still get global mechanism right due to compensating contributions.

Biology and thermodynamic can forgive modest modeling errors.

The Associative-memory, Water-mediated, Structure and Energy Model (AWSEM) [76], which has been successfully applied to study globular protein folding [76], protein recognition and binding [98,106], aggregation [152], membrane proteins [97,153], protein-DNA association and functional transitions [101,110,111], provides a promising opportunity for predictive simulations of IDPs. It is a coarse-grained model that has been developed using concepts from the energy funnel theory of folding of globular proteins and structural data on well-folded proteins. It contains both physics-based potentials and bioinformatics-motivated local structural biasing terms [91]. The synergy among the biophysical and bioinformatic potentials provides the needed flexibility for AWSEM's further development to simulate IDPs. For instance, specific potential terms can be tuned to regulate the formation of protein secondary or tertiary structures. The local structure biasing term can be drawn from diverse data sources including experiments [76] or from *in silico* simulations of more elaborate fully atomistic models [94, 154].

In this article, we introduce "AWSEM-IDP", a new coarse-grained model specialized for simulating IDPs. It is based on the standard AWSEM, but three major changes have been made: (1) the weights of the hydrogen bonding potentials have been modified to reflect the reduced propensities for secondary structure formation characteristic of IDPs; (2) the local fragment library is derived from either IDP experiments or structural ensembles obtained from atomistic simulations; (3) a novel radius of gyration (R_g) term is added into the AWSEM Hamiltonian to regulate finely the collapse of the chain, enabling delicate control of its size fluctuations.

We tested the performance of AWSEM-IDP on two examples: the H4 histone tail (H4 tail) and ParE2-associated antitoxin 2 (PaaA2). H4 tail is 26-residues long and largely lacks any secondary structure. PaaA2 is 71-residues long but has clear secondary structural elements in an extended chain geometry. Both sets of simulations show significant agreement between the AWSEM-IDP generated ensembles and the corresponding experimental measurements or atomistic simulations. We also carried out energy-landscape analyses of these IDPs, comparing the energy distributions with those found for globular proteins. Finally, we used thermodynamic perturbation theory to calculate how IDPs' structural propensities depend on the details of the potential, finding the responses of structural variables to force field perturbations are at least an order of magnitude larger than the same responses for globular proteins.

Altogether, this study introduces AWSEM-IDP as a transferable model for simulating various types of IDPs, whose computational efficiency allows broad, well-converged sampling of the disordered ensemble. It should be particularly useful for simulating mixed biomolecular complexes that contain IDPs or IDRs along with well-folded structural segments.

2.2 Methods

2.2.1 AWSEM-IDP Hamiltonian

AWSEM-IDP is a specialization of AWSEM [76], a coarse-grained protein force field, where each amino acid is represented by the positions of C_α, C_β (H for

glycine) and O atoms. The coordinates of other heavy atoms are calculated following the ideal peptide geometry. The total Hamiltonian of AWSEM-IDP, which largely coincides with that of AWSEM, is given below,

$$V_{total_{IDP}} = V_{backbone} + V_{contact} + V_{burial} + V'_{Hbond} + V'_{FM} + V_{R_g}, \tag{2.1}$$

where $V_{backbone}$ ensures protein-like backbone connectivity and stereochemistry, $V_{contact}$ and V_{burial} describe water- and protein-mediated tertiary interactions and also the preferences for each amino acid to be buried or exposed. Detailed definitions of the first five terms are provided in the references [76,91]. In this work, we report on tuning the parameters for the V'_{Hbond} and V'_{FM} terms for IDP simulations, hence denoting these terms with a single prime notation. We also introduce here a new V_{R_g} term, which allows for the control of the collapse and the size fluctuations of an IDP chain. In the following subsections, these three terms are introduced in greater detail.

2.2.1.1 Hydrogen Bonding Potential

V'_{Hbond} is a sum of three hydrogen bonding terms, as shown in eq 2.2,

$$V'_{Hbond} = \lambda'_{\beta} V_{\beta} + \lambda'_{P-AP} V_{P-AP} + \lambda'_{helical} V_{helical}, \tag{2.2}$$

where V_{β} favors formation of well-structured hydrogen bonding networks in β-sheets, V_{P-AP} enables a protein chain to adopt approximate parallel or anti-parallel β-sheet conformations before more detailed hydrogen bonds are fully formed,

and $V_{helical}$ controls the formation of hydrogen bonds in α-helices. λ'_{β}, λ'_{P-AP} and $\lambda'_{helical}$ indicate the corresponding weights of these potentials. These terms have been described in detail elsewhere [76, 91].

IDPs show a lesser propensity to form secondary structure elements than do globular proteins. Collapsing itself tends to increase secondary structural content [151]. Doubtless, the tendency to form secondary structure is also reinforced by a minimally frustrated correlation between secondary and tertiary interactions [151, 155–157]. In our test simulations we found that with the default V_{Hbond} setup, IDPs already tend to form more stable secondary structures than are seen in experiments. Therefore, while for AWSEM-IDP we kept the functional forms of these hydrogen bonding terms, we have re-calibrated the relative weights of these terms, namely λ'_{β}, λ'_{P-AP} and $\lambda'_{helical}$, such that the resulting α-helix and β-sheet propensities are more appropriate for IDPs and IDRs (see Appendix A for further details of this calibration).

2.2.1.2 Fragment Memory Potential

V'_{FM} is a bioinformatic fragment memory ("FM") potential that structurally biases short fragments of the protein chain, typically 3 - 9 residues at a time, towards conformations that are based on "memory" structures. In AWSEM the latter memory terms have been selected by matching the fragment sequence to sequences of proteins in the globular protein structural database, usually selected from the PDB (see Figure 2.1),

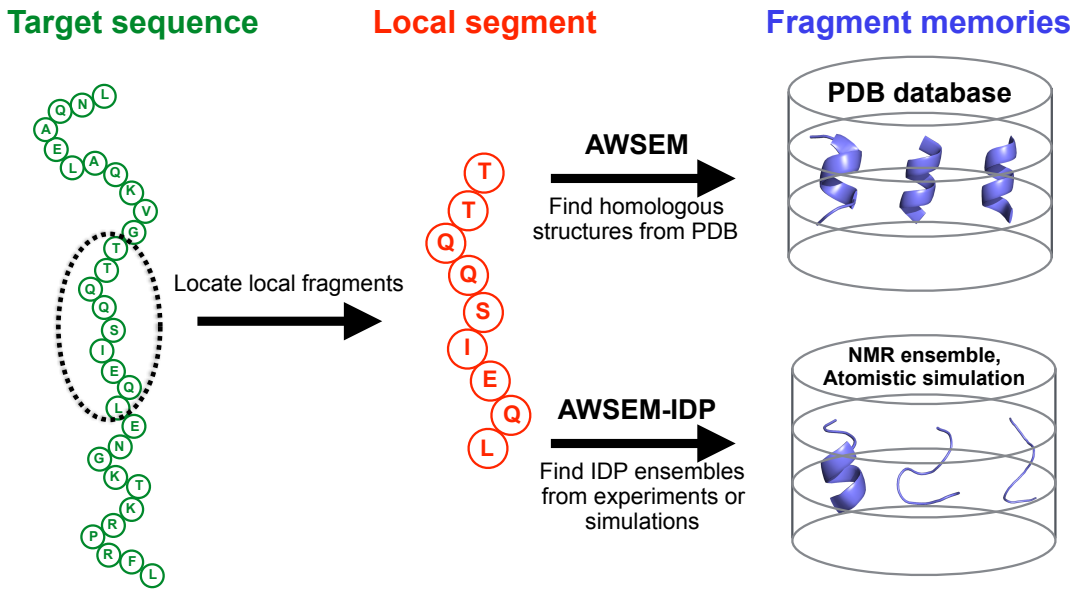


Figure 2.1: **Schematic diagram of the fragment memory terms in AWSEM and AWSEM-IDP.** In both AWSEM and AWSEM-IDP, the target sequence (green) is assigned into short local segments (red). Then structural fragments called "memories" (blue) are chosen to bias the local segment. The original forms of AWSEM search for fragment memories from the PDB database, while AWSEM-IDP utilizes NMR ensembles or atomistic simulation trajectories to construct the fragment library. The example sequence shown here is the amino-terminal domain of phage 434 repressor (PDB ID: 1R69).

$$V'_{FM} = -\lambda'_{FM} \sum_{m} \sum_{ij} \exp\left[-\frac{(r_{ij} - r^m_{ij})^2}{2\sigma^2_{ij}}\right]. \tag{2.3}$$

In eq 2.3, the outer summation is carried out over aligned fragment memories, while the inner summations are carried out over all possible pairs of C_α and C_β that are separated by two or more residues. r_{ij} is the distance between ith C_α and jth C_β atom in the target sequence and r^m_{ij} is the corresponding distance in the memories. $\sigma_{ij} = |i-j|^{0.15}$ is the tolerance factor for gauging similarity between two distances. λ_{FM} sets the overall weight of the FM term.

As shown in Figure 2.1, the fragment memory library in the standard form of AWSEM is constructed from structures in the PDB database, with the specific memory conformations selected based on the similarity between the target sequence and the individual memory sequences. This approach is not optimally suited for studying IDPs or the disordered regions in globular proteins [76], because most structures in the PDB, which serve as templates for potential fragment memories, belong to globular proteins having a significant amount of secondary structure that has been partially induced by the supporting tertiary structure. This bias, in turn, typically will result in an overestimation of secondary structure formation in IDPs or IDRs. Therefore, in AWSEM-IDP we have decided to rely instead on taking fragment memories either from the representative snapshots of the target protein carried out using atomistic simulations, similar in spirit to the way it was done in atomistic AWSEM [94, 158], or by taking them from the experimentally obtained structural ensembles for these peptide fragments, since these ensembles are expected

to describe more accurately the realistic conformational details [159].

2.2.1.3 R_g Potential

The standard AWSEM can accurately predict the size of globular proteins. However, for some disordered proteins highly extended in physiological conditions, the AWSEM Hamiltonian tends to over-collapse the IDP chain, especially for longer ones. To remedy this deficiency, we propose a new V_{R_g} term in AWSEM-IDP in the following expression,

$$V_{R_g} = \frac{DN + \alpha(R_g - \gamma R_g^0)^2}{1 + \beta(R_g - R_g^0)^4}, \tag{2.4}$$

where N is the number of residues in the target sequence, R_g^0 is the desired value for the average of the radius of gyration, which typically can be determined by the average R_g values from FRET or SAXS experiments. α and β modulate the width of the V_{R_g} curve, thus modulating the degree of allowed fluctuation in degree of collapse. The intensity of the potential is controlled by D. The values of α, β, and D should be carefully selected to reasonably modulate the R_g distribution of simulated molecules. In this paper, these parameters were determined by fitting the experimental or atomistic simulation data empirically (see more in the following part on force field parametrization).

The major advantage of this potential over the more commonly used alternatives, such as harmonic or Morse potentials for the radius of gyration, is the ability of this term to sculpt more flexibly the potential profile (Figure 2.2). In particu-

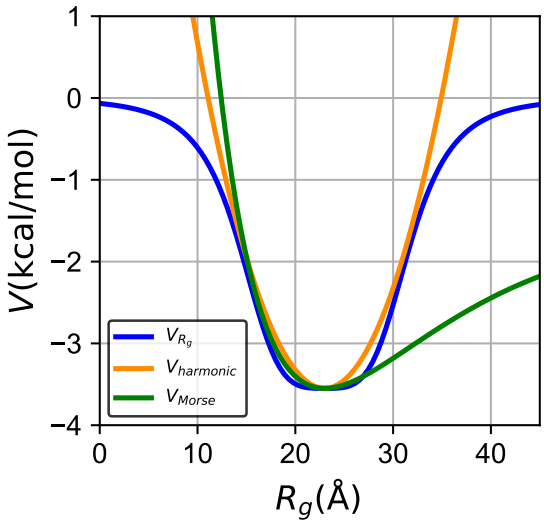


Figure 2.2: **The V_{R_g} potential introduced in this work allows for more powerful control of chain fluctuations of IDPs than the harmonic and Morse R_g potentials.** The harmonic potential (orange) and Morse potential (green) tend to restrain R_g in a narrow potential well with a steep energy barrier away from the ideal R_g value. In comparison, V_{R_g} (blue) shows a shallow bottom and allows the chain to escape the restraint.

lar, this potential allows the simulated chain to overcome the unrealistically large energy barriers for expansion of the chain that arise from the harmonic well much in the way the Morse potential does. The chain collapse potential therefore allows accessing extended chain conformations characteristic of IDPs, while separately controlling the extent of the fluctuations at the bottom of the potential profile. In this way, eq 2.4 goes beyond what can be done with the Morse potential. The width, depth, and slope of the V_{R_g} can all be carefully adjusted to regulate both the general collapse and the distribution of accessible conformations of the IDP chain. Hence, this potential could be useful not only in IDP simulations but also in computational studies of various other biological and artificial polymer chains, where more precise control of collapse dynamics is needed.

2.2.1.4 Force Field Parametrization

We used a two-step protocol to calibrate the modified and new parameters in AWSEM-IDP. Since V'_{Hbond} and V'_{FM} both account for local structure, these two terms were parameterized first. After local secondary structures were reproduced sufficiently and faithfully close to the targets from atomistic simulations or experimentally determined structural ensembles, the parameters in V_{R_g} were subsequently optimized. The parameters obtained for the V'_{Hbond}, V'_{FM}, and V_{R_g} terms in the current model are listed in Table 2.1. A detailed description of the parametrization procedure is provided in Appendix A.

Table 2.1: **Typical IDP parameters used in AWSEM-IDP**

Term	Parameter	Value	Unit
V'_{Hbond}	λ'_{β}	1.0	kcal/mol
	λ'_{P-AP}	1.0	kcal/mol
	$\lambda'_{helical}$	1.2	kcal/mol
V'_{FM}	$\|i-j\|_{min}$	3	
	$\|i-j\|_{max}$	12	
	λ'_{FM}	0.001 - 0.002	kcal/mol
V_{R_g}	D	-0.2 $\sim$ -0.8	kcal/mol
	α	0.001	kcal/mol·Å^{-2}
	β	0.0005 $\sim$ 0.003	Å^{-4}
	γ	1.0 $\sim$ 1.3	

2.2.2 Testing Models

The wild-type N-terminal H4 histone tail (H4 tail) and ParE2-associated antitoxin 2 (PaaA2) are both well-studied IDPs with important biological functions in regulating eukaryotic chromatin folding [160, 161] and prokaryotic cell growth and death [162, 163], respectively (Figure 2.3). These two IDPs were chosen as the test systems to evaluate the performance of AWSEM-IDP because they have quite distinct chain lengths, with rather different characteristics of their respective conformational ensembles. H4 tail is relatively short, with a small fraction of secondary

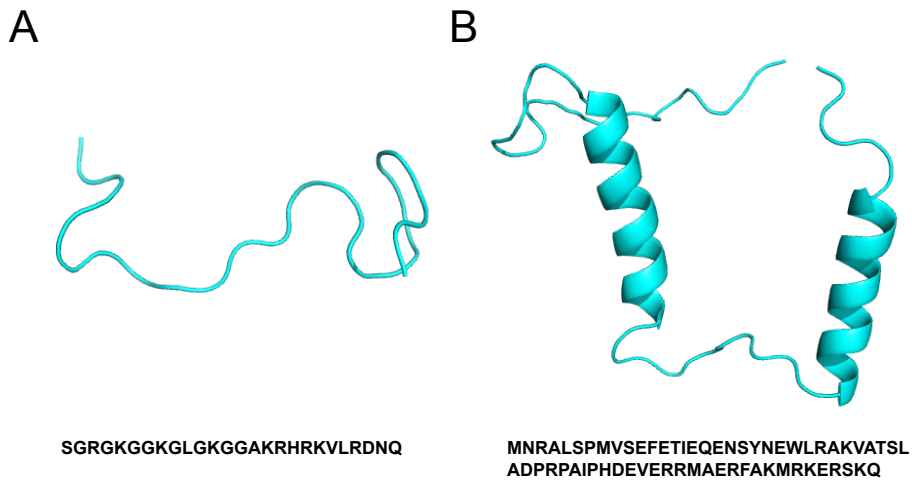


Figure 2.3: **The experimental structures and sequences of H4 tail and PaaA2.** (A) X-ray crystallography structure of the H4 tail in the context of the encompassing nucleosomal particle (PDB ID: 1KX5). (B) NMR ensemble structure of PaaA2 (PDB ID: 3ZBE). Amino acid sequences are shown under the corresponding structures.

structure elements, while PaaA2, on the other hand, is a longer IDP and is more extended with two preformed α-helices. The specific parameters for these targets are listed in Appendix A.

2.2.3 Simulation Details

We performed all molecular dynamics simulations using the open-source simulation package LAMMPS (Feb 2016 version), in which both the original AWSEM and AWSEM-IDP codes have been implemented [76]. We used non-periodic shrink-wrapped boundary conditions and the Nose-Hoover thermostat. The simulation timestep was set at 2 fs. We unfolded the initial structure at 800 K to generate a

random peptide chain as the initial conformation and then slowly cooled down the system from 800 to 300 K over 5×10^5 timesteps. Then we ran 10 production simulations at 300 K for 1.5×10^7 timesteps, recording snapshots every 1000 timesteps. The clock time to perform one production run on a single CPU core is around 1 hour (for H4 tail) and 4 hours (for PaaA2), ∼100 times faster than a standard atomistic MD simulation of the same molecule with explicit solvent. The first 5×10^6 timesteps of trajectories were discarded as the equilibration phase. All the analyses reported below are based on the final 1×10^7 timesteps. The convergence of all simulations was confirmed by the root mean square inner product analysis [164] (see Appendix A and Figure A.1 for details).

2.2.4 Analyses

Since AWSEM is based on a coarse-grained (CG) representation of amino acids, we converted the CG beads into more elaborate atomistic representation based on ideal peptide backbone geometry [76]. We determined secondary structure assignments in simulations by STRIDE [165] implemented in VMD (version 1.9.2). We also calculated the radius of gyration (R_g) and end-to-end distance (D_{e2e}) of structures in the ensembles as global structural metrics using C_α atoms coordinates.

Particularly for PaaA2, we compared our simulations with the NMR and SAXS experimental results that are available online [166]. We determined the secondary structure of PaaA2 from NMR chemical shifts data on the BioMagResBank database [167] (BMRB entry: 18841) with the δ2D method [168], which translates a set of

chemical shifts into probabilities of secondary structure elements. We also computed theoretical SAXS intensities from simulations with CRYSOL (version 2.8.2) [169] and compared with experimental results.

To measure the heterogeneity of ensembles, we employed the distribution of pairwise structural overlap values: q. This pairwise q quantifies the structural similarity between any two conformations and the formula for the pairwise q between structure i and j is given by:

$$q_{ij} = \frac{1}{N_{pairs}} \sum_{a,b} \exp\left[-\frac{(r^i_{ab} - r^j_{ab})^2}{2\sigma^2_{ab}}\right] \tag{2.5}$$

where r^i_{ab} represents the C_α distance between residue a and b for structure i, $\sigma_{ab} = (1 + |a - b|)^{0.15}$ is the resolution of this metric, and N_{pairs} is the number of a and b pairs summed for all possible choices. The range of pairwise q is from 0 to 1, the higher values indicating stronger structural similarity between conformations. Hence, the shape of the pairwise q distribution reflects the heterogeneity of the corresponding structural ensemble. Pairwise q distributions have been used to elucidate the intrinsic conformational preferences and the structural heterogeneities of histone tail conformation ensembles in previous simulation studies [130, 131].

2.3 Results and Discussion

We first describe the results of AWSEM-IDP calculations for the two test systems, H4 tail (Section 2.3.1) and PaaA2 (Section 2.3.2), comparing our results with either atomistic simulations or experimental data. In the third subsection, we then

characterize the secondary and tertiary structural properties of H4 tail and PaaA2 and some well-folded globular proteins from the energy landscapes perspective (Section 2.3.3).

2.3.1 Coarse-Grained Simulations of H4 Tail

We first applied AWSEM-IDP to the H4 histone tail, which is 26-residues long and has no prominent secondary structure elements (Figure 2.3). Winogradoff et al. [132] previously performed atomistic replica-exchange molecular dynamic (REMD) [170] simulations of the H4 tail at 300 K for 6 μs in total, using the amber99SB* [171] and ions94 [172] force fields with TIP3P water model. We randomly selected 100 conformational snapshots from those atomistic simulation trajectories to construct the fragment memory database for AWSEM-IDP.

We characterized the distribution of the chain sizes as measured by the radius of gyration, R_g of the H4 tail, calculated from both the atomistic [132] and the AWSEM-IDP simulations (Figure 2.4A). The average R_g value from AWSEM-IDP (8.2 $\pm$ 0.8 Å) reproduces well its atomistic simulation counterpart (8.6 $\pm$ 1.4 Å). Furthermore, the R_g probability distributions from atomistic and AWSEM-IDP simulations significantly overlap. Both distributions exhibit long tails stretching towards larger R_g values that correspond to extended chain conformations. Note that the R_g biasing potential, introduced in this work, provides enough flexibility to control the complete R_g distribution, not only the average value of the radius of gyration, enabling more accurate modeling of the H4 tail in more extended con-

formations. Interestingly, previous *in silico* studies revealed that histone tails can change their degree of chain condensation with different post-translational modifications and salt concentrations [131,132,173]. By tuning the R_g potential, we can thus nudge histone tails to explore specific regions of chain extension (Figure A.2), providing a basis for more accurate coarse-grained modeling of polynucleosomal arrays in future studies.

We examined the heterogeneity of the structures sampled in AWSEM-IDP and all-atom MD simulations by measuring the distributions of the pairwise q (Figure 2.4B). A similar level of structural heterogeneity was found in both the atomistic and the AWSEM-IDP simulations. The average pairwise q obtained from AWSEM-IDP (0.33 ± 0.07) however is slightly larger than that found from atomistic MD (0.27 ± 0.07), possibly resulting from AWSEM's tendency to over-structure protein chains. [174]

In addition to comparing the global characteristics of chain conformations, we also analyzed the local propensities for the secondary structure formation. Because the H4 tail is intrinsically disordered, lacking a well-defined secondary structure [130–132], we used the combination of coil and turn probabilities as a metric of local structural disorder and heterogeneity (Figure 2.4C). This comparison indicates that AWSEM-IDP replicates the amount of flickering secondary structures observed in atomistic simulations with relatively high fidelity. In both atomistic and AWSEM-IDP simulations, the coil + turn probabilities fluctuate around 90%. This particularly high level of disorder is not surprising because of the high proportion of positively-charged (lysine and arginine) and flexible (glycine) residues in the H4 tail

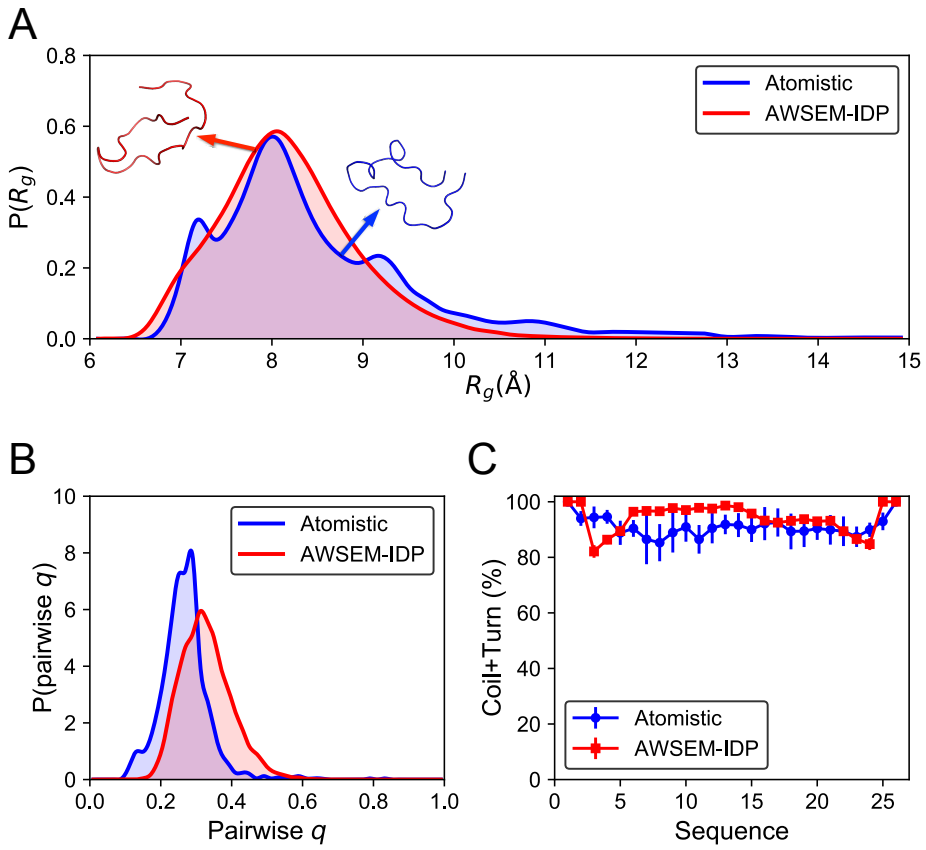


Figure 2.4: **AWSEM-IDP captures reasonably well the structural features of the H4 tail obtained from atomistic simulations.** (A) The probability distributions of R_g indicate similar global overall properties of AWSEM-IDP (red) and atomistic (blue) simulated ensembles. Representative snapshots at the average R_g values of the corresponding distributions are displayed for the atomistic (blue) and AWSEM-IDP (red) simulations. (B) The probability distributions of pairwise q demonstrate a somewhat shifted, but roughly similar structural heterogeneity in AWSEM-IDP and atomistic MD. (C) Local disordered secondary structural propensities (coil + turn) from atomistic and AWSEM-IDP results are close.

amino acid sequence (Figure 2.3A). Overall, these comparisons reveal robust agreement between the conformational ensembles sampled by atomistic simulations and those sampled by AWSEM-IDP simulations. In particular, the results obtained from AWSEM-IDP simulations of the H4 histone tail more faithfully reflect the atomistic results than do those found using the standard AWSEM force field (Figure A.3).

2.3.2 Coarse-Grained Simulations of PaaA2

We also tested the performance of AWSEM-IDP on another disordered protein, PaaA2. PaaA2 is relatively longer (71 residues) than H4 histone tails and has more stable secondary structural elements, namely two α-helices (Figure 2.3B). Sterckx et al. [166] calculated a PaaA2 ensemble based on NMR and SAXS experimental results. We used all 50 structures from their ensemble as the fragment memory library in our subsequent simulations.

We first analyze the AWSEM-IDP sampled ensemble by projecting the conformational space onto two collective variables, R_g and end-to-end distance (D_{e2e}), (Figure 2.5A). The resulting two-dimensional landscape topography reveals three well-connected conformational basins (labeled as i, ii, and iii in Figure 2.5A), with moderate energy barriers of $\sim 2\ k_BT$, suggesting high conformational lability.

To quantify further the simulation results, we compare the locations of the free energy basins to those inferred from the experimentally guided structural ensemble [166]. All three free energy basins are located near the average R_g (20.8 ± 3.2 Å) of the latter ensemble (the vertical green dotted line in Figure 2.5A), show-

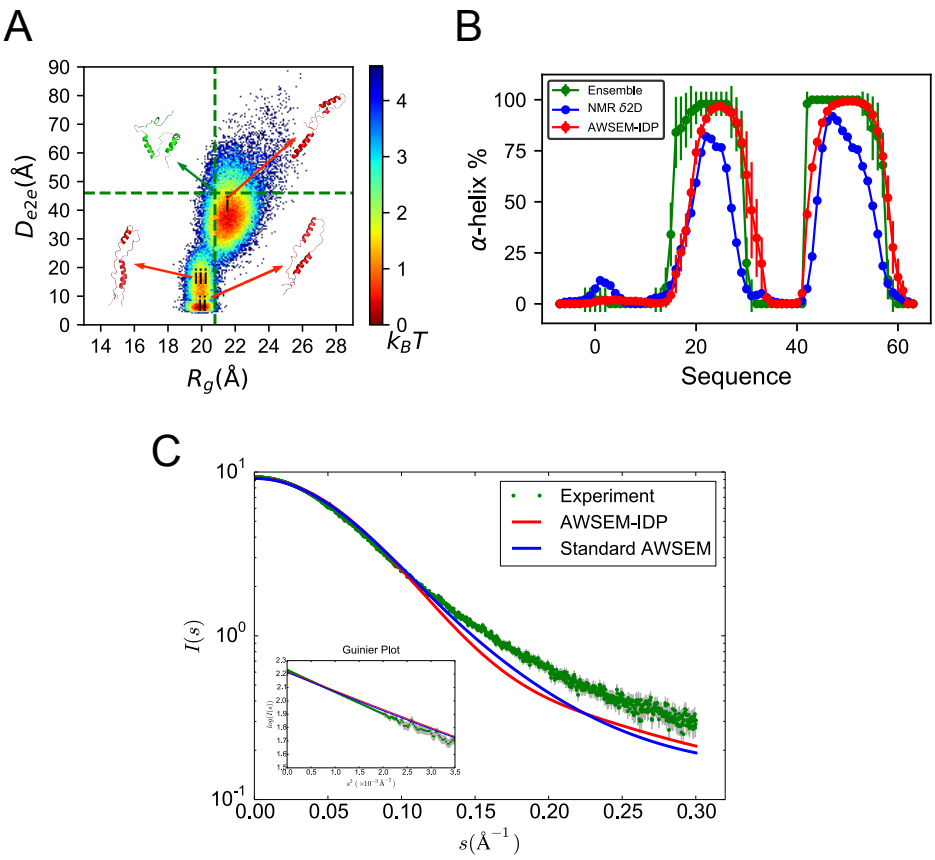


Figure 2.5: **AWSEM-IDP simulations agree well with experiments in the global and local structures of PaaA2.** (A) The free energy landscape of PaaA2 is projected on the coordinates of R_g and D_{e2e}. The vertical and horizontal lines in the figure are the average R_g and D_{e2e} from the experimental ensemble calculated based on NMR and SAXS data in Sterckx et al. [166] Representative structures are shown for the experimental ensemble (green) and different basins in AWSEM-IDP simulations (red). (B) The two helical structures in both experimental ensemble (green) and δ2D calculation from NMR chemical shifts data (blue) are well replicated by AWSEM-IDP simulations (red), with similar positions and probabilities. (C) The AWSEM predicted SAXS curves and the related Guinier plot (inset figure). Experimental errors are labeled by the gray shade.

ing consistency between the chain dimensions of the simulated and experimental ensembles. The experimentally guided structural ensemble gives a range of 30 $\sim$ 62 Å for the end-to-end distance D_{e2e}, covering the largest free energy basin explored by AWSEM-IDP. The average value 46.0 Å is marked with a horizontal green dotted line in Figure 2.5A. Notice that the two other free energy basins have lower D_{e2e} values than the experimental reference. This again could arise from the tendency of AWSEM to over-collapse or represents subpopulations that are too small for experiments to see or absent under a particular experimental condition. As shown in the most probable simulated ensembles, the R_g potential mitigates the over-collapse tendency of the standard AWSEM. It still cannot entirely avoid molecular artificial collapse, but significantly improves model performance.

Beyond global analyses of the chain conformations, we also looked into the local structural details of ensemble members. The PaaA2 experimental ensemble [166] indicates two prominent α-helices, connected by a highly flexible loop (Figure 2.3B). This topology is important for carrying out some of the significant biological functions of PaaA2, such as the molecular recognition driving toxin inhibition [166,175]. To analyze this structural feature, we calculated the average α-helical tendency of all the PaaA2 residues along the simulation trajectory. As seen in Figure 2.5B, PaaA2 has two well-defined α-helices in AWSEM-IDP simulations (shown in red). Moreover, both the positions and structural probabilities of these helices are quantitatively consistent with those in the experimental ensemble (green), as well as with the helical probabilities calculated directly from the NMR chemical shifts data [166] by the δ2D method [168] (blue). This agreement suggests that AWSEM-IDP can

reproduce reasonably well the local structural details obtained from experimental measurements. By contrast, in the standard AWSEM simulations, the first helix comes out as too long in comparison to the NMR determination (Figure A.5B). Reducing the weight of the helical structure formation term ($\lambda'_{helical}$) in AWSEM-IDP is necessary to improve the modeling of secondary structures of disordered proteins. Besides the experiment-guided ensemble and NMR chemical shifts signal, we also compared the simulation results to the SAXS experimental data [166] (Figure 2.5C). The experimental and simulated curve overlap with high precision for $s < 0.10$ Å^{-1}. For greater s, the small deviations indicate less extended structures in simulations than found in experiments. The slopes of the Guinier plot ($log(I(s))$ versus s^2) show similar global structures in experimental and simulated ensembles with close R_g values. This comparison indicates both ensembles have similar global size.

2.3.3 Analyzing IDPs from the AWSEM-Specific Energy Landscapes Perspective

We discuss and quantify in this section the role of those AWSEM-IDP energy terms that provide the most important contributions to the formation of secondary and tertiary structure. In the AWSEM-IDP Hamiltonian, the formation of protein secondary structures primarily results from the effects of two potential terms, V_{Hbond} and V_{rama}. V_{FM} also commonly contributes to local structure formation, but we must remember the fragment memories for IDPs do not necessarily carry directly the signals for conventionally well-defined helical or extended secondary structure.

Residual tertiary interactions in IDPs, on the other hand, arise largely from the terms $V_{contact}$ and V_{burial}, where $V_{contact}$ indicates water-mediated or protein-mediated interactions between pairs of amino acids distant in sequence, and V_{burial} governs the burial preference a particular residue. We define the secondary and tertiary average energies per residue using eq 2.6 and eq 2.7 as:

$$\langle E_{secondary} \rangle = \frac{1}{N} \langle V_{rama} + V_{Hbond} \rangle \tag{2.6}$$

$$\langle E_{tertiary} \rangle = \frac{1}{N} \langle V_{contact} + V_{burial} \rangle \tag{2.7}$$

Following these definitions, we calculated $\langle E_{secondary} \rangle$ and $\langle E_{tertiary} \rangle$ for H4 histone tail and PaaA2, along with these analyses for two globular proteins (PDB: 1R69, 1UBQ) and one mostly globular protein with a disordered tail (PDB: 1UZC). These data are plotted in Figure 2.6. As one could have anticipated, both the H4 histone tail and PaaA2 have higher $\langle E_{secondary} \rangle$ and $\langle E_{tertiary} \rangle$ than the three ordered proteins. Between the two IDPs, PaaA2 has lower average secondary structure energy compared with the H4 tail. The average tertiary energy of the H4 tail is approximately equal to PaaA2, however, with a similar level of fluctuations. This comparison suggests that PaaA2 and H4 tail may potentially belong to two different classes of IDPs: PaaA2 has stable secondary structural elements but is lacking tertiary organization, while H4 tail may be relatively collapsed but lacks stable secondary structures. 1R69 and 1UBQ, which are well-folded globular proteins, are characterized by lower secondary and tertiary structure energies than IDPs have, as expected. 1UZC, which has unstructured segments, shows correspondingly an inter-

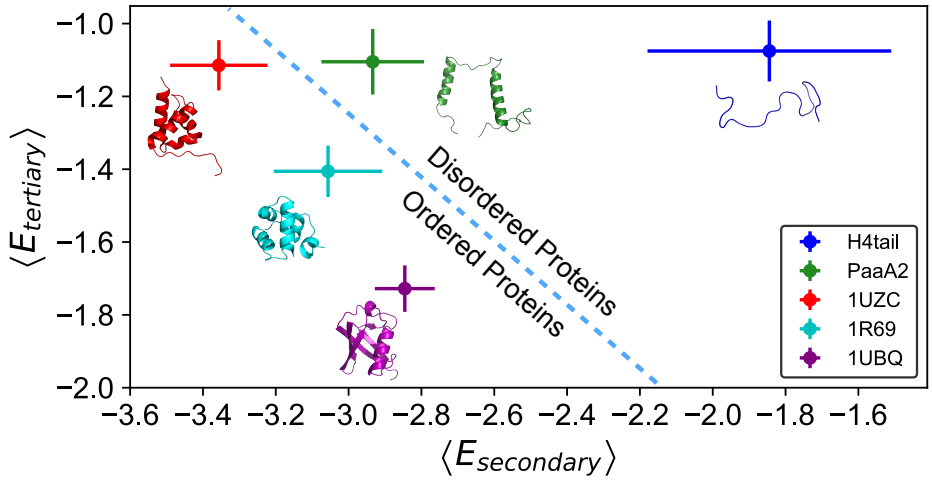


Figure 2.6: **Means and variances of AWSEM-specific energies corresponding to secondary and tertiary structures can efficiently demarcate protein disorder.** The average energies and corresponding standard deviations for secondary and tertiary structures are shown for all simulated proteins (H4 tail: blue; PaaA2: green; 1UZC: red; 1R69: cyan; 1UBQ: purple). Initial conformations of each protein are illustrated accordingly. The dashed line serves as a qualitative border between the ordered and disordered proteins. All energies are in the units of kcal/mol.

mediate behavior: low secondary structure energy but destabilized tertiary structure energy. Table A.1 and Figure A.6 elaborate on additional term-by-term contributions from other terms of the AWSEM Hamiltonian and also temporal evolution of $E_{secondary}$ and $E_{tertiary}$ during MD runs.

We see that the analyses more or less correspond to our intuitions about the differences between IDPs and well-structured globular proteins. Still more telling differences can be seen in the fluctuations and corresponding modeling sensitivities as monitored by susceptibilities or response functions. We calculate the sensitivities of the radius of gyration and the helical occupation probabilities. Both of these, as

we have seen, can be monitored experimentally. The susceptibility of R_g to potential variation is computed as:

$$\chi_{R_g,V_k} \equiv \frac{\partial \langle R_g \rangle}{\partial \gamma_k} = -\beta \langle \delta R_g \delta V_k \rangle, \tag{2.8}$$

while the sensitivity of helical occupations along the sequence,

$$\chi_{h_i,V_k} = -\beta \langle (P_{h,i} - \overline{P}) \delta V_k \rangle \tag{2.9}$$

can be computed on an individual residue basis (where $P_{h,i}$ is an indicator function, being 1 if the residue i is found in helical conformation, and 0 otherwise). Here, the AWSEM potential is assumed to have the following form

$$V_{AWSEM} = \sum_k \gamma_k V_k, \tag{2.10}$$

where the V_k terms represent various types of interactions, and γ_k parameters indicate the corresponding weights. We see in Figure 2.7A that the R_g modeling sensitivities for the IDPs studied are more than an order of magnitude larger than those are for globular proteins. This seems to trace back to considerable sensitivity of secondary structure occupation to the model terms as shown in Figure 2.7B. We see that the fraying ends of helices in PaaA2 are especially sensitive to energy modeling errors. This is where the structure of this IDP fluctuates most strongly.

2.4 Conclusions

In this paper, we introduce AWSEM-IDP, a coarse-grained model tailored for modeling intrinsically disordered proteins. Two terms from the standard AWSEM

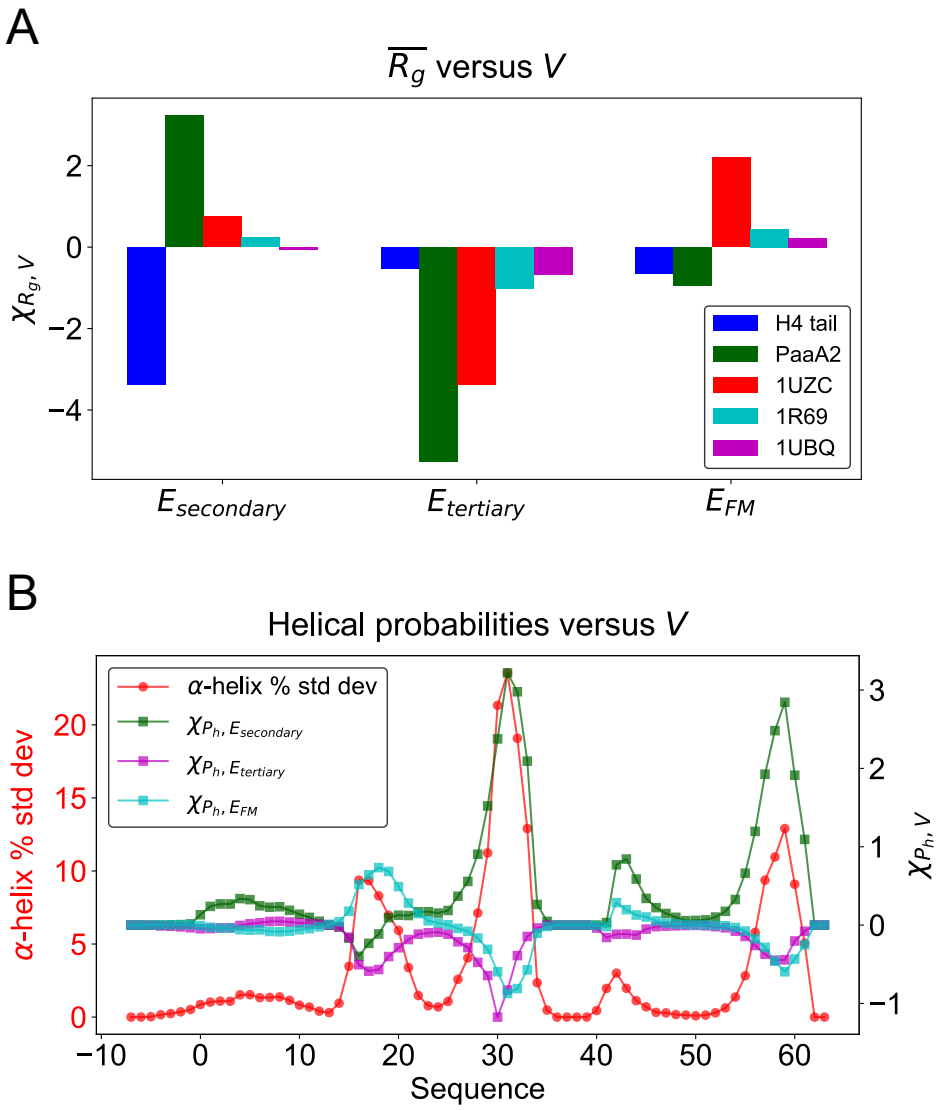


Figure 2.7: **Structural features of IDPs are highly sensitive to the force field potential variation compared with globular proteins.** (A) The susceptibilities of R_g to variation of the potential for IDPs and globular proteins are shown. (B) The susceptibilities of helical occupations along the PaaA2 sequence are primarily determined by the covariance with $E_{secondary}$ (green squares), whose peaks coincide with the locations where the helical probability fluctuation (red circles) reaches maximum. A similar plot with the raw value of helical probability as a reference is given in Appendix A (Figure A.7).

Hamiltonian, V_{Hbond} and V_{FM}, were modified and one new term, V_{R_g}, was added. Lowering the weight of V_{Hbond} diminishes secondary structure formation, thereby better representing the amount of secondary structure observed in IDPs. The V_{FM} term can be constructed and tuned using the structural ensembles obtained from either experiments or long-timescale atomistic simulations, allowing AWSEM-IDP accurately to replicate the known structural features of any given IDP. Finally, the new V_{R_g} term provides fine control over the chain's global fluctuations, being important for reproducing the average chain radius as well as variance and tails of the R_g distribution that may be known either from experiments or atomistic simulations.

The quality of predictions from AWSEM-IDP will depend on the quality of the available experimental input data or the accuracy of atomistically generated ensembles. Experimental databases for IDPs, such as pE-DB [176], are rapidly evolving and will become more useful. We must bear in mind however that obtaining accurate descriptions of IDPs by atomistic MD simulations alone will remain a challenge, both due to the intrinsic sensitivity of IDP structure to force field error and the incomplete samplings of fully atomistic landscapes. In particular, our calculations of such sensitivities indicate an-order-of-magnitude amplification of errors compared to globular proteins, which has profound implications in the context of recent attempts to improve atomistic force fields to better model IDPs and unfolded protein chains.

In summary, AWSEM-IDP enables the exploration of large conformational spaces of IDPs while still maintaining sufficient chemical accuracy. The present work should provide the foundation for simulating large protein complexes that

include both ordered and disordered protein segments, such as nucleosomes.

Chapter 3: Binding Dynamics of Disordered Linker Histone H1 with a Nucleosomal Particle

This chapter is based on the unpublished work of the authors: *Hao Wu, Yamini Dalal, and Garegin A. Papoian; Binding Dynamics of Disordered Linker Histone H1 with a Nucleosomal Particle; In Preparation; (2020)*

3.1 Introduction

Eukaryotic chromatin consists of histone proteins and DNA [177]. Four types of histone proteins, namely H2A, H2B, H3, and H4, are assembled as an octamer (two H2A-H2B dimers and one H3-H4 tetramer) and wrapped around by ∼147 base pairs (bp) of DNA [9]. These elementary repeating units, called nucleosomes, are connected by short linker DNA and the linker histone protein H1, forming a "zigzagging ladder" architecture. These nucleosomal arrays are folded in several hierarchical levels to form chromatin in the nucleus [19].

The fundamental mechanisms of this intricate folding process, however, remain poorly understood. How does chromatin change its condensation levels rapidly and accurately, so that gene can be precisely transcribed while being well protected from DNA damage [178]? The linker histone H1 is believed to play an important role in

the fast dynamics of chromatin folding [179–182]. As a developmentally regulated protein, H1 has a family of variants that are specific to distinct species or tissue [183]. All these H1 variants consist of three parts: a short (20 - 40 amino acids (AA)) disordered amino-terminal domain (NTD), a highly-conserved globular domain (∼80 AA) with rigid structure, and a long (100-125 AA) highly disorganized carboxyl-terminal domain (CTD) with varying lysine and arginine contents, making various binding affinities possible amongst the variants. [54, 55, 184]. The high mobility of H1 when binding to nucleosome is directly related to many biological processes [185–189]. Meanwhile, H1 depletion levels are found to alter global nucleosome spacing and local chromatin compaction [63, 190]. H1 variants are modified and regulated by various post-translational modifications (PTMs), which in turn are thought to confer distinct chromatin structures [20].

The first crystal structure of the linker histone globular domain at high-resolution (2.6 Å) discovered its "winged-helix" folding motif over two decades ago [56]. However, the high-resolution structural insights of the linker histone-nucleosome complex, referred as to chromatosome, remained elusive until five years ago. Consequently, compared with the four core histone proteins, the structure and dynamics of linker histones are less studied and understood. Interestingly, various H1 subtypes are found to bind on different locations of nucleosome under different experimental conditions [62]. *Drosophila* H1 [68] and human H1.4 [12] bind to nucleosome near dyad – the center bp defining pseudo-two-fold symmetry axis of the nucleosome – in an asymmetric manner, while chicken linker histone H5 [61,191], *Xenopus Laevis* H1.0b and human H1.5 [66] bind right on the center of

dyad symmetrically. Zhou *et al.* proposed these "on-dyad" and "off-dyad" binding modes will lead to distinct chromatin condensation levels [61]. A series of atomistic computational studies revealed conformational selection mechanisms of these H1-nucleosome binding modes, as well as the effects of sequence and PTMs [192–194]. Further cryo-EM studies [12,69] and large-scale simulations with ∼1 nm spatial resolution [195, 196] (i.e. mesoscale simulations) of nucleosomal arrays explored how the linker histone's subtype and concentration determine inter-nucleosome relative positions and resulting distinct chromatin geometries. More experimental studies also discovered other important biological functions of H1 variants besides altering chromatin structure, such as regulating gene expression [63], generating epigenetic heterogeneity within tumor cells [197], and directly inhibiting transcriptions [198].

Nevertheless, most of the previous studies on linker histones have focused on the globular domain, while the structure and dynamics of the N- and C-terminus remain poorly understood, in part because of their intrinsically disordered nature. Recent studies reveal that H1 NTD and CTD are extremely disordered, even when bound with other proteins [199] or DNA [200]. Their disordered nature, in turn, imparts unique functions of liquid-like glue to promote chromatin folding [201]. In particular, the long and highly basic H1 C-terminal domain (CTD) was found to be more essential for the linker histone to bind onto nucleosome with high affinity compared to the NTD [182,202]. Similar to the core histone tails, which have been extensively studied by computer simulations [130–132], H1 CTD is also highly disordered and bound to DNA, but much longer than the ∼15-40 AA-long core histone tails. A recent cryo-EM experimental study validated the function of CTD in stabilizing the

H1-nucleosome complex by primarily binding to one of the linker DNA arms [66]. FRET experiments [70,203] and mesoscale simulations [195] demonstrated more H1 CTD conformations depending on the environment of nucleosome arrays and linker DNA length. Most recently, a series of computational studies [204, 205] used both atomistic molecular dynamics simulations and mesoscale Monte Carlo simulations to investigate the H1 disordered domains' conformation and dynamics in one and multiple chromatosome particles, as well as the regulatory roles of H1 phosphorylation and disorder-to-order transition on the nucleosome asymmetry and chromatin large-scale organization. However, it is challenging for experimental studies to investigate H1 disordered domains at the near-atomic spatial resolution, while computational results are limited by the accuracy of force fields and sampling rates. As a result, the detailed binding dynamics and mechanism of H1 disordered domains onto the chromatosome remain elusive.

In this work, we studied the structure and dynamics of an H1 variant, *Xenopus Laevis* H1.0b, in complex with a nucleosome particle, via molecular dynamics simulations (we use "H1" below to refer to the *Xenopus Laevis* H1.0b in all the subsequent parts of this study, unless otherwise specified). We used our coarse-grained protein model called AWSEM [76], in combination with our recently developed AWSEM-IDP [116] for dissecting disordered domains, and a DNA model 3SPN.2 [109], to simulate this large protein-DNA complex, including the disordered domains. To analyze potentially independent functions of H1 globular and disordered domains on chromatosome structure and dynamics, we simulated three combinations of H1-nucleosome systems: (1) nucleosome alone with linker DNA arms; (2) nucleosome

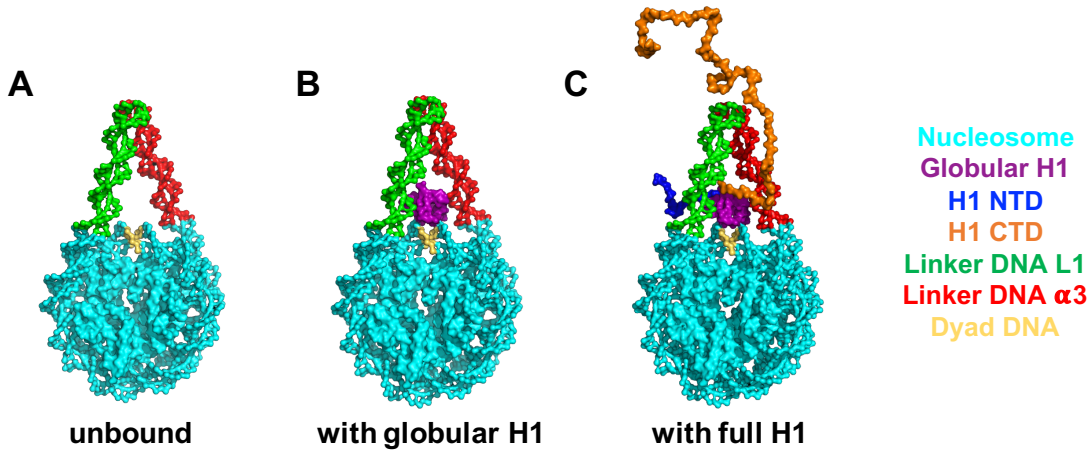


Figure 3.1: **The molecular systems simulated in this study.** From left to right: unbound nucleosome without H1 (A), globular H1-nucleosome (B), and full-length H1-nucleosome (C). All the models are based on a recent X-ray crystal structure (PDB: 5NL0) of the *Xenopus Laevis* H1.0b-nucleosome.

with linker DNA arms and the H1 globular domain (GH1); (3) nucleosome with linker DNA arms and the full-length H1 (see Figure 3.1 for graphic illustration). By comparative analyses of these three molecular systems, our study directly uncovers H1 disordered domains' restrictive functions on GH1 and linker DNA, shedding light on their functions in chromatosome compaction and chromatin condensation.

3.2 Methods

3.2.1 Hybrid Coarse-Grained Model for H1-Nucleosome

To investigate binding dynamics of the H1-nucleosome more efficiently and accurately, we model this large molecular complex (∼950 protein AA + 193 DNA

bp) with a coarse-grained protein force field AWSEM [76] and a DNA force field 3SPN.2 [109]. Inspired by the funneled energy landscape theory [81], AWSEM uses three beads (C_α, C_β, and O) to represent one residue and adopts both physical and bioinformatic potentials to account for amino acid interactions. AWSEM has been applied to study many different types of proteins [97–99,116], including histones [106, 107] successfully. Similarly, 3SPN.2 uses three sites - phosphate, deoxyribose sugar, and nitrogenous base - to represent a nucleotide and was carefully calibrated by structural and thermodynamic properties of single- and double-strand DNA. With comparable length scale and implicit solvent assumption, these two models have been combined to study several protein-DNA systems [101, 110, 206, 207], including the nucleosome [111].

Here, we elaborate on the detailed Hamiltonian of this protein-DNA force field. For protein-protein interactions, we used a strategy similar to Zhang *et al.* [111], which is mostly consistent with the original AWSEM but also introduces two modifications: an explicit Debye-Hückel (DH) electrostatic interactions between charged C_β atoms (+1 for arginine and lysine, -1 for aspartic acid and glutamic acid), and a weak Gō potential with fine-tuned parameters for the entire histone octamer to bias towards the crystal structure. Additionally, we applied the newly developed AWSEM-IDP force field [116] for the disordered H1 CTD and NTD. We reduced the helical potential weights specifically for these disordered domains to avoid artificial helices. We also performed extensive atomistic simulations for the structural segments of these disordered domains and used the resulting trajectories to bias local structure in AWSEM simulations, similar to Lin *et al.* [208] (see Appendix B

and Figures B.1, B.2, and B.3 for detailed atomistic simulation, sanity check, and related structure bias setup). All the DNA-DNA interactions are unchanged from the 3SPN.2 model.

For protein-DNA interactions, we first tested the simplified treatment as in Potoyan *et al.* [101] and Zhang *et al.* [111], which consists of non-residue-specific Lennard-Jones (LJ) potential and Debye-Hückel electrostatic interactions. We found these two forces are still too weak to keep a stable nucleosome structure within reasonable simulation time, where DNA tends to unwrap from the histone octamer (Figure B.4). In the nucleosome crystal structure, 14 arginine side chains from histone octamer insert deeply into the DNA minor grooves to further stabilize the nucleosome [9]. This important effect was not included in the previous simplified protein-DNA interactions. To mimic this special interaction, we added additional site-specific Lennard-Jones forces between these pairs of arginine C_β and DNA phosphate beads. With this new force, nucleosomal DNA stays wrapping around the histone core during the entire simulation (Figure B.4). Detailed formulas and parameters of this nucleosome-specific force are elaborated in Appendix B and Figure B.4.

As a summary, the formulae of AWSEM-DNA force field are listed below, where V stands for the potential energy of each term:

$$V_{\text{AWSEM-DNA}} = V_{\text{protein-protein}} + V_{\text{DNA-DNA}} + V_{\text{protein-DNA}} \tag{3.1}$$

$$V_{\text{protein-protein}} = V_{\text{AWSEM-original}} + V_{\text{DH}} + V_{\text{G\=o}} + V_{\text{AWSEM-IDP}}^{\text{H1NTD/CTD}} \tag{3.2}$$

$$V_{\text{DNA-DNA}} = V_{\text{3SPN.2}} \tag{3.3}$$

$$V_{\text{protein-DNA}} = V_{\text{DH}} + V_{\text{LJ}} + V_{\text{LJ}}^{\text{Arg-Phosphate}} \tag{3.4}$$

3.2.2 Simulation Details

All the coarse-grained simulations were performed with the open-source simulation package LAMMPS [92] (version 31Mar17), in which AWSEM-DNA was implemented. The initial conformation of the H1-nucleosome is a recent X-ray crystal structure [66] (*Xenopus Laevis* H1.0b, PDB: 5NL0). Three different molecular combinations were constructed to test the function of H1, including the nucleosome with linker DNA and (1) without H1 (unbound); (2) with the H1 globular domain; (3) with the full-length H1 (Figure 3.1). Note that the disordered H1 NTD and CTD are missing in the crystal structure. Hence we used MODELLER [67] (version 9.19) to generate their structure models as the initial conformation. Tails of core histones are not included in this study, because most of them are located far away from H1 or nucleosome dyad – the main region of interest in this study, and unlikely to interact with H1 or linker DNA. Moreover, including these disordered tails will significantly slow down the convergence of the simulations.

We set the simulation time step to 5 fs and used non-periodic shrink-wrapped boundary conditions. The thermostat is Langevin, with the damping parameter =

500 fs. We used a parameter set to mimic electrostatic interactions in 150 mM NaCl solution, which is close to a physiological cellular environment (see Appendix B for detailed parameters). With five different random initial velocity distributions, we heated up and annealed the system shortly (500 ps heating from 300 K to 330 K, 500 ps annealing from 330 K to 300 K) to generate five different initial configurations. Then for each initial configuration, we ran 10 independent simulations for 60 ns at 300 K and constant volume, each with different initial velocity distributions. In effect, for each molecular system, we obtained simulation trajectories summing up to 3 μs, a very long timescale considering the significantly faster sampling rate in our coarse-grained simulations than in atomistic ones.

3.2.3 Analysis

3.2.3.1 Linker DNA Geometry

We used several different metrics to quantify the conformation and dynamics of linker DNA arms. We computed α and β angles between the linker DNA arms and the vertical dyad axis to quantify the compaction of the chromatosome particle (see Figure 3.5A-B for graphical illustrations), following similar definitions from Bednar *et al.* [66] and Woods *et al.* [209]. Vectors representing linker DNA are defined as the center of mass (COM) of the beginning bp pointing to COM of the terminal bp. The vector representing the vertical dyad axis is defined as COM of bp at the center of nucleosome bottom pointing to the dyad. The "front" and "side" plane for angle α and β are defined as parallel and perpendicular to the nucleosome

core disk. Positive α indicates linker DNA arms bend "inward" while positive β means bending "outward". Similarly, we also defined and calculated angle θ and end-to-end distance between the linker DNA to quantify their relative geometry (see Figure 3.5C-D for graphical illustrations). We used VMD to conduct all the analyzes mentioned above.

3.2.3.2 Protein-DNA Regional Contact Map

We computed regional contact maps between H1 and DNA near the entry-exit site to describe their binding sites. We first divided H1 and DNA into small regions based on their secondary structure elements or location (see Table B.1 for detailed region definitions). A pair of protein-DNA beads in coarse-grained representation will be determined as "contact" if their inter-bead distance is smaller than a pre-determined threshold (= 8 Å, around 1.5 times of protein + DNA site radius for excluded volume effects). Based on this definition, we computed the contact probability by counting contact number within a certain H1/DNA region and normalizing it with the total number of protein-DNA bead pairs:

$$P_{\text{contact}}^{\text{protein}-\text{DNA}} = N_{\text{contact}}^{\text{protein}-\text{DNA}} / N_{\text{total}}^{\text{protein}-\text{DNA}} \tag{3.5}$$

3.3 Results

3.3.1 H1 Disordered Domains Confine GH1 Dynamics

To quantify how the presence of H1 disordered domains affects the dynamics of GH1, we tracked the trajectories of GH1's COM in our simulations (Figure 3.2). The trajectories in the absence of H1 disordered domains demonstrate that GH1 is very dynamic (Figure 3.2A). Starting from the initial position on the dyad, GH1 not only swings near the dyad but also drifts far away from the nucleosome. By contrast, the presence of H1 NTD/CTD significantly restricts GH1's range of activity (Figure 3.2B). As a part of the highly basic and disordered full-length H1, the globular domain still explores multiple conformations but all adjacent to the nucleosome. We also computed the radius of gyration (R_g) of GH1 COM trajectories to measure their average range of motion. The R_g without H1 NTD/CTD (23.5 Å) is larger than that with H1 NTD/CTD (20.9 Å), further validating H1 disordered domains' inhibitory role on GH1 dynamics.

The scatter plots of trajectories serve as an overview of GH1 conformations and dynamics, but the preferred GH1-nucleosome binding locations are not clearly shown. Therefore, we set up a three-dimensional reference coordinate system (x, y, z), and define the location of GH1 COM relative to the nucleosome core particle's COM as spherical coordinates (r, θ, ϕ) (see Figure 3.3A for detailed definition). Then we computed 2D histograms of all three spherical coordinates to identify the most probable GH1 binding modes. Here we found the effects brought by H1 disordered

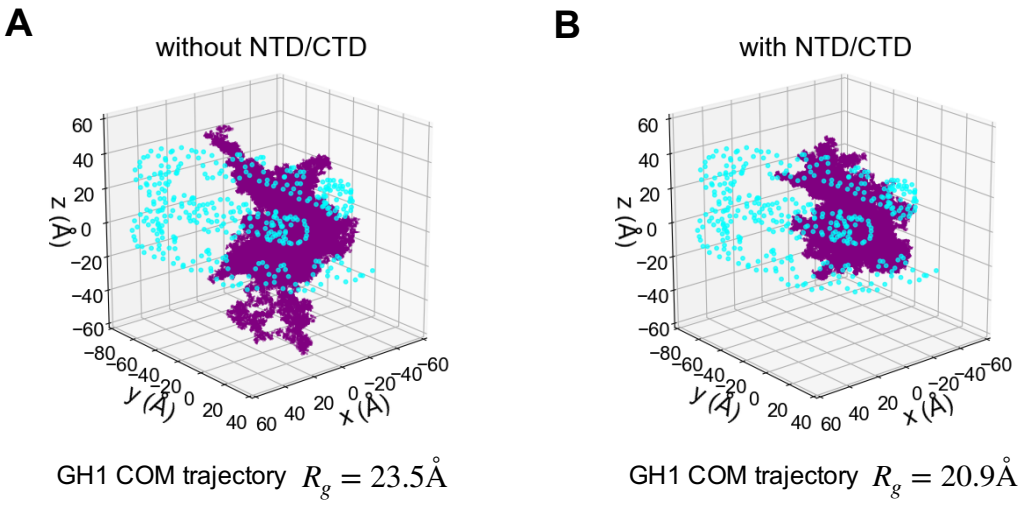


Figure 3.2: **GH1 dynamics are constrained by H1 NTD/CTD.** The trajectories of GH1 COM in the H1-nucleosome system without H1 NTD/CTD (A) and with H1 NTD/CTD (B) are plotted as purple dots. The DNA beads are represented as cyan circles. The core histone octamer and H1 NTD/CTD are not shown for clarity.

domains are best represented by the (ϕ, r) histogram (Figure 3.3B-C, the other two histograms of (θ, ϕ) and (θ, r) are shown in Figure B.5). For GH1 without the disordered domains, the (ϕ, r) distribution, especially ϕ values, is very dispersed. This means GH1 can almost freely rotate around the vertical y-axis across the dyad, whereas in the full-length H1-nucleosome simulations, we found the 2D distribution is more concentrated (Figure 3.3C). Almost all the GH1 conformations with H1 disordered domains have positive ϕ values, meaning GH1 is only allowed to be located on one side of the nucleosome disk. This comparison further indicates that the disordered domains restrict GH1's dynamics and prohibit many alternative binding conformations.

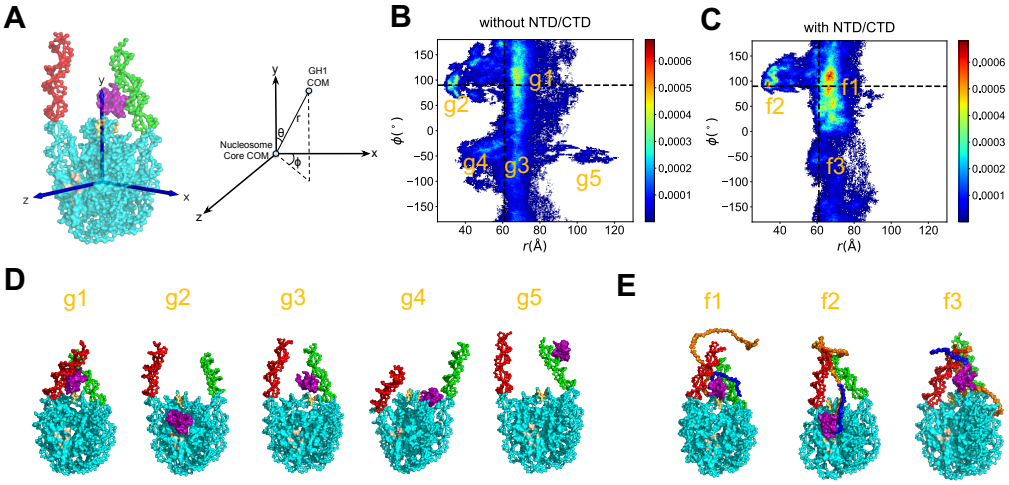


Figure 3.3: **GH1-nucleosome relative conformations.** (A) Definition of the 3D reference coordinate system and GH1 COM spherical coordinates (r, θ, ϕ). The 2D histogram of (ϕ, r) for GH1-nucleosome (B) and full-length H1-nucleosome (C) are shown as heat maps. The color bar from blue to red indicates a probability from low to high. The ϕ and r values at the beginning of simulations are labeled as black dashed lines. The representative snapshots of all the major basins labeled in (B-C) are shown in (D) and (E). The color code for different parts of molecules is the same as Figure 3.1, except for the gray region for the histone acidic patch.

3.3.2 H1 Disordered Domains Restrict GH1-Nucleosome Interaction Sites

Besides describing GH1 dynamics, the 2D histograms in Figure 3.3 also identify the preferred GH1-nucleosome binding modes. Here we found five major basins (g1, g2, g3, g4, g5) for GH1-nucleosome and three (f1, f2, f3) for full-length H1-nucleosome. The representative snapshots of each basin are shown in Figure 3.3D-E (see Table B.2 and Appendix B for definitions and population percentages of each basin, and the detailed procedure to select representative snapshots). In two major conformations g1 and g3 of H1-nucleosome without disordered domains, GH1 is still located near the dyad DNA minor groove (labeled as yellow). In basins g2 and g4, however, GH1 tends to slip off the dyad region and move towards the histone octamer (see Supplemental Movie 1 to demonstrate this behavior). With a small probability ($\sim$0.9%), GH1 even moves to the tip of one linker DNA arm to escape away from the dyad in basin g5. Meanwhile, four out of these five major basins (25.4%) have very different ϕ and r values from the on-dyad initial binding mode (black dashed lines). For full-length H1-nucleosome, two out of three basins (f1 and f3) are located more proximal to the initial binding mode. This contrast clearly shows that the presence of H1 NTD/CTD shifts the preferable GH1 binding positions closer to the dyad. Similar to g2, there is also an unignorable basin f2 representing a far-away-from-dyad binding conformation even with NTD/CTD. One possible reason might be the competitive electrostatic attraction from the acidic patch region on the histone core (shown as gray in Figure 3.3D-E, also see Figure B.6 for snapshots with

a zoomed-in view). Overall, this result demonstrates that H1 disordered domains tend to stabilize the GH1's position near the DNA entry-exit site.

Previous experiments have found that various types of linker histones recognize nucleosomal DNA via different parts of their "winged-helix" folding motif [61,66]. To further probe the DNA-binding preferences of these secondary structure elements, we divided GH1 and DNA into several regions and computed the contact probability of two arbitrary beads belonging to a certain pair of regions (Figure 3.4). For the GH1-nucleosome without NTD/CTD, the main regions of H1 in contact with dyad DNA are L1/β1 and β2 (Figure 3.4B, 2nd row). By contrast, when NTD/CTD are present, α2 and α3 helical regions form stronger contacts to the dyad DNA (Figure 3.4C, 2nd row). These two helices also have more residues proximal to the dyad DNA in the X-ray crystal structure [66] (labeled with purple stars), meaning our results with NTD/CTD have a similar near-dyad binding pattern to the experimental results. On the other hand, GH1's binding probabilities with the α3- and L1- linker DNA (Figure 3.4B-C, 1st, 3rd row) are mostly low with or without the disordered domains. But when NTD/CTD are present (Figure 3.4C, 1st row), the α3 helix forms significantly more contacts with the α3 DNA, consistent with the crystal structure as well. Our contact map analyses provide a detailed description of the GH1-DNA interface in the presence of H1 NTD/CTD: GH1 recognizes nucleosome mainly through dyad DNA by α2/α3 helix. But instead of being located in the exact dyad center, GH1 is tilted towards α3 DNA. The absence of H1 disordered domains, on the contrary, changes this stable interaction network and makes GH1 more flexible.

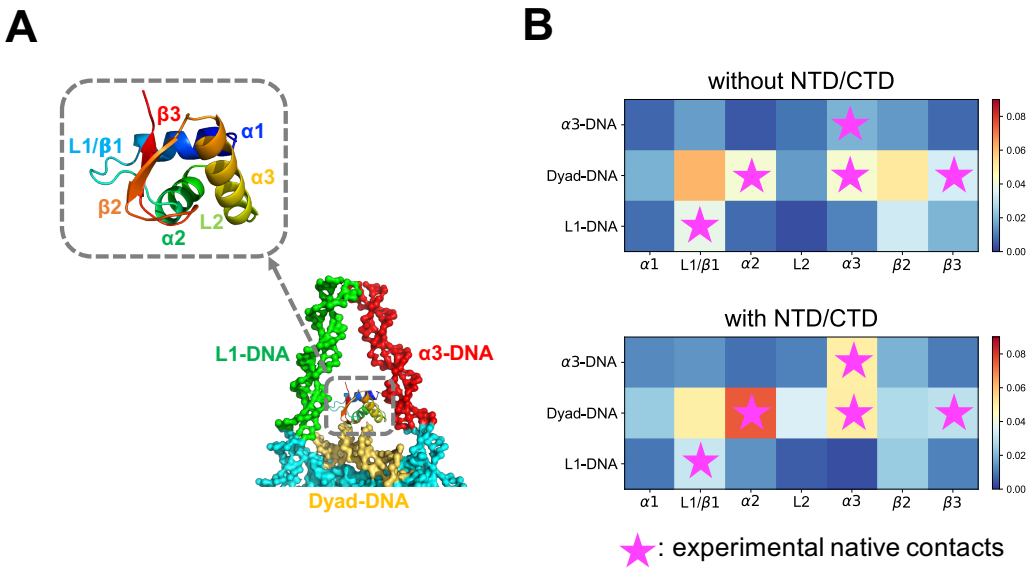


Figure 3.4: **H1 disordered domains regulate and stabilize GH1-DNA binding interface.** GH1 and nearby DNA structures are represented in (A), where colors represent different protein or DNA regions. (B) GH1-DNA contact maps without and with H1 NTD/CTD. Horizontal and vertical axes represent GH1 and DNA regions respectively. Color code from blue to red indicates contact probability of a protein-DNA beads pair in this region from low to high. Purple stars represent the experimental "native contact regions", where more than two residues in this GH1 region are close to a DNA region.

3.3.3 H1 Globular and Disordered Domains Converge Linker DNA

The geometry of linker DNA arms is another important feature to evaluate chromatosome structural compaction. Previous cryo-EM experiments [66] show that H1 induces a more compact and rigid nucleosome conformation by keeping the linker DNA arms convergent or "closed". Here we computed angles α and β to quantify the linker DNA's geometry parallel and perpendicular to the nucleosomal disk. We found the α angle distribution of full-length H1-nucleosome ($29.3° \pm 14.8°$) shifts toward higher values compared to the unbound results ($15.6° \pm 17.7°$) (Figure 3.5A), while β angle ($0.9° \pm 16.7°$) moves to lower values than the unbound ($9.6° \pm 16.3°$) (Figure 3.5B). Both trends show that full-length H1 keeps the linker DNA collapsed in our simulations, consistent with previous experimental measurements [66]. In particular, we found that both α ($20.0° \pm 16.7°$) and β ($5.2° \pm 16.3°$) distributions in GH1-nucleosome simulations are in between unbound and full-length H1 results. This demonstrates that while globular H1 by itself can bring the linker DNA together, the disordered domains will significantly reinforce this converging effect.

Analyses of individual linker DNA arms allow us to measure their relative conformations and distances to describe the compaction of chromatosome with further accuracy. Here we define an angle $\theta = \alpha 1 + \alpha 2$ between the two linker DNA arms and plot its distribution (Figure 3.5C). Its histogram demonstrates that the θ of GH1-nucleosome ($40.1° \pm 23.5°$) only increases by a small amount compared with the unbound nucleosome result ($30.2° \pm 27.0°$). While the binding of full-length

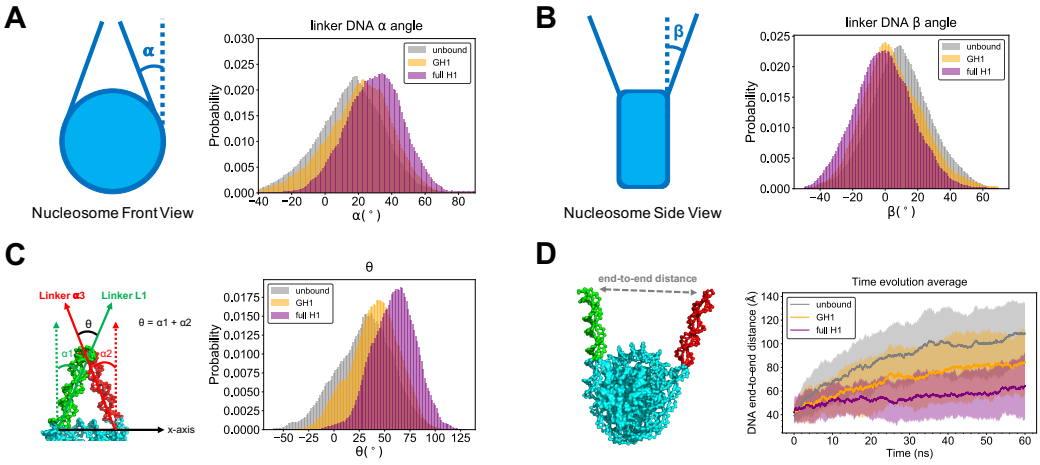


Figure 3.5: **Linker DNA arms are more converged with full-length H1 than globular H1 or unbound.** The definition and histogram of α, β, and θ angles are shown in (A, B, C). (D) The time evolution of DNA end-to-end distance, where solid lines with shaded regions represent average values with standard deviations. The same set of legends is used in all the subset figures above (gray: unbound nucleosome; yellow: GH1-nucleosome; purple: full-length H1-nucleosome).

H1 increases θ significantly ($58.6° \pm 21.4°$). This comparison reveals the disordered domains are more crucial than the GH1 to compact chromatosome structure.

We also calculated the end-to-end distance between the tips of linker DNA, and plot its average and standard deviation over all runs along the simulation time (Figure 3.5D). We found in absence of H1, the linker DNA arms start to separate from each other after 20 ns. The end-to-end distance reaches $\sim$ 100 Å at the end of simulations. Globular H1 slows down the linker DNA's separation but cannot fully prevent it. The end-to-end distance still rises to $\sim$ 80 Å by the end of simulations. On the other hand, the binding of full-length H1 remarkably inhibits this separation, as found in previous experiments [66]. The end-to-end distance does not show any apparent increasing trend beyond standard deviation. These results demonstrate that globular H1 alone is not sufficient to fully compact the chromatosome as found in previous experiments, while disordered domains act as a key player for this function.

As a more comprehensive comparison between our simulation results and cryo-EM experiments, we also ran simulations of the H1.5ΔC50 system, another chromatosome with H1.5 linker histone with the removal of the last 50 residues on CTD, as defined in Bednar *et al.* [66] (see Figure B.7 for its sequence and structure, the simulation details are elaborated in Appendix B). The statistics of all the DNA metrics used above (α, β, θ angles, and end-to-end distances) of H1.5ΔC50 are very similar to the full-length H1.0 results (Figure B.8). This similarity of DNA dynamics reveals that H1.5ΔC50, even though belonging to a different H1 subtype and having a shorter CTD, also leads to relatively compact linker DNA geometry and

dynamics. This result further consolidates the function of H1 disordered domains to restrict linker DNA.

3.3.4 H1 NTD and CTD are Tethered to Both Linker DNA Arms

The above analyses uncover restrictive functions of H1 disordered domains on the conformation and dynamics of both globular H1 and linker DNA. To further investigate the mechanism behind these functions, we computed a regional contact map for H1 NTD/CTD and DNA (Figure 3.6), using the same definitions as in Figure 3.4 and Table B.1. For H1 CTD, the GH1-proximal C1, C2, and C3 regions have higher contact probabilities (Figure 3.6B, 3rd-5th column), while the GH1-distal regions form very minimal contacts with DNA (Figure 3.6B, 6th-10th column). A similar trend is found for NTD, where the GH1-proximal N2 region is much more tightly bound than the distal N1 region (Figure 3.6B, 1st-2nd column). Both CTD and NTD have closely bound regions with the dyad DNA on N2, C1, and C2 (Figure 3.6B, 2nd-4th column, 2nd row). On the other hand, CTD prefers binding with L1-DNA via its C1 and C3 regions (Figure 3.6B, 3rd and 5th column, 1st row), while NTD is particularly bound with α3-DNA via the N2 region (Figure 3.6B, 2nd column, 3rd row). This contact map elucidates how the full-length H1 is steadily bound with DNA near the entry-exit site: the two disordered domains, especially the GH1-proximal parts, act as "hands" of GH1 to grasp the dyad and linker DNA. This entanglement between H1 disordered domains and DNA precludes GH1 from escaping far away from the dyad region and confines linker DNA dynamics.

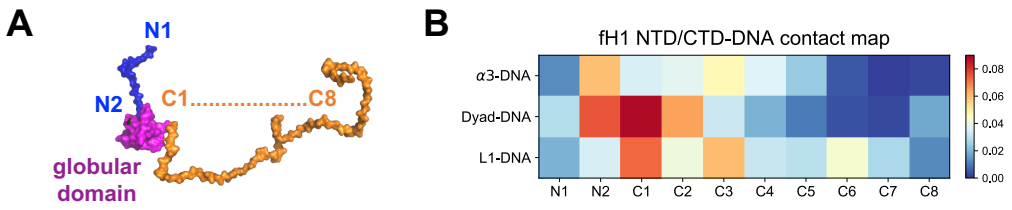


Figure 3.6: **H1 NTD and CTD are tightly bound with DNA mainly via the residues proximal to the globular domain.** (A) H1 NTD (blue) is divided into two regions (N1-N2, from GH1-distal to GH1-proximal). H1 CTD (orange) is divided into eight regions (C1-C8, from GH1-proximal to GH1-distal). (B) shows the contact map of DNA with NTD and CTD. The horizontal axis represents NTD (N1-N2) and CTD (C1-C8) regions. The vertical axis and color bar are the same as in Figure 3.4B.

We also observed that the contact probability of the GH1-distal region of the CTD (especially C4-C8) is much weaker, meaning it is only transiently bound to DNA. Interestingly, the entire NTD and CTD remain highly disordered in the simulation (see Figure B.9). This result agrees with previous NMR experimental results [200] and further verifies that even the tightly bound parts of NTD/CTD are too disordered, which may explain their absence in almost all the X-ray crystal structures of chromatosome.

3.4 Discussion

In this paper, we investigate the binding dynamics of the H1-nucleosome complex using extensive computer simulations using a state-of-art protein-DNA model. Our focus is on the regulatory function of H1 disordered domains on chromatosome

structure. By quantitative comparisons among different molecular combinations, we found H1 disordered domains, especially GH1-proximal regions, compact chromatosome structure by constraining the conformation and dynamics of linker DNA arms and GH1. By contrast, without the disordered domains, the binding affinity between the H1 globular domain and DNA is reduced notably. In this case, linker DNA arms become more separated, and GH1 may even escape away from the nucleosome, resulting in a more "open" chromatosome conformation.

As the first computational study of H1 disordered domains at this high resolution and extensive timescale, our results extend previous atomistic [193, 209] and mesoscale [195] simulations. Using our AWSEM-DNA force field, we can examine atomic details such as protein-DNA binding sites, while discovering the dynamics of the entire chromatosome particle for a long timescale. Moreover, the application of AWSEM-IDP [116] allows us to simulate H1 disordered domains with sufficient accuracy and efficiency. Many structural and dynamic insights of this highly disordered molecule are thereby discovered to explain their biological functions in chromatosome compaction. Meanwhile, our results are in general comparable to the most recent cryo-EM experiments [66], validating the sanity of this computational model for current and future protein-DNA simulations.

Previous studies suggest H1 can bind with nucleosome on- or off-dyad to regulate distinct chromatin higher-order structure [61]. As a recent review [62] points out, H1-nucleosome binding modes depend on H1 species and experimental conditions, and the resulting chromatosome structures should be viewed as an ensemble. Interestingly, our simulations not only replicate H1-nucleosome on/off-dyad binding

modes found in previous experiments and all-atom simulations, but also discover new H1 binding modes far from the dyad. One commonly observed new binding mode in our simulation is that GH1 totally escapes from dyad and reaches close to the acidic patch on histone core. In this case, the absence of GH1 near the dyad will weaken the interaction between nucleosomes directly connected by linker DNA. Meanwhile, the nucleosome disk packing might become more condensed because of GH1's electrostatic mediation. This finding indicates that within the same species, H1 can also bind to different locations and its modes should be treated as an ensemble as well. Here, the disordered domains play a balancing role to prevent the chromatosome from being too rigid or too flexible. The resulting suitable structural plasticity of chromatosome allows for H1's rapid transition among different modes, including binding/unbinding to the nucleosome, to alter chromatin compaction level within a desirable timescale.

In particular, our results uncover the significant roles of H1 disordered NTD and CTD in alternating GH1 binding modes and restricting linker DNA dynamics. These functions are highly relevant to the "zigzagging ladder" model of chromatin folding, where H1 acts as the "rung" between nucleosome particles. Based on our findings, we propose that H1 NTD/CTD will affect the location and strength of the connection between H1 and neighboring nucleosomes, thus regulating nucleosome array organization and chromatin higher-order structure. Compared with the globular domain, H1 disordered domains are more prone to various PTMs [20]. These PTMs would enable many important biological functions, including not only chromatin folding [210] but also apoptosis [211] and DNA transcription [212]. In

particular, the phosphorylations on H1 CTD have been proved to change the secondary structure of CTD and regulate chromatin condensation level [213, 214]. We expect that the phosphorylation will drive the dynamics of H1 to a moderate level between the two systems in this study (globular H1 and full-length H1) because its binding affinity to the linker DNA is weaker than the former but stronger than the latter. It would be thereby very meaningful to directly study the structural effects from PTMs on H1 disordered domains, by computer simulations with feasible accuracy and efficiency and compare them with high-speed atomic force microscopy data analyzing dynamics in real-time.

We also observed that GH1 can even totally escape from the dyad region and move towards the histone core acidic patch (Figure 3.3), with less probable but unignorable occurrence. The reason this behavior is both novel and potentially biologically meaningful is that it might explain the mechanistic basis for decades-old observations that *in vitro*, H1 can reposition nucleosomes without the use of ATP or remodelers [215, 216]. The spacing function of H1s remains mysterious, but our observation of H1 breaking symmetry at the center of the nucleosome might subtly lower the energetic barrier at the pseudo-dyad, allowing nucleosomal looping in advance of sliding and repositioning along the DNA fiber. Our contact analyses in Figure 3.4 reveal that $\alpha 2$ and $\alpha 3$ helices are the major GH1-dyad binding regions, no matter whether H1 disordered domains are present or not. Thus, it will be promising for further experiments to mutate some key residues in these regions, such as S49, K53, K69, and K74, to test how they affect this H1 escaping-dyad motion.

Shortly after the observation of nucleosome ladders by Hewish and Burgoyne

[217], coupled with classic EM experiments visualizing regularly spaced beads on a string by Woodcock [6] and the Olinses [5], early workers in the chromatin field [218] discovered the still inexplicable phenomenon of species-specific nucleosome repeat lengths. There NRLs were thought to derive from species-specific H1 variants. Our work here provides a testable theoretical framework to explore how species-specific H1 residue changes over evolutionary time [219], in the NTD, the CTD and the GD might subtly alter the motions of H1 we observed in this study, thereby potentially contributing to the global spacing of nucleosomes. We note with excitement that the advent of high-speed AFM [220–223] provides precisely the kinetic handle needed to complement H1 fast dynamics observed in FRAP studies [179, 224] and H1 off/on dyad classic steady-state biochemistry experiments [225, 226] glean insights into linker histone biology as it relates to chromatin spacing and folding.

One limitation of our study is the lack of individual H1 NTD/CTD's functions. Previous FRAP experiments found deletion of the longer and less conserved CTD reduces H1-nucleosome binding affinity to a much greater extent than the removal of NTD [202]. It is thus promising to computationally model an H1-nucleosome system with only NTD or CTD and investigate their independent effect in regulating chromatosome structure and dynamics. Meanwhile, for different variants, H1 NTD and CTD have much less conserved sequences and distinct lengths. Our study already shows H1.0 and H1.5ΔC50 have similar structural properties, but it would be more interesting to conduct a systematic investigation on how the sequence and length of H1 disordered domains regulate chromatosome structure and dynamics.

3.5 Conclusions

In conclusion, our study supports the indispensable role of disordered linker histone to compact a single chromatosome particle. This work sheds light on the direction of further researches, such as H1's regulatory function on nucleosomal array structure, to better understand the mechanism of chromatin folding.

Chapter 4: Folding-Upon-Binding Mechanism Widely Exists in Histone Fold Structures

This chapter is based on the unpublished work of the authors: *Haiqing Zhao, Hao Wu, Dulith Abeykoon, Alex Guseman, Christina M. Camara, Yamini Dalal, David Fushman and Garegin A. Papoian; Folding-Upon-Binding Mechanism Widely Exists in Histone Fold Structures; In Preparation; (2020)*

Author contributions: H.Z., Y.D., D.F., and G.A.P. designed research. H.Z. performed computer simulations and analyses for histone dimer dynamics, thermodynamics, and histone polymer scaling law (Sections 4.3.2, 4.3.3, 4.3.5, Figures 4.2, 4.3, 4.4, and 4.6). H.W. performed computer simulations and analyses to predict structures for histone monomer and histone fold proteins monomer and dimer (Sections 4.3.1, 4.3.7, Figures 4.1 and 4.7). D.A., A.G., and C.M.C. performed NMR and CD experiments and related analyses (Section 4.3.4 and Figure 4.5). All the authors wrote the manuscript.

4.1 Introduction

In eukaryotic cells, histone proteins package the genomic DNA into chromatin in the form of the basic subunit called the nucleosome. The histone proteins in

the nucleosomal core are assembled by four pairs of heterodimers among which two H3/H4 dimers form a tetramer while H2A/H2B participates as two dimers [9]. Hence, the histone dimer, or more precisely the histone heterodimer, is found to be the smallest protein unit in eukaryotic chromatin. Besides these canonical histones that compose the majority of nucleosomes, variant histones also evolved for diverse functions in the nucleus [227]. Interestingly, despite these functional and sequence diversities, all histones possess the same structural motif, known as a histone fold [228], where two histone monomers, each consisting of a helix-loop-helix frame, fold into a "handshake" motif to form a dimer in an intertwined, head-to-tail manner [28].

Extensive studies in biochemistry and cell biology have focused on the structure and function of canonical and variant histone nucleosomes, interrogating the relationships among sequence, structure, and function of histones in different types of nucleosomal environments [229–232]. Previously, the thermal stability of H2A/H2B and $(H3/H4)_2$ tetramer were studied by a series of denaturation experiments [25, 233–235]. Karantza *et al.* reported that during the unfolding processes of either H2A/H2B or $(H3/H4)_2$, individually folded monomers are not detectable indicating the direct transition from a folded histone dimer to unfolded monomers. However, whether or not this unfolding or folding principle widely applies to other histone-fold structures [42, 236], especially the histone variants, remains a critical question. A comprehensive understanding of the folding dynamics of these proteins may shed light on understanding their functions and their higher-level structural organization such as tetramer or octamer formation.

In this work, we used molecular dynamics (MD) simulations together with

nuclear magnetic resonance (NMR) spectroscopy and circular dichroism (CD) experiments to further explore the enigmatic mechanism of histone folding from a biophysical perspective. With annealing simulations, we predicted the structure of both canonical and variant histone monomers and reported that all of them tend to fold into collapsed states that are far from the native conformations. However, in the presence of their appropriate binding partner, two histones fold into a native-like dimer, revealing a folding-upon-binding dynamics. We tested these computational predictions using NMR measurements, which demonstrate close-zero signal for single types of histone monomers, while new signals appear and increase in intensity when hetero histones are paired as 1:1. We also performed CD experiments and the results demonstrate a significant increase in the helical content of these proteins. Finally, we extended these observations by simulating two histone-like transcription factors and found similar folding dynamics. Overall, these comprehensive data suggest that the folding-upon-binding principle may apply generally for histone-like structures.

4.2 Methods

4.2.1 MD Simulation

MD simulations here were carried out using the LAMMPS package, with AWSEM [76] model, under a non-periodic shrink-wrapped boundary condition and Nose-Hoover thermostat control. AWSEM is a coarse-grained protein force field inspired by the free energy landscape theory [81]. The Hamiltonian includes both

physical terms and bioinformatics-inspired term (Eq. 4.1). Details of every term are covered in Davtyan *et al.* [76] and Papoian *et al.* [91]. AWSEM has been successfully applied to predict protein monomer structures [76], protein-protein binding interfaces [98], and, in particular, histone proteins [106, 107, 111, 116].

$$V_{AWSEM} = V_{backbone} + V_{contact} + V_{burial} + V_{Hbond} + V_{FM} \tag{4.1}$$

Annealing simulations were conducted to predict the monomeric and dimer structures. By decreasing the simulation temperature from 600 K to 200 K, the annealing procedure would be able to help proteins search for the global minima of the energy landscape. The final minimum energy conformation is typically the prediction outcome after this optimization. Replica exchange [170] coupled with umbrella sampling [237] simulations around Q value (Eq. 4.2) were used for the dimeric proteins to estimate their binding free energy. All technical details and parameters are provided in Appendix C.

$$Q = \frac{1}{N} \sum_{i<j-2} exp[-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}}] \tag{4.2}$$

4.2.2 NMR and CD Experiments

Unlabeled and ^{15}N labeled histones H2A and H2B were expressed in *E. coli* and purified from inclusion bodies using cation exchange chromatography. Their correct mass was confirmed by mass spectrometry. All NMR experiments were performed at 23°C on Bruker Avance-III NMR spectrometer equipped with TCI cryoprobe.

Proteins were dissolved at 100-200 uM in 20 mM sodium phosphate buffer (pH 6.8) containing 7% D_2O and 0.02% NaN_3. NMR data were processed using TopSpin (Bruker Inc.)

Circular dichroism (CD) spectra of histone proteins (20 mM sodium phosphate buffer pH 6.8, 0.40 mg/mL concentration) were acquired on a Jasco J810 Spectro-Polarimeter using a Peltier-based temperature-controlled chamber, at 25°C and a scanning speed of 50 nm/min. A quartz cell (1.0 mm path length) was used. All measurements were performed in triplicate. To determine the secondary structure content, the CD data were analyzed using the DichroWeb server [238]. Two methods were used in parallel: (i) CDSSTR, a singular value decomposition (SVD)-based approach employing two datasets (7 and SMP180) from the DichroWeb server, and (ii) K_2D, a neural network-based algorithm trained using reference CD data [239]. All the NMR and CD experiments were performed in Prof. David Fushman's laboratory.

4.3 Results and Discussion

4.3.1 Monomers Fail to Fold on Their Own

A previous denaturation experiment reported individual folded histone monomers were not detectable during the unfolding process of H2A/H2B dimer or $(H3/H4)_2$ tetramer. Therefore, we first wanted to examine if histone monomers can fold on their own. To address this question, we performed AWSEM annealing simulations from the random coil state, for nine types of histone-fold proteins (HFP) including

canonical histone monomers H2A, H2B, H3, H4, variant histone type CENP-A, and four types of transcription factor protein $dTAF_{II}42$, $dTAF_{II}62$, NY-FB, and NY-FC. Each type was studied by ten separate simulations from 600 K to 200 K. We measured the structural similarity between simulated and native structures by Q value. Unlike the direct atomic position comparison metrics, such as root-mean-square deviation (RMSD), the Q quantifies residue-residue contacts so it is widely used in protein folding studies. As defined in Eq. 4.2, the Q ranges from 0 to 1. A higher Q value indicates similar contacts to native ones are formed in the simulated conformations.

We plotted the mean value and standard deviation of Q as a function of the decreasing temperature in all the annealing simulations (Figure 4.1A). None of the histone monomers' Q values increase significantly as the temperature is decreased. At simulation ends, the final Q values are only 0.28 - 0.36, indicating that corresponding structures are at best partially folded, but most remarkably different from the native conformations. In contrast, at almost identical setup, more than ten globular protein structures could be folded using AWSEM simulations at much higher accuracy ($Q \sim$ 0.60 - 0.65) in Davtyan *et al.* [76]. This comparison reveals that folding patterns of histone monomers are distinct from regular monomeric globular proteins. A further look into the representative snapshots of the predicted structures for each histone monomer (Figure 4.1B) shows that they do not form the stereotypical "hand-shake" motif as in dimer or tetramer states in the crystal structures. Although the α-helices are formed in the sequential regions roughly similar to the native structure, the conformation and orientation of the $\alpha 1$ and $\alpha 3$ helices signif-

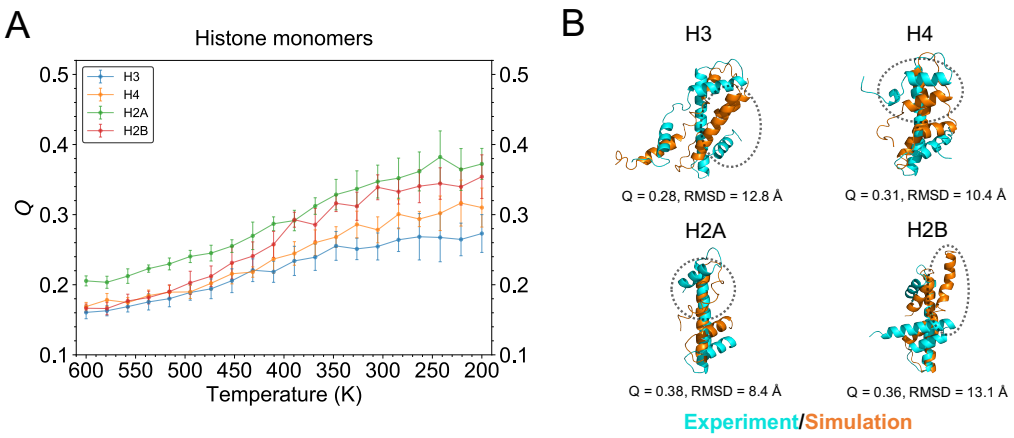


Figure 4.1: **Histone monomers cannot fold by themselves.** (A) shows the Q values as a function of the decreasing temperature in all the histone monomer annealing simulations. The average and standard deviation of all the Q values within each temperature window are represented as circles and error bars. (B) displays representative snapshots of each histone near the end of simulation with the final mean Q value (orange), superimposed with their corresponding X-ray crystal structures (cyan). Conformational differences in short α-helices are highlighted by dashed line circles.

icantly deviate from the histone fold (circled by dashed lines), which, in turn, may block the binding interface with its dimerization partner and disrupt the formation of histone dimer/tetramer assembly. All these computational results suggest that histone monomers are not globular proteins and cannot adopt the histone fold motif without help from their binding partner. These findings are consistent with previous experiments [25, 233].

4.3.2 Dynamics of the Histone Dimer Folding

After finding that histone monomers cannot fold by themselves, we next examine if monomers can fold in the presence of their binding partner. This part of the

work was performed by Haiqing Zhao. Following a similar strategy, we performed annealing simulations for the mixed monomer pairs for H3 and H4, H2A and H2B, CENP-A and H4, dTAF$_{II}$42 and dTAF$_{II}$42, NY-FB and NY-FC. We would like to note that the bioinformatic term in AWSEM only uses local fragment memories that are less than nine residues long. No structural biasing potential towards the dimeric structure is included.

To evaluate our simulations' prediction accuracy, we first analyzed the contact map of the predicted dimer conformation compared to those of the native structure. Overall, 98% of H2A/H2B's (Figure 4.2A) and 96% of H3/H4's (Figure 4.2B) native contacts were correctly predicted. The contacts within each monomer as well as those between two monomeric partners have been reproduced with high accuracy, indicating that our simulations can successfully predict intrachain and interchain structural information when histones are near their dimerization partners.

To further investigate the folding and binding mechanisms of the histone dimers, we also calculated the Q values of the entire dimer (Q_{dimer}) and the component monomers ($Q_{monomer}$) relative to their corresponding crystal structure and plotted them as a function of the annealing temperature (Figure 4.3). These plots show that around 410 - 380 K, there is a clear transition, wherein the Q_{dimer} value rapidly rises from 0.25 to 0.45 for H2A/H2B (Figure 4.3A), and from 0.25 to 0.4 for H3/H4 (Figure 4.3B). The transition of Q_{dimer} occurs roughly at the same time and temperature as that of $Q_{monomer}$, indicating that the two composing monomers in both H2A/H2B and H3/H4 dimer fold and bind simultaneously. Contrary to expectation, the two composing monomers' contributions to their binding are not

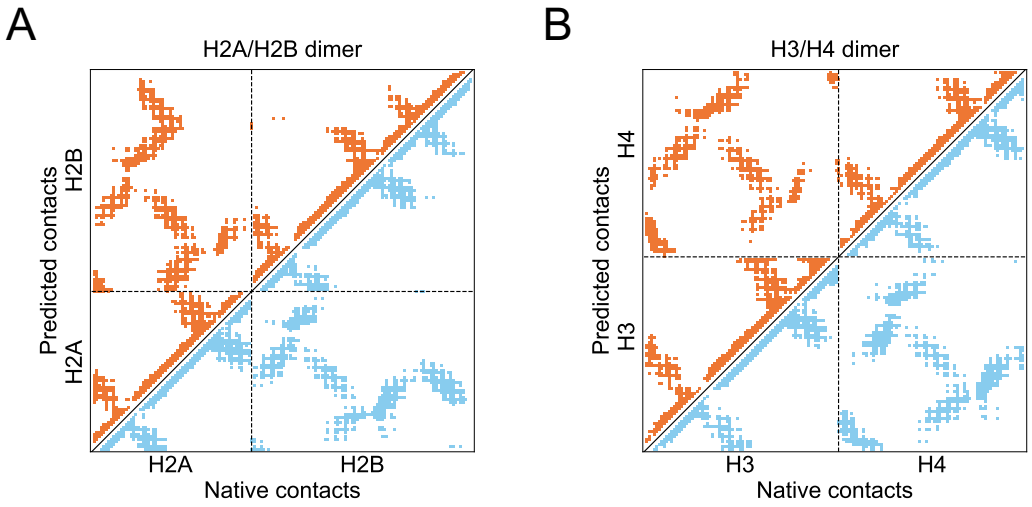


Figure 4.2: **Contact maps are precisely predicted in histone dimer simulations.** Predicted contacts (orange) versus native ones (blue) are plotted for H2A/H2B (A) and H3/H4 (B). Each colored dot in the figure represents a pair of residues in contact. The sequential regions of each monomer in the dimer are indicated by the horizontal and vertical axes text (separated by the dashed lines). Adapted from Haiqing Zhao's original figure with permission.

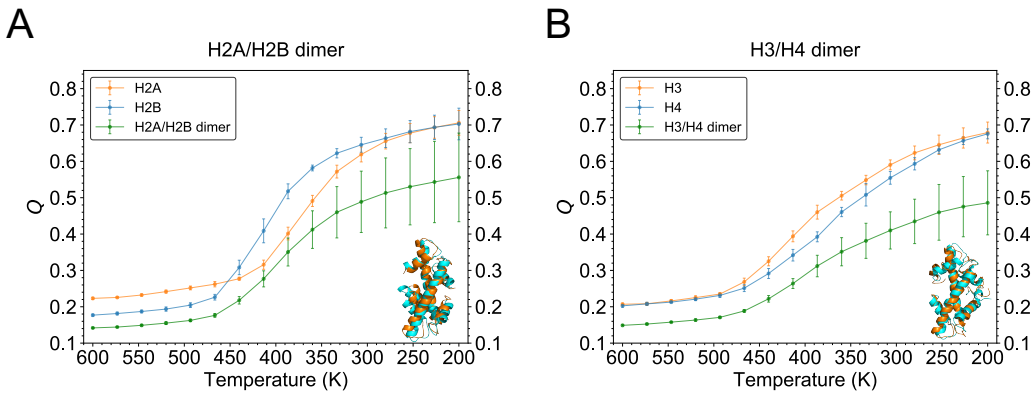


Figure 4.3: **Histone monomers help each other to fold in dimer annealing simulations.** (A) Q value analysis for H2A/H2B shows that the monomer H2A (orange), H2B (blue), and the histone dimer (green) fold simultaneously as the annealing temperature is cooled. (B) H3/H4 annealing simulations also displays a simultaneous folding and binding process between H3 (orange) and H4 (blue) monomer, resulting in the dimer H3/H4 (green). The final folded dimeric conformations of H2A/H2B and H3/H4 (orange) are aligned to the corresponding crystal structures (cyan). Adapted from Haiqing Zhao's original figure with permission.

the same. For instance, as shown in Figure 4.3A, H2B is on average slightly better folded than H2A during the annealing simulations. On the contrary, in the above monomer simulations, Q_{H2A} is generally higher than Q_{H2B}. This comparison suggests the folding behavior of H2A and H2B depends on the presence of their dimerization partner. On the other hand, in the H3/H4 dimer simulations, H4 maintains relatively more native-like conformation than H3 (Figure 4.3B). This observation is consistent with our previous finding that H4 preferentially maintains native-like stability in the presence of various binding partners [106].

4.3.3 Thermodynamics of Histone Dimer Folding

To further characterize the thermodynamical features of histone folding, we carried out coupled replica-exchange and umbrella sampling simulations for the H2A/H2B dimer. This part of the work was performed by Haiqing Zhao. The calculated free energies were projected onto $Q_{monomer}$ and Q_{dimer}. In Figure 4.4, the free energies of H2A/H2B are plotted as a function of Q_{dimer} (1D green curve in both Figure 4.4A and 4.4B), and as a function of Q_{H2A} and Q_{dimer} (2D contour map in Figure 4.4A), and Q_{H2B} and Q_{dimer} (2D contour map in Figure 4.4B). From the 1D free energy curve, it is clear that the simulated H2A/H2B dimer has two states: an unfolded state at $Q_{dimer} \sim 0.23$ and a folded state at $Q_{dimer} \sim 0.5$. The energy barrier between these two states is about 4 kcal/mol, located at $Q_{dimer} \sim 0.3$. This result is consistent with our simulated annealing simulations, where the folding transition of H2A/H2B also occurs at $Q_{dimer} \sim 0.3$ (Figure 4.3A).

Furthermore, on the 2D free energy surfaces, we observe two basins (reddish regions), with a saddle region near $Q_{monomer} \sim 0.5$ (orange region), representing the most favorable transition zone between the two minima. However, if the individual monomers were well folded as $Q_{monomer} \sim 0.6 - 0.7$, they would have to overcome a large energy barrier to form the native dimer. This result shows that to form the native histone dimer with "histone fold", the two monomers need to assemble cooperatively.

Meanwhile, the two histone monomers also make different thermodynamical contributions to the dimer formation. As seen in the 2D free energy surface, both

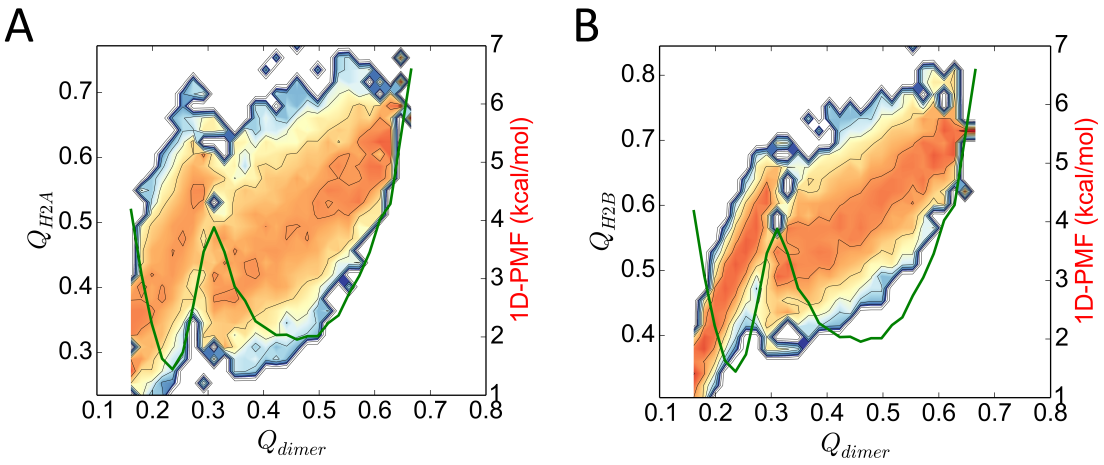


Figure 4.4: **Free energy profiles of H2A/H2B histone dimer folding and binding.** The 1D potential of mean force (PMF) along Q_{dimer} is plotted as the green curve in both (A) and (B), whose value is indicated by the rightmost vertical axis. 2D free energies along Q_{H2A} versus Q_{dimer} (A) and Q_{H2B} versus Q_{dimer} (B) are displayed as contour maps. The blue-to-red color legend represents free energy values from high to low. Adapted from Haiqing Zhao's original figure with permission.

of the two energy minima of H2B are deeper than those of H2A. On the other hand, the free energy landscape of H2A is relatively more rugged and frustrated. This finding suggests that H2B provides more drive for the transition between the unfolded and folded state, compared with H2A. In other words, our thermodynamical analyses indicate that during the binding/folding process of H2A/H2B, H2A's folding dynamics is more glass-like than H2B, which has a more funneled folding landscape [240].

4.3.4 Experimental Confirmation by NMR and Circular Dichroism

We then tested our computational predictions using NMR and circular dichroism (CD) measurements on H2A and H2B. ^{1}H-^{15}N NMR spectra of H2A alone (Figure 4.5A) shows a narrow spread of NMR signals resulting in signal crowding in the region typical for amide signals of unstructured/unfolded proteins. This experimental work was performed by Dulith Abeykoon and Christina Camara, graduate students in Prof. David Fushman's laboratory. The negative or close to zero signal intensities observed in the heteronuclear NOE spectrum of ^{15}N-labeled H2A recorded upon pre-saturation of amide protons (Figure 4.5C) are a clear indication that the protein is unstructured and highly flexible. Upon addition of unlabeled H2B we observed a dramatic change in the ^{1}H-^{15}N NMR spectra of ^{15}N-labeled H2A, wherein new signals (corresponding to the bound state) appear and increase in intensity until they saturate at ca. 1:1 H2B:H2A molar ratio (Figure 4.5B). Concomitantly, the unbound signals reduce in intensity and practically disappear at the saturation point. This behavior of the NMR signals, which exhibit essentially no shifts, indicates that the binding is in slow exchange regime on the NMR chemical shift time scale. In contrast to the unbound state, the signals of ^{15}N-labeled H2A in complex with H2B (Figure 4.5D) show a significant spread, indicating that the bound state of H2A is well structured. Also, many H2A signals in the heteronuclear steady-state NOE spectra recorded at these conditions have positive intensities, characteristic of a well-folded state of the protein (see Fushman *et al.* [241]). Similar behavior was observed for ^{15}N-labeled H2B, which is unstructured in the unbound state and folds

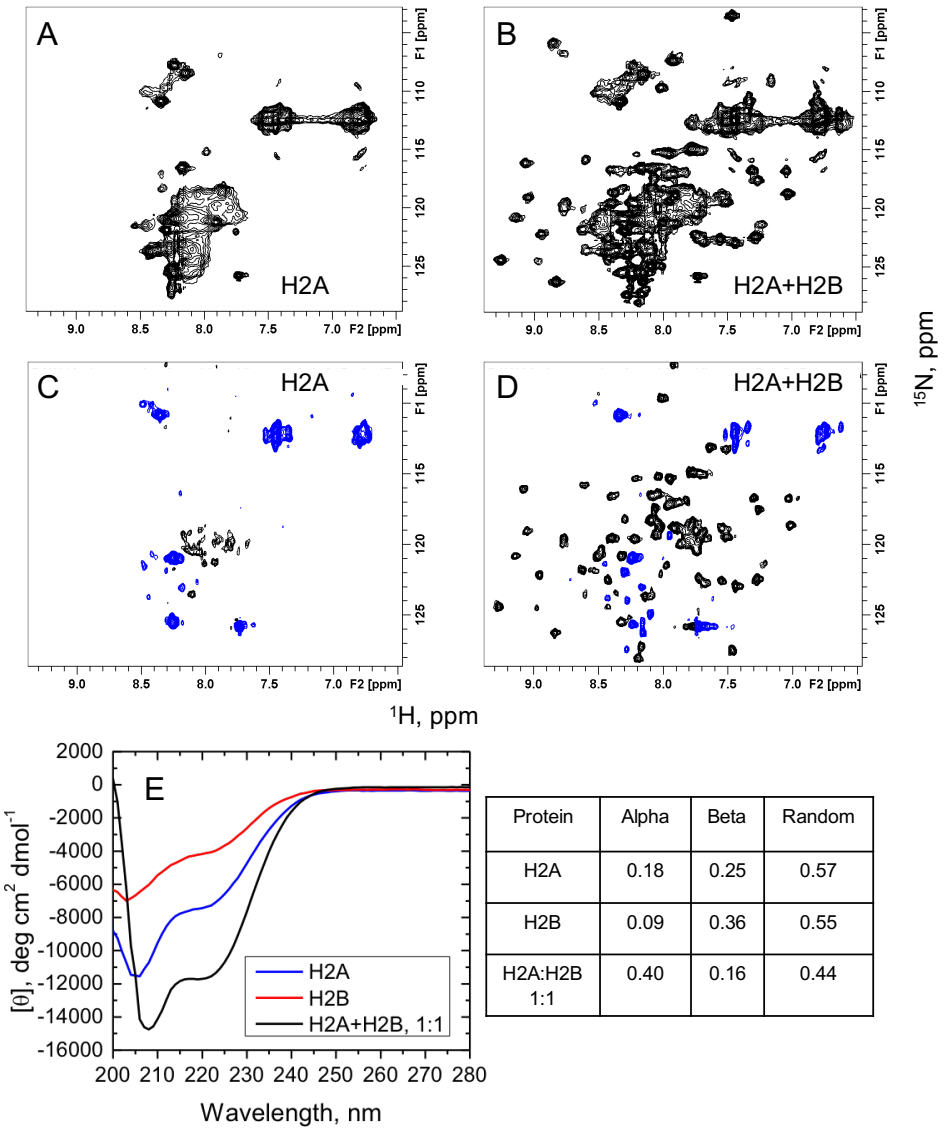
upon complex formation with H2A (data not shown).

The NMR data above suggested that only with each other can H2A and H2B fold into a dimer with well-defined structures. To extend this analysis further, we performed CD, which allows one to assess the folded helical content of a protein experimentally. Here, the CD results demonstrate a significant increase in the helical content of these proteins upon the formation of the H2A/H2B heterodimer (Figure 4.5E). Together, these experimental results indicate that in isolation H2A and H2B are intrinsically disordered but adopt a well-defined tertiary structure upon binding to each other, which is consistent with a previous experimental study on H2A/H2B's thermodynamical stability [25]. Overall, these experiments serve as strong support of the previous computational hypotheses.

4.3.5 Polymer Scaling Law

Our results indicate that histone dimers H2A/H2B and H3/H4 have similar "binding coupled to folding" mechanisms, which indicate that monomers cannot be well folded on its own but can be folded with the other histone partner. This suggests that histone dimers should be considered as an independent folding unit, similar to standard globular protein monomers. In the following, we further discuss this point of view from the perspective of polymer biophysics.

For many classes of polymer, the radius of gyration for a polymer chain (R_g) approximately follows the scaling relation: $R_g \sim \alpha N^{\nu}$, where is the radius of gyration of the polymer. N is the number of bond segments (*i.e.* the degree of polymer-



Protein	Alpha	Beta	Random
H2A	0.18	0.25	0.57
H2B	0.09	0.36	0.55
H2A:H2B 1:1	0.40	0.16	0.44

Figure 4.5: **NMR and CD studies of H2A and H2B upon complex formation.** (A-B) ^{1}H-^{15}N SOFAST-HMQC spectra of 15N-labeled H2A alone (A) and in the presence of unlabeled H2B at a 1:1 molar ratio (B). (C-D) Heteronuclear steady-state ^{15}N{^{1}H} NOE spectra recorded with amide proton presaturation [241] for ^{15}N-labeled H2A alone (C) and in the presence of unlabeled H2B at a 1:1 molar ratio (D). In these spectra, contours with positive intensities are colored black while negative intensities are blue. (E) CD spectra of H2A and H2B alone and in a 1:1 mixture. The concentrations of the proteins are the same in all three cases. The table on the right shows the percentage of the secondary structure obtained from these data. Panels (A-D) were provided by Dulith Abeykoon, panel (E) by Christina Camara.

ization) of the chain α is the linear slope, and ν is the scaling exponent [242]. After analyzing a large structural dataset of monomeric proteins, Dima *et al.* [243] verified this scaling relation and concluded the empirical parameters $\alpha \simeq 3$ and $\mu \simeq 1/3$ achieving a correlation coefficient of 0.90 for globular proteins.

With this in mind, we calculated the R_g of the crystal structures of histone monomers and dimers and fitted them to the empirical relation from Dima *et al.* (Figure 4.6A) [243]. This part of the work was performed by Haiqing Zhao. We found that all the histone monomers have a higher R_g than the expected value of a globular protein with the same residue length, while the R_g and N of both histone dimers fit closely with the empirical relation (the black line). Together with the geometry of histone fold structural motif, where three helices of one histone cross and bind with another three helices from the partner histone, this analysis supports the point that structurally, histone dimers represent a single folding unit.

4.3.6 Evolution of the Histone Fold

After confirming that histone heterodimers are a single folding unit, we next discuss from an evolutionary perspective one possible reason why they are split into two monomers in eukaryotes, instead of remaining a single-chain protein. Indeed, it is found that an ancient archaeon, *Methanopyrus kandleri*, produces a 154-residue single-chain histone (HMk) which is homologous to the eukaryotic histone heterodimers and shares the similar histone-fold structural motif [244] (Figure 4.6B). It is possible that during evolution, the eukaryotic histone dimers inherited the main

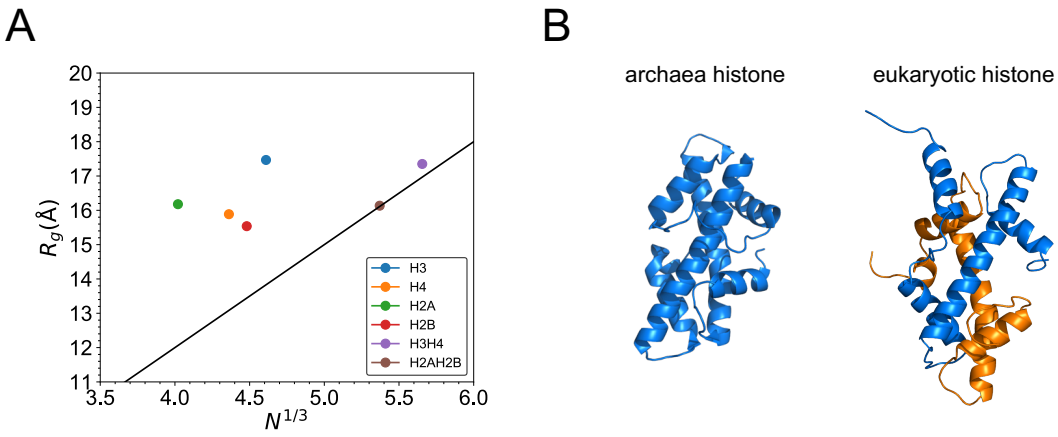


Figure 4.6: **Dimeric histones may originate from a single unit.** (A) R_g versus the residue number N is plotted for monomeric and dimeric histones H3, H4, H2A, H2B, and H3/H4, H2A/H2B. The black line is the empirical correlation between R_g and N for globular monomeric proteins [243]. (B) The archaeon *Methanopyrus kandleri* histone (left) folds as a monomeric chain (PDB ID: 1F1E [244]), while the eukaryotic histone (right) displays a dimeric structure (blue: H3, orange: H4, PDB ID: 1AOI [9]). Adapted from Haiqing Zhao's original figure with permission.

structural motif and folding mechanism from their ancestor proteins, but diversified into two different components to allow for more diverse biological functions needed for higher organisms [41, 43]. Possible functions may include but not limited to distinctive post-translational modifications on each monomer [48,131,132] and different structural and functional roles of the two composing partners as revealed here and in our previous works [106, 107]. On the other hand, this structural diversity may affect the folding rate of each monomer in a dimeric unit, which requires chaperones to interact with the dominant monomer and assist its folding, such as HJURP for CENP-A [106, 245]. Moreover, the disassembly kinetics of histone dimeric pairs may play an important role in the inherently asymmetric chromatin remodeling process. [246, 247]

4.3.7 Histone Fold Proteins Share a Similar Folding Mechanism

Despite the diverse evolution, the histone fold structural motif is surprisingly well conserved, not only among eukaryotic histones, but also in many other DNA-binding proteins that participate in a wide variety of biological processes, such as transcription, translation, and DNA metabolism [28, 42]. A better understanding of the folding mechanisms of these histone fold proteins (HFP), which has been enigmatic, may help elucidate their detailed functions. To address this question, we simulated two representative HFP systems, namely the TFIID TATA box-binding protein associated factors $dTAF_{II}42/dTAF_{II}62$ [248], and the transcription factor NF-YB/NF-YC [249], using the same protocol as in the previous histone annealing

simulations. The histone fold motif can be recognized from the native structures of these two HFP, where dTAF$_{II}$42/dTAF$_{II}$62 resembles H3/H4 and NF-YB/NF-YC is similar to H2A/H2B. As shown in Figure 4.7AB, these HFP monomers alone can only fold into an intermediate state, with a $Q \sim 0.32 - 0.39$, but not the native state, which is similar to the eukaryotic histones' behavior. In the presence of the binding partner, dTAF$_{II}$42/dTAF$_{II}$62 can fold and bind into the nearly perfect native dimer, with very high $Q_{monomer} \sim 0.75$ and $Q_{dimer} \sim 0.56$ (Figure 4.7C). In the NF-YB/NF-YC dimer binding simulations, each monomer folds into a near-native state ($Q_{monomer} \sim 0.68$), while the entire dimer conformation is roughly correct ($Q_{dimer} \sim 0.48$) with small discrepancies (Figure 4.7D). The reason for this behavior might be that the sequence length of NF-YB/NF-YC (165 AA) is longer than that of dTAF$_{II}$42/dTAF$_{II}$62 (138 AA), making the former's structure more difficult to predict. But overall, these results demonstrate that these two HFP fold and bind similarly to the eukaryotic histones, despite their relatively low sequence similarity (see Figure C.1 for their sequence alignments). We also measured the R_g versus N relation of these two HFP and found the dimers, but not the monomers, agree with the monomeric protein trend line and should be treated as a single folding unit (Figure C.2). Altogether, these data point to the fascinating possibility that HFP and histones share similar folding/binding mechanisms, which further suggests their common evolutionary origins and may help explain their universal protein-DNA interaction pattern in the diverse functions [41].

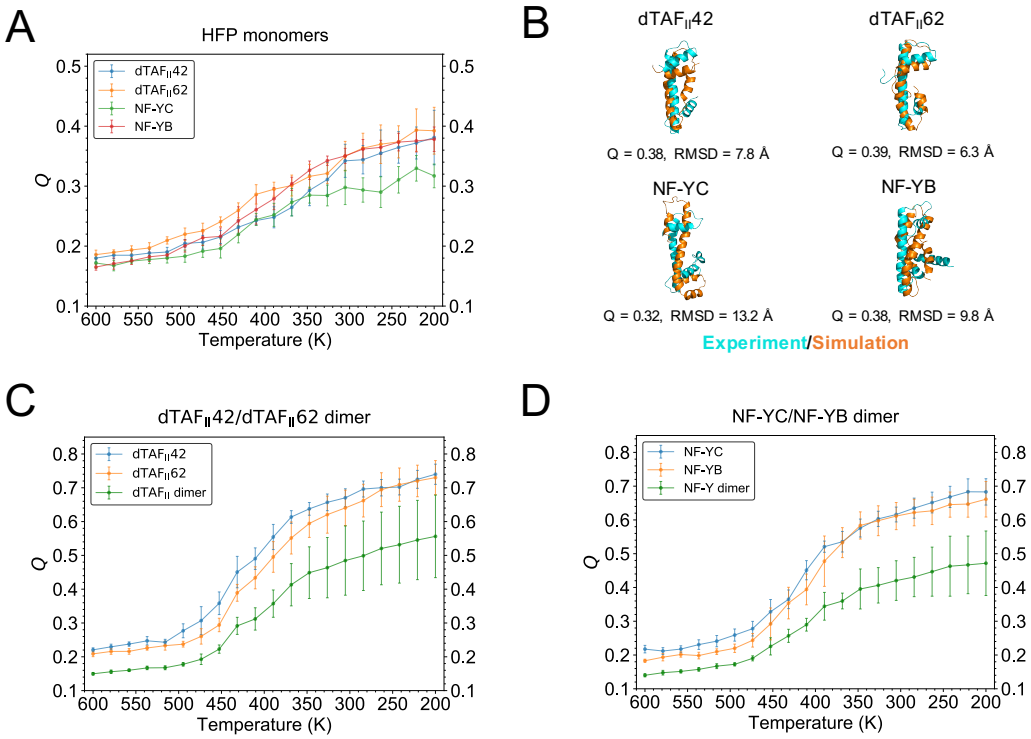


Figure 4.7: **Histone fold proteins have similar folding mechanism as the eukaryotic histones.** (A) Q analyses on annealing simulations of the four HFP monomers ($dTAF_{II}42$, $dTAF_{II}62$, NY-FC, NF-YB), with the representative snapshots of the predicted conformation (orange) superimposed with the crystal structure (cyan) in (B). (C-D) shows the $Q_{monomer}$ and Q_{dimer} analyses in the dimer annealing simulations on $dTAF_{II}42$/$dTAF_{II}62$ (C) and NF-YB/NF-YC (D). Subsets are the representative snapshots in the simulation (red) aligned with the crystal structure (cyan).

4.4 Conclusions

In this report, we studied the structure formation mechanisms of histones. Using computer simulations, NMR, and CD measurements, we showed that histones only fold upon binding. Furthermore, we extended our computational study to other proteins with the histone fold motif and found that the same folding upon binding principle widely applies to this structural motif. Besides, our work shows that two histone monomeric components contribute to their binding process asymmetrically, which may shed light on understanding the evolutionary origins of histone variants.

Chapter 5: Summary and Future Prospects

In this thesis, we have reported on the structure and dynamics of disordered histones to understand their important biological functions. The major approach we have used is an advanced computational model called AWSEM, guided by physical principles and verified by experimental results. A series of theoretical and computational studies included in this thesis have revealed essential biophysical mechanisms of various disordered histones, such as histone tail dynamics, H1-nucleosome binding conformations, and histone monomer/dimer folding thermodynamics. As a comprehensive complement and extension to a series of histone and chromatin studies performed at the Papoian lab in the past decade [250–253], this thesis provides new insights on disordered histone's unique functions in chromatin condensation and gene regulation.

All the research projects in this thesis would not be successfully performed without the development and modifications of AWSEM to model IDP and protein-DNA interactions. As an open-source simulation package, AWSEM-IDP (described in Chapter 2) has been widely used by many research groups to study not only disordered histones but also other IDPs in general. Even researchers with less experience in MD simulations can easily benefit from AWSEM-IDP's clear source

code, documentation, and example files, all available for free on GitHub (https://github.com/adavtyan/awsemmd/wiki/AWSEM_IDP). The release of AWSEM-IDP has also inspired further development and application of coarse-grained models for IDP [95, 254–258]. The improved AWSEM-DNA branch (to be released after the publication of Chapter 3) is a more systematic and robust approach to model protein-DNA complexes such as nucleosome [111], transcription factors [101], and other DNA-binding proteins [110]. The successful models of these protein-DNA complexes will enable large scale simulations to investigate many crucial biological processes related to gene transcription and regulation, such as molecular stripping [259] and breakpoint resection [260]. We expect these simulations to improve our understanding of related biological mechanisms by providing substantial dynamic information at near-atomic molecular details for a long timescale, which is beyond the reach of many current computational and experimental approaches. Meanwhile, there have been consistent efforts to simulate very large biomolecular complexes with AWSEM using parallel computing strategies [261]. In particular, a new efficient AWSEM branch implemented in OpenMM [262] with GPU acceleration, are under active development and tests. When these upgrades are successfully combined with IDP and DNA branches, it would be very promising in near future to computationally model the large-scale dynamics of massive biomolecular complexes, such as chromosome organization proteins and nucleosomal arrays, for a sufficient timescale without sacrificing near-atomic structural details.

One of the most exciting scientific outlook extended from this thesis is to simulate the nucleosomal array with disordered tails and the linker histone H1.

Based on our past comprehensive study on the dynamics of H1 bound with a single nucleosomal particle in Chapter 3, it is desirable to add more nucleosomes and H1 so that the inter-nucleosomal conformation and dynamics can be directly modeled. The X-ray crystal and cryo-EM structures of di-nucleosome [263], tetra-nucleosome [264], hexa-nucleosome [69], and chromatin fiber of tetra-nucleosomal units [12] can all serve as useful initial conformations for the future simulations. Thereby, we will be able to precisely measure how the disordered histone tails and linker histones affect the distance and orientation among nucleosomes for the first time. This proposed research will explore the configuration landscape of nucleosome arrays and reveal how the histone tail deletion [265] and H1 depletion [63] regulate chromatin condensation level, from a high-resolution computational perspective.

There are still some key biophysical questions in the chromatin research field that are beyond the current capability of the AWSEM model. Post-translational modifications (PTM) on histone are among one of these topics. As described in Chapter 1, histone modifications have been proved to play important roles in many fundamental epigenetic functions such as chromatin condensation [266] and gene expression [267]. However, most PTM modify a small functional group on amino acid side chains, which are only represented by one single bead in the coarse-grained AWSEM. Limited by its resolution, the current version of AWSEM cannot precisely mimic PTM in general. Some progress has been made to implement phosphorylation, one of the key PTM in many biological processes, in AWSEM and its predecessors by replacing the phosphorylated residues with "super-charged" glutamic acids [208, 268, 269]. Similar or more advanced approaches are expected to realisti-

cally mimic other PTM on histones, especially the most ones such as acetylation, methylation, and ubiquitylation.

Another challenging topic for AWSEM to study is how the DNA sequence affects nucleosome compaction and chromatin structure. As the essential genetic information, different DNA sequences also require distinct free energy cost to wrap around nucleosomes, thus affecting their positioning and stability [270]. Although the new AWSEM-DNA introduced in this thesis can model the nucleosome with comparable kinetics to experiments, its protein-DNA interactions are still simplified as sequence-independent Leonard-Jones potential and electrostatics. Therefore, a well-calibrated sequence-dependent protein-DNA interaction is essential for AWSEM to investigate related key properties such as nucleosome repeating length (NRL) [271] and breathing [272] more accurately. Continuous efforts have been made to develop more realistic protein-DNA interactions in AWSEM, which is dependent not only on DNA's nucleic acid sequence but also on protein's amino acid sequence. This new AWSEM-DNA branch under development and testing relies on another CG DNA model developed by our lab [108, 273] that is possibly more compatible with AWSEM than the currently used 3SPN.2 DNA model. The successful release of this next-generation AWSEM-DNA will enable more accurate simulations of nucleosome and chromatin to realistically investigate their sequence-related kinetic and thermodynamic properties.

Overall, these future upgrades and improvements on the AWSEM model will allow us to accurately simulate nucleosomal arrays with disordered histones and different DNA sequences and PTM types to probe their regulatory roles on chromatin

condensation. The future prospects based on the existing results in this thesis will push the boundary of chromatin research by shedding light on the mechanisms of gene expression regulation and epigenetic diseases.

Appendix A: Supporting Information for Chapter 2

This appendix is based on the supporting information of the published work of the authors: *Hao Wu, Peter G. Wolynes, and Garegin A. Papoian; AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins; The Journal of Physical Chemistry B, 122(49):11115-11125 (2018)* [116]

A.1 Parametrization Procedure

In Chapter 2 we discussed the general steps to calibrate the parameters in AWSEM-IDP. Here we elaborate on the detailed procedure and take H4 tail and PaaA2 as examples. V_{R_g} is system-dependent because the residue number (N) and the target radius of gyration (R_g^0) can be very different among various systems. Hence we performed parameter calibration for V_{Hbond} and V_{FM} first. In V_{Hbond} the scaling factors of V_β, V_{P-AP}, and $V_{helical}$ (λ_β, λ_{P-AP}, and $\lambda_{helical}$) were modified to control the overall secondary structure propensity. It turned out that the default values of λ_β and λ_{P-AP} (both 1.0) matches the benchmark β structure level well. $\lambda_{helical}$ was reduced to 1.2 to fit the benchmark α-helical structure level. In V_{FM}, we tuned the scaling factor (λ_{FM}) and cutoff range of ij separation along the sequence ($|i-j|_{min}$ and $|i-j|_{max}$), which set the relative intensity of V_{FM} and the range that

i and j go over in the calculation of r_{ij} and r_{ij}^m. In terms of memory selection, we chose 100 snapshots from the ~85000-frame replica-exchange atomistic simulation trajectory [132] for H4 tail and all the 50 structures generated with SAXS and NMR ensemble restrictions [166] for PaaA2. Noticing that λ_{FM} also depends on the number of fragment memories, we used 0.001 for H4tail and 0.002 for PaaA2 to keep their relative weights the same. As for V_{R_g}, the parameters generally vary among different systems. The range of parameters D, α, and β used in this study are described in Chapter 2 (Table 2.1). γ is tuned to compensate for the over-compact effect from other terms in AWSEM-IDP and normally ranges from 1.1 to 1.2. The rest two parameters N and R_g^0 depend on the residue number and target R_g value. For the two IDPs we studied in this report, $D = -0.2$, $\alpha = 0.001$, $\beta = 0.003$, $\gamma = 1.16$, $N = 26$, $R_g^0 = 8.6$ for H4tail, and $D = -0.8$, $\alpha = 0.001$, $\beta = 0.0005$, $\gamma = 1.11$, $N = 71$, $R_g^0 = 20.8$ for PaaA2. We tested multiple sets of parameters on both systems and compared results with atomistic simulation or experiments concerning various structural properties. With the current set of parameters, we can obtain results comparable with atomistic simulation and experiments.

A.2 RMSIP Analysis

The convergence of all the simulations is confirmed by the root mean square inner product (RMSIP) analysis [164], which quantifies the overlap between essential subspaces with the inner product of the first ten principal eigenvectors of CA atom coordinates.

$$\text{RMSIP} = \left(\frac{1}{10}\sum_{i=1}^{10}\sum_{j=1}^{10}(\boldsymbol{\eta_i}\cdot\boldsymbol{\nu_j})^2\right)^{1/2} \tag{A.1}$$

To calculate RMSIP, we selected subparts with increasing time length from the whole trajectory, divided each subpart into two halves, and calculated RMSIP of these pairs. Figure A.1 shows that an early convergence appears even in the beginning of the simulation, with RMSIP around 0.7. Then the RMSIP curves gradually increase and become saturated at around 0.8. These results are strong proof of convergence of both H4tail and PaaA2 simulations.

A.3 Energy Analysis

In the Results and Discussion section, we analyzed different energy terms in AWSEM-IDP Hamiltonian responsible for the protein secondary and tertiary structures. Here we provide more data on the detailed contribution from each energy term (Table A.1), as well as the time-evolution of $E_{secondary}$ and $E_{tertiary}$ as defined in A.2 (Figure A.6). For each protein, ten separate simulation runs were performed. After cutting off the first 10 ns, those simulation runs are combined for analysis.

$$\langle E_{secondary}\rangle = \frac{1}{N}\langle V_{Rama} + V_{Hbond}\rangle, \langle E_{tertiary}\rangle = \frac{1}{N}\langle V_{contact} + V_{burial}\rangle \tag{A.2}$$

Table A.1: **Detailed AWSEM Hamiltonian for all the simulated proteins**

	H4 tail	**PaaA2**	**1UZC**	**1R69**	**1UBQ**
Residue #	26	71	69	63	76
$\boldsymbol{E_{con}}$	30.91 ± 4.32	84.33 ± 7.1	81.88 ± 7.02	74.76 ± 6.68	90.26 ± 7.37
$\boldsymbol{E_{chain}}$	15.19 ± 2.96	48.64 ± 5.11	44.93 ± 4.81	39.87 ± 4.56	49.28 ± 5.03
$\boldsymbol{E_{\chi}}$	2.63 ± 1.09	10.96 ± 2.18	10.86 ± 2.19	9.13 ± 2.01	12.42 ± 2.42
$\boldsymbol{E_{excl}}$	1.34 ± 1.10	4.10 ± 1.80	4.32 ± 1.82	4.24 ± 1.82	6.3 ± 2.25
$\boldsymbol{E_{rama}}$	-32.47 ± 3.13	-159.74 ± 5.19	-157.65 ± 4.29	-129.87 ± 4.07	-161.04 ± 3.95
$\boldsymbol{E_{contact}}$	-5.31 ± 2.00	-17.02 ± 4.76	-18.78 ± 4.36	-35.34 ± 3.95	-64.29 ± 4.62
$\boldsymbol{E_{burial}}$	-22.86 ± 0.62	-61.43 ± 0.98	-58.12 ± 1.24	-53.21 ± 1.28	-67.03 ± 1.00
$\boldsymbol{E_{\beta}}$	-3.34 ± 4.27	-0.82 ± 1.51	-0.73 ± 1.82	-0.03 ± 0.36	-22.17 ± 4.00
$\boldsymbol{E_{P-AP}}$	-12.32 ± 6.20	-5.9 ± 3.49	-6.31 ± 3.71	-12.76 ± 3.18	-18.73 ± 1.89
$\boldsymbol{E_{helix}}$	-0.41 ± 1.08	-41.84 ± 5.81	-66.9 ± 6.63	-49.9 ± 5.62	-14.33 ± 2.26
$\boldsymbol{E_{FM}}$	-13.93 ± 0.84	-118.41 ± 3.21	-171.34 ± 3.36	-136.1 ± 2.89	-295.8 ± 3.61
$\boldsymbol{E_{R_g}}$	-20.54 ± 0.27	-55.93 ± 0.78	-54.63 ± 0.61	-50.38 ± 0.04	-60.8 ± 0.00

[a] Energies in kcal/mol.

[b] All simulations performed at 300 K.

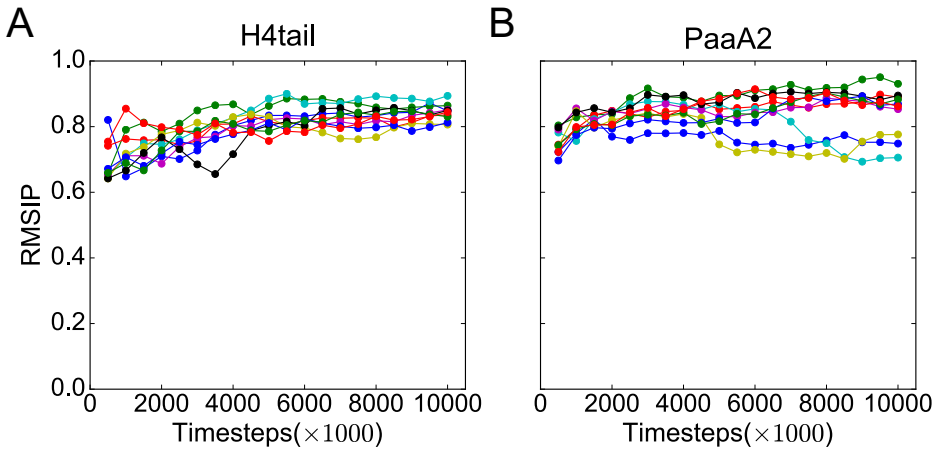


Figure A.1: **RMSIP analysis demonstrates the convergence of AWSEM-IDP simulations.** RMSIP curves of both H4tail (A) and PaaA2 (B) rises steadily with increasing time length and all the RMSIP values above 0.6, showing all the simulations are converged. The results of all the 10 runs are labeled with different colors.

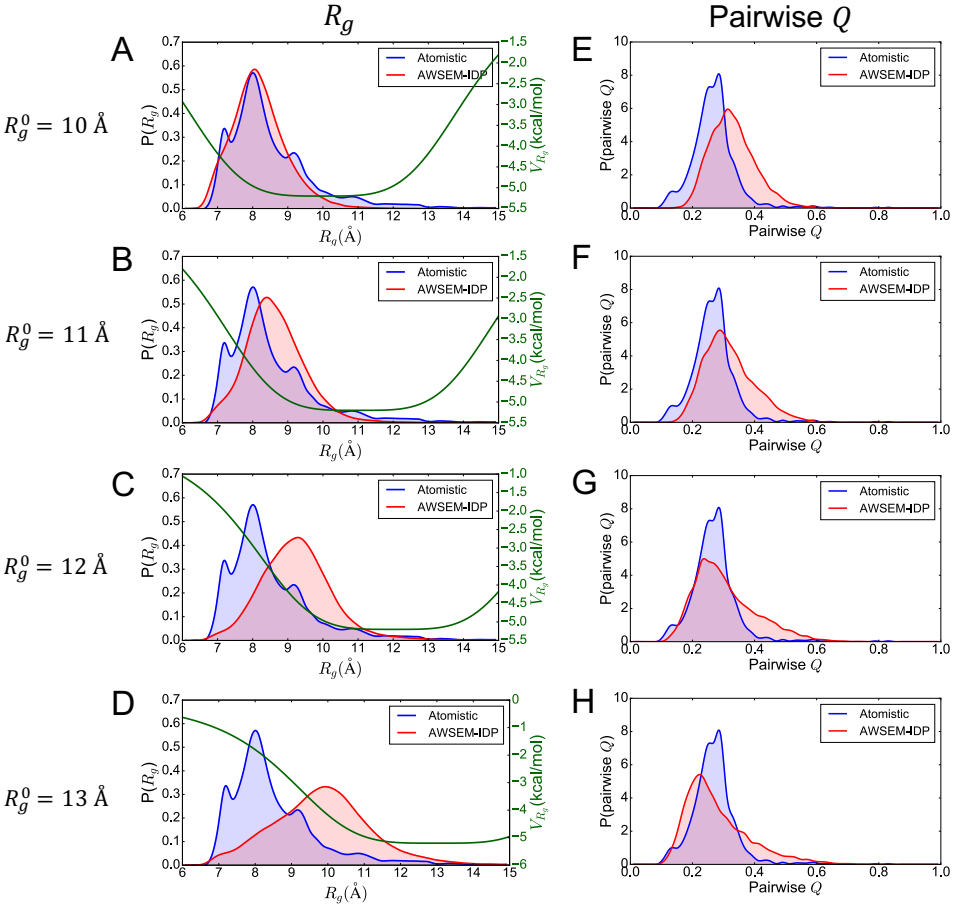


Figure A.2: **A wide range of conformations can be sampled via parameter tuning of V_{R_g}.** We simulated H4 tail with AWSEM-IDP with different $R_g^0 = 10, 11, 12, 13$ in the R_g potential and calculated the corresponding R_g (A-D) and pairwise q (E-H) distributions. When R_g^0 increases, we can observe a higher average value and wider distribution of R_g, accompanied by a smaller average value of pairwise q. All the rest parameters in V_{R_g} are the same. The corresponding V_{R_g} curves are shown in green solid lines (A-D).

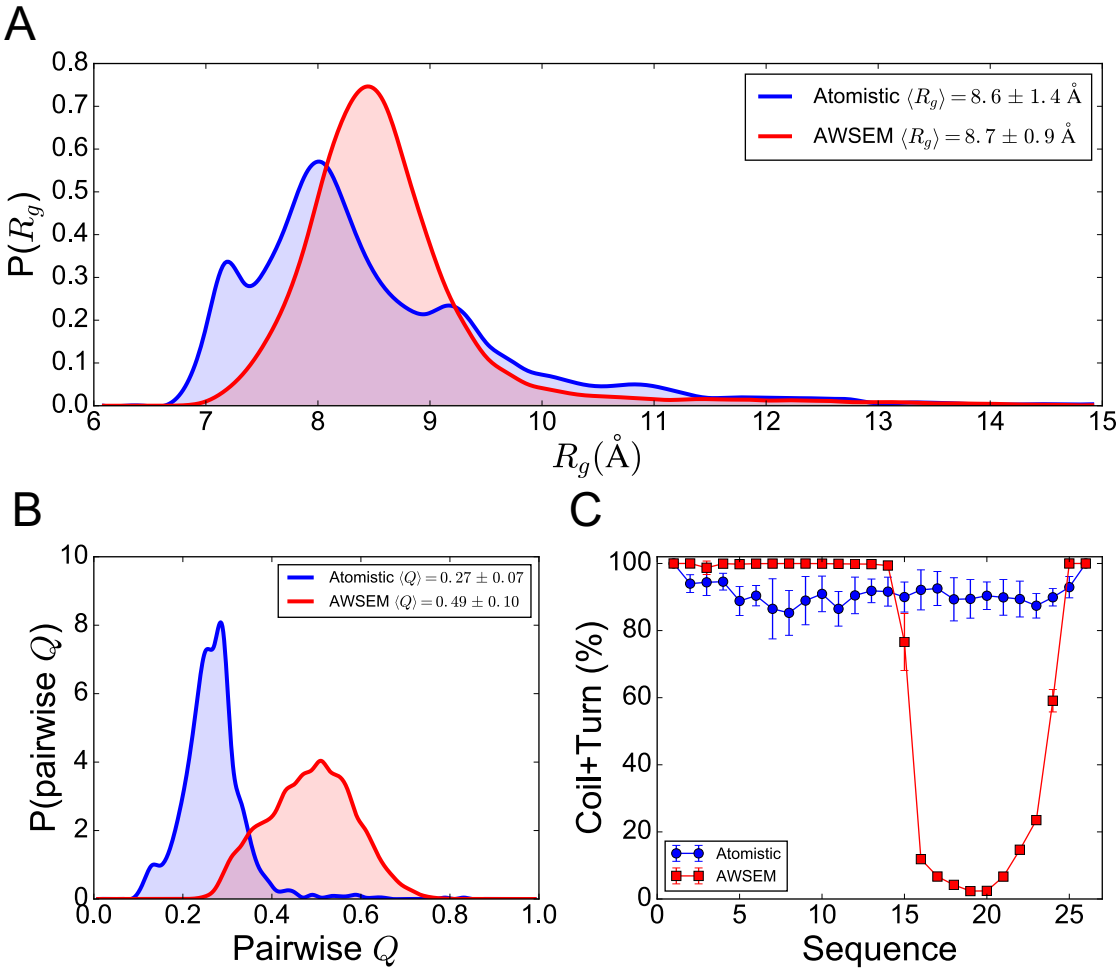


Figure A.3: **Simulations with the standard AWSEM are less accurate in describing the structure of the H4 tail.** The structural metrics and analyzing approaches are the same as in Chapter 2 (Figure 2.4).

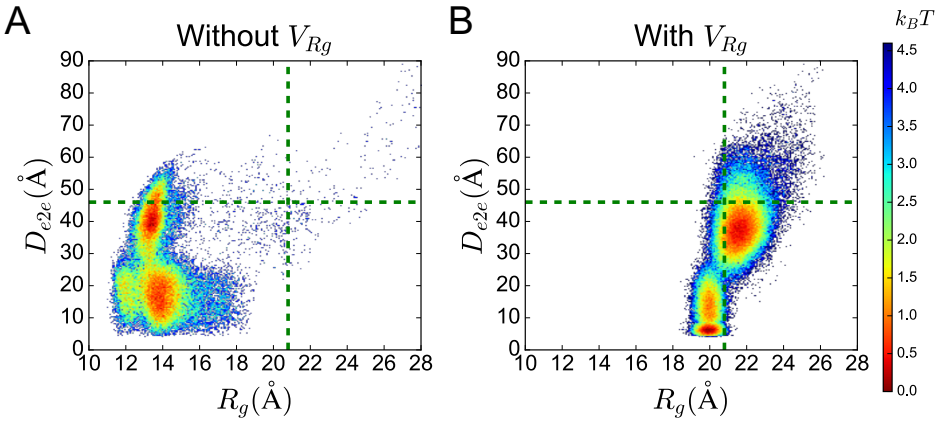


Figure A.4: V_{R_g} **efficiently prevents artificially collapsed conformations of PaaA2.** The effect of V_{R_g} is highlighted by the comparison between free energy landscape with R_g and D_{e2e} as reaction coordinates simulated with (A) and without (B) V_{R_g}. After V_{R_g} with proper parameters is applied, the locations of the major energy minimums shift closer to the NMR average values [166] (green dotted lines).

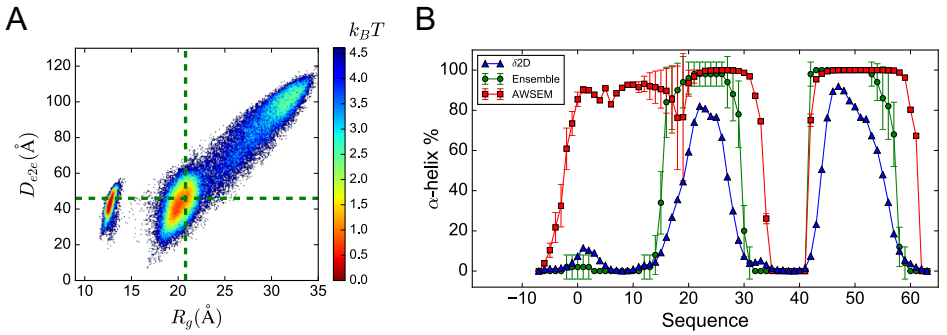


Figure A.5: **Standard AWSEM simulations are less accurate in describing the structure of PaaA2.** (A) Two of the three free energy minima in simulations are distant from NMR average values (green dotted lines). (B) The helical probabilities near the N-terminal region in simulations are much higher than those in experiments.

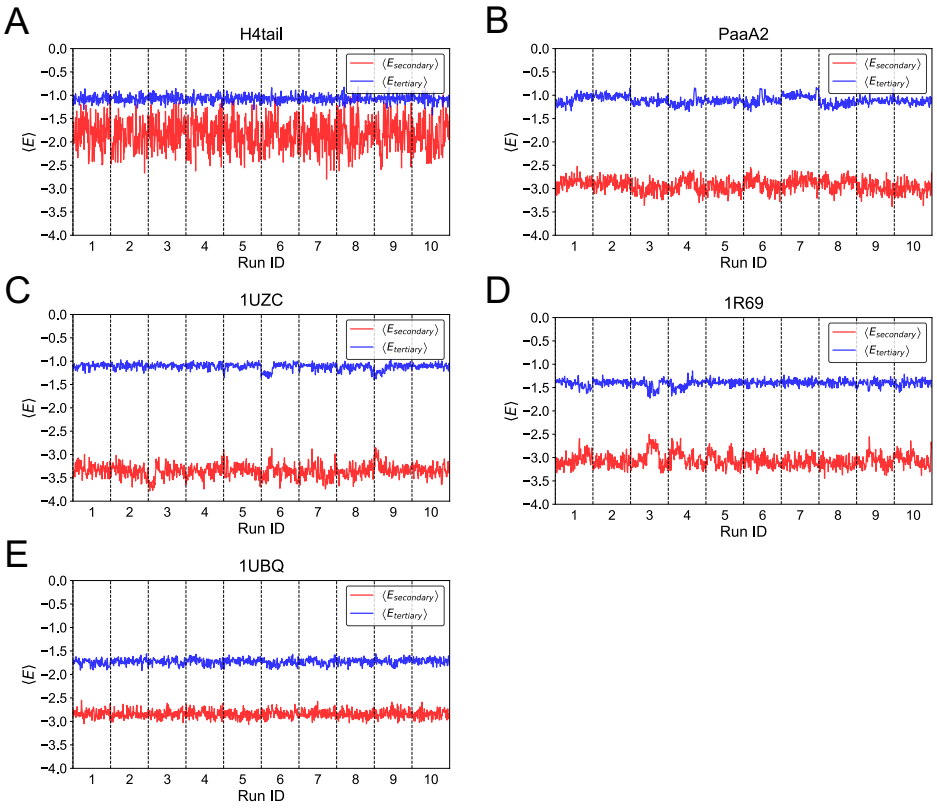


Figure A.6: **The secondary and tertiary structure energy vs time in the simulations of the two IDPs and three ordered proteins.** (A, B) As IDPs, H4tail and PaaA2 have either higher secondary or tertiary structural energy. (C) As a comparison, the overall ordered 1UZC, with a disordered tail, has a similar level of tertiary structure, but much lower secondary structure energy. (D, E) The two entirely ordered proteins (1R69 and 1UBQ) have both lower secondary and tertiary structural energy.

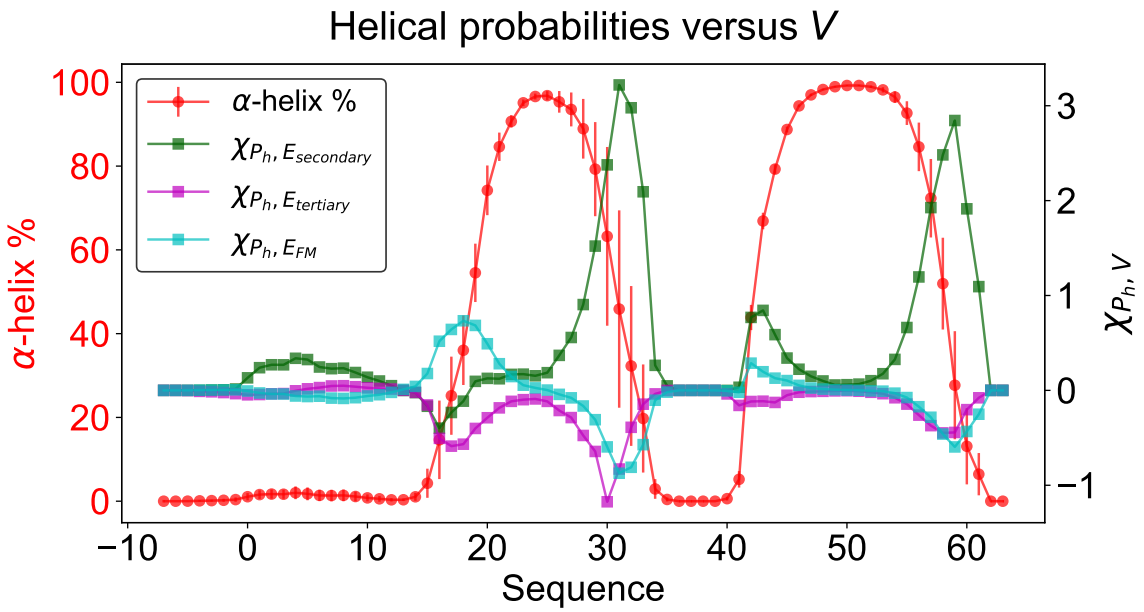


Figure A.7: **Sensitivities versus helical propensities for PaaA2** The sensitivities of helical probability on $E_{secondary}$ (green squares), $E_{tertiary}$ (purple squares), and E_{FM} (cyan squares) are not highly correlated with helical occupations (red circles), but their fluctuations (shown in Figure 2.7B).

Appendix B: Supporting Information for Chapter 3

This appendix is based on the supporting information of the unpublished work of the authors: *Hao Wu, Yamini Dalal, and Garegin A. Papoian; Binding Dynamics of Disordered Linker Histone H1 with a Nucleosomal Particle; In Preparation; (2020)*

B.1 H1 NTD/CTD Atomistic Simulations

Particularly for the full-length H1-nucleosome system, we performed atomistic simulations of the disordered NTD and CTD and used them as bioinformatic local structural bias in the subsequent coarse-grained (CG) simulations. To simulate the relatively large CTD (99 residues) efficiently and obtain local structural information, we separated its sequence into six short overlapping segments and ran individual atomistic simulations for each segment, as done in Lin *et al.* [208]. The 24-residue NTD is much shorter so we simulate its entire structure. See Figure B.1 for their sequences and detailed segmentation.

We used replica-exchange molecular dynamics to enhance sampling when simulating these two intrinsically disordered regions. As for the force field, we used the recently developed a99SB-disp [138] for protein and the modified TIP4P-D [274]

for water. These two force fields are specially optimized for simulating IDPs and prove more accurate than many other atomistic force fields. The temperatures of all replicas were determined by the T-REMD server [275]. The temperature range is 300 - 400 K and the target exchange probability is 0.25, generating 30 - 60 replicas at different temperatures for each simulated system. All the atomistic simulations were performed with GROMACS [276] (version 2018.2) package, with a timestep of 2 fs, periodic boundary conditions, and particle mesh Ewald method for long-range electrostatics. The initial structures (same as the CG simulation) were solvated in a dodecahedral box. Then sodium and chloride ions were added into the solvent to neutralize the system and create a 150 mM physiological salt concentration.

The systems first underwent a short energy minimization of 100 ps using the steepest descent algorithm at 300 K. The NVT equilibration for 500 ps and NPT equilibration for 1 ns were performed subsequently at the unique temperatures of each replica, both with position constraints and Berendsen temperature or pressure coupling methods. Then we performed the production runs (100 ns for NTD and 30 ns for CTD segments), adding up to 1.8 - 3 μs total simulation time of all the replicas for each system. The trajectories at 300 K were used for following analyses and structural bias for CG simulations, excluding the first 20 ns of NTD simulation and 5 ns of CTD simulations for equilibration. The resulting NTD/CTD structures have very low ordered secondary structure propensity (Figure B.2) and extended global size (Figure B.3). These disordered features agree with previous studies in general, proving the sanity of using these structures to bias subsequent CG simulations.

B.2 AWSEM-IDP Potential for H1 Disordered Domains

As mentioned in Chapter 3, we used AWSEM-IDP [116] to model H1 disordered domains, with some recent modifications. Here we elaborate on the details.

B.2.1 Fragment Memory

We used the atomistic simulation trajectories mentioned above to bias the local structure of H1 NTD and CTD in CG simulations. This structural bias is called "fragment memory" and its formula is:

$$V_{\mathrm{FM}} = -\lambda \sum_{m} \omega_m \sum_{ij} \gamma_{ij} \exp[-\frac{(r_{ij} - r_{ij}^{m})^2}{2\sigma^2}] \tag{B.1}$$

where ω_m represents the weight for each memory m. Detailed definitions of other parameters and significance of this potential can be found in previous AWSEM literature [76, 116].

The atomistic simulation trajectories were clustered by the simple linkage algorithm in GROMACS with RMSD = 2.5 Å as the threshold. The representative structures of all the clusters are then used as fragment memory. The weights for each memory are calculated based on the logarithm of cluster size based on the Boltzmann equation, and normalized to be on the same scale with weights ($\omega = 1$) of other globular proteins with single structure as memory in the system, *i.e.* for memory from cluster$_m$, its weight ω_m is:

$$\omega_m = \log(\text{size}_m) / \sum_{j}^{N_{\text{cluster}}} \log(\text{size}_j) \tag{B.2}$$

B.2.2 Sequence-Specific Helical Propensity

$V_{helical}$ is a term in responsible for the formation of α-helices in AWSEM. In the original version of AWSEM-IDP, we reduced its weight to 1.2 for the entire IDP chain. The disordered H1 NTD and CTD, however, are connected with the globular domain in the same protein chain. To solve this inconsistency, we implemented a modification for $V_{helical}$ so that AWSEM user can assign weights for each amino acid in a chain. Here, we reduced $V_{helical}$ weight to 1.0 for H1 NTD and CTD and kept using 1.5 for all the other ordered proteins, including the H1 globular domain.

B.2.3 R_g Potential

We did not use the newly introduced R_g potential from AWSEM-IDP in this study. Because there is no experimentally or computationally measured R_g available for the H1 disordered domains, especially when bound with nucleosomal DNA.

B.3 Electrostatic interactions

All the AWSEM-DNA simulations use two types of electrostatic interactions: DNA-DNA and protein-DNA. There is no protein-protein electrostatics because the contact term in the original AWSEM for protein has already incorporated the short-range electrostatic interactions.

The Debye-Hückel style DNA-DNA electrostatics are directly from 3SPN.2 without modification:

$$V_{\text{elec}}^{\text{DNA-DNA}} = \sum_{n_{\text{elec}}}^{i<j} \frac{q_i q_j e^{-r_{ij}/\lambda_D}}{4\pi\epsilon_o \epsilon(T,C) r_{ij}} \quad r_{ij} < r_c \tag{B.3}$$

where q_i and q_j are the charges on site i and j, r_{ij} the distance between these two sites, λ_D is the Debye length, ϵ_o is the dielectric permittivity of vacuum, $\epsilon(T,C)$ is the dielectric permittivity of solution. λ_D and $\epsilon(T,C)$ are both dependent on ionic concentration and temperature. In this study, we set ionic concentration = 150 mM, temperature = 300 K and r_c = 50 Å to mimic DNA-DNA electrostatics in the physiological cellular environment. The detailed formulae of λ_D and $\epsilon(T,C)$ and values of other parameters in DNA-DNA interactions can be found in Hinckley *et al.* [109]

Similarly, the protein-DNA electrostatics are also modeled in Debye-Hückel style in a non-sequence-specific manner:

$$V_{\text{elec}}^{\text{protein-DNA}} = \sum_{n_{\text{elec}}}^{i<j} \frac{q_i q_j e^{-r_{ij}/\lambda_D}}{4\pi\epsilon_o \epsilon r_{ij}} \quad r_{ij} < r_c \tag{B.4}$$

where all the parameters have the same definition as in Eq. B.3. Here we set λ_D = 10 Å, ϵ = 78, r_c = 40 Å, which results in proper electrostatics in a 150 mM NaCl solution.

B.4 Nucleosome Specific Arginine-Phosphate Potential

To prevent nucleosomal DNA unwrapping from the histone octamer core, we applied an additional Lennard-Jones (LJ) potential between certain C_β beads from arginine residues on histones and phosphate beads from DNA (both in CG representation). Note that this force is only applied to 14 protein-DNA pairs to mimic the effect that some arginines are deeply inserted into DNA minor grooves [9]. See Figure B.4A for the residues and base pairs involved in this force. Its formula is the standard 12/6 LJ potential:

$$V_{\mathrm{LJ}}^{\mathrm{Arg-Phosphate}} = \sum_{\mathrm{pair}}^{n_{\mathrm{pairs}}} 4\epsilon[(\frac{\sigma}{r_{\mathrm{pair}}})^{12} - (\frac{\sigma}{r_{\mathrm{pair}}})^{6}] \quad r < r_c \tag{B.5}$$

where r_{pair} is the C_β-phosphate inter-bead distance for a certain pair of arginine-DNA, ϵ is the strength of the potential, σ is the finite distance at which the potential is zero, and r_c is the cutoff distance. After fine-tuning the parameters, we eventually set $\epsilon = 5$ kcal/mol, $\sigma = 5$ Å and $r_c = 15.5$ Å. With this parameter set, this nucleosome-specific LJ potential provides stronger protein-DNA attraction (see Figure B.4B for its energy compared with other protein-DNA interaction terms) so that nucleosomal DNA can keep wrapping around the histone core. This potential is implemented as a new fix style in LAMMPS as "fix/lj/cut", inspired by the pair style "lj/cut".

To test the effect of this new potential, we run some short simulations (5 ns) for the canonical nucleosome without H1 (PDB: 1KX5) before and after applying the arginine-phosphate potential. All the other setup and parameters are the same

as the H1-nucleosome simulations in Chapter 3. We found without this arginine-phosphate potential, the DNA is highly dynamic and starts to unwrap from the histone core at the end of simulations in two out of five runs (Figure B.4C). By contrast, with the $V_{\mathrm{LJ}}^{\mathrm{Arg-Phosphate}}$, the nucleosomal DNA keeps wrapping around in all the five simulation runs (Figure B.4D).

B.5 Representative Snapshots from 3D Spherical Coordinates

We established a set of spherical coordinates (r, θ, ϕ) of the H1 globular domain's center of mass, as shown in Chapter 3, to quantify their dynamics. We identified all the major basins from their 2D (ϕ, r) histogram and selected the representative snapshots as follows. We first estimated ϕ and r values at the center of each basin. Then we computed the most probable θ value with similar ϕ and r and searched for the corresponding snapshot with this (r, θ, ϕ) as representative ones.

B.6 H1.5ΔC50 Simulations and Analyses

To further comprehensively compare our computational results with the cryo-EM data [66], we also performed simulations for H1.5, another subtype of linker histones, binding with the nucleosome. The C-terminal 50 amino acids of H1.5 in our study were deleted to be consistent with the molecule in the cryo-EM experiment. Therefore this system is called "H1.5ΔC50" (see Figure B.7 for its sequence aligned with H1.0 and structure).

Similar to H1.0, H1.5 also has disordered NTD and CTD. Therefore, we first ran atomistic simulations for H1.5ΔC50 NTD/CTD structural segments. Then we simulated the H1.5ΔC50-nucleosome complex by the AWSEM-DNA force field, with previous atomistic simulations as a structural bias for disordered domains. All the simulation details are the same as in the H1.0 case, except that we only performed 8 independent runs (total 480 ns in CG time scale), limited by computational resource and speed. Then we performed the same linker DNA analyses and compared them with the cryo-EM experiments (Figure B.8). We found our results are in general quantitatively consistent with Bednar *et al.* [66]

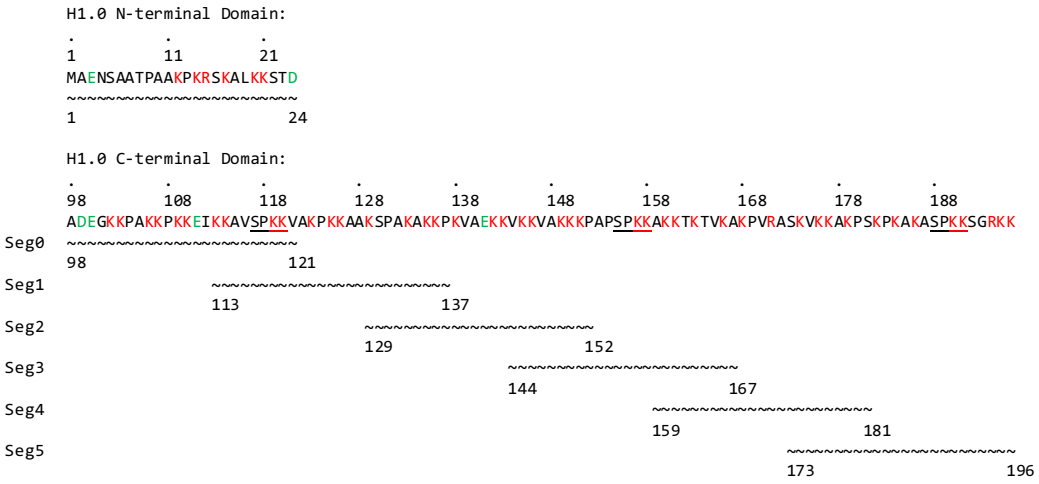


Figure B.1: **H1 NTD (residue 1-24) and CTD (residue 98-196) sequences used in the simulations.** Residues with positively and negatively charged side chains are labeled in green and red respectively. The SPKK repeating motifs in CTD are underlined. Segments for atomistic simulations are represented by $\sim$.

Table B.1: **H1 region definitions for contact analysis**

Domain	Regions	AA/BP_{start}	AA/BP_{end}	Length (AA/BP)
H1NTD	N1	1	12	12
	N2	13	24	12
GH1	α1	28	38	11
	L1/β1	39	46	7
	α2	47	57	11
	L2	58	62	5
	α3	63	78	16
	β2	81	87	7
	β3	90	95	6
H1CTD	C1	98	109	12
	C2	110	121	12
	C3	122	133	12
	C4	134	145	12
	C5	146	157	12
	C6	158	169	12
	C7	170	181	12
	C8	182	196	15
DNA	L1	1	23	23
	dyad	87	107	21
	α3	171	193	23

Table B.2: ϕ-r **basin definitions and population percentages**

Basin ID	r_{min} (Å)	r_{max} (Å)	ϕ_{min} (°)	ϕ_{max} (°)	Population %
g1	60	75	0	150	33.1
g2	30	45	60	140	8.5
g3	55	75	-100	-140	7.9
g4	40	60	-80	-20	8.0
g5	100	120	-60	-30	0.9
f1	60	75	0	150	53.3
f2	30	45	60	140	11.7
f3	55	75	-100	-50	4.0

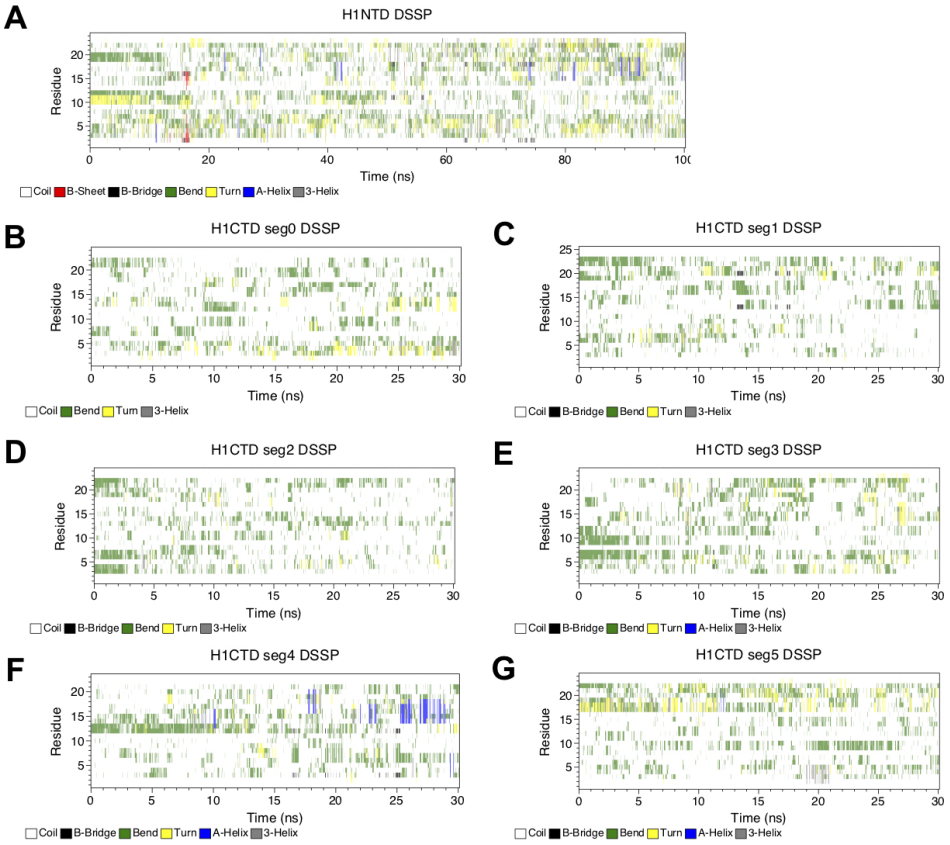


Figure B.2: **H1 NTD and CTD have disordered secondary structures in atomistic simulations.** Secondary structure elements for each residue (vertical axis) along the simulations (horizontal axis) are plotted for H1 NTD (A) and CTD (B-G) segments. Color code represents different types of secondary structure. Analyses were performed using DSSP [277] in GROMACS.

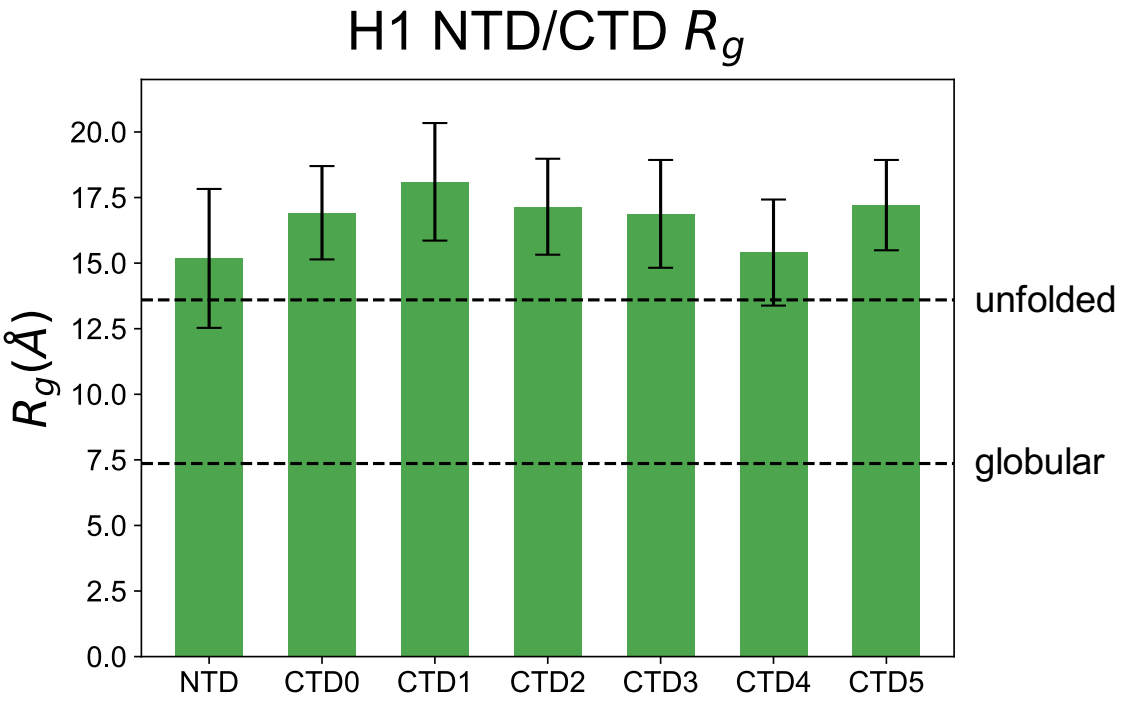


Figure B.3: **H1 NTD and CTD are extended in atomistic simulations.** The radius of gyration (R_g) was computed for all the H1 NTD/CTD segments. The average and standard deviation of R_g are shown as a bar plot with error bars. The theoretical R_g for globular protein [278] and unfolded random coil [279] with the same number of residues (~ 24) are plotted as horizontal dashed lines.

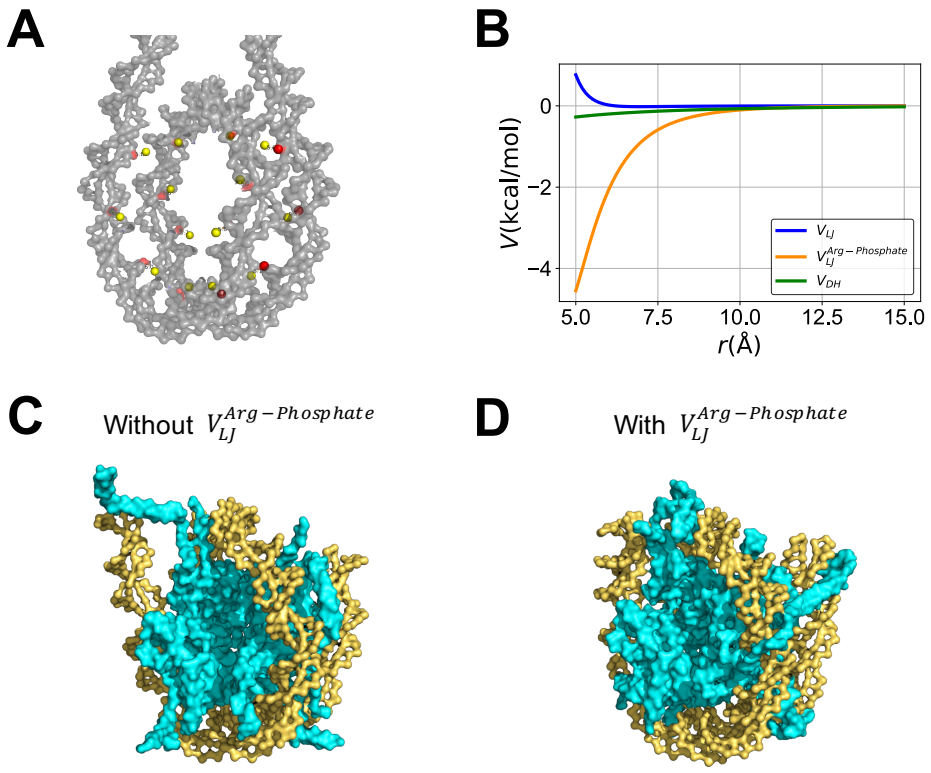


Figure B.4: **A special Lennard-Jones potential is applied between protein-DNA to avoid nucleosomal DNA unwrapping.** (A) shows the locations of all 14 pairs of arginine C_β (yellow) and DNA phosphate beads (red). Nucleosomal DNA is shown in gray, while the histone core and H1 are not shown for clarity. (B) Different protein-DNA energy terms V as a function of inter-bead distance r with the parameter set used in this study. The new potential ($V_{\mathrm{LJ}}^{\mathrm{Arg-Phosphate}}$) creates a significant energy barrier ($\sim$ 3 - 4 kcal/mol) from $r = 5$ Å to $r = 10$ Å, stronger than the standard LJ (V_{LJ}) and Debye-Hückel (V_{DH}) terms. The representative final snapshots of the test canonical nucleosome simulations without (C) and with $V_{\mathrm{LJ}}^{\mathrm{Arg-Phosphate}}$ (D) illustrate the effect of this new potential to avoid DNA unwrapping (gold: DNA; cyan: histone).

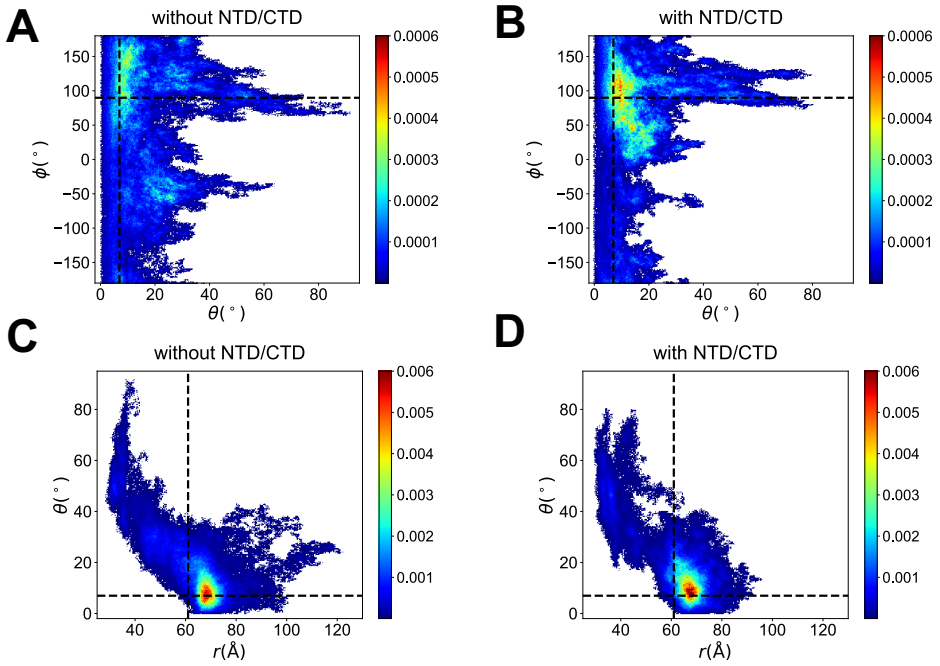


Figure B.5: **Additional 2D histograms of the 3D coordinates for GH1 COM** (A-B): (ϕ, θ) without (A) and with disordered domains (B). (C-D): (θ, r) without (C) and with disordered domains (D). The corresponding values of the chromatosome crystal structure [66] (with GH1 bound on dyad) are labeled with dashed lines.

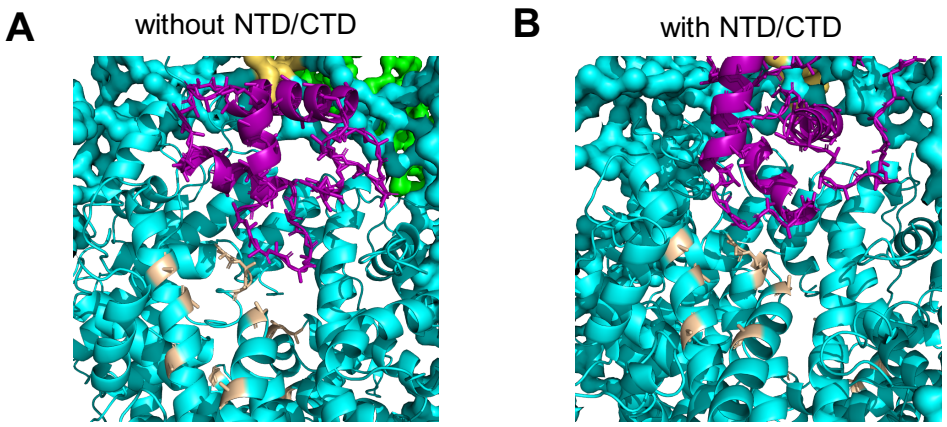


Figure B.6: **Zoomed-in snapshots in which the GH1 is proximal to the histone core acidic patch** GH1 (purple) moves close to the acidic patch (brown) on the histone core in both systems without (A) or with (B) the H1 disordered domains. These two figures as examples are taken from the g2 and f2 representative snapshots in the main text. All the "sticks" representation only show a small part of the amino acid side chains because of the coarse-grained nature of our AWSEM-DNA model.

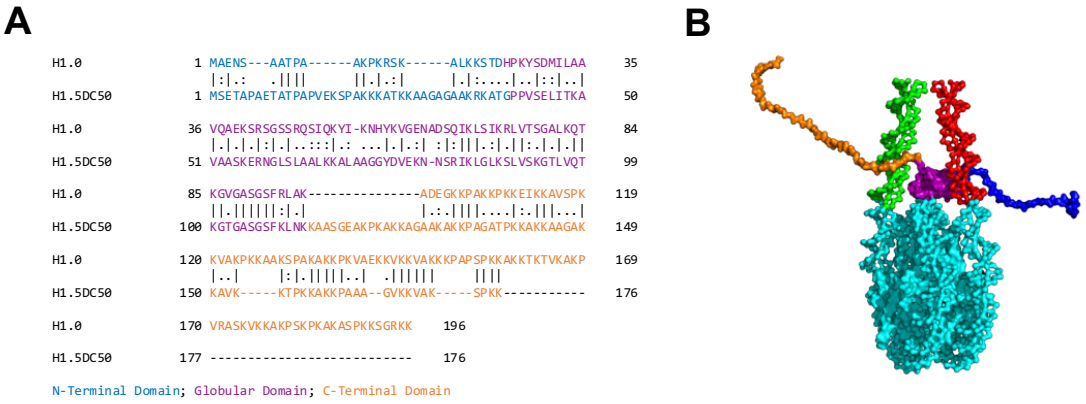


Figure B.7: **H1.5ΔC50 sequence and structure** (A) Sequence alignment of H1.0 (196 AA) and H1.5ΔC50 (176 AA) analyzed with EMBOSS NEEDLE web server [280]. The color code indicates different domains. The identical, similar, and different amino acids are labeled by "|", ".", and ":". (B) H1.5ΔC50-nucleosome complex structure. The color code is the same as in Chapter 3.

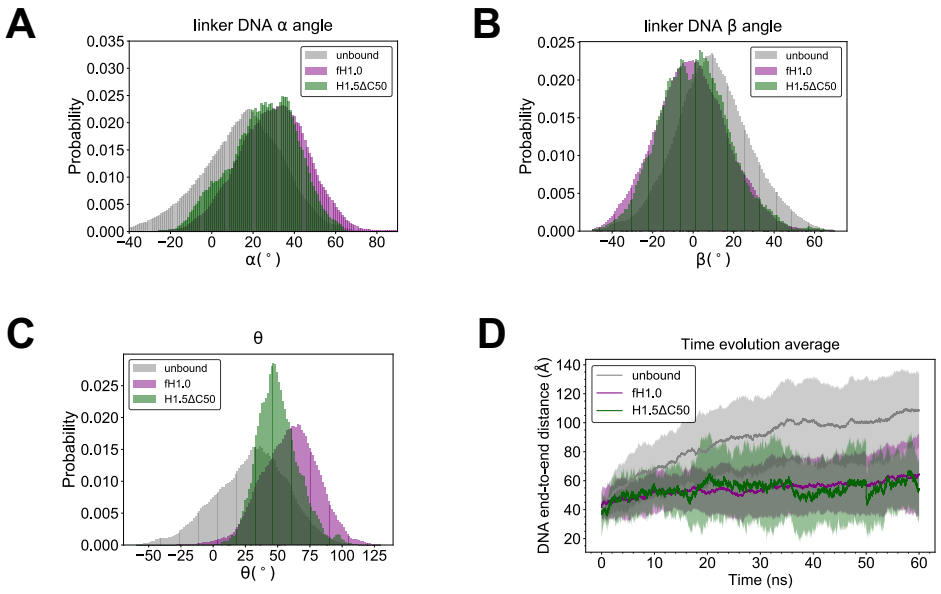


Figure B.8: **H1.5ΔC50 DNA conformation and dynamics are consistent with the previous cryo-EM study.** The DNA related metrics α, β, θ and end-to-end distance for unbound nucleosome, full-length H1.0-nucleosome and H1.5ΔC50-nucleosome are plotted using the same definition as in Chapter 3. (A-B) qualitatively agree with the same measurements in Figure 1E of Bednar *et al.* [66]

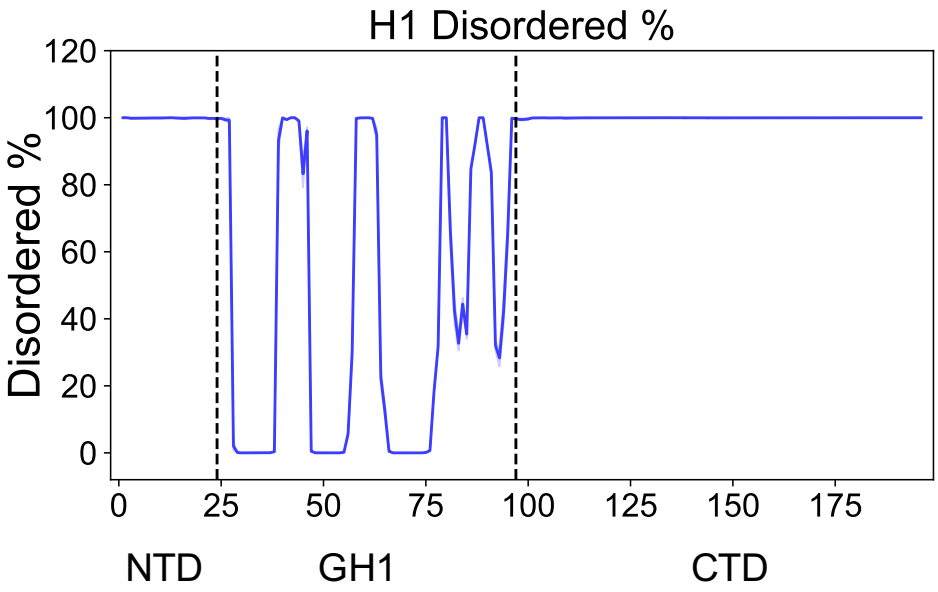


Figure B.9: **H1 remains disordered in the nucleosomal context.** The disordered probability at each residue is defined as the percentage that "turn" or "coil" occurs. Secondary structure elements in this analysis were determined by STRIDE [165] in VMD.

Appendix C: Supporting Information for Chapter 4

This appendix is based on the supporting information of the unpublished work of the authors: *Haiqing Zhao, Hao Wu, Dulith Abeykoon, Alex Guseman, Christina M. Camara, Yamini Dalal, David Fushman, and Garegin A. Papoian; Folding-Upon-Binding Mechanism Widely Exists in Histone Fold Structures; In Preparation; (2020)*

C.1 AWSEM Simulation Details

We used the AWSEM [76] model to simulate all the histone and histone fold protein (HFP) systems in this study. In this study, the parameters in the AWSEM model were tuned such that the simulated melting temperature of histone dimers is around 350 K, as observed in experiments. We also employed an AWSEM-featured bioinformatic term called "fragment memory", using available protein segments as the local structural bias. In histone/HFP monomer annealing simulations, the biasing segments were selected from proteins in the PDB which share similar local amino acid sequences to the histone monomers, equivalent to the "homolog allowed" structural library used in Davtyan *et al.* [76]. In the dimer simulations, the local memory fragments were selected from the X-ray crystal structures, which still only provide

local structure information but not tertiary contacts within each monomer and between monomers (as used in previous protein binding studies with AWSEM [98]). The length of a fragment is typically from 3 to 9 residues.

We ran AWSEM simulations using the open-source molecular dynamics software, LAMMPS [92] (version 9Oct12), with a non-periodic shrink-wrapped boundary condition and the Nose-Hoover thermostat. The simulation time step was set as 5 femtoseconds. The native conformations were taken from the corresponding X-ray crystal structures (PDB: 1AOI [9] for histone proteins; PDB: 1TAF [248] for dTAF_{II}; PDB: 1N1J [249] for NF-Y). All annealing simulations started from the completely unfolded state, and then were slowly cooled down from 600 K to 200 K. The simulation time of a production run is 1×10^7 steps. Ten independent runs with different initial states and velocities were performed for each system.

For histone dimers, we also performed coupled replica-exchange and umbrella sampling simulations to collect sufficient conformational statistics for estimating folding free energies. We first set 10 umbrella windows linearly distributed along the chosen collective variable Q_{dimer}. At every umbrella window of Q_{dimer}, 10 different temperature replicas were run in parallel. The replica that is close to the folding temperature was then collected from every umbrella window, following which WHAM [237] was used to compute the unbiased free energies.

C.2 Q Value Definition

To quantitatively describe the similarity between simulated and native structures, we used the order parameter Q defined as in Davtyan *et al.* [76]:

$$Q = \frac{1}{N_{pairs}} \sum_{i<j-2} \exp[-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}}] \tag{C.1}$$

where N_{pairs} is the number of pairs in the summation, r_{ij} is the instantaneous distance between C_α atoms of residues i and j, r_{ij}^N is the same distance in the native structure, and $\sigma_{ij} = (1 + |i - j|)^{0.15}$ represents the resolution of distance difference.

The range of Q is from 0 to 1. A higher Q value means that the simulated conformation is more similar to the native structure. Note that the group of atoms included for computing Q can be customized. In Chapter 4, we computed $Q_{monomer}$ using the C_α atoms only within a monomer, while Q_{dimer} was calculated using the entire dimer.

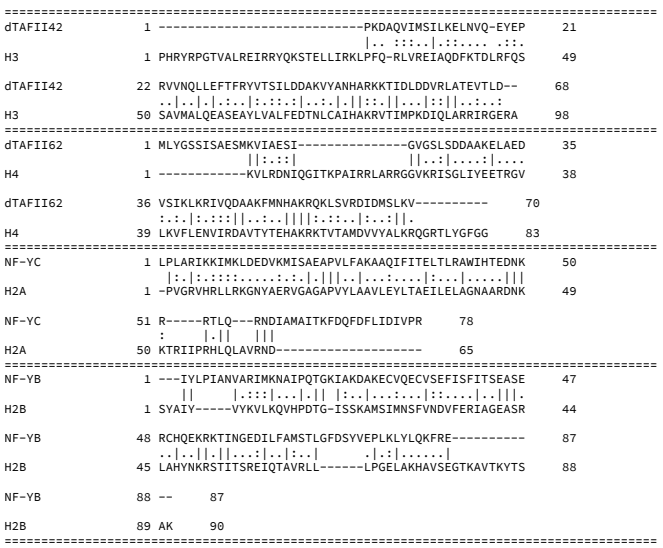


Figure C.1: **Sequence alignments between two HFP systems studied in this work and the similar eukaryotic histones.** All the four HFP monomer sequences are aligned with the most similar eukaryotic histone. Similarities of the aligned sequences are calculated on the right panel. All the alignments are performed via the EMBOSS Needle online server [280].

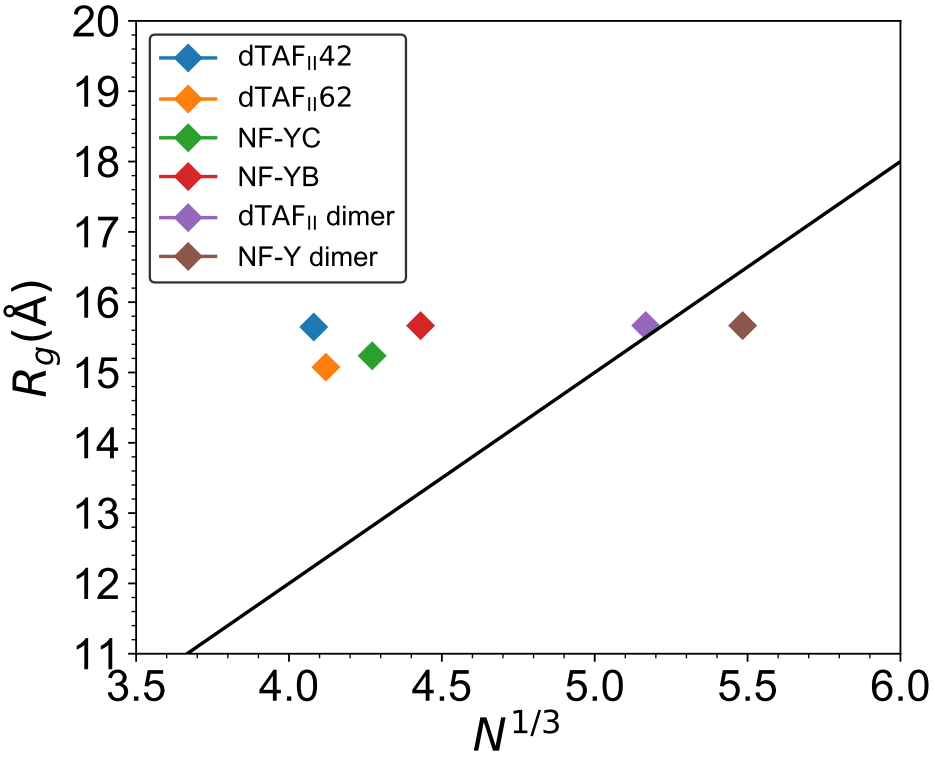


Figure C.2: **Polymer scaling fit of HFP monomers and dimers.** R_g versus the residue number N is plotted for the HFP monomers dTAF$_{II}$42, dTAF$_{II}$62, NY-FC, NF-YB, and dimers dTAF$_{II}$42/dTAF$_{II}$62, NF-YB/NF-YC with different colors respectively. The black line is the empirical relation of R_g and N for globular monomeric proteins (same as in Figure 4.6A in Chapter 4).

Bibliography

[1] Walther Flemming. Zur kenntnis der zelle und ihrer theilungserscheinungen. *Schriften Naturwiss. Vereins Schl.-Holsk*, 3(1):26, 1878.

[2] Edmund Beecher Wilson. The cell in development and heredity. Technical report, Macmillan New York, 1928.

[3] Donald E Olins and Ada L Olins. Chromatin history: our view from the bridge. *Nature Reviews Molecular Cell Biology*, 4(10):809, 2003.

[4] James D Watson and Francis HC Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[5] Ada L Olins and Donald E Olins. Spheroid chromatin units (ν bodies). *Science*, 183(4122):330–332, 1974.

[6] Christopher L Woodcock. Ultrastructure of inactive chromatin. In *Journal of Cell Biology*, volume 59, pages A368–A368. ROCKEFELLER UNIV PRESS 1114 FIRST AVE, 4TH FL, NEW YORK, NY 10021, 1973.

[7] Christopher L Woodcock, JP Safer, and JE Stanchfield. Structural repeating units in chromatin: I. evidence for their general occurrence. *Experimental Cell Research*, 97(1):101–110, 1976.

[8] TJ Richmond, JT Finch, B Rushton, D Rhodes, and A Klug. Structure of the nucleosome core particle at 7 Å resolution. *Nature*, 311(5986):532, 1984.

[9] Karolin Luger, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260, 1997.

[10] Xiao-Chen Bai, Greg McMullan, and Sjors HW Scheres. How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*, 40(1):49–57, 2015.

[11] Michael J Rust, Mark Bates, and Xiaowei Zhuang. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods*, 3(10):793–796, 2006.

[12] Feng Song, Ping Chen, Dapeng Sun, Mingzhu Wang, Liping Dong, Dan Liang, Rui-Ming Xu, Ping Zhu, and Guohong Li. Cryo-EM study of the chromatin fiber reveals a double helix twisted by tetranucleosomal units. *Science*, 344(6182):376–380, 2014.

[13] Horng D Ou, Sébastien Phan, Thomas J Deerinck, Andrea Thor, Mark H Ellisman, and Clodagh C O'Shea. ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, 357(6349):eaag0025, 2017.

[14] Bogdan Bintu, Leslie J Mateo, Jun-Han Su, Nicholas A Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N Boettiger, and Xiaowei Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413):eaau1783, 2018.

[15] JT Finch and A Klug. Solenoidal model for superstructure in chromatin. *Proceedings of the National Academy of Sciences*, 73(6):1897–1901, 1976.

[16] Rachel A Horowitz, David A Agard, John W Sedat, and Christopher L Woodcock. The three-dimensional architecture of chromatin in situ: electron tomography reveals fibers composed of a continuously variable zig-zag nucleosomal ribbon. *The Journal of Cell Biology*, 125(1):1–10, 1994.

[17] Jan Bednar, Rachel A Horowitz, Sergei A Grigoryev, Lenny M Carruthers, Jeffrey C Hansen, Abraham J Koster, and Christopher L Woodcock. Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proceedings of the National Academy of Sciences*, 95(24):14173–14178, 1998.

[18] Kazuhiro Maeshima, Saera Hihara, and Mikhail Eltsov. Chromatin structure: does the 30-nm fibre exist in vivo? *Current Opinion in Cell Biology*, 22(3):291–297, 2010.

[19] Christopher L Woodcock and Rajarshi P Ghosh. Chromatin higher-order structure and dynamics. *Cold Spring Harbor Perspectives in Biology*, page a000596, 2010.

[20] Dmitry V Fyodorov, Bing-Rui Zhou, Arthur I Skoultchi, and Yawen Bai. Emerging roles of linker histones in regulating chromatin structure and function. *Nature Reviews Molecular Cell Biology*, 19(3):192, 2018.

[21] Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

[22] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019.

[23] M Ryan Corces, Jeffrey M Granja, Shadi Shams, Bryan H Louie, Jose A Seoane, Wanding Zhou, Tiago C Silva, Clarice Groeneveld, Christopher K Wong, Seung Woo Cho, et al. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413):eaav1898, 2018.

[24] Thomas H Eickbush and Evangelos N Moudrianakis. The histone core complex: an octamer assembled by two sets of protein-protein interactions. *Biochemistry*, 17(23):4955–4964, 1978.

[25] Vassiliki Karantza, Andreas D Baxevanis, Ernesto Freire, and Evangelos N Moudrianakis. Thermodynamic studies of the core histones: ionic strength and pH dependence of H2A-H2B dimer stability. *Biochemistry*, 34(17):5988–5996, 1995.

[26] Alonso J Pardal, Filipe Fernandes-Duarte, and Andrew J Bowman. The histone chaperoning pathway: from ribosome to nucleosome. *Essays in Biochemistry*, 63(1):29–43, 2019.

[27] Gina Arents, Rufus W Burlingame, Bi-Cheng Wang, Warner E Love, and Evangelos N Moudrianakis. The nucleosomal core histone octamer at 3.1 Å resolution: a tripartite protein assembly and a left-handed superhelix. *Proceedings of the National Academy of Sciences*, 88(22):10148–10152, 1991.

[28] Gina Arents and Evangelos N Moudrianakis. The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proceedings of the National Academy of Sciences*, 92(24):11170–11174, 1995.

[29] V Ramakrishnan. Histone structure and the organization of the nucleosome. *Annual Review of Biophysics and Biomolecular Structure*, 26(1):83–112, 1997.

[30] Albert Serra-Cardona and Zhiguo Zhang. Replication-coupled nucleosome assembly in the passage of epigenetic information and cell identity. *Trends in Biochemical Sciences*, 43(2):136–148, 2018.

[31] Kyo Yamasu and Tatsuo Senshn. Conservative segregation of tetrameric units of H3 and H4 histones during nuclesome replication. *The Journal of Biochemistry*, 107(1):15–20, 1990.

[32] Paul D Kaufman, Ryuji Kobayashi, Naama Kessler, and Bruce Stillman. The p150 and p60 subunits of chromatin assemblyfactor I: A molecular link between newly synthesized histories and DNA replication. *Cell*, 81(7):1105–1114, 1995.

[33] Jessica K Tyler, Christopher R Adams, Shaw-Ree Chen, Ryuji Kobayashi, Rohinton T Kamakaka, and James T Kadonaga. The RCAF complex mediates chromatin assembly during DNA replication and repair. *Nature*, 402(6761):555–560, 1999.

[34] Laura J Benson, Yongli Gu, Tatyana Yakovleva, Kevin Tong, Courtney Barrows, Christine L Strack, Richard G Cook, Craig A Mizzen, and Anthony T Annunziato. Modifications of H3 and H4 during chromatin replication, nucleosome assembly, and histone exchange. *Journal of Biological Chemistry*, 281(14):9287–9296, 2006.

[35] Larry Louters and Roger Chalkley. Exchange of histones H1, H2A, and H2B in vivo. *Biochemistry*, 24(13):3080–3085, 1985.

[36] Young-Jun Park, Jayanth V Chodaparambil, Yunhe Bao, Steven J McBryant, and Karolin Luger. Nucleosome assembly protein 1 exchanges histone H2A-H2B dimers and assists nucleosome sliding. *Journal of Biological Chemistry*, 280(3):1817–1825, 2005.

[37] Jack A Vincent, Tracey J Kwong, and Toshio Tsukiyama. ATP-dependent chromatin remodeling shapes the DNA replication landscape. *Nature Structural & Molecular Biology*, 15(5):477–484, 2008.

[38] Tejas Yadav and Iestyn Whitehouse. Replication-coupled nucleosome assembly and positioning by ATP-dependent chromatin-remodeling enzymes. *Cell Reports*, 15(4):715–723, 2016.

[39] Curt A Davey, David F Sargent, Karolin Luger, Armin W Maeder, and Timothy J Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *Journal of Molecular Biology*, 319(5):1097–1113, 2002.

[40] Andreas D Baxevanis and David Landsman. Histone Sequence Database: a compilation of highly-conserved nucleoprotein sequences. *Nucleic Acids Research*, 24(1):245–247, 1996.

[41] Kathleen Sandman and John N Reeve. Archaeal histones and the origin of the histone fold. *Current Opinion in Microbiology*, 9(5):520–525, 2006.

[42] Andreas D Baxevanis, Gina Arents, Evangelos N Moudrianakis, and David Landsman. A variety of DNA-binding and multimeric proteins contain the histone fold motif. *Nucleic Acids Research*, 23(14):2685–2691, 1995.

[43] Vikram Alva, Moritz Ammelburg, Johannes Söding, and Andrei N Lupas. On the origin of the histone fold. *BMC Structural Biology*, 7(1):17, 2007.

[44] Vladimir N Uversky. Intrinsically disordered proteins and their environment: effects of strong denaturants, temperature, pH, counter ions, membranes,

binding partners, osmolytes, and macromolecular crowding. *The Protein Journal*, 28(7-8):305–325, 2009.

[45] Sharon Pepenella, Kevin J Murphy, and Jeffrey J Hayes. Intra-and inter-nucleosome interactions of the core histone tail domains in higher-order chromatin structure. *Chromosoma*, 123(1-2):3–13, 2014.

[46] Craig L Peterson and Marc-André Laniel. Histones and histone modifications. *Current Biology*, 14(14):R546–R551, 2004.

[47] VG Allfrey, R Faulkner, and AE Mirsky. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Sciences*, 51(5):786–794, 1964.

[48] Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381, 2011.

[49] Shelley L Berger. Histone modifications in transcriptional regulation. *Current Opinion in Genetics & Development*, 12(2):142–148, 2002.

[50] Andreas Lennartsson and Karl Ekwall. Histone modification patterns and epigenetic codes. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1790(9):863–868, 2009.

[51] Michael Shogren-Knaak, Haruhiko Ishii, Jian-Min Sun, Michael J Pazin, James R Davie, and Craig L Peterson. Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science*, 311(5762):844–847, 2006.

[52] Manuel Rodríguez-Paredes and Manel Esteller. Cancer epigenetics reaches mainstream oncology. *Nature Medicine*, 17(3):330, 2011.

[53] Mario F Fraga, Esteban Ballestar, Ana Villar-Garea, Manuel Boix-Chornet, Jesus Espada, Gunnar Schotta, Tiziana Bonaldi, Claire Haydon, Santiago Ropero, Kevin Petrie, et al. Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. *Nature Genetics*, 37(4):391–400, 2005.

[54] Sonja P Hergeth and Robert Schneider. The H1 linker histones: Multifunctional proteins beyond the nucleosomal core particle. *EMBO Reports*, 16(11):1439–1453, 2015.

[55] AV Lyubitelev, DV Nikitin, AK Shaytan, VM Studitsky, and MP Kirpichnikov. Structure and functions of linker histones. *Biochemistry (Moscow)*, 81(3):213–223, 2016.

[56] V Ramakrishnan, JT Finch, V Graziano, PL Lee, and RM Sweet. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature*, 362(6417):219, 1993.

[57] Saadi Khochbin. Histone H1 diversity: bridging regulatory signals to linker histone function. *Gene*, 271(1):1–12, 2001.

[58] Annalisa Izzo, Kinga Kamieniarz-Gdula, Fidel Ramírez, Nighat Noureen, Jop Kind, Thomas Manke, Bas van Steensel, and Robert Schneider. The genomic landscape of the somatic linker histone subtypes H1. 1 to H1. 5 in human cells. *Cell Reports*, 3(6):2142–2154, 2013.

[59] Robert T Simpson. Structure of the chromatosome, a chromatin particle containing 160 base pairs of DNA and all the histones. *Biochemistry*, 17(25):5524–5531, 1978.

[60] Marissa Vignali and Jerry L Workman. Location and function of linker histones. *Nature Structural & Molecular Biology*, 5(12):1025, 1998.

[61] Bing-Rui Zhou, Jiansheng Jiang, Hanqiao Feng, Rodolfo Ghirlando, T Sam Xiao, and Yawen Bai. Structural mechanisms of nucleosome recognition by linker histones. *Molecular Cell*, 59(4):628–638, 2015.

[62] Mehmet Ali Öztürk, Vlad Cojocaru, and Rebecca C Wade. Toward an ensemble view of chromatosome structure: A paradigm shift from one to many. *Structure*, 26(8):1050–1057, 2018.

[63] Yuhong Fan, Tatiana Nikitina, Jie Zhao, Tomara J Fleury, Riddhi Bhattacharyya, Eric E Bouhassira, Arnold Stein, Christopher L Woodcock, and Arthur I Skoultchi. Histone H1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation. *Cell*, 123(7):1199–1212, 2005.

[64] Christopher L Woodcock, Arthur I Skoultchi, and Yuhong Fan. Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosome Research*, 14(1):17–25, 2006.

[65] Christopher Wood, Ambrosius Snijders, James Williamson, Colin Reynolds, John Baldwin, and Mark Dickman. Post-translational modifications of the linker histone variants and their association with cell mechanisms. *The FEBS Journal*, 276(14):3685–3697, 2009.

[66] Jan Bednar, Isabel Garcia-Saez, Ramachandran Boopathi, Amber R Cutter, Gabor Papai, Anna Reymer, Sajad H Syed, Imtiaz Nisar Lone, Ognyan Tonchev, Corinne Crucifix, Hervé Menoni, Christophe Papin, Dimitrios A Skoufias, Hitoshi Kurumizaka, Richard Lavery, Ali Hamiche, Jeffrey J Hayes, Patrick Schultz, Dimitar Angelov, Carlo Petosa, and Stefan Dimitrov. Structure and dynamics of a 197 bp nucleosome in complex with linker histone H1. *Molecular Cell*, 66(3):384–397, 2017.

[67] Benjamin Webb and Andrej Sali. Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics*, 54(1):5.6.1–5.6.37, 2016.

[68] Bing-Rui Zhou, Hanqiao Feng, Hidenori Kato, Liang Dai, Yuedong Yang, Yaoqi Zhou, and Yawen Bai. Structural insights into the histone H1-nucleosome complex. *Proceedings of the National Academy of Sciences*, 110(48):19390–19395, 2013.

[69] Isabel Garcia-Saez, Hervé Menoni, Ramachandran Boopathi, Manu S Shukla, Lama Soueidan, Marjolaine Noirclerc-Savoye, Aline Le Roy, Dimitrios A Skoufias, Jan Bednar, Ali Hamiche, Dimitar Angelov, Carlo Petosa, and Stefan Dimitrov. Structure of an H1-bound 6-nucleosome array reveals an untwisted two-start chromatin fiber conformation. *Molecular Cell*, 72(5):902–915, 2018.

[70] He Fang, Sijie Wei, Tae-Hee Lee, and Jeffrey J Hayes. Chromatin structure-dependent conformations of the H1 CTD. *Nucleic Acids Research*, 44(19):9131–9141, 2016.

[71] Luda S Shlyakhtenko, Alexander Y Lushnikov, and Yuri L Lyubchenko. Dynamics of nucleosomes revealed by time-lapse atomic force microscopy. *Biochemistry*, 48(33):7842–7848, 2009.

[72] Nynke L Van Berkum, Erez Lieberman-Aiden, Louise Williams, Maxim Imakaev, Andreas Gnirke, Leonid A Mirny, Job Dekker, and Eric S Lander. Hi-C: a method to study the three-dimensional architecture of genomes. *JoVE (Journal of Visualized Experiments)*, (39):e1869, 2010.

[73] Nikolay Korolev, Yanping Fan, Alexander P Lyubartsev, and Lars Nordenskiöld. Modelling chromatin structure and dynamics: status and prospects. *Current Opinion in Structural Biology*, 22(2):151–159, 2012.

[74] Gungor Ozer, Antoni Luque, and Tamar Schlick. The chromatin fiber: multiscale problems and approaches. *Current Opinion in Structural Biology*, 31:124–139, 2015.

[75] Jaewoon Jung, Wataru Nishima, Marcus Daniels, Gavin Bascom, Chigusa Kobayashi, Adetokunbo Adedoyin, Michael Wall, Anna Lappala, Dominic Phillips, William Fischer, Chang-Shung Tung, Tamar Schlick, Yuji Sugita, and Karissa Y Sanbonmatsu. Scaling molecular dynamics beyond 100,000 processor cores for large-scale biophysical simulations. *Journal of Computational Chemistry*, 40(21):1919–1930, 2019.

[76] Aram Davtyan, Nicholas P Schafer, Weihua Zheng, Cecilia Clementi, Peter G Wolynes, and Garegin A Papoian. AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *The Journal of Physical Chemistry B*, 116(29):8494–8503, 2012.

[77] Mark S Friedrichs and Peter G Wolynes. Toward protein tertiary structure recognition by means of associative memory hamiltonians. *Science*, 246(4928):371–373, 1989.

[78] Masaki Sasai and Peter G Wolynes. Molecular theory of associative memory Hamiltonian models of protein folding. *Physical Review Letters*, 65(21):2740, 1990.

[79] Corey Hardin, Michael P Eastwood, Zaida Luthey-Schulten, and Peter G Wolynes. Associative memory Hamiltonians for structure prediction without homology: alpha-helical proteins. *Proceedings of the National Academy of Sciences*, 97(26):14235–14240, 2000.

[80] Garegin A Papoian, Johan Ulander, Michael P Eastwood, Zaida Luthey-Schulten, and Peter G Wolynes. Water in protein structure prediction. *Proceedings of the National Academy of Sciences*, 101(10):3352–3357, 2004.

[81] José N Onuchic, Zaida Luthey-Schulten, and Peter G Wolynes. Theory of protein folding: the energy landscape perspective. *Annual Review of Physical Chemistry*, 48(1):545–600, 1997.

[82] Gopalasamudram Narayana Ramachandran. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7:95–99, 1963.

[83] Cyrus Levinthal. Are there pathways for protein folding? *Journal de Chimie Physique*, 65:44–45, 1968.

[84] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[85] Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.

[86] Peter E Leopold, Mauricio Montal, and José N Onuchic. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proceedings of the National Academy of Sciences*, 89(18):8721–8725, 1992.

[87] José N Onuchic, Peter G Wolynes, Z Luthey-Schulten, and Nicholas D Socci. Toward an outline of the topography of a realistic protein-folding funnel. *Proceedings of the National Academy of Sciences*, 92(8):3626–3630, 1995.

[88] Ken A Dill and Hue Sun Chan. From levinthal to pathways to funnels. *Nature Structural Biology*, 4(1):10, 1997.

[89] Richard A Goldstein, Zaida A Luthey-Schulten, and Peter G Wolynes. Optimal protein-folding codes from spin-glass theory. *Proceedings of the National Academy of Sciences*, 89(11):4918–4922, 1992.

[90] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[91] Garegin A Papoian and Peter G Wolynes. *AWSEM-MD: From Neural Networks to Protein Structure Prediction and Functional Dynamics of Complex Biomolecular Assemblies*, chapter 4, pages 121–189. CRC Press, Taylor & Francis Group, Boca Raton, FL, 2017.

[92] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995.

[93] Faruck Morcos, Nicholas P Schafer, Ryan R Cheng, José N Onuchic, and Peter G Wolynes. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proceedings of the National Academy of Sciences*, 111(34):12408–12413, 2014.

[94] Mingchen Chen, Xingcheng Lin, Weihua Zheng, José N Onuchic, and Peter G Wolynes. Protein folding and structure prediction from the ground up: The atomistic associative memory, water mediated, structure and energy model. *The Journal of Physical Chemistry B*, 120(33):8557–8565, 2016.

[95] Andrew P Latham and Bin Zhang. Improving coarse-grained protein force fields with small-angle X-ray scattering data. *The Journal of Physical Chemistry B*, 123(5):1026–1034, 2019.

[96] Xingcheng Lin, Nicholas P Schafer, Wei Lu, Shikai Jin, Xun Chen, Mingchen Chen, José N Onuchic, and Peter G Wolynes. Forging tools for refining predicted protein structures. *Proceedings of the National Academy of Sciences*, 116(19):9400–9409, 2019.

[97] Bobby L Kim, Nicholas P Schafer, and Peter G Wolynes. Predictive energy landscapes for folding α-helical transmembrane proteins. *Proceedings of the National Academy of Sciences*, 111(30):11031–11036, 2014.

[98] Weihua Zheng, Nicholas P Schafer, Aram Davtyan, Garegin A Papoian, and Peter G Wolynes. Predictive energy landscapes for protein–protein association. *Proceedings of the National Academy of Sciences*, 109(47):19244–19249, 2012.

[99] Weihua Zheng, Nicholas P Schafer, and Peter G Wolynes. Frustration in the energy landscapes of multidomain protein misfolding. *Proceedings of the National Academy of Sciences*, 110(5):1680–1685, 2013.

[100] Weihua Zheng, Min-Yeh Tsai, Mingchen Chen, and Peter G Wolynes. Exploring the aggregation free energy landscape of the amyloid-β protein (1–40). *Proceedings of the National Academy of Sciences*, 113(42):11835–11840, 2016.

[101] Davit A Potoyan, Weihua Zheng, Elizabeth A Komives, and Peter G Wolynes. Molecular stripping in the NF-κB/IκB/DNA genetic regulatory network. *Proceedings of the National Academy of Sciences*, 113(1):110–115, 2016.

[102] David Winogradoff, Haiqing Zhao, Yamini Dalal, and Garegin A Papoian. Shearing of the CENP-A dimerization interface mediates plasticity in the octameric centromeric nucleosome. *Scientific Reports*, 5:17038, 2015.

[103] Alexey K Shaytan, Grigoriy A Armeev, Alexander Goncearenco, Victor B Zhurkin, David Landsman, and Anna R Panchenko. Coupling between histone conformations and DNA geometry in nucleosomes on a microsecond timescale: atomistic insights into nucleosome functions. *Journal of Molecular Biology*, 428(1):221–237, 2016.

[104] Gero Wedemann and Jörg Langowski. Computer simulation of the 30-nanometer chromatin fiber. *Biophysical Journal*, 82(6):2847–2859, 2002.

[105] Nikolay Korolev, Alexander P Lyubartsev, and Lars Nordenskiöld. Computer modeling demonstrates that electrostatic attraction of nucleosomal dna is mediated by histone tails. *Biophysical Journal*, 90(12):4305–4316, 2006.

[106] Haiqing Zhao, David Winogradoff, Minh Bui, Yamini Dalal, and Garegin A Papoian. Promiscuous histone mis-assembly is actively prevented by chaperones. *Journal of the American Chemical Society*, 138(40):13207–13218, 2016.

[107] Haiqing Zhao, David Winogradoff, Yamini Dalal, and Garegin A Papoian. The oligomerization landscape of histones. *Biophysical Journal*, 116(10):1845–1855, 2019.

[108] Alexey Savelyev and Garegin A Papoian. Chemically accurate coarse graining of double-stranded DNA. *Proceedings of the National Academy of Sciences*, 107(47):20340–20345, 2010.

[109] Daniel M Hinckley, Gordon S Freeman, Jonathan K Whitmer, and Juan J De Pablo. An experimentally-informed coarse-grained 3-site-per-nucleotide model of DNA: Structure, thermodynamics, and dynamics of hybridization. *The Journal of Chemical Physics*, 139(14):1–16, 2013.

[110] Min-Yeh Tsai, Bin Zhang, Weihua Zheng, and Peter G Wolynes. Molecular mechanism of facilitated dissociation of Fis protein from DNA. *Journal of the American Chemical Society*, 138(41):13497–13500, 2016.

[111] Bin Zhang, Weihua Zheng, Garegin A Papoian, and Peter G Wolynes. Exploring the free energy landscape of nucleosomes. *Journal of the American Chemical Society*, 138(26):8126–8133, 2016.

[112] Elaine M Dunleavy, Daniele Roche, Hideaki Tagami, Nicolas Lacoste, Dominique Ray-Gallet, Yusuke Nakamura, Yataro Daigo, Yoshihiro Nakatani, and Geneviève Almouzni-Pettinotti. HJURP is a cell-cycle-dependent maintenance and deposition factor of CENP-A at centromeres. *Cell*, 137(3):485–497, 2009.

[113] Thuy TM Ngo, Qiucen Zhang, Ruobo Zhou, Jaya G Yodh, and Taekjip Ha. Asymmetric unwrapping of nucleosomes under tension directed by DNA local flexibility. *Cell*, 160(6):1135–1144, 2015.

[114] Ty C Voss and Gordon L Hager. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics*, 15(2):69–81, 2014.

[115] A Keith Dunker, J David Lawson, Celeste J Brown, Ryan M Williams, Pedro Romero, Jeong S Oh, Christopher J Oldfield, Andrew M Campen, Catherine M Ratliff, Kerry W Hipps, et al. Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, 19(1):26–59, 2001.

[116] Hao Wu, Peter G Wolynes, and Garegin A Papoian. AWSEM-IDP: A coarse-grained force field for intrinsically disordered proteins. *The Journal of Physical Chemistry B*, 122(49):11115–11125, 2018.

[117] Vladimir N Uversky and A Keith Dunker. Understanding protein non-folding. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1804(6):1231–1264, 2010.

[118] Albert H Mao, Scott L Crick, Andreas Vitalis, Caitlin L Chicoine, and Rohit V Pappu. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences*, 107(18):8183–8188, 2010.

[119] M Madan Babu, Richard W Kriwacki, and Rohit V Pappu. Versatility from protein disorder. *Science*, 337(6101):1460–1461, 2012.

[120] Robin Van Der Lee, Marija Buljan, Benjamin Lang, Robert J Weatheritt, Gary W Daughdrill, A Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T Jones, et al. Classification of intrinsically disordered regions and proteins. *Chemical Reviews*, 114(13):6589–6631, 2014.

[121] Johnny Habchi, Peter Tompa, Sonia Longhi, and Vladimir N Uversky. Introducing protein intrinsic disorder. *Chemical Reviews*, 114(13):6561–6588, 2014.

[122] Garegin A Papoian. Proteins with weakly funneled energy landscapes challenge the classical structure–function paradigm. *Proceedings of the National Academy of Sciences*, 105(38):14237–14238, 2008.

[123] Garegin A Papoian and Peter G Wolynes. The physics and bioinformatics of binding and folding—an energy landscape perspective. *Biopolymers: Original Research on Biomolecules*, 68(3):333–349, 2003.

[124] Peter E Wright and H Jane Dyson. Linking folding and binding. *Current Opinion in Structural Biology*, 19(1):31–38, 2009.

[125] Peter Tompa. Intrinsically disordered proteins: a 10-year recap. *Trends in Biochemical Sciences*, 37(12):509–516, 2012.

[126] Pau Bernado and Dmitri I Svergun. Structural analysis of intrinsically disordered proteins by small-angle x-ray scattering. *Molecular Biosystems*, 8(1):151–167, 2012.

[127] Malene Ringkjøbing Jensen, Rob WH Ruigrok, and Martin Blackledge. Describing intrinsically disordered proteins at atomic resolution by nmr. *Current Opinion in Structural Biology*, 23(3):426–435, 2013.

[128] Christopher M Baker and Robert B Best. Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(3):182–198, 2014.

[129] Christopher K Materese, Alexey Savelyev, and Garegin A Papoian. Counterion atmosphere and hydration patterns near a nucleosome core particle. *Journal of the American Chemical Society*, 131(41):15005–15013, 2009.

[130] Davit A Potoyan and Garegin A Papoian. Energy landscape analyses of disordered histone tails reveal special organization of their conformational dynamics. *Journal of the American Chemical Society*, 133(19):7405–7415, 2011.

[131] Davit A Potoyan and Garegin A Papoian. Regulation of the H4 tail binding and folding landscapes via Lys-16 acetylation. *Proceedings of the National Academy of Sciences*, 109(44):17857–17862, 2012.

[132] David Winogradoff, Ignacia Echeverria, Davit A Potoyan, and Garegin A Papoian. The acetylation landscape of the H4 histone tail: disentangling the interplay between the specific and cumulative effects. *Journal of the American Chemical Society*, 137(19):6245–6253, 2015.

[133] Robert Schneider, Jie-rong Huang, Mingxi Yao, Guillaume Communie, Valéry Ozenne, Luca Mollica, Loïc Salmon, Malene Ringkjøbing Jensen, and Martin Blackledge. Towards a robust description of intrinsic protein disorder using nuclear magnetic resonance spectroscopy. *Molecular BioSystems*, 8(1):58–68, 2012.

[134] Virginia M Burger, Thomas Gurry, and Collin M Stultz. Intrinsically disordered proteins: where computation meets experiment. *Polymers*, 6(10):2684–2719, 2014.

[135] Asmit Bhowmick, David H Brookes, Shane R Yost, H Jane Dyson, Julie D Forman-Kay, Daniel Gunter, Martin Head-Gordon, Gregory L Hura, Vijay S Pande, David E Wemmer, et al. Finding our way in the dark proteome. *Journal of the American Chemical Society*, 138(31):9730–9742, 2016.

[136] Wei Wang, Wei Ye, Cheng Jiang, Ray Luo, and Hai-Feng Chen. New force field on modeling intrinsically disordered proteins. *Chemical Biology & Drug Design*, 84(3):253–269, 2014.

[137] Dong Song, Ray Luo, and Hai-Feng Chen. The IDP-specific force field ff14IDPSFF improves the conformer sampling of intrinsically disordered proteins. *Journal of Chemical Information and Modeling*, 57(5):1166–1178, 2017.

[138] Paul Robustelli, Stefano Piana, and David E Shaw. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences*, 115(21):E4758–E4766, 2018.

[139] Austin Huang and Collin M Stultz. The effect of a ΔK280 mutation on the unfolded state of a microtubule-binding repeat in tau. *PLoS Computational Biology*, 4(8):e1000155, 2008.

[140] Joseph A Marsh and Julie D Forman-Kay. Ensemble modeling of protein disordered states: experimental restraint contributions and validation. *Proteins: Structure, Function, and Bioinformatics*, 80(2):556–572, 2012.

[141] Robert B Best, Wenwei Zheng, and Jeetain Mittal. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *Journal of Chemical Theory and Computation*, 10(11):5113–5124, 2014.

[142] Joao Henriques, Carolina Cragnell, and Marie Skepö. Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *Journal of Chemical Theory and Computation*, 11(7):3420–3431, 2015.

[143] Sarah Rauscher, Vytautas Gapsys, Michal J Gajda, Markus Zweckstetter, Bert L de Groot, and Helmut Grubmüller. Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. *Journal of Chemical Theory and Computation*, 11(11):5513–5524, 2015.

[144] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L de Groot, Helmut Grubmüller, and Alexander D MacKerell Jr. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1):71, 2017.

[145] Sushant Kumar, Scott A Showalter, and William G Noid. Native-based simulations of the binding interaction between RAP74 and the disordered FCP1 peptide. *The Journal of Physical Chemistry B*, 117(11):3074–3085, 2013.

[146] Michael Knott and Robert B Best. Discriminating binding mechanisms of an intrinsically disordered protein via a multi-state coarse-grained model. *The Journal of Chemical Physics*, 140(17):05B603_1, 2014.

[147] Mateusz Kurcinski, Andrzej Kolinski, and Sebastian Kmiecik. Mechanism of folding and binding of an intrinsically disordered protein as revealed by *ab initio* simulations. *Journal of Chemical Theory and Computation*, 10(6):2224–2231, 2014.

[148] Agustí Emperador, Pedro Sfriso, Marcos Ariel Villarreal, Josep Lluis Gelpí, and Modesto Orozco. PACSAB: Coarse-grained force field for the study of protein–protein interactions and conformational sampling in multiprotein systems. *Journal of Chemical Theory and Computation*, 11(12):5929–5938, 2015.

[149] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical Reviews*, 116(14):7898–7936, 2016.

[150] Kuo Hao Lee and Jianhan Chen. Multiscale enhanced sampling of intrinsically disordered protein conformations. *Journal of Computational Chemistry*, 37(6):550–557, 2016.

[151] Zaida Luthey-Schulten, Benjamin E Ramirez, and Peter G Wolynes. Helix-coil, liquid crystal, and spin glass transitions of a collapsed heteropolymer. *The Journal of Physical Chemistry*, 99(7):2177–2185, 1995.

[152] Weihua Zheng, Nicholas P Schafer, and Peter G Wolynes. Free energy landscapes for initiation and branching of protein aggregation. *Proceedings of the National Academy of Sciences*, 110(51):20515–20520, 2013.

[153] Ha H Truong, Bobby L Kim, Nicholas P Schafer, and Peter G Wolynes. Predictive energy landscapes for folding membrane protein assemblies. *The Journal of Chemical Physics*, 143(24):243101, 2015.

[154] Ki-Jeong Kwac and Peter G Wolynes. Protein structure prediction using an associated memory hamiltonian and all-atom molecular dynamics simulations. *Bulletin of the Korean Chemical Society*, 29(11):2172–2182, 2008.

[155] Nobuhiro Gō. The consistency principle in protein structure and pathways of folding. *Advances in Biophysics*, 18:149–164, 1984.

[156] Joseph D Bryngelson and Peter G Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences*, 84(21):7524–7528, 1987.

[157] Jeffery G Saven and Peter G Wolynes. Local conformational signals and the statistical thermodynamics of collapsed helical proteins. *Journal of molecular biology*, 257(1):199–216, 1996.

[158] Mingchen Chen, Xingcheng Lin, Wei Lu, José N Onuchic, and Peter G Wolynes. Protein folding and structure prediction from the ground up II: AAWSEM for α/β proteins. *The Journal of Physical Chemistry B*, 121(15):3473–3482, 2016.

[159] Jane R Allison, Peter Varnai, Christopher M Dobson, and Michele Vendruscolo. Determination of the free energy landscape of α-synuclein using spin label nuclear magnetic resonance measurements. *Journal of the American Chemical Society*, 131(51):18314–18326, 2009.

[160] Brian D Strahl and C David Allis. The language of covalent histone modifications. *Nature*, 403(6765):41, 2000.

[161] Thomas Jenuwein and C David Allis. Translating the histone code. *Science*, 293(5532):1074–1080, 2001.

[162] Finbarr Hayes and Laurence Van Melderen. Toxins-antitoxins: diversity, evolution and function. *Critical Reviews in Biochemistry and Molecular Biology*, 46(5):386–408, 2011.

[163] Régis Hallez, Damien Geeraerts, Yann Sterckx, Natacha Mine, Remy Loris, and Laurence Van Melderen. New toxins homologous to ParE belonging to three-component toxin–antitoxin systems in Escherichia coli O157: H7. *Molecular Microbiology*, 76(3):719–732, 2010.

[164] Andrea Amadei, Marc A Ceruso, and Alfredo Di Nola. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, 36(4):419–424, 1999.

[165] Dmitrij Frishman and Patrick Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4):566–579, 1995.

[166] Yann GJ Sterckx, Alexander N Volkov, Wim F Vranken, Jaka Kragelj, Malene Ringkjøbing Jensen, Lieven Buts, Abel Garcia-Pino, Thomas Jové, Laurence Van Melderen, Martin Blackledge, et al. Small-angle X-ray scattering- and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin PaaA2. *Structure*, 22(6):854–865, 2014.

[167] Eldon L Ulrich, Hideo Akutsu, Jurgen F Doreleijers, Yoko Harano, Yannis E Ioannidis, Jundong Lin, Miron Livny, Steve Mading, Dimitri Maziuk, Zachary Miller, et al. BioMagResBank. *Nucleic Acids Research*, 36(suppl_1):D402–D408, 2007.

[168] Carlo Camilloni, Alfonso De Simone, Wim F Vranken, and Michele Vendruscolo. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry*, 51(11):2224–2231, 2012.

[169] Dmitri Svergun, Claudio Barberato, and Michel HJ Koch. CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of Applied Crystallography*, 28(6):768–773, 1995.

[170] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1-2):141–151, 1999.

[171] Robert B Best and Gerhard Hummer. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *The Journal of Physical Chemistry B*, 113(26):9004–9015, 2009.

[172] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[173] Hiroo Kenzaki and Shoji Takada. Partial unwrapping and histone tail dynamics in nucleosome revealed by coarse-grained molecular simulations. *PLoS computational biology*, 11(8):e1004443, 2015.

[174] Nicholas P Schafer, Bobby L Kim, Weihua Zheng, and Peter G Wolynes. Learning to fold proteins using energy landscape theory. *Israel Journal of Chemistry*, 54(8-9):1311–1337, 2014.

[175] Yann G-J Sterckx, Thomas Jové, Alexander V Shkumatov, Abel Garcia-Pino, Lieselotte Geerts, Maia De Kerpel, Jurij Lah, Henri De Greve, Laurence Van Melderen, and Remy Loris. A unique hetero-hexadecameric architecture displayed by the Escherichia coli O157 PaaA2–ParE2 antitoxin–toxin complex. *Journal of Molecular Biology*, 428(8):1589–1603, 2016.

[176] Mihaly Varadi, Simone Kosol, Pierre Lebrun, Erica Valentini, Martin Blackledge, A Keith Dunker, Isabella C Felli, Julie D Forman-Kay, Richard W Kriwacki, Roberta Pierattelli, et al. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Research*, 42(D1):D326–D335, 2013.

[177] Roger D Kornberg. Chromatin structure: A repeating unit of histones and DNA. *Science*, 184(4139):868–871, 1974.

[178] Hideaki Takata, Tomo Hanafusa, Toshiaki Mori, Mari Shimura, Yutaka Iida, Kenichi Ishikawa, Kenichi Yoshikawa, Yuko Yoshikawa, and Kazuhiro Maeshima. Chromatin compaction protects genomic DNA from radiation damage. *PloS One*, 8(10), 2013.

[179] Tom Misteli, Akash Gunjan, Robert Hock, Michael Bustin, and David T Brown. Dynamic binding of histone H1 to chromatin in living cells. *Nature*, 408(6814):877–881, 2000.

[180] Frédéric Catez, David T Brown, Tom Misteli, and Michael Bustin. Competition between histone H1 and HMGN proteins for chromatin binding sites. *EMBO Reports*, 3(8):760–766, 2002.

[181] Frédéric Catez, Huan Yang, Kevin J Tracey, Raymond Reeves, Tom Misteli, and Michael Bustin. Network of dynamic interactions between histone H1 and high-mobility-group proteins in chromatin. *Molecular and Cellular Biology*, 24(10):4321–4328, 2004.

[182] David T Brown, Tina Izard, and Tom Misteli. Mapping the interaction surface of linker histone H1.0 with the nucleosome of native chromatin in vivo. *Nature Structural & Molecular Biology*, 13(3):250, 2006.

[183] Annalisa Izzo, Kinga Kamieniarz, and Robert Schneider. The histone H1 family: Specific members, specific functions? *Biological Chemistry*, 389(4):333–343, 2008.

[184] J Allan, PG Hartman, C Crane-Robinson, and FX Aviles. The structure of histone H1 and its location in chromatin. *Nature*, 288(5792):675, 1980.

[185] Michael Bustin, Frédéric Catez, and Jae-Hwan Lim. The dynamics of histone H1 function in chromatin. *Molecular Cell*, 17(5):617–620, 2005.

[186] Frédéric Catez, Tetsuya Ueda, and Michael Bustin. Determinants of histone H1 mobility and chromatin binding in living cells. *Nature Structural & Molecular Biology*, 13(4):305, 2006.

[187] Jan Bednar, Ali Hamiche, and Stefan Dimitrov. H1–nucleosome interactions and their functional implications. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1859(3):436–443, 2016.

[188] Thomas W Flanagan and David T Brown. Molecular dynamics of histone H1. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1859(3):468–475, 2016.

[189] Alicia Roque, Inma Ponte, and Pedro Suau. Interplay between histone H1 structure and function. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1859(3):444–454, 2016.

[190] Andrew Routh, Sara Sandin, and Daniela Rhodes. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proceedings of the National Academy of Sciences*, 105(26):8872–8877, 2008.

[191] Bing-Rui Zhou, Hanqiao Feng, Rodolfo Ghirlando, Shipeng Li, Charles D Schwieters, and Yawen Bai. A small number of residues can determine if linker histones are bound on or off dyad in the chromatosome. *Journal of Molecular Biology*, 428(20):3948–3959, 2016.

[192] Georgi V Pachov, Razif R Gabdoulline, and Rebecca C Wade. On the structure and dynamics of the complex of the nucleosome and the linker histone. *Nucleic Acids Research*, 39(12):5255–5263, 2011.

[193] Mehmet Ali Öztürk, Georgi V Pachov, Rebecca C Wade, and Vlad Cojocaru. Conformational selection and dynamic adaptation upon linker histone binding to the nucleosome. *Nucleic Acids Research*, 44(14):6599–6613, 2016.

[194] Mehmet Ali Öztürk, Vlad Cojocaru, and Rebecca C Wade. Dependence of chromatosome structure on linker histone sequence and posttranslational modification. *Biophysical Journal*, 114(10):2363–2375, 2018.

[195] Ognjen Perišić and Tamar Schlick. Dependence of the linker histone and chromatin condensation on the nucleosome environment. *The Journal of Physical Chemistry B*, 121(33):7823–7832, 2017.

[196] Ognjen Perišić, Stephanie Portillo-Ledesma, and Tamar Schlick. Sensitive effect of linker histone binding mode and subtype on chromatin condensation. *Nucleic Acids Research*, 47(10):4948–4957, 2019.

[197] Cristina Morales Torres, Alva Biran, Matthew J Burney, Harshil Patel, Tristan Henser-Brownhill, Ayelet-Hashahar Shapira Cohen, Yilong Li, Rotem Ben-Hamo, Emma Nye, Bradley Spencer-Dene, Probir Chakravarty, Sol Efroni, Nik Matthews, Tom Misteli, Eran Meshorer, and Paola Scaffidi. The linker histone H1. 0 generates epigenetic and functional intratumor heterogeneity. *Science*, 353(6307):aaf1644, 2016.

[198] Miho Shimada, Wei-Yi Chen, Tomoyoshi Nakadai, Takashi Onikubo, Mohamed Guermah, Daniela Rhodes, and Robert G Roeder. Gene-specific H1 eviction through a transcriptional activator $\rightarrow$ p300 $\rightarrow$ NAP1 $\rightarrow$ H1 pathway. *Molecular Cell*, 74(2):268–283, 2019.

[199] Alessandro Borgia, Madeleine B Borgia, Katrine Bugge, Vera M Kissling, Pétur O Heidarsson, Catarina B Fernandes, Andrea Sottini, Andrea Soranno, Karin J Buholzer, Daniel Nettels, Birthe B Kragelund, Robert B Best, and Benjamin Schuler. Extreme disorder in an ultrahigh-affinity protein complex. *Nature*, 555(7694):61, 2018.

[200] Abigail L Turner, Matthew Watson, Oscar G Wilkins, Laura Cato, Andrew Travers, Jean O Thomas, and Katherine Stott. Highly disordered histone H1-DNA model complexes and their condensates. *Proceedings of the National Academy of Sciences*, 115(47):11964–11969, 2018.

[201] Eric B Gibbs and Richard W Kriwacki. Linker histones as liquid-like glue for chromatin. *Proceedings of the National Academy of Sciences*, 115(47):11868–11870, 2018.

[202] Michael J Hendzel, Melody A Lever, Ellen Crawford, and John PH Th'ng. The C-terminal domain is the primary determinant of histone H1 binding to chromatin in vivo. *Journal of Biological Chemistry*, 279(17):20028–20034, 2004.

[203] He Fang, David J Clark, and Jeffrey J Hayes. DNA and nucleosomes direct distinct folding of a linker histone H1 C-terminal domain. *Nucleic Acids Research*, 40(4):1475–1484, 2011.

[204] Akshay Sridhar, Stephen E Farr, Guillem Portella, Tamar Schlick, Modesto Orozco, and Rosana Collepardo-Guevara. Emergence of chromatin hierarchical loops from protein disorder and nucleosome asymmetry. *Proceedings of the National Academy of Sciences*, 117(13):7216–7224, 2020.

[205] Akshay Sridhar, Modesto Orozco, and Rosana Collepardo-Guevara. Protein disorder-to-order transition enhances the nucleosome-binding affinity of H1. *Nucleic Acids Research*, 48(10):5318–5331, 2020.

[206] Davit A Potoyan, Carlos Bueno, Weihua Zheng, Elizabeth A Komives, and Peter G Wolynes. Resolving the NFκB heterodimer binding paradox: Strain and frustration guide the binding of dimeric transcription factors. *Journal of the American Chemical Society*, 139(51):18558–18566, 2017.

[207] Min-Yeh Tsai, Weihua Zheng, Mingchen Chen, and Peter G Wolynes. Multiple binding configurations of Fis protein pairs on DNA: Facilitated dissociation versus cooperative dissociation. *Journal of the American Chemical Society*, 2019.

[208] Xingcheng Lin, Susmita Roy, Mohit Kumar Jolly, Federico Bocci, Nicholas P Schafer, Min-Yeh Tsai, Yihong Chen, Yanan He, Alexander Grishaev, Keith Weninger, John Orban, Prakash Kulkarni, Govindan Rangarajan, Herbert Levine, and José N Onuchic. PAGE4 and conformational switching: Insights from molecular dynamics simulations and implications for prostate cancer. *Journal of Molecular Biology*, 430(16):2422–2438, 2018.

[209] Dustin C Woods and Jeff Wereszczynski. Elucidating the influence of linker histone variants on chromatosome dynamics and energetics. *Nucleic Acids Research*, 2020. gkaa121.

[210] Maria A Christophorou, Gonçalo Castelo-Branco, Richard P Halley-Stott, Clara Slade Oliveira, Remco Loos, Aliaksandra Radzisheuskaya, Kerri A Mowen, Paul Bertone, José CR Silva, Magdalena Zernicka-Goetz, et al. Citrullination regulates pluripotency and histone H1 binding to chromatin. *Nature*, 507(7490):104, 2014.

[211] Kyunghwan Kim, Kwang Won Jeong, Hyunjung Kim, Jongkyu Choi, Wange Lu, Michael R Stallcup, and Woojin An. Functional interplay between p53 acetylation and H1. 2 phosphorylation in p53-regulated transcription. *Oncogene*, 31(39):4290, 2012.

[212] Heribert Talasz, Bettina Sarg, and Herbert H Lindner. Site-specifically phosphorylated forms of H1. 5 and H1. 2 localized at distinct regions of the nucleus are related to different processes during the cell cycle. *Chromosoma*, 118(6):693–709, 2009.

[213] Alicia Roque, Inma Ponte, Jose Luis R Arrondo, and Pedro Suau. Phosphorylation of the carboxy-terminal domain of histone H1: effects on secondary structure and DNA condensation. *Nucleic Acids Research*, 36(14):4719–4726, 2008.

[214] Rita Lopez, Bettina Sarg, Herbert Lindner, Salvador Bartolomé, Inma Ponte, Pedro Suau, and Alicia Roque. Linker histone partial phosphorylation: effects on secondary structure and chromatin condensation. *Nucleic Acids Research*, 43(9):4463–4476, 2015.

[215] Arnold Stein and Minou Bina. A model chromatin assembly system: factors affecting nucleosome spacing. *Journal of Molecular Biology*, 178(2):341–363, 1984.

[216] Akash Gunjan, Barbara T Alexander, Donald B Sittman, and David T Brown. Effects of H1 histone variant overexpression on chromatin structure. *Journal of Biological Chemistry*, 274(53):37950–37956, 1999.

[217] Dean R Hewish and Leigh A Burgoyne. Chromatin sub-structure. the digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. *Biochemical and Biophysical Research Communications*, 52(2):504–510, 1973.

[218] Kensal E Van Holde. *Chromatin*. Springer-Verlag New York, 1989.

[219] Paul B Talbert, Kami Ahmad, Geneviève Almouzni, Juan Ausió, Frederic Berger, Prem L Bhalla, William M Bonner, W Zacheus Cande, Brian P Chadwick, Simon WL Chan, et al. A unified phylogeny-based nomenclature for histone variants. *Epigenetics & Chromatin*, 5(1):7, 2012.

[220] Atsushi Miyagi, Toshio Ando, and Yuri L Lyubchenko. Dynamics of nucleosomes assessed with time-lapse high-speed atomic force microscopy. *Biochemistry*, 50(37):7901–7908, 2011.

[221] Yuri L Lyubchenko. Nanoscale nucleosome dynamics assessed with time-lapse AFM. *Biophysical Reviews*, 6(2):181–190, 2014.

[222] Rifka Vlijm, Mina Lee, Jan Lipfert, Alexandra Lusser, Cees Dekker, and Nynke H Dekker. Nucleosome assembly dynamics involve spontaneous fluctuations in the handedness of tetrasomes. *Cell Reports*, 10(2):216–225, 2015.

[223] Sung Hyun Kim, Rifka Vlijm, Jaco van der Torre, Yamini Dalal, and Cees Dekker. CENP-A and H3 nucleosomes display a similar stability to force-mediated disassembly. *PloS One*, 11(11):e0165078, 2016.

[224] Robert D Phair, Paola Scaffidi, Cem Elbi, Jaromíra Vecerová, Anup Dey, Keiko Ozato, David T Brown, Gordon Hager, Michael Bustin, and Tom Misteli. Global nature of dynamic protein-chromatin interactions in vivo: three-dimensional genome scanning and dynamic interaction networks of chromatin proteins. *Molecular and Cellular Biology*, 24(14):6393–6402, 2004.

[225] Danielle Vermaak, Oliver C Steinbach, Stephan Dimitrov, Ralph AW Rupp, and Alan P Wolffe. The globular domain of histone H1 is sufficient to direct specific gene repression in early Xenopus embryos. *Current Biology*, 8(9):533–S2, 1998.

[226] Danielle Vermaak and Alan P Wolffe. Chromatin and chromosomal controls in development. *Developmental Genetics*, 22(1):1–6, 1998.

[227] Steven Henikoff, Takehito Furuyama, and Kami Ahmad. Histone variants, nucleosome assembly and epigenetic inheritance. *Trends in Genetics*, 20(7):320–326, 2004.

[228] V Ramakrishnan. The histone fold: evolutionary questions. *Proceedings of the National Academy of Sciences*, 92(25):11328, 1995.

[229] Steven Henikoff and M Mitchell Smith. Histone variants and epigenetics. *Cold Spring Harbor perspectives in biology*, 7(1):a019364, 2015.

[230] Yamini Dalal, Takehito Furuyama, Danielle Vermaak, and Steven Henikoff. Structure, dynamics, and evolution of centromeric nucleosomes. *Proceedings of the National Academy of Sciences*, 104(41):15974–15981, 2007.

[231] Natalia Conde e Silva, Ben E Black, Andrei Sivolob, Jan Filipski, Don W Cleveland, and Ariel Prunell. CENP-A-containing nucleosomes: easier disassembly versus exclusive centromeric localization. *Journal of Molecular Biology*, 370(3):555–573, 2007.

[232] Minh Bui, Emilios K Dimitriadis, Christian Hoischen, Eunkyung An, Delphine Quénet, Sindy Giebe, Aleksandra Nita-Lazar, Stephan Diekmann, and Yamini Dalal. Cell-cycle-dependent structural transitions in the human CENP-A nucleosome in vivo. *Cell*, 150(2):317–326, 2012.

[233] Vassiliki Karantza, Ernesto Freire, and Evangelos N Moudrianakis. Thermodynamic studies of the core histones: pH and ionic strength effects on the stability of the $(H3-H4)/(H3-H4)_2$ system. *Biochemistry*, 35(6):2037–2046, 1996.

[234] Vassiliki Karantza, Ernesto Freire, and Evangelos N Moudrianakis. Thermodynamic studies of the core histones: stability of the octamer subunits is not altered by removal of their terminal domains. *Biochemistry*, 40(43):13114–13123, 2001.

[235] Douglas D Banks and Lisa M Gloss. Folding mechanism of the (H3–H4) 2 histone tetramer of the core nucleosome. *Protein Science*, 13(5):1304–1316, 2004.

[236] Yann-Gaël Gangloff, Christophe Romier, Sylvie Thuault, Sebastiaan Werten, and Irwin Davidson. The histone fold is a key structural motif of transcription factor TFIID. *Trends in Biochemical Sciences*, 26(4):250–257, 2001.

[237] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[238] Lee Whitmore and Bonnie A Wallace. Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers: Original Research on Biomolecules*, 89(5):392–400, 2008.

[239] MA Andrade, P Chacon, JJ Merelo, and F Moran. Evaluation of secondary structure of proteins from UV circular dichroism spectra using an unsupervised learning neural network. *Protein Engineering, Design and Selection*, 6(4):383–390, 1993.

[240] Nicholas D Socci, José Nelson Onuchic, and Peter G Wolynes. Protein folding mechanisms and the multidimensional folding funnel. *Proteins: Structure, Function, and Bioinformatics*, 32(2):136–158, 1998.

[241] David Fushman, Sean Cahill, and David Cowburn. The main-chain dynamics of the dynamin pleckstrin homology (PH) domain in solution: analysis of 15N relaxation with monomer/dimer equilibration 1. *Journal of Molecular Biology*, 266(1):173–194, 1997.

[242] Paul J Flory. *Principles of polymer chemistry*. Cornell University Press, 1953.

[243] Ruxandra I Dima and Devarajan Thirumalai. Asymmetry in the shapes of folded and denatured states of proteins. *The Journal of Physical Chemistry B*, 108(21):6564–6570, 2004.

[244] Richard L Fahrner, Duilio Cascio, James A Lake, and Alexei Slesarev. An ancestral nuclear protein assembly: crystal structure of the Methanopyrus kandleri histone. *Protein Science*, 10(10):2002–2007, 2001.

[245] Jonathan Nye, David Sturgill, Rajbir Athwal, and Yamini Dalal. HJURP antagonizes CENP-A mislocalization driven by the H3. 3 chaperones HIRA and DAXX. *PloS One*, 13(10), 2018.

[246] Yahli Lorch, Barbara Maier-Davis, and Roger D Kornberg. Chromatin remodeling by nucleosome disassembly in vitro. *Proceedings of the National Academy of Sciences*, 103(9):3090–3093, 2006.

[247] Naoko Yoshida, Manjula Brahmajosyula, Shisako Shoji, Manami Amanai, and Anthony CF Perry. Epigenetic discrimination by mouse metaphase II oocytes mediates asymmetric chromatin remodeling independently of meiotic exit. *Developmental Biology*, 301(2):464–477, 2007.

[248] Xiaoling Xie, Tetsuro Kokubo, Steven L Cohen, Urooj A Mirza, Alexander Hoffmann, Brian T Chait, Robert G Roeder, Yoshihiro Nakatani, and

Stephen K Burley. Structural similarity between TAFs and the heterotetrameric core of the histone octamer. *Nature*, 380(6572):316–322, 1996.

[249] Christophe Romier, Fabienne Cocchiarella, Roberto Mantovani, and Dino Moras. The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. *Journal of Biological Chemistry*, 278(2):1336–1345, 2003.

[250] Christopher K Materese. *Atomistic simulations uncover microscopic details of nucleosomal electrostatics, energy landscapes of proteins and photovoltaic polymer dynamics*. PhD thesis, University of North Carolina at Chapel Hill, 2011.

[251] Davit Potoyan. *Using energy landscape theory to uncover the organization of conformational space of proteins in their native states*. PhD thesis, University of Maryland, College Park, 2011.

[252] David Winogradoff. *Molecular dynamic simulations of nucleosomes and histone tails: The effects of histone variance and post-translational modification*. PhD thesis, University of Maryland, College Park, 2015.

[253] Haiqing Zhao. *Uncovering the biophysical mechanisms of histone complex assembly*. PhD thesis, University of Maryland, College Park, 2018.

[254] Markus R Hermann and Jochen S Hub. SAXS-restrained ensemble simulations of intrinsically disordered proteins with commitment to the principle of maximum entropy. *Journal of Chemical Theory and Computation*, 15(9):5103–5115, 2019.

[255] Andrew P Latham and Bin Zhang. Maximum entropy optimized force field for intrinsically disordered proteins. *Journal of Chemical Theory and Computation*, 2019.

[256] Upayan Baul, Debayan Chakraborty, Mauro L Mugnai, John E Straub, and Dave Thirumalai. Sequence effects on size, shape, and structural heterogeneity in intrinsically disordered proteins. *The Journal of Physical Chemistry B*, 123(16):3462–3474, 2019.

[257] Julien Roche and Davit A Potoyan. Disorder mediated oligomerization of DISC1 proteins revealed by coarse-grained molecular dynamics simulations. *The Journal of Physical Chemistry B*, 123(45):9567–9575, 2019.

[258] Xingcheng Lin, Prakash Kulkarni, Federico Bocci, Nicholas P Schafer, Susmita Roy, Min-Yeh Tsai, Yanan He, Yihong Chen, Krithika Rajagopalan, Steven M Mooney, Yu Zeng, Keith Weninger, Alex Grishaev, José N Onuchic, Herbert Levine, Peter G Wolynes, Ravi Salgia, Govindan Rangarajan, Vladimir Uversky, John Orban, and Mohit Kumar Jolly. Structural and dynamical order of a disordered protein: Molecular insights into conformational switching of page4 at the systems level. *Biomolecules*, 9(2):77, 2019.

[259] Vera Alverdi, Byron Hetrick, Simpson Joseph, and Elizabeth A Komives. Direct observation of a transient ternary complex during IκBα-mediated dissociation of NF-κB from DNA. *Proceedings of the National Academy of Sciences*, 111(1):225–230, 2014.

[260] Lan N Truong, Yongjiang Li, Linda Z Shi, Patty Yi-Hwa Hwang, Jing He, Hailong Wang, Niema Razavian, Michael W Berns, and Xiaohua Wu. Microhomology-mediated end joining and homologous recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. *Proceedings of the National Academy of Sciences*, 110(19):7720–7725, 2013.

[261] Dana Krepel, Aram Davtyan, Nicholas P Schafer, Peter G Wolynes, and José N Onuchic. Braiding topology and the energy landscape of chromosome organization proteins. *Proceedings of the National Academy of Sciences*, 117(3):1468–1477, 2020.

[262] Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, Rafal P Wiewiora, Bernard R Brooks, and Vijay S Pande. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology*, 13(7):e1005659, 2017.

[263] Daiki Kato, Akihisa Osakabe, Yasuhiro Arimura, Yuka Mizukami, Naoki Horikoshi, Kazumi Saikusa, Satoko Akashi, Yoshifumi Nishimura, Sam-Yong Park, Jumpei Nogami, Kazumitsu Maehara, Yasuyuki Ohkawa, Atsushi Matsumoto, Hidetoshi Kono, Rintaro Inoue, Masaaki Sugiyama, and Hitoshi Kurumizaka. Crystal structure of the overlapping dinucleosome composed of hexasome and octasome. *Science*, 356(6334):205–208, 2017.

[264] Thomas Schalch, Sylwia Duda, David F Sargent, and Timothy J Richmond. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature*, 436(7047):138, 2005.

[265] Benedetta Dorigo, Thomas Schalch, Kerstin Bystricky, and Timothy J Richmond. Chromatin fiber folding: requirement for the histone H4 N-terminal tail. *Journal of Molecular Biology*, 327(1):85–96, 2003.

[266] Bryan J Wilkins, Nils A Rall, Yogesh Ostwal, Tom Kruitwagen, Kyoko Hiragami-Hamada, Marco Winkler, Yves Barral, Wolfgang Fischle, and Heinz Neumann. A cascade of histone modifications induces chromatin condensation in mitosis. *Science*, 343(6166):77–80, 2014.

[267] Howard Cedar and Yehudit Bergman. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295, 2009.

[268] Tongye Shen, Chenghang Zong, Donald Hamelberg, J Andrew McCammon, and Peter G Wolynes. The folding energy landscape and phosphorylation: modeling the conformational switch of the NFAT regulatory domain. *The FASEB Journal*, 19(11):1389–1395, 2005.

[269] Joachim Lätzer, Tongye Shen, and Peter G Wolynes. Conformational switching upon phosphorylation: a predictive framework based on energy landscape principles. *Biochemistry*, 47(7):2110–2122, 2008.

[270] Jonathan Widom. Role of DNA sequence in nucleosome stability and dynamics. *Quarterly Reviews of Biophysics*, 34(3):269–324, 2001.

[271] Daria A Beshnova, Andrey G Cherstvy, Yevhen Vainshtein, and Vladimir B Teif. Regulation of the nucleosome repeat length in vivo by the DNA sequence, protein concentrations and long-range interactions. *PLoS Computational Biology*, 10(7):e1003698, 2014.

[272] Jamie Culkin, Lennart De Bruin, Marco Tompitak, Rob Phillips, and Helmut Schiessel. The role of DNA sequence in nucleosome breathing. *The European Physical Journal E*, 40(11):106, 2017.

[273] Alexey Savelyev and Garegin A Papoian. Molecular renormalization group coarse-graining of polymer chains: application to double-stranded DNA. *Biophysical Journal*, 96(10):4044–4052, 2009.

[274] Stefano Piana, Alexander G Donchev, Paul Robustelli, and David E Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *The Journal of Physical Chemistry B*, 119(16):5113–5123, 2015.

[275] Alexandra Patriksson and David van der Spoel. A temperature predictor for parallel tempering simulations. *Physical Chemistry Chemical Physics*, 10(15):2073–2077, 2008.

[276] Herman JC Berendsen, David van der Spoel, and Rudi van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56, 1995.

[277] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[278] Jeffrey Skolnick, Andrzej Kolinski, and Angel R Ortiz. MONSSTER: A method for folding globular proteins with a small number of distance restraints. *Journal of Molecular Biology*, 265(2):217–241, 1997.

[279] Feng Ding, Ramesh K Jha, and Nikolay V Dokholyan. Scaling behavior and structure of denatured proteins. *Structure*, 13(7):1047–1054, 2005.

[280] Fábio Madeira, Joon Lee, Nicola Buso, Tamer Gur, Nandana Madhusoodanan, Prasad Basutkar, Adrian Tivey, Simon C Potter, Robert D Finn, Rodrigo Lopez, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, 47(15):W636–W641, 2019.